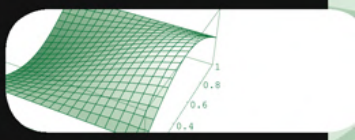
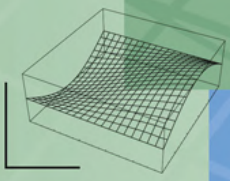
BIRKHÄUSER

Integral Methods in Science and Engineering



Volume 2 COMPUTATIONAL METHODS

C. Constanda, M.E. Pérez, EDITORS



BIRKHÄUSER

Integral Methods in Science and Engineering
Volume 2

Computational Methods

C. Constanda
M.E. Pérez
Editors

Birkhäuser
Boston • Basel • Berlin

Editors
C. Constanda
Department of Mathematical
and Computer Sciences
University of Tulsa
800 South Tucker Drive
Tulsa, OK 74104
USA
christian-constanda@utulsa.edu

M.E. Pérez
Departamento de Matemática Aplicada
y Ciencias de la Computación
Universidad de Cantabria
Avenida de los Castros s/n
39005 Santander
Spain
meperez@unican.es

ISBN 978-0-8176-4896-1 e-ISBN 978-0-8176-4897-8
DOI 10.1007/978-0-8176-4897-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009939427

Mathematics Subject Classification (2000): 34-06, 35-06, 40-06, 40C10, 45-06, 65-06, 74-06, 76-06

Cover design: Joseph Sherman

Printed on acid-free paper

Birkhäuser Boston is part of Springer Science+Business Media (www.birkhauser.com)

Contents

Preface

The international conferences on Integral Methods in Science and Engineering (IMSE) are biennial opportunities for academics and other researchers whose work makes essential use of analytic or numerical integration methods to discuss their latest results and exchange views on the development of novel techniques of this type.

The first two conferences in the series, IMSE1985 and IMSE1990, were hosted by the University of Texas–Arlington. At the latter, the IMSE consortium was created and charged with organizing these conferences under the guidance of an International Steering Committee. Subsequently, IMSE1993 took place at Tohoku University, Sendai, Japan, IMSE1996 at the University of Oulu, Finland, IMSE1998 at Michigan Technological University, Houghton, MI, USA, IMSE2000 in Banff, AB, Canada, IMSE2002 at the University of Saint-Étienne, France, IMSE2004 at the University of Central Florida, Orlando, FL, USA, and IMSE2006 at Niagara Falls, ON, Canada. The IMSE conferences are now recognized as an important forum where scientists and engineers working with integral methods express their views about, and interact to extend the practical applicability of, a very elegant and powerful class of mathematical procedures.

A distinguishing characteristic of all the IMSE meetings is their general atmosphere—a blend of utmost professionalism and a strong collegial-social component. IMSE2008, organized at the University of Cantabria, Spain, and attended by delegates from 27 countries on 5 continents, maintained this tradition, marking another unqualified success in the history of the IMSE consortium. For the smoothness and detail-perfect arrangements throughout the conference, the participants and the Steering Committee would like to express their special thanks to the Local Organizing Committee:

M. Eugenia Pérez (Departamento de Matemática Aplicada y Ciencias de la Computación, ETSI Caminos, Canales y Puertos), *Chairman;*

Miguel Lobo (Departamento de Matemáticas, Estadística y Computación, Facultad de Ciencias);

Delfina Gómez (Departamento de Matemáticas, Estadística y Computación, Facultad de Ciencias).

The Local Organizing Committee and the Steering Committee also wish to acknowledge the financial support received from the following institutions:

Universidad de Cantabria (in particular, Vicerrectorado de Investigación y Transferencia del Conocimiento, Facultad de Ciencias, ETSI Caminos, Canales y Puertos, Departamento de Matemáticas, Estadística y Computación, and Departamento de Matemática Aplicada y Ciencias de la Computación);

Ministerio de Ciencia e Innovación (Ref. MTM2007-30182-E);

Sociedad Regional Cantabra de I+D+i (IDICAN. Ref. 25-2-2007);

i-MATH Consolider (MEC, Ref. C3-0087);

Caja de Burgos;

Consejería de Cultura, Turismo y Deporte del Gobierno de Cantabria;

Ayuntamiento de Santander;

Sociedad Española de Matemática Aplicada (SeMA).

Last but not least, they would like to express their thanks to MICINN (MTM2005-07720) for partial support, to Antonio José González for his work on the graphical design of the conference, to the colleagues involved in the coordination of the monographic sessions, and to all the participants, whose presence and scientific activity in Santander ensured the success of this meeting.

The next IMSE conference will be held in July 2010 in Brighton, UK. Details concerning this event are posted on the conference web page,

http://www.cmis.brighton.ac.uk/imse2010

This volume contains 2 invited papers and 32 contributed peer-reviewed papers, arranged in alphabetical order by (first) author's name. The editors would like to thank the staff at Birkhäuser-Boston for their efficient handling of the publication process.

Tulsa, Oklahoma, USA *Christian Constanda, IMSE Chairman*

The International Steering Committee of IMSE:

C. Constanda (University of Tulsa), *Chairman*
M. Ahues (University of Saint-Étienne)
B. Bodmann (Federal University of Rio Grande do Sul)
I. Chudinovich (University of Tulsa)
H. de Campos Velho (INPE, Saõ José dos Campos)

P. Harris (University of Brighton)
A. Largillier (University of Saint-Étienne)
S. Mikhailov (Brunel University)
A. Mioduchowski (University of Alberta)
D. Mitrea (University of Missouri-Columbia)
Z. Nashed (University of Central Florida)
A. Nastase (Rhein.-Westf. Technische Hochschule, Aachen)
M.E. Pérez (University of Cantabria)
S. Potapenko (University of Waterloo)
K. Ruotsalainen (University of Oulu)
S. Seikkala (University of Oulu)
O. Shoham (University of Tulsa)

List of Contributors

Mario Ahues
Université de Saint-Étienne
23 rue du Dr. Paul Michelon
Saint-Étienne 42023, cedex 2, France
mario.ahues@univ-st-etienne.fr

Juan A. Alcober
Universidad Politécnica de Valencia
Camino de Vera s/n
Valencia 46022, Spain
juaalau@mat.upv.es

Filomena D. d'Almeida
Universidade do Porto
Rua Roberto Frias
Porto 4200-465, Portugal
falmeida@fe.up.pt

Lino J. Alvarez-Vázquez
Universidad de Vigo
ETSI Telecomunicación
Vigo 36310, Spain
lino@dma.uvigo.es

Ricardo C. de Barros
Universidade do Estado do Rio de Janeiro
Rua Alberto Rangel s.n.
Nova Friburgo, RJ 28630-050, Brazil
rcbarros@pq.cnpq.br

Bardo E.J. Bodmann
Universidade Federal do Rio Grande do Sul
Av. Osvaldo Aranha 99/4
Porto Alegre, RS 90035-190, Brazil
bardo.bodmann@ufrgs.br

Daniela Buske
Universidade Federal de Pelotas
Campus Capão do Leão
Caixa Postal 354
Pelotas, RS 96010-900, Brazil
danielabuske@gmail.com

Haroldo F. de Campos Velho
Instituto Nacional de Pesquisas Espaciais
P.O. Box 515
São José dos Campos, SP 12245-970, Brazil
haroldo@lac.inpe.br

Juan Cardona
Universidad de Cantabria
Dique de Gamazo, 1
Santander 39004, Spain
cardonaj@unican.es

Juan Casado-Díaz
Universidad de Sevilla
C/ Tarfia s/n
Sevilla 41012, Spain
jcasadod@us.es

Carlos Castro
Universidad Politécnica de Madrid
Professor Aranguren
Madrid 28040, Spain
carlos.castro@upm.es

Manuel J. Castro Díaz
Universidad de Málaga
Campus de Teatinos
Málaga 29071, Spain
castro@anamat.cie.uma.es

David J. Chappell
University of Nottingham
University Park
Nottingham NG7 2RD, UK
david.chappell@nottingham.ac.uk

Andrea S. Charão
Universidade Federal de Santa Maria
Avenida Roraima, 1000
Santa Maria, RS 97105-900, Brazil
andrea@inf.ufsm.br

Subhashini Chitta
Flow Analysis Inc.
411 B.H. Goethert Parkway
Tullahoma, TN 37388, USA
subha@flowanalysis.com

Igor Chudinovich
University of Tulsa
800 S. Tucker Drive,
Tulsa, OK 74104, USA
igor-chudinovych@utulsa.edu

Christian Constanda
University of Tulsa
800 S. Tucker Drive,
Tulsa, OK 74104, USA
christian-constanda@utulsa.edu

Alicia Cordero
Universidad Politécnica de Valencia
Camino de Vera s/n
Valencia 46022, Spain
acordero@mat.upv.es

Aldina I.A. Correia
Instituto Politécnico do Porto
and Universidade de Trás-os-Montes
e Alto Douro
Rua do Curral, Casa do Curral
Margaride, 4610-156 Felgueiras,
Portugal
aldinacorreia@eu.ipp.pt

Camila Pinto da Costa
Universidade Federal de Pelotas
Campus Capão do Leão
Caixa Postal 354
Pelotas, RS 96010-900, Brazil
camiladacosta@gmail.com

Julio Couce-Calvo
Universidad de Sevilla
C/ Tarfia s/n
Sevilla 41012, Spain
couce@us.es

Teresa Diogo
Instituto Superior Técnico
Av. Rovisco Pais, 1
Lisbon 1049-001, Portugal
tdiogo@math.ist.utl.pt

Dale R. Doty
University of Tulsa
800 S. Tucker Drive
Tulsa, OK 74104, USA
dale-doty@utulsa.edu

Rosário Fernandes
Universidade do Minho
Campus de Gualtar
Braga 4710-057, Portugal
rosario@math.uminho.pt

Francisco J. Fernández
Universidad de Santiago de
Compostela
Campus Universitario Sur
Santiago de Compostela 15782 Spain
franfdz@usc.es

Enrique D. Fernández-Nieto
Universidad de Sevilla
Avda. Reina Mercedes, 2
Sevilla 41012, Spain
edofer@us.es

Donald A. French
University of Cincinnati
2855 Campus Way
Cincinnati, OH 45221-0025, USA
french@math.uc.edu

José M. González-Vida
Universidad de Málaga
Campus de Teatinos
Málaga 29071, Spain
jgv@uma.es

Evgeniy Gorodnitskiy
St. Petersburg University
Ulyanovskaya 1-1, Petrodvorets
St. Petersburg 198904, Russia
eugy@yandex.ru

Laurence Grammont
Université de Saint-Étienne
23, rue du Dr. Paul Michelon
Saint-Étienne 42023, cedex 2, France
laurence.grammont
@univ-st-etienne.fr

Charles W. Groetsch
The Citadel
171 Moultrie St.
Charleston, SC 29409-6420, USA
charles.groetsch@citadel.edu

Andrey V. Gusarov
École Nationale d'Ingénieurs de Saint-Étienne
58 rue Jean Parot
Saint-Étienne 42023, France
gusarov@enise.fr

Carl Hardwidge
Princess Royal Hospital
Lewes Road
Haywards Heath RH16 4EX, UK
carl.hardwidge@bsuh.nhs.uk

Paul J. Harris
University of Brighton
Lewes Road
Brighton BN2 4GJ, UK
p.j.harris@brighton.ac.uk

Ismael R. Heinen
Universidade Federal do Rio Grande do Sul
Av. Osvaldo Aranha 99/4
Porto Alegre, RS 90046-900, Brazil
heire@bol.com.br

Carlos F. Jerez-Hanckes
ETH Zürich
Rämistrasse 101
Zürich 8092, Switzerland
cjerez@math.ethz.ch

Jukka Kemppainen
University of Oulu
PO Box 4500
Oulu 90014, Finland
jukemppa@paju.oulu.fi

Marek Kolk
University of Tartu
Liivi 2
Tartu 50409, Estonia
marek.kolk@ut.ee

Alexander F. Koshchii
International Solomon University
Grazhdanskaya 22/26
Kharkiv 61057, Ukraine
af-koshchii@rambler.ru

Vesa Kotila
University of Oulu
PO Box 4500
Oulu 90014, Finland
vesa.kotila@oulu.fi

Alexander V. Kudrin
University of Nizhny Novgorod
23 Gagarin Avenue
Nizhny Novgorod 603950, Russia
kud@rf.unn.ru

Rekha P. Kulkarni
Indian Institute of Technology Bombay
Powai
Mumbai 400076, India
rpk@math.iitb.ac.in

Sulo Lahdelma
University of Oulu
PO Box 4200
Oulu 90014, Finland
sulo.lahdelma@oulu.fi

Pedro Lima
Instituto Superior Técnico
Av. Rovisco Pais, 1
Lisbon 1049-001, Portugal
plima@math.ist.utl.pt

Carlo Lovadina
Università di Pavia
Via Ferrata 1
Pavia 27100, Italy
carlo.lovadina@unipv.it

Manuel Luna-Laynez
Universidad de Sevilla
C/ Tarfia s/n
Sevilla 41012, Spain
mllaynez@us.es

Anne Mangeney
Institut de Physique du Globe de Paris
4, place Jussieu
Paris 75252, cedex 05, France
mangeney@ipgp.jussieu.fr

Osni Marques
Lawrence Berkeley National Laboratory
1 Cyclotron Road, MS 50F-1650
Berkeley, CA 94720-8139, USA
oamarques@lbl.gov

J.D. Martín-Gómez
Universidad de Sevilla
c/Tartia s/n
Sevilla 41011, Spain
jdmartin@us.es

João L.H. Matias
Universidade de Trás-os-Montes e Alto Douro
Edifício das Ciências Florestais
Quinta de Prados
5001-801 Vila Real, Portugal
j_matias@utad.pt

Pedro M.M.A. Mestre
Universidade de Trás-os-Montes e Alto Douro
Edifício Engenharias II
5001-801 Vila Real, Portugal
pmestre@utad.pt

Davidson M. Moreira
Universidade Federal do Pampa
Rua Carlos Barbosa s/n
B. Getúlio Vargas
Bagé, RS 96412-420, Brazil
davidson@pq.cnpq.br

Rafael Muñoz-Sola
Universidad de Santiago de Compostela
Campus Universitario Sur
Santiago de Compostela 15782
Spain
rafams@usc.es

Adriana Nastase
Rhein.-Westf. Technische Hochschule
Templergraben 55
Aachen 52062, Germany
nastase@lafaero.rwth-aachen.de

Phillippe O.A. Navaux
Universidade Federal do Rio Grande do Sul
Av. Bento Gonąlves, 9500
Porto Alegre 91501-970, Brazil
navaux@inf.ufrgs.br

Jean-Claude Nédélec
École Polytechnique
Route de Saclay
Palaiseau 91128, France
nedelec@cmapx.polytechnique.fr

Fabiana F. Paes
Instituto Nacional de Pesquisas Espaciais
P.O. Box 515
São José dos Campos, SP 12245-970
Brazil
fabiana.paes@lac.inpe.br

Francisco Palacios
Universidad Politécnica de Madrid
Avda Complutense
Madrid 28040, Spain
fpalacios@gmail.com

Carlos Parés Madroñal
Universidad de Málaga
Campus de Teatinos
Málaga 29071, Spain
pares@anamat.cie.uma.es

Arvet Pedas
University of Tartu
Liivi 2
Tartu 50409, Estonia
arvet.pedas@ut.ee

Maria Perel
St. Petersburg University
Ulyanovskaya 1-1, Petrodvorets
St. Petersburg 198904, Russia
and Ioffe Physical-Technical Institute
Politekhnicheskaya 26
St. Petersburg 194021, Russia
perel@mph.phys.spbu.ru

Cláudio Z. Petersen
Universidade Federal do Rio Grande do Sul
Sarmento Leite, 425/3
Porto Alegre, RS 90046-900, Brazil
claudiopetersen@yahoo.com.br

Evgeny Yu. Petrov
University of Nizhny Novgorod
23 Gagarin Avenue
Nizhny Novgorod 603950, Russia
epetrov@rf.unn.ru

Bárbara D.A. Rodriguez
Universidade Federal do Rio Grande
Avenida Itália km 8
Campus Carreiros
Rio Grande, RS 96201-900, Brazil
barbara.rodriguez@gmail.com

José E. Roman
Universidad Politécnica de Valencia
Camino de Vera s/n
Valencia 46022, Spain
jroman@dsic.upv.es

Keijo Ruotsalainen
University of Oulu
PO Box 4500
Oulu 90014, Finland
keijo.ruotsalainen@ee.oulu.fi

Avelino Samartin
Universidad Politécnica de Madrid
Ciudad Universitaria s/n
Madrid 28040, Spain
avelino.samartin@upm.es

Carlos M.J.A. Serôdio
Universidade de Trás-os-Montes e Alto Douro
Edifício Engenharias II
5001-801 Vila Real, Portugal
cserodio@utad.pt

Mikhail Sidorenko
St. Petersburg University
Ulyanovskaya 1-1, Petrodvorets
198904 St. Petersburg, Russia
m-sidorenko@yandex.ru

László Simon
L. Eötvös University
Pázmány P. sétány 1/C
Budapest 1117, Hungary
simonl@ludens.elte.hu

Igor Smurov
École Nationale d'Ingénieurs
58 rue Jean Parot
Saint-Étienne 42023, France
smurov@enise.fr

Roberto P. Souto
Instituto Nacional de
Pesquisas Espaciais
São José dos Campos, SP 12245-970
Brazil
rpsouto@gmail.com

John Steinhoff
University of Tennessee
411 B.H. Goethert Parkway
Tullahoma, TN 37388, USA
jsteinho@utsi.edu

Stephan Stephany
Instituto Nacional de
Pesquisas Espaciais
São José dos Campos, SP 12245-970
Brazil
stephan@lac.inpe.br

Pedro Tabuenca
Universidad de Cantabria
Dique de Gamazo, 1
Santander 39004, Spain
tabuencp@orange.es

Tiziano Tirabassi
Istituto di Scienze dell'Atmosfera e
del Clima–CNR
Via P. Gobetti 101
Bologna 40129, Italy
t.tirabassi@isac.cnr.it

Igor M. Tkachenko
Universidad Politécnica de Valencia
Camino de Vera s/n
Valencia 46022, Spain
imtk@mat.upv.es

Juan R. Torregrosa
Universidad Politécnica de Valencia
Camino de Vera s/n
Valencia 46022, Spain
jrtorre@mat.upv.es

Marcel Urrea
Universidad Politécnica de Valencia
Camino de Vera s/n
Valencia 46022, Spain
murrea@mat.upv.es

Paulo B. Vasconcelos
Universidade do Porto
Rua Dr. Roberto Frias
Porto 4200-464, Portugal
pjv@fep.up.pt

Marco T.M.B. de Vilhena
Universidade Federal do Rio Grande
do Sul
Rua Sarmento Leite, 425/3
Porto Alegre, RS 90046-900
Brazil
vilhena@pq.cnpq.br

Juliana K. Vizzotto
Centro Universitário Franciscano
Rua dos Andradas, 1614
Santa Maria, RS 97010-032, Brazil
juvizzotto@gmail.com

Tatyana M. Zaboronkova
Technical University of Nizhny Novgorod
24 Minin Street
Nizhny Novgorod 603950,
Russia
zabr@nirfi.sci-nnov.ru

Enrique Zuazua
Basque Center for Applied Mathematics
Bizkaia Technology Park
Zamudio (Bilbao) 48170,
Basque Country, Spain
zuazua@bcamath.org

1

Error Bounds for L^1 Galerkin Approximations of Weakly Singular Integral Operators

M. Ahues,[1] F.D. d'Almeida,[2] and R. Fernandes[3]

[1] Université de Lyon, Laboratoire de Mathématiques de l'Université de Saint-Étienne, France; mario.ahues@univ-st-etienne.fr
[2] Universidade do Porto, Portugal; falmeida@fe.up.pt
[3] Universidade do Minho, Portugal; rosario@math.uminho.pt

1.1 Introduction

From all standard projection approximations of a bounded linear operator in a Banach space, a general (i.e., not necessarily orthogonal) Galerkin scheme ([At97] and [ALL01]) is the simplest one from a computational point of view. In this chapter, we give an upper bound of the relative error in terms of the mesh size of the underlying discretization grid on which no regularity assumptions are made. A weakly singular second kind Fredholm integral equation is used as an application to illustrate the actual sharpness of the error estimates. As is usual in the case of weakly singular error bounds, the sharpness of our bound is rather poor compared with practical results.

We consider the space $L^1([0,\tau^*],\mathbb{C})$ of complex-valued Lebesgue-integrable (classes of) functions on $[0,\tau^*]$. For $x \in L^1([0,\tau^*],\mathbb{C})$, define

$$(Tx)(s) := \int_0^{\tau^*} g(|s-t|)x(t)\,dt, \quad s \in [0,\tau^*], \tag{1.1}$$

where $g :]0,\infty[\to \mathbb{R}$ is a weakly singular function at 0 in the following sense:

$$g(0^+) = \infty,\ g \in L^1([0,\infty[,\mathbb{R}) \cap C^0(]0,\infty[,\mathbb{R}),\ g \geq 0,\ g \searrow \text{ in }]0,\infty[. \tag{1.2}$$

It can be checked that $Tx \in L^1([0,\tau^*],\mathbb{C})$, and that T is compact as an operator on $L^1([0,\tau^*],\mathbb{C})$ (see [ALL01]). Let $z \in \mathrm{re}(T)$, the resolvent set of T, so $T - zI$ is bijective and has a bounded inverse. Since T is compact, $z \neq 0$. This implies that, for any $f \in L^1([0,\tau^*],\mathbb{C})$, the Fredholm integral equation of the second kind

$$(T - zI)\xi \quad = \quad f \tag{1.3}$$

has a unique solution $\xi \in L^1([0,\tau^*],\mathbb{C})$.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_1,

The resolvent operator $R(z) := (T - zI)^{-1}$ allows us to write this solution as $\xi = R(z)f$.

Concerning the derivative of Tx we have the following theorem proved in [AAF09]:

Theorem 1. *For any $x \in L^1([0,\tau^*],\mathbb{C})$ such that $x' \in L^1([0,\tau^*],\mathbb{C})$, Tx is a differentiable function at all $s \in]0,\tau^*[$, and its derivative is given by*

$$(Tx)'(s) = x(0)g(s) - x(\tau^*)g(\tau^* - s) + (Tx')(s), \quad s \in]0,\tau^*[.$$

Since the solution ξ of (1.3) satisfies $\xi = \dfrac{1}{z}(T\xi - f)$, we may expect boundary layers for ξ at the end points and at points where f has a discontinuity. Boundary layers lead us to decompose the interval $[0,\tau^*]$ into subdomains. Those including the boundary layers will be discretized with finer grids than the ones used elsewhere.

1.2 Numerical Approximations

Let us consider the operator (1.1) in an arbitrary interval $[a,b]$ and let the underlying complex Banach space be $X := L^1([a,b],\mathbb{C})$.

$$(Tx)(s) := \int_a^b g(|s-t|)x(t)\,dt, \quad s \in [a,b], \quad x \in X,$$

where $g :]0,+\infty[\to \mathbb{R}$ satisfies (1.2). We describe the general Galerkin scheme. To compute a numerical solution φ_n of the exact solution φ of the equation

$$(T - zI)\varphi = f \tag{1.4}$$

we use a sequence of bounded projections $(\pi_n)_{n\geq 1}$ each one having finite rank, and the corresponding sequence of operators $(T_n)_{n\geq 1}$ given by $T_n := \pi_n T \pi_n$ and we assume that $\mathrm{re}(T) \subseteq \mathrm{re}(T_n)$. We replace the exact equation (1.4) with the approximate problem of solving exactly the following equation for φ_n:

$$(T_n - zI)\varphi_n \quad = \quad \pi_n f. \tag{1.5}$$

The approximate resolvent $R_n(z) := (T_n - zI)^{-1}$ allows us to write the unique solution of the approximate equation as $\varphi_n := R_n(z)\pi_n f$. The second resolvent identities,

$$R_n(z) - R(z) = R_n(z)(T - T_n)R(z) = R(z)(T - T_n)R_n(z),$$

will be useful in the sequel.

Proposition 1 (See [ALL01], [Ch83]). *If $(\pi_n)_{n\geq 1}$ is pointwise convergent to I, then there exists n_0 such that*

$$\beta := |z| \sup_{n\geq n_0} \|(\pi_n T - zI)^{-1}\|$$

is finite, and there exists a constant $\alpha > 0$ such that, for large enough n,

$$\alpha\|(I-\pi_n)\varphi\| \leq \|\varphi_n - \varphi\| \leq \beta\|(I-\pi_n)\varphi\|.$$

Theorem 2. *For $f \neq 0$, the solution of the Galerkin approximation satisfies*

$$\frac{\|\varphi_n - \varphi\|}{\|\varphi\|} \leq C(\|(I-\pi_n)T\| + \frac{\|(I-\pi_n)f\|}{\|\varphi\|}), \tag{1.6}$$

for n large enough and any $C \geq \sup_n \|(\pi_n T - zI)^{-1}\|$.

The proof can be found in [AAF09].

Let us consider a general grid $\mathcal{G}_n := (\tau_j)_{j=0}^n$ such that

$$\tau_0 := a,\ \tau_n := b,\ h_j := \tau_j - \tau_{j-1} > 0,\ h_{\max} := \max_{1\leq j\leq n} h_j,\ h_{\min} := \min_{1\leq j\leq n} h_j.$$

We associate to this grid the local mean functionals e_j^* defined by

$$\langle x\,,\, e_j^*\rangle := \frac{1}{h_j}\int_{\tau_{j-1}}^{\tau_j} x(t)\,dt,$$

and the piecewise constant canonical functions e_j given by

$$e_j(s) : \begin{cases} 1 & \text{for } s \in \,]\tau_{j-1}, \tau_j], \\ 0 & \text{otherwise.} \end{cases}$$

Since the families $(e_j)_{j=1}^n$ and $\left(e_j^*\right)_{j=1}^n$ are adjoint to one another, these linearly independent families lead to a sequence of projections with finite rank n:

$$\pi_n x := \sum_{j=1}^n \langle x\,,\, e_j^*\rangle e_j \quad \text{for } x \in X.$$

Recall that the oscillation of $x \in X$ is given by

$$\omega_1(x,\delta) := \sup_{0\leq h\leq\delta} \int_a^{b-h} |x(s+h) - x(s)|\,ds.$$

Theorem 3. *For all $x \in X$, $\|(I-\pi_n)x\| \leq 2\sum_{j=1}^n \omega_1(x_{|[\tau_{j-1},\tau_j]}, h_j)$.*

A proof of this estimate can be found in [AAF09], and a proof of a similar bound can be found in [AAL09].

Theorem 3 with $x = f$ gives a bound on one part of (1.6):

$$\|(I - \pi_n)f\| \leq 2 \sum_{j=1}^{n} \omega_1(f_{|[\tau_{j-1},\tau_j]}, h_j). \tag{1.7}$$

The following theorem establishes a bound on the other part of (1.6).

Theorem 4. *If g satisfies (1.2) then*

$$\begin{aligned}\|(I - \pi_n)T\| \leq 2h_{\max}(g(h_{\min}/2) + g(h_{\min}) - 2g(b-a)) \\ + 4\int_0^{h_{\max}/2} g(\sigma)d\sigma + 4\int_0^{h_{\max}} g(\sigma)d\sigma + 4\int_0^{3h_{\max}/2} g(\sigma)d\sigma \end{aligned} \tag{1.8}$$

in the subordinated operator norm.

Proof. If we write the bound of Theorem 3 with the definition of ω_1 and perform the change of variable $\tau = \alpha h_j$, $d\tau = h_j d\alpha$, we get

$$\|(I - \pi_n)x\| \leq 2\sum_{j=1}^{n} \int_0^1 \int_{\tau_{j-1}}^{\tau_j - \alpha h_j} |x(s + \alpha h_j) - x(s)|\, ds\, d\alpha.$$

Replacing x with Tx, for all $x \in L^1([a,b], \mathbb{C})$, and changing the order of the integrals, we can easily prove that

$$\begin{aligned}\|(I - \pi_n)Tx\| &\leq 2\int_a^b \sum_{j=1}^{n} \int_0^1 \int_{\tau_{j-1}-t}^{\tau_j - \alpha h_j - t} |g(|\tau + \alpha h_j|) - g(|\tau|)| d\tau d\alpha\, |x(t)|\, dt \\ &\leq 2\, \|x\| \sup_{t\in[a,b]} \int_0^1 \sum_{j=1}^{n} \int_{\tau_{j-1}-t}^{\tau_j - \alpha h_j - t} |g(|\tau + \alpha h_j|) - g(|\tau|)| d\tau d\alpha.\end{aligned}$$

Let

$$A_j(t) \quad := \quad \int_0^1 \int_{t_{j-1}}^{t_j - \alpha h_j} |g(|\tau + \alpha h_j|) - g(|\tau|)| d\tau d\alpha$$

and $t_j := \tau_j - t$, $t \in [a,b]$. We estimate an upper bound of $\sup\limits_{t\in[a,b]} \sum\limits_{j=1}^{n} A_j(t)$.

This proof is based on the geometry of the underlying discretization grid and it includes the dependence on the possible subdomains $[a,b]$ of the interval $[0, \tau^\star]$.

Any t in $[a,b]$ belongs to a certain subinterval of the grid, say $[\tau_{k-1}, \tau_k]$ and it may be located in the second half of it—case (A), or in the first—case (B).

(A) In this case $t_{k-1} \leq -h_k/2$ (see Figure 1.1) and we have four subcases for j:

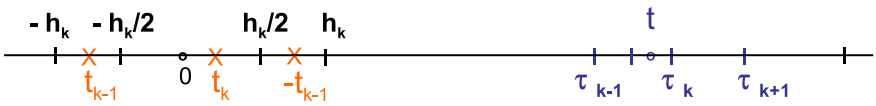


Fig. 1.1. Location of t, t_k,$-t_{k-1}$ and t_{k-1}.

(A1) For all j such that $\tau_j \leq \tau_{k-1}$, that is, $j \leq k-1$ (see Figure 1.1), $t_j, t_{j-1}, t_j - \alpha h_j$, and τ are negative. So $-t_j < -t_j + \alpha h_j < -t_{j-1}$. As τ is negative, $|\tau + \alpha h_j| < |\tau|$ and since g is a decreasing function,

$$|g(|\tau + \alpha h_j|) - g(|\tau|)| = g(|\tau + \alpha h_j|) - g(|\tau|) = g(-\tau - \alpha h_j) - g(-\tau).$$

Replacing in the first term $g(-\tau - \alpha h_j)$ with a larger value, $g(-t_j + \alpha h_j - \alpha h_j)$, and in the second term $g(-\tau)$ with a smaller value, $g(-t_{j-1})$, we have a larger value for the integral:

$$A_j(t) \leq \int_0^1 \int_{t_{j-1}}^{t_j - \alpha h_j} (g(-t_j) - g(-t_{j-1})) d\tau d\alpha,$$

and enlarging the interval for τ to $[t_{j-1}, t_j]$, we have

$$A_j(t) \leq h_j(g(-t_j) - g(-t_{j-1})) \leq h_{\max}(g(-t_j) - g(-t_{j-1})).$$

Hence,

$$\sum_{j=1}^{k-1} A_j(t) \leq h_{\max}(g(-t_{k-1}) - g(-t_0)) \leq h_{\max}(g(h_k/2) - g(b-a)),$$

because $t_{k-1} \leq -h_k/2$ implies that $-t_{k-1} \geq h_k/2$ and g is a decreasing function.

(A2) For all j such that $\tau_{j-1} > \tau_k$, i.e., $j > k+1$, t_j and t_{j-1} are positive. Also, $t_j - \alpha h_j \geq \tau_{j-1}$ is positive, τ is positive, and so is $\tau + \alpha h_j$. As g is a decreasing function,

$$|g(|\tau + \alpha h_j|) - g(|\tau|)| = g(|\tau|) - g(|\tau + \alpha h_j|).$$

Using the same arguments as in the previous case, we get

$$\sum_{j=k+2}^{n} A_j(t) \leq h_{\max}(g(t_{k+1}) - g(t_n)) \leq h_{\max}(g(h_{\min}) - g(b-a)),$$

because $t_k \geq 0$ implies $t_{k+1} \geq h_{k+1} \geq h_{\min}$, and g is a decreasing function.

(A3) For the interval $[\tau_{k-1}, \tau_k]$, $\tau_{k-1} + h_k/2 \le t \le \tau_k$ and we consider that

$$|g(|\tau + \alpha h_k|) - g(|\tau|)| \le |g(|\tau + \alpha h_k|)| + |g(|\tau|)|.$$

We decompose, accordingly, A_k into two integrals:

$$A_k(t) \le \int_0^1 \int_{t_{k-1}}^{t_k - \alpha h_k} g(|\tau + \alpha h_k|) d\tau d\alpha + \int_0^1 \int_{t_{k-1}}^{t_k - \alpha h_k} g(|\tau|) d\tau d\alpha.$$

With the change of variable $\sigma = \tau + \alpha h_k$ in the first integral, and enlarging all the intervals of τ to $[t_{k-1}, t_k]$, we get

$$\begin{aligned} A_k(t) &\le \int_0^1 \int_{t_{k-1}+\alpha h_k}^{t_k} g(|\sigma|) d\sigma d\alpha + \int_0^1 \int_{t_{k-1}}^{t_k} g(|\tau|) d\tau d\alpha \\ &\le 2 \int_0^1 \int_{t_{k-1}}^{t_k} g(|\sigma|) d\sigma d\alpha \le 2 \int_0^1 \int_{-h_k}^{h_k/2} g(|\sigma|) d\sigma d\alpha, \end{aligned}$$

since $0 \le t_k \le h_k/2$ implies that $-h_k \le t_k - h_k = t_{k-1}$ and $t_k \le h_k/2$. Hence,

$$A_k(t) \le 2\Big(\int_0^{h_k/2} g(\sigma) d\sigma + \int_0^{h_k} g(\sigma) d\sigma \Big).$$

(A4) For the interval $[\tau_k, \tau_{k+1}]$, $\tau_k + h_k/2 \le t \le \tau_{k+1}$. We consider a similar decomposition of A_{k+1} into two integrals:

$$A_{k+1}(t) \le \int_0^1 \int_{t_k}^{t_{k+1} - \alpha h_{k+1}} g(|\tau + \alpha h_{k+1}|) d\tau d\alpha + \int_0^1 \int_{t_k}^{t_{k+1} - \alpha h_{k+1}} g(|\tau|) d\tau d\alpha.$$

With the change of variable $\sigma = \tau + \alpha h_{k+1}$ in the first integral, and enlarging the intervals of τ to $[t_k, t_{k+1}]$, we have

$$\begin{aligned} A_{k+1}(t) &\le \int_0^1 \int_{t_k+\alpha h_{k+1}}^{t_{k+1}} g(|\sigma|) d\sigma d\alpha + \int_0^1 \int_{t_k}^{t_{k+1} - \alpha h_{k+1}} g(|\tau|) d\tau d\alpha \\ &\le 2 \int_0^1 \int_{t_k}^{t_{k+1}} g(|\sigma|) d\sigma d\alpha \le 2 \int_0^1 \int_0^{h_{k+1} + h_k/2} g(\sigma) d\sigma d\alpha, \end{aligned}$$

since $0 \le t_k$ and $t_{k+1} \le h_k/2 + h_{k+1}$; hence,

$$A_{k+1}(t) \le 2 \int_0^{3h_{\max}/2} g(\sigma) d\sigma.$$

(B) The case $\tau_{k-1} + h_k/2 \ge t \ge \tau_{k-1}$, that is $t_{k-1} \ge -h_k/2$, gives the same partial bounds.

So for all cases of t we have

$$\sum_{j=1}^{n} A_j(t) \le h_{\max}(g(h_{\min}/2) + g(h_{\min}) - 2g(b-a))$$
$$+ 2\int_0^{h_{\max}/2} g(\sigma)d\sigma + 2\int_0^{h_{\max}} g(\sigma)d\sigma + 2\int_0^{3h_{\max}/2} g(\sigma)d\sigma, \quad (1.9)$$

and the bound (1.8) follows by considering the supremum of (1.9) when $t \in [a, b]$ and by multiplying it by 2.

1.3 Computational Experiments

We consider the function $g(s) = -\ln(s/2)$, $s \in]0, 2]$, $z = 4$ and the following right-hand side function:

$$f(s) := \begin{cases} -1 & \text{if } 0 \le s \le 1, \\ 0 & \text{if } 1 < s \le 2. \end{cases}$$

In this example we will compute the Galerkin approximate solution with uniform grids of 501 and 1001 points, respectively.

As we do not know the exact solution, we will take as reference solution the one obtained with a uniform grid of 4001 nodes, in Figure 1.2, and use it in the computation of the absolute errors of solutions corresponding to the two, much coarser, grids built with $n = 500$ and $n = 1000$ subintervals, respectively.

Figure 1.2, the reference solution, and Figure 1.3, the approximate coarser one, look very similar, and so the error with respect to this reference solution is plotted in Figure 1.4 for a uniform grid with 501 nodes. In Figure 1.5 we plot the error corresponding to an approximation with a uniform grid with 1001 nodes.

As we can see, the error reduces by a factor of approximately 2, when we double the number of subintervals in the grid. We can also see that the error is larger where the kernel has a logarithmic discontinuity (near 0) and where the right-hand side function f has a discontinuity (near 1).

Elementary computations and [ALL01], [AALT05], and [ALT01] show that, in Theorem 2,

$$0 < C \le \frac{1}{2 - 2\ln 2} < 1.63 \quad \text{and} \quad (I - \pi_n)f = 0,$$

and so the error bound in Theorem 4 can be computed explicitly as given in Table 1.1. This table also contains the values of the L^1-norm of the relative error (using the reference solution) and, as expected, it shows that the bound is somewhat pessimistic, in this example. It also shows that doubling the number of subintervals, the error bound reduces correspondingly.

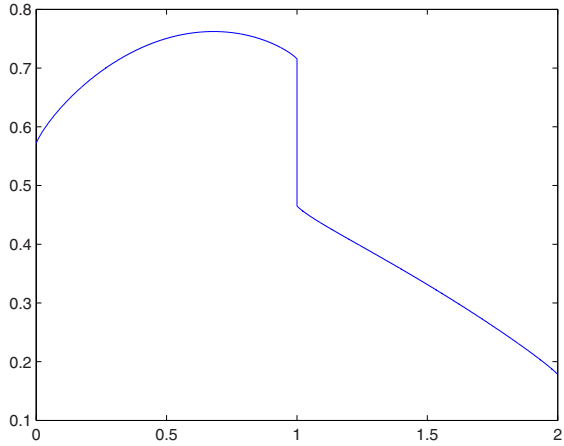


Fig. 1.2. Reference solution with uniform grid of 4001 nodes.

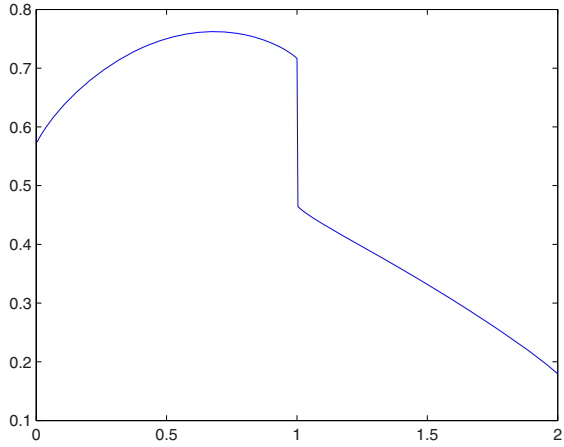


Fig. 1.3. Approximate solution with uniform grid of 501 nodes.

1.4 Bibliographical Comments and Conclusions

The Galerkin approximation to a compact integral operator is the cheapest one among projection discretizations (see [At97], [ALL01], and [Ch83]). The L^1 class of functions is the largest space among the Lebesgue ones. Weakly singular kernels define the most general integral operators among the com-

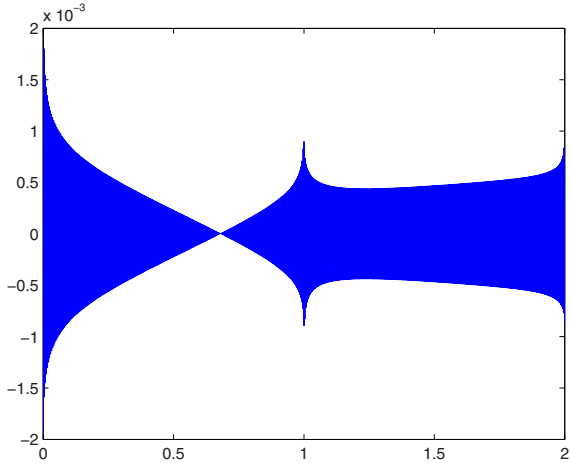


Fig. 1.4. Error of a 501-node-uniform-grid solution with respect to the reference solution.

Table 1.1. L^1-norm relative errors in uniform grids.

n	$h_{\max}$	Error Bound	L^1-norm Relative Error
500	0.004	0.729	0.000736
1000	0.002	0.401	0.000251

pact ones. Hence, the framework of this paper is as general and as weak as possible in the domain of numerical resolution of Fredholm integral equations of the second kind. The main theoretical result is Theorem 4, in which a relative error bound is produced. Other efforts in this sense have been accomplished in [AALT05] and [ALT01], where the condition of quasi-uniformity is imposed to the underlying grid, and in [AAL09] where other Banach spaces and other projection-type discretizations are considered. The investigation of the existence of possible boundary layers in the solution thus deserving grid refinements has been studied in [AAL09] and [AAF09]. The numerical experiments presented in this paper are done with uniform grids and show that for a kernel with a logarithmic singularity and an equation whose right-hand side is a piecewise constant discontinuous function the Galerkin discretization studied in this chapter gives significantly better approximations than the ones expected by theory. The shape of the relative error function shows that the predicted boundary layers have occurred in practice and that the numerical solution is less accurate in those subdomains.

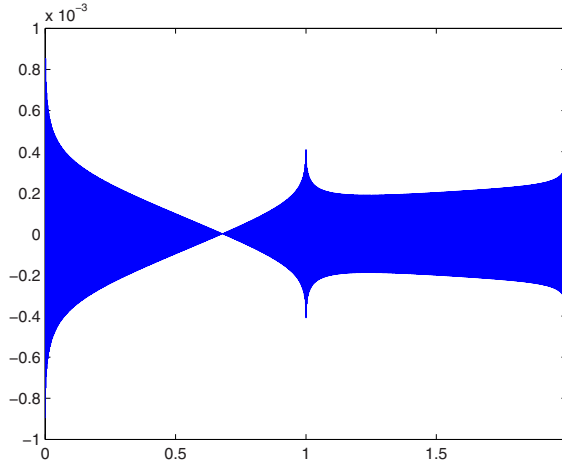


Fig. 1.5. Error of a 1001-node-uniform-grid solution with respect to the reference solution.

Acknowledgement. This research was partially supported by FCT (Fundação para a Ciência e a Tecnologia) through the doctoral scholarship SFRH/BD/30826/2006.

References

[At97] Atkinson, K.: *The Numerical Solution of Integral Equations of the Second Kind,* Cambridge University Press, London (1997).

[AAL09] Amosov, A., Ahues, M., Largillier, A.: Superconvergence of some projection approximations for weakly singular integral equations using general grids, *SIAM J. Numer. Anal.,* **47**, 646–674 (2009).

[AALT05] Ahues, M., Amosov, A., Largillier, A., Titaud, O.: L^p error estimates for projection approximations. *Appl. Math. Lett.,* **18**, 381–386 (2005).

[AAF09] Ahues, M., D'Almeida, F., Fernandes, R.: Piecewise constant Galerkin approximations of weakly singular integral equations, accepted for publication in *Internat. J. Pure Appl. Math.* (to appear).

[ALL01] Ahues, M., Largillier, A., Limaye, B.V.: *Spectral Computations with Bounded Operators,* Chapman & Hall/CRC, Boca Raton, FL (2001).

[ALT01] Ahues, M., Largillier, A., Titaud, O.: The roles of a weak singularity and the grid uniformity in relative error bounds. *Numer. Functional Anal. Optimization,* **22**, 789–814 (2001).

[Ch83] Chatelin, F.: *Spectral Approximations of Linear Operators,* Academic Press, New York (1983).

2

Construction of Solutions of the Hamburger–Löwner Mixed Interpolation Problem for Nevanlinna Class Functions

J.A. Alcober, I.M. Tkachenko, and M. Urrea

Universidad Politécnica de Valencia, Spain; juaalau@mat.upv.es, imtk@mat.upv.es, murrea@mat.upv.es

2.1 Formulation of the Problem

By definition, a Nevanlinna class function $\varphi \in \mathfrak{R}$ is holomorphic and has a nonnegative imaginary part in the half-plane $\operatorname{Im} z > 0$. In this chapter we also consider Nevanlinna functions which belong to the subclass $\mathfrak{R}_0 \subset \mathfrak{R}$ such that if $\varphi(z) \in \mathfrak{R}_0$, $\lim_{z\to\infty}(\varphi(z)/z) = 0$, $\operatorname{Im} z > 0$. Then, due to the Riesz–Herglotz theorem,

$$\varphi(z) = \int_{-\infty}^{\infty} \frac{d\sigma(t)}{t-z}, \quad \operatorname{Im} z > 0, \tag{2.1}$$

where $\sigma(t)$ is a nondecreasing function such that $\int_{-\infty}^{\infty}\left(1+t^2\right)^{-1} d\sigma(t) < \infty$. Consider the mixed Löwner–Nevanlinna problem [Löw34, KrNu77, Akh65, KaSt66, CuFi91, CuFi96, AdTk00, UrTkFC01, AdAlTk03], see also [AdTk01(a)] and (for the matrix version of the problem) [AdTk01(b)].

Problem 1. Given a set of real numbers $(c_0, ..., c_{2n})$, a finite set of points $(t_1, ..., t_p)$ on the real axis, and a set of complex numbers $(w_1, ..., w_p)$ with nonnegative imaginary parts, find a function of the Nevanlinna class $\varphi \in \mathfrak{R}_0$ such that asymptotically, for $z \to \infty$ inside any angle $\delta < \arg z < \pi - \delta$, $\delta > 0$,

$$\varphi(z) = -\sum_{r=1}^{2n+1} c_{r-1} z^{-r} + o\left(|z|^{-2n-1}\right) \tag{2.2}$$

possesses continuous boundary values in some vicinities of the points $(t_1, ..., t_p)$ and

$$\varphi(t_s + i0) = w_s, \quad s = 1, ..., p. \tag{2.3}$$

Remark 1. Notice that by virtue of the representation (2.1) [Akh65], condition (2.2) is equivalent to the moment conditions

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_2,

$$\int_{-\infty}^{\infty} t^k d\sigma(t) = c_k, \quad k = 0, 1, ..., 2n, \tag{2.4}$$

for the generating distribution $\sigma(t)$.

Remark 2. The suggested Problem 1 is a mixture of the truncated Hamburger moment problem [KrNu77, CuFi91, AdTk00] with the Löwner-type interpolation problem in the class of Nevanlinna functions [Löw34].

We describe and test numerically an algorithm for finding irrational solutions of this problem. The rational solutions of a similar problem were discussed in [AdAlTk03]. Errors of such approximations depending on the number and distribution of the interpolation nodes on the real axis will be discussed elsewhere.

These kind of problems occur when a distribution density reconstruction from scarce experimental data is attempted. In other words, we are interested in the possibility of solving the problem when only a very small number of moments and constraints (data at the interpolation nodes) is known.

The studies of convergence as the number of moments and/or interpolation nodes grows are out of the scope of this work. Untruncated moment problems are solved in the classical theory of moments, see [KrNu77] and [Akh65]. The behavior of the problem solution when the number of interpolation nodes grows is treated in [DeDy81].

2.2 The Mixed Problem Solution

2.2.1 Solvability and Contractive Functions

Recall that the truncated Hamburger moments problem is solvable [KrNu77, CuFi96, AdTk00] if and only if the block-Hankel matrix $(c_{k+l})_{k,l=0}^n$ is nonnegative. If, in addition, we exclude from our consideration the nonnegative block-Hankel matrices like

$$\begin{pmatrix} 0 & 0 \\ 0 & \gamma \end{pmatrix}, \quad \gamma > 0,$$

which cannot be generated by power moments of nonnegative measures, and if the set $(c_0, ..., c_{2n})$ is positive definite, there exists an infinite set of nonnegative measures σ on the real axis satisfying (2.4).

Let $(D_k(t))_{k=0}^n$ be the finite set of polynomials constructed according to the formulas

$$D_0 = \frac{1}{\sqrt{c_0}}\ , \quad D_k(t) = \frac{1}{\sqrt{\Delta_{k-1}\Delta_k}} \det \begin{pmatrix} c_0 & \cdots & c_{k-1} & 1 \\ c_1 & \cdots & c_k & t \\ \vdots & \vdots & \vdots & \vdots \\ c_k & \cdots & c_{2k-1} & t^k \end{pmatrix}\ ,$$

$$\Delta_{-1} = 1\ , \quad \Delta_0 = c_0\ , \quad \Delta_k = \det \begin{pmatrix} c_0 & \cdots & c_k \\ \vdots & \vdots & \vdots \\ c_k & \cdots & c_{2k} \end{pmatrix}\ , k = 1, 2, ..., n. \tag{2.5}$$

Polynomials D_k form an orthogonal system with respect to each σ-measure satisfying (2.4). Let

$$E_0 \equiv 0,\ E_k(t) = \int_{-\infty}^{\infty} \frac{D_k(t) - D_k(s)}{t - s}\ d\sigma(s),\ k = 1, ..., n,$$

be the corresponding set of conjugate polynomials.

Then the formula

$$\varphi(z) = \int_{-\infty}^{\infty} \frac{d\sigma(t)}{t - z} = -\frac{E_n(z)(\zeta(z) + z) - E_{n-1}(z)}{D_n(z)(\zeta(z) + z) - D_{n-1}(z)}, \quad \operatorname{Im} z > 0, \quad n = 1, 2, \ldots \tag{2.6}$$

according to the Nevanlinna theorem [KrNu77, UrTkFC01], establishes a one-to-one correspondence between the set of all Nevanlinna functions $\varphi(z)$ satisfying (2.2) and the elements $\zeta(z)$ of the subclass $\mathfrak{R}_0$.

Notice that the zeros of each orthogonal polynomial $D_k(z)$ are real and by virtue of the Schwarz–Christoffel identity [KrNu77]

$$D_{n-1}(z)E_n(z) - D_n(z)E_{n-1}(z) \equiv \Xi_n = \frac{\Delta_{n-1}}{\sqrt{\Delta_{n-2}\Delta_n}} > 0\ , \quad n = 1, 2, \ldots, \tag{2.7}$$

the zeros of $D_{n-1}(z)$ alternate with the zeros of $D_n(z)$ as well as with the zeros of $E_{n-1}(z)$. Therefore, *any function $\varphi(z)$ given by the expression on the right-hand side of (2.6) has a continuous boundary value on the real axis if and only if the corresponding Nevanlinna function $\zeta \in \mathfrak{R}_0$ is continuous in the closed upper half-plane and such that $\zeta(z) + z$ has no joint zeros with $D_{n-1}(z)$.*

To meet the constraints (2.3) it suffices to substitute into the right-hand side of (2.6) any function $\zeta(z) \in \mathfrak{R}_0$ which is continuous in the closed upper half-plane and satisfies the following conditions:

$$\xi_s = \zeta\left(t_s\right) = -t_s + \frac{w_s D_{n-1}(t_s) + E_{n-1}(t_s)}{w_s D_n(t_s) + E_n(t_s)},\ s = 1, ..., p. \tag{2.8}$$

Note that by (2.7), $\operatorname{Im}\ \xi_s = \Xi_n \operatorname{Im}\ w_s \left|w_s D_n(t_s) + E_n(t_s)\right|^{-2} > 0,\ s = 1, ..., p.$ Thus, Problem 1 reduces to the following.

Problem 2. Given a finite number of distinct points $t_1, \ldots, t_p$ of the real axis and a set of complex numbers $w_1, \ldots, w_p$ with positive imaginary parts, find the set of functions $\zeta(z) \in \mathfrak{R}_0$ continuous in the closed upper half-plane which satisfy conditions (2.8).

Each Nevanlinna function $\zeta(z)$ in the upper half-plane admits the Cayley representation

$$\zeta(z) = i\frac{1+\theta(z)}{1-\theta(z)}, \tag{2.9}$$

where

$$\theta(z) = \frac{\zeta(z)-i}{\zeta(z)+i} \tag{2.10}$$

is a holomorphic function on the upper half-plane with *contractive* values, i.e., $|\theta(z)| \leq 1$, $\operatorname{Im} z > 0$. The function $\theta(z)$ connected with the Nevanlinna function $\zeta(z)$ by the linear fractional transformation (2.10) is continuous in the closed upper half-plane if $\zeta(z)$ satisfies this condition. On the other hand, the Nevanlinna function $\zeta(z)$ given as the linear fractional transformation (2.9) of a function $\theta(z)$ which is holomorphic on the upper half-plane, continuous in its closure, and has contractive values, is continuous at the points of the closed upper half-plane where $\theta(z) \neq 1$. Therefore, Problem 2 is equivalent to the following problem for contractive functions.

Let $\mathfrak{B}$ be the set of all contractive functions which are holomorphic on the upper half-plane and continuous on its closure.

Problem 3. Given a finite number of distinct points $t_1, ..., t_p$ of the real axis and a set of points $\lambda_1, ..., \lambda_p$,

$$\lambda_s = \frac{\xi_s - i}{\xi_s + i},\ |\lambda_s| \leq 1,\ s = 1, ..., p,$$

find a set of functions $\theta \in \mathfrak{B}$ such that

$$\theta(t_s) = \lambda_s,\ s = 1, ..., p. \tag{2.11}$$

Remark 3. Problem 3 is a limiting case of the Nevanlinna–Pick problem [Akh65, KrNu77] with interpolation nodes on the real axis. Its solvability for any interpolation data $\lambda_1, ..., \lambda_p$ inside the unit circle was actually proven in [KhaTa85]. The point is that the associated Pick matrix is automatically positive definite for given contractive interpolation values once the interpolation nodes are close enough to the axis; this guarantees that the approximate Nevanlinna–Pick problem is solvable once the interpolation nodes are close enough to the real line. Then one applies the Vitali–Montel theorem to take the limit as the interpolation nodes go to the real line. This implies also that the Nevanlinna–Pick problem is solvable even if some or all $|\lambda_s| = 1$.

We describe below an algorithm of the solution of Problem 3 when all $|\lambda_s| < 1$, which is a simple modification of the Schur algorithm. An alternative algorithm [AdAlTk03], similar to the Lagrange method of interpolation theory, can be applied if some or even all $|\lambda_s| = 1$.

2.2.2 Schur Algorithm

Note that a function $\theta \in \mathfrak{B}$ satisfies the condition $\theta(t_1) = \lambda_1,\ |\lambda_1| < 1$, if and only if it admits the representation

$$\theta(z) = \frac{\phi(z) + \lambda_1}{\overline{\lambda_1}\phi(z) + 1}\ ,$$

where $\phi \in \mathfrak{B}$ and $\phi(t_1) = 0$. In the case of the Nevanlinna–Pick problem, i.e., when t_1 belongs to the upper half-plane, the function $\phi(z)$ admits the representation $\phi(z) = \left((z - t_1) / \left(z - \overline{t_1}\right)\right) \chi(z)$, where $\chi(z)$ is an arbitrary contractive function in the upper half-plane. There is no such simple form for the contractive function $\phi(z)$ when $t_1 \in \mathbb{R}$.

Here we wish to carry out the reconstruction procedure using irrational functions. To this end, we propose to use the function

$$\phi(z) = \theta_1(z) \exp\left\{ \frac{\alpha}{\pi i} \int\limits_{t_1-1}^{t_1+1} \frac{1 + tz}{t - z} \ln|t - t_1| \frac{dt}{t^2 + 1} \right\} := \theta_1(z)\, u_1(z)\,,$$

with a unique free parameter $\alpha \in (0, 1)$. Here θ_1 is any function from $\mathfrak{B}$ such that

$$\theta_1(t_s) = \lambda'_s = \frac{1}{u_1(t_s)} \frac{\lambda_s - \lambda_1}{1 - \overline{\lambda_1}\lambda_s},\ s = 2, ..., p. \tag{2.12}$$

Such a choice of $\theta_1(z)$ guarantees the verification of all of the conditions (2.11). Hence, Problem 3 with p nodes of interpolation on the real axis and strictly contractive values of the functions to find at these nodes, reduces to the same problem but with $p - 1$ nodes of interpolation and modified values at these nodes given by (2.12). Repeating the above procedure $p - 1$ times with a suitable choice of the parameter α and modifying the values of emerging contractive functions at the remaining points $t_{s+1}, ..., t_p$ according to (2.12), permits us to obtain some solution of Problem 3. Observe that contrary to the Nevanlinna–Pick problem with nodes in the open upper half-plane, our Problem 3 is always solvable if the values of the function to reconstruct are contractive at the nodes of interpolation.

Let $\theta_{s-1} \in \mathfrak{B}$ be a contractive function emerging after the $s - 1$ step in the course of the Problem 3 solution by the above method, and let $\lambda_s^{(s-1)} = \theta_{s-1}(t_s),\ \lambda_1^{(0)} = \lambda_1$. It follows from the above arguments that should the initial parameters $\lambda_1, ..., \lambda_p$ be strictly contractive, there exists a set of solutions of Problem 3 described by the formula

$$\theta(z) = \frac{a(z)\mu(z) + b(z)}{c(z)\mu(z) + d(z)}, \tag{2.13}$$

where the elements of the matrix of the linear fractional transformation (2.13) are irrational functions constructed as above and $\mu(z)$ runs the subset of all

functions from $\mathfrak{B}$ satisfying the condition $\mu(t_p) = \lambda_p^{(p-1)}$. This matrix can be calculated as

$$\begin{pmatrix} a(z) & b(z) \\ c(z) & d(z) \end{pmatrix} = \overset{\curvearrowright}{\prod_{s=1}^{p-1}} \begin{pmatrix} u_s(z) & \lambda_s^{(s-1)} \\ \overline{\lambda_s^{(s-1)}} u_s(z) & 1 \end{pmatrix} ,$$

where the numbers s in the matrix factors on the right-hand side increase from left to right.

Observe that the simplest choice for the function $\mu(z)$ in (2.13) is just $\mu(z) \equiv \lambda_p^{(p-1)}$. Hence, if initial parameters $\lambda_1, ..., \lambda_p$ in Problem 3 are strictly contractive, then among the solutions of this problem there are irrational functions of the type we consider.

A numerical testing of this representation is given in the next section.

2.3 Numerical Results

To check the quality of the reconstruction technique we suggest, we carried out an extensive study of the present approach as applied to a number of distribution densities: $\exp\left(-t^2\right)$, $\exp\left(-t^4 + 2t^2\right)$, $\exp\left(-3t^4 - 5t^3 + \frac{3}{4}t^2 + 3t + 1\right)$, and $\exp\left(-0.16t^6 - 0.15t^5 + 0.75t^4 + 0.5t^3 - t^2 - 0.25t + 0.1\right)$ with the latter two selected to possess clear extrema and to be unsymmetrical with nonzero odd-order moments.

Since we try to reconstruct certain nonnegative densities, the solvability of the moment problem is not an issue. In each case the absolutely continuous nonnegative measure with this density is just one of the solutions of the moment problem. We use a finite, very small number of moments, which can be easily estimated numerically, i.e., we want to solve the truncated problem which, since the sought measure has a nonzero density, has infinitely many solutions.

Remark 4. Notice that earlier [AdAlTk03] we tested the numerical viability of an algorithm of reconstruction (of the densities 1 and 2) using the moment technique without local constraints. It turned out that one needed to know the values of hundreds of power moments to obtain some acceptable agreement between the numerically generated density and the one whose moments were used. It is clear that such an approach is of no practical importance.

To apply the Schur-like algorithm described above, one has to know not only the values of some power moments of the distribution density $f(t)$ under investigation,

$$c_k = \int_{-\infty}^{\infty} t^k f(t) dt \ , \quad k = 0, 1, \ldots, 2n \ , \quad n = 1, 2, \ldots,$$

but also the values of the Nevanlinna function,

$$w_s = \varphi(t_s) = P.V. \int_{-\infty}^{\infty} \frac{f(t)dt}{t - t_s} + i\pi f(t_s) \ ,$$

at the set of points $(t_1, ..., t_p) \subset \mathbb{R}$.

In all four cases we consider, the latter principal value integrals were computed numerically and the sets of orthogonal polynomials (2.5) were calculated directly, while the conjugate polynomials were generated using the recurrence relations stemming from the Schwarz–Christoffel identity (2.7).

To find the value of the parameter $\alpha \in (0, 1)$ of the auxiliary function

$$u_s(z) = \exp\left\{ \frac{\alpha}{\pi i} \int_{t_s-1}^{t_s+1} \frac{1+tz}{t-z} \ln|t - t_s| \frac{dt}{t^2+1} \right\}, \quad s = 1, 2, \ldots, p,$$

we made use of the Shannon entropy [TkUr99]

$$\mathfrak{S}(\alpha) = -\int_{-\infty}^{+\infty} \psi(\alpha, t) \ln(\psi(\alpha, t))\, dt,$$

where the density $\psi(\alpha, t)$ is the one reconstructed within the algorithm, i.e., it is the imaginary part (divided by π) of the Nevanlinna model function obtained by our algorithm. The density $\psi(\alpha, t)$ has no real poles and is positive over the whole real axis, hence it is quite easy to solve the maximization procedure equation: $d\mathfrak{S}(\alpha)/d\alpha = 0$.

Our numerical results can be summarized in the following way. In all figures the dashed lines correspond to the original distributions. Some averaging procedure was applied to minimize the influence of the choice of the initial point. Precisely, first the lowest of the points of interpolation, $t_1, \ldots t_p$, was chosen as the initial point (the case 123) and then the points were chosen in the inverse order, in each case a certain value of α labeled $(1 \cdots p)$ or $(p \cdots 1)$ was obtained and, finally, we took the average of these two groups of data and plotted the figure. In Figure 2.1 we display the Gauss distribution reconstructed using $p = 3$ points and $n = 2$ (which are 3 nonzero moments in this case). Here we have $\alpha_{123} = 0.793437$ and $\alpha_{321} = 0.7965$. Figure 2.2 corresponds to $f(t) = \exp\left(-t^4 + 2t^2\right)$ ($p = 3$ and $n = 2$); we have $\alpha_{123} = 0.526876$, $\alpha_{321} = 0.524844$. In Figure 2.3 we display our results for $f(t) = \exp\left(-3t^4 - 5t^3 + \frac{3}{4}t^2 + 3t + 1\right)$ obtained with $p = 3$ and $n = 2$ (5 nonzero moments in this case) and with $\alpha_{123} = 0.264258$, $\alpha_{321} = 0.3675$. Figures 2.4 and 2.5 correspond to the function $f(t) = \exp\left(-0.16t^6 - 0.15t^5 + 0.75t^4 + 0.5t^3 - t^2 - 0.25t + 0.1\right)$ (the first one with $p = 3$ and $n = 2$ (5 moments), and the second one with $p = 3$ and $n = 1$ (3 moments)). They were obtained choosing $\alpha_{12345} = 0.650153$ and $\alpha_{54321} = 0.69748$ in the first case, and $\alpha_{12345} = 0.622827$ and $\alpha_{54321} = 0.613504$ in the second one.

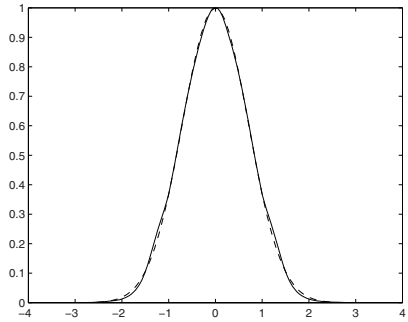


Fig. 2.1. An example of irrational reconstruction (solid line) of a symmetric distribution density (dashed line) by 3 local constraints and 5 (3 nonzero) power moments.

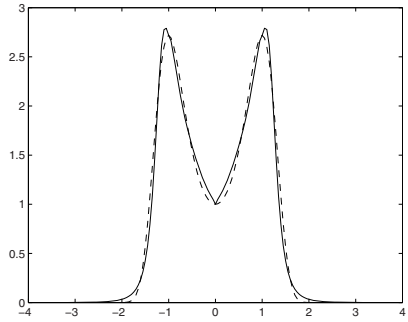


Fig. 2.2. Another example of irrational reconstruction (solid line) of a symmetric distribution density (dashed line) by 3 local constraints and 5 (3 nonzero) power moments.

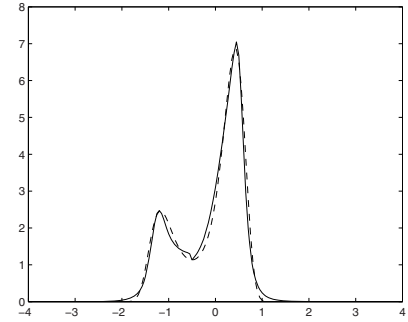


Fig. 2.3. An example of irrational reconstruction (solid line) of an asymmetric distribution density (dashed line) by 3 local constraints and 5 power moments.

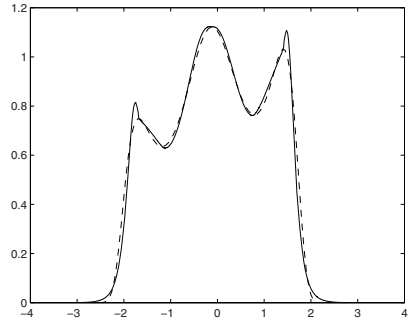


Fig. 2.4. An example of irrational reconstruction (solid line) of an asymmetric distribution density (dashed line) by 3 local constraints and 5 power moments.

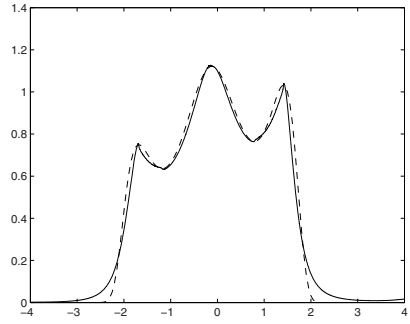


Fig. 2.5. Another example of irrational reconstruction (solid line) of an asymmetric distribution density (dashed line) by 3 local constraints and only 3 power moments.

2.4 Conclusions

An algorithm is presented which permits us to obtain, at least in the cases we consider, a reasonable agreement between the selected densities and their irrational counterparts reconstructed by a few integral characteristics, the power moments, and the local constraints. Further applicability and convergence properties of the approach are to be considered elsewhere.

Acknowledgement. Valuable discussions with V. Adamyan and P. Kurasov are gratefully acknowldeged.

References

[AdTk00] Adamyan, V.M., Tkachenko, I.M.: Solution of the truncated Hamburger moment problem according to M.G. Krein. *Operator Theory Adv. Appl.,* **OT-118**, 33–51 (2000).

[AdTk01(a)] Adamyan, V.M., Tkachenko, I.M.: Truncated Hamburger moment problems with constraints. *Math. Studies,* **189**, 321–333 (2001).

[AdTk01(b)] Adamyan, V.M., Tkachenko, I.M.: Truncated Hamburger matrix moment problems with constraints. *Proc. Appl. Math. Mech.,* **1**, 420–421 (2001).

[AdAlTk03] Adamyan, V.M., Alcober, J.A., Tkachenko, I.M.: Reconstruction of distributions by their moments and local constraints. *Appl. Math. Research eXpress,* **2003**, 33–70.

[Akh65] Akhiezer, N.I.: *The Classical Moment Problem and Some Related Questions in Analysis,* Hafner, New York (1965).

[CuFi91] Curto, R.E., Fialkow, L.A.: Recursiveness, positivity, and truncated moment problems. *Houston J. Math.,* **17**, 603–635 (1991).

[CuFi96] Curto, R.E., Fialkow, L.A.: Solutions of the truncated moment problem for flat data. *Mem. Amer. Math. Soc.,* **119** (1996).

[DeDy81] Dewilde, P., Dym, H.: Schur recursions, error formulas, and convergence of rational estimators for stationary stochastic sequences. *IEEE Trans. Information Theory,* **IT-27**, 446–461 (1981).

[KrNu77] Krein, M.G., Nudel'man, A.A.: *The Markov Moment Problem and Extremal Problems,* American Mathematical Society, Providence, RI (1977).

[KaSt66] Karlin, S., Studden, W.S.: *Tschebyscheff Systems with Applications in Analysis and Statistics,* Interscience, New York (1966).

[KhaTa85] Khargonekar, P., Tannenbaum, A.: Non-Euclidian metrics and the robust stabilization of systems with parameter uncertainty. *IEEE Trans. Automat. Contr.,* **AC-30**, 1005–1013 (1985).

[Löw34] Löwner, K.: Über monotone Matrixfunktionen. *Math. Z.,* **38**, 177 (1934).

[TkUr99] Tkachenko, I.M., Urrea, M.: Determination of the adjustable moment model parameter by the minimization of the Shannon entropy. *Z. Angew. Math. Mech.,* **79**, 789–790 (1999).

[UrTkFC01] Urrea, M., Tkachenko, I.M., Fernández de Córdoba, P.: The Nevanlinna theorem of the classical theory of moments revisited. *J. Appl. Anal.,* **7**, 209–224 (2001).

3

A Three-Dimensional Eutrophication Model: Analysis and Control

L.J. Alvarez-Vázquez,[1] F.J. Fernández,[2] and R. Muñoz-Sola[2]

[1] Universidad de Vigo, Spain; lino@dma.uvigo.es
[2] Universidad de Santiago de Compostela, Spain; franfdz@usc.es, rafams@usc.es

3.1 The Environmental Problem

Eutrophication of lakes, reservoirs, streams, and coastal areas is one of the most widespread environmental problems of large water bodies. Eutrophication consists of unnatural enrichment with two plant nutrients: nitrogen and phosphorus. This overnutrification causes undesirable changes in water resources: excessive production of algae, deterioration of water quality and availability, fish kills, health hazards for humans, etc. Controlling the eutrophication is important in order to mitigate and remedy the problem.

The basic idea of a bioreactor consists holding up hypernutrified water (rich, for instance, in nitrogen) in large tanks where we add a certain quantity of phytoplankton, that we let freely grow to absorb nitrogen from the water. In the particular case analyzed in this chapter we have considered only two large shallow tanks with the same capacities (but possibly different geometries). Water rich in nitrogen fills the first tank Ω_1, where we add a quantity ρ^1 of phytoplankton (which we let grow for a permanence time T^1) to drop, nitrogen level down to a desired threshold. We are also interested in obtaining a certain quantity of organic detritus (very desirable for use as agricultural fertilizer) in this first tank. Once we reach the desired levels of nitrogen and organic detritus (settled in the bottom of the tank, and then reclaimed for agricultural use), we drain this first tank and pass water to the second tank Ω_2, where the same operation is repeated, by adding a new amount ρ^2 of phytoplankton. Water leaving this second fermentation tank after a time period T^2 will usually be poor in nitrogen, but rich in detritus (settled in the bottom) and phytoplankton (recovered from a final filtering). At this point, we are interested (both for economical and ecological reasons) in minimizing this final quantity of phytoplankton. Thus, the optimal control problem consists of finding the quantities (ρ^1, ρ^2) of phytoplankton that we must add to each tank during the respective times so that nitrogen levels are lower than the maximum thresholds and the detritus levels are higher than the minimum

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_3,

thresholds, and in such a way that the final phytoplankton concentration is as reduced as possible.

From a mathematical viewpoint, this problem can be formulated as an optimal control problem with state-control constraints, where the controls (ρ^1, ρ^2) are the quantities of phytoplankton added at each tank, the state variables are the concentrations of representative species, the objective function to be minimized is the phytoplankton concentration of water leaving the second tank, the state constraints stand for the thresholds required for nitrogen and detritus concentrations, and the control constraints are related to technological bounds. A detailed formulation of the problem is presented in the next section, along with results for existence and characterization of solutions. Finally, a numerical algorithm is proposed for computing the solution of the state system, and is applied to a realistic example.

3.2 The Analytical Problem

Recent mathematical models for the simulation of a eutrophication process are based in systems of partial differential equations with a high complexity due to the large variety of internal phenomena that they include. In this chapter we consider a realistic model with four biological variables involved (the meaning of the biochemical interaction terms can be found, for instance, in [Ca76] or [DCI01]). So, we consider the state $\mathbf{u} = (u^1, u^2, u^3, u^4)$, where $u^1(t,x)$ stands for a generic nutrient concentration (for instance, nitrogen), $u^2(t,x)$ for phytoplankton concentration, $u^3(t,x)$ for zooplankton concentration, and $u^4(t,x)$ for organic detritus concentration.

Interaction of these four species into a given still water domain $\Omega \subset \mathbb{R}^3$ (with a smooth enough boundary $\partial\Omega$) and along a time interval $I = (0,T)$ can be described by the following system of coupled partial differential equations for diffusion-reaction systems with Michaelis–Menten kinetics:

$$\begin{cases} \frac{\partial u^1}{\partial t} - \nabla \cdot (\mu_1 \nabla u^1) + C_{nc} L \frac{u^1}{K_N + u^1} u^2 - C_{nc} K_r u^2 - C_{nc} K_{rd} \Theta^{\theta - 20} u^4 = g^1, \\ \frac{\partial u^2}{\partial t} - \nabla \cdot (\mu_2 \nabla u^2) - L \frac{u^1}{K_N + u^1} u^2 + K_r u^2 + K_{mf} u^2 + K_z \frac{u^2}{K_F + u^2} u^3 = g^2, \\ \frac{\partial u^3}{\partial t} - \nabla \cdot (\mu_3 \nabla u^3) - C_{fz} K_z \frac{u^2}{K_F + u^2} u^3 + K_{mz} u^3 = g^3, \\ \frac{\partial u^4}{\partial t} - \nabla \cdot (\mu_4 \nabla u^4) - K_{mf} u^2 - K_{mz} u^3 + K_{rd} \Theta^{\theta - 20} u^4 = g^4, \end{cases}$$

in $Q = I \times \Omega$, with suitable boundary conditions on $\Sigma = I \times \partial\Omega$ and initial conditions in Ω, and where $\theta(t,x)$ is the water temperature (in degrees Celsius), $L(t,x)$ the luminosity function (related to incident light intensity and phytoplankton growth rate), $\mu_i,\ i = 1, \ldots, 4$, the diffusion coefficients of each species, C_{nc} the nitrogen-carbon stoichiometric relation, C_{fz} the grazing efficiency factor, Θ the detritus regeneration thermic constant, K_N and K_F the nitrogen and phytoplankton half-saturation constants, K_{mf} and K_{mz} the phytoplankton and zooplankton death rates (including predation), K_{rd} the detritus regeneration rate, K_r the phytoplankton endogenous respiration

rate, and K_z the zooplankton predation (grazing). The existence and uniqueness of solutions for this system have been previously obtained by the authors in the recently published paper [AFM09].

To present a simpler expression for the system, we consider the mapping $\mathbf{A} = (A^1, A^2, A^3, A^4) : \mathbb{R}_+ \times \Omega \times \mathbb{R}^4_+ \longrightarrow \mathbb{R}^4$, given by

$$\mathbf{A}(t,x,\mathbf{u}) = \begin{bmatrix} -C_{nc}\left[L(t,x)\frac{u^1}{K_N+u^1}u^2 - K_r u^2\right] + C_{nc}K_{rd}\Theta^{\theta(t,x)-20}u^4 \\ \left[L(t,x)\frac{u^1}{K_N+u^1}u^2 - K_r u^2\right] - K_{mf}u^2 - K_z\frac{u^2}{K_F+u^2}u^3 \\ C_{fz}K_z\frac{u^2}{K_F+u^2}u^3 - K_{mz}u^3 \\ K_{mf}u^2 + K_{mz}u^3 - K_{rd}\Theta^{\theta(t,x)-20}u^4 \end{bmatrix}.$$

Thus, the eutrophication system can be written in the following equivalent way:

$$\frac{\partial u^i}{\partial t} - \nabla\cdot(\mu_i\nabla u^i) = A^i(t,x,\mathbf{u}) + g^i \quad \text{in } Q, \quad \text{for } i = 1,\dots,4. \tag{3.1}$$

With this notation in mind we can formulate the bioreactor control problem with the following items.

- *Controls:* As already mentioned, we will control the system by means of two design variables: the quantities $\rho^j(t,x)$, $j = 1,2$, of phytoplankton added in the tank Ω_j along the time intervals $I_j = (0,T^j)$.
- *State systems:* We consider two state systems giving the concentrations of nitrogen-phytoplankton-zooplankton-organic detritus in each tank. Since both tanks are isolated, no transference for any of the four species is considered through the boundaries (i.e., Neumann boundary conditions are assumed to be null). Both systems will be coupled by means of the initial-final conditions: when water is passed from the first tank to the second one, it is natural to assume that water is mixed up, and this is the reason for considering the initial conditions for the concentrations inside the second tank as given by the corresponding averaged final concentrations in the first tank. These two state systems are given by
  - *First tank Ω_1:* The state variables for the first tank will be denoted $\mathbf{u}^1 = (u^{1,1}, u^{2,1}, u^{3,1}, u^{4,1})$ with $u^{1,1}$ (nitrogen), $u^{2,1}$ (phytoplankton), $u^{3,1}$ (zooplankton), and $u^{4,1}$ (organic detritus). The permanence time of water inside this first tank will be T^1, and the initial concentrations will be given by $\mathbf{u}^1_0 = (u^{1,1}_0, u^{2,1}_0, u^{3,1}_0, u^{4,1}_0)$. Thus, for $Q_1 = I_1 \times \Omega_1$ and $\Sigma_1 = I_1 \times \partial\Omega_1$, we have the system, for $i = 1,\dots,4$,

$$\begin{cases} \frac{\partial u^{i,1}}{\partial t} - \nabla\cdot(\mu_i\nabla u^{i,1}) = A^i(t,x,\mathbf{u}^1) + \delta_{2i}\rho^1 & \text{in } Q_1, \\ \frac{\partial u^{i,1}}{\partial n} = 0 & \text{on } \Sigma_1, \\ u^{i,1}(0) = u^{i,1}_0 & \text{in } \Omega_1, \end{cases} \tag{3.2}$$

where δ_{ji} denotes the Kronecker delta, that is, $\delta_{ji} = 1$ if $j = i$, and $\delta_{ji} = 0$ otherwise.

– *Second tank Ω_2 :* The state variables for the second tank will be denoted $\mathbf{u}^2 = (u^{1,2}, u^{2,2}, u^{3,2}, u^{4,2})$ with $u^{1,2}$ (nitrogen), $u^{2,2}$ (phytoplankton), $u^{3,2}$ (zooplankton), and $u^{4,2}$ (organic detritus). The permanence time of water inside this second tank will be T^2. Thus, for $Q_2 = I_2 \times \Omega_2$ and $\Sigma_2 = I_2 \times \partial\Omega_2$, we have, for $i = 1, \ldots, 4$,

$$\begin{cases} \frac{\partial u^{i,2}}{\partial t} - \nabla \cdot (\mu_i \nabla u^{i,2}) = A^i(t, x, \mathbf{u}^2) + \delta_{2i}\rho^2 & \text{in } Q_2, \\ \frac{\partial u^{i,2}}{\partial n} = 0 & \text{on } \Sigma_2, \\ u^{i,2}(0) = \frac{1}{\text{meas}(\Omega_1)} M_1^i(\mathbf{u}^1(T^1)) & \text{in } \Omega_2, \end{cases} \tag{3.3}$$

where $\mathbf{M}_j = (M_j^1, M_j^2, M_j^3, M_j^4)$, for $j = 1, 2$, are the functionals, defined from $[L^1(\Omega_j)]^4$ to $\mathbb{R}^4$, given by

$$\mathbf{M}_j(\mathbf{v}^j) = \begin{bmatrix} \int_{\Omega_j} v^{1,j} dx \\ \int_{\Omega_j} v^{2,j} dx \\ \int_{\Omega_j} v^{3,j} dx \\ 0 \end{bmatrix}.$$

(Note here that, since detritus settle before water passes to the second tank, the initial detritus concentration $u^{4,2}(0)$, i.e., the fourth component of $\mathbf{M}_1(\mathbf{u}^1(T^1))$, is considered null.)

- *Objective function:* Since we are interested in minimizing the final phytoplankton concentration of water leaving the second tank, we are led to consider the cost functional J given by

$$J(\rho^1, \rho^2) = \frac{1}{\text{meas}(\Omega_2)} \int_{\Omega_2} u^{2,2}(T^2) dx. \tag{3.4}$$

- *State constraints:* The final nitrogen concentration in each tank must be lower than a given threshold, and the final organic detritus concentration in each tank must be greater than another given threshold. These constraints translate into the relations given by $\mathbf{B} = (B^1, B^2, B^3, B^4)$, where

$$\begin{cases} B^1(\rho^1, \rho^2) = \frac{1}{\text{meas}(\Omega_1)} \int_{\Omega_1} u^{1,1}(T^1) dx \leq \sigma_1, \\ B^2(\rho^1, \rho^2) = \frac{1}{\text{meas}(\Omega_2)} \int_{\Omega_2} u^{1,2}(T^2) dx \leq \sigma_2, \\ B^3(\rho^1, \rho^2) = \frac{1}{\text{meas}(\Omega_1)} \int_{\Omega_1} u^{4,1}(T^1) dx \geq \theta_1, \\ B^4(\rho^1, \rho^2) = \frac{1}{\text{meas}(\Omega_2)} \int_{\Omega_2} u^{4,2}(T^2) dx \geq \theta_2, \end{cases} \tag{3.5}$$

for certain given values σ_1, σ_2, θ_1, $\theta_2 > 0$.

- *Control constraints:* Finally, for technological reasons, the quantities ρ^1, ρ^2 of phytoplankton added to the tanks must be nonnegative and bounded by a maximal admissible value $C > 0$, that is, they must lie in the set

$$\begin{aligned} \mathcal{U}_{ad} \quad = \quad & \{(\rho^1, \rho^2) \in L^2(Q_1) \times L^2(Q_2) \ : \\ & 0 \leq \rho^j(t, x) \leq C \ \text{ a.e. } \ (t, x) \in Q_j, \ j = 1, 2\}, \end{aligned}$$

which is a closed, bounded, convex, and nonempty subset of $L^2(Q_1) \times L^2(Q_2)$.

Thus, the formulation of the optimal control problem, denoted by $(\mathcal{P})$, will be the following:

$$\inf\big\{J(\rho^1,\rho^2) \text{ s.t. } (\rho^1,\rho^2)\in\mathcal{U}_{ad} \text{ and } (\mathbf{u}^1,\mathbf{u}^2) \text{ satisfies (3.2)–(3.3), and (3.5)}\big\}.$$

As proved by the authors in [AFM09], the eutrophication system (3.1) admits a solution under nonsmooth hypotheses. To be exact, if we assume that the fluid temperature $\theta\in L^2(Q)$ satisfies the boundedness condition $0\le\theta(t,x)\le M$ a.e. $(t,x)\in Q$, then the eutrophication system admits a unique solution $\mathbf{u}\in W^{1,2,2}(I;[H^1(\Omega)]^4,[H^1(\Omega)']^4)\cap[L^\infty(Q)]^4$, where

$$W^{1,p,q}(I;V,V')=\{v\in L^p(I;V)\ :\ \frac{du}{dt}\in L^q(I;V')\},$$

for any Banach space V and for $1\le p,q\le\infty$. Moreover, this solution $\mathbf{u}$ is nonnegative and bounded (in the previous space norm) by a value only depending on time T, second member $\mathbf{g}=(g^1,g^2,g^3,g^4)$, and initial-boundary values.

We say that $(\rho^1,\rho^2)\in\mathcal{U}_{ad}$ is a feasible control for problem $(\mathcal{P})$ if the associated state $(\mathbf{u}^1,\mathbf{u}^2)$, solution of (3.2)–(3.3), satisfies the constraints (3.5). Then, by standard minimizing sequences arguments, and taking into account that the solutions of the state systems are bounded and that $\mathcal{U}_{ad}$ is weakly closed with the topology of $L^2(Q_1)\times L^2(Q_2)$, we can prove the following existence result.

Theorem 1. *Let us assume that the set of feasible controls is nonempty. Let* $\mathbf{u}_0^1\in[L^\infty(\Omega_1)]^4$ *be such that* $0\le u_0^{i,1}(x)\le M$ *a.e.* $x\in\Omega_1$, $i=1,\dots,4$. *Then, there exist elements* $(\widetilde{\rho}^1,\widetilde{\rho}^2,\widetilde{\mathbf{u}}^1,\widetilde{\mathbf{u}}^2)\in\mathcal{U}_{ad}\times(W^{1,2,2}(I_1;[H^1(\Omega_1)]^4,[H^1(\Omega_1)']^4)\cap[L^\infty(Q_1)]^4)\times(W^{1,2,2}(I_2;[H^1(\Omega_2)]^4,[H^1(\Omega_2)']^4)\cap[L^\infty(Q_2)]^4)$ *such that* $(\widetilde{\rho}^1,\widetilde{\rho}^2)$ *is a solution of the control problem* $(\mathcal{P})$ *with associated state* $(\widetilde{\mathbf{u}}^1,\widetilde{\mathbf{u}}^2)$.

Finally, by classical adjoint state techniques, we can also derive the following necessary first order optimality condition, which characterizes the optimal solutions of the control problem $(\mathcal{P})$.

Theorem 2. *Let* $(\widetilde{\rho}^1,\widetilde{\rho}^2)\in\mathcal{U}_{ad}$ *be a solution of the control problem* $(\mathcal{P})$ *with associated state* $(\widetilde{\mathbf{u}}^1,\widetilde{\mathbf{u}}^2)\in(W^{1,2,2}(I_1;[H^1(\Omega_1)]^4,[H^1(\Omega_1)']^4)\cap[L^\infty(Q_1)]^4)\times(W^{1,2,2}(I_2;[H^1(\Omega_2)]^4,[H^1(\Omega_2)']^4)\cap[L^\infty(Q_2)]^4)$. *Then, there exist elements* $\gamma\ge0$ *and* $\lambda=(\lambda^1,\lambda^2,\lambda^3,\lambda^4)\in\mathbb{R}^4$ *such that*

$$\langle\lambda,\mathbf{B}(\widetilde{\rho}^1,\widetilde{\rho}^2)-\mu\rangle_{\mathbb{R}^4}\ge0,\quad\forall\mu\in[0,\sigma_1]\times[0,\sigma_2]\times[\theta_1,\infty)\times[\theta_2,\infty)\subset\mathbb{R}^4 \tag{3.6}$$

and

$$\sum_{j=1}^{2}\int_{Q_j}(\rho^j-\widetilde{\rho}^j)p^{2,j}\,dxdt\ge0,\quad\forall(\rho^1,\rho^2)\in\mathcal{U}_{ad}, \tag{3.7}$$

with $\mathbf{p}^j=(p^{1,j},p^{2,j},p^{3,j},p^{4,j})\in W^{1,2,2}(I_j;[H^1(\Omega_j)]^4,[H^1(\Omega_j)']^4)\cap[L^\infty(Q_j)]^4$, $j=1,2$, *the solutions of the following coupled linear adjoint systems, for* $i=1,\dots,4$,

$$\begin{cases} -\frac{\partial p^{i,2}}{\partial t} - \nabla \cdot (\mu_i \nabla p^{i,2}) = [D_{\mathbf{u}}\mathbf{A}(t,x,\widetilde{\mathbf{u}}^2)^T \mathbf{p}^2]_i & in \;\; Q_2, \\ \frac{\partial p^{i,2}}{\partial n} = 0 & on \;\; \Sigma_2, \\ \mathbf{p}^2(T^2) = \begin{bmatrix} 0 \\ \frac{\gamma}{\text{meas}(\Omega_2)} \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{\lambda^2}{\text{meas}(\Omega_2)} \\ 0 \\ 0 \\ \frac{\lambda^4}{\text{meas}(\Omega_2)} \end{bmatrix} & in \;\; \Omega_2, \end{cases} \tag{3.8}$$

$$\begin{cases} -\frac{\partial p^{i,1}}{\partial t} - \nabla \cdot (\mu_i \nabla p^{i,1}) = [D_{\mathbf{u}}\mathbf{A}(t,x,\widetilde{\mathbf{u}}^1)^T \mathbf{p}^1]_i & in \;\; Q_1, \\ \frac{\partial p^{i,1}}{\partial n} = 0 & on \;\; \Sigma_1, \\ \mathbf{p}^1(T^1) = \frac{1}{\text{meas}(\Omega_1)}\mathbf{M}_2(\mathbf{p}^2(0)) + \begin{bmatrix} \frac{\lambda^1}{\text{meas}(\Omega_1)} \\ 0 \\ 0 \\ \frac{\lambda^3}{\text{meas}(\Omega_1)} \end{bmatrix} & in \;\; \Omega_1. \end{cases} \tag{3.9}$$

3.3 The Numerical Problem

This final section is devoted to the computation of a numerical approximation to the unique (nonnegative and bounded) solution of the eutrophication system (3.1) in $Q = (0,T) \times \Omega$. In order to obtain it, we will use a first order implicit time discretization (based on a finite difference scheme) and a standard space discretization based on the Lagrange finite element method.

3.3.1 The Time Semi-Discretization

For the time semi-discretization we will consider a finite set of discrete times $\{t_n\}_{n=0}^{N_T} \subset [0,T]$ such that $t_0 = 0$, $t_{N_T} = T$, and $t_n - t_{n-1} = \Delta t$, $\forall n = 1,\ldots,N_T$, with a time step $\Delta t > 0$. Associated to the above set we construct the following time semi-discretization of the state system (3.1), where $\alpha = \frac{1}{\Delta t} > 0$:

- $\mathbf{u}_0 \in [L^\infty(\Omega)]^4$ such that $0 \le u_0^i(x) \le M$ a.e. $x \in \Omega$, $i = 1,\ldots,4$, given.
- $\forall n = 1,\ldots,N_T$, $\mathbf{u}_n \in [H^1(\Omega)]^4$ such that $0 \le u_n^i(x) \le C(\alpha, M)$ a.e. $x \in \Omega$, $i = 1,\ldots,4$, solution of the steady state problem:

$$\begin{cases} \alpha \mathbf{u}_n - \nabla \cdot (\Lambda_\mu \nabla \mathbf{u}_n) = \mathbf{A}(t_n, x, \mathbf{u}_n) + \alpha \mathbf{u}_{n-1} + \mathbf{g}(t_n) & \text{in} \;\; \Omega, \\ \frac{\partial \mathbf{u}_n}{\partial n} = 0 & \text{on} \;\; \partial\Omega, \end{cases} \tag{3.10}$$

where Λ_μ is a diagonal matrix with diagonal elements $(\mu_1, \mu_2, \mu_3, \mu_4)$.

By using fixed point techniques, we can easily prove that under the assumption on α,

$$\alpha > \max\{\|L\|_{L^\infty(Q)} - K_r - K_{mf}, C_{fz}K_z - K_{mz}\}, \tag{3.11}$$

there exists a constant $C(\alpha, M)$—depending only on α and M—such that the unique solution $\mathbf{u}_n \in [H^1(\Omega)]^4$ of (3.10) is nonnegative and bounded by $C(\alpha, M)$.

To deal with the nonlinear part $\mathbf{A}(t_n, x, \mathbf{u}_n)$ of the semi-discretized system (3.10), we propose for each discrete time $n = 1, \ldots, N_T$ a fixed point scheme of the following type.

For a given $\mathbf{u}_{n,k} = (u_k^1, u_k^2, u_k^3, u_k^4) \in [H^1(\Omega)]^4$, we compute $\mathbf{u}_{n,k+1} = (u_{k+1}^1, u_{k+1}^2, u_{k+1}^3, u_{k+1}^4) \in [H^1(\Omega)]^4$ obtained by the following algorithm.

1) First we compute $u_{k+1}^2 \in H^1(\Omega)$ as the solution of the boundary value problem:

$$\begin{cases} (\alpha + K_r + K_{mf})u_{k+1}^2 - \nabla \cdot (\mu_2 \nabla u_{k+1}^2) \\ \qquad + \; K_z \frac{u_k^3}{K_F + u_k^2} u_{k+1}^2 - L \frac{u_k^1}{K_N + u_k^1} u_{k+1}^2 = g^2 & \text{in } \Omega, \\ \frac{\partial u_{k+1}^2}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

2) Next we compute $u_{k+1}^3 \in H^1(\Omega)$ as the solution of the problem:

$$\begin{cases} (\alpha + K_{mz})u_{k+1}^3 - \nabla \cdot (\mu_3 \nabla u_{k+1}^3) \\ \qquad - \; C_{fz} K_z \frac{u_{k+1}^2}{K_F + u_{k+1}^2} u_{k+1}^3 = g^3 & \text{in } \Omega, \\ \frac{\partial u_{k+1}^3}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

3) Then we compute $u_{k+1}^4 \in H^1(\Omega)$ as the solution of the problem:

$$\begin{cases} (\alpha + K_{rd}\Theta^{\theta-20})u_{k+1}^4 - \nabla \cdot (\mu_4 \nabla u_{k+1}^4) \\ \qquad = K_{mf} u_{k+1}^2 + K_{mz} u_{k+1}^3 + g^4 & \text{in } \Omega, \\ \frac{\partial u_{k+1}^4}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

4) Finally we compute $u_{k+1}^1 \in H^1(\Omega)$ as the solution of the problem:

$$\begin{cases} \alpha u_{k+1}^1 - \nabla \cdot (\mu_1 \nabla u_{k+1}^1) + C_{nc} L \frac{u_{k+1}^2}{K_N + u_k^1} u_{k+1}^1 \\ \qquad = C_{nc} K_r u_{k+1}^2 + C_{nc} K_{rd} \Theta^{\theta-20} u_{k+1}^4 + g^1 & \text{in } \Omega, \\ \frac{\partial u_{k+1}^1}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

Again, under assumption (3.11) on α, this algorithm will be monotonic and convergent for any initial iterate that is nonnegative and bounded by $C(\alpha, M)$.

Several alternative techniques have also been tried for the treatment of the nonlinear term (fully explicit formulation of the second terms, defect correction principle, and so on), but all of them have shown a worse behavior in the numerical examples.

3.3.2 The Space Discretization

For the fully discretized formulation of the eutrophication system (3.1)—and due to the fact that we have used a first order time semi-discretization—we only present here a standard P_1-Lagrange finite element method (for details see, for instance, the classical monograph [ZT00]). However, Q_1-Lagrange and P_2-Lagrange finite elements have also been tested in the numerical examples with very satisfactory results.

Then, for the domain Ω (assumed to be polygonal), we consider a family of regular meshes $\{\mathcal{T}_h\}_{h\to 0}$. Associated to each mesh $\mathcal{T}_h$ we consider the finite-dimensional vector subspace $V_h \subset H^1(\Omega)$ given by

$$V_h = \{u_h \in \mathcal{C}^0(\overline{\Omega}) \ : \ u_{h|_T} \in P_1(T), \ \forall T \in \mathcal{T}_h\},$$

where $P_1(T)$ denotes the space of degree one polynomials on T.

If we denote N_h the number of nodes in the mesh $\mathcal{T}_h$, $\{b_j\}_{j=1}^{N_h}$ the set of nodes of the mesh $\mathcal{T}_h$, and $\{\phi_i\}_{i=1}^{N_h}$ the standard basis of the space V_h (i.e., $\phi_i(b_j) = \delta_{ji}, \ \forall i, j = 1, \ldots, N_h$), we have that any element $u_h \in V_h$ admits a unique representation in the basis $\{\phi_i\}_{i=1}^{N_h}$ in the following way:

$$u_h = \sum_{i=1}^{N_h} u_h(b_i)\phi_i = [u_h] \cdot [\phi], \tag{3.12}$$

with vectors $[u_h] = (u_h(b_1), \ldots, u_h(b_{N_h}))$ and $[\phi] = (\phi_1, \ldots, \phi_{N_h})$. So, if we define the following matrices and vectors:

$$\begin{array}{rcl}
[M_h] \in \mathbb{R}^{N_h \times N_h} & : & [M_h]_{ij} = \int_\Omega \phi_i \phi_j dx, \\
[R_h] \in \mathbb{R}^{N_h \times N_h} & : & [R_h]_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j dx, \\
[A_h^1(K, u_h, v_h)] \in \mathbb{R}^{N_h \times N_h} & : & [A_h^1(K, u_h, v_h)]_{ij} = \int_\Omega \frac{v_h}{K+u_h} \phi_i \phi_j dx, \\
[A_h^2(K, u_h, v_h)] \in \mathbb{R}^{N_h \times N_h} & : & [A_h^2(K, u_h, v_h)]_{ij} = \int_\Omega L \frac{v_h}{K+u_h} \phi_i \phi_j dx, \\
[B_h^k] \in \mathbb{R}^{N_h}, \ k = 1, \ldots, 4 & : & [B_h^k]_i = \int_\Omega g^k \phi_i dx,
\end{array}$$

for $i, j = 1, \ldots, N_h$, we can obtain the following full discretization of the fixed point algorithm introduced in the previous subsection.

For a given $\mathbf{u}_{h,n,k} = (u_{h,k}^1, u_{h,k}^2, u_{h,k}^3, u_{h,k}^4) \in [V_h]^4$, compute $\mathbf{u}_{h,n,k+1} = (u_{h,k+1}^1, u_{h,k+1}^2, u_{h,k+1}^3, u_{h,k+1}^4) \in [V_h]^4$ obtained by the following algorithm.

1) First we compute $u_{h,k+1}^2 \in V_h$ as the solution of the linear system:

$$\begin{aligned}
&\{(\alpha + K_r + K_{mf})[M_h] + \mu_2[R_h] + K_z[A_h^1(K_F, u_{h,k}^2, u_{h,k}^3)] \\
&\quad - \ [A_h^2(K_N, u_{h,k}^1, u_{h,k}^1)]\}[u_{h,k+1}^2] = [B_h^2].
\end{aligned}$$

2) Next we compute $u_{h,k+1}^3 \in V_h$ as the solution of the linear system:

$$\begin{aligned}
&\{(\alpha + K_{mz})[M_h] + \mu_3[R_h] \\
&\quad - \ C_{fz}K_z[A_h^1(K_F, u_{h,k+1}^2, u_{h,k+1}^2)]\}[u_{h,k+1}^3] = [B_h^3].
\end{aligned}$$

3) Then we compute $u^4_{h,k+1} \in V_h$ as the solution of the linear system:

$$\begin{aligned}&\{(\alpha + K_{rd}\Theta^{\theta-20})[M_h] + \mu_4[R_h]\}[u^4_{h,k+1}]\\ &\quad = K_{mf}[M_h][u^2_{h,k+1}] + K_{mz}[M_h][u^3_{h,k+1}] + [B^4_h].\end{aligned}$$

4) Finally we compute $u^1_{h,k+1} \in V_h$ as the solution of the linear system:

$$\begin{aligned}&\{\alpha[M_h] + \mu_1[R_h] + C_{nc}[A^2_h(K_N, u^1_{h,k}, u^2_{h,k+1})]\}[u^1_{h,k+1}]\\ &\quad = C_{nc}K_r[M_h][u^2_{h,k+1}] + C_{nc}K_{rd}\Theta^{\theta-20}[M_h][u^4_{h,k+1}] + [B^1_h].\end{aligned}$$

Once more, under assumption (3.11) on α, the above algorithm will be convergent. In particular, the constraint on α (or, equivalently, the corresponding constraint on Δt) ensures the positive definiteness of the matrices in the previous linear systems and, consequently, their solvability.

3.3.3 Numerical Example

In this final subsection we present the numerical results obtained for a realistic example consisting of two tanks: the first tank Ω_1 is a shallow tank of dimensions $16\,m \times 16\,m \times 4\,m$, and the second one Ω_2 is a deeper tank of dimensions $8\,m \times 8\,m \times 16\,m$. (We must remark that both tanks have different sizes, but the same capacities: 1024 m^3.) Permanence times will be the same for both tanks: $T^1 = T^2 = 200$ hours.

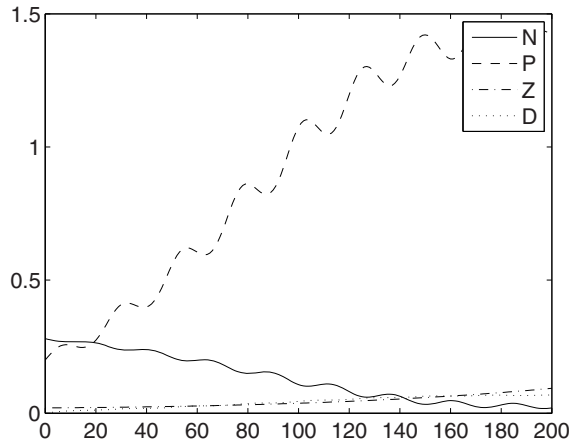


Fig. 3.1. Averaged concentrations of nitrogen (N), phytoplankton (P), zooplankton (Z), and organic detritus (D) in the first tank Ω_1.

The averaged concentrations (in mg/l) of the four species in both tanks are shown in Figures 3.1 and 3.2, obtained with our own code (completely developed by the authors in MATLAB and C++).

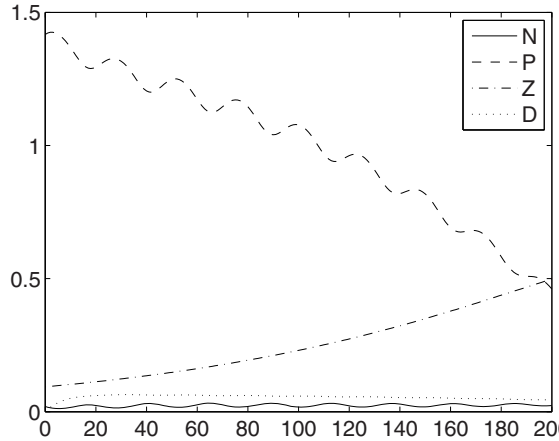


Fig. 3.2. Averaged concentrations of nitrogen (N), phytoplankton (P), zooplankton (Z), and organic detritus (D) in the second tank Ω_2.

In the first tank Ω_1 nitrogen decreases as phytoplankton increases (showing the typical oscillatory behavior due to cyclical night/day luminosity variations during the approximately 8-day period). A moderate increase in zooplankton (natural predator of phytoplankton) and organic detritus (produced by its death) can also be observed. We mention here that the final concentrations in the first tank are used as initial concentrations in the second tank (except for the organic detritus that is recovered from the water; its initial concentration will be considered null). The deeper shape of the second tank Ω_2—carrying a limitation on light availability—and the low level of nitrogen promote a decrease in phytoplankton concentration. Phytoplankton death causes an increase in organic detritus from zero up to a maximum level, until it begins to decrease due to decomposition (re-injecting nitrogen into the water column). Finally, in this second tank, we can see how the zooplankton concentration keeps growing, but the nitrogen level remains controlled.

References

[AFM09] Alvarez-Vázquez, L.J., Fernández, F.J., Muñoz-Sola, R.: Mathematical analysis of a three-dimensional eutrophication model. *J. Math. Anal. Appl.,* **349**, 135–155 (2009).

[Ca76] Canale, R.P.: *Modeling Biochemical Processes in Aquatic Ecosystems,* Ann Arbor Science Publishers, Ann Arbor, MI (1976).

[DCI01] Drago, M., Cescon, B., Iovenitti, L.: A three-dimensional numerical model for eutrophication and pollutant transport. *Ecological Modelling,* **145**, 17–34 (2001).

[ZT00] Zienkiewicz, O.C., Taylor, R.L.: *The Finite Element Method, Vol. 1,* Butterworth–Heinemann, London (2000).

4

An Analytical Solution for the Transient Two-Dimensional Advection–Diffusion Equation with Non-Fickian Closure in Cartesian Geometry by the Generalized Integral Transform Technique

D. Buske,[1] M.T. Vilhena,[2] D. Moreira,[3] and T. Tirabassi[4]

[1] Universidade Federal de Pelotas, Brazil; danielabuske@gmail.com
[2] Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil; vilhena@pq.cnpq.br
[3] Universidade Federal do Pampa, Bagé, RS, Brazil; davidson@pq.cnpq.br
[4] Istituto di Scienze dell'Atmosfera e del Clima, Bologna, Italy; t.tirabassi@isac.cnr.it

4.1 Introduction

Analytical solutions of equations are of fundamental importance in understanding and describing physical phenomena, since they are able to take into account all the parameters of a problem and investigate their influence. In a recent work, [Bus07] reported an analytical solution for the stationary two-dimensional advection–diffusion equation with Fickian closure by the Generalized Integral Laplace Transform Technique (GILTT). The main idea of this method consists of: construction of an auxiliary Sturm–Liouville problem, expansion of the contaminant concentration in a series in terms of the obtained eigenfunctions, replacement of the expansion in the original equation, and finally after taking moments, resulting a set of ordinary differential equations which are then solved analytically by the Laplace transform technique.

In this chapter, pursuing the task of searching analytical solutions, we start by presenting an analytical solution for the transient two-dimensional advection–diffusion equation with non-Fickian closure in Cartesian geometry by the GILTT method. We specialize the application of this methodology to the simulation of pollutant dispersion in the planetary boundary layer (PBL) under low wind conditions. We also present numerical results and comparison with experimental data.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_4,

4.2 The Analytical Solution

The advection–diffusion equation of air pollution in the atmosphere is essentially a statement of conservation of a suspended material, and it can be written as

$$\frac{\partial \overline{c}}{\partial t} + \overline{u}\,\frac{\partial \overline{c}}{\partial x} + \overline{v}\,\frac{\partial \overline{c}}{\partial y} + \overline{w}\,\frac{\partial \overline{c}}{\partial z} = -\frac{\partial \overline{u'c'}}{\partial x} - \frac{\partial \overline{v'c'}}{\partial y} - \frac{\partial \overline{w'c'}}{\partial z} + S, \tag{4.1}$$

in which $\overline{c}$ denotes the mean concentration of a passive contaminant (g/m^3) and $\overline{u}$, $\overline{v}$ and $\overline{w}$ are the Cartesian components of the mean wind (m/s) in the directions x ($-\propto < x < \propto$), y ($-\propto < y < \propto$), and z ($0 < z < h$). The terms $\overline{u'c'}$, $\overline{v'c'}$, $\overline{w'c'}$ are, respectively, the contaminant turbulent fluxes (g/sm^2) in the longitudinal, lateral, and vertical directions and S is the source term.

One of the most widely used closures for equation (4.1) is based on the gradient transport hypothesis which, by analogy with molecular diffusion, assumes that turbulence causes a net movement of material down the gradient of material concentration at a rate which is proportional to the magnitude of the gradient:

$$\overline{u'c'} = -K_x\,\frac{\partial \overline{c}}{\partial x};\ \overline{v'c'} = -K_y\,\frac{\partial \overline{c}}{\partial y};\ \overline{w'c'} = -K_z\frac{\partial \overline{c}}{\partial z},$$

where K_x, K_y, and K_z are the Cartesian components of the turbulent diffusion coefficients (m^2/s). In the first-order closure of the turbulence all the information of the turbulence complexity is contained in the eddy diffusivity (see the work [Bus07] for the solution of (4.1) with first-order closure).

To take into account the nonhomogeneous character of the turbulence in the convective boundary layer (CBL), [Ert42] and [Dea66] proposed to modify the usual application of the flux gradient in the K-theory approximation in such a way that

$$\overline{w'c'} = -K_z\left(\frac{\partial \overline{c}}{\partial z} - \gamma\right), \tag{4.2}$$

where γ represents the countergradient term. In the literature we can find many parameterizations for γ and, in this chapter, without losing generality, we used that proposed in [Van01]:

$$\left(1 + \left(\frac{S_k \sigma_w T_{l_w}}{2}\right)\frac{\partial}{\partial z} + \tau\frac{\partial}{\partial t}\right)\overline{w'c'} = -K_z\frac{\partial \overline{c}}{\partial z}, \tag{4.3}$$

where S_k is the skewness, σ_w is the vertical turbulent velocity variance (m/s), T_{l_w} is the vertical Lagrangian time scale (s) and τ is the relaxation time (s). Using equations (4.2) and (4.3), the turbulence closure problem was solved without obeying Fick's law, which is called non-Fickian closure. The non-Fickian closure allows the investigation of more energy eddies at different heights and the effect of the asymmetric transport in the computation of

the pollutant concentration, considering in a more complete way the complex structure of the turbulent dispersion.

Considering the Eulerian framework for a Cartesian coordinate system in which the x-direction coincides with that of the average wind and substituting the above equations in (4.1), we write the crosswind integrated transient advection–diffusion equation in the form

$$\begin{aligned}\frac{\partial c(x,z,t)}{\partial t}+\overline{u}\frac{\partial c(x,z,t)}{\partial x}+\overline{w}\frac{\partial c(x,z,t)}{\partial z}&=\frac{\partial}{\partial x}\left(K_x\frac{\partial c(x,z,t)}{\partial x}\right)\\ +\frac{\partial}{\partial z}\left(K_z\frac{\partial c(x,z,t)}{\partial z}\right)&-\frac{\partial}{\partial z}\left(\beta\frac{\partial c(x,z,t)}{\partial t}\right)-\frac{\partial}{\partial z}\left(\beta\overline{u}\frac{\partial c(x,z,t)}{\partial x}\right)\\ -\frac{\partial}{\partial z}\left(\beta\overline{w}\frac{\partial c(x,z,t)}{\partial z}\right)&-\tau\frac{\partial^2 c(x,z,t)}{\partial t^2}-\frac{\partial}{\partial t}\left(\tau\overline{u}\frac{\partial c(x,z,t)}{\partial x}\right)\\ -\frac{\partial}{\partial t}\left(\tau\overline{w}\frac{\partial c(x,z,t)}{\partial z}\right)&+\frac{\partial}{\partial z}\left(\beta\frac{\partial}{\partial x}\left(K_x\frac{\partial c(x,z,t)}{\partial x}\right)\right)\\ &+\frac{\partial}{\partial t}\left(\tau\frac{\partial}{\partial x}\left(K_x\frac{\partial c(x,z,t)}{\partial x}\right)\right),\end{aligned}\tag{4.4}$$

with $\beta = 0.5S_k\sigma_w T_{l_w}$. Equation (4.4) is subjected to the null flux concentration at the ground and at the top of the boundary layer as well initial condition $\overline{u}c(0,z,t) = Q\delta(z-H_s)$ and $\frac{\partial c(L_*,z,t)}{\partial x} = 0$ far away from the source. Q is the emission rate (g/s), h the height of the CBL (m), H_s the height of the source (m), and δ represents the Dirac delta function.

To solve problem (4.4), we apply the Laplace transformation with respect to the time variable:

$$\begin{aligned}r\,\overline{C}(x,z,r)+\overline{u}\frac{\partial\overline{C}(x,z,r)}{\partial x}+\overline{w}\frac{\partial\overline{C}(x,z,r)}{\partial z}&=\frac{\partial}{\partial x}\left(K_x\frac{\partial\overline{C}(x,z,r)}{\partial x}\right)\\ +\frac{\partial}{\partial z}\left(K_z\frac{\partial\overline{C}(x,z,r)}{\partial z}\right)&-\frac{\partial}{\partial z}\left(\beta\,\overline{u}\,\frac{\partial\overline{C}(x,z,r)}{\partial x}\right)\\ -\frac{\partial}{\partial z}\left(\beta\,r\,\overline{C}(x,z,r)\right)-\frac{\partial}{\partial z}\left(\beta\,\overline{w}\,\frac{\partial\overline{C}(x,z,r)}{\partial z}\right)&-\tau\,r^2\,\overline{C}(x,z,r)\\ -\tau\,\overline{w}\,r\,\frac{\partial\overline{C}(x,z,r)}{\partial z}&+\tau\,r\,\frac{\partial}{\partial x}\left(K_x\frac{\partial\overline{C}(x,z,r)}{\partial x}\right)\\ -\tau\,\overline{u}\,r\,\frac{\partial\overline{C}(x,z,r)}{\partial x}&+\frac{\partial}{\partial z}\left(\beta\frac{\partial}{\partial x}\left(K_x\frac{\partial\overline{C}(x,z,r)}{\partial x}\right)\right),\end{aligned}\tag{4.5}$$

where $\overline{C}(x,z,r) = \mathfrak{L}\{c(x,z,t); t \to r\}$ and r is complex. Remember that in this chapter the terms u and w are functions of height z, and K_x and K_z are also functions of distance x.

Following the works of [Bus07] and [Bus07a], we pose that the solution of problem (4.5) has the form

$$\overline{C}(x,z,r) = \sum_{n=0}^{N} \overline{c_n}(x,r)\,\zeta_n(z), \tag{4.6}$$

where $\zeta_n(z)$ are the eigenfunctions of the associated Sturm–Liouville problem ($\zeta_n(z) = \cos(\lambda_n z)$ where $\lambda_n = n\pi/h$ (n =0,1,2,...) are the eigenvalues) and the dependent variable of the problem $\overline{c_n}(x,r)$ needs to be encountered.

By substituting equation (4.6) in (4.5) and taking moments, we obtain an ordinary differential equation with variable coefficients (because the eddy diffusivity depends on the x and z variables). Taking an average on the x variable (performing a stepwise approximation), we can rewrite the resultant equation in matrix form as

$$Y''(x,r) + F \cdot Y'(x,r) + G \cdot Y(x,r) = 0, \tag{4.7}$$

where $Y(x,r)$ is the column vector whose components are $\{\overline{c_n}(x,r)\}$. The matrix F is defined as $F = B_1^{-1}B_2$ and the matrix G as $G = B_1^{-1}B_3$. The entries of matrices B_1, B_2, and B_3 are, respectively, given by

$$\begin{aligned}(b_1)_{n,m} = &\int_0^h K_x\,\zeta_n(z)\,\zeta_m(z)dz + \int_0^h \beta K_x\,\zeta_n'(z)\,\zeta_m(z)dz \\ &+ \int_0^h (\beta K_x)'\,\zeta_n(z)\,\zeta_m(z)dz + \tau\, r \int_0^h K_x\,\zeta_n(z)\,\zeta_m(z)dz,\end{aligned}$$

$$\begin{aligned}(b_2)_{n,m} = &-\int_0^h \overline{u}\,\zeta_n(z)\,\zeta_m(z)dz - \int_0^h \beta\,\overline{u}\,\zeta_n'(z)\,\zeta_m(z)dz \\ &- \int_0^h (\beta\,\overline{u})'\,\zeta_n(z)\,\zeta_m(z)dz + \int_0^h K_x'\,\zeta_n(z)\,\zeta_m(z)dz \\ &+ \int_0^h \beta K_x'\,\zeta_n'(z)\,\zeta_m(z)dz + \int_0^h (\beta K_x')'\,\zeta_n(z)\,\zeta_m(z)dz \\ &- \tau\, r \int_0^h \overline{u}\,\zeta_n(z)\,\zeta_m(z)dz + \tau\, r \int_0^h K_x'\,\zeta_n(z)\,\zeta_m(z)dz,\end{aligned}$$

and

$$\begin{aligned}(b_3)_{n,m} = &\int_0^h K_z'\,\zeta_n'(z)\,\zeta_m(z)dz - \lambda_n^2 \int_0^h K_z\,\zeta_n(z)\zeta_m(z)dz \\ &- \int_0^h \overline{w}\,\zeta_n'(z)\,\zeta_m(z)dz - r \int_0^h \zeta_n(z)\,\zeta_m(z)dz \\ &- r \int_0^h \beta\,\zeta_n'(z)\,\zeta_m(z)dz - r \int_0^h \beta'\,\zeta_n(z)\,\zeta_m(z)dz \\ &- \tau\, r^2 \int_0^h \zeta_n(z)\,\zeta_m(z)dz + \lambda_n^2 \int_0^h \beta\,\overline{w}\,\zeta_n(z)\,\zeta_m(z)dz \\ &- \int_0^h (\beta\,\overline{w})'\,\zeta_n'(z)\,\zeta_m(z)dz - \tau\, r \int_0^h \overline{w}\,\zeta_n'(z)\,\zeta_m(z)dz.\end{aligned}$$

Applying the standard procedure of order reduction to equation (4.7), we obtain the result

$$Z'(x,r) + H.Z(x,r) = 0, \tag{4.8}$$

subjected to the boundary conditions

$$Z(0,r) = \frac{Q}{r}\,\zeta_m(H_s)A^{-1} \quad and \quad Z(L_*,r) = 0$$

where A^{-1} is the inverse of matrix A with entries $a_{n,m} = \int_0^h \overline{u}\,\zeta_n(Z)\,\zeta_m(Z)dZ$.

Following the procedure of [Bus07], one obtains the following solution for problem (4.8):

$$Z(x,r) = X\exp(Dx)X^{-1}Z(0) = M(x,r)\xi,$$

where $M(x,r) = X\exp(Dx)$ and $\xi = X^{-1}Z(0)$. X is the matrix of eigenvectors of matrix H and $\exp(Dx)$ is the diagonal matrix of the respective eigenvalues. For more details see the work [Mor08].

Once the coefficients of the series solution are determined, we are in position to invert the Laplace Transform solution using the fixed Talbot algorithm [VaAb04]:

$$c(x,z,t) = \frac{r_*}{M^*}\big[\frac{1}{2}\overline{C}(x,z,r)e^{rt} + \sum_{k=1}^{M^*-1}\mathrm{Re}[e^{tS(\theta_k)}\overline{C}(x,z,S(\theta_k))(1+i\sigma(\theta_k))]\big], \tag{4.9}$$

where $S(\theta_k) = r_*\theta(\cot\theta + i)$, $-\pi < \theta < \pi$, $\sigma(\theta_k) = \theta_k + (\theta_k\cot\theta_k - 1)\cot\theta_k$, $\theta_k = k\pi/M^*$, and $r_* = 2M^*/5t$ is a parameter based on numerical experiments. The control of the round-off error in the computation of (4.9) is specified by the accuracy requirement, i.e., the number of decimal digits accuracy (M).

4.3 Experimental Data and Turbulent Parameterization

In order to illustrate the aptness of the discussed formulation to simulate pollutant contaminant dispersion in the atmosphere, we evaluate the performance of the discussed solution against experimental ground-level concentration. The data used to evaluate the performance of the model in unstable conditions were constituted by a series of diffusion tests conducted at the Indian Institute of Technology (IIT Delhi) for surface releases, with light winds over flat, even terrain. The tracer used was SF_6 released from a height of 1 m and observed near the ground (0.5 m). In all cases the wind velocity was less than 2 ms^{-1} at a height of 15 m. For more details see the work [Sha96].

To compare the results obtained with the GILTT with the experimental data the time-dependent three-dimensional solution is written in terms of the

time-dependent two-dimensional solution (obtained in Section 4.2) multiplied by the Gaussian function in the y-direction as

$$\overline{c}(x,y,z,t) = c(x,z,t)\frac{e^{(-y^2/2\sigma_y^2)}}{\sqrt{2\pi}\sigma_y},$$

where σ_y is the lateral dispersion and is expressed as [Deg98]

$$\frac{\sigma_y^2}{h^2} = \frac{0.21}{\pi}\int_0^\infty \sin^2(2,26\psi^{1/3}X^* n')\frac{dn'}{(1+n')^{5/3}n'^2},$$

where $X^* = xw_*/\overline{u}h$ is the nondimensional distance, h the top of the convective boundary layer, n' is the nondimensional frequency, w_* the convective velocity scale, and $\psi = 0.4$ is the molecular dissipation of turbulent velocity.

To represent the near-source diffusion in weak winds, the eddy diffusivities should be considered as functions of not only turbulence but also of distance from the source [Ary95]:

$$K_\alpha = \frac{0.583 w_* h c_i \psi^{2/3}(z/h)^{4/3}X^*[0.55(z/h)^{2/3} + 1.03c_i^{1/2}\psi^{1/3}(f_m^*)_i^{2/3}X^*]}{[0.55(z/h)^{2/3}(f_m^*)_i^{1/3} + 2.06c_i^{1/2}\psi^{1/3}(f_m^*)_i X^*]^2}, \tag{4.10}$$

where $c_{v,w} = 0.36$, $c_u = 0.3$, and $(f_m^*)_i$ is the normalized frequency of the spectral peak independent of the stratification with $(f_m^*)_u = 0.67$ for the longitudinal component and $(f_m^*)_w = 0.55\left(\frac{z}{h}\right)\left[1 - \exp\left(-\frac{4z}{h}\right) - 0.0003\exp\left(\frac{8z}{h}\right)\right]^{-1}$ for the vertical component.

The expressions used to evaluate the term $\beta = S_k\sigma_w T_{l_w}$ are obtained in [Deg97]:

$$\sigma_w^2 = 1.06c_w\frac{\psi^{2/3}}{(f_m^*)_w^{2/3}}\left(\frac{z}{h}\right)^{2/3}w_*^2,$$

$$\psi = 1.5 - 1.2\left(\frac{z}{h}\right)^{1/3},\ \ T_{L_w} = \frac{0.55}{4}\frac{1}{\sigma_w}\frac{z}{(f_m^*)_w}.$$

The wind velocity profile was described by a power law expressed as follows [PD88]:

$$\frac{u_z}{u_1} = \left(\frac{z}{z_1}\right)^n,$$

where u_z and u_1 are the mean wind velocities at the heights z and z_1, while $n = 0.1$ under unstable conditions.

4.4 Numerical Results

We now specialize the application of this solution for the experimental data of IIT Delhi and we report the statistical numerical comparison in Table 4.1,

using the statistical evaluation procedure described by [Han89]. The statistical index FB indicates whether the predicted quantity underestimates or overestimates the observed ones. The statistical index NMSE represents the quadratic error of the predicted quantities related to the observed ones. The best results are expected to have values near zero for the indices NMSE, FB, and FS, and near 1 in the indices COR and FA2. In Table 4.1 we present the statistical results for four different simulations: Case 1, stationary problem considering $S_k = 0.0$; Case 2, stationary problem considering $S_k = 1.0$; Case 3, transient problem considering $S_k = 0.0$ with time of one hour; Case 4, transient problem considering $S_k = 1.0$ with time of one hour.

Table 4.1. Statistical results obtained with the GILTT method compared with the IIT Delhi experiment.

GILTT	NMSE	COR	FA2	FB	FS
Case 1	0.32	0.71	0.81	0.08	-0.11
Case 2	0.27	0.71	0.81	-0.07	-0.08
Case 3	0.33	0.71	0.94	-0.02	-0.21
Case 4	0.27	0.71	0.94	-0.01	-0.18

Taking a closer look at the results appearing in Table 4.1, we promptly note the good agreement between the results attained with those predicted from the experimental data.

4.5 Conclusions

To summarize, we stress the relevant aspect concerning the aptness of the proposed method to solve, analytically, the transient two-dimensional advection–diffusion equation either for Fickian or non-Fickian flow. By analytical, we mean that no approximation is made along the solution derivation. Here we recall the Cauchy–Kovalevsky theorem, which guarantees that the proposed solution is a solution of equation (4.1). From the previous discussion, we are confident in affirming that besides the elegance inherent in analytical solutions, the GILTT approach is a promising methodology to generate benchmark results, i.e., a relevant technique to validate computational codes. Furthermore, we must emphasize that this method can be directly applied to others fields of science, like heat and mass transfer problems, for instance. To complete the study of the capabilities of this technique, we focus our future attention on extending the application of this approach to three-dimensional, time-dependent, advection–diffusion equations.

References

[Ary95] Arya, S.P.: Modeling and parameterization of near-source diffusion in weak winds. *J. Appl. Meteor.,* **34**, 1112–1122 (1995).

[Bus07] Buske, D., Vilhena, M.T., Moreira, D.M., Tirabassi, T.: Simulation of pollutant dispersion for low wind conditions in stable and convective Planetary Boundary Layer. *Atmos. Environ.,* **41**, 5496–5501 (2007).

[Bus07a] Buske, D., Vilhena, M.T., Moreira, D.M., Tirabassi, T.: An analytical solution of the advection–diffusion equation considering non-local turbulence closure. *Environ. Fluid Mech.,* **7**, 43–54 (2007).

[Dea66] Deardorff, J.W.: The countergradient heat flux in the lower atmosphere and in the laboratory. *J. Atmospheric Sci.,* **23**, 503–506 (1966).

[Deg97] Degrazia, G.A., Campos Velho, H.F., Carvalho, J.C.: Nonlocal exchange coefficients for the convective boundary layer derived from spectral properties. *Cont. Atm. Phys.,* 57–64 (1997).

[Deg98] Degrazia, G.A., Mangia, C., Rizza U.: A comparison between different methods to estimate the lateral dispersion parameter under convective conditions. *J. Appl. Met.,* **37**, 227–231 (1998).

[Ert42] Ertel, H.: Der vertikale Turbulenz-wärmestrom in der Atmosphäre. *Meteor. Z.,* **59**, 250–253 (1942).

[Han89] Hanna, S.R.: Confidence limit for air quality models as estimated by bootstrap and jacknife resampling methods. *Atmos. Environ.,* **23**, 1385–1395 (1989).

[Mor08] Moreira, D.M., Vilhena, M.T., Buske, D., Tirabassi, T.: The state-of-the-art of the GILTT method to simulate pollutant dispersion in the atmosphere. *Atmospheric Research,* http://dx.doi.org/10.1016/j.atmosres.2008.07.004 (2008) (in press).

[PD88] Panofsky, A.H., Dutton, J.A.: *Atmospheric Turbulence,* Wiley, New York (1988).

[Sha96] Sharan, M., Singh, M.P., Yadav, A.K.: A mathematical model for the atmospheric dispersion in low winds with eddy diffusivities as linear functions of downwind distance. *Atmos. Environ.,* **30**, 1137–1145 (1996).

[VaAb04] Valkó, P.P., Abate, J.: Comparison of sequence accelerators for the Gaver method of numerical Laplace transform inversion. *Computers Math. Appl.,* **48**, 629–636 (2004).

[Van01] van Dop, H., Verver, G.S.: Countergradient transport revisited. *J. Atmos. Sci.,* **58**, 2240–2247 (2001).

5

A Numerical Solution of the Dispersion Equation of Guided Wave Propagation in N-Layered Media

J. Cardona,[1] P. Tabuenca,[1] and A. Samartin[2]

[1] Universidad de Cantabria, Spain; cardonaj@unican.es, tabuencp@orange.es
[2] Universidad Politécnica de Madrid, Spain; avelino.samartin@upm.es

5.1 Introduction

The study of wave propagation through elastic solid media can be used to carry out non-destructive tests (NDT) of structures. These tests can be used to detect and identify, in many cases, both the actual elastic properties and possible geometric imperfections included in the material damage of a structure [SaGa04]. One important application of high-frequency waves is the characterization of composite materials.

A composite material consists of several thin layers or laminae, and the resultant solid acts as a full plate. The layers can be of different materials, but normally the same material is used across the plate. Within the framework of the science of materials, one important issue is to design and estimate the mechanical properties of a composite material. There is an extensive literature on this topic (see [Jo75], [TsPa68], [Wh87], and [ViSi89]). The subject of composite materials optimal design is also of great interest and has been treated in [GuHa99].

In this chapter, the general theory of high-frequency wave propagation in layered media will be summarized and the dispersion equation will be obtained. The dispersion equation is presented as a result of a generalized nonlinear eigenvalue–eigenvector problem.

The chapter is organized as follows. In Section 5.2, the simple case $N = 1$, i.e., the well-known Lamb wave propagation is summarized. The dispersion curves obtained there are used to validate some results of the multi-layered general theory described in Section 5.3. Finally, in Sections 5.4 and 5.5, some computational procedures and examples are presented.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_5,

5.2 Lamb Guided Waves

An excellent summary of the general theory of Lamb wave propagation is given in [LaLi59]. A more detailed description is presented in the recent texts [RoDi00] and [Ro04]. Here only final results are shown.

Lamb waves correspond to a propagation of elastic waves throughout a homogenous and isotropic elastic infinite plate bounded by two parallel planes separated by a small distance $2h$. In this case, very often wave reflections along the faces of the plate occur, and therefore the propagation of the waves modifies its direction. The adopted system of coordinate axes is such that the equations of the plate free faces are $x_2 = \pm h$ and axes x_1, x_3 are contained in the plate middle plane.

According to the general theory of wave propagation, the displacement vector $\mathbf{u}$ of a material point can be derived from a potential scalar ϕ and a vector potential $\boldsymbol{\psi}$ as follows: $\mathbf{u} = \boldsymbol{\nabla}\phi + \boldsymbol{\nabla} \times \boldsymbol{\psi}$. In this expression, the two potentials fulfill the two wave equations.

It is assumed that Lamb waves travel along the x_1 axis, and diffraction in the x_3 direction is ignored. In the case of an isotropic and homogenous elastic solid, the scalar and vector potentials are trigonometric functions of time t with the same circular frequency ω. Then, they can be expressed in the following way, with k the wave number:

$$\phi = \phi_0(x_2)e^{i(\omega t - kx_1)} \quad \text{and} \quad \boldsymbol{\psi} = [\psi_{0j}(x_2)]e^{i(\omega t - kx_1)}, \quad j = 1, 2, 3. \tag{5.1}$$

A boundary value problem of the waves can be defined by the wave equations for each potential function and the boundary conditions $\sigma_{2i} = 0$, $i = 1, 2, 3$ on the free faces $x_2 = \pm h$. Lamb waves occur if the dispersion equation is satisfied, i.e., a relationship between the circular frequency ω and the wave number k. This dispersion equation, known as the Rayleigh–Lamb equation, is

$$\frac{\omega^4}{v_T^4} = 4kq^2 \left[1 - \frac{p\tan(ph + \alpha)}{q\tan(qh + \alpha)}\right] \quad \text{with} \quad \alpha = 0 \text{ and } \alpha = \frac{\pi}{2}, \tag{5.2}$$

where the wave constants p and q are $p^2 = \frac{\omega^2}{v_L^2} - k^2$ and $q^2 = \frac{\omega^2}{v_T^2} - k^2$ and the constant angle α can take the values 0 and $\frac{\pi}{2}$ depending on the type of symmetry of the Lamb wave, as will be discussed later.

If the relation (5.2) is satisfied, then the potential functions can be found but they are multiplied by an arbitrary constant factor. Once the functions $\phi(x_1, x_2, t)$ and $\boldsymbol{\psi}(x_1, x_2, t)$ are found, the displacements at time t of any material point (x_1, x_2) of the plate can be obtained up to a constant factor.

The equation (5.2) can be represented in the plane (ω, k) and then it defines a curve known as *dispersion curve*. On this curve three regions can be distinguished, according to the value of the phase velocity $V = \frac{\omega}{k}$. This value can be greater than the longitudinal wave velocity v_L (region 1), or it can lie between the velocity v_L and the transverse velocity velocity v_T (region 2),

or it can be smaller than v_T and therefore also than v_L (region 3). Then the wave constants can be written as $p^2 = \omega^2 \left(\frac{1}{v_L^2} - \frac{1}{V^2}\right)$, $q^2 = \omega^2 \left(\frac{1}{v_T^2} - \frac{1}{V^2}\right)$ and therefore the following boundaries for the regions of dispersion space can be defined:

- Region 1. $V > v_L > v_T$, i.e., $k < \frac{\omega}{v_L} < \frac{\omega}{v_T}$. Then p and q are both real.
- Region 2. $v_L > V > v_T$, i.e., $\frac{\omega}{v_L} < k < \frac{\omega}{v_T}$. Then q is real and p is imaginary.
- Region 3. $v_L > v_T > V$, i.e., $\frac{\omega}{v_L} < \frac{\omega}{v_T} < k$. Then p and q are both imaginary.

A particular case of special interest of Lamb waves corresponds to the values $q^2 = k^2$. Details of this special Lamb wave are given in [RoDi00].

5.3 Guided Waves in N-Layered Media

The objective is to obtain the dispersion curves, i.e., to find the wave number k for each angular frequency or pulsation ω of the wave propagation through an elastic solid of thickness H composed by N layers. From this result one can obtain the wavelength $\lambda = \frac{2\pi}{k}$ as well as the phase velocity $c = \frac{\lambda}{T}$, in which the period T is defined by $T = \frac{2\pi}{\omega} = \frac{1}{f}$ with frequency f.

5.3.1 General Equations

An infinite elastic solid is considered, bounded by two parallel horizontal planes separated a distance H (Figure 5.1). The thickness H of the solid is divided into a set of N layers. The layer n has a thickness h_n, and a Cartesian coordinate system $Ox_1x_2x_3$ is introduced. The following notation will be used: $H_n = \sum_{j=1}^{j=n} h_j$ and $H = H_N$. Each layer n is constituted by an isotropic elastic material of density ρ^n and elastic Lamé constants λ^n and μ^n. In the analysis of the wave propagation through the solid, the following variables for each layer n are data: h_n, λ^n, μ^n, and ρ^n. Therefore, the wave propagation longitudinal v_L^n and transversal v_T^n velocities are also data. The problem to be solved consists in computing the wave number k corresponding to each specified pulsation value ω. The resultant transcendental equation, relating k and ω, is known as the *dispersion equation,* and it is derived from the condition of existence of a nontrivial solution of a system of $4N$ simultaneous equations.

The dynamic equilibrium equations of layer n are

$$\mu^n \nabla^2 \mathbf{U}^n + (\lambda^n + \mu^n)\nabla(\nabla \cdot \mathbf{U}^n) = \rho^n \frac{\partial^2 \mathbf{U}^n}{\partial t^2} \tag{5.3}$$

and the stationary harmonic solution $\mathbf{u}(\mathbf{x})$ of these equations, with $\mathbf{x} = (x_1, x_2, x_3)$, is found if this solution is assumed to be expressed as $\mathbf{U}(\mathbf{x}, t) = \mathbf{u}(\mathbf{x})e^{i\omega t}$, in which case (5.3) becomes

$$\mu^n \nabla^2 \mathbf{u}^n + (\lambda^n + \mu^n)\nabla(\nabla \cdot \mathbf{u}^n) = \rho^n \omega^2 \mathbf{u}^n.$$

If the Helmholtz decomposition is introduced,

$$\mathbf{u}^n = \nabla \cdot \phi^n + \nabla \times \boldsymbol{\psi}^n, \tag{5.4}$$

and in each layer we assume that the displacement $u_1^n = 0$ and the other displacement components u_2^n and u_3^n are dependent only on x_2 and x_3, i.e., the following conditions are fulfilled:

$$\phi^n = \phi^n(x_2, x_3), \quad \psi_2^n = \psi_3^n = 0 \quad \text{and} \quad \psi_1^n = \psi^n(x_2, x_3),$$

then

$$u_1^n = 0, \quad u_2^n = \phi_{,2}^n + \psi_{,3}^n \quad \text{and} \quad u_3^n = \phi_{,3}^n - \psi_{,2}^n. \tag{5.5}$$

Substituting equations (5.5) into (5.3), the well-known uncoupled equations for longitudinal and transversal wave propagation are obtained:

$$\left[\nabla^2 - \frac{1}{(v_L^n)^2}\frac{\partial^2}{\partial t^2}\right]\phi^n = 0, \quad \left[\nabla^2 - \frac{1}{(v_T^n)^2}\frac{\partial^2}{\partial t^2}\right]\psi^n = 0,$$

in which

$$\nabla^2 \equiv \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}, \quad v_L^n = \sqrt{\frac{\lambda^n + 2\mu^n}{\rho^n}}, \quad v_T^n = \sqrt{\frac{\mu^n}{\rho^n}}.$$

5.3.2 Solution

The solution of the spatial part of equation (5.3) is

$$\phi^n = C_1^n \exp[ik_L^n(x_3 \sin\theta_L^n + x_2 \cos\theta_L^n)] + C_2^n \exp[ik_L^n(x_3 \sin\theta_L^n - x_2 \cos\theta_T^n)], \tag{5.6}$$

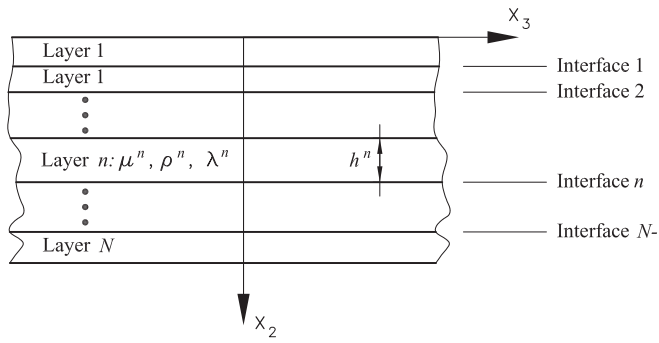


Fig. 5.1. N-layered elastic solid.

$$\psi^n = C_3^n \exp[ik_T^n(x_3 \sin\theta_T^n + x_2 \cos\theta_T^n)] + C_4^n \exp[ik_T^n(x_3 \sin\theta_T^n - x_2 \cos\theta_T^n)], \quad (5.7)$$

with $i = \sqrt{-1}$, $k_L^n = \frac{\omega}{v_L^n}$, and $k_T^n = \frac{\omega}{v_T^n}$.

Each of the two preceding expressions is a sum of two terms, one representing a downward propagating plane wave and the other representing an upward propagating term, according to the sign (positive or negative) of the exponential term x_2. The terms of the former expressions are called *partial waves* and are represented in Figures 5.1 and 5.2.

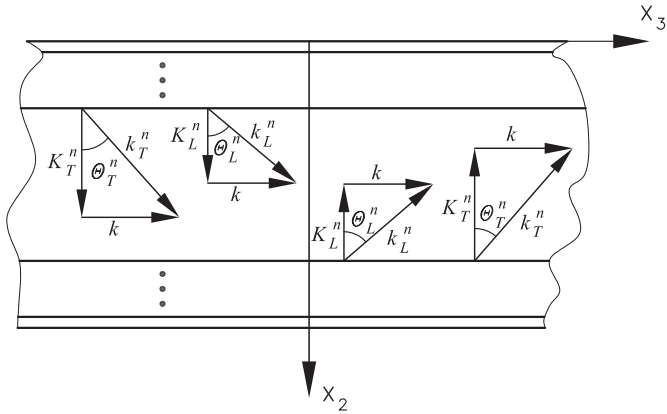


Fig. 5.2. Four partial waves in layer n.

The values of the arbitrary constants C_i^n with $i = 1, 2, 3, 4$ are found from the boundary conditions at $x_3 = 0$ and $x_3 = H$ in which $H = \sum_{n=1}^{N} h_n$ and from the continuity conditions between the $N - 1$ layer interfaces as well. In the following the different variables involved in these conditions are expressed in terms of the unknown solution.

- **Displacements.** The displacement expressions can be simplified if the following notation is introduced:

$$J_L^n = k_L \sin\theta_L^n, \quad K_L^n = k_L \cos\theta_L^n, \quad J_T^n = k_T \sin\theta_L^n, \quad K_T^n = k_T \cos\theta_L^n.$$

Substituting (5.6) and (5.7) into (5.4) the following expressions are obtained:

$$u_2^n = iK_L^n \left[F_{L1}^n C_1^n - F_{L2}^n C_2^n\right] + iJ_T^n \left[F_{T1}^n C_3^n + F_{T2}^n C_4^n\right], \quad (5.8)$$

$$u_3^n = iJ_L^n \left[F_{L1}^n C_1^n + F_{L2}^n C_2^n\right] + iK_T^n \left[F_{T1}^n C_3^n - F_{T2}^n C_4^n\right], \quad (5.9)$$

where

$$F_{Lj}^n = \exp[iJ_L^n x_3 + (-)^j iK_L^n x_2], \; F_{Tj}^n = \exp[iJ_T^n x_3 + (-)^j iK_T^n x_2], \quad j = 1, 2.$$

- **Strains.** The strains are found by using the formulae $\varepsilon = \frac{1}{2}\left(u_{i,j} + u_{j,i}\right)$ and taking into account expressions (5.8) and (5.9).
  It is convenient to obtain the expression of the volumetric deformation e^n defined as $e^n = \varepsilon_{11}^n + \varepsilon_{22}^n = \nabla \mathbf{u}^n$, i.e., according to [Ro04]:

$$e = -\left(k_L^n\right)^2 \left[C_1^n \exp(iJ_L^n x_3 + iK_L^n x_2) + C_2^n \exp(iJ_L^n x_3 - iK_L^n x_2)\right],$$

  where

$$J_L^n = k_L^n \sin\theta_L^n = k, \;\; J_T^n = k_T^n \sin\theta_T^n = k, \tag{5.10}$$

$$K_L^n = k_L^n \cos\theta_L^n = \sqrt{(k_L^n)^2 - k^2}, \;\; K_T^n = k_T^n \sin\theta_T^n = \sqrt{(k_T^n)^2 - k^2}. \tag{5.11}$$

- **Stresses.** The stresses acting on the faces of each layer are found from the Lamé constitutive equations as follows:

$$\sigma_{ij}^n = \lambda^n e^n \delta_{ij} + 2\mu^n \varepsilon_{ij}^n.$$

5.3.3 Boundary Conditions

From the displacement and stress expressions, the boundary conditions can be set up. In the following discussion it is assumed that each of the N layers is a solid, i.e., an intermediate liquid does not exist.[1]

- **Free upper face.** The free face is $(x_2 = 0, -\infty < x_3 < \infty)$ and the boundary conditions to be imposed are $\sigma_{22}^1 = \sigma_{23}^1 = 0$ and these equations can be written in the form [Ro04]:

$$\left[\lambda^1 \left(k_L^1\right)^2 + 2\mu^1 \left(K_L^1\right)^2\right] \left[C_1^1 + C_2^1\right] + 2\mu^1 k K_T^1 \left[C_3^1 - C_4^1\right] = 0,$$

$$-2k\mu^1 K_L^1 \left[C_1^1 - C_2^1\right] + \mu^1 \left[\left(K_T^1\right)^2 - k^2\right] \left[C_3^1 + C_4^1\right] = 0.$$

- **Layers n and $n+1$. Interface n.** The interface n is defined as $(x_2 = H_n, \quad -\infty < x_3 < \infty)$ with $H_n = \sum_{j=1}^n h_j$ and the conditions to be imposed are

$$u_2^n = u_2^{n+1}, \quad u_3^n = u_3^{n+1}, \quad \sigma_{22}^n = \sigma_{22}^{n+1}, \quad \sigma_{23}^n = \sigma_{23}^{n+1} \tag{5.12}$$

  with $n = 1, 2, \ldots, N-1$. The following notation is introduced for each layer n, with $H = H_n$:

$$\mathbf{F}_r^n = (a_r^{1n} F_1^n, a_r^{2n} F_2^n, a_r^{3n} F_3^n, a_r^{4n} F_4^n) \quad r = u_2,\, u_3,\, \sigma_{22},\, \sigma_{23},$$

$$\text{where } F_1^n = e^{iK_L^n H},\; F_2^n = e^{-iK_L^n H},\; F_3^n = e^{iK_T^n H},\; F_4^n = e^{-iK_T^n H}$$

[1] In the existence of a liquid layer between two solid layers, the boundary conditions on the liquid layer faces can be expressed by zero shear stresses and displacements and normal stress continuity.

and for the different values of r the coefficients a_r^{jn}, $j = 1, 2, 3, 4$ are

$$\begin{aligned}
r = u_2: &\quad a_r^{1n} = -a_r^{2n} = K_L^n,\ a_r^{3n} = a_r^{4n} = k \\
r = u_3: &\quad a_r^{1n} = a_r^{2n} = k,\ a_r^{3n} = -a_r^{4n} = K_T^n \\
r = \sigma_{22}: &\quad a_r^{1n} = a_r^{2n} = \lambda^n (k_L^n)^2 + 2\mu^n (K_L^n)^2,\ a_r^{3n} = -a_r^{4n} = 2\mu^n k K_T^n \\
r = \sigma_{23}: &\quad a_r^{1n} = -a_r^{2n} = -2\mu^n k K_L^n,\ a_r^{3n} = a_r^{4n} = \mu^n [(K_T)^2 - k^2].
\end{aligned}$$

With this notation the conditions (5.12) can be written for each index r as follows:

$$\mathbf{F}_r^n \mathbf{C}^n - \mathbf{F}_r^{n+1} \mathbf{C}^{n+1} = 0 \quad \text{for } r = u_2,\ u_3,\ \sigma_{22},\ \sigma_{23} \tag{5.13}$$

in which for $n = 1, 2, \ldots, N$ the vector of unknown constants is defined as

$$\mathbf{C}^n = \begin{pmatrix} C_1^n & C_2^n & C_3^n & C_4^n \end{pmatrix}^T$$

- **Layer N. Free lower face.** This surface is defined by the expression ($x_2 = H, \quad -\infty < x_3 < \infty$) and the boundary conditions to be imposed are $\sigma_{22}^N = \sigma_{23}^N = 0$, i.e., these resulting equations are similar to the ones corresponding to the free upper face.

5.3.4 Dispersion Equation

The former boundary and continuity conditions can be written in matrix form:

$$\mathbf{A}_{11}^0 \mathbf{C}^1 = \mathbf{0},\ \ldots,\ \mathbf{A}_{n,n} \mathbf{C}^n + \mathbf{A}_{n,n+1} \mathbf{C}^{n+1} = \mathbf{0},\ \ldots,\ \mathbf{A}_{N,N}^0 \mathbf{C}^N = \mathbf{0} \tag{5.14}$$

with $\mathbf{A}_{11}^0$ and $\mathbf{A}_{N,N}^0$ coefficient matrices of 2×4 dimension. The dimension of coefficient matrices $\mathbf{A}_{n,n}$ and $\mathbf{A}_{n,n+1}$ is 4×4. All elements of these matrices are functions of the problem data. The unknown to be found, for each specified circular frequency ω, is the wave number k. The remaining variables of the coefficients of the former matrices can be expressed as functions of the unknown k, according to equations (5.10) and (5.11).

The number of unknowns of the system of homogenous equations (5.14) is 2 for each end surface (lower and upper faces) and 4 for each interface; then the total number of unknowns is $2 \times 2 + 4(N - 1) = 4N$ plus four constants C_i^n for each layer n, i.e., the total number is $4N$. Therefore, the dimension of the system (5.14) is $4N \times 4N$ and it can be written as $\mathbf{AC} = \mathbf{0}$, in which $\mathbf{C} = (\mathbf{C}^{1T}, \mathbf{C}^{2T}, \ldots, \mathbf{C}^{NT})^T$ and the coefficients of the matrix $\mathbf{A}$ are found from the expressions (5.14).

In order for the solution of the system of homogenous equations (5.14) to be nontrivial, it is necessary that the determinant of matrix $\mathbf{A}$ be zero, i.e., the following *dispersion equation* must be satisfied:

$$\det(\mathbf{A}) = 0, \quad \text{that is } |\mathbf{A}(\omega, k, \lambda^n, \mu^n, h_n)| = 0. \tag{5.15}$$

In equation (5.15) the data are the constants λ^n, μ^n, h_n for each layer n, with $n = 1, 2, \ldots, N$. The wave numbers K_L^n and K_T^n can be expressed as functions of the unknowns k and ω, according to formulae (5.10) and (5.11). Then by solving equation (5.15) it is possible to compute for each circular frequency ω the infinite values of k, although a finite number of k are real values, i.e., nonimaginary numbers. Each pair of solutions defines the phase velocity $c_p = \frac{\omega}{k}$.

Several numerical procedures exist to find solutions k for each value of ω, and some of them will be discussed in a later section.

5.3.5 Results

By sweeping of the pair k and ω, it is possible to represent the frequency spectrum. In addition, the dispersion curves defined by $c_p = \frac{\omega}{k}$ as a function of the circular frequency ω is another result of interest. Once the pair of values, k_j and ω_j, have been obtained as a solution of the dispersion equation (5.15), it is necessary to compute the column vector $\mathbf{C} = \mathbf{C}_j$ of dimension $4N \times 1$ containing the values of the constants, assuming the system of equations (5.14) particularized for the values k_j and ω_j, i.e.,

$$\mathbf{C}_j = \left[\mathbf{C}_j^n\right] \quad \text{with} \quad \mathbf{C}_j^{nT} = \left[\begin{array}{cccc} C_{j1}^n & C_{j2}^n & C_{j3}^n & C_{j4}^n \end{array}\right], \quad n = 1, 2, 3, \ldots, N.$$

From the knowledge of the values of these constants, the following results of interest as a function of x_2 using the corresponding formulae can be found: (a) Displacement values $u_1(x_2, x_3)$ and $u_2(x_2, x_3)$; (b) values of the strains $\varepsilon_{22}(x_2, x_3)$, $\varepsilon_{33}(x_2, x_3)$, and $\varepsilon_{23}(x_2, x_3)$, where x_3, and t are parameters; (c) values of the stresses $\sigma_{22}(x_2, x_3)$, $\sigma_{33}(x_2, x_3)$, and $\sigma_{23}(x_2, x_3)$, where x_3 and t are parameters.

5.4 Numerical Solution of the Dispersion Curve

A general numerical procedure has been developed to build up the matrix $\mathbf{A}$ of complex elements, to solve the nonlinear equation (5.15), and to generate the dispersion curve $c_p = \frac{\omega}{k}$.

Algorithms with global convergence, i.e., those aiming to obtain an approximate value of the root, are used for an initial guess of an interval of approximated solutions [FoMa77]. Among them the bisection procedure has been used (if simple zeros are considered) and also the algorithm based on the change of slope detection and minimum value of the function $\det(\mathbf{A})$ (if multiple zeros are taken into consideration). The obtained values are then starting values for the local convergence algorithm. Due to the difficulty of obtaining an explicit derivative of the function, a method of linear interpolation has been selected using the secant procedure. The convergence order in this case for simple roots has been 1.618 [StBu80] but for multiple roots the

convergence order is deteriorated to become of order 1. However, in this last case, the Aitken convergence acceleration technique [MaFi98] can be applied. The algorithm of quadratic inverse interpolation (method of Muller) [MaFi98] has also been tested and has similar convergence results computationally but is more efficient than the Aitken method. However, this interpolation method demands three initial values to be applied.

Using the programming environment MATLAB, several subroutines have been written.

The main computational steps of this analysis are summarized as follows.

1. **Computation of Lamb modes**. For a given frequency, within the interval $(k1, k2)$ the graph $\det(A) - k$ is represented and the zeros of the dispersion equation can be computed. In the case of simple zeros, the bisection algorithm is efficient for obtaining a preliminary approximation of the root. In the case of a multiple zero, it is necessary to detect a slope change and also the minimum of the absolute value of the function $\det \mathbf{A}$ in order to find an approximation of the root.
2. **The selection of the propagation mode of interest**. The mode is selected by choosing a value close, obtained already in step 1, to the exact solution either for k (or c_p), and this value is used as a starting point for the generation program (program of local convergence) of the dispersion curve. In this case the secant method or an algorithm based on quadratic inverse interpolation (Muller [MaFi98], Dekker–Brent [FoMa77]) can be used.

5.5 Results Validation

A first validation example will be the simulation of a Lamb wave propagation through a plate of total thickness h considering this plate as a composite material composed of N layers of identical thickness $\frac{h}{N}$ and equal properties. The example uses a steel plate defined by $2h = 0.02$ m, $E = 1.962 \times 10^8$ MPa, $\nu = 0.3093$, and $\rho = 7.797$ t/m^3. A train of symmetric Lamb waves ($\alpha = 0$) is introduced with frequency $f = 200 \times 10^3$. Then from (5.1) the circular frequency is $\omega = 1256637.061$ rad/s and the velocities are $v_T = 3099.9248$ m/s and $v_L = 5889.5724$ m/s. The dispersion equation (5.2) can be used to find the wave number $k = 218.1658$ m^{-1} and the wave velocity $V = 5760.01$ m/sec. In this particular case, the Lamb wave is located in the region 2, i.e., p is an imaginary number and q is a real number. For comparative purposes, the plan dimensions of the plate are supposed to be very large in order to simulate a plane strain as is assumed in the Lamb wave propagation.

This plate has also been modeled as an N-layered plate, with $N = 2, 3, 4$, and 5 layers, and the analysis of the wave propagation through the multilayer plate has given the same result $k = 218.1658$ as the one found using the dispersion equation (5.2). The displacements of the plate subjected to

Lamb waves are computed assuming the values p and q are in region 2. These displacements are compared with the ones obtained using the expressions found from the analysis of multi-layered elements, once the constants $\mathbf{C}$ are computed, using formulae (5.8) and (5.9) for the region 2. The dispersion curves obtained for the former plate modeled as an N-layered plate with $N = 2, 3$, and 4 are shown in Figures 5.3–5.5.

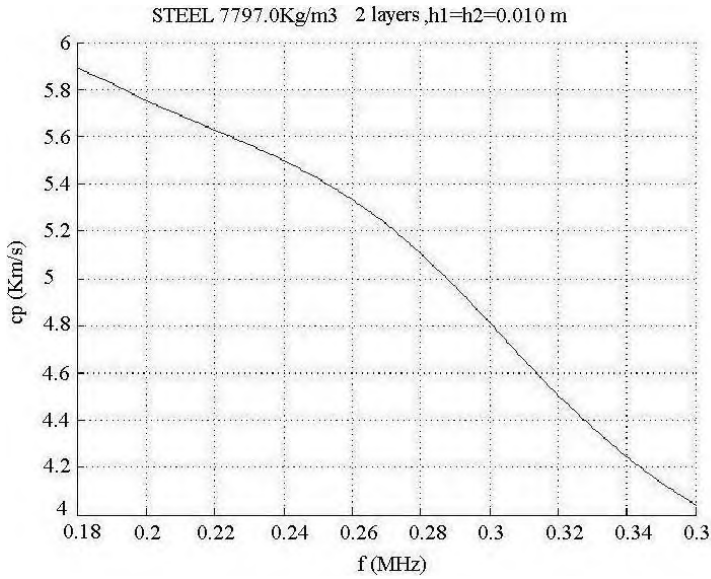


Fig. 5.3. Steel plate modeled as a 2-layered plate.

In Figure 5.6 the dispersion curve corresponding to an aluminum plate of thickness $h = 40$ mm, $\rho = 2698.4$ kg/m^3, and $v_L = v_T = 6300$ m/s is shown. In Figure 5.6, the dispersion curve for the former plate with an additional upper ice layer $h = 3$ mm is shown. The ice layer properties are $\rho = 917$ kg/m^3, $v_L = 3980$ m/s, and $v_T = 1990$ m/s. In the following Figure 5.7 the effect of the thickness variation of a steel plate is described. The plate is defined by $h = 20$ mm, $\rho = 7797$ kg/m^3, $v_L = 5889.57$ m/s, and $v_T = 3099.92$ m/s. Thickness variation curves for 1%, 5%, and 10% are given in the figure.

Acknowledgement. Financial support provided by the Spanish Ministry of Education and Science (Research contract DPI2005-09203-02) is gratefully acknowledged by the authors.

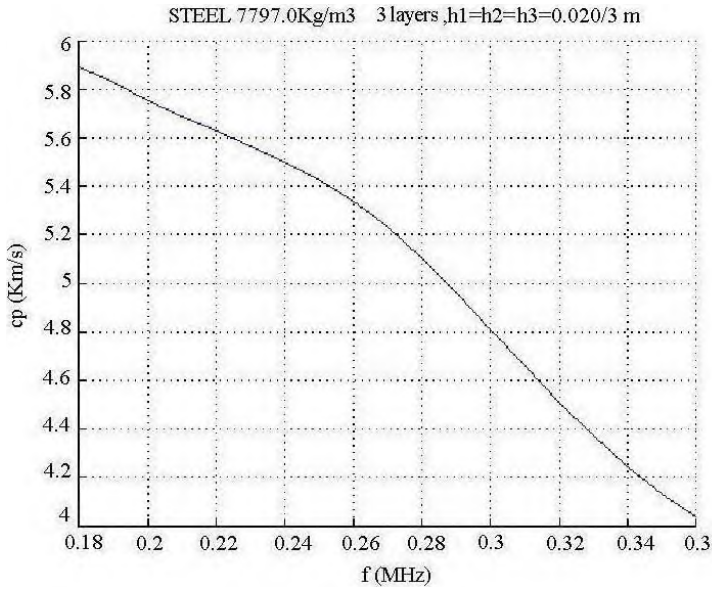


Fig. 5.4. Steel plate modeled as 3-layered plate.

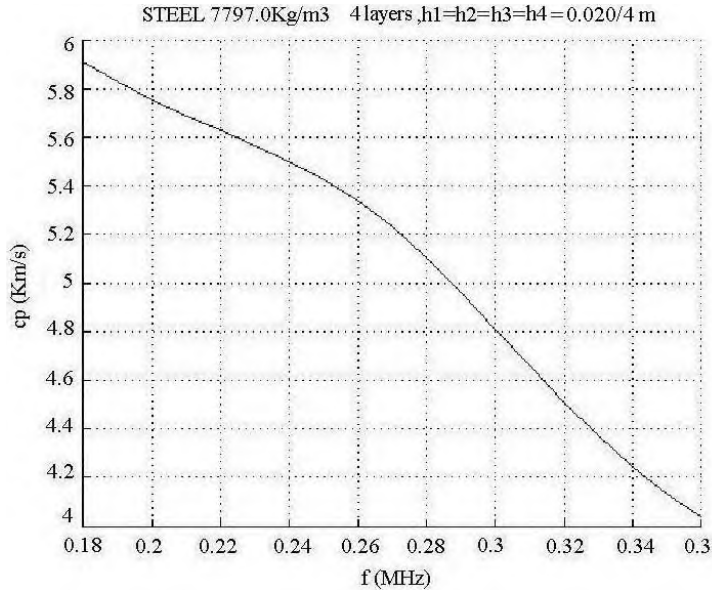


Fig. 5.5. Steel plate modeled as 4-layered plate.

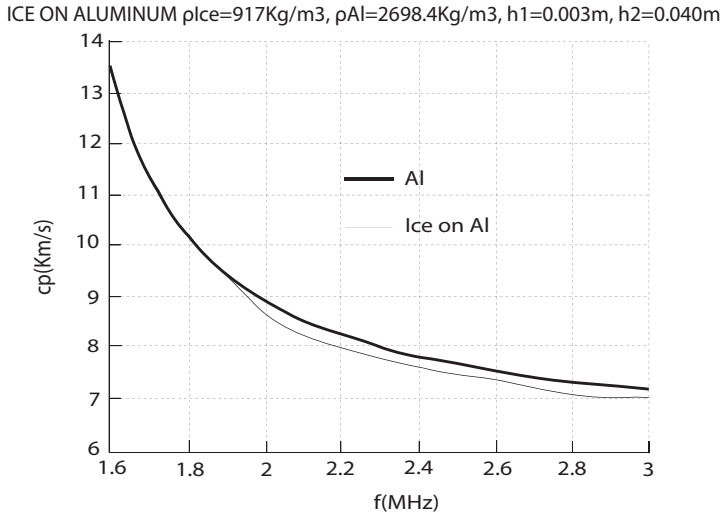


Fig. 5.6. Aluminum plate with $h = 40$ mm.

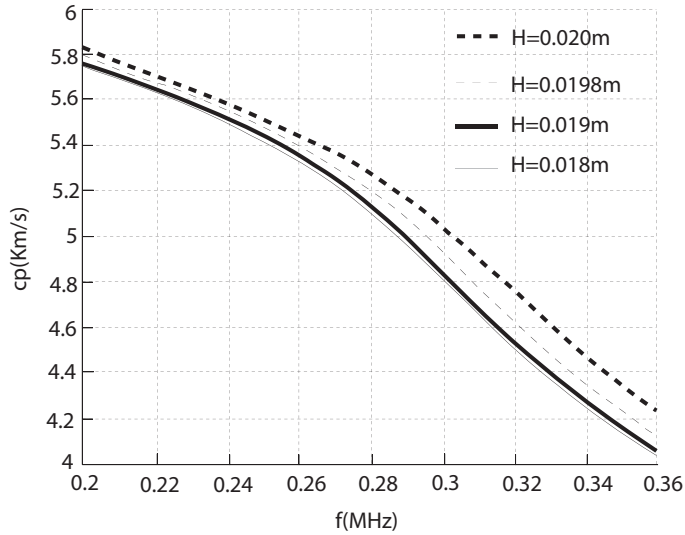


Fig. 5.7. Sensitivity of the steel dispersion curve to plate thickness.

References

[FoMa77] Forsythe, G.E., Malcolm, M.A., Moler, C.B.: *Computer Methods for Mathematical Computations,* Prentice-Hall, Englewood Cliffs, NJ (1977).

[GuHa99] Gurdal, Z., Haftka, R.T., Hajela, P.: *Design and Optimization of Laminated Composite Materials,* Wiley, Princeton University Press, Princeton, NJ (1999).

[Jo75] Jones, R.M.: *Mechanics of Composite Materials,* McGraw-Hill, New York (1975).

[LaLi59] Landau, L.D., Lifshitz, E.M.: *Theory of Elasticity, Vol. 7,* Pergamon Press, New York (1959).

[MaFi98] Mathews, K.D., Fink, K.D.: *Numerical Methods Using Matlab,* Prentice-Hall, Englewood Cliffs, NJ (1998).

[Ro04] Rose, J.L.: *Ultrasonic Waves in Solid Media,* Cambridge University Press, London (2004).

[RoDi00] Royer, D., Dieulesaint, E.: *Elastic Waves in Solids, Vols. I and II,* Springer, Berlin–Heidelberg (2000).

[SaGa04] Samartín, A., García-Palacios, J., Tabuenca, P.: Structural damage identification using dynamic numerical methods, in *Proceedings of the 2004 IASS Symposium on Shell and Spatial Structures from Models to Realization,* Montpellier, France (2004), 20–24.

[StBu80] Stoer, J., Bulirsch, R.: *Introduction to Numerical Analysis,* Springer, New York (1980).

[TsPa68] Tsai, S.W., Pagano, N.J.: Invariant properties of composite materials, in *Composite Materials Workshop,* Technomic, Westport, CT (1968), 233–253.

[ViSi89] Vinson, J.R., Sierakowski, R.L.: *The Behavior of Structures Composed of Composite Materials,* Martinus Nijhoff, Leiden, The Netherlands (1989).

[Wh87] Whitney, J.M.: *Structural Analysis of Laminate Anisotropic Plates,* Technomic, Westport, CT (1987).

6

Discretization of Coefficient Control Problems with a Nonlinear Cost in the Gradient

J. Casado-Díaz, J. Couce-Calvo, M. Luna-Laynez, and J.D. Martín-Gómez

Universidad de Sevilla, Spain; jcasadod@us.es, couce@us.es, mllaynez@us.es, jdmartin@us.es

6.1 Introduction

We consider a control problem of optimal design consisting in mixing two electric phases in order to minimize a given objective function. For simplicity, we assume that the two phases are isotropic, although the results still hold true for more general composites (see [CaEtAl08]). Mathematically the problem can be formulated as follows.

Let Ω be a bounded smooth open set in $\mathbb{R}^N$, $N \geq 2$, let α, β, κ be positive constants such that $\alpha < \beta$, $\kappa < |\Omega|$, and let f be in $L^2(\Omega)$. We look for a measurable set $\omega \subset \Omega$ with $|\omega| = \kappa$ such that the solution u of

$$\begin{cases} -\mathrm{div}(\alpha\chi_\omega + \beta\chi_{\Omega\setminus\omega})\nabla u = f \quad \text{in } \Omega, \\ u \in H_0^1(\Omega), \end{cases} \tag{6.1}$$

minimizes the functional

$$J(u) = \int_\Omega F(\nabla u)\, dx + G(u),$$

where $F : \mathbb{R}^N \to \mathbb{R}^N$ has a growth of order two at infinity and $G : H_0^1(\Omega) \to \mathbb{R}$ is sequentially continuous in the weak topology of $H_0^1(\Omega)$.

The constants α and β represent the conductivity of the two phases. The restriction $|\omega| = \kappa$ on the volume of the sets ω comes from the fact that usually one of the phases has better properties than the other one, but it is also more expensive and we can only use a limited quantity of it.

It is well known ([Mu71], [Mu72]) that this control problem does not have a solution in general. In fact, when we try to prove the existence of a solution using the direct method of the calculus of variations, two main difficulties appear. If we consider a minimizing sequence of sets $\omega_n \subset \Omega$, it is easy to prove that the corresponding solution u_n of (6.1) with ω replaced by ω_n is bounded in $H_0^1(\Omega)$. Therefore, there exists $u \in H_0^1(\Omega)$ such that, up to a

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_6,

subsequence, u_n converges weakly to u in $H_0^1(\Omega)$. However, in general there does not exist a measurable set $\omega \subset \Omega$ such that this weak limit u is the solution of a problem such as (6.1). Moreover, since we only have the weak convergence in $H_0^1(\Omega)$ of u_n to u and the cost functional depends nonlinearly on the gradient of the state, we cannot ensure that $J(u_n)$ is converging to $J(u)$.

From the above considerations, it follows that it is necessary to introduce a relaxation of the problem. For $F \equiv 0$, this relaxation is obtained by replacing the materials of the form $\alpha\chi_\omega + \beta\chi_{\Omega\setminus\omega}$ by the mixtures of α and β obtained by homogenization (see [MuTa85]). For a general F, it is proved in [CaCoMa08] that this relaxation is obtained using as above the materials constructed by homogenization, but taking instead of $F(\nabla u)$ a function $H(\nabla u, M\nabla u, \theta)$, where M is the homogenized matrix and θ the proportion of material α used in the mixture (related results can be found in [AlGu07], [BePe02], [Gr01], [LiVe02], [Pe06], and [Ta94] in the case where $F(\xi) = |\xi|^2$). This function $H : \mathbb{R}^N \times \mathbb{R}^N \times [0,1] \to (-\infty, +\infty]$ is defined in [CaCoMa08] by means of a minimization problem. We only have an explicit expression of H in the boundary of its domain $\mathcal{D}$, which makes it very difficult to deal with the relaxed problem.

The main goal of this chapter is to show how the solutions of the relaxed control problem can be numerically approximated replacing H by an upper or lower function which agrees with H on $\partial\mathcal{D}$, taking the proportions and the materials constant in the components of a partition of Ω whose diameter tends to zero, and solving the partial differential equations using the usual finite elements. The results given in this chapter are proved in [CaEtAl08], where the anisotropic case is also considered. In that paper, we also provide some numerical experiments.

6.2 Statement of the Problem and Prerequisites

Throughout the chapter, we consider a bounded open set $\Omega \subset \mathbb{R}^N$, $N \geq 2$, sufficiently smooth so that Meyer's theorem ([Me63]) holds, three positive constants α, β, κ such that $\alpha < \beta$ and $\kappa < |\Omega|$, a function $F : \mathbb{R}^N \to \mathbb{R}^N$ satisfying the existence of $L > 0$ such that

$$|F(\xi) - F(\xi')| \leq L(1 + |\xi| + |\xi'|)|\xi - \xi'|, \quad \forall\, \xi, \xi' \in \mathbb{R}^N,$$

a functional $G : H_0^1(\Omega) \to \mathbb{R}$ sequentially continuous for the weak topology of $H_0^1(\Omega)$, and a function $f \in L^2(\Omega)$. For these data, we are interested in the numerical solution of the coefficients control problem

$$\inf \left\{ \int_\Omega F(\nabla u)\, dx + G(u) \right\}, \tag{6.2}$$

where u is such that there exists $\omega \subset \Omega$ satisfying

$$\begin{cases} -\mathrm{div}(\alpha\chi_\omega + \beta\chi_{\Omega\setminus\omega})\nabla u = f \quad \text{in } \Omega, \\ u \in H_0^1(\Omega), \quad \omega \subset \Omega \text{ measurable, } |\omega| = \kappa. \end{cases} \tag{6.3}$$

Definition 1. *For $p \in [0,1]$, we define $\mathcal{K}(p)$ as the set of materials constructed by homogenization using the phase α with proportion p, and the phase β with proportion $1-p$.*

The set $\mathcal{K}(p)$ is characterized in [MuTa85] (see also [LuCh86]), where the following assertion is proved.

Theorem 1. *For $p \in [0,1]$, denoting λ_p^-, $\lambda_p^+ \in \mathbb{R}$ by*

$$\lambda_p^- = \left(\frac{p}{\alpha} + \frac{1-p}{\beta}\right)^{-1}, \quad \lambda_p^+ = p\alpha + (1-p)\beta,$$

we have

$$\mathcal{K}(p) = \Bigg\{ M \in \mathbb{R}^{N\times N} : \ M \text{ symmetric with eigenvalues } \lambda_1, \dots, \lambda_N \text{ satisfying}$$
$$\lambda_p^- \le \lambda_i \le \lambda_p^+, \ \forall i \in \{1,\dots,N\}, \ \sum_{i=1}^N \frac{1}{\lambda_i - \alpha} \le \frac{1}{\lambda_p^- - \alpha} + \frac{N-1}{\lambda_p^+ - \alpha},$$
$$\sum_{i=1}^N \frac{1}{\beta - \lambda_i} \le \frac{1}{\beta - \lambda_p^-} + \frac{N-1}{\beta - \lambda_p^+} \Bigg\}.$$

Definition 2. *For every $p \in [0,1]$, $\xi \in \mathbb{R}^N$, we denote*

$$\mathcal{K}(p)\xi = \big\{ M\xi : \ M \in \mathcal{K}(p) \big\} = \big\{ \eta \in \mathbb{R}^N : \ (\eta - \lambda_p^-\xi)\cdot(\eta - \lambda_p^+\xi) \le 0 \big\}.$$

We also define

$$\mathcal{D} = \big\{ (\xi, \eta, p) \in \mathbb{R}^N \times \mathbb{R}^N \times [0,1] : \ \eta \in \mathcal{K}(p)\xi \big\}.$$

The following result is proved in [CaCoMa08], and it gives the relaxation of problem (6.2)–(6.3).

Theorem 2. *There exists a continuous function $H : \mathcal{D} \to \mathbb{R}$, which only depends on α, β, and F such that a relaxation of problem (6.2)–(6.3) is given by*

$$\min \left\{ \int_\Omega H(\nabla u, M\nabla u, \theta)\, dx + G(u) \right\}, \tag{6.4}$$

with $(u, M, \theta) \in H_0^1(\Omega) \times L^\infty(\Omega)^{N\times N} \times L^\infty(\Omega)$ satisfying

$$\begin{cases} -\mathrm{div}\, M\nabla u = f \quad \text{in } \Omega \\ 0 \le \theta \le 1 \text{ a.e. in } \Omega, \quad \displaystyle\int_\Omega \theta\, dx = \kappa, \quad M \in \mathcal{K}(\theta) \quad \text{a.e. in } \Omega. \end{cases} \tag{6.5}$$

Assuming H extended by $+\infty$ outside $\mathcal{D}$, it satisfies the following properties:
(i) First,

$$\liminf_{n\to\infty}\int_\Omega H(\nabla u_n,\sigma_n,\theta_n)\,dx \geq \int_\Omega H(\nabla u,\sigma,\theta)\,dx,$$
$$\forall\,(u_n,\sigma_n,\theta_n) \text{ such that } u_n \rightharpoonup u \text{ in } H_0^1(\Omega), \quad (6.6)$$

$$\theta_n \overset{*}{\rightharpoonup} \theta \text{ in } L^\infty(\Omega),\ \sigma_n \rightharpoonup \sigma \text{ in } L^2(\Omega)^N,\quad \operatorname{div}\sigma_n \to \operatorname{div}\sigma \text{ in } H^{-1}(\Omega).$$

(ii) For every $(\xi,\eta,p)\in\mathcal{D}$ we have

$$|H(\xi,\eta,p)| \leq \frac{L\beta}{\alpha}|\xi|\left(1+\frac{\beta}{\alpha}|\xi|\right).$$

(iii) The value of H on the boundary of $\mathcal{D}$ is given by

$$H(\xi,\eta,p) = pF\left(\frac{\beta\xi-\eta}{(\beta-\alpha)p}\right)+(1-p)F\left(\frac{\eta-\alpha\xi}{(\beta-\alpha)(1-p)}\right),$$
$$\forall\,(\xi,\eta,p)\in\partial\mathcal{D},\ p\neq 0,1, \quad (6.7)$$

$$H(\xi,\alpha\xi,1) = H(\xi,\beta\xi,0) = F(\xi),\quad \forall\,\xi\in\mathbb{R}^N.$$

Moreover, if F is convex, then

$$H(\xi,\eta,p) \geq pF\left(\frac{\beta\xi-\eta}{(\beta-\alpha)p}\right)+(1-p)F\left(\frac{\eta-\alpha\xi}{(\beta-\alpha)(1-p)}\right),$$
$$\forall\,(\xi,\eta,p)\in\mathcal{D},\ p\neq 0,1.$$

Remark 1. The function H is defined in [CaCoMa08] using periodic homogenization. By Theorem 4.5 in [CaCoMa08], taking $O\subset\mathbb{R}^N$ open, bounded, and smooth, it can also be defined by

$$H(\xi,\eta,p) = \min\Big\{\liminf_{n\to\infty}\frac{1}{|O|}\int_O F(\nabla u_n)\,dx:\ u_n-\xi\cdot x \rightharpoonup 0 \text{ in } H_0^1(O),$$
$$\omega_n\subset O,\ \chi_{\omega_n}\overset{*}{\rightharpoonup} p \text{ in } L^\infty(O),\ \big(\alpha\chi_{\omega_n}+\beta\chi_{\Omega\setminus\omega_n}\big)\nabla u_n \rightharpoonup \eta \text{ in } L^2(O)^N,$$
$$-\operatorname{div}\big(\alpha\chi_{\omega_n}+\beta\chi_{\Omega\setminus\omega_n}\big)\nabla u_n = 0 \text{ in } O\Big\}, \quad (6.8)$$

for every $(\xi,\eta,p)\in\mathcal{D}$.

Remark 2. Restrictions (6.5) are equivalent to

$$\begin{cases} -\operatorname{div}\sigma = f \text{ in } \Omega \\ 0\leq\theta\leq 1 \text{ a.e. in } \Omega,\quad \displaystyle\int_\Omega \theta\,dx = \kappa,\quad \sigma\in\mathcal{K}(\theta)\nabla u \text{ a.e. in } \Omega, \end{cases} \quad (6.9)$$

in the sense that (u, σ, θ) satisfies (6.9) if and only if there exists $M \in \mathcal{K}(\theta)$ such that $\sigma = M\nabla u$ and thus (u, M, θ) satisfies (6.5). Thus, sometimes in this chapter we will say that $(\hat{u}, \hat{\sigma}, \hat{\theta})$ is a solution of (6.4)–(6.5) to mean that it gives a minimum of

$$\int_\Omega H(\nabla u, \sigma, \theta)\, dx + G(u),$$

with (u, σ, θ) satisfying (6.9).

6.3 Discrete Approximations

For H defined by (6.8), we consider another function $\overline{H} : \mathbb{R}^N \times \mathbb{R}^N \times [0,1] \to \mathbb{R}$ such that

$$\begin{cases} \overline{H} = +\infty \ \text{ in } (\mathbb{R}^N \times \mathbb{R}^N \times [0,1]) \setminus \mathcal{D}, \\ \overline{H}(\xi, \eta, p) \geq H(\xi, \eta, p), \quad \forall\, (\xi, \eta, p) \in \mathcal{D}, \\ \overline{H} \text{ is lower semicontinuous}, \\ \overline{H}(\xi, \alpha\xi, 1) = \overline{H}(\xi, \beta\xi, 0) = F(\xi). \end{cases} \tag{6.10}$$

Let us show how this function $\overline{H}$ can be used to solve numerically problem (6.4)–(6.5).

For $h > 0$, let us consider a partition of Ω given by $T_{j,h} \subset \Omega$, $1 \leq j \leq n_h$, such that

$$\begin{cases} \Omega = \displaystyle\bigcup_{j=1}^{n_h} T_{j,h}, \ |T_{j,h}| > 0, \ \operatorname{diam}(T_{j,h}) < h, \\ |T_{j,h} \cap T_{k,h}| = 0, \quad 1 \leq j, k \leq n_h, \ j \neq k, \end{cases} \tag{6.11}$$

and a closed subspace $V_h \subset H_0^1(\Omega)$, which satisfies the following properties:
(i) First,

$$\lim_{h\to 0} \min_{v_h \in V_h} \|v_h - v\|_{H_0^1(\Omega)} = 0, \quad \forall v \in H_0^1(\Omega). \tag{6.12}$$

(ii) Second,

$$\begin{gathered} \lim_{h\to 0} \min_{v_h \in V_h} \|v_h - w_h\varphi\|_{H_0^1(\Omega)} = 0, \quad \forall \varphi \in C_c^\infty(\Omega), \\ \forall w_h \in V_h \text{ bounded in } H_0^1(\Omega). \end{gathered} \tag{6.13}$$

(iii) Third,

$$\liminf_{h\to 0} \int_\Omega H(\nabla u_h, \sigma_h, \theta_h)\, dx \geq \int_\Omega H(\nabla u, \sigma, \theta)\, dx, \tag{6.14}$$

for every $(u_h, \sigma_h, \theta_h) \in V_h \times L^2(\Omega)^N \times L^\infty(\Omega)$ such that

$$\begin{cases} u_h \rightharpoonup u \text{ in } H_0^1(\Omega), \quad \theta_h \overset{*}{\rightharpoonup} \theta \text{ in } L^\infty(\Omega), \quad \sigma_h \rightharpoonup \sigma \text{ in } L^2(\Omega)^N, \\ \lim\limits_{h\to 0} \max\limits_{v_h \in V_h \setminus \{0\}} \dfrac{1}{\|v_h\|_{H_0^1(\Omega)}} \displaystyle\int_\Omega (\sigma_h - \sigma)\cdot \nabla v_h \, dx = 0. \end{cases} \tag{6.15}$$

With these definitions we consider the discrete control problem

$$\min \left\{ \int_\Omega \overline{H}(\nabla u, M\nabla u, \theta)\, dx + G(u) \right\}, \tag{6.16}$$

with $(u, M, \theta) \in V_h \times L^\infty(\Omega)^{N\times N} \times L^\infty(\Omega)$ satisfying

$$\begin{cases} \displaystyle\int_\Omega M\nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx, \quad \forall\, v \in V_h, \\ \displaystyle\int_\Omega \theta \, dx = \kappa, \quad 0 \le \theta \le 1, \ M \in \mathcal{K}(\theta) \ \text{ a.e. in } \Omega, \\ \theta, \ M \text{ constants in } \ T_{j,h}, \ 1 \le j \le n_h. \end{cases} \tag{6.17}$$

Under these assumptions, we have the following result.

Theorem 3. *We take $\overline{H}$, $T_{j,h}$, and V_h as above. Then problem (6.16)–(6.17) attains its minimum in some $(\hat{u}_h, \hat{M}_h, \hat{\theta}_h)$. The functions $(\hat{u}_h, \hat{M}_h\nabla\hat{u}_h, \hat{\theta}_h)$ are bounded in $H_0^1(\Omega)\times L^2(\Omega)^N \times L^\infty(\Omega)$ uniformly in h. Taking any subsequence of h (still denoted by h) such that $\hat{u}_h$ converges weakly in $H_0^1(\Omega)$ to some $\hat{u}$, $\hat{M}_h\nabla\hat{u}_h$ converges weakly in $L^2(\Omega)^N$ to some $\hat{\sigma}$, and $\hat{\theta}_h$ converges weakly-$*$ in $L^\infty(\Omega)$ to some $\hat{\theta}$, we get that $(\hat{u}, \hat{\sigma}, \hat{\theta})$ is a solution of (6.4)–(6.5). Moreover, we have*

$$\lim_{h\to 0} \left(\int_\Omega \overline{H}(\nabla\hat{u}_h, \hat{M}_h\nabla\hat{u}_h, \hat{\theta}_h)\, dx + G(\hat{u}_h) \right) = \int_\Omega H(\nabla\hat{u}, \hat{M}\nabla\hat{u}, \hat{\theta})\, dx + G(\hat{u}).$$

Remark 3. The most simple example of function $\overline{H}$ satisfying (6.10) is

$$\overline{H}(\xi, \alpha\xi, 1) = \overline{H}(\xi, \beta\xi, 0) = F(\xi) \ \ \forall \xi \in \mathbb{R}^N, \quad \overline{H}(\xi, \eta, p) = +\infty \text{ otherwise.}$$

However, it would be desirable to take $\overline{H}$ close to H. In this sense a choice of $\overline{H}$ which seems to be a lot better than the previous one is to take $\overline{H} = H$ in $\partial\mathcal{D}$ and $+\infty$ in another case. Recall that the expression of H in $\partial\mathcal{D}$ is given by (6.7).

Remark 4. Taking V_h as the whole space $H_0^1(\Omega)$ for every h (i.e., solving the state equation $-\operatorname{div} M\nabla u = f$ in Ω exactly), property (6.6) of H implies that V_h satisfies (6.14), while (6.12) and (6.13) are trivially satisfied. So, Theorem 3 shows the convergence of the numerical method consisting in taking in problem (6.4)–(6.5) the function H replaced by any $\overline{H}$ satisfying (6.10), and the controls θ, M constants in the elements of a partition $\{T_{j,h}\}$ of Ω satisfying (6.11). On the other hand, in the few cases where we know H,

property (6.14)–(6.15) is in fact satisfied for any sequence of closed subspaces $V_h \subset H_0^1(\Omega)$. So, in these cases only properties (6.12) and (6.13) must be verified. Indeed, assuming Ω a polyhedron, these properties hold for the usual Lagrange finite elements.

In order to solve numerically problem (6.4)–(6.5), in Theorem 3 we have replaced H by an upper function. Next we show an analogous result replacing H by a lower function.

We consider a function $\underline{H} : \mathbb{R}^N \times \mathbb{R}^N \times [0,1] \to (0,+\infty]$ which satisfies the following properties:

$$\underline{H} \text{ is continuous in } \mathcal{D}, \tag{6.18}$$

$$\underline{H}(\xi,\eta,p) \le H(\xi,\eta,p) \quad \forall\,(\xi,\eta,p) \in \mathcal{D}, \tag{6.19}$$

$$\underline{H}(\xi,\eta,p) = H(\xi,\eta,p) \quad \forall\,(\xi,\eta,p) \in \partial\mathcal{D}, \tag{6.20}$$

and for every $p \in (0,1)$, the function $(\xi,\eta) \in \mathbb{R}^N \times \mathbb{R}^N \mapsto \underline{H}(\xi,\eta,p)$ is Fréchet differentiable and there exists $c > 0$ such that

$$\begin{cases} |\partial_\xi \underline{H}(\xi,\eta,p)| + |\partial_\eta \underline{H}(\xi,\eta,p)| \le c\left(1 + |\xi| + |\eta| + \dfrac{|\beta\xi - \eta|}{p} + \dfrac{|\eta - \alpha\xi|}{1-p}\right), \\ \quad \forall\,(\xi,\eta,p) \in \mathbb{R}^N \times \mathbb{R}^N \times (0,1). \end{cases} \tag{6.21}$$

Theorem 4. *We consider $\underline{H} : \mathbb{R}^N \times \mathbb{R}^N \times [0,1] \to (0,+\infty]$ as above.*

For $h > 0$, let us consider a partition of Ω given by $T_{j,h} \subset \Omega$, $1 \le j \le n_h$, which satisfies (6.11). Assume also that there exists a sequence $V_h \subset H_0^1(\Omega)$ of closed subspaces satisfying (6.12), (6.13), and such that

$$\liminf_{h\to 0} \int_\Omega \underline{H}(\nabla u_h, \sigma_h, \theta_h)\,dx \ge \int_\Omega \underline{H}(\nabla u, \sigma, \theta)\,dx, \tag{6.22}$$

for every $(u_h, \sigma_h, \theta_h) \in V_h \times L^2(\Omega)^N \times L^\infty(\Omega)$ which satisfies (6.15).

We consider the discrete control problem

$$\min\left\{\int_\Omega \underline{H}(\nabla u, M\nabla u, \theta)\,dx + G(u)\right\}, \tag{6.23}$$

with $(u, M, \theta) \in V_h \times L^\infty(\Omega)^{N\times N} \times L^\infty(\Omega)$ satisfying

$$\begin{cases} \displaystyle\int_\Omega M\nabla u \cdot \nabla v dx = \int_\Omega f v dx, \quad \forall\, v \in V_h, \\ (M,\theta) \in \bar{co}(\{(M,p) \in \mathbb{R}^{N\times N} \times [0,1] :\ M \in \mathcal{K}(p)\}) \text{ a.e. in } \Omega, \\ \displaystyle\int_\Omega \theta dx = \kappa, \quad \theta,\ M \text{ constants in } T_{j,h},\ 1 \le j \le n_h. \end{cases} \tag{6.24}$$

Then problem (6.23)–(6.24) attains its minimum in some $(\hat{u}_h, \hat{M}_h, \hat{\theta}_h)$. The sequence $(\hat{u}_h, \hat{M}_h\nabla\hat{u}_h, \hat{\theta}_h)$ is bounded in $H_0^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega)$ uniformly

in h. Taking a subsequence of h (still denoted by h) such that $\hat{u}_h$ converges weakly in $H_0^1(\Omega)$ to some $\hat{u}$, $\hat{M}_h\nabla\hat{u}_h$ converges weakly in $L^2(\Omega)^N$ to some $\hat{\sigma}$ and $\hat{\theta}_h$ converges weakly-$*$ in $L^\infty(\Omega)$ to some $\hat{\theta}$, we have that $(\hat{u},\hat{\sigma},\hat{\theta})$ is a solution (see Remark 2) of

$$\min\left\{\int_\Omega \underline{H}(\nabla u, M\nabla u,\theta)\,dx + G(u)\right\},$$

with $(u,M,\theta)\in H_0^1(\Omega)\times L^\infty(\Omega)^{N\times N}\times L^\infty(\Omega)$ satisfying (6.5). Moreover,

$$\lim_{h\to 0}\left(\int_\Omega \underline{H}(\nabla\hat{u}_h,\hat{M}_h\nabla\hat{u}_h,\hat{\theta}_h)\,dx + G(\hat{u}_h)\right)=\int_\Omega \underline{H}(\nabla\hat{u},\hat{M}\nabla\hat{u},\hat{\theta})\,dx + G(\hat{u}).$$

Finally, if F is Fréchet differentiable and the solution q of

$$\begin{cases} -\mathrm{div}\,\hat{M}\nabla q = -\mathrm{div}\,Z & \text{in } \Omega,\\ q\in H_0^1(\Omega), \end{cases}$$

with

$$Z = \nabla F(\nabla\hat{u})\chi_{\{\hat{\theta}=0,1\}} + \big(\partial_\xi\underline{H}(\nabla\hat{u},\hat{M}\nabla\hat{u},\hat{\theta}) + \hat{M}\partial_\eta\underline{H}(\nabla\hat{u},\hat{M}\nabla\hat{u},\hat{\theta})\big)\chi_{\{0<\hat{\theta}<1\}},$$

satisfies

$$\partial_\eta\underline{H}(\nabla\hat{u},\hat{M}\nabla\hat{u},\hat{\theta})\neq\nabla q \quad \text{a.e. in } \Omega,$$

then $(\hat{u},\hat{M},\hat{\theta})$ is a solution of (6.4)–(6.5) and

$$\underline{H}(\nabla\hat{u},\hat{M}\nabla\hat{u},\hat{\theta}) = H(\nabla\hat{u},\hat{M}\nabla\hat{u},\hat{\theta}) \quad \text{a.e. in } \Omega.$$

The assumptions imposed on the function $\underline{H}$ are justified by the following result.

Proposition 1. *Assume F convex and Fréchet differentiable. For $(\xi,\eta,p)\in\mathbb{R}^N\times\mathbb{R}^N\times[0,1]$, we define $\underline{H}(\xi,\eta,p)$ by*

$$\begin{cases} pF\left(\dfrac{\beta\xi-\eta}{(\beta-\alpha)p}\right)+(1-p)F\left(\dfrac{\eta-\alpha\xi}{(\beta-\alpha)(1-p)}\right) & \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } (\xi,\eta,p)\in\mathcal{D},\ p\neq 0,1,\\ F(\xi) \quad \text{if} \quad p=0,\ \eta=\beta\xi \quad \text{or} \quad p=1,\ \eta=\alpha\xi,\\ +\infty \quad \text{otherwise.} \end{cases}$$

Then $\underline{H}$ satisfies the properties (6.18), (6.19), (6.20), and (6.21). Moreover, property (6.22) is satisfied by any sequence of closed subspaces $V_h\subset H_0^1(\Omega)$.

References

[AlGu07] Allaire, G., Gutiérrez, S.: Optimal design in small amplitude homogenization. *ESAIM:M2AN,* **41**, 543–574 (2007).

[BePe02] Bellido, J.C., Pedregal, P.: Explicit quasiconvexification for some cost functionals depending on derivatives of the state in optimal designing. *Discr. Contin. Dyn. Syst.,* **8**, 967–982 (2002).

[CaCoMa08] Casado-Díaz, J., Couce-Calvo, J., Martín-Gómez, J.D.: Relaxation of a control problem in the coefficients with a functional of quadratic growth in the gradient. *SIAM J. Control Optim.,* **47**, 1428–1459 (2008).

[CaEtAl08] Casado-Díaz, J., Couce-Calvo, J., Luna-Laynez, M., Martín-Gómez, J.D.: Optimal design problems for a non-linear cost in the gradient: numerical results. *Applicable Anal.,* **87**, 1461-1487 (2008).

[Gr01] Grabovsky. Y.: Optimal design for two-phase conducting composites with weakly discontinuous objective functionals. *Adv. Appl. Math.,* **27**, 683–704 (2001).

[LiVe02] Lipton, R., Velo., A.P.: Optimal design of gradient fields with applications to electrostatics, in *Nonlinear Partial Differential Equations and Their Applications, Vol. XIV,* Cioranescu, D., Lions, J.-L., eds., North-Holland, Amsterdam (2002), 509–532.

[LuCh86] Lurie, K.A., Cherkaev, A.V.: Exact estimates of the conductivity of a binary mixture of isotropic materials. *Proc. Roy. Soc. Edinburgh A,* **104**, 21–38 (1986).

[Me63] Meyers, N.G.: An L^p-estimate for the gradient of solutions of second order elliptic divergence equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.,* **17**, 189–206 (1963).

[Mu71] Murat, F.: Un contre-example pour le problème du contrôle dans les coefficients. *C.R. Acad. Sci. Paris A,* **273**, 708–711 (1971).

[Mu72] Murat, F.: Théorèmes de non existence pour des problèmes de contrôle dans les coefficients. *C.R. Acad. Sci. Paris A,* **274**, 395–398 (1972).

[MuTa85] Murat, F., Tartar, L.: Calculus of variations and homogenization, in *Topics in the Mathematical Modelling of Composite Materials,* Cherkaev, L., Kohn, R.V., eds., Birkhaüser, Boston (1998), 139–174.

[Pe06] Pedregal, P.: Optimal design in two-dimensional conductivity for a general cost depending on the field. *Arch. Rational Mech. Anal.,* **182**, 367–385 (2006).

[Ta83] Tartar, L.: Estimations fines de coefficients homogénéisés, in *Research Notes in Math.,* **125**, Pitman, London (1985), 168–187.

[Ta94] Tartar, L.: Remarks on optimal design problems, in *Calculus of Variations, Homogenization and Continuum Mechanics,* Buttazzo, G., Bouchitte, G., Suquet, P., eds., World Scientific, Singapore (1994), 279–296.

7

Optimal Control and Vanishing Viscosity for the Burgers Equation

C. Castro,[1] F. Palacios,[2] and E. Zuazua[3]

[1] Universidad Politécnica de Madrid, Spain; carlos.castro@upm.es
[2] Universidad Politécnica de Madrid, Spain; fpalacios@gmail.com
[3] Basque Center for Applied Mathematics, Bilbao, Spain; zuazua@bcamath.org

7.1 Introduction

We revisit an optimization strategy recently introduced by the authors to compute numerical approximations of minimizers for optimal control problems governed by scalar conservation laws in the presence of shocks. We focus on the one-dimensional (1-D) Burgers equation. This new descent strategy, called the *alternating descent method*, in the inviscid case, distinguishes and alternates descent directions that move the shock and those that perturb the profile of the solution away from it. In this chapter we analyze the optimization problem for the viscous version of the Burgers equation. We show that optimal controls of the viscous equation converge to those of the inviscid one as the viscosity parameter tends to zero and discuss how the alternating descent method can be adapted to this viscous frame.

Optimal control for hyperbolic conservation laws is a difficult topic which requires a considerable analytical effort and is computationally expensive in practice. In the last years a number of methods have been proposed to reduce the computational cost and to render this type of problem affordable.

In particular, recently, the authors have introduced the *alternating descent method*, which takes into account possible shock discontinuities. This chapter is devoted to revisit this method in the context of the *viscous Burgers equation*.

We focus on the 1-D Burgers equation although most of our results extend to more general equations with convex fluxes. Most of the ideas we develop here, although they need further developments at a technical level, apply to multi-dimensional scenarios, too.

To be more precise, given a finite time horizon $T > 0$, we consider the following inviscid Burgers equation:

$$\begin{cases} \partial_t u + \partial_x(\frac{u^2}{2}) = 0, & \text{in } \mathbb{R} \times (0,T), \\ u(x,0) = u^0(x), & x \in \mathbb{R}. \end{cases} \tag{7.1}$$

We also consider its viscous counterpart

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_7,

$$\begin{cases} \partial_t u - \nu u_{xx} + \partial_x(\frac{u^2}{2}) = 0, & \text{in } \mathbb{R} \times (0,T), \\ u(x,0) = u^0(x), \quad x \in \mathbb{R}, \end{cases} \tag{7.2}$$

where $\nu > 0$.

Given a target $u^d \in L^2(\mathbb{R})$ we consider the cost functional to be minimized $J : L^1(\mathbb{R}) \to \mathbb{R}$, defined by

$$J(u^0) = \int_{\mathbb{R}} |u(x,T) - u^d(x)|^2 \, dx, \tag{7.3}$$

where $u(x,t)$ is the unique entropy solution of (7.1) in the inviscid case or the unique weak solution of the viscous model (7.2). Sometimes, to make the dependence on the viscosity parameter ν more explicit, the functional J will be denoted by J_ν, although its definition is the same as that of J, but rather for the solutions u_ν of (7.2) instead of (7.1). Note that the functional above is well defined in both cases, inviscid and viscous, because of the effect on the gain of integrability of both equations: When the initial data belongs to $L^1(\mathbb{R})$, the solutions of both (7.1) and (7.2), for any positive time $t > 0$, belong to $L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$.

Although this chapter is devoted to this particular choice of J, most of our analysis can be adapted to many other functionals and control problems (we refer for instance to [JaSe99] and [CaZu08], where the control variable is the nonlinearity of the scalar conservation law).

We also introduce the set of admissible initial data $\mathcal{U}_{ad} \subset L^1(\mathbb{R})$, that we define later in order to guarantee the existence of minimizers for the following optimization problem: *Find* $u^{0,\min} \in \mathcal{U}_{ad}$ *such that*

$$J(u^{0,\min}) = \min_{u^0 \in \mathcal{U}_{ad}} J(u^0). \tag{7.4}$$

Similarly, we consider the same minimization problem for the viscous model (7.2): *Find* $u_\nu^{0,\min} \in \mathcal{U}_{ad}$ *such that*

$$J_\nu(u_\nu^{0,\min}) = \min_{u^0 \in \mathcal{U}_{ad}} J_\nu(u^0). \tag{7.5}$$

This is one of the model optimization problems that is often addressed in the context of optimal aerodynamic design, the inverse design problem (see, for example, [GiPi01]).

As we will see, the existence of minimizers for both models, the inviscid and the viscous one, is easily established under some natural assumptions on the class of admissible data $\mathcal{U}_{ad}$ using well-known well-posedness and compactness properties of the Burgers equation. However, uniqueness is false, in general.

The first result of this chapter is a Γ-convergence result guaranteeing that any sequence of minimizers $\{u_\nu^{0,\min}\}_{\nu>0}$, as $\nu \to 0$, has a minimizer $u^{0,\min}$ of J as an accumulation point.

Obviously, when $\nu > 0$, which is the common situation in practice, solutions are smooth; therefore, the alternating descent method, based on the fact

that solutions have shock discontinuities, cannot be applied as such. But for ν small enough, solutions present quasi-shock configurations. It is therefore natural to analyze how the method can be adapted to this situation to take advantage of the presence of these quasi-shocks.

The closely related issue of numerical approximations in the inviscid case has already been discussed in [CaPa08]. Indeed, in practical applications and in order to perform numerical computations and simulations, one has to replace the continuous optimization problems above by discrete ones. In particular, in what concerns the inviscid model (7.1), it is natural to consider a discretization of system (7.1) and the functional J. If this is done in an appropriate way, the discrete optimization problem has minimizers that are often taken, for small enough mesh sizes, as approximations of the continuous minimizers. This convergence result was proved in [CaPa08] in a suitable class of monotone finite difference schemes, satisfying the one-sided Lipschitz condition (OSLC). These schemes introduce artificial numerical viscosity. But the analysis in [CaPa08] showed that if, in the optimization process, the fact that discrete solutions may be close to shock configurations is not used, the discrete gradient algorithm shows a very slow convergence rate. Accordingly, the method proposed in [CaPa08] combines the discrete optimization approach and continuous shock analysis to derive the alternating descent method, which performs better. It is therefore natural to address the optimal control of the viscous model (7.2) similarly by viewing it as an approximation of the inviscid one (7.1) as $\nu \to 0$ and trying to take advantage of the quasi-shock configurations when they arise. This is precisely the goal of this chapter.

Our first result guarantees the convergence of the minimizers, based on the fact that the OSLC is satisfied uniformly with respect to the vanishing viscosity parameter, which ensures compactness.

The rest of this chapter is organized as follow. In Section 7.2 we recall the basic results in [CaPa08] on the existence of minimizers for the continuous problem (7.4). In Section 7.3 we analyze the convergence of the viscous minimizers as $\nu \to 0$. In Section 7.4 we recall some known results on the sensitivity of the continuous functional by linearizing system (7.1) in the presence of a shock. In Section 7.5 we briefly recall the *alternating descent method*. In Section 7.6 we develop an adaptation of the method of alternating descent directions to the viscous case. In Section 7.7 we present some numerical experiments that show the efficiency of the method we have developed.

7.2 Existence of Minimizers

In this section we prove that, under certain conditions on the set of admissible initial data $\mathcal{U}_{ad}$, there exists at least one minimizer of J and J_ν for all $\nu > 0$.

To simplify the presentation, we consider the class of admissible initial data $\mathcal{U}_{ad}$:

$$\mathcal{U}_{ad} = \{f \in L^\infty(\mathbb{R}),\ supp(f) \subset K,\ ||f||_\infty \leq C\},$$

where $K \subset \mathbb{R}$ is a bounded interval and $C > 0$ a constant. Obviously, $\mathcal{U}_{ad}$ as above is a bounded set of $L^1(\mathbb{R})$.

The analysis we shall develop here can be extended to a much wider class of admissible sets.

Theorem 1. *Assume that $\mathcal{U}_{ad}$ is as above and $u^d \in L^2(\mathbb{R})$. Then the minimization problems (7.4) and (7.5) have at least one minimizer $u^{0,\min} \in \mathcal{U}_{ad}$.*

Proof. The proof is simpler when $\nu > 0$ because of the regularizing effect of the viscous Burgers equation. But, in order to develop arguments that are uniform on the viscosity parameter ν, it is better to give a proof for the inviscid case, which applies in the viscous one as well. Thus, in what follows, we refer to the functional J although the same arguments apply for J_ν too.

Let $u_n^0 \in \mathcal{U}_{ad}$ be a minimizing sequence of J. Then, by definition of $\mathcal{U}_{ad}$, u_n^0 is bounded in L^∞ and there exists a subsequence, still denoted by u_n^0, such that $u_n^0 \rightharpoonup u_*^0$ weakly-* in L^∞. Moreover, $u_*^0 \in \mathcal{U}_{ad}$.

Let $u_n(x,t)$ and $u_*(x,t)$ be the entropy solutions of (7.1) with initial data u_n^0 and u_*^0, respectively, and assume for the moment that

$$u_n(\cdot, T) \to u_*(\cdot, T), \qquad \text{in } L^2(\mathbb{R}). \tag{7.6}$$

Then, clearly,

$$\inf_{u^0 \in \mathcal{U}_{ad}} J(u^0) = \lim_{n\to\infty} J(u_n^0) = J(u_*^0),$$

and we deduce that u_*^0 is a minimizer of J.

Thus, the key point is to prove the strong convergence result (7.6). Two main steps are necessary to do it. *a) The relative compactness of $u_n(\cdot, T)$ in L^2.* Taking the structure of $\mathcal{U}_{ad}$ into account and using the maximum principle and the finite velocity of propagation that entropy solutions fulfill, it is easy to see that the support of all solutions at time $t = T$ is uniformly included in the same compact set of $\mathbb{R}$. Thus, it is sufficient to prove compactness in L^2_{loc}. This is obtained from Oleinik's one-sided Lipschitz condition

$$\frac{u(x,t) - u(y,t)}{x - y} \le \frac{1}{t}, \tag{7.7}$$

which guarantees in fact a uniform bound of the BV-norm of $u_n(\cdot, T)$, locally in space (see [BrOs88]). The needed compactness property is then a consequence of the compactness of the embedding $BV(I) \subset L^2(I)$, for all bounded intervals I. *b) The identification of the limit as the entropy solution of (7.1) with initial datum u_*^0.* This can be proved using this compactness property and passing to the limit in the variational formulation of (7.1). We refer to [EsVa93] for a detailed description of this limit process in the more delicate case where the initial data converge to a Dirac delta.

This completes the proof of the existence of minimizers in the inviscid case.

In the viscous one, one cannot use the finite velocity of propagation. However, it is easy to get uniform bounds on the queues of solutions as $|x| \to \infty$, which allow us to reduce the global compactness problem to a local one. Locally, the same argument as above, based on the one-sided estimate (7.7), which is also true for the viscous equations, applies.

Remark 1. The above proof is in fact quite general and it can be adapted to other optimization problems with different functionals and admissible sets. In particular, using Oleinik's one-sided Lipschitz condition (7.7), one can also consider admissible sets of the form

$$\mathcal{U}_{ad} = \{f \in L^1(\mathbb{R}),\ supp(f) \subset K, ||f||_1 \leq C\}.$$

7.3 Vanishing Viscosity

The purpose of this section is to show that the minimizers of the vicous problem ($\nu > 0$) converge to a minimizer of the inviscid problem as the viscosity tends to zero, $\nu \to 0$.

Theorem 2. *Any accumulation point as $\nu \to 0$ of $u_\nu^{0,\min}$, the minimizers of (7.5), with respect to the weak topology in L^2, is a minimizer of the continuous problem (7.4).*

Proof of Theorem 2. We follow a standard Γ-convergence argument, as in [CaPa08], in the context of the convergence of minimizers for the numerical approximation schemes.

The proof is similar to the one in Theorem 1, although, this time, $\nu \to 0$.

The key ingredient is the following continuity property. Assume that $u_\nu^0 \in \mathcal{U}_{ad}$ satisfies $u_\nu^0 \rightharpoonup u^0$ weakly in $L^2(\mathbb{R})$. Then

$$J_\nu(u_\nu^0) \to J(u^0). \tag{7.8}$$

This is due to the fact that the OSLC condition, together with the uniform L^1-bound, guarantees uniform local BV bounds on the viscous solutions. For the viscous problem we do not have a finite velocity of propagation property but, as mentioned above, the uniform control of the queues allows us to reduce the compactness problem to a local one and then the local BV bounds suffice.

The limit process, as the viscosity parameter tends to zero, to recover in the limit the weak entropy solution of the inviscid model, can be conducted in a classical way. This is, for instance, done in [EsVa93].

Now, let $\hat{u}^0 \in \mathcal{U}_{ad}$ be an accumulation point of $u_\nu^{0,\min}$ with respect to the weak topology of L^2. To simplify the notation, we still denote by $u_\nu^{0,\min}$ the subsequence for which $u_\nu^{0,\min} \rightharpoonup \hat{u}^0$, weakly-* in $L^\infty(\mathbb{R})$, as $\nu \to 0$. Let $v^0 \in \mathcal{U}_{ad}$ be any other function. We are going to prove that

$$J(\hat{u}^0) \leq J(v^0). \tag{7.9}$$

To do this we construct a sequence $v_\nu^0 \in \mathcal{U}_{ad}^\nu$ such that $v_\nu^0 \to v^0$, in $L^2(\mathbb{R})$, as $\nu \to 0$. In this particular case, taking into account that the set of admissible controls $\mathcal{U}_{ad}^\nu$ is independent of $\nu > 0$, i.e., $\mathcal{U}_{ad} = \mathcal{U}_{ad}^\nu$, it is sufficient to choose, in particular, $v_\nu^0 = v^0$.

Taking into account the continuity property (7.8), we have

$$J(v^0) = \lim_{\nu \to 0} J_\nu(v_\nu^0) \geq \lim_{\nu \to 0} J_\nu(u_\nu^{0,\min}) = J(\hat{u}^0),$$

which proves (7.9).

Remark 2. Theorem 2 concerns global minima. However, both the continuous and discrete functionals may possibly have local minima as well. Extending this kind of Γ-convergence result for local minima requires important further developments.

7.4 Sensitivity Analysis: The Inviscid Case

In this section we collect the results in [CaPa08] for the sensitivity of the functional J in the presence of shocks, which follows previous works, e.g., [BrMa95a], [BaPi02], [BoJa98], [BoJa99], [Ul03], and [GoRa99].

We focus on the particular case of solutions having a single shock, but the analysis can be extended to consider more general one-dimensional systems of conservation laws with a finite number of noninteracting shocks (see [BrMa95a]).

7.4.1 Linearization of the Inviscid Burgers Equation

Following [CaPa08], we introduce the following hypothesis.

(H) Assume that $u(x,t)$ is a weak entropy solution of (7.1) with a discontinuity along a regular curve $\Sigma = \{(t, \varphi(t)), t \in [0,T]\}$, which is Lipschitz continuous outside Σ. In particular, it satisfies the Rankine–Hugoniot condition on Σ,

$$\varphi'(t)[u]_{\varphi(t)} = \left[u^2/2\right]_{\varphi(t)}, \tag{7.10}$$

or, simply,

$$\varphi'(t) = (u(\varphi(t)^+, t) + u(\varphi(t)^-, t))/2. \tag{7.11}$$

Here we have used the notation $[v]_{x_0} = v(x_0^+) - v(x_0^-)$ for the jump at x_0 of any piecewise continuous function v with a discontinuity at $x = x_0$, $v(x_0^\pm)$ standing for the values of v to both sides of x_0.

Note that Σ divides $\mathbb{R} \times (0,T)$ into two parts: Q^- and Q^+, the subdomains of $\mathbb{R} \times (0,T)$ to the left and to the right of Σ, respectively (see Figure 7.1).

As explained in [CaPa08], in the presence of shocks, for correctly dealing with optimal control and design problems, the state of the system (7.1) has

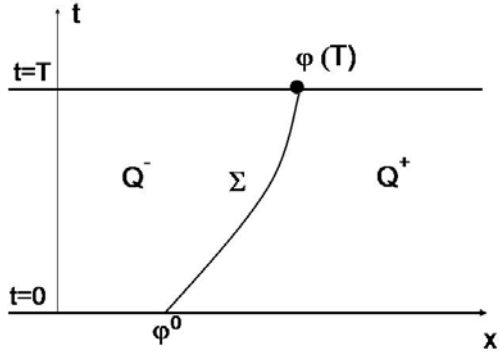


Fig. 7.1. The subdomains Q^- and Q^+.

to be viewed as being a pair (u, φ) combining the solution of (7.1) and the shock location φ.

Then the pair (u, φ) satisfies the system

$$\begin{cases} \partial_t u + \partial_x(\frac{u^2}{2}) = 0, & \text{in } Q^- \cup Q^+, \\ \varphi'(t)[u]_{\varphi(t)} = \left[u^2/2\right]_{\varphi(t)}, & t \in (0,T), \\ \varphi(0) = \varphi^0, & \\ u(x,0) = u^0(x), & \text{in } \{x < \varphi^0\} \cup \{x > \varphi^0\}. \end{cases} \tag{7.12}$$

We now analyze the sensitivity of (u, φ) with respect to perturbations of the initial datum, in particular, with respect to variations δu^0 of the initial profile u^0 and $\delta\varphi^0$ of the shock location φ^0. To be precise, we adopt the functional framework based on the generalized tangent vectors introduced in [BrMa95a].

Definition 1 ([BrMa95a]). *Let $v : \mathbb{R} \to \mathbb{R}$ be a piecewise Lipschitz continuous function with a single discontinuity at $y \in \mathbb{R}$. We define Σ_v as the family of all continuous paths $\gamma : [0, \varepsilon_0] \to L^1(\mathbb{R})$ with*

1. $\gamma(0) = v$ and $\varepsilon_0 > 0$ possibly depending on γ.

2. For any $\varepsilon \in [0, \varepsilon_0]$ the functions $u^\varepsilon = \gamma(\varepsilon)$ are piecewise Lipschitz with a single discontinuity at $x = y^\varepsilon$ depending continuously on ε and there exists a constant L independent of $\varepsilon \in [0, \varepsilon_0]$ such that

$$|v^\varepsilon(x) - v^\varepsilon(x')| \le L|x - x'|,$$

whenever $y^\varepsilon \notin [x, x']$.

Furthermore, we define the set T_v of generalized tangent vectors of v as the space of $(\delta v, \delta y) \in L^1 \times \mathbb{R}$ for which the path $\gamma_{(\delta v, \delta y)}$ given by

$$\gamma_{(\delta v, \delta y)}(\varepsilon) = \begin{cases} v + \varepsilon\delta v + [v]_y \ \chi_{[y+\varepsilon\delta y, y]} & \text{if } \delta y < 0, \\ v + \varepsilon\delta v - [v]_y \ \chi_{[y, y+\varepsilon\delta y]} & \text{if } \delta y > 0, \end{cases} \tag{7.13}$$

satisfies $\gamma_{(\delta v,\delta y)} \in \Sigma_v$.

Finally, we define the equivalence relation $\sim$ *defined on* Σ_v *by*

$$\gamma \sim \gamma' \text{ if and only if } \lim_{\varepsilon\to 0} \frac{\|\gamma(\varepsilon)-\gamma'(\varepsilon)\|_{L^1}}{\varepsilon} = 0,$$

and we say that a path $\gamma \in \Sigma_v$ *generates the generalized tangent vector* $(\delta v, \delta y) \in T_v$ *if* γ *is equivalent to* $\gamma_{(\delta v,\delta y)}$ *as in (7.13).*

Remark 3. The path $\gamma_{(\delta v,\delta y)} \in \Sigma_v$ in (7.13) represents, at first order, the variation of a function v by adding a perturbation function $\varepsilon\delta v$ and by shifting the discontinuity by $\varepsilon\delta y$.

Note that, for a given v (piecewise Lipschitz continuous function with a single discontinuity at $y \in \mathbb{R}$) the associated generalized tangent vectors $(\delta v, \delta y) \in T_v$ are those pairs for which δv is Lipschitz continuous with a single discontinuity at $x = y$.

Let u^0 be the initial datum in (7.12) that we assume to be Lipschitz continuous to both sides of a single discontinuity located at $x = \varphi^0$, and consider a generalized tangent vector $(\delta u^0, \delta\varphi^0) \in L^1(\mathbb{R}) \times \mathbb{R}$. Let $u^{0,\varepsilon} \in \Sigma_{u^0}$ be a path which generates $(\delta u^0, \delta\varphi^0)$. For ε sufficiently small the solution $u^\varepsilon(\cdot, t)$ of (7.12) is Lipschitz continuous with a single discontinuity at $x = \varphi^\varepsilon(t)$, for all $t \in [0, T]$. Thus, $u^\varepsilon(\cdot, t)$ generates a generalized tangent vector $(\delta u(\cdot, t), \delta\varphi(t)) \in L^1(\mathbb{R}) \times \mathbb{R}$. Moreover, in [BrMa95b] it is proved that it satisfies the following linearized system:

$$\begin{cases} \partial_t \delta u + \partial_x(u\delta u) = 0, \quad \text{in } Q^- \cup Q^+, \\ \delta\varphi'(t)[u]_{\varphi(t)} + \delta\varphi(t)\left(\varphi'(t)[u_x]_{\varphi(t)} - [u_x u]_{\varphi(t)}\right) \\ \qquad + \ \varphi'(t)[\delta u]_{\varphi(t)} - [u\delta u]_{\varphi(t)} = 0, \quad \text{in } (0,T), \\ \delta u(x,0) = \delta u^0, \quad \text{in } \{x < \varphi^0\} \cup \{x > \varphi^0\}, \\ \delta\varphi(0) = \delta\varphi^0, \end{cases} \tag{7.14}$$

with the initial data $(\delta u^0, \delta\varphi^0)$.

Remark 4. In this way, we can obtain formally the expansion

$$(u_\varepsilon, \varphi_\varepsilon) = (u, \varphi) + \varepsilon(\delta u, \delta\varphi) + \mathcal{O}(\varepsilon^2).$$

Remark 5. The linearized system (7.14) has a unique solution which can be computed in two steps. The method of characteristics determines δu in $Q^- \cup Q^+$, i.e., outside Σ, from the initial data δu^0 (note that system (7.14) has the same characteristics as (7.12)). This yields the value of u and u_x at both sides of the shock Σ and allows the determination of the coefficients of the ordinary different equation (ODE) that $\delta\varphi$ satisfies. This ODE yields $\delta\varphi$.

Remark 6. We have assumed that the discontinuity of the solution of the Burgers equation u is present in the whole time interval $t \in [0, T]$. But the discontinuities may appear at time $\tau \in (0, T)$ for some regular initial data. We refer to [CaPa08] for the linearization in this case.

7.4.2 Sensitivity in the Presence of Shocks

In this section we study the sensitivity of the functional J with respect to variations associated with the generalized tangent vectors defined in the previous section. We first define an appropriate generalization of the Gateaux derivative.

Definition 2 ([BrMa95a]). *Let $J : L^1(\mathbb{R}) \to \mathbb{R}$ be a functional and $u^0 \in L^1(\mathbb{R})$ be Lipschitz continuous with a discontinuity at $x = \varphi^0$, an initial datum for which the solution of (7.1) satisfies hypothesis (H). We say that J is Gateaux differentiable at u^0 in a generalized sense if for any generalized tangent vector $(\delta u^0, \delta\varphi^0)$ and any family $u^{0,\epsilon} \in \Sigma_{u^0}$ associated to $(\delta u^0, \delta\varphi^0)$ the following limit exists:*

$$\delta J = \lim_{\epsilon \to 0} \frac{J(u^{0,\epsilon}) - J(u^0)}{\epsilon},$$

and it depends only on (u^0, φ^0) and $(\delta u^0, \delta\varphi^0)$, i.e., it does not depend on the particular family $u^{0,\epsilon}$ which generates $(\delta u^0, \delta\varphi^0)$.

The limit δJ is the generalized Gateux derivative of J in the direction $(\delta u^0, \delta\varphi^0)$.

The following result provides an easy characterization of the generalized Gateaux derivative of J in terms of the solution of the associated adjoint system.

Proposition 1. *The Gateaux derivative of J can be written as*

$$\delta J = \int_{\{x<\varphi^0\}\cup\{x>\varphi^0\}} p(x,0)\delta u^0(x)\, dx + q(0)[u^0]_{\varphi^0}\delta\varphi^0, \tag{7.15}$$

where the adjoint state pair (p, q) satisfies the system

$$\begin{cases} -\partial_t p - u\partial_x p = 0, \quad \text{in } Q^- \cup Q^+, \\ [p]_\Sigma = 0, \\ q(t) = p(\varphi(t), t), \ \text{in } t \in (0,T) \\ q'(t) = 0, \ \text{in } t \in (0,T) \\ p(x,T) = u(x,T) - u^d, \quad \text{in } \{x < \varphi(T)\} \cup \{x > \varphi(T)\} \\ q(T) = \frac{\frac{1}{2}\left[(u(x,T)-u^d)^2\right]_{\varphi(T)}}{[u]_{\varphi(T)}}. \end{cases} \tag{7.16}$$

Remark 7. System (7.16) has a unique solution. In fact, to solve the backwards system (7.16) we first define the solution q on the shock Σ from the condition $q' = 0$, with the final value $q(T)$ given in (7.16). This determines the value of p along the shock. We then propagate this information, together with the datum of p at time $t = T$ to both sides of $\varphi(T)$, by characteristics. As both systems (7.1) and (7.16) have the same characteristics, any point $(x,t) \in \mathbb{R} \times (0,T)$ is reached backwards in time by a unique characteristic line coming

either from the shock Σ or the final data at (x,T) (see Figure 7.2). The solution obtained this way coincides with the reversible solutions introduced in [BoJa98] and [BoJa99].

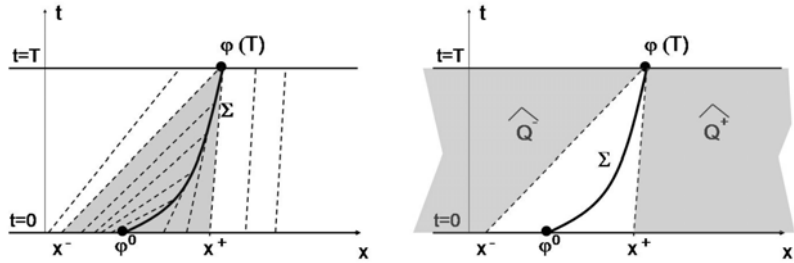


Fig. 7.2. Characteristic lines entering on a shock (left) and subdomains $\hat{Q}^-$ and $\hat{Q}^+$ (right).

Remark 8. Solutions of (7.16) can also be obtained as the limit of solutions of the transport equation with artificial viscosity depending on a small parameter $\varepsilon \to 0$,

$$\begin{cases} -\partial_t p - u\partial_x p = \varepsilon \partial_{xx} p, & \text{in } x \in \mathbb{R},\ t \in (0,T), \\ p(x,T) = p_n^T(x), & \text{in } x \in \mathbb{R}, \end{cases} \tag{7.17}$$

and a suitable choice of the initial data $p_n^T(x)$, depending on $n \to \infty$. To be more precise, let $p_n^T(x)$ be any sequence of Lipschitz continuous functions, uniformly bounded in $BV_{loc}(\mathbb{R})$, such that

$$p_n^T(x,T) \to p^T(x) = u(x,T) - u^d(x), \quad \text{in } L^1_{loc}(\mathbb{R}),$$

and

$$p_n^T(\varphi(T),T) = \frac{\frac{1}{2}\left[(u(x,T)-u^d)^2\right]_{\varphi(T)}}{[u]_{\varphi(T)}}.$$

We first take the limit of the solutions $p_{\varepsilon,n}$ of (7.17) as $\varepsilon \to 0$, to obtain the solution p_n of

$$\begin{cases} -\partial_t p - u\partial_x p = 0, & \text{in } x \in \mathbb{R},\ t \in (0,T), \\ p(x,T) = p_n^T(x), & \text{in } x \in \mathbb{R}, \end{cases}$$

which is called the *reversible solution* (see [BoJa98]). These solutions can be characterized by the fact that they take the value $p_n(\varphi(T),T)$ in the whole region occupied by the characteristics that meet the shock (see [BoJa98], Theorem 4.1.12). Thus, in particular, they satisfy the 2nd, 3rd, 4th, and 6th equations in (7.16). Moreover, $p_n \to p$ as $n \to \infty$, and p takes a constant value

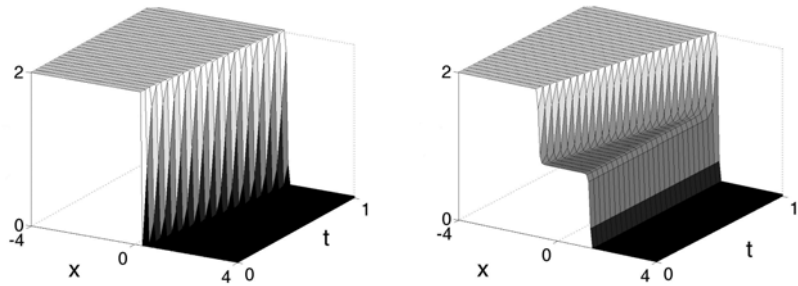


Fig. 7.3. Solution $u(x,t)$ of the Burgers equation with an initial datum having a discontinuity (left) and adjoint solution which takes a constant value in the region occupied by the characteristics that meet the shock (right).

in the region occupied by the characteristics that meet the shock. Note that, by construction, this constant is the same value for all p_n in this region. Thus, this limit solution p coincides with the solution of (7.16) constructed above.

Formula (7.15) provides an obvious way to compute a first descent direction of J at u^0. We just take

$$(\delta u^0, \delta\varphi^0) = (-p(x,0), -q(0)[u]_{\varphi^0}). \tag{7.18}$$

Here, the value of $\delta\varphi^0$ must be interpreted as the optimal infinitesimal displacement of the discontinuity of u^0.

But this $(\delta u^0, \delta\varphi^0)$ is not a generalized tangent vector in T_{u^0} since $p(x,0)$ is not continuous away from $x \neq \varphi^0$. A typical example is shown in Figure 7.3. In [CaPa08] we have solved this drawback by introducing the alternating descent algorithm, which only uses generalized tangent vectors, distinguishing those that move the shock and those that do not.

7.5 The Method of Alternating Descent Directions

In this section we briefly present the alternating descent algorithm introduced in [CaPa08] in the inviscid case.

Motivated by the above discussion, we introduce a decomposition of the generalized tangent vectors. This requires us first to introduce some notation. Let

$$x^- = \varphi(T) - u^-(\varphi(T))T, \qquad x^+ = \varphi(T) - u^+(\varphi(T))T, \tag{7.19}$$

and consider the following subsets (see Figure 7.2):

$$\hat{Q}^- = \{(x,t) \in \mathbb{R} \times (0,T) \text{ such that } x < \varphi(T) - u^-(\varphi(T))t\},$$

$$\hat{Q}^+ = \{(x,t) \in \mathbb{R} \times (0,T) \text{ such that } x > \varphi(T) - u^+(\varphi(T))t\}.$$

The classification of the generalized tangent vectors in two classes is motivated by the following result (see [CaPa08]).

Proposition 2. *Consider the paths in Σ_{u^0} for which the associated generalized tangent vectors $(\delta u^0, \delta\varphi^0) \in T_{u^0}$ satisfy*

$$\delta\varphi^0 = -\frac{\int_{x^-}^{\varphi^0} \delta u^0 + \int_{\varphi^0}^{x^+} \delta u^0}{[u^0]_{\varphi^0}}. \tag{7.20}$$

Then, the solution $(\delta u, \delta\varphi)$ of system (7.14) satisfies $\delta\varphi(T) = 0$ and the generalized Gateaux derivative of J in the direction $(\delta u^0, \delta\varphi^0)$ can be written as

$$\delta J = \int_{\{x<x^-\}\cup\{x>x^+\}} p(x,0)\delta u^0(x)\, dx, \tag{7.21}$$

where p satisfies the system

$$\begin{cases} -\partial_t p - u\partial_x p = 0, & \text{in } \hat{Q}^- \cup \hat{Q}^+, \\ p(x,T) = u(x,T) - u^d, & \text{in } \{x < \varphi(T)\} \cup \{x > \varphi(T)\}. \end{cases} \tag{7.22}$$

Analogously, when considering paths in Σ_{u^0} for which the associated generalized tangent vectors $(\delta u^0, \delta\varphi^0) \in T_{u^0}$ satisfy $\delta u^0 = 0$, then $\delta u(x,T) = 0$ and the generalized Gateaux derivative of J in the direction $(\delta u^0, \delta\varphi^0)$ can be written as

$$\delta J = -\left[\frac{(u(x,T) - u^d(x))^2}{2}\right]_{\varphi(T)} \frac{[u^0]_{\varphi^0}}{[u(\cdot,T)]_{\varphi(T)}} \delta\varphi^0. \tag{7.23}$$

Remark 9. Formula (7.21) provides a simplified expression for the generalized Gateaux derivative of J when considering directions $(\delta u^0, \delta\varphi^0)$ that do not move the shock position at $t = T$. These directions are characterized by formula (7.20) which determines the infinitesimal displacement of the shock position $\delta\varphi^0$ in terms of the variation of u^0 to both sides of $x = \varphi^0$. Note, in particular, that to any value δu^0 to both sides of the jump φ^0 there corresponds a unique infinitesimal translation $\delta\varphi^0$ of the initial shock position that does not move it at $t = T$.

Note also that the value of p outside the region $\hat{Q}^- \cup \hat{Q}^+$ is not needed to evaluate the generalized Gateaux derivative in (7.21). Solving system (7.22) is particularly easy since the potential u is smooth in the region where the system is formulated.

Analogously, formula (7.23) provides a simplified expression of the generalized Gateaux derivative of J when considering directions $(\delta u^0, \delta\varphi^0)$ that uniquely move the shock position at $t = T$ and which correspond to purely translating the shock.

The method of *alternating descent directions* can then be implemented as follows, applying in each step of the descent, the following two substeps:

1. Use generalized tangent vectors that move the shock to search its optimal placement.
2. Use generalized tangent vectors to modify the value of the solution at time $t = T$ to both sides of the discontinuity, leaving the shock location unchanged.

One of the main advantages of this method is that the complexity of the solutions is preserved without introducing artificial shocks that are unnecessary to approximate the target u^d.

The efficiency of this method compared to the classical one based on purely discrete approaches or continuous ones that do not make an optimal use of the shock analysis has been illustrated in [CaPa08] through several numerical experiments.

Note also that this method is, in some sense, close to the methods employed in shape design in elasticity in which topological derivatives (that in the present setting would correspond to controlling the location of the shock) are combined with classical shape deformations (that would correspond to simply shaping the solution away from the shock in the present setting) ([GaGu01]).

7.6 Alternating Descent Directions in the Viscous Case

The linearized Burgers equation reads as follows:

$$\begin{cases} \partial_t \delta u - \nu \delta u_{xx} + \partial_x(u\delta u) = 0, & \text{in } \mathbb{R} \times (0, \infty), \\ \delta u(x, 0) = \delta u^0, & \text{in } \mathbb{R}. \end{cases} \tag{7.24}$$

In this case the derivation of this linearized equation is straightforward because of the smoothness of solutions.

Moreover, the Gateaux derivative of the functional J is as follows:

$$\delta J =< \delta J(u^0), \delta u^0 >= \int_{\mathbb{R}} p(x, 0)\delta u^0(x)\, dx, \tag{7.25}$$

where the adjoint state $p = p_\nu$ solves the adjoint system

$$\begin{cases} -\partial_t p - \nu p_{xx} - u\partial_x p = 0, & \text{in } \mathbb{R},\ 0 < t < T, \\ p(x, T) = u(x, T) - u^d, & \text{in } \mathbb{R}. \end{cases} \tag{7.26}$$

In this case, unlike the inviscid one, the adjoint state has only one component. Indeed, since the state does not present shocks, there is no adjoint shock variable. Similarly, the derivative of J in (7.15) exhibits only one term. According to this, the straightforward application of a gradient method for the optimization of J would lead, in each step of the iteration, to the use of variations pointing in the direction

$$\delta u^0 = -p(x,0),$$

p being the solution of this viscous adjoint system. But, proceeding in this way, we would not exploit the possibilities that the alternate descent method provides.

To do this we must also consider the effect of possible infinitesimal translations of the initial data. Indeed, the previous calculus is valid when the variations of the initial data are of the form

$$u^0_\varepsilon(x) = u^0(x) + \varepsilon\delta u^0(x). \tag{7.27}$$

But, in order to consider the possible effect of the infinitesimal translations, we have to use rather variations of the form

$$u^0_\varepsilon(x) = u^0(x + \varepsilon\delta\varphi^0) + \varepsilon\delta u^0, \tag{7.28}$$

where, now, φ^0 stands for a reference point on the profile of u^0, not necessarily a point of discontinuity. When u^0 has a point of discontinuity, φ^0 could be its location and $\delta\varphi^0$ an infinitesimal variation of it. But φ^0 could also stand for another singular point on the profile of u^0, e.g., an extremal point, or a point where the gradient of u^0 is large, i.e., a smeared discontinuity.

Note that, by a Taylor expansion, when considering variations of this form, to first order, this corresponds to

$$u^0_\varepsilon(x) \sim u^0(x) + \varepsilon\delta u^0(x) + \varepsilon\delta\varphi^0 u^0_x(x). \tag{7.29}$$

This indicates that the result of these combined variations $(\delta u^0, \delta\varphi^0)$ is equivalent to a classical variation in the direction of $\delta u^0 + \delta\varphi^0 u^0_x$. When u^0 is smooth enough, for instance, $u^0 \in H^1$, then, this yields a standard variation in an L^2 direction. But when u^0 lacks regularity, for instance, when u^0 has a point of discontinuity, this yields variations that are singular and contain Dirac deltas. Similarly, when u^0 is smooth but has a large gradient, we see that the effect of a small $\delta\varphi^0$ is amplified by the size of the gradient.

The corresponding linearization of the Burgers equation is as follows:

$$\begin{cases} \partial_t \delta u - \nu\delta u_{xx} + \partial_x(u\delta u) = 0, & \text{in } \mathbb{R}\times(0,\infty), \\ \delta u(x,0) = \delta u^0(x) + \delta\varphi^0 u^0_x(x), & \text{in } \mathbb{R}. \end{cases} \tag{7.30}$$

Again, the derivation of this linearized equation is straightforward because of the smoothness of solutions.

In view of (7.30) the linearization of the functional is as follows:

$$\delta J = \int_{\mathbb{R}} p(x,0)\delta u^0(x)\,dx + \delta\varphi^0 \int_{\mathbb{R}} p(x,0)u^0_x(x)\,dx, \tag{7.31}$$

where the adjoint state $p = p_\nu$ is as above.

When u^0 is piecewise smooth but it has a discontinuity at $x = \varphi^0$, this variation can be written as follows:

$$\delta J = \int_{\mathbb{R}} p(x,0)\delta u^0(x)\,dx + \delta\varphi^0 \int_{\mathbb{R}-\{\varphi^0\}} p(x,0)u_x^0(x)\,dx + \delta\varphi^0[u^0]_{x=\varphi^0}p(\varphi^0,0), \tag{7.32}$$

where $[u^0]_{x=\varphi^0}$ stands for the jump of u^0 at $x=\varphi^0$.

When comparing this expression with (7.15), we see that there is an extra term, namely,

$$\delta\varphi^0 \int_{\mathbb{R}-\{\varphi^0\}} p(x,0)u_x^0(x)\,dx,$$

which corresponds to the fact that the variations considered in the inviscid case by means of the generalized tangents and (7.30) only coincide with those considered here when u_0 is piecewise constant with a shock at φ^0. When the initial datum has a discontinuity at $x=\varphi^0$, a slight change in the way the variations (7.28) are defined, considering the vectors in (7.13), leads to an expression which is closer to (7.15). This is done translating the point of discontinuity by adding as in (7.13) a characteristic function of the amplitude of the jump of u^0 so that the jump point is shifted infinitesimally to the left or to the right, but without adding any extra variation on the initial profile u^0 due to this shift.

But, let us continue our analysis by keeping the class of variations (7.28), supposing that u^0 is continuous. We can rewrite the first variation of J as follows:

$$\delta J = \int_{\mathbb{R}} p(x,0)[\delta u^0(x) + \delta\varphi^0 u_x^0(x)]\,dx. \tag{7.33}$$

In the inviscid case, to develop the method of alternating descent, we distinguished the variations of the initial datum moving the shock and those that did not move it by modifying the profile away from it. This discussion does not make sense as such in the present setting since the solutions of the viscous state equation are now smooth. However, from a computational viewpoint, it is interesting to develop the analogue of the alternating descent method.

For this to be done, one needs to distinguish two classes of possible variations. But this time one has to do it without having, as in the inviscid case, the shock location and its region of influence at $t=0$ which, in that case, we identified with the interval $[x^-,x^+]$ as in (7.19).

Let us however assume that the viscosity parameter ν is small enough, so that viscous solutions are close to the inviscid ones, and develop a strategy inspired in the way that the alternating descent direction was built in the inviscid case. For it to be meaningful, we need to identify a class of initial data for which the alternating descent method might be more efficient than the classical one, which consists in simply applying a descent algorithm based on the adjoint calculus above. We will identify this class as the one in which the initial data u^0 leads to a discontinuous solution in the inviscid case.

Assume, to begin with, that u^0 has a discontinuity at φ^0 and that it is smooth to both sides of it. The viscosity parameter ν being positive, even if

it is small, the solution is smooth and, therefore, it may not develop shocks. However, taking into account that solutions are close to the inviscid ones, when the latter exhibit shocks, the viscous ones will develop regularized quasi-shocks. Therefore, one could try to mimic the same procedure for the viscous case. The first thing to be done is to identify the region of influence $[x^-, x^+]$ of the inner boundary of the inviscid adjoint system. But, of course, this should be done without solving the inviscid adjoint system which, on the other hand, would require solving the inviscid state equation. We therefore need an alternate definition of the interval $[x^-, x^+]$ to that in (7.19) which might be easy to compute. To do that it is necessary to compute the curve where the shock is located in the inviscid case. This can be done by solving the ODE given by the Rankine–Hugoniot condition:

$$\varphi'(t) = [u^+(\varphi(t), t) + u^-(\varphi(t), t)]/2, \quad t \in (0, T). \tag{7.34}$$

Here u^+ and u^- stand for the value of the solution u of the inviscid problem at both sides of the shock. They can be computed by the method of characteristics so that

$$u^\pm(\varphi(t), t) = u^0(s^\pm(t)), \quad t \in (0, T), \tag{7.35}$$

where $s^\pm(t)$ is the spatial trajectory of the characteristic which arrives to $(\varphi(t), t)$ starting from $t = 0$, and we solve

$$s^\pm(t) + tu^{0,\pm}(s^\pm(t)) = \varphi(t), \quad t \in (0, T). \tag{7.36}$$

Substituting (7.35) and (7.36) into (7.34), the ODE for the shock then reads

$$\varphi'(t) = [u^0(s^+(t)) + u^0(s^-(t))]/2, \quad t \in (0, T), \tag{7.37}$$

and

$$x^\pm = s^\pm(T). \tag{7.38}$$

Once this is done, we need to identify the variations $(\delta u^0, \delta\varphi^0)$ such that

$$\int_{x^-}^{x^+} p(x, 0)[\delta u^0(x) + \delta\varphi^0 u_x^0(x)]\, dx = 0. \tag{7.39}$$

If $p(x, 0)$ were constant within the interval $[x^-, x^+]$ as in the inviscid case, this would amount to considering variations such that

$$\delta\varphi^0 = -\frac{\int_{x^-}^{x^+} \delta u^0(x)dx}{u^0(x^+) - u^0(x^-)}. \tag{7.40}$$

There is no unique way of doing this. One possibility would be to consider variations δu^0 in $[x^-, x^+]$ such that $\int_{x^-}^{x^+} \delta u^0(x)dx = 0$ and $\delta\varphi^0 = 0$.

The variation of the functional J would then be

$$\delta J = \int_{\{x<x^-\}\cap\{x>x^+\}} p(x,0)\delta u^0(x)dx, \tag{7.41}$$

and the optimal descent direction

$$\delta u^0(x) = -p(x,0), \quad \text{in } \{x < x^-\} \cap \{x > x^+\}. \tag{7.42}$$

This discussion leads to considering "descent directions" of the form (7.42), where p is the solution of the adjoint viscous system and the extremes of the interval $x^\pm$ are computed according to (7.36)–(7.38). Furthermore, δu^0 has to be extended to $[x^-, x^+]$ so that $\int_{x^-}^{x^+} \delta u^0(x)dx = 0$ and $\delta\varphi^0 = 0$. Note also that, as observed in [CaPa08], it is convenient to choose δu^0 which is continuous away from φ^0 to guarantee that the deformations under consideration do not increase the number of possible discontinuities of u^0. Obviously, this is possible within the class of variations we have identified.

This class of deformations has been identified under the assumption that $p(x,0)$ is constant within the interval $[x^-, x^+]$, a property that is indeed true in the inviscid case but not in the viscous one. The rigorous analysis of the convergence of the adjoint states as the viscosity parameter ν tends to zero, and the possible improvement of the class of variations above, is an interesting topic for future research.

The second class of variations is the one that takes advantage of the infinitesimal translations $\delta\varphi^0$. We can then set $\delta u^0 \equiv 0$ and choose $\delta\varphi^0$ such that

$$\delta\varphi^0 = -\int_{\mathbb{R}-\{\varphi^0\}} p(x,0)u_x^0(x)\, dx - [u^0]_{x=\varphi^0} p(\varphi^0, 0).$$

As mentioned above, we could consider slightly different variations of the initial data of the form

$$\delta\varphi^0 = -[u^0]_{x=\varphi^0} p(\varphi^0, 0),$$

as in [CaPa08].

On the other hand, in the inviscid case, $p(\varphi^0, 0)$ coincides with the value of p at time $t = T$ at the shock location and, therefore, this descent direction can be computed without performing any numerical approximation of p. This is no longer the case in the present viscous setting in which $p(\varphi^0, 0)$ is a priori unknown. To simplify the choice, we can use the proximity of the inviscid adjoint state and the viscous one. When doing this and using the above (slightly different) variations of the initial data, the choice for $\delta\varphi^0$, inspired by (7.23), would be

$$\delta\varphi^0 = \left[\frac{(u(x,T) - u^d(x))^2}{2}\right]_{\varphi(T)} \frac{[u^0]_{\varphi^0}}{[u(\cdot,T)]_{\varphi(T)}},$$

where $\varphi(T)$ is the location of the shock in the inviscid case which, in view of (7.36)–(7.38), is given by

$$\varphi(T) = x^- + Tu^0(x^-) = x^+ + Tu^0(x^+).$$

Similarly, in the inviscid case, the computation of the jump of $u(\cdot, T)$ and $(u(x,T) - u^d(x))^2$ at $x = \varphi(T)$ can be greatly simplified since the values of u at $t = T$ at both sides of the discontinuity $x = \varphi(T)$ can be computed by the method of characteristics and coincide with $u^0(x^{\pm})$.

In this way, we have identified two classes of variations and its approximate values inspired in the structure of the state and the adjoint state in the inviscid case, allowing us to implement the method of alternating descent in the inviscid case when u^0 is discontinuous.

This analysis can be extended to the case where u^0 is smooth but the corresponding solution of the inviscid problem develops shock discontinuities in some time $0 \leq \tau < T$. This can be fully characterized in terms of u^0, as is well known. Then, the analysis of the previous case can be applied with the possible variant discussed in [CaPa08] when the shock does not start at $t = 0$ but rather appears in a time $0 < \tau < T$.

In this way one can handle, for instance, the prototypical solutions of the viscous Burgers equation that, as $\nu \to 0$, converge to shock solutions ([Wh74]). These are the smooth traveling wave solutions of the viscous Burgers equation (7.2) taking values $u_{\pm}$ at $\pm\infty$, of the form,

$$u_\nu(x,t) = u_+ + \frac{u_- - u_+}{1 + \exp[(u_- - u_+)(x - \overline{u}t)/2\nu]}, \tag{7.43}$$

where

$$\overline{u} = (u_- + u_+)/2. \tag{7.44}$$

When $u_- > u_+$ and $\nu \to 0$, they converge to the shock solution of the inviscid Burgers equation taking values u_+ for $x > \overline{u}t$ and u_- for $x < \overline{u}t$.

The efficiency of the method developed in this section is illustrated by several numerical experiments in the following section.

7.7 Numerical Experiments

In this section we focus on the numerical approximation of the optimization problem described in this chapter. The first natural question is the choice of the numerical method to approximate both the Burgers equation and its adjoint.

Let us introduce a mesh in $\mathbb{R} \times [0,T]$ given by $(x_j, t^n) = (j\Delta x, n\Delta t)$ ($j = -\infty, ..., \infty$; $n = 0, ..., N+1$ so that $(N+1)\Delta t = T$), and let u_j^n be a numerical approximation of $u(x_j, t^n)$ obtained as a solution of a suitable discretization of the Burgers equation.

As we are assuming the viscosity parameter ν to be small, it seems natural to consider a viscous perturbation of the most common numerical schemes for conservation laws. Accordingly, let us introduce a 3-point conservative numerical approximation scheme for the nonlinearity and an explicit scheme for the viscosity:

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{\Delta t}{\Delta x}\left(g_{j+1/2}^n - g_{j-1/2}^n\right) + \nu\frac{\Delta t}{\Delta x^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n), \\ & j \in \mathbb{Z},\ n = 0, ..., N, \end{aligned} \tag{7.45}$$

where

$$g_{j+1/2}^n = g(u_j^n, u_{j+1}^n),$$

and g is the numerical flux. These schemes are consistent with the viscous Burgers equation when $g(u,u) = u^2/2$ since, in this case, both the nonlinear part and the viscous perturbation are consistent.

In order to analyze the scheme (7.45), we note that it can be written as a conservative numerical scheme with the modified numerical flux,

$$g_{vis}(u,v) = g(u,v) - \frac{\nu}{\Delta x}(v-u). \tag{7.46}$$

In particular, the stability analysis can be obtained from the classical analysis for conservative schemes.

It is interesting to observe that the stability of these numerical schemes is not granted from the stability of the underlying conservative scheme for the inviscid Burgers equation. To clarify this issue, we divide the rest of this section into two more subsections. In the first one we analyze the stability of the numerical schemes written in the form (7.45) and we introduce a convergent numerical scheme. The second subsection is devoted to illustrate the numerical results for the optimization problem.

7.7.1 Discussion of the Stability of the Viscous Versions of Hyperbolic Conservative Schemes

We first focus on the von Neumann analysis for the stability of the simpler linear equation,

$$u_t + au_x = \nu u_{xx}, \quad \text{with } a \text{ constant.} \tag{7.47}$$

We follow the analysis in [GoRa91] for conservative schemes. It is well known that any 3-point conservative numerical scheme can be written in viscosity form as

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}a\frac{u_{j+1}^n - u_{j-1}^n}{2} + \tilde{q}\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2}, \tag{7.48}$$

for some viscosity coefficient $\tilde{q}$. Therefore, the numerical scheme (7.45) can also be written as (7.48) with the new viscosity coefficient,

$$q = \tilde{q} + 2\nu\frac{\Delta t}{\Delta x^2}.$$

Taking into account that $u_{j+1}^n \sim u(x_j + \Delta x, t_n)$, if we write $x_j = x$ and consider the Fourier transform in x, we obtain

$$\widehat{u}^{n+1}(\eta) = h(\eta)\widehat{u}^{n}(\eta).$$

The value $h(\eta)$ represents the amplification factor that must be smaller than one in modulus to guarantee stability. In this case,

$$h(\eta) \quad = \quad 1 - q(1 - \cos\eta\Delta x) - i\frac{\Delta t}{\Delta x}a\sin\eta\Delta x.$$

If we write

$$y = \sin(\eta\Delta x/2)^2,$$

then

$$|h(\eta)|^2 = (1 - 2qy)^2 + 4\left(\frac{\Delta t}{\Delta x}a\right)^2 y(1-y).$$

It is not difficult to show that a necessary and sufficient condition for the L^2-stability, i.e., $|h(\eta)| \leq 1$, is to have

$$\frac{\Delta t^2}{\Delta x^2} \leq q = \tilde{q} + 2\nu\frac{\Delta t}{\Delta x^2} \leq 1.$$

From this condition we easily deduce that not all convergent numerical methods for solving the inviscid Burgers equation are stable when adding the dissipative term $\nu\frac{\Delta t}{\Delta x^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n)$, even for arbitrarily small Δt. For example, in the Lax–Friedrichs scheme the numerical flux is given by

$$g^{lf}(u,v) = a\frac{u+v}{2} - \frac{v-u}{2\Delta t/\Delta x},$$

and $\tilde{q} = 1$. Therefore, it becomes unstable as soon as $\nu > 0$, whatever the choice of Δt is.

In the following experiments we have chosen the numerical flux associated to the Engquist–Osher scheme. For the linear equation (7.47) the numerical flux is reduced to

$$g^{eo}(u,v) = \frac{u(a+|a|)}{4} + \frac{v(a-|a|)}{4}.$$

In this case, $q = |a|\frac{\Delta t}{\Delta x}$ and the scheme is stable as soon as

$$\Delta t \leq \frac{\Delta x^2}{\Delta x|a| + 2\nu}.$$

In the nonlinear case, the numerical flux associated to the Engquist–Osher scheme is given by

$$g^{eo}(u,v) = \frac{u(u+|u|)}{4} + \frac{v(v-|v|)}{4}.$$

Generally speaking, the stability of these schemes for the nonlinear viscous Burgers equation can be obtained from the stability analysis for general

conservative schemes, since they can be written in conservative form with the modified flux (7.46).

It is well known that such schemes admit the following viscous form:

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\frac{f(u_{j+1}^n) - f(u_j^n)}{2} + \frac{Q_{j+1/2}(v_{j+1} - v_j) - Q_{j-1/2}(v_j - v_{j-1})}{2},$$

where

$$Q_{j+1/2} = \frac{\Delta t}{\Delta x}(f(u_{j+1}^n) + f(u_j^n) - 2g_{vis}^{eo}(u_j^n, u_{j+1}^n)),$$

and that the scheme is total variation diminishing (TVD) if (see [GoRa91], p. 136)

$$\frac{\Delta t}{\Delta x}\left|\frac{f(u_{j+1}^n) + f(u_j^n)}{u_{j+1}^n - u_j^n}\right| \leq Q_{j+1/2} \leq 1.$$

Thus, this numerical scheme is stable if the following condition is satisfied:

$$(\max_j |u_j^0|)^2 \frac{\Delta t}{\Delta x} + 2\nu \frac{\Delta t}{\Delta x^2} \leq 1. \tag{7.49}$$

7.7.2 Numerical Experiments for the Optimization Problem

In this section we present some numerical experiments to illustrate the results of the previous sections. We pay special attention to showing the applicability of the alternating descent method.

We emphasize that the solutions obtained with each method may correspond to global minima or local ones since the gradient algorithm does not distinguish them.

We consider an exact solution of the Burgers equation obtained as a traveling wave

$$u(x,t) = \frac{1}{2}\left(1 - \tanh\frac{x - t/2}{4\nu}\right).$$

This solution is smooth for $\nu > 0$ but, as $\nu \to 0$, it approaches a piecewise constant function with a discontinuity at $x = t/2$, $t \in [0,1]$. We choose the final time $T = 1$ and the target u^d, different for each value of the viscosity parameter, given by

$$u^d = \frac{1}{2}\left(1 - \tanh\frac{x - T/2}{4\nu}\right). \tag{7.50}$$

Note that the functional attains its minimum value, $J = 0$, and a minimizer is given by

$$u^{0,min} = \frac{1}{2}\left(1 - \tanh\frac{x}{4\nu}\right). \tag{7.51}$$

The interval $(-6, 6)$ has been chosen as the computational domain, and we have taken as boundary conditions, at each time step $t = t^n$, the value of the known target at the boundary.

To illustrate the efficiency of the alternating descent method, we have solved the optimization problem with a descent method using the usual adjoint formulation and with the alternating descent method, for different values of the viscosity parameter $\nu = 0.5, 0.1$ and $\nu = 0.01$.

We also consider $\Delta x = 0.02$ and $\Delta t = \Delta x^2/2$, which satisfies the stability condition (7.49).

It is interesting to compare the relation between the physical viscosity parameter ν and the numerical viscosity introduced by the Engquist–Osher scheme itself. Observe that the last term in (7.48) can be written as

$$\left(|a|\frac{\Delta x}{2} + \nu\right)\frac{\Delta t}{\Delta x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n),$$

which allows us to compare the influence of these two quantities on the numerical solution. In the case $\nu = 0.01$ and when $|a| = 1$, the physical viscosity is of the order of the numerical viscosity introduced by the Engquist–Osher scheme for the inviscid Burgers equation, i.e., $|a|\frac{\Delta x}{2} = 0.01|a| = \nu$. Thus, $\nu = 0.01$ can be interpreted as the numerical solution of the inviscid Burgers equation.

Note that this is not an unusual situation in transonic aerodynamic applications of fluid dynamics problems. In those problems, the thickness of the shock wave is too small to be resolved by a computational mesh. The numerical dissipation dominates the physical one, unless an exceptionally fine mesh is set up. In these cases, it is natural to obtain approximate solutions using numerical methods for inviscid flows (see [Hi88], Chapter 22).

We solve the optimization problem starting either with $u^0 \equiv 0$ or the following:

$$u^0 = \begin{cases} 2 \text{ if } x < 1/4, \\ 0 \text{ if } x \geq 1/4, \end{cases} \tag{7.52}$$

which has a discontinuity at $x = 1/4$. A discontinuous function is suitable for the alternating method while, for the classical adjoint method, a smooth initialization is a priori more natural.

In Figure 7.4 we show numerical experiments for three different values of the viscosity parameters $\nu = 0.5, 0.1$ and $\nu = 0.01$ in different rows. At each row, the left figure corresponds to the initial data u^0 obtained after optimization when the gradient is computed with the adjoint method, initialized with $u^0 \equiv 0$ and the u^0 given in (7.52), as well as the alternating method initialized with the discontinuous function in (7.52). In the figure on the right, the solutions at the final time $t = 1$ are drawn.

In Figure 7.5 the values of the functional versus the iteration are shown for each method and the different values of ν described before.

We see that for large values of the viscosity ν the classical adjoint method starting with the smooth data $u^0 \equiv 0$ is preferable. When ν becomes smaller, the efficiency of the algorithm does not depend very much on the initialization, and both $u^0 \equiv 0$ and the one in (7.52) provide similar results.

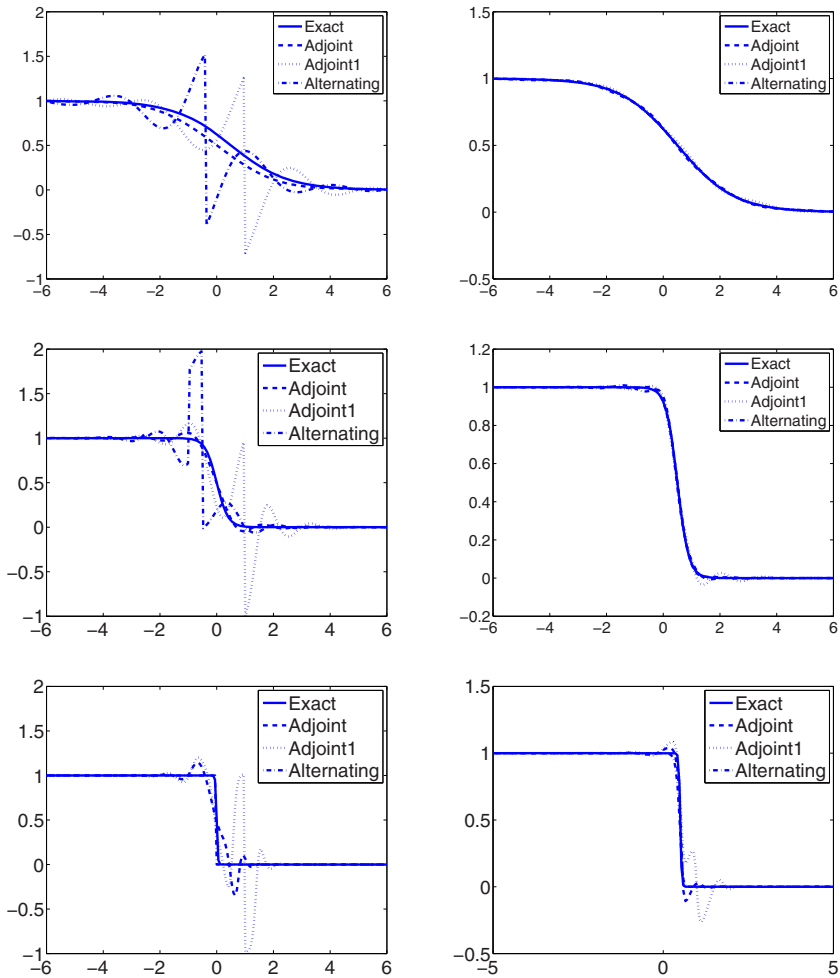


Fig. 7.4. The upper figures correspond to the value of the viscosity $\nu = 0.5$, the middle ones correspond to $\nu = 0.1$ and the lower ones correspond to $\nu = 0.01$. The left figure of each row contains: Exact initial data (Exact), initial data obtained from the descent algorithm with the classical adjoint method initialized with $u^0 = 0$ (Adjoint), the same initialized with (7.52) (Adjoint1) and the alternating descent method described in this paper and initialized with (7.52) (Alternating). The right figure of each row contains: Exact solution at $t = 1$ (Exact) and solutions obtained with the different methods described.

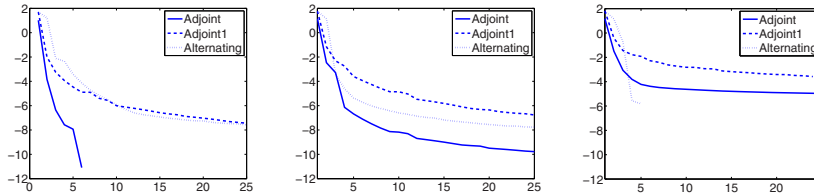


Fig. 7.5. Values of the functional versus iterations of the descent method, for the different methods with viscosities $\nu = 0.5$ (left), $\nu = 0.1$ (middle), and $\nu = 0.01$ (right).

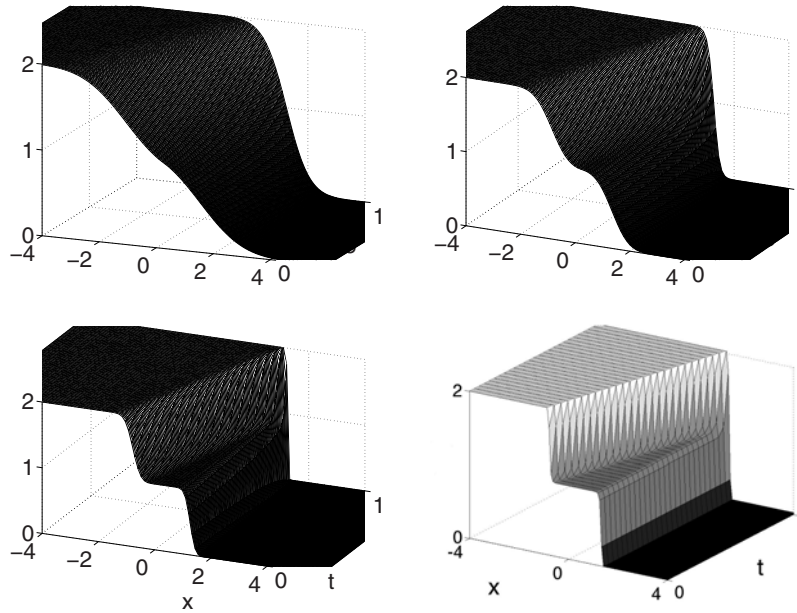


Fig. 7.6. Adjoint solutions corresponding to the solution u in Figure 7.3 for different viscous values $\nu = 0.5$ (upper left), $\nu = 0.1$ (upper right), and $\nu = 0.01$ (lower left) and the exact adjoint solution (lower right).

On the other hand, the alternating descent method is more efficient when the viscosity becomes sufficiently small, especially in those cases where ν is of the order of the numerical dissipation. Let us briefly explain this. In this nonlinear situation, the numerical dissipation is given by $|u|\frac{\Delta x}{2}$. Taking into account that our target is a function which takes values in the interval $[0, 1]$, it is natural to assume that the numerical optimal solution will take values in a neighborhood of $[0, 1]$. Thus, according to our choice of $\Delta x = 0.02$, the

numerical dissipation will be at most of the order of 0.01, depending on the value of the numerical solution u_j^n at each point of the mesh. If we incorporate a physical viscosity $\nu = 0.01$, we introduce a perturbation which is of the order of the maximum value of the numerical dissipation.

In Figure 7.6 it is shown that, in this case, the solutions of the adjoint system are closer to the solutions of the adjoint system for the inviscid equation given in Figure 7.3.

Acknowledgement. This work is supported by Grant MTM2008-03541 of the MEC (Spain).

References

[BaPi02] Bardos, C., Pironneau, O.: A formalism for the differentiation of conservation laws. *C.R. Acad. Sci. Paris Sér I,* **335**, 839–845 (2002).

[BaPi03] Bardos, C., Pironneau, O.: Derivatives and control in presence of shocks. *J. Comput. Fluid Dynamics,* **11**, 383–392 (2003).

[BoJa98] Bouchut, F., James, F.: One-dimensional transport equations with discontinuous coefficients. *Nonlinear Anal. Theory Appl.,* **32**, 891–933 (1998).

[BoJa99] Bouchut, F., James, F.: Differentiability with respect to initial data for a scalar conservation law, in *Proceedings Seventh Internat. Conf. on Hyperbolic Problems,* Birkhäuser, Basel (1999), 113–118.

[BoJa05] Bouchut, F., James, F., Mancini, S.: Uniqueness and weak stability for multi-dimensional transport equations with one-sided Lipschitz coefficient. *Ann. Sc. Norm. Super. Pisa Cl. Sci.,* **4**, 1–25 (2005).

[BrOs88] Brenier, Y., Osher, S.: The discrete one-sided Lipschitz condition for convex scalar conservation laws. *SIAM J. Numer. Anal.,* **25**, 8–23 (1988).

[BrMa95a] Bressan, A., Marson, A.: A variational calculus for discontinuous solutions of systems of conservation laws. *Comm. Partial Diff. Equations,* **20**, 1491–1552 (1995).

[BrMa95b] Bressan, A., Marson, A.: A maximum principle for optimally controlled systems of conservation laws. *Rend. Sem. Mat. Univ. Padova,* **94**, 79–94 (1995).

[CaPa08] Castro, C., Palacios, F., Zuazua, E.: An alternating descent method for the optimal control of the inviscid Burgers equation in the presence of shocks. *Math. Models Methods Appl. Sci.,* **18**, 369–416 (2008).

[CaZu08] Castro, C., Zuazua, E.: On the flux identification problem for scalar conservation laws in the presence of shocks (preprint, 2008).

[DaLe95] Dal Maso, G., Le Floch, P., Murat, F.: Definition and weak stability of nonconservative products. *J. Math. Pures Appl.,* **74**, 458–483 (1995).

[EsVa93] Escobedo, M., Vázquez, J.L., Zuazua, E.: Asymptotic behavior and source-type solutions for a diffusion–convection equation. *Arch. Rational Mech. Anal.,* **124**, 43–65 (1993).

[GaGu01] Garreau, S., Guillaume, P., Masmoudi, M.: The topological asymptotic for PDE systems: the elasticity case. *SIAM J. Control Optim.,* **39**, 1756–1778 (2001).

[GiPi01] Giles, M.B., Pierce, N.A.: Analytic adjoint solutions for the quasi-one-dimensional Euler equations. *J. Fluid Mech.,* **426**, 327–345 (2001).

[Gl03] Glowinski, R.: *Numerical Methods for Fluids. Part 3,* Handbook of Numerical Analysis IX, Ciarlet, P., Lions, J.-L., eds., Elsevier, Amsterdam (2003).

[GoRa99] Godlewski, E., Raviart, P.A.: The linearized stability of solutions of nonlinear hyperbolic systems of conservation laws. A general numerical approach. *Math. Comp. Simulations,* **50**, 77–95 (1999).

[GoRa91] Godlewski, E., Raviart, P.A.: *Hyperbolic Systems of Conservation Laws,* Ellipses, Paris (1991).

[GoOl98] Godlewski, E., Olazabal, M., Raviart, P.A.: On the linearization of hyperbolic systems of conservation laws. Application to stability, in *Équations Différentielles et Applications,* Gauthier-Villars, Paris (1998), 549–570.

[GoRa96] Godlewski, E., Raviart, P.A.: *Numerical Approximation of Hyperbolic Systems of Conservation Laws,* Springer, Berlin (1996).

[GoJa00] Gosse, L., James, F.: Numerical approximations of one-dimensional linear conservation equations with discontinuous coefficients. *Math. Comput.,* **69**, 987–1015 (2000).

[Hi88] Hirsch, C.: *Numerical Computation of Internal and External Flows. Vols. 1 and 2,* Wiley, New York (1988).

[JaSe99] James, F., Sepúlveda, M.: Convergence results for the flux identification in a scalar conservation law. *SIAM J. Control Optim.,* **37**, 869–891 (1999).

[Le02] LeVeque, R.: *Finite Volume Methods for Hyperbolic Problems,* Cambridge University Press, London (2002).

[Ma83] Majda, A.: *The Stability of Multidimensional Shock Fronts,* American Mathematical Society, Providence, RI (1983).

[Me03] Métivier, G.: Stability of multidimensional shocks. Course notes, `http://www.math.u-bordeaux.fr/~metivier/cours.html` (2003).

[MoPi04] Mohammadi, B., Pironneau, O.: Shape optimization in fluid mechanics. *Annual Rev. Fluids Mech.,* **36**, 255–279 (2004).

[NaJa00] Nadarajah, S., Jameson, A.: A comparison of the continuous and discrete adjoint approach to automatic aerodynamic optimization. AIAA Paper 2000-0667, 38th Aerospace Sciences Meeting and Exhibit, January 2000, Reno, NV.

[Ol57] Oleinik, O.: Discontinuous solutions of nonlinear differential equations. *Amer. Math. Soc. Transl.,* **26**, 95–172 (1957).

[Ul03] Ulbrich, S.: Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws. *Systems Control Lett.,* **48**, 313–328 (2003).

[Wh74] Whitham, G.B.: *Linear and Nonlinear Waves,* Wiley, New York (1974).

8

A High-Order Finite Volume Method for Nonconservative Problems and Its Application to Model Submarine Avalanches

M.J. Castro Díaz,[1] E.D. Fernández-Nieto,[2] J.M. González-Vida,[1] A. Mangeney,[3] and C. Parés[1]

[1] Universidad de Málaga, Spain; castro@anamat.cie.uma.es, jgv@uma.es, pares@anamat.cie.uma.es
[2] Universidad de Sevilla, Spain; edofer@us.es
[3] Institut de Physique du Globe de Paris, France; mangeney@ipgp.jussieu.fr

8.1 Introduction

In this chapter we investigate how to apply a high-order finite volume method to discretize the model proposed in [FeBo08] to study submarine avalanches.

The model proposed by Fernández-Nieto et al. in [FeBo08] is an integrated two-layer model of Savage–Hutter type. The upper layer models the fluid, and the second layer is assumed to be constituted by sediment or rocks. The derivation of the model is done by taking into account some physical properties of both layers: density, porosity, friction angle in a Coulomb law, internal friction angle between particles, and buoyancy.

The previous model reduces to the one proposed by Savage and Hutter to study avalanches of granular materials when the height of the water layer tends to zero. In the pioneering works of Savage and Hutter (see [SaHu91]) a model to study avalanches over an inclined slope is proposed. They derive their model by integration of Euler equations and assuming a Coulomb friction law. Bouchut et al. in [BoMa03] propose a generalization of the model in order to take into account more general topographies. In particular, the angle of the bottom with the horizontal is not constant and depends on spatial variables. They show that a new term depending on the curvature is necessary to be introduced in the model in order to preserve stationary solutions and to verify an entropy inequality.

In [FeBo08] a first-order finite volume method is also proposed. In this work we propose a high-order finite volume method to discretize the two-layer Savage–Hutter model. We also study numerically the dependency of the sediment layer profile and the generated tsunami with respect to some of the parameters of the model, such as the friction angle in the Coulomb law and

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_8,

the ratio of densities between both layers. The effective angle of repose of the sediment layer after an avalanche is also measured at the stationary state.

This chapter is organized as follows. In Section 8.2, the submarine avalanche model is briefly presented. In Section 8.3 we summarize how to discretize the model by using a high-order finite volume method by state reconstructions. Finally, in Section 8.4, we present a battery of numerical tests, to investigate the effective angle of repose of the material after an avalanche, and to study the buoyancy effect in the final stationary solution.

8.2 Submarine Avalanche Model

In this section we present the model proposed in [FeBo08]. For brevity, we consider only the case of a bottom with constant slope, and the porosity of the sediment layer is set to zero.

We use the following notation: by h_1 we denote the height of the fluid layer, by q_1 the discharge of the fluid layer, h_2 denotes the height of the sediment layer and q_2 its discharge (see Figure 8.1). We consider that the bottom is an

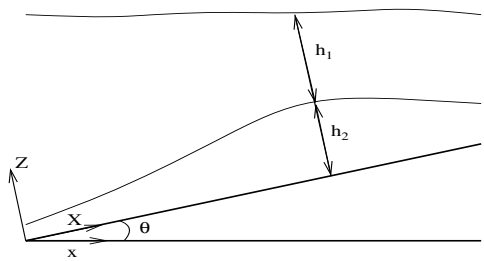


Fig. 8.1. Sketch of the domain.

inclined plane with a constant slope defined by the angle θ. Let us consider local coordinates over the inclined plane, if X is the local spatial variable and x the Cartesian coordinates, $X = x\cos(\theta)$, with $X \in [0, L]$, where L is the length of the domain. t denotes the time variable, usually, $t \in [0, T]$ where T is the final time. The model is defined by

$$\begin{cases} \partial_t h_1 + \partial_X q_1 = 0, \\ \partial_t q_1 + \partial_X \left(\dfrac{q_1^2}{h_1} + g\dfrac{h_1^2}{2}\cos(\theta) \right) = -gh_1\sin(\theta) - gh_1\cos(\theta)\partial_X h_2, \\ \partial_t h_2 + \partial_X q_2 = 0, \\ \partial_t q_2 + \partial_X \left(\dfrac{q_2^2}{h_2} + g\dfrac{h_2^2}{2}\cos(\theta) \right) = -gh_2\sin(\theta) - rgh_2\cos(\theta)\partial_X h_1 + \mathcal{T}, \end{cases} \tag{8.1}$$

where by $\mathcal{T}$, we denote the Coulomb friction term. This term must be understood as follows:

$$\text{if } |\mathcal{T}| \geq \sigma_c \qquad \Rightarrow \qquad \mathcal{T} = -g(1-r)h_2\cos(\theta)\frac{q_2}{|q_2|}\tan(\delta_0), \tag{8.2}$$

$$\text{if } |\mathcal{T}| < \sigma_c \qquad \Rightarrow \qquad q_2 = 0, \tag{8.3}$$

where $\sigma_c = g(1-r)h_2\cos(\theta)$. Moreover, we have denoted $r = \frac{\rho_1}{\rho_2}$, where ρ_1 is the density of the fluid and ρ_2 is the mean density of the sediment layer. Finally, the Coulomb friction term is defined in the function of the friction angle δ_0.

Observe that the presence of the term $(1-r)$ in the definition of the Coulomb friction term is due to the buoyancy effects, which must be taken into account only in the case that the sediment layer is submerged in the fluid. Otherwise, this term must be replaced by 1.

8.3 High-Order Finite Volume Method

In this section we briefly describe how to discretize (8.1) by using a high-order finite volume method with state reconstructions. We apply the method proposed in [CaGa06] with a special treatment of the Coulomb friction term.

Let us rewrite model (8.1) as a hyperbolic system with conservative terms, source terms, and nonconservative products:

$$\partial_t W + \partial_X F(W) = S(W) + B(W)\partial_X W, \tag{8.4}$$

where by W we denote the vector of unknowns, $F(W)$ is a flux function, and $S(W)$ is the source term, which contains the topography terms and the Coulomb friction term. Finally, $B(W)\partial_X W$ contains the coupling terms. Concretely,

$$W = \begin{pmatrix} h_1 \\ q_1 \\ h_2 \\ q_2 \end{pmatrix}, \quad F(W) = \begin{pmatrix} q_1 \\ q_1^2/h_1 + gh_1^2\cos(\theta)/2 \\ q_2 \\ q_2^2/h_2 + gh_2^2\cos(\theta)/2 \end{pmatrix},$$

$S(W) = S_B(W) + S_{\mathcal{T}}(W)$,

$$S_B(W) = \begin{pmatrix} 0 \\ -gh_1\sin(\theta) \\ 0 \\ -gh_2\sin(\theta) \end{pmatrix}, \; S_{\mathcal{T}}(W) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \mathcal{T} \end{pmatrix}$$

and

$$B(W) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -gh_1\cos(\theta) & 0 \\ 0 & 0 & 0 & 0 \\ -rgh_2\cos(\theta) & 0 & 0 & 0 \end{pmatrix}.$$

Due to the nondivergence form of the equations, the notion of solutions in the sense of distributions cannot be used. The theory introduced by Dal Maso, LeFloch, and Murat [DaLe95] is followed here to define weak solutions. This theory is based on the introduction of a family of paths. Therefore, the numerical scheme is also based on the choice of a family of paths. For more details see [Pa06].

To construct the numerical scheme, (8.4) is rewritten as follows: first, the bottom topography $b(X) = \sin(\theta)X$ is defined. Note that S_B can be rewritten in terms of $\partial_X b(X)$. Next, we define $\widetilde{W} = (W, b)^T$ as

$$\partial_t \widetilde{W} + \mathcal{A}(\widetilde{W})\partial_X \widetilde{W} = \widetilde{S}_{\mathcal{T}}(\widetilde{W}), \tag{8.5}$$

where $\mathcal{A}$ is defined in terms of the Jacobian matrix of $F(W)$, $B(W)$, and $S_B/\sin(\theta)$, and $\widetilde{S}_{\mathcal{T}} = (S_{\mathcal{T}}, 0)^T$.

The left-hand side of (8.5) is discretized using a high-order Roe method with state reconstruction introduced in [CaGa06], while $\widetilde{S}_{\mathcal{T}}$ is descretized in a centered way as described in [FeBo08]. Let us consider a partition of the interval $[0, L]$ in cells defined by $I_i = \left[X_{i-1/2}, X_{i+1/2}\right]$. Let us suppose that all of them have the same length ΔX and $X_{i+\frac{1}{2}} = i\Delta X$. $X_i = (i - 1/2)\Delta X$ is the center of the control volume I_i. Let Δt be the time step and $t^n = n\Delta t$.

Then, we denote by $\widetilde{W}_i^n$ an approximation of the mean value of $\widetilde{W}$ over I_i at time $t = t^n$,

$$\widetilde{W}_i^n = \frac{1}{\Delta X}\int_{X_{i-1/2}}^{X_{i+1/2}} \widetilde{W}(X, t^n)dX.$$

Let us define a Roe matrix for system (8.5) (see [To92], [Pa06] for details).

Definition 1. *For a given family of paths Ψ, a function $\mathcal{A} : \Omega \times \Omega \to \mathcal{M}_N$ is a Roe linearization of system (8.5) if it satisfies the following properties:*

1. *For each $\widetilde{W}_L$, $\widetilde{W}_R \in \Omega$, $\mathcal{A}_\Psi(\widetilde{W}_L, \widetilde{W}_R)$ has N real and different eigenvalues.*
2. *$\mathcal{A}_\Psi(\widetilde{W}, \widetilde{W}) = \mathcal{A}(\widetilde{W})$, for all $\widetilde{W} \in \Omega$.*
3. *For each $\widetilde{W}_L$, $\widetilde{W}_R \in \Omega$,*

$$\mathcal{A}_\Psi(\widetilde{W}_L, \widetilde{W}_R)(\widetilde{W}_R - \widetilde{W}_L) = \int_0^1 \mathcal{A}\left[\Psi(s; \widetilde{W}_L, \widetilde{W}_R)\right] \frac{\partial \Psi}{\partial s}(s; \widetilde{W}_L, \widetilde{W}_R)ds. \tag{8.6}$$

Let us denote

$$\mathcal{A}_{i+1/2} = \mathcal{A}_\Psi\left(\widetilde{W}_i^n, \widetilde{W}_{i+1}^n\right) \tag{8.7}$$

the Roe matrix associated to the states $\widetilde{W}_i$ and $\widetilde{W}_{i+1}$, with eigenvalues

$$\lambda_1^{i+1/2} < \lambda_2^{i+1/2} < \cdots < \lambda_N^{i+1/2},$$

and $\left\{R_l^{i+1/2}\right\}_{l=1}^N$ is the base of the associated eigenvectors. By $\mathcal{K}_{i+1/2}$ we denote the $N \times N$ matrix whose columns are eigenvectors and by $\mathcal{L}_{i+1/2}$, the diagonal matrix of eigenvalues. We will also use the following matrices: $\mathcal{L}_{i+1/2}^{\pm} = \mathrm{diag}(\lambda_l^{i+1/2})^{\pm}, l = 1, \ldots, N$, $\mathcal{A}_{i+1/2}^{\pm} = \mathcal{K}_{i+1/2}\mathcal{L}_{i+1/2}^{\pm}\mathcal{K}_{i+1/2}^{-1}$.

A state reconstruction operator P^t is considered, that is, an operator that associates, to a given sequence $\left\{\widetilde{W}_i(t)\right\}$, two new sequences $\left\{\widetilde{W}_{i+1/2}^{-}(t)\right\}$, $\left\{\widetilde{W}_{i+1/2}^{+}(t)\right\}$ in such a way that, whenever

$$\widetilde{W}_i(t) = \frac{1}{\Delta X}\int_{I_i} \widetilde{W}(X,t)dX, \ \forall i$$

for some regular function $\widetilde{W}$, then

$$\widetilde{W}_{i+1/2}^{\pm}(t) = \widetilde{W}(X_{i+1/2}, t) + O(\Delta X^p), \ \forall i.$$

Over each control volume I_i, at each instant $t > 0$, we define a regular function P_i^t such that

$$\lim_{X \to X_{i-1/2}^+} P_i^t(X) = \widetilde{W}_{i-1/2}^{+}(t), \quad \lim_{X \to X_{i+1/2}^-} P_i^t(X) = \widetilde{W}_{i+1/2}^{-}(t). \tag{8.8}$$

The following numerical scheme is considered (see [CaGa06]):

$$\begin{aligned} \widetilde{W}_i' \quad = \quad & -\frac{1}{\Delta X}\Bigg(\mathcal{A}_{i-1/2}^{+}(\widetilde{W}_{i-1/2}^{+}(t) - \widetilde{W}_{i-1/2}^{-}(t)) \\ & + \mathcal{A}_{i+1/2}^{-}(\widetilde{W}_{i+1/2}^{+}(t) - \widetilde{W}_{i+1/2}^{-}(t)) \\ & + \int_{X_{i-1/2}}^{X_{i+1/2}} \mathcal{A}\left[P_i^t(X)\right] \frac{d}{dX} P_i^t(X) dX\Bigg) + \widetilde{S}_{\mathcal{T},i}, \end{aligned} \tag{8.9}$$

where $\widetilde{S}_{\mathcal{T},i}$ is a centered discretization of the Coulomb friction term $\mathcal{T}$ defined by (8.2)–(8.3). See [FeBo08] for more details about the definition of $\widetilde{S}_{\mathcal{T},i}$.

Here, Marquina's local piecewise hyperbolic reconstruction in space (see [Ma94]) is used. For the time discretization a Runge–Kutta third-order total variation diminishing (TVD) scheme has been used. The resulting scheme is third-order accurate in space and time and linearly stable under the usual CFL condition:

$$\frac{\Delta t}{\Delta X} \max\{|\lambda_j^{i+1/2}|, j = 1, \ldots, N\} \leq CFL, \quad \forall i, \tag{8.10}$$

where $CFL \in (0, 1]$.

8.4 Numerical Tests

A battery of numerical tests is presented here to study numerically the dependency of the sediment layer profile and the generated tsunami with respect to the friction angle δ_0 and the ratio of densities, r. The effective angle of repose of the sediment layer after an avalanche is also measured at the stationary state. Let us consider a rectangular channel of 10 m length, centered at the origin, with a flat bottom topography, that is, $\theta = 0$. As an initial condition, we set $q_1 = q_2 = 0$ and

$$h_2(X,0) = \begin{cases} 1, & \text{if } -1 \leq X \leq 1, \\ 0, & \text{otherwise,} \end{cases} \qquad h_1(X,0) = \begin{cases} 1, & \text{if } -1 \leq X \leq 1, \\ 2, & \text{otherwise.} \end{cases}$$

Free boundary conditions are imposed at both channels ends. The CFL parameter is set to 0.8. In the simulations, wet/dry fronts appear. Here, we use the numerical treatment proposed by Castro et al. in [CaFe05].

In Figure 8.2 we compare the final stationary interface that we obtain for three different meshes with $\Delta X \in \{0.1, 0.05, 0.02\}$ for $r = 0.4$ and $\delta_0 = 20^o$. Only some small differences near the "wet/dry" fronts can be observed.

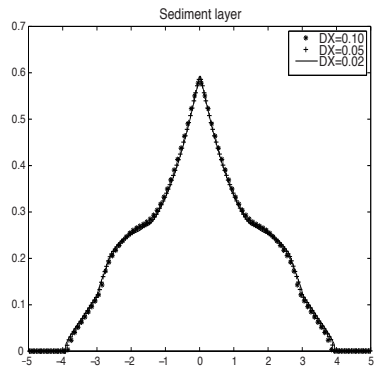


Fig. 8.2. Sediment layer at stationary state $\Delta X \in \{0.1, 0.05, 0.02\}$, ($\delta_0 = 20^o$, $r = 4$).

Table 8.1 shows the maximum and the mean effective angle of repose of the sediment layer after the avalanche at the stationary state. As expected, the maximum value is under $\delta_0 = 20^o$, while the mean value is close to 8.5^o.

Figure 8.3(a) shows the profiles of the sediment layer at the stationary state for $r = 0.4$, $\Delta X = 0.05$, and $\delta_0 \in \{10^o, 15^o, 20^o, 25^o, 30^o\}$, and Table 8.2 shows the maximum and mean effective angle of repose of the sediment layer after the avalanche. As expected, the maximum value is always under δ_0. Figure 8.3(b) shows the maximum of the free surface, $\eta = h_1 + h_2 - 2.0$, vs.

Table 8.1. Effective angle of repose ($r = 0.4$, $\delta_0 = 20^o$).

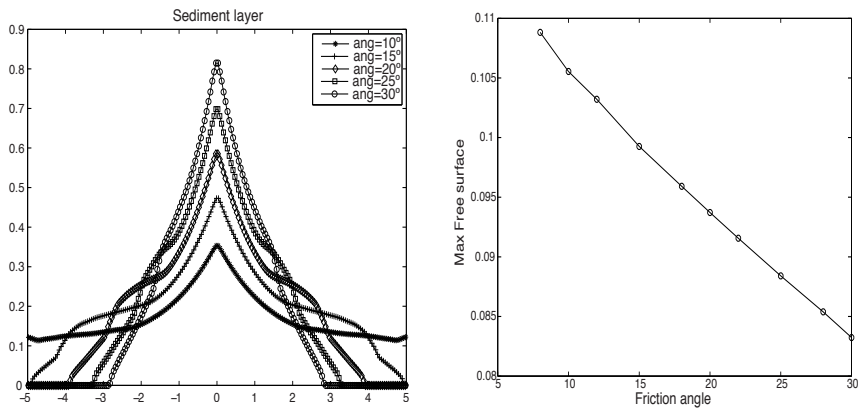
ΔX	max	mean	ΔX	max	mean	ΔX	max	mean
0.1	18.03^o	8.43^o	0.05	19.16^o	8.46^o	0.02	19.69^o	8.46^o

Table 8.2. Effective angle of repose ($r = 0.4$).

δ_0	max	mean	δ_0	max	mean	δ_0	max	mean
10°	9.82^o	2.84^o	15°	14.51^o	5.35^o	20°	19.16^o	8.46^o
25°	23.90^o	12.05^o	30°	29.42^o	16.13^o			

δ_0. Figure 8.3(b) gives an idea of the amplitude of the generated tsunami. Note that the amplitude decreases for bigger values of the parameter δ_0. Finally, Figure 8.4 shows the free surface evolution from $t = 0.2$ s to $t = 0.8$ s, for $\delta_0 \in \{10^o, 15^o, 20^o, 25^o, 30^o\}$. As mentioned before, the amplitudes of the waves are bigger for smaller values of the friction angle δ_0, while the wave speeds are approximately the same for the different values of δ_0 (see Figure 8.4).

Now, the parameter δ_0 is set to 20° and $r \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$. Figure 8.5(a) shows the profiles of the sediment layer at the stationary state for

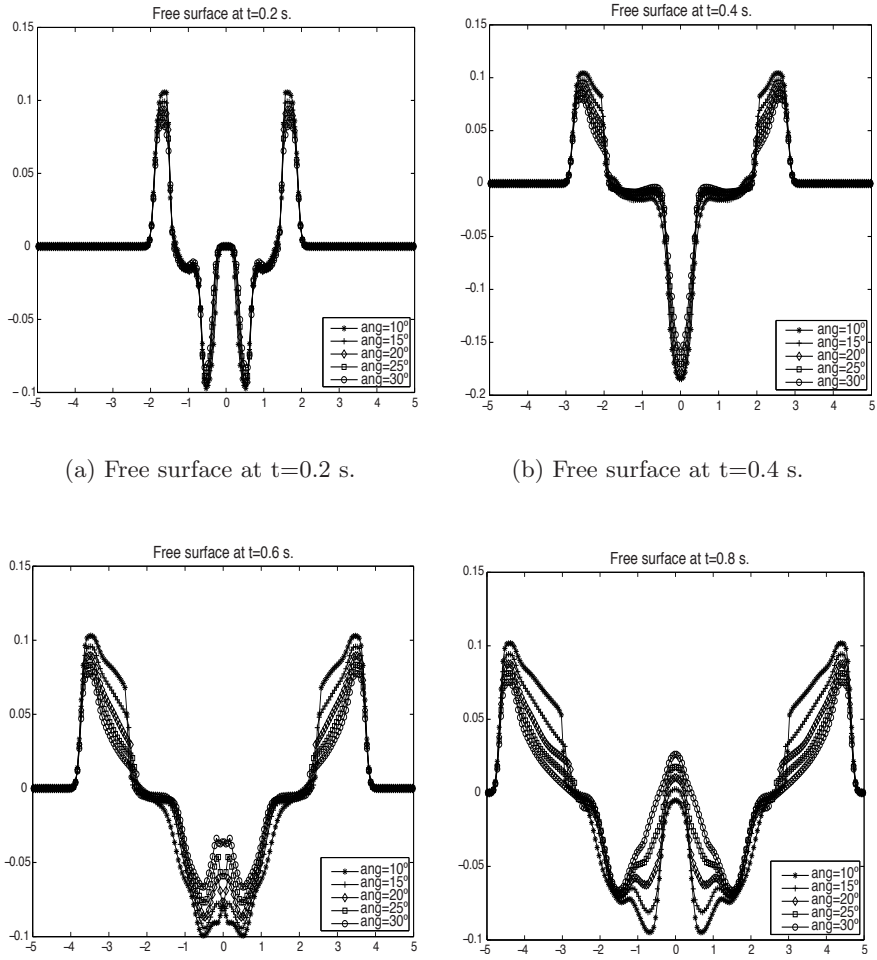


(a) Sediment layer depth at stationary state

(b) Maximal height of the free surface

Fig. 8.3. Sediment layer depth at the stationary state and maximal height of the free surface for $r = 0.4$ and $\delta_0 \in \{10^o, 15^o, 20^o, 25^o, 30^o\}$.

$\Delta X = 0.05$, and Table 8.3 shows the maximum and mean effective angle of repose of the sediment layer after the avalanche.



(a) Free surface at t=0.2 s.

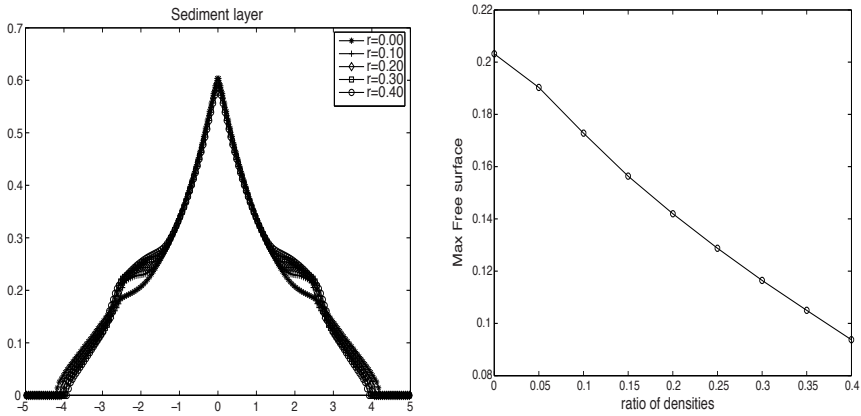
(b) Free surface at t=0.4 s.

(c) Free surface at t=0.6 s.

(d) Free surface at t=0.8 s.

Fig. 8.4. Free surface evolution for $r = 0.4$ and $\delta_0 \in \{10^o, 15^o, 20^o, 25^o, 30^o\}$.

Again, the maximum value is always under δ_0. Note that the maximum value decreases with r while the mean increases with respect to r. Nevertheless, the variations are not significant. More differences can be observed in the stationary profile of the second layer (see Figure 8.5(a)), in particular, the



(a) Sediment layer depth at stationary state

(b) Maximal height of the free surface

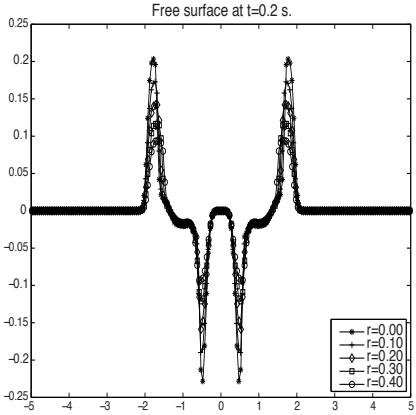
Fig. 8.5. Sediment layer depth at the stationary state and maximal height of the free surface for $\delta_0 = 20^o$ and $r \in \{0, 0.1, 0.2, 0.3, 0.4\}$.

Table 8.3. Effective angle of repose ($\delta_0 = 20^o$).

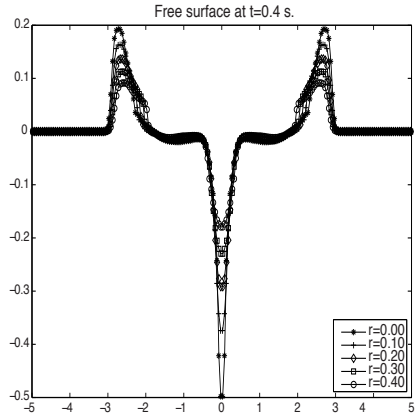
r	max	mean	r	max	mean	r	max	mean
0.0	19.64^o	8.07^o	0.1	19.56^o	8.22^o	0.2	19.37^o	8.30^o
0.3	19.23^o	8.41^o	0.4	19.16^o	8.46^o			

position of the front decreases with r, as well as the maximum height of the sediment layer. Figure 8.5(b) shows the maximum of the free surface, $\eta = h_1 + h_2 - 2.0$, vs. r. As expected, the amplitude of the generated tsunami is bigger for smaller values of r. Finally, Figure 8.6 shows the free surface evolution from $t = 0.2$ s to $t = 0.8$ s, for $r \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$. Note that the wave speeds at the free surface are quite similar, being a bit bigger than those corresponding to $r = 0.0$.

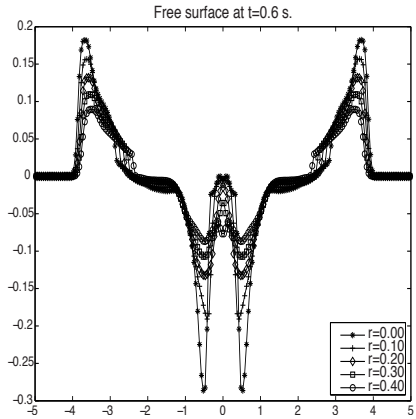
Acknowledgement. This research has been partially supported by the Spanish Government Research project MTM2006-08075. The numerical computations have been performed at the Laboratory of Numerical Methods of the University of Málaga.



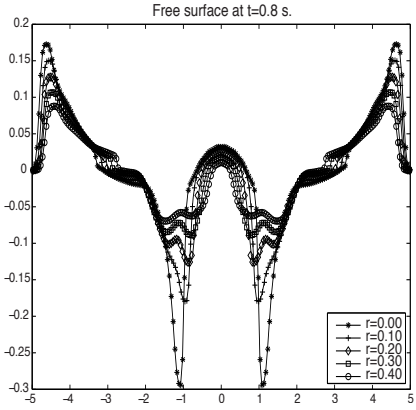
(a) Free surface at t=0.2 s.



(b) Free surface at t=0.4 s.



(c) Free surface at t=0.6 s.



(d) Free surface at t=0.8 s.

Fig. 8.6. Free surface evolution for $\delta_0 = 20^o$ and $r \in \{0, 0.1, 0.2, 0.3, 0.4\}$.

References

[BoMa03] Bouchut, F., Mangeney-Castelnau, A., Perthame, B., Vilotte, J.P.: A new model of Saint–Venant and Savage–Hutter type for gravity driven shallow flows. *C.R. Acad. Sci. Paris Sér I,* **336**, 531–536 (2003).

[CaFe05] Castro, M.J., Ferreiro, A.M., García, J.A., González, J.M., Macías, J., Parés, C., Vázquez, M.E.: On the numerical treatment of wet/dry fronts

in shallow flows: application to one-layer and two-layer systems. *Match Comput. Model.,* **42**, 419–439 (2005).

[CaGa06] Castro, M.J., Gallardo, J.M. Parés, C.: High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. Applications to shallow water systems. *Math. Comput.,* **75**, 1103–1134 (2006).

[DaLe95] Dal Maso, G., LeFloch, P.G., Murat, F.: Definition and weak stability of nonconservative products. *J. Math. Pures Appl.,* **74**, 483–548 (1995).

[FeBo08] Fernández-Nieto, E.D., Bouchut, F., Bresch, D., Castro, M.J., Mangeney, A.: A new Savage–Hutter type model for submarine avalanches and generated tsunami. *J. Comput. Phys.,* **227**, 7720–7754 (2008).

[Ma94] Marquina, A.: Local piecewise hyperbolic reconstructions for nonlinear scalar conservation laws. *SIAM J. Sci. Comput.,* **15**, 892–915 (1994).

[Pa06] Parés, C.: Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM J. Numer. Anal.,* **44**, 300–321 (2006).

[SaHu91] Savage, S.B., Hutter, K.: The dynamics of avalanches of granular materials from initiation to run-out. *Acta Mech.,* **86**, 201–223 (1991).

[To92] Toumi, I.: A weak formulation of Roe approximate Riemann solver. *J. Comput. Phys.,* **102**, 360–373 (1992).

9

Convolution Quadrature Galerkin Method for the Exterior Neumann Problem of the Wave Equation

D.J. Chappell

University of Nottingham, UK; david.chappell@nottingham.ac.uk

9.1 Introduction

The wave equation is important for many real-world applications in time-domain linear acoustics, including scattering from aircraft components and submarines and radiation from loudspeakers. The latter example forms the underlying motivation for the present study. Here the problem is to be solved on an unbounded exterior domain, and so the boundary integral method is a powerful tool for reducing this to an integral equation on the boundary of the radiating or scattering object.

Time-domain boundary integral methods have been employed to solve wave propagation problems since the 1960s [Fr62]. Since then, increasing computer power has made numerical solutions possible over longer run times, and so long-time instabilities in the time marching numerical solutions have become evident [Bi99, Ry85]. A number of methods have been suggested to resolve this such as time averaging [Ry90] and modified time stepping [Bi99]. Using an implicit formulation with high order interpolation and quadrature was also found to give stable results for all practical purposes [Bl96, Do98]. Terrasse *et al.* [HaD03] obtained stable results using a Galerkin approach and used an energy identity to prove stability of the Galerkin approximation. A stable Burton–Miller type integral equation formulation has also been developed in the time domain [ChHa06, Er99]. In addition, the convolution quadrature method of Lubich [Lu88a, Lu88b] has been applied to a number of problems [Ab06, Ban08, Sc01] and has been shown to give stable numerical results. However, computations for the wave equation tend to be for either two-dimensional or very simple three-dimensional cases such as spheres.

In this chapter we consider the convolution quadrature method for the Neumann problem of the wave equation as was recently studied in [Ch08]. Here we summarize the application of this method and give numerical results comparing it with a direct collocation-based Burton–Miller type method. The numerical experiments are given for transient acoustic radiation from a range of axisymmetric structures.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_9,

9.2 Boundary Integral Formulation

Let $\Omega \subset \mathbb{R}^3$ be a finite Lipschitz domain with boundary Γ and let $\Omega_+ = \mathbb{R}^3 \setminus \bar{\Omega}$ denote the unbounded exterior domain, which we assume is filled with a homogeneous acoustic medium with speed of sound c. In this chapter we consider the numerical solution of the wave equation

$$\frac{\partial^2 u}{\partial t^2}(x,t) = c^2 \Delta u(x,t), \quad x \in \Omega_+,\ t \in (0,T) \tag{9.1}$$

with initial conditions

$$u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0, \quad x \in \Omega_+ \tag{9.2}$$

and the Neumann boundary condition

$$\frac{\partial u}{\partial \nu_x}(x,t) = f(x,t), \quad x \in \Gamma,\ t \in (0,T), \tag{9.3}$$

where f denotes the given boundary data and ν_x is the outward unit normal vector to Γ at x. The existence and uniqueness of solutions to this initial-boundary value problem (IBVP) has been established for f belonging to a suitable anisotropic Sobolev space [Bam86]. In order to define these spaces, we first define the Sobolev space $H^\alpha(D)$ in the usual way for $\alpha \in [-k,k]$ with k a positive integer (see, for example, [Mc00]). The value of k depends on the global smoothness of the domain $D \subset \mathbb{R}^3$, with $k = 1$ for the Lipschitz case considered here. Let us denote the norm on these spaces by $\|\cdot\|_{H^\alpha(D)}$. The anisotropic Sobolev space $H^r(\mathbb{R}; H^\alpha(D))$ of order $r \in \mathbb{R}$ is given by $H^r(\mathbb{R}; H^\alpha(D)) = \{g : D \times \mathbb{R} \to \mathbb{R} \mid \int_{\mathbb{R}} (1+|\omega|)^{2r} \|\mathcal{F}g(\cdot,\omega)\|^2_{H^\alpha(D)} < \infty\}$, with $\mathcal{F}$ denoting the Fourier transformation on $\mathbb{R}$. This may be restricted to finite time intervals of the form $(0,T)$ by denoting

$$H^r_0(0,T; H^\alpha(D)) = \{g|_{D\times(0,T)} | g \in H^r(\mathbb{R}; H^\alpha(D)) \text{with} g \equiv 0 \text{ on} D \times (-\infty, 0)\}, \tag{9.4}$$

with notation as in Lubich [Lu94]. Given boundary data

$$f \in H^{r+1}_0(0,T; H^{-1/2}(\Gamma)),$$

there exists a unique solution to the IBVP

$$u \in H^r_0(0,T; H^1(\Omega_+))$$

depending continuously on the data [Bam86, Ch08]. Note that for $r \in \mathbb{Z}$ the condition on the data implies that f and its first r time derivatives vanish at $t = 0$.

Let us consider the solution of the IBVP (9.1)–(9.3) by means of a double-layer potential,

$$\mathcal{D}\varphi(x,t) := \int_0^t \int_\Gamma \frac{\partial G}{\partial \nu_\xi}(x-\xi, t-\tau)\varphi(\xi,\tau)d\Gamma_\xi d\tau, \quad x \in \Omega_+,\, t \in (0,T), \tag{9.5}$$

where G is the fundamental solution of the wave equation

$$G(x,t) = \frac{1}{4\pi|x|}\,\delta\left(t - \frac{|x|}{c}\right), \tag{9.6}$$

δ denotes the Dirac delta distribution, and φ is the unknown layer density. The double-layer potential (9.5) satisfies the wave equation (9.1) and initial conditions (9.2). Let $\varepsilon > 0$, $x \in \Gamma$, and $x' = x + \varepsilon\nu_x \in \Omega_+$. It is well known that the operator

$$W\varphi(x,t) := \nu_x \cdot \lim_{\varepsilon\to 0} \nabla_{x'}\mathcal{D}\varphi(x',t) \tag{9.7}$$

is continuous across Γ [Co04]. Combining this fact with (9.5) and the boundary condition (9.3) yields the following boundary integral equation for the layer density:

$$W\varphi(x,t) = f(x,t), \quad x \in \Gamma,\, t \in (0,T). \tag{9.8}$$

In the definition of W given by (9.7) we may only take the derivative terms inside the integrals in the double-layer potential $\mathcal{D}$ if the resulting integrals are interpreted as a finite part in the sense of Hadamard. This is so because the integrand would then contain an $O(|x-\xi|^{-3})$ singularity. However, this may be reduced to the weakly singular case $O(|x-\xi|^{-1})$ when the Galerkin method is employed for the spatial semi-discretization [Ch08]. Once the boundary integral equation (9.8) has been solved for the layer density, the solution of the IBVP follows from the representation formula (9.5).

9.3 Discretization Methods

9.3.1 Time Discretization: Convolution Quadrature

A direct space-time discretization of equation (9.8) involves the treatment of the Dirac delta distribution. The resulting integration domains are given by the intersection of a light cone of finite width with the spatial boundary elements. Since these integration regions can be of quite general shape, numerical quadrature can become very complicated. These methods also have well-known stability problems. The convolution quadrature method for the time discretization leads to an unconditionally stable scheme, and the integration regions are simply the spatial boundary elements. We do not detail the theoretical framework here (see [Ch08, Lu88a, Lu88b, Lu94]), but summarize the application of the method.

For the time discretization of (9.8) we split the time interval $[0,T]$ into $N+1$ equal time steps of length $\Delta t = T/N$ and compute an approximate solution

at the discrete time levels $t_n = n\Delta t$ for $n = 0, 1, 2, \ldots, N$. Following [Ch08, Lu88a, Lu88b, Lu94], the convolution quadrature method is based on a linear multistep method which, for differential equations $\Phi'(t) = g(\Phi(t))$, can be formulated as

$$\sum_{j=0}^{k} \alpha_j \Phi_{n+j-k} = \Delta t \sum_{j=0}^{k} \beta_j g(\Phi_{n+j-k}), \tag{9.9}$$

where $\Phi_n = \Phi(t_n)$. Let

$$\gamma(\zeta) := \frac{\sum_{j=0}^{k} \alpha_j \zeta^{k-j}}{\sum_{j=0}^{k} \beta_j \zeta^{k-j}} \tag{9.10}$$

be the quotient of generating polynomials of the linear multistep method (9.9).

The continuous convolution operator W is replaced by the discrete convolution operator

$$(W_{\Delta t}\varphi_{\Delta t})(\cdot, n\Delta t) = \sum_{j=0}^{n} w_{n-j}(\Delta t, \widetilde{W})\varphi_j, \tag{9.11}$$

where $\varphi_j = \varphi_{\Delta t}(\cdot, j\Delta t)$ and $\widetilde{W}$ denotes the Laplace transform of W. Here the "quadrature weights" or convolution coefficients are linear operators

$$w_n(\Delta t, \widetilde{W}) : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$$

defined by their z-transform

$$\sum_{n=0}^{\infty} w_n(\Delta t, \widetilde{W})\zeta^n = \widetilde{W}\left(\cdot, \frac{\gamma(\zeta)}{\Delta t}\right), \qquad |\zeta| < 1. \tag{9.12}$$

We employ the second order accurate, A-stable backward difference formula (BDF2) method with

$$\gamma(\zeta) = \frac{1}{2}(\zeta^2 - 4\zeta + 3).$$

9.3.2 Space Discretization: Galerkin Boundary Element Method

In the previous section we derived the semi-discrete problem: For $n = 1, 2, ..., N$, find $\varphi_n \in H^{1/2}(\Gamma)$ such that

$$\sum_{j=0}^{n} w_{n-j}(\Delta t, \widetilde{W})\varphi_j = f(\cdot, n\Delta t). \tag{9.13}$$

For the space discretization we let Γ be discretized by a regular boundary element mesh in the sense of Ciarlet [Ci87] with element diameter Δx. Let $X_{\Delta x} \subset H^{1/2}(\Gamma)$ denote a family of finite approximation spaces consisting of piecewise polynomials of degree $k \geq 1$.

For the Galerkin boundary element method we replace φ_n in (9.13) by some $\varphi^n = \varphi_\Delta(\cdot, n\Delta t) \in X_{\Delta x}$ of the form

$$\varphi^n = \sum_{i=1}^{N_x} \phi_i^n b_i, \tag{9.14}$$

where b_i, $i = 1, \ldots, N_x$ are the basis functions chosen for $X_{\Delta x}$ and $(\phi_i^n)_{i=1}^{N_x} = \phi^n \in \mathbb{R}^{N_x}$ is the vector of coefficients to be determined. Applying the Galerkin method, we impose the integral equation in a weak form, which yields the fully discrete problem: For $n = 1, 2, ..., N$, find $\varphi^n \in X_{\Delta x}$ of the form (9.14) such that

$$\sum_{j=0}^{n} \sum_{i=1}^{N_x} \phi_i^j \int_\Gamma w_{n-j}(\Delta t, (\widetilde{W} b_i)(x,s))\, b_k(x)\, d\Gamma_x = \int_\Gamma f(x, n\Delta t) b_k(x) d\Gamma_x \tag{9.15}$$

for all $k = 1, \ldots, N_x$ and $n = 0, \ldots, N$. Here s is the Laplace transform frequency parameter. This can be written in the form of a recursion

$$\sum_{j=0}^{n} A^{n-j} \phi^j = f^n, \qquad n = 0, \ldots, N, \tag{9.16}$$

where the entries of the matrices A^n are given by

$$(A^n)_{k\,i} = w_n \left(\Delta t, \int_\Gamma (\widetilde{W} b_i)(x,s) b_k(x) d\Gamma_x \right), \tag{9.17}$$

and

$$(f^n)_k = \int_\Gamma f(x, n\Delta t) b_k(x) d\Gamma_x.$$

We may evaluate the integral

$$\int_\Gamma (\widetilde{W} b_i)(x,s) b_k(x) d\Gamma_x$$

using the weakly singular formula given in [Bam86]. We find that this integral is equivalent to

$$\int_\Gamma \int_\Gamma \left\{ \mathrm{curl}_\Gamma\, b_i(\xi) \cdot \mathrm{curl}_\Gamma\, b_k(x) + \frac{s^2}{c^2} (\nu_x \cdot \nu_\xi) b_i(\xi) b_k(x) \right\} \frac{e^{-s|x-\xi|/c}}{4\pi|x-\xi|}\, d\Gamma_\xi\, d\Gamma_x, \tag{9.18}$$

where the tangential curl operator is defined by

$$\mathrm{curl}_\Gamma\, \psi(x) = \nu_x \wedge \nabla \hat{\psi}(x). \tag{9.19}$$

Here $\hat{\psi}$ is defined in a tubular neighborhood of Γ, constant along each line normal to Γ and equal to ψ at the intersection point. Hence, the hypersingular operators W and $\widetilde{W}$ do not appear directly in our method, since we can use the weakly singular formula (9.18) instead.

9.3.3 Error Analysis

We give a theorem from [Ch08, Lu94] giving the convergence rate of the space-time discretization.

Theorem 1. *For smooth compatible data f, the fully discrete method (9.15) is unconditionally convergent of optimal order*

$$\|\varphi^n - \varphi(\cdot, n\Delta t)\|_{H^{1/2}(\Gamma)} \leq C_T(\Delta t^2 + \Delta x^{k+\frac{1}{2}}) \tag{9.20}$$

uniformly on $[0, T]$.

Therefore, choosing $\Delta x^{k+\frac{1}{2}} \propto \Delta t^2$ means that the convergence rates in the temporal and spatial discretizations match. The convolution coefficients are evaluated approximately using the trapezoidal rule to approximate the Cauchy integral given by the inverse z-transform

$$w_n(\Delta t, A) = \frac{1}{2\pi i} \int_{|\zeta|=\rho} A\left(\frac{\gamma(\zeta)}{\Delta t}\right) \zeta^{-n-1} d\zeta \tag{9.21}$$

as suggested in [Lu94], where $\rho \in (0, 1)$ is a parameter to be fixed. This computation may be done very efficiently using fast Fourier transform techniques. Applying the trapezoidal rule to (9.21) yields

$$w_n(\Delta t, A) \approx \hat{w}_n(\Delta t, A) = \frac{\rho^{-n}}{L} \sum_{l=0}^{L-1} A\left(\frac{\gamma(\zeta_l)}{\Delta t}\right) e^{-2\pi i n l/L}, \tag{9.22}$$

for $n = 0, \ldots, N$ with $\zeta_l = \rho e^{2\pi i l/L}$. The errors due to the trapezoidal approximation, numerical integration procedures, and for the approximate evaluation of the exterior solution u have all been considered in [Ch08]. This work is based on similar perturbation analysis to that in [Ha08], and the result is that the optimal $O(\Delta t^2)$ convergence rate is attainable with sufficiently small Δt for a suitably designed numerical scheme. Supporting computations are also given to demonstrate the results in practice.

9.4 Numerical Experiments: Comparison with an Alternative Method

We present results comparing the convolution quadrature approach with a direct Burton–Miller type integral equation model as detailed in [ChHa06]. The discretization of the Burton–Miller approach was done using a full space-time collocation method with piecewise cubic polynomials in time and piecewise constants in space. The discretization of the convolution quadrature method is done using the BDF2 multistep method in time, as detailed earlier, and piecewise linear polynomials in space. Clearly, these choices are not designed

to compare equivalent orders of interpolation, but simply compare two existing models in terms of their accuracy and efficiency when computed using MATLAB on a Pentium 4 PC. In both cases the spatial order of interpolation is the simplest permitted, and the temporal order of interpolation is chosen to give a good level of accuracy. For the convolution quadrature method this is the highest order temporal approximation that is permitted by the theory.

The examples studied in this work are axisymmetric, as this simplifies the calculations. Figure 9.1 shows the generating curves for the three surfaces considered. The generating curve for the peanut is defined by three unit circles whose centers lie on an equilateral triangle. Starting from the point $(0, 2)$ at the top of the curve and moving along it in a clockwise direction, the first $2/5$ of the arc length is formed by an arc from a circle centered at $(0, 1)$, the next $1/5$ from a circle centered at $(\sqrt{3}, 0)$, and the final $2/5$ from a circle centered at $(0, -1)$. All meshes used to approximate Γ are exact geometrical representations defined in terms of straight line segments and arcs of circles. Note that the element diameter Δx should be restricted to the generating curves, since this is the only part of Γ where boundary element approximation is employed directly. The solution on the rest of Γ is calculated as a consequence of the axisymmetry assumption.

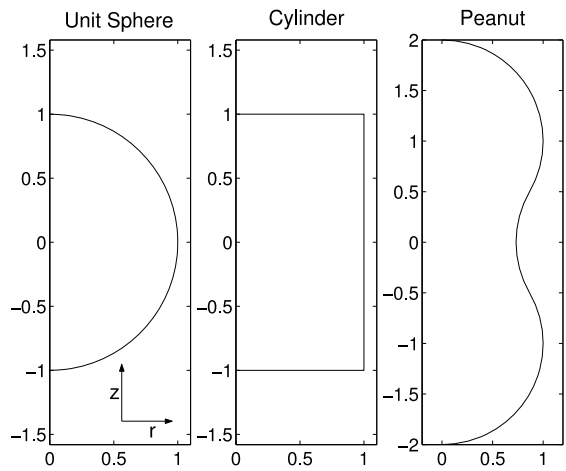


Fig. 9.1. Axisymmetric surface generating curves.

Consider the radiation of a spherically symmetric wave defined by

$$u(R,t) = \frac{H(ct+a-R)}{R}\left(\frac{2}{3} - \cos\left(\frac{\pi(R-ct+5a)}{3a}\right) + \frac{2}{5}\cos\left(\frac{2\pi(R-ct+5a)}{3a}\right)\right.$$
$$\left. -\frac{1}{15}\cos\left(\frac{\pi(R-ct+5a)}{a}\right)\right) \quad e^{-((R-ct+2a)/(3a))^4}. \tag{9.23}$$

Here a is the radius of some sphere $\mathbb{S} \subseteq \Omega$, R is the distance from the center of $\mathbb{S}$ to some point $x \in \Omega_+$, and H is the Heaviside step function. It may be verified that (9.23) satisfies equations (9.1) and (9.2) from the initial-boundary value problem. The required boundary data may be calculated from (9.23) using the chain rule and simplifications due to the axisymmetric geometry to give

$$\frac{\partial u}{\partial r} = \frac{\partial u}{\partial R}\sin\theta, \qquad \frac{\partial u}{\partial z} = \frac{\partial u}{\partial R}\cos\theta,$$

where r and z are the coordinate axes shown in Figure 9.1 and θ is the angle measured clockwise from the positive z-axis. Hence, the boundary data is given by

$$f = \frac{\partial u}{\partial \nu} = \nu_r \frac{\partial u}{\partial r} + \nu_z \frac{\partial u}{\partial z} = \frac{\partial u}{\partial R}[\nu_r \sin\theta + \nu_z \cos\theta], \tag{9.24}$$

where ν_r and ν_z are the components of ν in the r and z directions, respectively.

The numerical approximations for u are compared with the exact solution at two different points $x \in \Omega_+$. The error is calculated using

$$\mathrm{Err}(x) = \left(\Delta t \sum_{n=0}^{N} |\bar{u}_\Delta(x, n\Delta t) - u(x, n\Delta t)|^2 \right)^{1/2}, \tag{9.25}$$

where $\bar{u}_\Delta$ is the approximation to u. Figure 9.2 shows a plot of the exact solution u at the points $x_1 = (0, 0, 3)$ and $x_2 = (2, 0, 2)$.

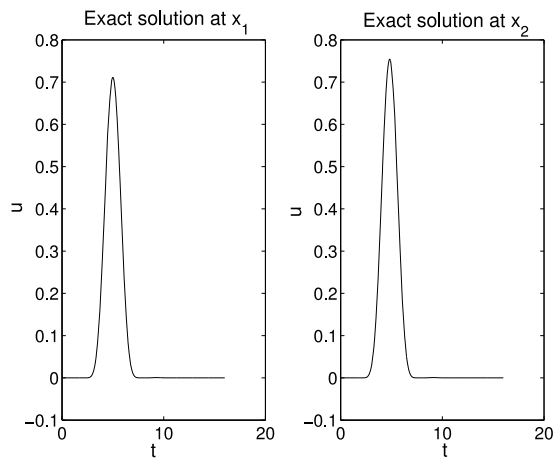


Fig. 9.2. $u(R, t)$ at x_1 and x_2 for $t \in [0, 16]$ and $a = c = 1$.

The first example to be considered is that of a unit sphere. Table 9.1 gives the results for Err with $T = 25.6$, $a = c = 1$, $\rho = \Delta t^{9/N}$, $L = N$, and $N_x = 10$.

Table 9.1. Results for radiation by the unit sphere.

Method	Δt	Err(x_1)	Rate	Err(x_2)	Rate	Time (s)
CQM	0.4	0.668	-	0.645	-	41
B&M	0.4	0.0686	-	0.0611	-	6000
CQM	0.2	0.235	1.51	0.219	1.56	86
B&M	0.2	0.0113	2.60	0.0112	2.45	11000
CQM	0.1	0.0618	1.93	0.0571	1.94	180
B&M	0.1	0.00401	1.49	0.00512	1.13	22000 (*)

(*) 20700 s computing the integral operator containing the data—this may be reduced if the data is zero for $t > t_0$ say.

The results for the other structures are given in Table 9.2. In both cases $c = 1$, $L = N$, and $\rho = \Delta t^{9/N}$. For the cylinder, $T = 25.6$, $N_x = 20$, $a = 1$, and $\Delta t = 0.1$. For the peanut-shaped object, $T = 128/7$, $a = \sqrt{3}-1$, $N_x = 30$, and $\Delta t = 1/14$.

Table 9.2. Results for radiation by the cylinder and peanut.

Geometry	Method	Err(x_1)	Err(x_2)	Time (s)
Cylinder	CQM	0.0563	0.0564	1700
Cylinder	B&M	0.0247	0.0146	58000
Peanut	CQM	0.0376	0.0664	5400
Peanut	B&M	0.0138	0.00823	109000

Considering both methods with the same time step and number of boundary elements, we see that the convolution quadrature method (CQM) is faster with known convergence rates, whereas the Burton–Miller method (B&M) is more accurate for the examples considered here. The difference in accuracy is less pronounced in the non-spherical cases. The Burton–Miller method could perform faster in the case of time-limited data, although there is currently no available error analysis.

References

[Ab06] Abreu, A.I., Mansur, W.J., Carrer, J.A.: Initial conditions contribution in a BEM formulation based on the convolution quadrature method. *Internat. J. Numer. Methods Engng.,* **67**, 417–434 (2006).

[Bam86] Bamberger, A., Ha-Duong, T.: Formulation variationnelle pour le calcul de la diffraction d'une onde acoustique par une surface rigide. *Math. Methods Appl. Sci.,* **8**, 598–608 (1986).

[Ban08] Banjai, L., Sauter, S.: Rapid solution of the wave equation on unbounded domains. *SIAM J. Numer. Anal.,* **47**, 227–249 (2008).

[Bi99] Birgisson, B., Siebrits, E., Pierce, A.P.: Elastodynamic direct boundary element methods with enhanced numerical stability properties. *Internat. J. Numer. Methods Engng.,* **46**, 871–888 (1999).

[Bl96] Bluck, M.J., Walker, S.P.: Analysis of three-dimensional transient acoustic wave propagation using the boundary integral equation method. *Internat. J. Numer. Methods Engng.,* **39**, 1419–1431 (1996).

[Ch08] Chappell, D.J.: A convolution quadrature Galerkin boundary element method for the exterior Neumann problem of the wave equation. *Math. Meth. Appl. Sci.,* **32**:1585–1608 (2009).

[ChHa06] Chappell, D.J., Harris, P.J., Henwood, D., Chakrabarti, R.: A stable boundary element method for modeling transient acoustic radiation. *J. Acoust. Soc. Amer.,* **120**, 74–80 (2006).

[Ci87] Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam (1987).

[Co04] Costabel, M.: Time-dependent problems with the boundary integral equation method, in *Encyclopedia of Computational Mechanics,* Stein, E., De Borst, R., Hughes, T.J.R., eds., Wiley, New York (2004).

[Do98] Dodson, S.J., Walker, S.P., Bluck, M.J.: Implicitness and stability of time domain integral equation scattering analysis. *Appl. Computat. Electromagnetics Soc. J.,* **13**, 291–301 (1998).

[Er99] Ergin, A.A., Shanker, B., Michielssen, E.: Analysis of transient wave scattering from rigid bodies using a Burton-Miller approach. *J. Acoust. Soc. Amer.,* **106**, 2396–2404 (1999).

[Fr62] Friedman, M.B., Shaw, R.: Diffraction of pulses by cylindrical objects of arbitrary cross section. *J. Appl. Mech.,* **29**, 40–46 (1962).

[Ha08] Hackbusch, W., Kress, W., Sauter, S.: Sparse convolution quadrature for time domain boundary integral formulations of the wave equation. *IMA J. Numer. Anal.,* **29**:158–179 (2009).

[HaD03] Ha-Duong, T., Ludwig, B., Terrasse, I.: A Galerkin BEM for transient acoustic scattering by an absorbing obstacle. *Internat. J. Numer. Methods Engng.,* **57**, 1845–1882 (2003).

[Lu88a] Lubich, C.: Convolution quadrature and discretized operational calculus. I. *Numer. Math.,* **52**, 129–145 (1988).

[Lu88b] Lubich, C.: Convolution quadrature and discretized operational calculus. II. *Numer. Math.,* **52**, 413–425 (1988).

[Lu94] Lubich, C.: On the multistep time discretization of linear initial-boundary value problems and their boundary integral equations. *Numer. Math.,* **67**, 365–369 (1994).

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

[Ry85] Rynne, B.P.: Stability and convergence of time marching methods in scattering problems. *IMA J. Appl. Math.,* **35**, 297–310 (1985).

[Ry90] Rynne, B.P., Smith, P.D.: Stability of time marching algorithms for the electric field integral equation. *J. Electromagnetic Waves Appl.,* **4**, 1181–1205 (1990).

[Sc01] Schantz, M.: Application of 3D time domain boundary element formulation to wave propagation in poroelastic solids. *Engng. Anal. Boundary Elements,* **25**, 363–376 (2001).

10

Solution Estimates in Classical Bending of Plates

I. Chudinovich,[1] C. Constanda,[1] D. Doty,[1] and A. Koshchii[2]

[1] University of Tulsa, OK, USA; igor-chudinovych@utulsa.edu, christian-constanda@utulsa.edu, dale-doty@utulsa.edu
[2] International Solomon University, Kharkiv, Ukraine; af-koshchii@rambler.ru

10.1 Introduction

Kirchhoff's classical theory of bending of elastic plates is widely used in mechanical engineering for the mathematical modeling of structures consisting of thin elements. Since most of the solutions in such problems are found computationally, it is very useful to have a tool that provides tight a priori estimates for the error. In this chapter, we construct an algorithm that generates such estimates by means of what is called a dual functional. The argument is constructed variationally and is illustrated by means of a numerical example.

Similar methods for the plate model with transverse shear deformation have been developed in [ChCoKo00] and [ChEtAl06]. A full mathematical study of the static and dynamic bending within the framework of this model can be found in [ChCo00] and [ChCo05], respectively.

10.2 Formulation of the Problem

Consider a thin plate occupying a region $\bar{S}\times[-h/2,h/2]$, where S is a bounded domain in $\mathbb{R}^2$ with a simple, closed, smooth boundary ∂S and $0<h\ll \operatorname{diam} S$. In Kirchhoff's model, the displacement field is of the form

$$v(x') = \big(v_1(x'), v_2(x'), v_3(x')\big)^T,$$
$$v_\alpha(x') = -x_3\partial_\alpha u(x), \quad v_3(x') = u(x),$$

where $x' = (x, x_3)$, $x = (x_1, x_2)$, and the superscript T indicates transposition.

We examine the boundary value problem (D_0) [TiWo87], which consists in finding $u \in C^4(S)\cap C^1(\bar{S})$ such that

$$\begin{gathered} D\Delta^2 u(x) - h\operatorname{div}(T(x)\nabla u(x)) + \varkappa u(x) = q(x), \quad x\in S, \\ u(x) = \partial_\nu u(x) = 0, \quad x\in\partial S. \end{gathered} \tag{10.1}$$

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_10,

In the first equation (10.1), the second term on the left-hand side accounts for a middle-plane pre-existing stress system in equilibrium, the third one represents the response of an elastic foundation, D, h, and $\varkappa$ are material constants, q is the load, Δ is the Laplacian, and ∇ is the gradient operator. The second line in (10.1) describes clamped-edge boundary conditions, with ∂_ν denoting the derivative in the direction of the outward normal to ∂S.

Equation (10.1) can be written in the simpler form

$$\Delta^2 u(x) - \operatorname{div}(T(x)\nabla u(x)) + \varkappa u(x) = q(x), \tag{10.2}$$

where T incorporates the factor h/D and $\varkappa$ and q incorporate $1/D$. It is equation (10.2) that we are referring to in the subsequent analysis.

10.3 Function Spaces

We denote the inner product and norm in $L^2(S)$ by $(\cdot\,,\cdot)_{0;S}$ and $\|\cdot\|_{0;S}$, and by $(\cdot\,,\cdot)_0$ and $\|\cdot\|_0$ if $S = \mathbb{R}^2$. The following spaces of distributions are essential in our considerations.

$H_m(\mathbb{R}^2)$, $m \in \mathbb{R}$: the standard Sobolev space, with norm

$$\|u\|_m^2 = \int\limits_{\mathbb{R}^2} (1 + |\xi|^2)^m |\tilde{u}(\xi)|^2 \, d\xi;$$

$H_{-m}(\mathbb{R}^2)$: the dual of $H_m(\mathbb{R}^2)$ with respect to the duality generated by the inner product in $L^2(\mathbb{R}^2)$;

$\mathring{H}_m(S)$: the subspace of $H_m(\mathbb{R}^2)$ of all u with $\operatorname{supp} u \subset \bar{S}$;

$H_m(S)$: the space of the restrictions to S of all $u \in H_m(\mathbb{R}^2)$, with norm

$$\|u\|_{m;S} = \inf_{v \in H_m(\mathbb{R}^2): v|_S = u} \|v\|_m;$$

$H_{-m}(S)$: the dual of $\mathring{H}_m(S)$ with respect to the duality generated by the inner product $(\cdot\,,\cdot)_{0;S}$;

$H_m(\partial S)$: the standard space of distributions on ∂S, with norm $\|\cdot\|_{m;\partial S}$;

$H_{-m}(\partial S)$: the dual of $H_m(\partial S)$ with respect to the duality generated by the inner product $(\cdot\,,\cdot)_{0;\partial S}$ in $L^2(\partial S)$.

Let γ be the trace operator that maps $H_2(S)$ continuously to $H_{3/2}(\partial S) \times H_{1/2}(\partial S)$ according to the formula

$$(\gamma u)(x) = \{u(x), \partial_\nu u(x)\}, \quad x \in \partial S.$$

The bilinear form of the energy density is

$$a(u, v) = \int\limits_S \{(\Delta u)(\Delta v) + (T\nabla u, \nabla v) + \varkappa uv\} \, dx.$$

We consider the variational version of (D_0), which consists in finding $u \in \mathring{H}_2(S)$ such that

$$a(u,v) = (q,v)_{0;S} \quad \forall v \in \mathring{H}_2(S).$$

If the elements $t_{\alpha\beta}$ of the (2×2)-matrix T satisfy $t_{\alpha\beta} = t_{\beta\alpha} \in L^\infty(S)$, then the following assertions hold.

Theorem 1. *The bilinear form $a(u,v)$ is*

(i) symmetric:

$$a(u,v) = a(v,u) \quad \forall u,v \in \mathring{H}_2(S);$$

(ii) continuous on $\mathring{H}_2(S)$:

$$|a(u,v)| \le M\|u\|_2\|v\|_2 \quad \forall u,v \in \mathring{H}_2(S);$$

(iii) coercive on $\mathring{H}_2(S)$:

$$a(u,u) \ge \delta\|u\|_2^2 \quad \forall u \in \mathring{H}_2(S).$$

Theorem 2. *For any $q \in H_{-2}(S)$, problem (D_0) has a unique solution $u \in \mathring{H}_2(S)$, which satisfies*

$$\|u\|_2 \le c\|q\|_{-2;S}.$$

The classical nonhomogeneous problem (D) consists in finding $u \in C^4(S) \cap C^1(\bar{S})$ such that

$$\begin{gathered}\Delta^2 u(x) - \operatorname{div}(T(x)\nabla u(x)) + \varkappa u(x) = q(x), \quad x \in S,\\ u(x) = f_1(x), \quad \partial_\nu u(x) = f_2(x), \quad x \in \partial S.\end{gathered}$$

In the corresponding variational problem (D), we seek $u \in H_2(S)$ such that

$$\begin{gathered}a(u,v) = (q,v)_{0;S} \quad \forall v \in \mathring{H}_2(S),\\ \gamma u = f.\end{gathered}$$

Theorem 3. *For any $q \in H_{-2}(S)$ and $f \in H_{3/2}(\partial S) \times H_{1/2}(\partial S)$, problem* (D) *has a unique solution $u \in H_2(S)$, which satisfies*

$$\|u\|_{2;S} \le c(\|q\|_{-2;S} + \|f_1\|_{3/2;\partial S} + \|f_2\|_{1/2;\partial S}).$$

We now consider a general abstract variational problem (D_0) in which, for a bilinear form $a(u,v)$ and a linear functional $L(v)$ on a real separable Hilbert space H, we want to find $u \in H$ such that

$$a(u,v) = L(v) \quad \forall v \in H.$$

Theorem 4. *If $a(u,v)$ is continuous and coercive on H, the abstract problem (D_0) has a unique solution $u_0 \in H$, which minimizes the energy functional*

$$J^0(u) = \tfrac{1}{2}\, a(u,u) - L(u).$$

Corollary 1. *For approximate solutions* $\{u_n\}_{n=1}^{\infty}$ *constructed by means of the Galerkin method,*

$$\|u_n - u_0\|^2 \le 2c_2^{-1}[J^0(u_n) - J^0(u_0)],$$
$$a(u - u_0, u - u_0) \le 2[J^0(u_n) - J^0(u_0)].$$

The difficulty in this procedure is that $J^0(u_0)$ is not known. To eliminate it, we design a dual functional for $J^0(u)$ whose maximum (dual extremal problem) coincides with $J^0(u_0)$, and then construct a sequence

$$\{J_n\}_{n=1}^{\infty}, \quad J_n \le J^0(u_0), \quad J_n \to J^0(u_0).$$

10.4 The Dual Functional for $T = 0$ and $\varkappa > 0$

In problem (D_0), we seek $u_0 \in \mathring{H}_2(S)$ such that

$$a(u_0, v) = (q, v)_{0;S} \quad \forall v \in \mathring{H}_2(S),$$

where

$$a(u, v) = \int_S \{(\Delta u)(\Delta v) + \varkappa uv\}\, dx.$$

The solution u_0 minimizes the energy functional

$$J^0(u) = \frac{1}{2}\int_S \{(\Delta u)^2 + \varkappa u^2 - 2qu\}\, dx.$$

On the set

$$\mathcal{U} = \{v \in L^2(S) : \Delta v - q \in L^2(S)\},$$

we define the dual functional

$$I_0(v) = -\frac{1}{2}\int_S \{v^2 + \varkappa^{-1}(q - \Delta v)^2\}\, dx.$$

Theorem 5. *If* $v_0 = \Delta u_0$, *then*

$$\inf_{u \in \mathring{H}_2(S)} J^0(u) = J^0(u_0) = I_0(v_0) = \sup_{v \in \mathcal{U}} I_0(v).$$

In problem (D) with $f = \{f_1, f_2\} \in H_{3/2}(\partial S) \times H_{1/2}(\partial S)$, on $\mathcal{U}$ we define the dual functional

$$J_0(v) = -\frac{1}{2}\int_S \{v^2 + \varkappa^{-1}(q - \Delta v)^2\}\, dx + (\tau_\nu v, f)_{0;\partial S},$$
$$v \in \mathcal{U}, \quad \tau_\nu v = \{-\partial_\nu v, v\}.$$

Theorem 6. *If* $u_0 \in H_2(S)$ *is a solution of* (D) *with* $q \in L^2(S)$ *and* $v_0 = \Delta u_0$, *then*

$$\inf_{u \in \mathcal{U}_f} J^0(u) = J^0(u_0) = J_0(v_0) = \sup_{v \in \mathcal{U}} J_0(v).$$

10.5 The Dual Functional for $\varkappa > 0$ and T Positive Definite

The solution $u_0 \in \mathring{H}_2(S)$ of (D_0) minimizes on $\mathring{H}_2(S)$ the energy functional

$$J^0(u) = \frac{1}{2}\int\limits_S \{(\Delta u)^2 + (T\nabla u, \nabla u) + \varkappa u^2 - 2qu\}\, dx.$$

Here, on the set

$$\mathcal{U} = \{(\lambda, \mu),\ \mu = (\mu_1, \mu_2)^T,\ \lambda \in L^2(S),$$
$$\mu \in [L^2(S)]^2,\ q - \Delta\lambda + \operatorname{div}\mu \in L^2(S)\}$$

we define the dual functional

$$I_0(\lambda, \mu) = -\frac{1}{2}\int\limits_S \{\lambda^2 + (T^{-1}\mu, \mu) + \varkappa^{-1}(q - \Delta\lambda + \operatorname{div}\mu)^2\}\, dx.$$

Theorem 7. *If $\lambda_0 = \Delta u_0$ and $\mu_0 = T\nabla u_0$, then*

$$\inf_{u \in \mathring{H}_2(S)} J^0(u) = J^0(u_0) = I_0(\lambda_0, \mu_0) = \sup_{(\lambda,\mu)\in\mathcal{U}} I_0(\lambda, \mu).$$

In problem (D), on $\mathcal{U}$ we define the dual functional

$$J_0(\lambda, \mu) = -\frac{1}{2}\int\limits_S \{\lambda^2 + (T^{-1}\mu, \mu)$$
$$+ \varkappa^{-1}(q - \Delta\lambda + \operatorname{div}\mu)^2\}\, dx + (\theta_\nu\{\lambda, \mu\}, f)_{0;\partial S},$$

$$\theta_\nu\{\lambda, \mu\} = (-\partial_\nu\lambda + \mu_\nu, \lambda), \mu_\nu = \mu_1\nu_1 + \mu_2\nu_2.$$

Theorem 8. *If $u_0 \in H_2(S)$ is the solution of* (D) *with*

$$q \in L^2(S), \quad \lambda_0 = \Delta u_0, \quad \mu_0 = T\nabla u_0,$$

then

$$\inf_{u \in \mathcal{U}_f} J^0(u) = J^0(u_0) = J_0(\lambda_0, \mu_0) = \sup_{\{\lambda,\mu\}\in\mathcal{U}} J_0(\lambda, \mu).$$

10.6 The Dual Functional in the Absence of an Elastic Foundation ($\varkappa = 0$)

We assume that T is uniformly positive definite in S. In problem (D_0), on the set

$$\mathcal{U} = \big\{\{\lambda, \mu\},\ \mu = (\mu_1, \mu_2)^T,\ \lambda \in L^2(S),\ \mu \in [L^2(S)]^2,\ \Delta\lambda - \operatorname{div}\mu = q\big\}$$

we define the dual functional

$$I_0(\lambda,\mu) = -\frac{1}{2}\int\limits_S \{\lambda^2 + (T^{-1}\mu,\mu)\}\,dx.$$

The exact solution u_0 minimizes on $\mathring{H}_2(S)$ the energy functional

$$J^0(u) = \frac{1}{2}\int\limits_S \{(\Delta u)^2 + (T\nabla u,\nabla u) - 2qu\}\,dx.$$

Theorem 9. *If $\lambda_0 = \Delta u_0$ and $\mu_0 = T\nabla u_0$, then*

$$\inf_{u\in\mathring{H}_2(S)} J^0(u) = J^0(u_0) = I_0(\lambda_0,\mu_0) = \sup_{(\lambda,\mu)\in\mathcal{U}} I_0(\lambda,\mu).$$

In problem (D) with $\gamma u = f \in H_{3/2}(S)\times H_{1/2}(S)$, on $\mathcal{U}$ subjected to the new restrictions

$$\Delta\lambda - \operatorname{div}\mu = Q,$$
$$Q = q - \Delta^2 F + \operatorname{div}(T\nabla F),$$

where F is an extension of f to S, we define the dual functional

$$J_0(\lambda,\mu) = -\frac{1}{2}\int\limits_S \{\lambda^2 + (T^{-1}\mu,\mu)\}\,dx + (\theta_\nu\{\lambda,\mu\}, f)_{0;\partial S}.$$

Theorem 10. *If $u_0 \in H_2(S)$ is the solution of* (D) *and*

$$q \in L^2(S), \quad \lambda_0 = \Delta u_0, \quad \mu_0 = T\nabla u_0,$$

then

$$\inf_{u\in\mathcal{U}_f} J^0(u) = J^0(u_0) = J_0(\lambda_0,\mu_0) = \sup_{\{\lambda,\mu\}\in\mathcal{U}} J_0(\lambda,\mu).$$

Suppose now that $T = 0$. Then the solution u_0 of problem (D_0) minimizes on $\mathring{H}_2(S)$ the energy functional

$$J^0(u) = \frac{1}{2}\int\limits_S \{(\Delta u)^2 - 2qu\}\,dx.$$

On the set

$$\mathcal{U} = \{v \in L^2(S) : \Delta v = q\},$$

we define the dual functional

$$I_0(v) = -\frac{1}{2}\int\limits_S v^2\,dx.$$

Theorem 11. *If $u_0 \in \mathring{H}_2(S)$ is the solution of* (D_0) *and $v_0 = \Delta u_0$, then*

$$\inf_{u \in \mathring{H}(S)} J^0(u) = J^0(u_0) = J_0(v_0) = \sup_{v \in \mathcal{U}} J_0(v).$$

Problem (D) is treated as above.

Remark 1. When $\varkappa = 0$, the set $\mathcal{U}$ is defined with differential restrictions. This inconvenience is removed by developing nonclassical dual methods, as in the case of the transverse shear deformation model.

10.7 Numerical Example

Working in kilograms and centimeters, we consider a square steel floor with

$$S = [-100, 100] \times [-100, 100], \quad h = 0.5, \quad D = 44048,$$

an elastic foundation characterized by

$$\varkappa = 10^{-6},$$

and a pre-existing stress given by

$$T = \frac{8}{10^4} \begin{pmatrix} 2x_1 + x_2 + 1000 & x_1 - 2x_2 \\ x_1 - 2x_2 & 3x_1 - x_2 + 1000 \end{pmatrix}.$$

Both the direct and dual problems are solved by means of the Galerkin method. In the former, we consider the subspaces spanned by

$$\left\{P_{i,j}(x_1, x_2)\right\}_{i+j=0}^{i+j=n},$$
$$P_{i,j}(x_1, x_2) = x_1^i x_2^j (10^4 - x_1^2)(10^4 - x_2^2).$$

In the latter, we work with the subspaces spanned by

$$\left\{(\Delta P_{i,j}, 0, 0),\ (0, \partial_1 P_{i,j}, 0),\ (0, 0, \partial_2 P_{i,j})\right\}_{i+j=0}^{i+j=n}.$$

In the computation, performed with Mathematica, the operative error estimate is

$$\|u_n - u_0\|_2 \le c[J^0(u_n) - I_0(v_n)]^{1/2} \to 0 \quad \text{as } n \to \infty.$$

A test problem is used to estimate the value

$$c = 100.$$

The results for a uniform load

$$q(x) \equiv -2 \times 10^{-8}$$

are shown in Table 10.1.

The approximate solution u_{30} is graphed in Figure 10.1.

Table 10.1. Results for a uniform load $q(x) \equiv -2 \times 10^{-8}$.

n	$J^0(u_n)$	$I_0(v_n)$	$c[J^0(u_n) - I_0(v_n)]^{1/2}$
2	-2.291844×10^{-6}	-2.681685×10^{-6}	6.24373×10^{-2}
8	-2.297886×10^{-6}	-2.420752×10^{-6}	3.50523×10^{-2}
14	-2.297889×10^{-6}	-2.375689×10^{-6}	2.78927×10^{-2}
20	-2.297890×10^{-6}	-2.327588×10^{-6}	1.72333×10^{-2}
26	-2.297890×10^{-6}	-2.307677×10^{-6}	9.89276×10^{-3}
30	-2.297890×10^{-6}	-2.302678×10^{-6}	6.91943×10^{-3}

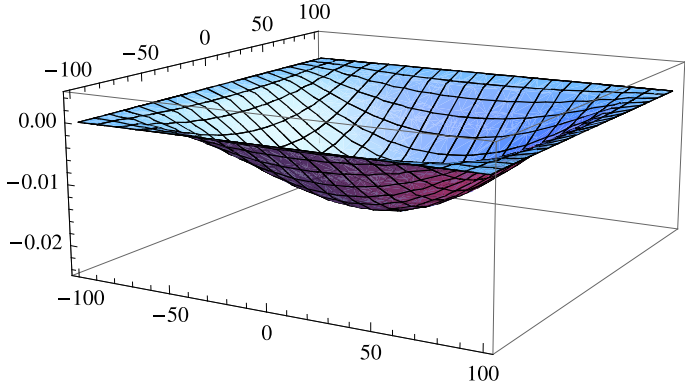


Fig. 10.1. Approximate solution u_{30}.

References

[TiWo87] Timoshenko, S., Woinowsky-Krieger, S.: *Theory of Plates and Shells,* 2nd ed., McGraw-Hill, New York (1987).

[ChCoKo00] Chudinovich, I., Constanda, C., Koshchii, A.: The classical approach to dual methods for plates. *Quart. J. Mech. Appl. Math.,* **53**, 497–510 (2000).

[ChEtAl06] Chudinovich, I., Constanda, C., Doty, D., Koshchii, A.: Non-classical dual methods in equilibrium problems for thin elastic plates. *Quart. J. Mech. Appl. Math.,* **59**, 125–137 (2006).

[ChCo00] Chudinovich, I., Constanda, C.: *Variational and Potential Methods in the Theory of Bending of Plates with Transverse Shear Deformation,* Chapman & Hall/CRC, Boca Raton, FL (2000).

[ChCo05] Chudinovich, I., Constanda, C.: *Variational and Potential Methods for a Class of Linear Hyperbolic Evolutionary Processes,* Springer, London (2005).

11

Modified Newton's Methods for Systems of Nonlinear Equations

A. Cordero and J.R. Torregrosa

Universidad Politécnica de Valencia, Spain; acordero@mat.upv.es, jrtorre@mat.upv.es

11.1 Introduction

The main goal of this chapter is to obtain new iterative formulas in order to solve systems of nonlinear equations. They are proved to be modifications on classical Newton's method which accelerate the convergence of the iterative process.

In previous works, the authors have obtained variants on Newton's method based on quadrature formulas whose truncation error was up to $O(h^5)$ (see [CoTo06] and [CoTo07]). Nevertheless, the approach used in this paper to solve a nonlinear system is different: by using Adomian polynomials, we obtain a family of multipoint iterative formulas, which include Newton and Traub (see [Tr82]) methods in the simplest cases.

The decomposition method using Adomian polynomials is used to solve different problems in applied mathematics in [Ad88]. Indeed, Babolian et al. (see [BaBiVa04]) apply this general method to a concrete nonlinear system. Nevertheless, with a different system, it is necessary to reconstruct the entire process.

We deduce in Section 11.2, by means of Adomian decomposition, a family of iterative formulas that can be applied to solve any nonlinear system without knowledge about Adomian polynomials. These iterative formulas involve classical methods, like those of Newton (order $p = 2$) and Traub (order $p = 3$), and also new methods whose convergence order is proved to be higher.

In Section 11.3, we study the convergence of the different methods by using the following result.

Theorem 1. *(see [Tr82]) Let $G(x)$ be a fixed point function with continuous partial derivatives of order p with respect to all components of x. The iterative method $x^{(k+1)} = G(x^{(k)})$ is of order p if*

$$G(\overline{x}) = \overline{x};$$

$$\frac{\partial^k g_i(\overline{x})}{\partial x_{j_1} \partial x_{j_2} \ldots \partial x_{j_k}} = 0, \quad \textit{for all} \;\; 1 \le k \le p-1, \quad 1 \le i, j_1, \ldots, j_k \le n;$$

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_11,

$$\frac{\partial^p g_i(\overline{x})}{\partial x_{j_1}\partial x_{j_2}\dots\partial x_{j_p}} \neq 0, \quad \textit{for at least one value of } \ i, j_1, \dots, j_p,$$

where g_i are the component functions of G.

Finally, numerical tests are made in Section 11.4 comparing the classical and new methods and confirming the theoretical results.

11.2 Description of the Methods

Let $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $n > 1$, be a sufficiently differentiable function whose coordinate functions are $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, 2, \dots, n$. Let $\overline{x}$ be a zero of the nonlinear system $F(x) = 0$ and $\alpha \in \mathbb{R}^n$ an estimation of $\overline{x}$. Then, this system is equivalent to

$$F(\alpha) + J_F(\alpha)(x - \alpha) + K(x) = 0,$$

where $J_F(\alpha)$ is the Jacobian matrix of the function F evaluated in the estimation α and $K : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ verifies

$$K(x) = F(x) - F(\alpha) - J_F(\alpha)(x - \alpha).$$

Then, $x = \alpha - J_F^{-1}(\alpha)F(\alpha) - J_F^{-1}(\alpha)K(x)$. Let us denote the linear component as $c \equiv \alpha - J_F^{-1}(\alpha)F(\alpha) \in \mathbb{R}^n$, and by $P(x)$ the nonlinear one, $P(x) = -J_F^{-1}(\alpha)K(x)$, $P : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ with coordinate functions $P_i : \mathbb{R}^n \longrightarrow \mathbb{R}$. So, $x = c + P(x)$.

Let us suppose that each one of the respective ith components of the approximation x of the solution $\overline{x}$ and also of $P(x)$ can be written as $x_i = \sum_{l=0}^{\infty} x_l^i$ and $P_i(x) = \sum_{l=0}^{\infty} A_l^i$, $i = 1, 2, \dots, n$, where $A_l^i : \mathbb{R}^n \longrightarrow \mathbb{R}$ are Adomian polynomials. Subsequently, a first estimation of $\overline{x}$ is $x^0 \equiv (x_0^1, x_0^2, \dots, x_0^n)^T$, where $x_0^i = c_i = \alpha_i - \sum_{j=1}^{n} H_{ij}(\alpha)f_j(\alpha)$, $H_{ij}(x)$ being the (i, j)-entry of the inverse of the Jacobian matrix. So,

$$x^0 = \alpha - J_F^{-1}(\alpha)F(\alpha) \tag{11.1}$$

and $x \simeq x^0$, which corresponds to the classical Newton's method (CN).

If a better approximation is needed, a new term in the series development of x is used, $x_i \simeq x_0^i + x_1^i$, where $x_1^i = A_0^i = P_i(x_0^i) = -\sum_{j=1}^{n} H_{ij}(\alpha)k_j(x_0^i)$ and $k_j(x)$ is the jth coordinate function of $K(x)$. Then,

$$x^1 = -J_F^{-1}(\alpha)K(x^0) \tag{11.2}$$

and

$$x \simeq x^0 + x^1 = \alpha - J_F^{-1}(\alpha)\left(F(\alpha) + F(x^0)\right).$$

This method, whose convergence order is 3, was described by Traub in [Tr82].

The term x_2^i in the series development of x_i is obtained as

$$\begin{aligned} x_2^i &= A_1^i = \frac{d}{d\lambda}\left(P_i\left(\sum_{l=0}^{\infty}\lambda^l x_l^1,\ldots,\sum_{l=0}^{\infty}\lambda^l x_l^n\right)\right)_{\lambda=0} \\ &= -\sum_{j=1}^{n} H_{ij}(\alpha)\frac{d}{d\lambda}\left(f_j\left(\sum_{l=0}^{\infty}\lambda^l x_l^1,\ldots,\sum_{l=0}^{\infty}\lambda^l x_l^n\right)\right)_{\lambda=0} \\ &\quad +\sum_{j=1}^{n} H_{ij}(\alpha)\frac{d}{d\lambda}\left(f_j(\alpha)+\sum_{m=1}^{n}\frac{\partial f_j(\alpha)}{\partial x_m}\left(\sum_{l=0}^{\infty}\lambda^l x_l^m-\alpha_m\right)\right)_{\lambda=0}. \end{aligned}$$

Let us denote $\left(\sum_{l=0}^{\infty}\lambda^l x_l^1,\ldots,\sum_{l=0}^{\infty}\lambda^l x_l^n\right)$ by $\mu(\lambda,x)$, whose mth component is denoted by $\mu_m(\lambda,x)$, and note that $\mu(0,x)=(x_0^1,\ldots,x_0^n)^T=x^0$. Moreover, taking into account that $\frac{\partial\mu_m(\lambda,x)}{\partial\lambda}=\sum_{l=1}^{\infty} l\lambda^{l-1}x_l^m$,

$$\begin{aligned} x_2^i &= -\sum_{j=1}^{n} H_{ij}(\alpha)\left(\sum_{m=1}^{n}\frac{\partial f_j(x^0)}{\partial x_m}x_1^m\right)+\sum_{j=1}^{n} H_{ij}(\alpha)\left(\sum_{m=1}^{n}\frac{\partial f_j(\alpha)}{\partial x_m}x_1^m\right) \\ &= x_1^i-\sum_{j=1}^{n}\sum_{m=1}^{n} H_{ij}(\alpha)\frac{\partial f_j(x^0)}{\partial x_m}x_1^m, \end{aligned}$$

by using

$$\sum_{j=1}^{n} H_{ij}(x)J_{jm}(x)=\delta_{im}, \tag{11.3}$$

where δ_{im} is the Kronecker symbol. Then, in vectorial notation:

$$x^2=x^1-J_F^{-1}(\alpha)J_F(x^0)x^1 \tag{11.4}$$

and, using (11.1), (11.2), and (11.4), a new estimation of the solution $\overline{x}$ is

$$x\simeq\alpha-J_F^{-1}(\alpha)F(\alpha)-\left[2J_F^{-1}(\alpha)-J_F^{-1}(\alpha)J_F(x^0)J_F^{-1}(\alpha)\right]F(x^0).$$

This expression corresponds to a new method which involves only a new function evaluation with respect to the previously described methods and whose convergence order will be proved to be 4. We call this new method NAd1, as it is a variant of Newton's method that use Adomian polynomials of sub-index 1. Nevertheless, an iterative expression of NAd1 can be used with no knowledge of Adomian polynomials, only in terms of previous estimations and Newton's approximation. So, $x^{(0)}$ being the initial estimation of the iterative process and $\overline{x}^{(k+1)}=x^{(k)}-J_F^{-1}(x^{(k)})F(x^{(k)})$ being the $(k+1)$th approximation of Newton's method, a new estimation $x^{(k+1)}$ can be obtained by means of the following expression:

$$x^{(k+1)} = \overline{x}^{(k+1)} - \left[2J_F^{-1}(x^{(k)}) - M^{(k)}\right] F(\overline{x}^{(k+1)}), \tag{11.5}$$

where $M^{(k)} = J_F^{-1}(x^{(k)})J_F(\overline{x}^{(k+1)})J_F^{-1}(x^{(k)})$. Indeed, other new methods can be obtained if more terms are added in the truncation of the theoretical series developments of each ith component of the estimation x. In particular, calculating the components of $x^3 = (x_3^1, \ldots, x_3^n)^T$ by means of the respective Adomian polynomial A_2^i, $i = 1, 2, \ldots, n$,

$$\begin{aligned}
x_3^i &= A_2^i = \frac{1}{2}\frac{d^2}{d\lambda^2}\left(P_i\left(\mu(\lambda, x)\right)\right)_{\lambda=0} \\
&= -\frac{1}{2}\sum_{j=1}^{n} H_{ij}(\alpha)\sum_{m=1}^{n}\left\{\sum_{a=1}^{n}\frac{\partial^2 f_j(\mu(\lambda,x))}{\partial\mu_m(\lambda,x)\partial\mu_a(\lambda,x)}\frac{\partial\mu_m(\lambda,x)}{\partial\lambda}\frac{\partial\mu_a(\lambda,x)}{\partial\lambda}\right\}_{\lambda=0} \\
&\quad -\frac{1}{2}\sum_{j=1}^{n} H_{ij}(\alpha)\sum_{m=1}^{n}\left\{\frac{\partial f_j(\mu)}{\partial\mu_m(\lambda,x)}\frac{\partial\mu_m(\lambda,x)}{\partial\lambda}\right\}_{\lambda=0} \\
&\quad +\sum_{m=1}^{n} H_{ij}(\alpha)\left\{\frac{\partial f_j(\alpha)}{\partial x_m}\frac{\partial^2\mu_m(\lambda,x)}{\partial\lambda^2}\right\}_{\lambda=0}.
\end{aligned}$$

As $\frac{\partial^2\mu_m(\lambda,x)}{\partial\lambda^2} = \sum_{l=2}^{\infty}(l-1)l\lambda^{l-2}x_l^m$ and using (11.3), each component x_3^i is defined by

$$\begin{aligned}
x_3^i &= -\frac{1}{2}\sum_{j=1}^{n}\sum_{m=1}^{n}\sum_{a=1}^{n} H_{ij}(\alpha)\frac{\partial^2 f_j(x^0)}{\partial x_m\partial x_a}x_1^m x_1^a \\
&\quad -\sum_{j=1}^{n} H_{ij}(\alpha)\sum_{m=1}^{n}\frac{\partial f_j(x^0)}{\partial x_m}x_2^m + \sum_{m=1}^{n}\delta_{im}x_2^m.
\end{aligned}$$

Then, in vector notation, x^3 can be expressed as

$$x^3 = x^2 - \frac{1}{2}J_F^{-1}(\alpha)B - J_F^{-1}(\alpha)J_F(x^0)x^2,$$

where B is a vector whose jth component is $B_j = \sum_{m=1}^{n}\sum_{a=1}^{n}\frac{\partial^2 f_j(x^0)}{\partial x_m\partial x_a}x_1^m x_1^a$.

A new approximation of the solution is obtained, $x \simeq x^0 + x^1 + x^2 + x^3$. This is a new method which involves the functional evaluation of vector B including second-order partial derivatives of $f_j, j = 1, \ldots, n$, whose convergence order is 5. We call this new method NAd2, as it uses Adomian polynomials of sub-index 2. So, $x^{(0)}$ being the initial estimation and $\overline{x}^{(k+1)} = x^{(k)} - J_F^{-1}(x^{(k)})F(x^{(k)})$ being the $(k+1)$th approximation of Newton's method, a new estimation $x^{(k+1)}$ can be obtained by means of

$$\begin{aligned}
x^{(k+1)} &= \overline{x}^{(k+1)} - 3J_F^{-1}(x^{(k)})F(\overline{x}^{(k+1)}) + 3M^{(k)}F(\overline{x}^{(k+1)}) \\
&\quad - M^{(k)}J_F(\overline{x}^{(k+1)})J_F^{-1}(x^{(k)})F(\overline{x}^{(k+1)}) - \frac{1}{2}J_F^{-1}(x^{(k)})B^{(k)},
\end{aligned} \tag{11.6}$$

where the jth component of $B^{(k)}$ is $B_j^{(k)} = \sum_{m=1}^{n}\sum_{a=1}^{n} \frac{\partial^2 f_j(\overline{x}^{(k+1)})}{\partial x_m \partial x_a} x_1^m x_1^a$.

11.3 Convergence Analysis

Let $\overline{x}$ be a zero of the nonlinear system $F(x) = 0$. It can be easily proved that

$$\sum_{i=1}^{n} \frac{\partial H_{ji}(x)}{\partial x_l} \frac{\partial f_i(x)}{\partial x_r} = -\sum_{i=1}^{n} H_{ji}(x) \frac{\partial^2 f_i(x)}{\partial x_l \partial x_r}, \tag{11.7}$$

$$\begin{aligned}\sum_{i=1}^{n} \frac{\partial^2 H_{ji}(x)}{\partial x_s \partial x_l} \frac{\partial f_i(x)}{\partial x_r} &= -\sum_{i=1}^{n} \frac{\partial H_{ji}(x)}{\partial x_l} \frac{\partial^2 f_i(x)}{\partial x_s \partial x_r} - \sum_{i=1}^{n} \frac{\partial H_{ji}(x)}{\partial x_s} \frac{\partial^2 f_i(x)}{\partial x_r \partial x_l} \\ &\quad - \sum_{i=1}^{n} H_{ji}(x) \frac{\partial^3 f_i(x)}{\partial x_s \partial x_r \partial x_l},\end{aligned} \tag{11.8}$$

and

$$\begin{aligned}\sum_{i=1}^{n} \frac{\partial^3 H_{ji}(x)}{\partial x_u \partial x_s \partial x_l} \frac{\partial f_i(x)}{\partial x_r} &= -\sum_{i=1}^{n} \frac{\partial^2 H_{ji}(x)}{\partial x_s \partial x_l} \frac{\partial^2 f_i(x)}{\partial x_u \partial x_r} - \sum_{i=1}^{n} \frac{\partial^2 H_{ji}(x)}{\partial x_u \partial x_l} \frac{\partial^2 f_i(x)}{\partial x_s \partial x_r} \\ &\quad - \sum_{i=1}^{n} \frac{\partial^2 H_{ji}(x)}{\partial x_u \partial x_s} \frac{\partial^2 f_i(x)}{\partial x_l \partial x_r} - \sum_{i=1}^{n} \frac{\partial H_{ji}(x)}{\partial x_l} \frac{\partial^3 f_i(x)}{\partial x_u \partial x_s \partial x_r} \\ &\quad - \sum_{i=1}^{n} \frac{\partial H_{ji}(x)}{\partial x_s} \frac{\partial^3 f_i(x)}{\partial x_u \partial x_r \partial x_l} - \sum_{i=1}^{n} \frac{\partial H_{ji}(x)}{\partial x_u} \frac{\partial^3 f_i(x)}{\partial x_s \partial x_r \partial x_l} \\ &\quad - \sum_{i=1}^{n} H_{ji}(x) \frac{\partial^4 f_i(x)}{\partial x_u \partial x_s \partial x_r \partial x_l}.\end{aligned} \tag{11.9}$$

The following result, partially proved in [Tr82], will be useful in the proof of the main theorem.

Lemma 1. *Let $\lambda(x)$ be the iteration function of the classical Newton's method, whose coordinates are $\lambda_j(x) = x_j - \sum_{i=1}^{n} H_{ji}(x) f_i(x)$, for $j = 1, \ldots, n$. Then,*

$$\frac{\partial \lambda_j(\overline{x})}{\partial x_l} = 0, \tag{11.10}$$

$$\frac{\partial^2 \lambda_j(\overline{x})}{\partial x_r \partial x_l} = \sum_{i=1}^{n} H_{ji}(\overline{x}) \frac{\partial^2 f_i(\overline{x})}{\partial x_r \partial x_l}, \tag{11.11}$$

$$\begin{aligned}\frac{\partial^3 \lambda_j(\overline{x})}{\partial x_s \partial x_r \partial x_l} &= \sum_{i=1}^{n} \left[\frac{\partial H_{ji}(\overline{x})}{\partial x_r} \frac{\partial^2 f_i(\overline{x})}{\partial x_s \partial x_l} + \frac{\partial H_{ji}(\overline{x})}{\partial x_s} \frac{\partial^2 f_i(\overline{x})}{\partial x_r \partial x_l} + \frac{\partial H_{ji}(\overline{x})}{\partial x_l} \frac{\partial^2 f_i(\overline{x})}{\partial x_s \partial x_r} \right] \\ &\quad + 2 \sum_{i=1}^{n} H_{ji}(\overline{x}) \frac{\partial^3 f_i(\overline{x})}{\partial x_s \partial x_r \partial x_l},\end{aligned} \tag{11.12}$$

and

$$\begin{aligned}\frac{\partial^4\lambda_j(\overline{x})}{\partial x_u\partial x_s\partial x_r\partial x_l} &= \sum_{i=1}^{n}\frac{\partial^2 H_{ji}(\overline{x})}{\partial x_s\partial x_l}\frac{\partial^2 f_i(\overline{x})}{\partial x_r\partial x_u}+\sum_{i=1}^{n}\frac{\partial^2 H_{ji}(\overline{x})}{\partial x_r\partial x_l}\frac{\partial^2 f_i(\overline{x})}{\partial x_s\partial x_u}\\ &+\sum_{i=1}^{n}\frac{\partial^2 H_{ji}(\overline{x})}{\partial x_r\partial x_s}\frac{\partial^2 f_i(\overline{x})}{\partial x_l\partial x_u}+\sum_{i=1}^{n}\frac{\partial^2 H_{ji}(\overline{x})}{\partial x_u\partial x_r}\frac{\partial^2 f_i(\overline{x})}{\partial x_s\partial x_l}\\ &+\sum_{i=1}^{n}\frac{\partial^2 H_{ji}(\overline{x})}{\partial x_u\partial x_l}\frac{\partial^2 f_i(\overline{x})}{\partial x_s\partial x_r}+\sum_{i=1}^{n}\frac{\partial^2 H_{ji}(\overline{x})}{\partial x_u\partial x_s}\frac{\partial^2 f_i(\overline{x})}{\partial x_l\partial x_r}\\ &+2\sum_{i=1}^{n}\frac{\partial H_{ji}(\overline{x})}{\partial x_r}\frac{\partial^3 f_i(\overline{x})}{\partial x_u\partial x_s\partial x_l}\\ &+2\sum_{i=1}^{n}\frac{\partial H_{ji}(\overline{x})}{\partial x_s}\frac{\partial^3 f_i(\overline{x})}{\partial x_u\partial x_r\partial x_l}\\ &+2\sum_{i=1}^{n}\frac{\partial H_{ji}(\overline{x})}{\partial x_l}\frac{\partial^3 f_i(\overline{x})}{\partial x_u\partial x_s\partial x_r}\\ &+2\sum_{i=1}^{n}\frac{\partial H_{ji}(\overline{x})}{\partial x_u}\frac{\partial^3 f_i(\overline{x})}{\partial x_s\partial x_r\partial x_l}\\ &+3\sum_{i=1}^{n}H_{ji}(\overline{x})\frac{\partial^4 f_i(\overline{x})}{\partial x_u\partial x_s\partial x_r\partial x_l},\end{aligned}\tag{11.13}$$

for $j,l,r,s,u\in\{1,2,\ldots,n\}$.

Let us note that, by applying Theorem 1 and using expressions (11.10) and (11.11) in Lemma 1, it can be concluded that the convergence order of Newton's method is $p=2$.

Lemma 2. *Let* $\lambda(x)$ *be the iteration function of the classical Newton's method. Moreover, let us denote by* $N_{ij}(x)$ *the* (i,j)*-entry of the matrix* $N(x)=J_F(\lambda(x))J_F^{-1}(x)$, $N_{ij}(x)=\sum_{q=1}^{n}J_{iq}(\lambda(x))H_{qj}(x)$. *Then,*

$$N_{ij}(\overline{x})=\delta_{ij},\tag{11.14}$$

$$\frac{\partial N_{ij}(\overline{x})}{\partial x_l}=-\sum_{q=1}^{n}H_{qj}(\overline{x})\frac{\partial^2 f_i(\overline{x})}{\partial x_q\partial x_l},\tag{11.15}$$

and

$$\begin{aligned}\frac{\partial^2 N_{ij}(\overline{x})}{\partial x_l\partial x_r} &= \sum_{q=1}^{n}\sum_{p=1}^{n}H_{qj}(\overline{x})\frac{\partial^2 f_i(\overline{x})}{\partial x_q\partial x_p}\frac{\partial^2 f_i(\overline{x})}{\partial x_l\partial x_r}\\ &- \sum_{q=1}^{n}\frac{\partial H_{qj}(\overline{x})}{\partial x_r}\frac{\partial^2 f_i(\overline{x})}{\partial x_q\partial x_l}-\sum_{q=1}^{n}H_{qj}(\overline{x})\frac{\partial^3 f_i(\overline{x})}{\partial x_q\partial x_l\partial x_r},\end{aligned}\tag{11.16}$$

for $i,j,l,r\in\{1,2,\ldots,n\}$.

Theorem 2. *Let $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be sufficiently differentiable at each point of an open neighborhood D of $\overline{x} \in \mathbb{R}^n$ that is a solution of the system $F(x) = 0$. Let us suppose that $J_F(x)$ is continuous and nonsingular in $\overline{x}$. Then, the sequence $\{x^{(k)}\}_{k\geq 0}(x^{(0)} \in D)$ obtained by using the iterative expressions of methods NAd1 (11.5) and NAd2 (11.6) converges to $\overline{x}$ with convergence order 4 and 5, respectively.*

Proof. Let us consider a solution $\overline{x} \in \mathbb{R}^n$ of the nonlinear system $F(x) = 0$ as a fixed point of the iteration function $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ associated witht the method described in (11.5). Let us denote by $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, 2, \ldots, n$, the coordinate functions of G.

We denote by $M_{ij}(x)$ the (i, j)-entry of $M(x) = J_F^{-1}(x)J_F(\lambda(x))J_F^{-1}(x)$. Thus, the ith component of the iteration function corresponding to method NAd1 is

$$g_i(x) = \lambda_i(x) - 2\sum_{j=1}^{n} H_{ij}(x)f_j(\lambda(x)) + \sum_{j=1}^{n} M_{ij}(x)f_j(\lambda(x)). \tag{11.17}$$

Since $H_{ij}(x)$ and $J_{ij}(x)$ are the elements of inverse matrices, (11.17) can be rewritten as

$$\sum_{j=1}^{n} J_{ij}(x)\left(g_j(x) - \lambda_j(x)\right) + 2f_i(\lambda(x)) - \sum_{j=1}^{n} N_{ij}f_j(\lambda(x)) = 0. \tag{11.18}$$

Now, by direct differentiation of (11.18), with i and l arbitrary and fixed,

$$\begin{aligned}
&\sum_{j=1}^{n} \frac{\partial J_{ij}(x)}{\partial x_l}\left(g_j(x) - \lambda_j(x)\right) + \sum_{j=1}^{n} J_{ij}(x)\left(\frac{\partial g_j(x)}{\partial x_l} - \frac{\partial \lambda_j(x)}{\partial x_l}\right)\\
&+ 2\sum_{q=1}^{n} \frac{\partial f_i(\lambda(x))}{\partial \lambda_q(x)}\frac{\partial \lambda_q(x)}{\partial x_l} - \sum_{j=1}^{n} \frac{\partial N_{ij}(x)}{\partial x_l} f_j(\lambda(x))\\
&- \sum_{j=1}^{n} N_{ij}(x)\left(\sum_{q=1}^{n} \frac{\partial f_j(\lambda(x))}{\partial \lambda_q(x)}\frac{\partial \lambda_q(x)}{\partial x_l}\right) = 0.
\end{aligned} \tag{11.19}$$

When $x = \overline{x}$, by applying Lemma 1, expression (11.10), and taking into account that $g(\overline{x}) = \overline{x}$, $\lambda(\overline{x}) = \overline{x}$, and $f_i(\overline{x}) = 0$, we have

$$\sum_{j=1}^{n} J_{ij}(\overline{x})\frac{\partial g_j(\overline{x})}{\partial x_l} = 0.$$

Then, as $J_F(\overline{x})$ is supposed to be nonsingular, and i and l are arbitrary,

$$\frac{\partial g_j(\overline{x})}{\partial x_l} = 0.$$

Now, by direct differentiation of (11.19), with r arbitrary and fixed,

$$\begin{aligned}
&\sum_{j=1}^{n} \frac{\partial^2 J_{ij}(x)}{\partial x_r \partial x_l} (g_j(x) - \lambda_j(x)) + \sum_{j=1}^{n} \frac{\partial J_{ij}(x)}{\partial x_l} \left(\frac{\partial g_j(x)}{\partial x_r} - \frac{\partial \lambda_j(x)}{\partial x_r} \right) \\
&+ \sum_{j=1}^{n} \frac{\partial J_{ij}(x)}{\partial x_r} \left(\frac{\partial g_j(x)}{\partial x_l} - \frac{\partial \lambda_j(x)}{\partial x_l} \right) + \sum_{j=1}^{n} J_{ij}(x) \left(\frac{\partial^2 g_j(x)}{\partial x_r \partial x_l} - \frac{\partial^2 \lambda_j(x)}{\partial x_r \partial x_l} \right) \\
&+ 2 \sum_{q=1}^{n} \sum_{p=1}^{n} \frac{\partial^2 f_i(x)}{\partial \lambda_p(x) \partial \lambda_q(x)} \frac{\partial \lambda_p(x)}{\partial x_r} \frac{\partial \lambda_q(x)}{\partial x_l} \\
&+ 2 \sum_{q=1}^{n} \frac{\partial f_i(x)}{\partial \lambda_q(x)} \frac{\partial^2 \lambda_q(x)}{\partial x_r \partial x_l} - \sum_{j=1}^{n} \frac{\partial^2 N_{ij}(x)}{\partial x_r \partial x_l} f_j(\lambda(x)) \\
&- \sum_{j=1}^{n} \frac{\partial N_{ij}(x)}{\partial x_l} \left(\sum_{q=1}^{n} \frac{\partial f_j(x)}{\partial \lambda_q(x)} \frac{\partial \lambda_q(x)}{\partial x_r} \right) - \sum_{j=1}^{n} \frac{\partial N_{ij}(x)}{\partial x_r} \left(\sum_{q=1}^{n} \frac{\partial f_j(x)}{\partial \lambda_q(x)} \frac{\partial \lambda_q(x)}{\partial x_l} \right) \\
&- \sum_{j=1}^{n} N_{ij}(x) \left(\sum_{q=1}^{n} \sum_{p=1}^{n} \frac{\partial^2 f_j(x)}{\partial \lambda_p(x) \partial \lambda_q(x)} \frac{\partial \lambda_p(x)}{\partial x_r} \frac{\partial \lambda_q(x)}{\partial x_l} \right) \\
&- \sum_{j=1}^{n} N_{ij}(x) \left(\sum_{q=1}^{n} \frac{\partial f_j(x)}{\partial \lambda_q(x)} \frac{\partial^2 \lambda_q(x)}{\partial x_r \partial x_l} \right) = 0. \qquad (11.20)
\end{aligned}$$

Let us substitute $x = \overline{x}$ and apply (11.10), (11.11) from Lemma 1, and (11.14), (11.15) from Lemma 2. Then,

$$\begin{aligned}
&\sum_{j=1}^{n} J_{ij}(\overline{x}) \frac{\partial^2 g_j(\overline{x})}{\partial x_r \partial x_l} - \sum_{j=1}^{n} \sum_{i=1}^{n} J_{ij}(\overline{x}) H_{ji}(\overline{x}) \frac{\partial^2 f_i(\overline{x})}{\partial x_r \partial x_l} \\
&- \sum_{j=1}^{n} \delta_{ij} \sum_{i=1}^{n} \sum_{q=1}^{n} J_{iq}(\overline{x}) H_{qi}(\overline{x}) \frac{\partial^2 f_i(\overline{x})}{\partial x_r \partial x_l} + 2 \sum_{q=1}^{n} J_{iq}(\overline{x}) H_{qi}(\overline{x}) \frac{\partial^2 f_i(\overline{x})}{\partial x_r \partial x_l} = 0.
\end{aligned}$$

Therefore, as $J_F(\overline{x})$ is nonsingular, and i, l, and r are arbitrary,

$$\frac{\partial^2 g_j(\overline{x})}{\partial x_r \partial x_l} = 0.$$

We now analyze the fourth order of convergence. To do this, it is necessary to differentiate (11.20) with respect to x_s, with s arbitrary and fixed, and evaluate the resulting expression in $x = \overline{x}$. Then, by using (11.3), (11.7), and (11.10), (11.11), and (11.12) from Lemma 1, (11.14) and (11.15) from Lemma 2, and simplifying, it is proved that

$$\frac{\partial^3 g_j(\overline{x})}{\partial x_s \partial x_r \partial x_l} = 0.$$

Again, with u arbitrary and fixed and using results from (11.7)–(11.9), Lemma 1 (expressions (11.10) through (11.13)) and Lemma 2 (expressions (11.14)–(11.16)), it can be proved that

$$J_{ij}(\overline{x})\frac{\partial^4 g_j(\overline{x})}{\partial x_u \partial x_s \partial x_r \partial x_l} + P(\overline{x}) = 0,$$

where $P(\overline{x})$ is a linear combination of partial derivatives of f_i of second order, evaluated in $\overline{x}$.

So, by Theorem 1 we conclude that the method NAd1 of iterative expression (11.5) converges to $\overline{x}$ with convergence order 4. The fifth-order convergence of method NAd2 can be proved in an analogous way.

11.4 Numerical Examples

In this section we give numerical examples and compare the effectiveness of the obtained iterative methods. In particular, the new methods NAd1 and NAd2 are analyzed, and also Traub's method (TM) and the classical Newton's method (CN), in order to estimate the zeros of several nonlinear functions.

(a) $F(x_1, x_2) = (\exp(x_1^2) - \exp(\sqrt{2}x_1), x_1 - x_2)$, $\overline{x}_1 = (\sqrt{2}, \sqrt{2})^T$, $\overline{x}_2 = (0, 0)^T$.
(b) $F(x_1, x_2) = (x_1 + \exp(x_2) - \cos(x_2), 3x_1 - x_2 - \sin(x_2))$, $\overline{x} = (0, 0)^T$.
(c) $F(x_1, x_2) = (x_1^2 - 2x_1 - x_2 + 0.5, x_1^2 + 4x_2^2 - 4)$, $\overline{x} = (1.9007, 0.3112)^T$.
(d) $F(x_1, x_2) = (x_1^2 + x_2^2 - 1, x_1^2 - x_2^2 + 0.5)$, $\overline{x}_1 = (\frac{1}{2}, \frac{\sqrt{3}}{2})^T$, $\overline{x}_2 = (-\frac{1}{2}, -\frac{\sqrt{3}}{2})^T$.
(e) $F(x_1, x_2) = (\sin(x_1) + x_2\cos(x_1), x_1 - x_2)$, $\overline{x} = (0, 0)^T$.

Table 11.1. Numerical results for nonlinear systems.

$F(x)$	$x^{(0)}$	Iterations				p				Sol.
		CN	**Tr**	**NAd1**	**NAd2**	**CN**	**Tr**	**NAd1**	**NAd2**	
(a)	$(2.3, 2.3)^T$	10	8	7	6	2.0	3.0	3.9	3.8	$\overline{x}_1$
	$(1.8, 1.8)^T$	7	5	5	4	2.0	3.0	3.6	4.2	$\overline{x}_1$
	$(0.8, 0.8)^T$	5	4	3	3	3.0	4.3	4.6	6.7	$\overline{x}_2$
(b)	$(1.5, 2)^T$	7	6	5	4	2.0	3.0	3.6	4.6	$\overline{x}$
	$(0.3, 0.5)^T$	5	4	4	3	2.0	3.0	3.7	4.6	$\overline{x}$
(c)	$(3, 2)^T$	7	5	5	4	2.0	2.6	2.5	3.1	$\overline{x}$
	$(1.6, 0)^T$	5	4	4	4	2.1	3.8	5.0	5.3	$\overline{x}$
(d)	$(0.7, 1.2)^T$	5	4	3	3	2.0	2.5	3.7	4.7	$\overline{x}_1$
	$(-1, -2)^T$	6	4	4	4	2.0	2.9	3.0	3.7	$\overline{x}_2$
(e)	$(1.2, -1.5)^T$	6	4	4	3	2.9	3.7	5.5	7.5	$\overline{x}$
	$(-0.6, 0.6)^T$	5	3	3	3	3.0	4.3	6.4	6.6	$\overline{x}$

The stopping criterion used is $\left\|x^{(k+1)} - x^{(k)}\right\| + \left\|F\left(x^{(k)}\right)\right\| < 10^{-12}$. For every method, Table 11.1 shows the number of iterations needed for convergence to the solution and the order of convergence estimated by means of the computational order of convergence p (see [WeFe00]).

References

[Ad88] Adomian, G.: A review of the decomposition method in applied mathematics. *J. Math. Anal. Appl.,* **135**, 501–544 (1988).

[BaBiVa04] Babolian, E., Biazar, J., Vahidi, A.R.: Solution of a system of nonlinear equations by Adomian decomposition method. *Appl. Math. Comput.,* **150**, 847–854 (2004).

[CoTo06] Cordero, A., Torregrosa, J.R.: Variants of Newton's method for functions of several variables. *Appl. Math. Comput.,* **183**, 199–208 (2006).

[CoTo07] Cordero, A., Torregrosa, J.R.: Variants of Newton's method using fifth-order quadrature formulas. *Appl. Math. Comput.,* **190**, 686–698 (2007).

[Tr82] Traub, J.F.: *Iterative Methods for the Solution of Equations,* Chelsea, New York (1982).

[WeFe00] Weerakoon, S., Fernando, T.G.I.: A variant of Newton's method with accelerated third-order convergence. *Appl. Math. Lett.,* **13** (8), 87–93 (2000).

12

Classification of Some Penalty Methods

A. Correia,[1] J. Matias,[2] P. Mestre,[2] and C. Serôdio[2]

[1] Instituto Politécnico do Porto, Felgueiras, Portugal; aldinacorreia@eu.ipp.pt
[2] Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal; j_matias@utad.pt, pmestre@utad.pt, cserodio@utad.pt

12.1 Introduction

Optimization problems arise in science, engineering, economy, etc. and we need to find the best solutions for each reality. The methods used to solve these problems depend on several factors, including the amount and type of accessible information, the available algorithms for solving them, and, obviously, the intrinsic characteristics of the problem.

There are many kinds of optimization problems and, consequently, many kinds of methods to solve them.

When the involved functions are nonlinear and their derivatives are not known or are very difficult to calculate, these methods are more rare. These kinds of functions are frequently called black box functions.

To solve such problems without constraints (unconstrained optimization), we can use direct search methods. These methods do not require any derivatives or approximations of them. But when the problem has constraints (nonlinear programming problems) and, additionally, the constraint functions are black box functions, it is much more difficult to find the most appropriate method. Penalty methods can then be used. They transform the original problem into a sequence of other problems, derived from the initial, all without constraints. Then this sequence of problems (without constraints) can be solved using the methods available for unconstrained optimization.

In this chapter, we present a classification of some of the existing penalty methods and describe some of their assumptions and limitations. These methods allow the solving of optimization problems with continuous, discrete, and mixing constraints, without requiring continuity, differentiability, or convexity.

Thus, penalty methods can be used as the first step in the resolution of constrained problems, by means of methods that typically are used by unconstrained problems.

We also discuss a new class of penalty methods for nonlinear optimization, which adjust the penalty parameter dynamically.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_12,

12.2 Formulation of the Problem

In the last years considerable investigation has been devoted to penalty methods (in 2006, Byrd [ByNoWa06], in 2005 Chen [ChGo05], in 2003 Gould [GoOrTo03], in 2006 Leyffer [LeLoNo06], in 2002 Klatte [KlKu02], in 1995 Mongeau [MoSa95], and in 2005 Zaslavski [Za05]), because of their capacity to solve degenerate problems with nonlinear constraints.

Penalty methods belong to a general approach that can solve continuous, discrete, and mixed constrained optimization problems, with no continuity, differentiability, and convexity requirements.

They were used to solve mathematical programs with complementarity constraints (MPCCs) by Benson [BeSeSh03] and by Leyffer [LeLoNo06] and were used by Byrd [ByNoWa06] and Chen [ChGo05] in constrained nonlinear programming to ensure sub-problems admissibility and increase the robustness of each iteration. Thus, penalty methods are the primary methods for solving constrained problems.

Consider the following general nonlinear programming problem (NLP), denoted by P:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x) \text{ subject to } & e_i(x) \geq 0, \quad i = 1,2,\ldots,s, \\ & d_j(x) = 0, \quad j = 1,2,\ldots,t, \\ & a_k \leq x_k, \quad k = 1,2,\ldots,n, \\ & x_l \leq b_l, \quad l = 1,2,\ldots,n, \end{aligned} \tag{12.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function, $e_i : \mathbb{R}^n \to \mathbb{R}$, with $i = 1,2,...,s$, are the s inequality constraints, $d_j : \mathbb{R}^n \to \mathbb{R}$, with $j = 1,2,...,t$, are the t equality constraints, and the two last conditions are the constraints of simple limits.

12.3 Penalty Methods

In a penalty method, the feasible region of P, R, defined by

$$\begin{aligned} & e_i(x) \geq 0, \quad i = 1,2,\ldots,s, \\ & d_j(x) = 0, \quad j = 1,2,\ldots,t, \\ & a_k \leq x_k, \quad k = 1,2,\ldots,n, \\ & x_l \leq b_l, \quad l = 1,2,\ldots,n, \end{aligned}$$

is expanded from R to $\mathbb{R}^n$, but a larger cost or penalty is added to the objective function for points that lie outside of the original feasible region, R.

Penalty methods construct a new objective function, Φ, that contains information about the initial objective function, f, and the problem constraints.

A sequence of unconstrained problems is constructed, dependent on the positive parameter r, with solutions $x^*(r)$ that converge to the solution of the initial problem x^*.

The new objective function Φ is

$$\Phi(x, r) = f(x) + \Theta(x, r),$$

where $\Theta : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is a function that depends on the positive parameter r, called the *penalty parameter* and $\Theta(x, r) = rp(x)$, where p is the *penalty function*.

Definition 1 ([FrRo04]). *The function $p : \mathbb{R}^n \to \mathbb{R}$ is the penalty function for P if*

- $p(x) = 0$ *if* $c_i(x) \leq 0$*;*
- $p(x) > 0$ *if* $c_i(x) > 0$*.*

Then we must solve a sequence of unconstrained problems P_m that replace problem P with the new objective function

$$P_m : \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) + r_m p(x),$$

where r_m is a sequence of constants such that $r_m \to +\infty$.

12.4 Classification of Some Penalty Methods

The goal of a penalty method is to find suitable penalty parameters in such a way that $x^*(r)$ minimizes P_m, corresponds to either a constrained global minimum (CGM) that is feasible and has the best objective value in the entire search space, or a constrained local minimum (CLM) that is feasible and has the best objective value in a pre-defined neighborhood.

Therefore, penalty methods can be classified as follows:

- *Global optimal penalty methods* (GOPM) if they look for CGM solutions of P;
- *Local optimal penalty methods* (LOPM) if they look for CLM solutions of P.

Penalty methods can also be classified in a different way [Be99]:

- *Inexact penalty methods,* in which the minimization of a penalty function, Φ, does not lead to exact CGM and CLM points, but instead successive minimizations of an inexact penalty function with increasing penalty values lead to points infinitely close to a CGM or CLM solution (converge to a CGM or CLM solution);
- *Exact penalty methods,* if they can find an exact CGM or CLM under finite penalty values.

Table 12.1. Classification of some penalty methods.

Global Optimization	Exact Penalty	Static Penalty
		Dynamic Penalty
	Inexact Penalty	Refuse Penalty
		Discrete Penalty
Local Optimization	Exact Penalty	Lagrangian
		Penalty l_1

Table 12.1 summarizes the classification of existing penalty methods.

In inexact penalty methods, a sequence of sub-problems with a divergent series of penalty parameters must be solved. For exact penalty methods, one choice of penalty parameters may be adequate for the entire minimization procedure. Consequently, exact penalty methods are less parameter dependent, which is their most appealing feature.

Definition 2 ([Za05]). *A penalty function is said to have the exact penalty property if there exists a penalty coefficient for which a solution of an unconstrained penalized problem is a solution of the corresponding constrained problem.*

12.4.1 Global Optimal Penalty Methods

Global optimal penalty methods (GOPM) can be exact or inexact methods. In this class are the static penalty methods and the dynamic penalty methods.

Static Penalty Methods

Static penalty methods were proposed by Homaifar [HoLaQi95]. In these methods, a set of violation levels is considered for each type of constraint and each violation level of constraints imposes a different level of penalty.

The disadvantage of these methods is that they require the setting of many parameters. The number of parameters grows faster when the number of constraints and violation levels increase. So they are computationally expensive because they involve finding a global minimum of a nonlinear penalty function.

We rewrite the constraints $a_i(x) \geq 0$ as $-a_i(x) \leq 0$, $i = 1, 2, ..., s$, and $a_k - x_k \leq 0, k = 1, 2, ..., n$ and $x_l - b_l \leq 0, l = 1, 2, ..., n$, as $g_i(x) \leq 0, i = 1, 2, ..., s + 2n$. Then problem (12.1) can be written as

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x) \text{ subject to } d_j(x) = 0, &\quad j = 1, 2, \ldots, t, \\ g_i(x) \leq 0, &\quad i = 1, 2, \ldots, s + 2n. \end{aligned} \tag{12.2}$$

With the penalty vectors α and β, an example of a static penalty problem for (12.2), $\rho \geq 1$, is

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \ L_s(x, \alpha, \beta), \tag{12.3}$$

where

$$L_s(x, \alpha, \beta) = f(x) + \sum_{j=1}^{t} \alpha_j \left| d_j(x) \right|^{\rho} + \sum_{i=1}^{s+2n} \beta_i \max(0, g_i(x))^{\rho}. \tag{12.4}$$

A static penalty method can be exact or inexact. For example, in (12.3), if $\rho = 1$ for (12.4), the method is an exact static penalty method; if $\rho > 1$, it is an inexact static penalty method [Be99].

That is, when $\rho = 1$, there exist penalty values α and β that ensure the minimum of the penalty function is exactly the CGM of P. However, when $\rho > 1$, the method is inexact because it converges to CGM as an infinite approximation of the penalty values.

The static penalty method of Homaifar [HoLaQi95] solves a similar problem to (12.3), but requires the choice of a very large number of parameters and it is also an inexact penalty method. Thus, the common limitation of all static penalty methods is that it is usually very difficult to choose statically the appropriate values of penalties. Moreover, these methods were developed to find CGM and do not allow finding a CLM for P. An alternative for finding penalty parameters is offered by dynamic penalty methods.

Dynamic Penalty Methods

Dynamic penalty methods were proposed initially by Wang [WaLi06]. These methods increase the penalty parameters gradually, instead of finding the penalty values by trial and error.

Like the static penalty methods, a dynamic penalty method can be an exact or inexact method, depending on the value of ρ. Moreover, it has the same limitation as all static penalty methods because it requires finding global minima of nonlinear functions.

There are many versions of dynamic penalty methods. A well-known one is the *nonstationary method,* which solves a sequence of problems like (12.3), with $C > 0$ and $\rho > 1$ constant parameters, at each iteration k, where

$$\alpha_j(k + 1) = \alpha_j(k) + C. \left| d_j(x) \right|,$$

$$\beta_i(k+1) = \beta_i(k) + C.\max(0, g_i(x)).$$

An advantage of this penalty method is that it requires only a few parameters to be tuned.

There are two other variations of global penalty methods that are exact methods: the refuse penalty methods and the discrete penalty methods.

Refuse Penalty Methods and Discrete Penalty Methods

Refuse penalty methods were proposed by Hu [HuEb02] and Zhang [Zh05]. A Refuse Penalty Method must start with a feasible point; it simply rejects all infeasible points. It begins with one or more points and searches for others, and if a new point is infeasible, it is rejected.

In this category, the main difficulty is to generate feasible initially points, particularly when the feasible region is small.

Thus, this method solves the problem (12.2) considering

$$L_p(x, \alpha, \beta) = f(x) + \alpha^T P(d(x)) + \beta^T Q(g(x)),$$

where

$$P(d(x)) = +\infty \text{ if } d(x) \neq 0, \quad P(d(x)) = 0 \text{ if } d(x) = 0,$$
$$Q(g(x)) = +\infty \text{ if } g(x) > 0, \quad Q(d(x)) = 0 \text{ if } g(x) \leq 0.$$

This is an exact penalty method. Given any finite penalty values α and β, the minimum point of the penalty function must be feasible and must have the minimum objective value, and therefore is exactly the CGM of P.

Another exact penalty method is the discrete penalty method, which uses the numbers of violated constraints instead of the degree of violations in the penalty function. This kind of method is often used in finite element methods [Da07].

It follows, therefore, that the methods of global optimization of (12.2) have limited practical application, because the search for the global minimum is computationally expensive; techniques of global optimization, such as the nonstationary method, are also slow, because they only get global optimality with asymptotic convergence [KiGeVe83].

12.4.2 Local Optimal Penalty Methods

To avoid costly global optimization methods, local optimal penalty methods (LOPM) were developed, to look for constrained local minima (CLM). These include, for example, methods of the Lagrange multipliers and l_1-penalty methods, which are both exact penalty methods. These methods were created to solve problems of nonlinear continuous optimization, that is, problems such as (12.2), where f is continuous and differentiable and g and d may be discontinuous and nondifferentiable.

In these methods, the goal is to find a local minimum $\breve{x}$ with respect to the neighborhood $N(\breve{x}) = \{x^* : \|x^* - \breve{x}\| \leq \epsilon \ e \ \epsilon \to 0\}$ of x^*.

Definition 3 ([GoOrTo03]). *A point $\breve{x}$ is a local minimum of P_m with respect to the neighborhood* $\mathrm{N}(\breve{x})$ *if $\breve{x}$ is feasible and $f(\breve{x}) \leq f(x)$ for all feasible $x \in \mathrm{N}(\breve{x})$.*

Lagrange Multiplier Methods

The Lagrangian function for (12.2) with Lagrange multipliers λ and μ is

$$L(x, \lambda, \mu) = f(x) + \lambda^T d(x) + \mu^T g(x).$$

Lagrange multiplier methods are designed for solving continuous nonlinear programming problems (CNLPs), so this approach is limited to solving CNLPs with continuous and differentiable functions and cannot be applied to solve discrete and mixed-integer problems.

This method limitation is due to the fact that the existence of Lagrange multipliers depends on the existence of the gradients of constraints and objective functions and the regularity conditions at the solution points.

Another local optimal exact penalty method is the l_1-penalty method.

l_1-Penalty Method

The l_1-*penalty method*, introduced by Pietrzykowski in 1969, solves minimization problems with the function [GoOrTo03]

$$l_1(x, \mu) = f(x) + \mu \sum_{j=1}^{t} |d_j(x)| + \mu \sum_{i=1}^{s+2n} \max[g_i(x), 0]. \tag{12.5}$$

The theory developed around this expression shows that there is a one-to-one correlation between the CLMs and the global minimum of l_1 function (12.5), when μ is large enough [Be99].

The most appealing feature of this method is that one choice of μ may be adequate for the entire minimization procedure; making it less dependent on the penalty parameter.

Function (12.5) forms the basis for many penalty methods proposed in the literature.

The difficulty of this method is the minimization of the l_1-penalty function because it is nonsmooth. As a result of these obstacles, this unconstrained approach is unlikely to be viable as a general-purpose technique for nonlinear programming. Techniques similar to Lagrange methods work for continuous and differentiable problems only.

12.5 Dynamic Penalty Methods

Unfortunately, the choice of suitable penalty parameters is, frequently, very difficult, because most of the strategies for choosing them are heuristic.

As an alternative to penalty functions, filter methods were introduced by Fletcher and Leyffer [FlLe02]. Since then, the filter technique has been mostly applied to sequential linear programming (SLP) and sequential quadratic programming (SQP) types of methods.

A filter algorithm introduces a function that aggregates constrained violations and constructs a biobjective problem. In this problem, the step is accepted if it either reduces the objective function or the constrained violation. This implies that the filter methods are less parameter dependent than a penalty function.

The difficulties of choosing appropriate values of penalty parameters in penalty methods caused nonsmooth penalty methods to fall out of favor during the early 1990s and stimulated the development of filter methods, which do not require a penalty parameter.

However, the new approach for updating the penalty parameter promises to solve these difficulties. It will automatically increase the penalty parameter and overcome this undesirable behavior, resulting in the dynamic penalty methods.

Dynamic penalty methods adjust the penalty parameter at every iteration so as to achieve a prescribed level of linear feasibility. The choice of the penalty parameter then ceases to be heuristic and becomes an integral part of the step computation.

An earlier form of the penalty update strategy is presented by Byrd et al. [ByNoWa06], in the context of a successive linear quadratic programming (SLQP) algorithm.

Other penalty strategies have been proposed recently. Chen and Goldfarb [ChGo05] propose rules that update the penalty parameter as optimality of the penalty problem is approached; they are based in part on feasibility and the size of the multipliers; Leyffer et al. [LeLoNo06] consider penalty methods for MPCCs and describe dynamic criteria for updating the penalty parameter based on the average decrease in the penalized constraints.

The methods of Byrd et al. [ByNoWa06] differ from these strategies in that they assess the effect of the penalty parameter on the step to be taken, based on the current model of the problem.

12.6 Conclusion

In this chapter, we review and classify some of the most popular existing penalty methods, and we discuss some of their assumptions and limitations.

This classification of the most popular existing Penalty Methods is essentially based on the type of minimum (global or local) that can found and the exactness of the solutions found (Exact or Inexact methods).

As future work we intend to create a web page with an application able to solve any constrained and unconstrained nonlinear problem.

As a first step we started to implement in the Java language the direct search methods. The next step is to implement the exact penalty methods and filter methods. It will then become possible to solve constrained nonlinear problems without the use of derivatives or their approximations.

References

[BeSeSh03] Benson, H.Y., Sen, A., Shanno, D.F., Vanderbei, R.J.: Interior-point algorithms, penalty methods and equilibrium problems. Technical Report ORFE-03-02, Princeton University, Princeton, NJ (2003).

[Be99] Bertsekas, D.P.: *Nonlinear Programming,* Athena Scientific, Belmont, MA (1999).

[ByNoWa06] Byrd, R.H., Nocedal, J., Waltz, R.A.: Steering exact penalty methods for optimization. Technical Report, Northwestern University, Chicago, IL (2006).

[ChGo05] Chen, L., Goldfarb, D.: Interior-point $l2$-penalty methods for nonlinear programming with strong global convergence properties. Technical Report, Columbia University, New York, NY (2005).

[Da07] Dai, X.: Finite element approximation of the pure Neumann problem using the iterative penalty method. *Appl. Math. Comput.,* **186**, 1367–1373 (2007).

[FlLe02] Fletcher, R., Leyffer, S.: Nonlinear programming without a penalty function. *Math. Programming,* **91**, 239–270 (2002).

[FrRo04] Freund, R.M.: *Penalty and Barrier Methods for Constrained Optimization,* Massachusetts Institute of Technology, Cambridge, MA (2004).

[GoOrTo03] Gould, N.I., Orban, D., Toint, P.L.: An interior-point l_1-penalty method for nonlinear optimization. Technical Report RAL-TR-2003-022, Rutherford Appleton Laboratory, Chilton, UK (2003).

[HoLaQi95] Homaifar, A., Lai, S.H.V., Qi, X.: Constrained optimization via generic algorithms. *Simulation,* **62**, 242–254 (1994).

[HuEb02] Hu, X., Eberhart, R.: Solving constrained nonlinear optimization problems with particle swarm optimization, in *Proceedings Sixth World Multiconf. on Systemics, Cybernetics and Informatics,* Orlando, FL (2002).

[KiGeVe83] Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. *Science,* **220**, 671–680 (1983).

[KlKu02] Klatte, D., Kummer, B.: Constrained minima and Lipschitzian penalties in metric spaces. *SIAM J. Optimization,* **13**, 619–633 (2002).

[LeLoNo06] Leyffer, S., López-Calva, G., Nocedal: Interior methods for mathematical programs with complementarity constraints. *SIAM J. Optimization,* **17**, 52–77 (2006).

[MoSa95] Mongeau, M., Sartenaer, A.: Automatic decrease of the penalty parameter in exact penalty function methods. *European J. Oper. Research,* **83**, 686–699 (1995).

[WaLi06] Wang, F.Y., Liu, D.: *Advances in Computational Intelligence: Theory and Applications,* World Scientific, Singapore (2006).

[Za05] Zaslavski, A.J.: A sufficient condition for exact penalty in constrained optimization. *SIAM J. Optimization,* **16**, 250–262 (2005).

[Zh05] Zhang, S.: Constrained optimization by ϵ constrained hybrid algorithm of particle swarm optimization and genetic algorithm, in *Advances in Artificial Intelligence,* Springer, Berlin (2005).

13

A Closed-Form Formulation for Pollutant Dispersion in the Atmosphere

C.P. Costa,[1] M.T. Vilhena,[2] and T. Tirabassi[3]

[1] Universidade Federal de Pelotas, Brazil; camiladacosta@gmail.com
[2] Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil; vilhena@pq.cnpq.br
[3] Istituto di Scienze dell'Atmosfera e del Clima, Bologna, Italy; t.tirabassi@isac.cnr.it

13.1 Introduction

Transport and diffusion models of air pollution are based either on simple techniques, such as the Gaussian approach, or on more complex algorithms, such as the K-theory differential equation. The Gaussian equation is an easy and fast method, which, however, cannot properly simulate complex nonhomogeneous conditions. The K-theory can accept virtually any complex meteorological input, but generally requires numerical integration, which is computationally expensive and is often affected by large numerical advection errors. Conversely, Gaussian models are fast, simple, do not require complex meteorological input, and describe the diffusive transport in an Eulerian framework, making easy use of the Eulerian nature of measurements.

For these reasons they are still widely used by environmental agencies all over the world for regulatory applications. However, because of its well-known intrinsic limits, the reliability of a Gaussian model strongly depends on the way the dispersion parameters are determined on the basis of the turbulence structure of the planetary boundary layer (PBL) and the model's ability to reproduce experimental diffusion data. The Gaussian model has to be completed by empirically determined standard deviations (the "sigmas"), while some commonly measurable turbulent exchange coefficient has to be introduced in the advection–diffusion equation.

To overcome this drawback, we propose an analytical solution of the advection–diffusion equation with any restriction to wind and eddy diffusion vertical profiles, which is believed to give a better representation of the effects due to the vertical stratification of the atmosphere and while maintaining the simplicity of an analytical formulation.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_13,

13.2 Description of the Model

The nonstationary advection–diffusion equation that models air pollution in the atmosphere is essentially a statement of conservation of the suspended material. In a Cartesian coordinate system with the x-axis aligned in the direction of the wind, the y-axis oriented in the horizontal crosswind direction, and the z-axis chosen vertically upwards, this equation has the form [Ar95]

$$\frac{\partial \overline{c}}{\partial t} + \overline{u}\frac{\partial \overline{c}}{\partial x} + \overline{v}\frac{\partial \overline{c}}{\partial y} + \overline{w}\frac{\partial \overline{c}}{\partial z} = -\frac{\partial \overline{u'c'}}{\partial x} - \frac{\partial \overline{v'c'}}{\partial y} - \frac{\partial \overline{w'c'}}{\partial z} + S, \tag{13.1}$$

where $\overline{c}$ denotes the average concentration, u, v, and w are the Cartesian components of the wind, and S is the source term. The terms $\overline{u'c'}$, $\overline{v'c'}$, and $\overline{w'c'}$ represent, respectively, the turbulent fluxes of contaminants in the longitudinal, crosswind, and vertical directions.

The concentration turbulent fluxes are assumed to be proportional to the mean concentration gradient, which is known as Fick theory or local turbulence closure:

$$\overline{u'c'} = -K_x\frac{\partial \overline{c}}{\partial x}, \quad \overline{v'c'} = -K_y\frac{\partial \overline{c}}{\partial y}, \quad \overline{w'c'} = -K_z\frac{\partial \overline{c}}{\partial z},$$

where K_x, K_y, and K_z are the Cartesian components of eddy diffusivity.

This assumption, combined with the continuity equation, leads to the advection–diffusion equation. For a Cartesian coordinate system in which z is the height, we rewrite the advection–diffusion equation in the form

$$\frac{\partial \overline{c}}{\partial t} + \overline{u}\frac{\partial \overline{c}}{\partial x} + \overline{v}\frac{\partial \overline{c}}{\partial y} + \overline{w}\frac{\partial \overline{c}}{\partial z} = \frac{\partial}{\partial x}\left(K_x\frac{\partial \overline{c}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial \overline{c}}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z\frac{\partial \overline{c}}{\partial z}\right) + S \tag{13.2}$$

for $t > 0$, $0 < z < h$, $0 < y < L_y$, and $x > 0$, where h is the height of the PBL and L_y is a limit on the y-axis which is positive and far from the source. In this chapter, we consider that the vertical (w) and lateral (v) components of the mean flow are null. Moreover, the mean horizontal flow is incompressible and horizontally homogeneous. We also neglect diffusion in the x-direction ($K_x = 0$). In view of this hypothesis, we recast (13.2) in the form

$$\frac{\partial \overline{c}}{\partial t} + \overline{u}\frac{\partial \overline{c}}{\partial x} = \frac{\partial}{\partial y}\left(K_y\frac{\partial \overline{c}}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z\frac{\partial \overline{c}}{\partial z}\right) + S. \tag{13.3}$$

We assume that when the pollutant is released, the dispersion pollutant domain is not contaminated; that is,

$$\overline{c}(x, y, z, 0) = 0 \quad \text{at } t = 0.$$

We also consider, respectively, zero flux in the z-direction at ground and PBL top as well as in the y-direction at $y = 0$, L_y:

$$K_z \frac{\partial \bar{c}}{\partial z} = 0 \quad \text{at } z = 0,\, z_i, \quad K_y \frac{\partial \bar{c}}{\partial y} = 0 \quad \text{at } y = 0,\, L_y.$$

Finally, we assume a continuous point source of constant emission rate Q, written as [Ar95]

$$\overline{u}\, \bar{c}\,(0, y, z, t) = Q\, \delta(z - H_s)\, \delta(y),$$

where δ is the Dirac delta, $x = 0$, $z = H_s$, and $y = y_0$ is the source position.

In order to solve the advection–diffusion equation for inhomogeneous turbulence, we must take into account the dependence of the eddy diffusivities and wind speed profiles on the height variable z. Therefore, we use the idea of the advection–diffusion multilayer model (ADMM) method, where we perform a stepwise approximation of these coefficients (see [CoEA06] and [MoEA06]). To reach this goal, we discretize the height h of the PBL into N subintervals in such a manner that inside each subregion the eddy diffusivities and wind velocities assume average values:

$$K_{\alpha_n} = \frac{1}{z_n - z_{n-1}} \int_{z_{n-1}}^{z_n} K_\alpha(z) dz, \quad \overline{u}_n = \frac{1}{z_n - z_{n-1}} \int_{z_{n-1}}^{z_n} \overline{u}_z(z) dz$$

for $n = 1, \ldots, N$.

We are now in a position to solve the advection–diffusion equation for each subinterval. Indeed, it is now possible to recast problem (13.2) as a set of advective-diffusive problems with constant parameters; specifically, for a generic sublayer we have

$$\frac{\partial \overline{c}_n}{\partial t} + \overline{u_n} \frac{\partial \overline{c}_n}{\partial x} = K_{y_n} \frac{\partial^2 \overline{c}_n}{\partial y^2} + K_{z_n} \frac{\partial^2 \overline{c}_n}{\partial z^2} \tag{13.4}$$

for $n = 1, \ldots, N$, where N denotes the number of sublayers and $\overline{c}_n$ denotes the concentration on the nth subinterval.

In addition, two boundary conditions are imposed at 0 and h, and continuity conditions for the concentration and flux of concentration at the interfaces; that is, we must have

$$\overline{c}_n = \overline{c}_{n+1}, \quad z = z_n, \quad n = 1, 2, \ldots, N-1,$$
$$K_{z_n} \frac{\partial \overline{c}_n}{\partial z} = K_{z_{n+1}} \frac{\partial \overline{c}_{n+1}}{\partial z}, \quad z = z_n, \quad n = 1, 2, \ldots, N-1,$$

in order to uniquely determine the $2N$ arbitrary constants appearing in the solution of the set of problems (13.4).

We now apply the Laplace transformation with respect to time in (13.4) and set $\pounds\,\{\overline{c}_n(x, y, z, t)\} = \Gamma_n(x, y, z, \gamma)$. This procedure leads to the stationary problem

$$\overline{u_n} \frac{\partial \Gamma_n}{\partial x} = K_{y_n} \frac{\partial^2 \Gamma_n}{\partial y^2} + K_{z_n} \frac{\partial^2 \Gamma_n}{\partial z^2} - \gamma \Gamma_n. \tag{13.5}$$

At this point, proceeding as in [CoEA06], we are in a position to solve problem (13.5) by the GIADMT (*generalized integral advection–diffusion multilayer technique*). To this end, we begin by expanding the solution in the series

$$\Gamma_n(x,y,z,\gamma) = \sum_{j=0}^{\infty} \frac{\overline{c}_{j\,n}(x,z,\gamma)\,\psi_j(y)}{\sqrt{N_j}}, \tag{13.6}$$

where, as in the generalized integral transform technique (GITT), $\psi_j(y) = \cos(\lambda_j y)$ and $\lambda_j = j\pi/Ly$ are, respectively, the eigenfunctions and eigenvalues.

Substituting (13.6) in (13.5), we arrive at

$$\sum_{j=0}^{\infty} \overline{u}_n \frac{\partial \overline{c}_{j\,n}(x,z,\gamma)}{\partial x} \frac{\psi_j(y)}{N_j^{1/2}} = \sum_{j=0}^{\infty} K_{y_n} \overline{c}_{j\,n}(x,z,\gamma) \frac{\psi_j''(y)}{N_j^{1/2}} + \sum_{j=0}^{\infty} K_{z_n} \frac{\partial^2 \overline{c}_{j\,n}(x,z,\gamma)}{\partial z^2} \frac{\psi_j(y)}{N_j^{1/2}} - \gamma \sum_{j=0}^{\infty} \overline{c}_{j\,n}(x,z,\gamma) \frac{\psi_j(y)}{N_j^{1/2}},$$

where the double prime denotes the second-order y-derivative.

Taking moments and solving the resulting transformed problem by the Laplace transformation technique with respect to the x variable, we find that

$$K_{z_n} \frac{d^2 \overline{\overline{c}_{j\,n}}(s,z,\gamma)}{dz^2} - \left(s\,u_n + K_{y_n}\,\lambda_j^2 + \gamma\right) \overline{\overline{c}_{j\,n}}(s,z,\gamma) = -\left(\frac{1}{\gamma}\frac{\psi_j(y_0)}{N_j^{1/2}}\right) Q\,\delta(z-H_s),$$

which has the well-known solution

$$\overline{\overline{c}_{j\,n}}(s,z,\gamma) = C_{1_n} e^{R_{jn} z} + C_{2_n} e^{-R_{jn} z} + \frac{Q}{2R_{a_{nj}}} + \left[e^{R_{jn}(z-H_s)} - e^{-R_{jn}(z-H_s)}\right] H(z-H_s), \tag{13.7}$$

where $\overline{\overline{c}_{jn}}$ denotes the Laplace transform of $\overline{c}_{j\,n}$ with respect to the x-variable and the parameters $R_{j\,n}$ and $R_{a_{nj}}$ are defined by

$$R_{jn} = \sqrt{\frac{s\,u_n + K_{y_n}\,\lambda_j^2 + \gamma}{K_{z_n}}},$$

$$R_{a_{nj}} = \frac{\gamma}{\psi_j(y_0)} \sqrt{N_j K_{z_n}\left(s\,u_n + K_{y_n}\,\lambda_j^2 + \gamma\right)}.$$

Therefore, we obtain the coefficients of the series solution given by (13.6) by performing the inverse Laplace transformation on the transformed solution appearing in (13.7).

Applying the boundary and interface conditions, we determine the unknown integration constants (C_{1n} and C_{2n}) from the resulting linear system. Once the coefficients $\overline{c}_{nj}$ are known, we find that

$$\begin{aligned} c_n(x,y,z,t) \\ = \frac{1}{2\pi i}\int_{\zeta-i\infty}^{\zeta+i\infty} e^{\gamma\, t}\sum_{j=0}^{\infty}\frac{\psi_j(y)}{\sqrt{N_j}}\Bigg\{\frac{1}{2\pi i}\int_{\xi-i\infty}^{\xi+i\infty} e^{s\,x}\big[\big(C_{1_n}e^{R_{jn}z}+C_{2_n}e^{-R_{jn}z}\big) \\ + \frac{Q}{2R_{a_{nj}}}\big(e^{R_{jn}(z-H_s)}-e^{-R_{jn}(z-H_s)}\big)H(z-H_s)\big]ds\Bigg\}d\gamma. \quad (13.8) \end{aligned}$$

To overcome the difficulty of evaluating the line integral appearing in the solution given by (13.8), we perform the double integration numerically by the fixed Talbot (FT) method [AbVa04] in the x-variable and by the Gaussian quadrature scheme [StSe66] in the time variable.

13.3 Nonlocal Closure

Already some decades ago it was noted that in the upper part of convectively driven boundary layers, the flux of potential temperature is counter to the gradient of the mean potential temperature profile [Er42] and [De72]. The mean potential temperature gradient and the flux change sign at different levels, introducing a certain region in the convective boundary layer where they have the same sign. This result was in contrast to the common view in first order turbulent closure that turbulent diffusion is down gradient. In order to also describe diffusion in these regions, [Er42] and [De72] proposed to modify the usual applied flux-gradient relationship in K-theory approach according to

$$\overline{w'c'} = -K_z\left(\frac{\partial\overline{c}}{\partial z}-\gamma\right),$$

where γ represents the counter gradient term.

Many schemes and parameterizations for the counter gradient term have been developed. In this chapter, we use the parameterization proposed by van Dop and Verver (2001) [VaVe01], based on the work in [WyWe91]; that is,

$$\left(1+\frac{S_k\,\sigma_w\,T_{L_w}}{2}\frac{\partial}{\partial z}+\tau\frac{\partial}{\partial z}\right)\overline{w'c'} = -K_z\,\frac{\partial\overline{c}}{\partial z}, \quad (13.9)$$

where S_k is the skewness, T_{L_w} is the vertical Lagrangian time scale, σ_w is the vertical turbulent velocity variance and τ is the relaxation time. The second term on the left-hand side of (13.9) represents the nonlocal counter gradient term as proposed in [VaVe01]; it is obtained by applying the Taylor expansion to the turbulent flux [WyWe91]. Substituting the above ansatz in (13.1), we arrive at the problem

$$\frac{\partial \overline{c}}{\partial t} + \tau \frac{\partial^2 \overline{c}}{\partial t^2} + \beta \frac{\partial^2 \overline{c}}{\partial z\, \partial t} + \overline{u} \frac{\partial \overline{c}}{\partial x} + \overline{u}\beta \frac{\partial^2 \overline{c}}{\partial z\, \partial x} + \overline{u}\tau \frac{\partial^2 \overline{c}}{\partial t\, \partial x}$$
$$= K_y \frac{\partial^2 \overline{c}}{\partial y^2} + K_y \beta \frac{\partial^3 \overline{c}}{\partial z\, \partial y^2} + K_y \tau \frac{\partial^3 \overline{c}}{\partial t\, \partial y^2} + K_z \frac{\partial^2 \overline{c}}{\partial z^2}, \quad (13.10)$$

where $\beta = \frac{S_k\, \sigma_w\, T_{L_w}}{2}$. In this equation, $t > 0$, $0 < z < h$, $0 < y < L_y$, where L_y is a large distance from the source, and $x > 0$.

We solve (13.10) by a procedure similar to the one above, using the solution of the stationary problem obtained by the GIADMT method as in the works of Costa et al. (2006) [CoEA06] and Moreira at al. (2006) [MoEA06], and applying the Laplace transformation technique with respect to the t variable. Finally, we obtain

$$c_n(x, y, z, t)$$
$$= \frac{1}{2\pi i} \int_{\zeta - i\infty}^{\zeta + i\infty} e^{\gamma t} \sum_{j=0}^{\infty} \frac{\psi_j(y)}{\sqrt{N_j}} \left\{ \frac{1}{2\pi i} \int_{\xi - i\infty}^{\xi + i\infty} e^{sx} \left[\left(C_{1_n} e^{(F_{jn} + G_{jn})z} + C_{2_n} e^{(F_{jn} - G_{jn})z} \right) \right.\right.$$
$$\left.\left. + \frac{Q}{2G_{a_{nj}}} \left(e^{(F_{jn} + G_{jn})(z - H_s)} - e^{(F_{jn} - G_{jn})(z - H_s)} \right) H(z - H_s) \right] ds \right\} d\gamma, \quad (13.11)$$

where

$$F_{jn} = \frac{\beta_n}{2K_{z_n}} \left(s\overline{u}_n + K_{y_n} \lambda_j^2 + \gamma \right),$$

$$G_{jn} = \sqrt{(F_{jn})^2 + \frac{4}{K_{z_n}} \left(s\overline{u}_n + K_{y_n} \lambda_j^2 + \gamma + \gamma \overline{u}_n \tau s + \tau \gamma^2 + \lambda_j^2 \gamma \tau K_{y_n} \right)},$$

$$G_{a_{nj}} = \frac{\gamma \sqrt{N_j K_{z_n}}}{2\,(1 + \gamma\tau)\, \psi_j(y_0)}\, G_{jn}.$$

13.4 Numerical Simulation

In order to show the performance of the present solution of the advection–diffusion equation for nonstationary conditions, and to evaluate the performance of the proposed PBL parameterization, we applied the model using the Copenhagen experimental datasets [GrLy84].

To do this, we had to introduce a boundary layer parameterization. The literature contains many, widely different formulas for the calculation of the vertical turbulent diffusion coefficient [SePa98]. As examples of applications of our new solution, we tested the vertical and lateral diffusion parameterization suggested by Degrazia et al. [DeEA97] for convective conditions. The wind speed profile was described by a power law expressed as in [PaDu88].

The Copenhagen dataset was chosen since most of the experiments were performed during moderately unstable atmospheric conditions, and without strong buoyancy, so that ground-level crosswind integrated concentration can be simulated by an advection–diffusion equation. The stability parameter z_i/L (L is the Monin–Obukhov length) indicates cases where the unstable PBL presents weak to moderate convection.

Figure 13.1 shows the observed and predicted scatter diagram of centerline

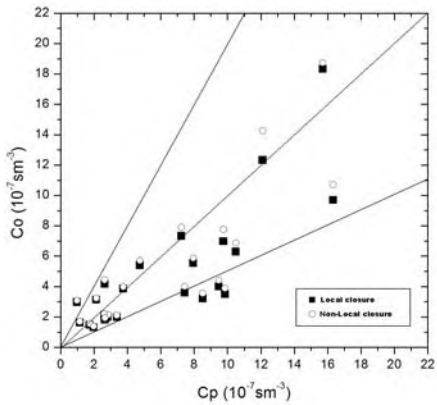


Fig. 13.1. Scatter plot of observed (Co) and computed (Cp) arc maximum level concentration normalized with emission rate (c/Q). The data between two external lines are within a factor of 2 ($Co/Cp \in [0.5; 2]$).

ground-level concentrations for the solutions (13.8) and (13.11), normalized with the emission source rate (c/Q). This figure points out that a good agreement is obtained between experimental data and the model considering the local and nonlocal closure.

In Figure 13.2, we show the time evolution of pollutant concentration for several source distances for the Copenhagen experiment assuming both local and nonlocal turbulence closure. For all cases, we quickly realized that, as time passes, the pollutant concentration reaches, as expected, the stationary regime.

For the non-Fickian problem, we must notice the influence of the nonlocal transport, once the asymmetry observed in the pollutant concentration plays an important role, because it is responsible for the dislocation of the maximum concentration peak.

Figure 13.3 shows an example of concentration distributions in the horizontal (x, y)-plane at ground level for the local and nonlocal turbulence closure. The lines represent isolines of equal concentration. Here once again we realize the effect of pollutant dispersion for nonlocal turbulence closure when compared with local closure, in the sense that we observe the increase of the

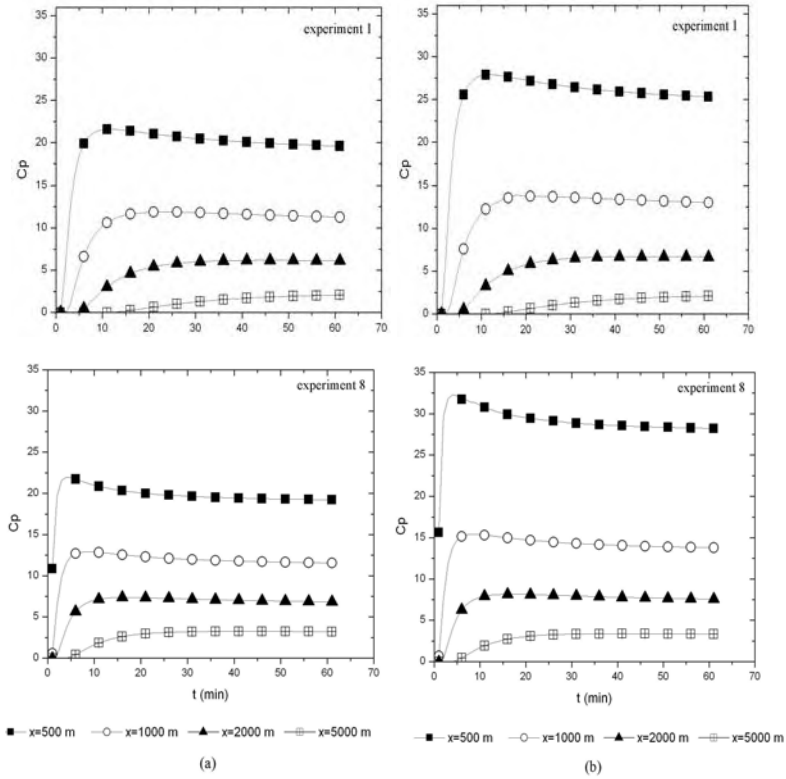


Fig. 13.2. Time evolution of pollutant concentration for several source distances for the Copenhagen experiment with (a) local turbulence closure and (b) nonlocal turbulence closure.

maximum value of the concentration as well as its shifting to the right. We also notice an asymmetry in the concentration, which does not occur for local turbulence closure.

The results of the statistical indices used to evaluate the models are shown in Table 13.1. The statistical indices are defined in [Ha89].

Table 13.1. Statistical indices used to evaluate the model performance.

Model	*Nmse*	*Cor*	*Fa2*	*Fb*	*Fs*
Local closure	0.29	0.81	0.78	0.26	0.13
Nonlocal closure	0.23	0.83	0.83	0.18	0.07

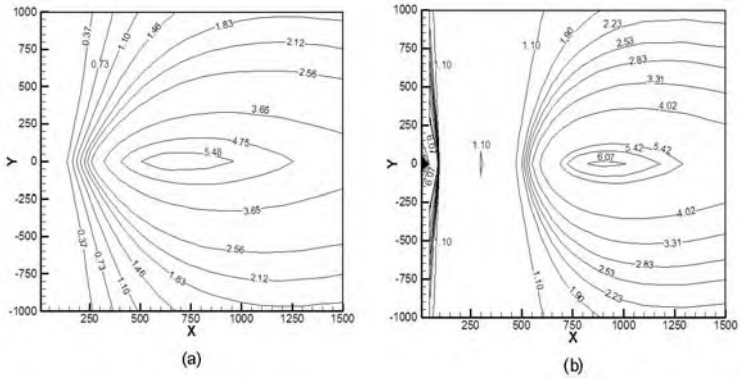


Fig. 13.3. (x, y)-concentration cross sections at ground-level concentration normalized with emission rate (c/Q): (a) local closure and (b) nonlocal closure.

13.5 Conclusions

The good computational results obtained together with the analytical feature of the solution encountered in the approximation considered for the advection-diffusion equation—except for the stepwise approximation of the eddy-diffusivity coefficient—give us confidence that the proposed solution is a promising and efficient tool for predicting air pollutant dispersion in the atmosphere.

References

[AbVa04] Abate, J., Valkó, P.P.: Multi-precision Laplace transform inversion. *Internat. J. Numer. Methods Engng.,* **60**, 979–993 (2004).

[Ar95] Arya, P.: Modeling and parameterization of near-source difusion in weak winds. *J. Appl. Meteorology,* **34**, 1112–1122 (1995).

[CoEA06] Costa, C.P., Vilhena, M.T., Moreira, D.M., Tirabassi, T.: Semi-analytical solution of the steady three-dimensional advection–diffusion equation in the planetary boundary layer. *Atmospheric Environment,* **40**, 5659–5669 (2006).

[De72] Deardoff, J.W.: Theoretical expression for the countergradient heat flux. *J. Geophys. Res. Pap.,* **59**, 5900–5904 (1972).

[DeEA97] Degrazia, G.A., Velho, H.F.C., Carvalho, J.C.: Nonlocal exchange coefficients for the convective boundary layer derived from spectral properties. *Contributions to Atmosph. Phys.,* **40**, 57–64 (1997).

[Er42] Ertel, H.: *Der Vertikale Turbulenz-Wärmestrom in der Atmosphäre, Meteor. Z.,* **59**, 250–253 (1942).

[Ha89] Hanna, S.R.: Confidence limit for air quality models as estimated by bootstrap and jacknife resampling methods. *Atmospheric Environment,* **23**, 1385–1395 (1989).

[MoEA06] Moreira, D.M., Vilhena, M.T., Tirabassi, T., Costa, C.P., Bodmann, B.: Simulation of pollutant dispersion in the atmosphere by the Laplace transform: the ADMM approach. *Water, Air, and Soil Pollution,* **177**, 411–439 (2006).

[PaDu88] Panofsky, H.A., Dutton, J.A.: *Atmospheric Turbulence,* Wiley, New York (1988).

[SePa98] Seinfeld, J.H., Pandis, S.N.: *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change,* Wiley, New York (1998).

[StSe66] Stroud, A.H., Secrest, D.: *Gaussian Quadrature Formulas,* Prentice-Hall, Englewood Cliffs, NJ (1966).

[VaVe01] van Dop, H., Verver, G.: Countergradient transport revisited. *J. Atmospheric Sci.,* **58**, 2240–2247 (2001).

[WyWe91] Wyngaard, J.C., Weil, J.C.: Transport asymmetry in skewed turbulence. *Phys. Fluids,* **A3**, 155–162 (1991).

[GrLy84] Gryning, S.E., Lyck, E.: Atmospheric dispersion from elevated sources in an urban area: comparison between tracer experiments and model calculations. *J. Climate Appl. Meteorology,* **23**, 651–660 (1984).

14

High-Order Methods for Weakly Singular Volterra Integro-Differential Equations

T. Diogo,[1] M. Kolk,[2] P. Lima,[1] and A. Pedas[2]

[1] Centro de Matemática e Aplicações, IST, UTL, Lisbon, Portugal; tdiogo@math.ist.utl.pt, plima@math.ist.utl.pt
[2] University of Tartu, Estonia; marek.kolk@ut.ee, arvet.pedas@ut.ee

14.1 Introduction

Let $\mathbb{R} = (-\infty, \infty)$, $b > 0$, $\Delta_b = \{(t,s) \in \mathbb{R}^2 : 0 \le t \le b, 0 \le s < t\}$, $\overline{\Delta}_b = \{(t,s) \in \mathbb{R}^2 : 0 \le s \le t \le b\}$. We consider a linear integro-differential equation of the form

$$y'(t) = p(t)y(t) + q(t) + \int_0^t K_0(t,s)y(s)ds + \int_0^t K_1(t,s)y'(s)ds\,, \quad 0 \le t \le b\,, \tag{14.1}$$

with given initial condition

$$y(0) = y_0\,, \quad y_0 \in \mathbb{R}\,. \tag{14.2}$$

We assume that $K_0, K_1 \in W^{m,\nu}(\Delta_b)$, $p, q \in C^{m,\nu}(0,b]$, $m \in \mathbb{N} = \{1, 2, ...\}$, $\nu \in \mathbb{R}$, $\nu < 1$.

For given $m \in \mathbb{N}$ and $-\infty < \nu < 1$ we define $W^{m,\nu}(\Delta_b)$ as the set of all m-times continuously differentiable functions $K : \Delta_b \to \mathbb{R}$ satisfying

$$\left|\left(\frac{\partial}{\partial t}\right)^i \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right)^j K(t,s)\right| \le c \begin{cases} 1 & \text{if} \quad \nu + i < 0\,, \\ 1 + |\log(t-s)| & \text{if} \quad \nu + i = 0\,, \\ (t-s)^{-\nu-i} & \text{if} \quad \nu + i > 0\,, \end{cases} \tag{14.3}$$

with a constant $c = c(K)$ for all $(t,s) \in \Delta_b$ and all nonnegative integers i and j such that $i + j \le m$.

It follows from (14.3) (with $i = j = 0$, $0 \le \nu < 1$) that the kernels $K_0(t,s)$ and $K_1(t,s)$ of equation (14.1) may possess a weak singularity as $s \to t$. In the case $\nu < 0$ the kernels K_0 and K_1 are bounded on Δ_b but their derivatives may be singular as $s \to t$. In particular, if $K_0 = 0$ and $K_1(t,s) = \kappa(t,s)(t-s)^{-\nu}$, $0 < \nu < 1$, where κ is an m-times continuously differentiable function on $\overline{\Delta}_b$, then (14.1) is a Basset-type equation (see [BrTa89], [McSt83]).

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_14,

The set $C^{m,\nu}(0,b]$ ($m \in \mathbb{N}$, $-\infty < \nu < 1$) consists of continuous functions $y : [0,b] \to \mathbb{R}$ which are m-times continuously differentiable in $(0,b]$ and whose derivatives satisfy

$$\left|y^{(j)}(t)\right| \le c \left\{ \begin{array}{lll} 1 & \text{if} & j < 1-\nu \\ 1+|\log t| & \text{if} & j = 1-\nu \\ t^{1-\nu-j} & \text{if} & j > 1-\nu \end{array} \right\}, \; 0 < t \le b, \; j = 0, 1, ..., m.$$

Throughout the text c denotes a positive constant which may have different values for different occurrences. Note that $C^m[0,b]$, the set of m times ($m \ge 1$) continuously differentiable functions $y : [0,b] \to \mathbb{R}$, is a subset of $C^{m,\nu}(0,b]$ for arbitrary $\nu < 1$. On the other hand, if $y \in C^{m,\nu}(0,b]$ and $\nu < 1-k$ for $k \in \{1, ..., m\}$, then the derivative $y^{(k)}$ is bounded on $(0,b]$ and the derivatives $y', ..., y^{(k-1)}$ of y can be extended so that $y \in C^{k-1}[0,b]$. Here and below, we denote by $C^0[0,b] \equiv C[0,b]$ the Banach space of continuous functions $y : [0,b] \to \mathbb{R}$ equipped with the usual norm $||y|| = \max\{|y(t)| : 0 \le t \le b\}$.

If $K_0, K_1 \in W^{m,\nu}(\Delta_b)$, $p, q \in C^{m,\nu}(0,b]$, $m \in \mathbb{N}$, $\nu \in \mathbb{R}$, $\nu < 1$, then problem (14.1),(14.2) has a unique solution $y \in C^{m+1,\nu-1}(0,b]$ (see [PaPe03], [Pe04]). Thus, the solution y of problem (14.1),(14.2) may not belong to $C^2[0,b]$. In collocation methods, the possible singular behavior of the solution of (14.1), (14.2) can be taken into account by using polynomial splines on special nonuniform grids [BrTa89], [Pe04]. A problem which may arise is that the use of strongly nonuniform grids may create significant round-off errors in calculations and therefore lead to unstable behavior of numerical results.

The purpose of this chapter is to construct high-order algorithms for the numerical solution of problem (14.1), (14.2) which do not need strongly graded grids. To this end, we first introduce an equivalent integral equation reformulation of the original problem. Then, following an approach used in [PeVa04], we apply a smoothing transformation so that the singularities of the derivatives of the exact solution of the resulting equation will be milder or disappear. After that, we solve the transformed equation by a piecewise polynomial collocation method on a mildly graded or uniform grid. Finally, some numerical results are presented.

We note that collocation and related methods for various weakly singular Volterra integro-differential equations of the form (14.1), with $K_1 = 0$, have been studied by many authors, see, e.g., [BaOr06], [Br04], [BrHo86], [BrPeVa01a], [BrPeVa01b], [KaPa03], [Pa05], [Ta92], and [Ta93]. We also refer to [CaEtAl07], [PeTa06], and [PeTa08], where the numerical solution of weakly singular Fredholm integro-differential equations is considered.

14.2 Reformulation of the Original Problem and Smoothing

Using the notation $y' = z$ and applying condition (14.2), we may rewrite equation (14.1) as a linear Volterra integral equation of the second kind with respect to z:

$$z(t) = \int_0^t K_0(t,s)\left(\int_0^s z(\tau)d\tau\right)ds + \int_0^t \Big[p(t) + K_1(t,s)\Big]z(s)ds + f(t), \quad (14.4)$$

$$f(t) = q(t) + y_0 p(t) + y_0 \int_0^t K_0(t,s)ds\,, \quad t \in [0,b].$$

Lemma 1. *Let $K_0, K_1 \in W^{m,\nu}(\Delta_b)$, $p, q \in C^{m,\nu}(0,b]$, $m \in \mathbb{N}$, $-\infty < \nu < 1$. Then equation* (14.4) *has a unique solution $z \in C^{m,\nu}(0,b]$.*

The proof of this lemma can be found in [PaPe03], [Pe04].

Let us now introduce the class of functions φ defined by

$$\varphi(x) = b^{1-d}x^d, \quad 0 \le x \le b, \quad d \in \mathbb{N}. \quad (14.5)$$

Clearly, $\varphi \in C[0,b]$, $\varphi(0) = 0$, $\varphi(b) = b$, and $\varphi'(x) > 0$ for $0 < x < b$. Thus, φ maps $[0,b]$ onto $[0,b]$ and has a continuous inverse $\varphi^{-1} : [0,b] \to [0,b]$, $\varphi^{-1}(t) = b^{(d-1)/d}\, t^{1/d}$, $0 \le t \le b$. Note that $\varphi(x) \equiv x$ for $d = 1$.

We are interested in transformations φ with $d > 1$ since they possess a smoothing property for $z(\varphi(x))$ with singularities of $z(t)$ at $t = 0$ (see Lemma 2).

Lemma 2 ([PeVa04]). *Assume $z \in C^{m,\nu}(0,b]$, $m \in \mathbb{N}$, $\nu \in \mathbb{R}$, $\nu < 1$. Let $z_\varphi(x) = z(\varphi(x))$, $x \in [0,b]$. Then $z_\varphi \in C^{m,\nu_d}(0,b]$ with $\nu_d = 1 - d(1-\nu)$.*

Using (14.5), we make a change of variables in equation (14.4), in order to obtain a new integral equation whose solution does not involve any more singularities in its derivatives up to a certain order. Introducing in (14.4) the variables transformation

$$t = \varphi(x),\ s = \varphi(\mu),\ \tau = \varphi(\sigma), \quad x,\, \mu,\, \sigma \in [0,b],$$

we obtain an integral equation of the form

$$z_\varphi(x) = \int_0^x K_{0,\varphi}(x,\mu)\Big(\int_0^\mu z_\varphi(\sigma)\varphi'(\sigma)d\sigma\Big)d\mu$$
$$+ \int_0^x \Big[p_\varphi(x)\varphi'(\mu) + K_{1,\varphi}(x,\mu)\Big]z_\varphi(\mu)d\mu + f_\varphi(x),\ 0 \le x \le b, \quad (14.6)$$

where

$$f_\varphi(x) = f(\varphi(x)), \quad p_\varphi(x) = p(\varphi(x)),$$
$$K_{0,\varphi}(x,\mu) = K_0(\varphi(x),\varphi(\mu))\varphi'(\mu), \quad K_{1,\varphi}(x,\mu) = K_1(\varphi(x),\varphi(\mu))\varphi'(\mu)$$

are the given functions and $z_\varphi(x) = z(\varphi(x))$ is a function which we have to find. Changing the order of integration in the double integral of (14.6) leads to

$$(I - T_\varphi)z_\varphi = f_\varphi\,, \tag{14.7}$$

where I is the identity mapping and

$$(T_\varphi z_\varphi)(x) = \int_0^x L_\varphi(x,\mu)z_\varphi(\mu)d\mu,\ x \in [0,b], \tag{14.8}$$

with

$$L_\varphi(x,\mu) = p_\varphi(x)\varphi'(\mu) + \varphi'(\mu)\int_\mu^x K_{0,\varphi}(x,\sigma)d\sigma + K_{1,\varphi}(x,\mu), \quad 0 \le \mu < x \le b.$$

Due to (14.3) we have

$$\left|K_{i,\varphi}(x,\mu)\right| = \left|K_i(\varphi(x),\varphi(\mu))\right|\varphi'(\mu) \le c\begin{cases} 1 & \text{for } \nu < 0, \\ 1 + |\log(x-\mu)| & \text{for } \nu = 0, \\ (x-\mu)^{-\nu} & \text{for } \nu > 0, \end{cases}$$

where $0 \le \mu < x \le b$, $i = 0$, $i = 1$. Since $K_0 \in \mathcal{W}^{m,\nu}(\Delta_b)$, $\int_\mu^x K_{0,\varphi}(x,\sigma)d\sigma$ is a continuous function for $0 \le \mu \le x \le b$. Then, since $p_\varphi, \varphi' \in C[0,b]$, it follows that $L_\varphi(x,\mu)$ is continuous for $0 \le \mu < x \le b$ and at most weakly singular as $\mu \to x$. Therefore, T_φ is compact as an operator from $L^\infty(0,b)$ into $C[0,b]$. Then, since $f_\varphi \in C[0,b]$, it follows that equation (14.7) (and also (14.6)) has a unique solution $z_\varphi \in C[0,b]$. Due to Lemmas 1 and 2, $z_\varphi \in C^{m,\nu_d}(0,b]$, $\nu_d = 1 - d(1-\nu)$.

14.3 Piecewise Polynomial Interpolation

For given $N \in \mathbb{N}$ and $r \ge 1$, let $\Pi_N^r = \{t_0, \ldots, t_N : 0 = t_0 < t_1 < \cdots < t_N = b\}$ be a partition (a grid) of the interval $[0,b]$ with the nodes

$$t_j = b(j/N)^r, \quad j = 0, ..., N. \tag{14.9}$$

Here the grading exponent $r \in [1,\infty)$ characterizes the nonuniformity of the grid Π_N^r: if $r > 1$, then the grid points (14.9) are more densely clustered near the left endpoint of the interval $[0,b]$. Further, let

$$S_{m-1}^{(k)}(\Pi_N^r) = \{u \in C^{(k)}[0,b] : u|_{[t_{j-1},t_j]} \in \pi_{m-1}, j = 1, ..., N\},\ k = 0,\ k = 1,$$
$$S_{m-1}^{(-1)}(\Pi_N^r) = \{u : u|_{[t_{j-1},t_j]} \in \pi_{m-1}, j = 1, ..., N\}$$

be the underlying spline spaces of piecewise polynomial functions on the grid Π_N^r. Here π_{m-1} denotes the set of polynomials of degree not exceeding $m-1$ and $u|_{[t_{j-1},t_j]}$ is the restriction of u to the subinterval $[t_{j-1},t_j]$, $j=1,...,N$.

In every subinterval $[t_{j-1},t_j]\subset[0,b]$, we introduce m interpolation points $t_{j1},\dots,t_{jm}$ as follows:

$$t_{jk}=t_{j-1}+\eta_k(t_j-t_{j-1}),\quad k=1,\dots,m,\ j=1,\dots,N, \tag{14.10}$$

where the parameters $\eta_1,\dots,\eta_m$ do not depend on j and N and satisfy

$$0\le\eta_1<\dots<\eta_m\le 1. \tag{14.11}$$

To a given continuous function $z:\ [0,b]\to\mathbb{R}$ we assign a piecewise polynomial interpolation function $P_Nz=P_N^{(m-1)}z\in S_{m-1}^{(-1)}(\Pi_N^r)$ such that $(P_Nz)(t_{jk})=z(t_{jk}),\quad k=1,\dots,m;\ j=1,\dots,N$. We also introduce an interpolation operator $P_N=P_N^{(m-1)}$ which assigns to every continuous function $z:\ [0,b]\to\mathbb{R}$ its piecewise polynomial interpolation function P_Nz.

In what follows, for given Banach spaces E and F we denote by $\mathcal{L}(E,F)$ the Banach space of linear bounded operators A: $E\to F$ with the norm $||A||=\sup\{||Az||:\ z\in E,\ ||z||\le 1\}$.

It follows from [Va93] that the norms of $P_N\in\mathcal{L}(C[0,b],L^\infty(0,b))$ are bounded by a constant c which is independent of N,

$$||P_N||_{\mathcal{L}(C[0,b],L^\infty(0,b))}\le c,\quad N\in\mathbb{N}, \tag{14.12}$$

and

$$||z-P_N\,z||_{L^\infty(0,b)}\to 0\quad\text{as}\quad N\to\infty, \tag{14.13}$$

for every $z\in C[0,b]$. Moreover, if $z\in C^{m,\nu}(0,b]$, $m\in\mathbb{N}$, $-\infty<\nu<1$, then

$$\sup_{x\in[0,b]}\left|z(x)-(P_Nz)(x)\right|\le c\,\varepsilon_N^{(m,\nu,r)}, \tag{14.14}$$

where

$$\varepsilon_N^{(m,\nu,r)}=\begin{cases} N^{-m} & \text{for}\ \ m<1-\nu,\ r\ge 1\,,\\ N^{-m}(1+\log N) & \text{for}\ \ m=1-\nu,\ r=1\,,\\ N^{-m} & \text{for}\ \ m=1-\nu,\ r>1\,,\\ N^{-r(1-\nu)} & \text{for}\ \ m>1-\nu,\ 1\le r<m/(1-\nu)\,,\\ N^{-m} & \text{for}\ \ m>1-\nu,\ r\ge m/(1-\nu). \end{cases} \tag{14.15}$$

14.4 Numerical Methods and Convergence Analysis

The approach proposed here for the numerical solution of problem (14.1),(14.2) can be described in the following three steps.

Step 1. We choose a function φ of the form (14.5) and, introducing in (14.4) the variables transformation $t = \phi(x)$, we obtain the new integral equation (14.6).

Step 2. We find an approximation v_N to z_φ, the solution of equation (14.6), determining $v_N = v_{N,m,r,\varphi} \in S_{m-1}^{(-1)}(\Pi_N^r)$ by the standard collocation method from the conditions

$$v_N(t_{jk}) = \int_0^{t_{jk}} K_{0,\varphi}(t_{jk},\mu)\Big(\int_0^{\mu} v_N(\sigma)\varphi'(\sigma)d\sigma\Big)d\mu + \int_0^{t_{jk}} \Big[p_\varphi(t_{jk})\varphi'(\mu) + K_{1,\varphi}(t_{jk},\mu)\Big]v_N(\mu)d\mu + f_\varphi(t_{jk}), \quad (14.16)$$

for $k = 1,\dots,m;\ j = 1,\dots,N$, with the points $\{t_{jk}\}$ defined by (14.10).

Step 3. We determine an approximation $u_N = u_{N,m,r,\varphi}$ to y, the solution of the Cauchy problem (14.1), (14.2), setting

$$u_N(t) = y_0 + \int_0^t v_N(\varphi^{-1}(s))ds, \quad 0 \le t \le b. \quad (14.17)$$

Remark 1. If we use the parameters $\eta_1 = 0$, $\eta_m = 1$ in (14.11), then the resulting collocation approximation v_N belongs to the smoother polynomial spline space $S_{m-1}^{(0)}(\Pi_N^r)$.

Remark 2. Conditions (14.16) lead to a system of linear equations whose exact form is specified by the choice of a basis in the space $S_{m-1}^{(-1)}(\Pi_N^r)$ (or in $S_{m-1}^{(0)}(\Pi_N^r)$ if $\eta_1 = 0$, $\eta_m = 1$).

Remark 3. If in (14.5) $d = 1$, then the method described above coincides with the standard collocation method studied in [Pe04] (see also [BrTa89]).

Theorem 1. *Assume that $p, q \in C^{m,\nu}(0,b]$, $K_0, K_1 \in \mathcal{W}^{m,\nu}(\Delta_b)$, $m \in \mathbb{N}$, $-\infty < \nu < 1$. Let φ be the transformation given by* (14.5)*. Finally assume that the collocation points* (14.10) *associated with the grid points (14.9) of the partition Π_N^r are used. Then, for all sufficiently large $N \in \mathbb{N}$, say $N \ge N_0$, the method* (14.17), (14.16) *determines unique approximations u_N and v_N to the solution y of the Cauchy problem* (14.1), (14.2) *and its derivative y', respectively. Moreover, for $N \ge N_0$ the following error estimates hold:*

$$\max_{0\le t\le b} |u_N(t) - y(t)| \le c\,\varepsilon_N^{(m,\nu_d,r)}, \quad (14.18)$$

$$\sup_{0\le t\le b} |v_N(\varphi^{-1}(t)) - y'(t)| \le c\,\varepsilon_N^{(m,\nu_d,r)}. \quad (14.19)$$

Here $\nu_d = 1 - d(1-\nu)$, $\varepsilon_N^{(m,\nu_d,r)}$ is defined by (14.15) *and c is a positive constant not depending on N.*

Proof. We consider (14.7) as an operator equation in $L^\infty[0,b)$. We already know (see Section 14.2) that equation (14.7) is uniquely solvable in $L^\infty[0,b)$ and its solution $z_\varphi \in C^{m,\nu_d}(0,b]$. Furthermore, conditions (14.16) allow the following operator equation representation:

$$v_N - P_N T_\varphi v_N = P_N f_\varphi, \tag{14.20}$$

with T_φ given by (14.8), and P_N introduced in Section 14.3. From (14.12) and (14.13) we obtain that $\|T - P_N T\|_{\mathcal{L}(L^\infty(0,b),L^\infty(0,b))} \to 0$ as $N \to \infty$. This together with the boundedness of $(I - T_\varphi)^{-1}$ in $L^\infty(0,b)$ yields that $I - P_N T_\varphi$ is invertible in $L^\infty(0,b)$ for all sufficiently large N, say $N \geq N_0$. Furthermore, it follows that the norms of $(I - P_N T_\varphi)^{-1}$ are uniformly bounded in N,

$$\|(I - P_N T_\varphi)^{-1}\|_{\mathcal{L}(L^\infty(0,b),L^\infty(0,b))} \leq c, \quad N \geq N_0, \tag{14.21}$$

for some constant c which is independent of N. Thus, equation (14.20) has a unique solution $v_N \in S_{m-1}^{(-1)}(\Pi_N^r)$ for $N \geq N_0$.

We have $v_N - z_\varphi = (I - P_N T_\varphi)^{-1}(P_N z_\varphi - z_\varphi)$, $N \geq N_0$, where z_φ is the solution of equation (14.7). Therefore, using (14.21) we obtain

$$\|v_N - z_\varphi\|_{L^\infty(0,b)} \leq c\|P_N z_\varphi - z_\varphi\|_{L^\infty(0,b)}. \tag{14.22}$$

Further, we have

$$\|v_N - z_\varphi\|_{L^\infty(0,b)} = \sup_{x\in[0,b]} |v_N(x) - z_\varphi(x)| = \sup_{t\in[0,b]} |v_N(\varphi^{-1}(t)) - y'(t)|,$$

where $z_\varphi \in C^{m,\nu_d}(0,b]$, $\nu_d = 1 - d(1-\nu)$. This together with (14.14) and (14.22) yields the estimate (14.19). Since

$$|u_N(t) - y(t)| \leq \int_0^t |v_N(\varphi^{-1}(s)) - y'(s)|ds, \quad 0 \leq t \leq b,$$

the estimate (14.18) is a consequence of (14.19).

Remark 4. According to Theorem 1, in the case $m > 1 - \nu_d = d(1-\nu)$, the estimate $\max_{0\leq t\leq b} |u_N(t) - y(t)| \leq cN^{-m}$ holds for $r \geq m/d(1-\nu)$. If ν is close to 1, this condition on r may be too restrictive. However, if $K_1 \in \mathcal{W}^{m,\nu-1}(\Delta_b)$, then the condition on r can be relaxed, as shown in the following theorem.

Theorem 2. *Let the conditions of Theorem 1 be fulfilled and let $K_1 \in \mathcal{W}^{m,\nu-1}(\Delta_b)$. Then, with the notation of Theorem 1, we have the following estimates, for $N \geq N_0$:*

1) if $1 \leq m < 2 - \nu_d = 1 + d(1-\nu)$, then

$$\max_{0\leq t\leq b} |u_N(t) - y(t)| \leq cN^{-m} \text{ for r} \geq 1,$$

2) if $m = 2 - \nu_d$, *then*

$$\max_{0\le t\le b} |u_N(t) - y(t)| \le c \begin{cases} N^{-m}(1+\log N) & \text{for } r = 1, \\ N^{-m} & \text{for } r > 1, \end{cases}$$

3) if $m > 2 - \nu_d$, *then*

$$\max_{0\le t\le b} |u_N(t) - y(t)| \le c \begin{cases} N^{-r(2-\nu_d)} & \text{for } 1 \le r < m/(2-\nu_d), \\ N^{-m}(1+\log N) & \text{for } \quad r = m/(2-\nu_d), \\ N^{-m} & \text{for } \quad r > m/(2-\nu_d). \end{cases}$$

The proof of Theorem 2 will be given in a forthcoming paper where the superconvergence properties of the method proposed in Section 14.4 will also be discussed.

14.5 Numerical Results

We considered the following initial value problem:

$$y'(t) = y(t) + q(t) + \int_0^t (t-s)^{-\nu} y(s)ds + \int_0^t (t-s)^{-\nu+1} y'(s)ds, \quad 0 \le t \le b, \tag{14.23}$$

with $0 < \nu < 1$ and the initial condition

$$y(0) = 0. \tag{14.24}$$

The forcing function q has been selected so that $y(t) = t^{2-\nu}$ is the exact solution to (14.23), (14.24). We note that this is a problem of the form (14.1), (14.2), with $y_0 = 0$, $p(t) \equiv 1$, $K_0(t,s) = (t-s)^{-\nu}$, $K_1(t,s) = (t-s)^{-\nu+1}$, and

$$q(t) = (2-\nu)t^{1-\nu} + t^{2-\nu}\int_0^1 (1-x)^{-\nu} x^{2-\nu} dx + (2-\nu)t^{3-2\nu}\int_0^1 (1-x)^{-\nu+1} x^{1-\nu} dx.$$

In this case it is easy to check that $K_0 \in \mathcal{W}^{m,\nu}(\Delta_b), K_1 \in \mathcal{W}^{m,\nu-1}(\Delta_b)$, $p, q \in C^{m,\nu}(0,b]$ for arbitrary $m \in \mathbb{N}$.

The problem (14.23), (14.24) was solved numerically by the method described in Section 14.4 for $b = 1/2$, $m = 2$, $\nu = 1/2$, $\eta_1 = 1/4$, $\eta_2 = 3/4$, and $r = 1$. In Table 14.1, some results for different values of the parameters N and d are displayed. The quantities $\delta_{N,d}$ and $\delta'_{N,d}$ are approximate values of the errors $\max\{|u_N(t) - y(t)| : 0 \le t \le b\}$ and $\sup\{|v_N(\varphi^{-1}(t)) - y'(t)| : 0 \le t \le b\}$, respectively. They have been calculated as follows:

$$\delta_{N,d} = \max\{|u_N(\tau_{jl}) - y(\tau_{jl})| : l = 1, ..., 10;\ j = 1, ..., N\},$$
$$\delta'_{N,d} = \max\{|v_N(\tau_{jl}) - y'(\tau^d_{jl})| : l = 1, ..., 10;\ j = 1, ..., N\},$$

where $\tau_{jl} = t_{j-1} + l(t_j - t_{j-1})/10$, $l = 1, ..., 10$, $j = 1, ..., N$, with $\{t_j\}$ defined by (14.9) for $b = 1/2$. The ratios

$$\rho_{N,d} = \delta_{N/2,d}/\delta_{N,d}\,, \quad \rho'_{N,d} = \delta'_{N/2,d}/\delta'_{N,d},$$

characterizing the observed convergence rate, are also presented.

Table 14.1. Numerical results for the problem (14.23), (14.24).

N	$\delta_{N,1}$ $\rho_{N,1}$	$\delta_{N,3}$ $\rho_{N,3}$	$\delta_{N,5}$ $\rho_{N,5}$	$\delta'_{N,1}$ $\rho'_{N,1}$	$\delta'_{N,3}$ $\rho'_{N,3}$	$\delta'_{N,5}$ $\rho'_{N,5}$
10	1.06×10^{-2} 2.829	5.72×10^{-6} 5.525	2.97×10^{-6} 4.332	2.44×10^{-2} 1.386	1.45×10^{-3} 2.716	9.15×10^{-4} 3.938
20	3.72×10^{-3} 2.829	1.42×10^{-6} 4.031	7.27×10^{-7} 4.086	1.74×10^{-2} 1.401	5.19×10^{-4} 2.789	2.30×10^{-4} 3.971
40	1.32×10^{-3} 2.829	3.52×10^{-7} 4.030	1.81×10^{-7} 4.022	1.24×10^{-2} 1.408	1.84×10^{-4} 2.815	5.78×10^{-5} 3.986
80	4.68×10^{-4} 2.828	8.77×10^{-8} 4.030	4.51×10^{-8} 4.013	8.75×10^{-3} 1.417	6.53×10^{-5} 2.818	1.45×10^{-5} 3.986
	2.828	4	4	1.414	2.828	4

If $m = 2$ and $\nu = 1/2$, then, for sufficiently large N, we obtain from Theorem 1 that

$$\delta'_{N,d} \approx \sup_{0 \le t \le b} |v_N(\varphi^{-1}(t)) - y'(t)| \le c \begin{cases} N^{-d/2} & \text{for } 1 \le d < 4, \\ N^{-2}(1 + \log N) & \text{for } d = 4, \\ N^{-2} & \text{for } d > 4. \end{cases}$$

Thus, the ratio $\rho'_{N,d}$ ought to be approximately $(N/2)^{-d/2}/N^{-d/2} = 2^{d/2}$ for $1 \le d < 4$ and 4 for $d > 4$. In a similar way, by applying Theorem 2 we would expect the ratio $\rho_{N,d}$ to be approximately $2^{3/2}$ for $d = 1$ and 4 for $d \ge 3$. In particular, the values of $\rho_{N,1}$, $\rho_{N,3}$, $\rho_{N,5}$, $\rho'_{N,1}$, $\rho'_{N,3}$, $\rho'_{N,5}$ ought to be approximately 2.828, 4.000, 4.000, 1.414, 2.828, 4.000. These values are given in the last row of the Table 14.1 for the case $m = 2$, $\nu = 1/2$, $\eta_1 = 1/4$, and $\eta_2 = 3/4$.

As we can see, the numerical results displayed in Table 14.1 are in good agreement with the corresponding theoretical estimates given in Theorems 1 and 2.

Acknowledgement. The work of M. Kolk and A. Pedas was partially supported by the Estonian Science Foundation (Research Grant No. 7353).

References

[BaOr06] Baratella, P., Orsi, A.P.: Numerical solution of weakly singular linear Volterra integro-differential equations. *Computing,* **77**, 77–96 (2006).

[Br04] Brunner, H.: *Collocation Methods for Volterra Integral and Related Functional Equations,* Cambridge University Press, London (2004).

[BrHo86] Brunner, H., van der Houwen, P.J.: *The Numerical Solution of Volterra Equations,* North-Holland, Amsterdam (1986).

[BrPeVa01a] Brunner, H., Pedas, A., Vainikko, G.: Piecewise polynomial collocation methods for linear Volterra integro-differential equations with weakly singular kernels. *SIAM J. Numer. Anal.,* **39**, 957–982 (2001).

[BrPeVa01b] Brunner, H., Pedas, A., Vainikko, G.: A spline collocation method for linear Volterra integro-differential equations with weakly singular kernels. *BIT Numer. Math.,* **41**, 891–900 (2001).

[BrTa89] Brunner, H., Tang, T.: Polynomial spline collocation methods for the nonlinear Basset equation. *Comput. Math. Appl.,* **18**, 449–457 (1989).

[CaEtAl07] Cao, Y., Huang, M., Liu, L., Xu, Y.: Hybrid collocation methods for Fredholm integral equations with weakly singular kernels. *Appl. Numer. Math.,* **57**, 549–561 (2007).

[KaPa03] Kangro, R., Parts, I.: Superconvergence in the maximum norm of a class of piecewise polynomial collocation methods for solving linear weakly singular Volterra integro-differential equations. *J. Integral Equations Appl.,* **15**, 403–427 (2003).

[McSt83] McKee, S., Stokes, A.: Product integration methods for the nonlinear Basset equation. *SIAM J. Numer. Anal.,* **20**, 143–160 (1983).

[Pa05] Parts, I.: Optimality of theoretical estimates for spline collocation methods for linear weakly singular Volterra integro-differential equations. *Proc. Estonian Acad. Sci. Phys. Math.,* **54**, 162–180 (2005).

[PaPe03] Parts, I., Pedas, A.: Collocation approximations for weakly singular Volterra integro-differential equations. *Math. Model. Anal.,* **8**, 315–328 (2003).

[Pe04] Pedas, A.: Piecewise polynomial approximations for linear Volterra integro-differential equations with nonsmooth kernels, in: *Numerical Mathematics and Advanced Applications,* Feistauer, M. et al., eds., Springer, Berlin-Heidelberg, (2004), 677–686.

[PeTa06] Pedas, A., Tamme, E.: Spline collocation method for integro-differential equations with weakly singular kernels. *J. Comput. Appl. Math.,* **197**, 253–269 (2006).

[PeTa08] Pedas, A., Tamme, E.: Discrete Galerkin method for Fredholm integro-differential equations with weakly singular kernels. *J. Comput. Appl. Math.,* **213**, 111–126 (2008).

[PeVa04] Pedas, A., Vainikko, G.: Smoothing transformation and piecewise polynomial collocation for weakly singular Volterra integral equations. *Computing,* **73**, 271–293 (2004).

[Ta92] Tang, T.: Superconvergence of numerical solutions to weakly singular Volterra integro-differential equations. *Numer. Math.,* **61**, 373–382 (1992).

[Ta93] Tang, T.: A note on collocation methods for Volterra integro-differential equations with weakly singular kernels. *IMA J. Numer. Anal.,* **13**, 93–99 (1993).

[Va93] Vainikko, G.: *Multidimensional Weakly Singular Integral Equations,* Springer, Berlin-Heidelberg-New York (1993).

15

Numerical Solution of a Class of Integral Equations Arising in a Biological Laboratory Procedure

D.A. French[1] and C.W. Groetsch[2]

[1] University of Cincinnati, OH, USA; `french@math.uc.edu`
[2] The Citadel, Charleston, SC, USA; `charles.groetsch@citadel.edu`

15.1 Introduction

We discuss a numerical method for certain integral equations of the form

$$I(t) = \int_0^L k(x,t)\rho(x)dx, \quad 0 \le t \le T, \tag{15.1}$$

where I is a smooth increasing function with $I(0) = 0$, L and T are positive constants, and the kernel $k(\cdot,\cdot)$ satisfies the following assumptions:

(i) $k(\cdot,\cdot)$ is continuous and bounded on $[0,L] \times [0,T] - \{(0,0)\}$, positive on $(0,L] \times (0,T]$, and $k(\cdot,\cdot) \in C^1((0,L) \times (0,T))$.
(ii) $k(x,0) = 0$ for $x > 0$ and there is a positive constant κ with $k(0,t) \ge \kappa$ for $t \in (0,T]$.
(iii) $\partial_1 k < 0 < \partial_2 k$ on $(0,L) \times (0,T)$ (we use the notation ∂_j to indicate the partial derivative with respect to the jth variable).

Our study of this class of Fredholm integral equations of the first kind is motivated by a mathematical model of an aspect of the olfactory system of frogs (see [FG06]). The function of the olfactory system is to transduce an odor stimulus into an electrical signal that is fed to the nervous system. This transduction is accomplished by a cascade of chemical processes that leads to an influx of ions through channels in very thin hair-like features, known as cilia, that reside in the nasal mucus. The potential difference across the membrane forming the lateral surface of the cilium resulting from this ion migration produces the electrical signal.

A morphological feature of interest is the distribution of ion channels along the length of the cilium. In an experimental procedure developed by S.J. Kleene, a single cilium is drawn into a recording pipette containing a solution of sodium ions and the cilium is detached at its base. The pipette is then emersed in a bath of a channel activating ligand (cAMP: cyclic adenosine

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_15,

monophosphate), allowing the agent to enter the cilium at its open base. The agent then diffuses along the cilium from its base to its closed end, opening ion channels as it goes. Sodium ions enter the interior of the cilium through the opened channels, inducing a potential difference between the exterior and interior of the cilium. The resulting current signal $I(t)$ is measured by the pipette and recorded. The integral equation model that motivates this work is intended to deduce the spatial distribution of ion channels along the length of the cilium from this electrical signal.

The concentration $c(x,t)$ of the activating agent satisfies a diffusion equation and the recorded current $I(t)$ is given by (15.1), where L is the length of the cilium, ρ is the density of ion channels along the length of the cilium ($x = 0$ corresponds to the base), and the kernel $k(\cdot,\cdot)$ is given by a Hill's function,

$$k(x,t) = J_0 \frac{c(x,t)^n}{c(x,t)^n + K_{1/2}^n}, \tag{15.2}$$

where n and J_0 are positive constants and $K_{1/2}$ is half the bulk concentration of cAMP in the bath (see [FG06] for details). Note that conditions (i)–(iii) are satisfied for kernels of the form (15.2), where c satisfies the diffusion equation with initial condition equal to the bulk concentration K of cAMP in the bath. See Figure 15.1.

It is well known that equations of the form (15.1) are ill posed, necessitating special care in their numerical solution (see, e.g., [G84]). As an indication of this difficulty, we illustrate a naive numerical method for (15.1). Here, we use uniform partitions of space and time, that is,

$$0 < y_1 < \cdots < y_N = L, \quad y_j = j\Delta y, \quad \Delta y = \frac{L}{N},$$

$$0 < t_1 < \cdots < t_N = T, \quad t_j = j\Delta t, \quad \Delta t = \frac{T}{N},$$

and define

$$A_{ij} = J_0 \int_{y_{j-1}}^{y_j} k(y,t_i)\, dy \quad \text{and} \quad F_i = I(t_i).$$

We are interested in finding a piecewise constant approximation of ρ, that is,

$$\rho^S(y) = \rho_j^S \quad \text{for} \quad y \in [y_{j-1}, y_j],$$

which satisfies $A\vec{\rho}^S \cong \vec{F}$. If this straightforward approach is used, the matrix A tends to be ill conditioned, as shown in Table 15.1. These condition numbers are an indication of the ill-posed nature of the problem (15.1), a phenomenon that has been well studied for decades.

15.2 Tail Clipping

Figure 15.1 depicts typical behavior of kernels of this type; the "fronts" $k(\cdot,t)$ move from left to right as t increases. It is evident that for a given t, the

Table 15.1. Condition numbers for the naive scheme.

N	10	15	20
$\mathrm{Cond}_2(A)$	5.4×10^5	1.3×10^{10}	2.3×10^{13}

relatively flat "tail" of the front $k(\cdot, t)$ contributes little information on ρ in the tail region. The method discussed below is a kind of marching scheme that attempts to mitigate this lack of information by systematically truncating the tail.

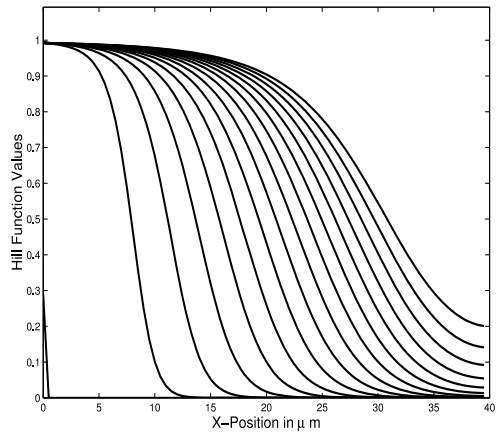


Fig. 15.1. Propagation of the kernel as the ligand diffuses into a cilium.

The proposed numerical method begins with a small positive parameter ϵ which plays a role akin to a regularization parameter. As will be seen below, the function of this parameter is to remove the flat tail of the kernel.

Lemma 1. *Suppose* $0 < \epsilon < k(L, T)$. *Then there is a unique* $T(\epsilon) \in (0, T)$ *satisfying*

$$k(L, T(\epsilon)) = \epsilon.$$

$T(\cdot)$ *is an increasing function,* $T(\epsilon) \to 0$ *as* $\epsilon \to 0^+$ *and* $T(\epsilon) \to T$ *as* $\epsilon \to k(L, T)^-$.

Proof. Since $k(L, 0) = 0$ and $k(L, \cdot)$ is strictly increasing, the existence of a unique such $T(\epsilon)$ is ensured. If $0 < \epsilon_1 < \epsilon_2 < k(L, T)$, then by (iii)

$$0 < \epsilon_2 - \epsilon_1 = k(L, T(\epsilon_2)) - k(L, T(\epsilon_1)) = \partial_2 k(L, \theta)(T(\epsilon_2) - T(\epsilon_1))$$

for some $\theta \in (0,T)$ and hence, by (iii), $T(\epsilon_1) < T(\epsilon_2)$. If for some sequence $\epsilon_n \to 0$, $T(\epsilon_n)$ clusters at $\alpha > 0$, then by continuity, $k(L,\alpha) = 0$, contradicting (i). On the other hand, if $\epsilon_n \uparrow k(L,T)$ and $\{T(\epsilon_n)\}$ clusters at $t < T$, then $k(L,t) = k(L,T)$, which contradicts (iii).

Remark 1. Note that for the front-tracking algorithm $T(\epsilon)$ will become the final value in our time partition.

Lemma 2. *Suppose $0 < \epsilon < min\{\kappa, k(L,T)\}$. For $0 < t \leq T(\epsilon)$ there is a unique $x_\epsilon(t)$ satisfying $k(x_\epsilon(t),t) = \epsilon$. Also,*

(a) for fixed $\epsilon > 0$, $x_\epsilon(\cdot)$ is an increasing function and $x_\epsilon(t) \to 0$ as $t \to 0^+$; $x_\epsilon(t) \to L$ as $t \to T(\epsilon)^-$, and

(b) for each $t \in (0,T)$, $x_\epsilon(t)$ is a decreasing function of ϵ and $x_\epsilon(t) \to L$ as $\epsilon \to 0^+$.

Proof. If $t < T(\epsilon)$, then by (iii), $k(L,t) < k(L,T(\epsilon)) = \epsilon$. Also, by (i), $k(0,t) > \epsilon$. As $k(\cdot,t)$ is continuous and strictly decreasing, a unique such $x_\epsilon(t)$ exists.
(a) If $t_1 < t_2$, then for some positive θ and ψ,

$$\begin{aligned} 0 = k(x_\epsilon(t_1),t_1) - k(x_\epsilon(t_2),t_2) \\ = \partial_1 k(\theta,t_1)(x_\epsilon(t_1) - x_\epsilon(t_2)) + \partial_2 k(x_\epsilon(t_2),\psi)(t_1 - t_2), \end{aligned}$$

and it follows from (iii) that $x_\epsilon(t_1) < x_\epsilon(t_2)$. Suppose $t_n \downarrow 0$ and $x_\epsilon(t_n) \downarrow x^* > 0$. Then, by continuity, $k(x^*,0) = \epsilon > 0$, contradicting (ii), and hence $x_\epsilon(t) \to 0$ as $t \to 0$.

Let $t_n \uparrow T(\epsilon)$. Then, since $x_\epsilon(t_n)$ is increasing and bounded, $x_\epsilon(t_n) \uparrow x^*$, say. By continuity, $k(x^*,T(\epsilon)) = \epsilon = K(L,T(\epsilon))$. But $k(\cdot,T(\epsilon))$ is one-to-one, and hence $x^* = L$.
(b) Let $\epsilon_1 < \epsilon_2$, then $k(x_{\epsilon_1}(t),t) < k(x_{\epsilon_2}(t),t)$ and hence for some positive θ,

$$\partial_1 k(\theta,t)(x_{\epsilon_1}(t) - x_{\epsilon_2}(t)) < 0$$

and hence, by (iii), $x_{\epsilon_1}(t) > x_{\epsilon_2}(t)$.

Suppose $x_{\epsilon_n} \to x_0 < L$ for some sequence $\epsilon_n \downarrow 0$. Then, by continuity, $k(x_0,t) = 0$. But then, $k(L,t) < k(x_0,t) = 0$ since $x_0 < L$, which contradicts (i).

Remark 2. Note that by the implicit function theorem (see, e.g., [HS74]), $x_\epsilon(\cdot) \in C^1(0,T)$ and, by (iii),

$$x'_\epsilon(\cdot) = -\frac{\partial_2 k}{\partial_1 k}(x_\epsilon(\cdot),\cdot) > 0.$$

In particular, x_ϵ, extended by continuity to $[0,T]$, maps $[0,T]$ onto $[0,L]$ in a one-to-one manner, is continuously differentiable on $(0,T)$, and has an inverse which is continuously differentiable on $(0,L)$.

Our numerical method is based on clipping the flat tail of the kernel in the following way. For $0 < t < T(\epsilon)$, let

$$k_\epsilon(x,t) = \begin{cases} k(x,t), & 0 \le x \le x_\epsilon(t), \\ 0, & x_\epsilon(t) \le x \le L. \end{cases}$$

Now define $\rho_\epsilon(\cdot)$ by

$$I(t) = \int_0^L k_\epsilon(x,t)\rho_\epsilon(x)dx =: (K_\epsilon \rho_\epsilon)(t). \tag{15.3}$$

Remark 3. The existence of a unique solution of (15.3) for suitable I and $k(\cdot,\cdot)$ is provided by Proposition 1 below. Equation (15.3) is, on the face of it, a Fredholm equation of the first kind. However, if ρ_ϵ satisfies (15.3), then it is also the solution of a Volterra equation of the *second* kind—a typically well-posed problem. This gives credence to (15.3) being a type of *regularization* method (but not quite: see the remark following Proposition 2).

We note that (15.3) may be written as

$$I(t) = \int_0^{x_\epsilon(t)} k(x,t)\rho_\epsilon(x)dx, \quad 0 \le t \le T(\epsilon). \tag{15.4}$$

Remark 4. If (15.1) is assumed to have a solution $\rho \in L^2$ for a given $I \in L^2$, then in fact I inherits the smoothness of k giving $I \in H^1$. So we may as well assume that $I \in H^1$.

Proposition 1. *If $I \in H^1(0,T)$ and $I(0) = 0$, then ρ_ϵ is a solution of the Fredholm equation of the first kind (15.3) if and only if it is a solution of the Volterra equation of the second kind*

$$\rho_\epsilon(z) = f(z) + \frac{1}{\epsilon}\int_0^z \hat{k}(z,x)\rho_\epsilon(x)dx, \quad 0 < z < L, \tag{15.5}$$

where

$$f(z) = \frac{d}{dz}I(x_\epsilon^{-1}(z))/\epsilon, \quad \hat{k}(z,x) = -\frac{d}{dz}k(x, x_\epsilon^{-1}(z)).$$

Proof. The substitution $z = x_\epsilon(t)$ shows that (15.4) is equivalent to

$$I(x_\epsilon^{-1}(z)) = \int_0^z \tilde{k}(z,x)\rho_\epsilon(x)dx$$

where $\tilde{k}(z,x) = k(x, x_\epsilon^{-1}(z))$. But, since $I \in H^1$ and $I(x_\epsilon^{-1}(0)) = I(0) = 0$, this is equivalent to

$$\frac{d}{dz}I(x_\epsilon^{-1}(z)) = \int_0^z \frac{d\tilde{k}}{dz}(z,x)\rho_\epsilon(x)dx + \tilde{k}(z,z)\rho_\epsilon(z).$$

But

$$\tilde{k}(z,z) = k(z, x_\epsilon^{-1}(z)) = k(x_\epsilon(t), t) = \epsilon,$$

and hence the result.

Now, by standard L^2-theory ([S70], Chapter 2), the Volterra second-kind equation (15.5) has for each $\epsilon > 0$ a unique solution ρ_ϵ that depends (L^2) continuously on f. In particular (using $I = 0$), we have the following.

Corollary 1. $N(K_\epsilon) = \{0\}$.

Proposition 2. *The range of K_ϵ is a proper dense subspace of $L^2[0,T]$.*

Proof. First note that

$$(K_\epsilon^* g)(x) = \int_0^T k_\epsilon(x,t)g(t)dt = \int_{x_\epsilon^{-1}(x)}^T k(x,t)g(t)dt.$$

Hence, if $g \in N(K_\epsilon^*)$, then

$$\int_T^{x_\epsilon^{-1}(x)} k(x,t)g(t)dt = 0.$$

Setting $\tau = x_\epsilon^{-1}(x)$, we then have

$$\begin{aligned} 0 = \frac{d}{d\tau}\int_T^\tau k(x_\epsilon(\tau),t)g(t)dt = k(x_\epsilon(\tau),\tau)g(\tau) + \int_0^\tau \partial_1 k(x_\epsilon(\tau),t)x_\epsilon'(\tau)g(t)dt \\ = \epsilon g(\tau) + \int_0^\tau k^\dagger(\tau,t)g(t)dt, \end{aligned}$$

where

$$k^\dagger(\tau,t) = \partial_1 k(x_\epsilon(\tau),t)x_\epsilon'(\tau).$$

Hence, by the standard Volterra theory, $g = 0$. Therefore, $\{0\} = N(K_\epsilon^*) = R(K_\epsilon)^\perp$. It follows that $R(K_\epsilon)$ is dense; however, $R(K_\epsilon)$ does not exhaust $L^2[0,T]$ since K_ϵ is compact and non-degenerate (see, e.g., [G77]).

Remark 5. So, (15.4) is not a regularization method in the full sense that Tikhonov regularization is, since a solution is guaranteed to exist only for I in a dense subspace of L^2, but there are functions $I \in L^2$ for which (15.4) has no solution.

First, we give an estimate for the *residual*,

$$I - K\rho_\epsilon = K\rho - K\rho_\epsilon.$$

Proposition 3. *If $\rho_\epsilon \in L^2[0,L]$ and $\{\|\rho_\epsilon\|_2\}_{\epsilon>0}$ is bounded, then*

$$\left|\int_0^L k(x,t)(\rho(x) - \rho_\epsilon(x))dx\right| = O(\epsilon).$$

Proof. We have

$$0 = \int_0^L k(x,t)\rho(x)dx - \int_0^L k_\epsilon(x,t)\rho_\epsilon(x)dx$$
$$= \int_0^L k(x,t)(\rho(x) - \rho_\epsilon(x))dx + \int_{x_\epsilon(t)}^L k(x,t)\rho_\epsilon(x)dx,$$

and hence, since $k(x,t) \le \epsilon$ for $x \ge x_\epsilon(t)$,

$$\left| \int_0^L k(x,t)(\rho(x) - \rho_\epsilon(x))dx \right| \le \epsilon \int_{x_\epsilon(t)}^L |\rho_\epsilon(x)|dx$$
$$\le \epsilon \int_0^L |\rho_\epsilon(x)|dx \le \epsilon\sqrt{L}\|\rho_\epsilon\|_2.$$

We now give a weak convergence result assuming that (15.1) has a unique solution.

Proposition 4. *If $N(K) = \{0\}$ and $\{\|\rho_\epsilon\|_2\}_{\epsilon>0}$ is bounded, then $\rho_\epsilon \rightharpoonup \rho$ (weak convergence) as $\epsilon \to 0$.*

Proof. $R(K^*)$ is dense since $N(K) = \{0\}$. For $\varphi \in R(K^*)$, say $\varphi = K^*\psi$, we have (using $\langle\cdot,\cdot\rangle$ for the L^2 inner product):

$$\langle \rho_\epsilon, \varphi\rangle = \langle K\rho_\epsilon, \psi\rangle = \langle K\rho_\epsilon - K\rho, \psi\rangle + \langle K\rho, \psi\rangle = O(\epsilon) + \langle \rho, K^*\psi\rangle = O(\epsilon) + \langle \rho, \varphi\rangle$$

and hence $\langle \rho_\epsilon, \varphi\rangle \to \langle \rho, \varphi\rangle$. Since $\{\rho_\epsilon\}$ is bounded and the convergence takes place for φ in a dense set, it follows from the Banach–Steinhaus theorem that

$$\rho_\epsilon \rightharpoonup \rho.$$

15.3 A Numerical Method

Our numerical algorithm is based on piecewise constant spline approximation and collocation applied to equation (15.3). We define a sequence of wavefront points (computed using the bisection method)

$$x_j = x(t_j) \ \text{ where } \ k(x(t_j), t_j) = \epsilon.$$

The existence of the x_j's is guaranteed by Lemma 2. Recall that, at time t_j, we define x_j as the point where the "wave" $k(\cdot, t_j)$ drops within ϵ of zero. We let

$$\rho^{FT} = \sum_{j=1}^n \rho_j^{FT} \chi_{(x_{j-1}, x_j]}, \tag{15.6}$$

where χ_S denotes the indicator function of a set S. The collocation equations are then

$$I(t_i) = \int_0^L k_\epsilon(x, t_i)\rho^{FT}(x)dx, \quad i = 1, \ldots, n. \tag{15.7}$$

Proposition 5. *Equations (15.7) have a unique solution of the form (15.6).*

Proof. For $i = 1$, (15.7) reads

$$I(t_1) = \sum_{j=1}^{n} \rho_j^{FT} \int_0^L k_\epsilon(x, t_i)\chi_{(x_{j-1},x_j]}(x)dx = \rho_1^{FT} \int_0^{x_1} k(x, t_1)dx,$$

by the definition of $k_\epsilon(\cdot, \cdot)$; therefore,

$$\rho_1^{FT} = I(t_1)\left(\int_0^{x_1} k(x, t_1)dx\right)^{-1}.$$

We find inductively that $\{\rho_j^{FT}\}$ are uniquely determined. Thus, for $i > 1$,

$$\begin{aligned} I(t_i) &= \int_0^L k_\epsilon(x, t_i) \sum_{j=1}^{n} \rho_j^{FT} \chi_{(x_{j-1},x_j]}(x)dx \\ &= \sum_{j=1}^{i-1} \rho_j^{FT} \int_0^{x_i} k(x, t_i)\chi_{(x_{j-1},x_j]}(x)dx + \rho_i^{FT} \int_{x_{i-1}}^{x_i} k(x, t_i)dx, \end{aligned}$$

which uniquely determines ρ_i^{FT}:

$$\begin{aligned} \rho_i^{FT} &= \left(I(t_i) - \sum_{j=1}^{i-1} \rho_j^{FT} \int_0^{x_i} k(x, t_i)\chi_{(x_{j-1},x_j]}(x)dx\right)\left(\int_{x_{i-1}}^{x_i} k(x, t_i)dx\right)^{-1} \\ &= \left(I(t_i) - \sum_{j=1}^{i-1} \rho_j^{FT} \int_{x_{j-1}}^{x_j} k(x, t_i)dx\right)\left(\int_{x_{i-1}}^{x_i} k(x, t_i)dx\right)^{-1}. \end{aligned}$$

Hence, the $\{\rho_j^{FT}\}$ are uniquely determined by an explicit marching procedure.

It is natural to define a related matrix equation for this process. Letting

$$H_{ij} = \int_{x_{j-1}}^{x_j} k(x, t_i)\, dx \quad \text{for } j \le i \quad \text{and} \quad H_{ij} = 0 \quad \text{otherwise},$$

we observe that $H\vec{\rho}^{FT} \cong \vec{F}$.

For comparison, a Tikhonov regularization method is developed. Here, we use the uniform partitions of space and time in variables y and t defined earlier. We are interested in finding a piecewise constant function

$$\rho^T(y) = \rho_j^T \quad \text{for} \quad y \in [y_{j-1}, y_j],$$

which gives $A\vec{\rho}_T \cong \vec{F}$. The Tikhonov approximation ρ_T, with parameter β_T, then satisfies

$$(A^T A + \beta_T I)\vec{\rho}_T = A^T \vec{F}, \quad \text{where} \quad \beta_T = \beta\|A\|_\infty^2.$$

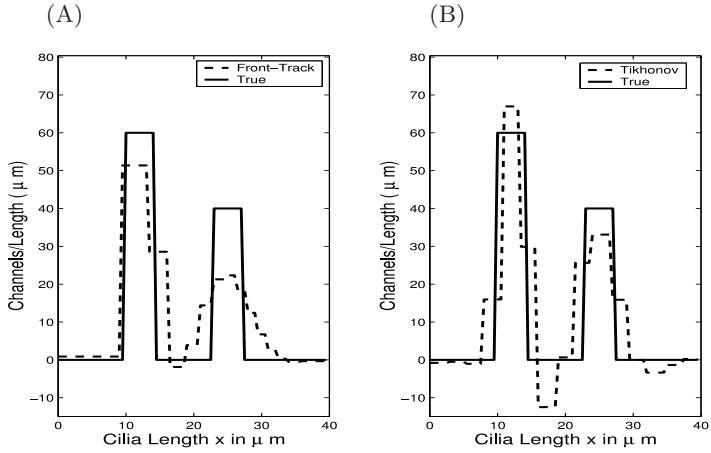


Fig. 15.2. Comparison of front-tracking (A) and Tikhonov (B) channel distribution solutions with $T = 0.475$ on synthetic data. In (A) we used $\epsilon = 0.2$ and in (B) we took $\beta = 1 \times 10^{-6}$.

We first treat a test problem. A "true" channel distribution $\rho_{True}(x)$ is created; the corresponding current is defined via the original continuous system of equations. The methods are then used to develop approximations to ρ_{True}. Figure 15.2 displays the resulting ρ's for our front-tracking method ρ^{FT} in (A) and the Tikhonov method ρ^{T} in (B). Approximation errors were computed by the formula

$$E_{Appx} = \frac{\sum_{i=1}^{N} |\rho(x_i) - \rho_i^{Appx}| \Delta x_i}{\sum_{i=1}^{N} |\rho(x_i)| \Delta x_i},$$

where "Appx" represents either the "FT" or "T" approximations.

Table 15.2. Errors (formula E_{Appx}) and condition numbers for front-tracking approximations of ρ with $N = 15$ at varying T values as defined by ϵ; see Figure 15.2.

ϵ	$\text{Cond}_2(H)$	FT Error	Time T
0.1	5.222×10^2	0.8313	0.418
0.2	2.580×10^2	0.7987	0.475
0.3	1.704×10^2	0.8691	0.522

We were interested in the effect that variations in ϵ and β had on the errors and condition number of our methods. We indeed find that this has an effect.

Table 15.3. Errors (formula E_{Appx}) and condition numbers for Tikhonov approximations of ρ with $N = 15$ at $T = 0.475$ at varying values for β; see Figure 15.2.

β	β_T	$\mathrm{Cond}_2(A^T A + \beta_T I)$	Tikhonov Error
1.0×10^{-5}	4.894×10^{-4}	7.265×10^4	1.022
1.0×10^{-6}	4.894×10^{-5}	7.265×10^5	0.995
1.0×10^{-7}	4.894×10^{-6}	7.265×10^6	1.013

Table 15.2 shows our results for the front-tracking approach and Table 15.3 shows them for the Tikhonov scheme (using the same final T as gives the best result for the front-tracking method).

Finally, Figures 15.3 and 15.4 display the results for the front-tracking and Tikhonov schemes in a sample case with experimental data. Front-tracking predicts 461 CNG channels and Tikhonov predicts 451 for the 70 μm cilium. In each we had $T = 2.099$.

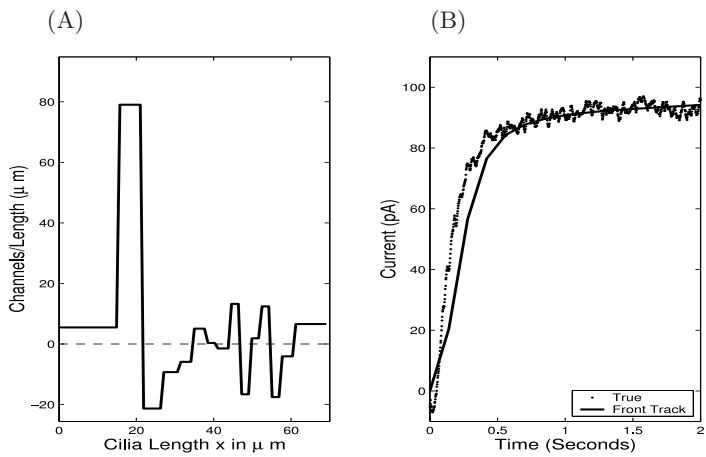


Fig. 15.3. Ion channel distribution and current from front-tracking method applied to experimental data with $\epsilon = 0.2$.

Acknowledgement. We thank our colleague Bill Krantz for suggesting this "front-tracking" algorithm as a possible numerical method for identifying ion channel distributions in frog olfactory systems. DAF was partially supported by an Interdisciplinary Grant in the Mathematical Sciences (IGMS) from the National Science Foundation (NSF DMS-0207145). The Mathematical Biosciences Institute (MBI) at

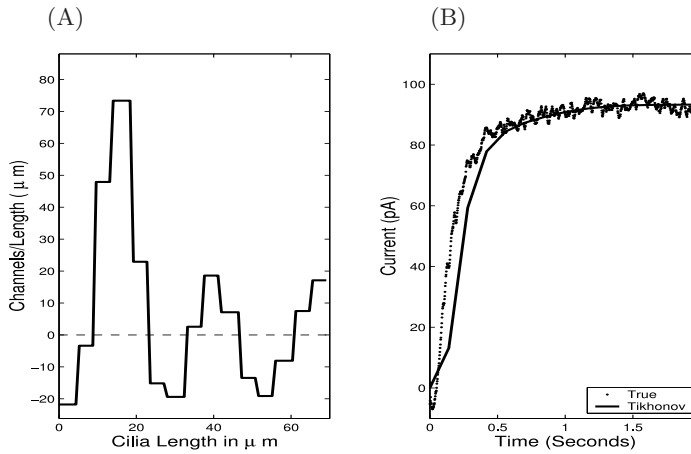


Fig. 15.4. Ion channel distribution and current from Tikhonov method applied to experimental data with $\beta = 1 \times 10^{-6}$.

Ohio State University also contributed in part to this work. His research, with Co-PI S.J. Kleene, is currently funded by a Mathematical-Biology grant from the National Science Foundation (NSF DMS-0515989). The work of CWG was supported by the Traubert Endowment at The Citadel.

References

[FG06] French, D.A., Flannery, R.J., Groetsch, C.W., Krantz, W.B., Kleene, S.J.: Numerical approximation of solutions of a nonlinear inverse problem arising in olfaction experimentation. *Math. Comput. Modelling,* **43**, 945–956 (2006).

[FG07] French, D.A., Groetsch, C.W.: Integral equation models for the inverse problem of biological ion channel distributions. *J. Phys. Conf. Ser.,* **73**, article 102006 (2007).

[G77] Groetsch, C.W.: *Generalized Inverses of Linear Operators: Representation and Approximation,* Marcel Dekker, New York (1977).

[G84] Groetsch, C.W.: *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind,* Pitman, London (1984).

[HS74] Hirsch, M.W., Smale, S.: *Differential Equations, Dynamical Systems, and Linear Algebra,* Academic Press, New York (1974).

[S70] Smithies, F.: *Integral Equations,* Cambridge University Press, London (1970).

16

A Mixed Two-Grid Method Applied to a Fredholm Equation of the Second Kind

L. Grammont

Université de Saint-Étienne, France; laurence.grammont@univ-st-etienne.fr

16.1 Introduction

The purpose of this chapter is to compute at a low cost an approximate solution of a Fredholm integral equation at a given accuracy. Vainikko proposed to compute the Nyström approximation of order n with quadrature two-grid iterations. We propose here to compute it with a two-grid method based on a projection method of a new type developed by Kulkarni. We will theoretically compare the absolute errors and the complexities of these two approximations.

Let T be the integral operator

$$Tu(t) = \int_0^1 k(t,x)u(x)dx,$$

where $k \in C^{m'}([0,1] \times [0,1])$.

Consider the integral equation of the second kind

$$u - Tu = f, \tag{16.1}$$

where $f \in C^m[0,1]$. We will assume that this equation has a unique solution.

We deal with the problem of looking for an approximate solution of the integral equation (16.1).

It has been stated in [We03] that, using $O(n)$ evaluations of f and k, the optimal accuracy which an algorithm can achieve is of the form:

$$\|u - u_n\|_0 \leq cn^{-m''}\|f\|_m, \quad m'' = \min(m, m'/2), \tag{16.2}$$

where c is a constant independent of n and f. An algorithm which produces a solution with an accuracy $O(\left(\frac{1}{n}\right)^{m''})$ using $O(n)$ evaluations of f and k is said to be optimal. As Werschulz wrote in [We03], one should use the term quasi-optimal because we ignore the constant multiplicative factor c. We will focus on the contribution of the multiplicative factor in the error estimation.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_16,

For the cases in which it is big, we cannot ignore it and the choice of the method shall take it into account even if the method is "quasi"-optimal.

The Nyström approximation provides a solution u_n whose error estimate is $\|u-u_n\|_0 = O(n^{-m})$, where u is the exact solution of (16.1) (see Theorem 2.1, p. 98 in [Va05]). This algorithm creates a linear system of size n which can be solved in $O(n^3)$ flops.

In [Va05], Vainikko shows how the amount of work can be reduced to $O(n^2)$ using two-grid iterations. Also, his third step consists in reducing the computational cost to $O(n)$ using a cheaper approximation of the kernel which maintains the accuracy (16.2) (Theorem 2.2, p. 99). In this chapter, we propose to perform a two-grid projection method proposed by Kulkarni in [Ku04] instead of the quadrature two-grid method. The third step is the same as in [Va05]. The accuracy of both methods is $\left(\frac{1}{n}\right)^m$ but the multiplicative factors are not the same. Their comparison is the purpose of this chapter.

We adopt the following notation:

$D^{i,j}k = \dfrac{\partial^{i+j}k}{\partial^i t \partial^j x}$.

$L^\infty = L^\infty[0,1]$: the space of all equivalence classes of essentially bounded Lebesgue measurable functions on the interval $[0,1]$, equipped with the norm $\|x\|_0 = \text{esssup}\{|x(t)| : t \in [0,1]\}$.

$C = C^0[0,1]$: the space of real-valued continuous functions defined on $[0,1]$, with the norm $\|x\|_0 = \sup\{|x(t)| : t \in [0,1]\}$.

$C^m = C^m[0,1]$: the space of all the real-valued functions defined on $[0,1]$, whose first m derivatives are continuous on $[0,1]$.

$x^{(i)}$: the ith derivative of x.

$\|x\|_m = \sum\limits_{i=0}^{m} \|x^{(i)}\|_0$: the norm on C^m.

Let $\mathcal{L}(X,Y)$ be the set of all bounded linear operators from the normed space $(X,\|.\|_X)$ into the normed space $(Y,\|.\|_Y)$, and let $\mathcal{L}(X)$ denote $\mathcal{L}(X,X)$. For $K \in \mathcal{L}(X,Y)$, we write $\|K\|_{\mathcal{L}(X,Y)} = \sup\limits_{x\in X} \dfrac{\|Kx\|_Y}{\|x\|_X}$.

$\|k\|_{p,q} := \sum_{i=0}^{p}\sum_{j=0}^{q} \|D^{i,j}k\|_0$.

The Nyström approximation relies on a quadrature formula, constructed as follows.

Let $(x_{j,n})_{j=1,\dots,n}$ be a partition of $[0,1]$. We assume that $\sup\{x_{j+1,n} - x_{j,n},\ j = 0,\dots,q_n-1\} \le n^{-1}$. Then

$$\int_0^1 v(y)dy = \sum_{j=1}^{q_n} w_{j,n} v(x_{j,n}) + \varphi_n(v),$$

where

$$|\varphi_n(v)| \le c_q \left(\frac{1}{n}\right)^m \|v\|_m \quad \text{for } v \in C^m[0,1], \quad \sum_{j=1}^{q_n} |w_{j,n}| \le a_q.$$

16.2 The Nyström Two-Grid Method

In Section 3 of [Va05], Vainikko recalls with short proofs some convergence results for the two-grid iteration method applied to the quadrature system. But he does not express the multiplicative factors in his estimations. That is the aim of this section: we resume the results of [Va05] incorporating the explicit expression of the multiplicative constants.

Let us recall the quadrature two-grid method, which corresponds to the iteration method 1 for the Nyström method in [At97] (p. 249). The coarse level is denoted by $\nu \ll n$ and the Nyström coarse approximate operator by T_ν:

$$u_n^{(0)} = 0, \quad u_n^{(k)} = S_{n,\nu} u_n^{(k-1)} + (I - T_\nu)^{-1} f,$$

where $S_{n,\nu} = (I - T_\nu)^{-1}(T_n - T_\nu)$.

We set

$$b_m = \sup_{0 \le k,p \le m} \frac{k!}{p!(k-p)!},$$

$$c_1 = \|(I-T)^{-1}\|_{\mathcal{L}(C)}, \quad c_2 = \|(I-T)^{-1}\|_{\mathcal{L}(C^m)}, \quad c_3 = \|k\|_{0,0}(1+a_q),$$
$$c_4 = 2c_1(1+c_1c_3), \quad c_5 = 1 + c_4 a_q \|k\|_{m,0}, \quad c_6 = 2c_2(1 + c_2\|T - T_n\|_{\mathcal{L}(C^m)}),$$
$$c_7 = 2c_q c_4 b_m \|k\|_{0,m}, \quad c_8 = 2c_q c_6 b_m \|k\|_{m,m}, \quad c_{10} = 2c_5 a_q \|k\|_{m,0}.$$

Proposition 1. *If $k \in C^m([0,1] \times [0,1])$, then*

$$\|u_n^{(2l)} - u_n\|_0 \le c_5 (c_7 c_{10})^l \|f\|_m \nu^{-lm},$$
$$\|u_n^{(2l+1)} - u_n\|_0 \le c_5 c_7 (c_7 c_{10})^l \|f\|_m \nu^{-(l+1)m}.$$

If $k \in C^{2m}([0,1] \times [0,1])$, then

$$\|u_n - u_n^{(k)}\|_m \le c_5 c_8^k \|f\|_m \nu^{-mk}.$$

In [At97], p. 257, Atkinson gives the computational cost of the quadrature two-grid method 1: the total cost in operations per iteration is approximately

$$q_n^2 + 2q_n q_\nu + q_\nu^2 = (q_n + q_\nu)^2.$$

16.3 A New Two-Grid Method

This method is described in [Ku04]. Let us define the "coarse" approximate operator by

$$\tilde{T}_\nu = P_\nu T_n + T_n P_\nu - P_\nu T_n P_\nu,$$

where P_ν is a projection into $\mathcal{P}_{r,\Delta_\nu}$, the set of piecewise polynomial functions of degree less than or equal to r on the partition $\Delta_\nu = (t_{i,\nu})_{i=1,\dots,q_\nu}$. Kulkarni's two-grid method can be written as

$$\tilde{u}_n^{(0)} = 0, \quad \tilde{u}_n^{(k)} = \tilde{S_{n,\nu}}\tilde{u}_n^{(k-1)} + (I - \tilde{T_\nu})^{-1} f,$$

where

$$\tilde{S}_{n,\nu} = (I - \tilde{T_\nu})^{-1}(T_n - \tilde{T_\nu}).$$

In what follows, we take $r = m$.

We have $T_n - \tilde{T_\nu} = (I - P_\nu)T_n(I - P_\nu)$ so that for any operator norm $\|.\|$, $\|T_n - \tilde{T_\nu}\| \leq \|(I - P_\nu)T_n\|(I \ - P_\nu)\|$. As P_ν is uniformly bounded and $\|(I - P_\nu)T_n\|$ tends towards zero because T_n is compact, $\|T_n - \tilde{T_\nu}\|$ tends towards zero. That is an advantage over Nyström.

As the range of P_ν belongs to the set of the piecewise continuous functions, we write $\|.\|_{\mathcal{L}(C)}$ instead of $\|.\|_{\mathcal{L}(L^\infty)}$.

If P_ν is the orthogonal projection or the interpolatory projection on $\mathcal{P}_{r,\Delta_\nu}$, we quote the following estimate from Chatelin–Lebbar [ChLe84]: "if $C^m_{\Delta_\nu}$ is the set of piecewise C^m-functions, then there is a constant c_I such that for all $u \in C^m_{\Delta_\nu}$,

$$\|(I - P_\nu)u\|_0 \leq c_I \nu^{-\beta} \|u^{(\beta)}\|_0, \tag{16.3}$$

where $\beta = \min(m, r + 1)$."

Proposition 2. *If $k \in C^m([0,1] \times [0,1])$, then*

$$\|T_n - \tilde{T_\nu}\|_{\mathcal{L}(C)} \leq c_I a_q \|k\|_{m,0} \|I - P_\nu\|_{\mathcal{L}(C)} \nu^{-m},$$
$$\|T_n - \tilde{T_\nu}\|_{\mathcal{L}(C^m,C)} \leq c_I^2 a_q \|k\|_{m,0} \nu^{-2m}.$$

Proof. As $k \in C^m([0,1] \times [0,1])$, we have $T_n(I - P_\nu) \in C^m[0,1]$, so, according to (16.3),

$$\|(I - P_\nu)T_n(I - P_\nu)u\|_0 \leq c_I \left(\frac{1}{\nu}\right)^m \|T_n(I - P_\nu)(u)^{(m)}\|_0.$$

Since

$$T_n(I - P_\nu)(u)^{(m)}(x) = \sum_{j=1}^{n} w_{j,n} D^{m,0} k(x, x_{j,n})(I - P_\nu)(u)(x_{j,n}),$$

it follows that

$$|T_n(I - P_\nu)(u)^{(m)}(x)| \leq a_q \|D^{m,0} k\|_0 \max_{j=1,\ldots,n} |(I - P_\nu)(u)(x_{j,n})|.$$

Setting

$$\|u - P_\nu u\|_{0,\Delta_j} = \max_{x \in [\frac{j-1}{\nu}, \frac{j}{\nu}]} |u(x) - P_\nu u(x)|, \quad \|u - P_\nu u\|_0 = \max_j \|u - P_\nu u\|_{0,\Delta_j},$$

we obtain

$$\|T_n(I - P_\nu)(u)^{(m)}\|_0 \leq a_q \|D^{m,0} k\|_0 \|u - P_\nu u\|_0;$$

hence,

$$\|T_n(I-P_\nu)^{(m)}\|_{\mathcal{L}(C)} \le a_q \|D^{m,0}k\|_0 \|I-P_\nu\|_{\mathcal{L}(C)}.$$

If $u \in C^m$, then, by (16.3), $\|(I-P_\nu)u\|_0 \le c_I \nu^{-m}\|u^{(m)}\|_0$, so,

$$\|T_n(I-P_\nu)(u)^{(m)}\|_0 \le a_q \|D^{m,0}k\|_0 c_I \left(\frac{1}{\nu}\right)^m \|u^{(m)}\|_0,$$

and we arrive at

$$\|(I-P_\nu)T_n(I-P_\nu)\|_{\mathcal{L}(C^m,C)} \le c_I^2 a_q \|D^{m,0}k\|_0 \left(\frac{1}{\nu}\right)^{2m}.$$

Proposition 3. *Suppose that $I-T$ is nonsingular. If $k \in C^m([0,1]\times[0,1])$, then for ν large enough, $I-\tilde{T}_\nu$ is nonsingular and*

$$\|(I-\tilde{T}_\nu)^{-1}\|_{\mathcal{L}(C)} \le 2c_4,$$

where c_4 is defined in the previous section.

Proof. We have $I-\tilde{T}_\nu = [I-(\tilde{T}_\nu - T_n)(I-T_n)^{-1}][I-T_n]$. Since

$$\|(\tilde{T}_\nu - T_n)(I-T_n)^{-1}\|_{\mathcal{L}(C)} \le \|(\tilde{T}_\nu - T_n)\|_{\mathcal{L}(C)} \|(I-T_n)^{-1}\|_{\mathcal{L}(C)},$$

and the right-hand side of the inequality tends to zero as ν tends to infinity, if follows that for ν large enough,

$$\|(\tilde{T}_\nu - T_n)(I-T_n)^{-1}\|_{\mathcal{L}(C)} \le \frac{1}{2} < 1.$$

According to Lemma 12.3, p. 198 in [Li06], $I-(\tilde{T}_\nu - T_n)(I-T_n)^{-1}$ is invertible and

$$\|\big(I-(\tilde{T}_\nu - T_n)(I-T_n)^{-1}\big)^{-1}\|_{\mathcal{L}(C)} \le \frac{1}{1-\|(\tilde{T}_\nu - T_n)(I-T_n)^{-1}\|_{\mathcal{L}(C)}} \le 2.$$

We have

$$\|(I-T_n)^{-1}\|_{\mathcal{L}(C)} \le c_4.$$

Indeed, $I-T_n = [I-(T_n-T)(I-T)^{-1}][I-T]$. As

$$\begin{aligned}\|(T_n-T)(I-T)^{-1}(T_n-T)(I-T)^{-1}\|_{\mathcal{L}(C)} \\ \le c_1\|(T_n-T)^2\|_{\mathcal{L}(C)} + c_1^2\|(T_n-T)T\|_{\mathcal{L}(C)}\|(T_n-T)\|_{\mathcal{L}(C)}\end{aligned}$$

and in view of the collectively compact convergence of T_n to T, $\|(T_n-T)(I-T)^{-1}(T_n-T)(I-T)^{-1}\|_{\mathcal{L}(C)}$ tends to zero, so, for n large enough,

$$\|(T_n-T)(I-T)^{-1}(T_n-T)(I-T)^{-1}\|_{\mathcal{L}(C)} \le \frac{1}{2} < 1.$$

By Lemma 12.3, p. 198 in [Li06], $I-(T_n-T)(I-T)^{-1}$ is invertible and

$$\begin{aligned}\|(I-(T_n-T)(I-T)^{-1})^{-1}\|_{\mathcal{L}(C)} \\ \leq \frac{1+\|(T_n-T)(I-T)^{-1}\|_{\mathcal{L}(C)}}{1-\|(T_n-T)(I-T)^{-1}(T_n-T)(I-T)^{-1}\|_{\mathcal{L}(C)}} \\ \leq 2c_1(1+c_1\|T_n-T\|_{\mathcal{L}(C)}).\end{aligned}$$

Since $\|T_n-T)\|_{\mathcal{L}(C)} \leq \|T_n\|_{\mathcal{L}(C)}+\|T\|_{\mathcal{L}(C)} \leq \|k\|_{0,0}(1+a_q)$, we now obtain our estimate. Thus,

$$\|(I-\tilde{T}_\nu)^{-1}\|_{\mathcal{L}(C)} \leq 2c_4.$$

Proposition 4. *If $k \in C^m([0,1]\times[0,1])$, then, for ν large enough,*

$$\|\tilde{S_{n,\nu}}\|_{\mathcal{L}(C^m,C)} \leq d_7\nu^{-2m}, \quad d_7 = 2c_4c_I^2a_q\|k\|_{m,0},$$
$$\|\tilde{S_{n,\nu}}\|_{\mathcal{L}(C)} \leq d_{9,\nu}\left(\frac{1}{\nu}\right)^m, \quad d_{9,\nu} = 2c_4c_Ia_q\|k\|_{m,0}\|I-P_\nu\|_{\mathcal{L}(C)}.$$

Proof. We have

$$\|\tilde{S}_{n,\nu}\|_{\mathcal{L}(C^m,C)} \leq \|(I-\tilde{T}_\nu)^{-1}\|_{\mathcal{L}(C)}\|(T_n-\tilde{T}_\nu)\|_{\mathcal{L}(C^m,C)},$$
$$\|\tilde{S}_{n,\nu}\|_{\mathcal{L}(C)} \leq \|(I-\tilde{T}_\nu)^{-1}\|_{\mathcal{L}(C)}\|(T_n-\tilde{T}_\nu)\|_{\mathcal{L}(C)}.$$

By Proposition 2,

$$\|\tilde{S}_{n,\nu}\|_{\mathcal{L}(C^m,C)} \leq 2c_4c_I^2a_q\|k\|_{m,0}\nu^{-2m},$$
$$\|\tilde{S}_{n,\nu}\|_{\mathcal{L}(C)} \leq 2c_4c_Ia_q\|k\|_{m,0}\|I-P_\nu\|_{\mathcal{L}(C)}\nu^{-m}.$$

Proposition 5. *If $k \in C^m([0,1]\times[0,1])$, then*

$$\|\tilde{S}^k_{n,\nu}\|_{\mathcal{L}(C^m,C)} \leq d_7d_{9,\nu}^{k-1}\nu^{-(k+1)m}, \quad k \in \mathbb{N}^*.$$

Proof. As $\tilde{S}^k_{n,\nu} = \tilde{S}^{k-1}_{n,\nu}\tilde{S}_{n,\nu}$, we have

$$\|\tilde{S}^k_{n,\nu}\|_{\mathcal{L}(C^m,C)} \leq \|\tilde{S}^{k-1}_{n,\nu}\|_{\mathcal{L}(C)}\|\tilde{S}_{n,\nu}\|_{\mathcal{L}(C^m,C)},$$

so,

$$\|\tilde{S}_{n,\nu}\|_{L(C^m,C)} \leq d_7d_{9,\nu}^{k-1}\nu^{-(k+1)m}.$$

Theorem 1. *If $k \in C^m([0,1]\times[0,1])$, $\tilde{u}_n^{(k)}$ is the new two-grid solution, and u_n is the Nyström approximation, then*

$$\|\tilde{u}_n^{(k)}-u_n\|_0 \leq c_5d_7d_{9,\nu}^{k-1}\|f\|_m\nu^{-(k+1)m}.$$

Proof. We have

$$u_n = \tilde{S_{n,\nu}} u_n + (I - T_\nu)^{-1} f, \quad \tilde{u}_n^{(k)} = \tilde{S}_{n,\nu} \tilde{u}_n^{(k-1)} + (I - T_\nu)^{-1} f;$$

so,

$$u_n - \tilde{u}_n^{(k)} = \tilde{S}_{n,\nu}(u_n - \tilde{u}_n^{(k-1)}).$$

Then

$$u_n - u_n^{(k)} = \tilde{S}_{n,\nu}^k (u_n - \tilde{u}_n^{(0)}) = S_{n,\nu}^k (I - T_n)^{-1} f,$$

and we deduce that

$$\|u - \tilde{u}_n^{(k)}\|_0 \leq \|S_{n,\nu}^k\|_{\mathcal{L}(C^m, C)} \|(I - T_n)^{-1}\|_{\mathcal{L}(C^m, C^m)} \|f\|_m,$$

from which,

$$\|u - \tilde{u}_n^{(k)}\|_0 \leq c_5 d_7 d_{9,\nu}^{k-1} \left(\frac{1}{\nu}\right)^{(k+1)m} \|f\|_m.$$

In [Ku04], p. 370, Kulkarni gives the computational cost of this two-grid projection method: it is, per iteration, approximately

$$q_n(q_n + 6q_\nu) + 2q_\nu^2.$$

Compared to the first two-grid method, there is an additional cost involved in generating matrices, but for $n \gg \nu$, the computational costs of both two-grid methods are comparable.

16.4 Comparison of Error Estimates

The aim of two-grid methods is to provide, at a lower cost, an approximation whose accuracy is the same as that of the Nyström approximation.

In this section, we will compare the errors for different assumptions on the data. $u_n^{(k)}$ will denote the quadrature two-grid iterate, $\tilde{u_n}^{(k)}$ the Kulkarni two-grid iterate, u_n the Nyström approximation, and u the exact solution of the equation $u = Tu + f$.

Using the previous notation, we have

$$\|u_n - u\|_0 \leq c_1 c_5 c_q b_m \|k\|_{0,m} \|f\|_m n^{-m}.$$

- In the case where $k \in C^m([0,1] \times [0,1])$, we have

$$\|u_n^{(2\ell+1)} - u_n\|_0 \leq c_5 c_7 (c_7 c_{10})^\ell \|f\|_m \nu^{-(\ell+1)m},$$

$$\|\tilde{u}_n^{(2\ell+1)} - u_n\|_0 \leq c_5 d_7 d_{9,\nu}^{2\ell} \|f\|_m \left(\frac{1}{\nu}\right)^{2(\ell+1)m}.$$

The order of accuracy of Kulkarni's two-grid iteration method is $2(\ell+1)m$, that is, twice the order of the quadrature iteration methods, which is $(\ell+1)m$.

Suppose that

$$\nu = n^{\rho}, \quad 0 < \rho < 1.$$

In the quadrature two-grid method, one has to perform at least k^* iterations with $k^* \geq 2\rho^{-1} - 1$ to reach the desired accuracy. In the Kulkarni two-grid method, one has to perform at least $\tilde{k}^*$ iterations with $\tilde{k}^* \geq \rho^{-1} - 1$ to reach the desired accuracy.

Theorem 2. *Let $k \in C^m([0,1] \times [0,1])$. If $k \geq 2\rho^{-1} - 1$, then*

$$\|u_n^{(k)} - u\|_0 \leq c_5 \left(c_7 (c_7 c_{10})^{\frac{k-1}{2}} + c_1 c_q b_m \|k\|_{0,m} \right) \|f\|_m n^{-m}.$$

If $k \geq 1\rho^{-1} - 1$, then

$$\|\tilde{u_n}^{(k)} - u\|_0 \leq c_5 \left(d_7 d_{9,\nu}^{k-1} + c_1 c_q b_m \|k\|_{0,m} \right) \|f\|_m n^{-m}.$$

- In the case where $k \in C^{2m}([0,1] \times [0,1])$, we have

$$\|u_n^{(k)} - u_n\|_m \leq c_5 (c_8)^k \|f\|_m \nu^{-km},$$
$$\|\tilde{u}_n^{(k)} - u_n\|_0 \leq c_5 d_7 d_{9,\nu}^{k-1} \|f\|_m \nu^{-(k+1)m}.$$

The order of accuracy of the Kulkarni two-grid iteration method is better than that of the quadrature two-grid method.

Let us compare the multiplicative factors. Let us define $\kappa^{(2\ell+1)}$ as the multiplicative factor of the Kulkarni two-grid $(2\ell+1)$th iterate divided by the multiplicative factor of the quadrature two-grid $(2\ell+1)$th iterate. In the case where $k \in C^m([0,1] \times [0,1])$, we have

$$\kappa^{(2\ell+1)} = \left(\frac{1}{b_m}\right)^{\ell+1} \left(\frac{c_4}{1 + c_4 a_q \|k\|_{m,0}}\right)^{\ell} \left(\frac{a_q c_I^2}{c_q}\right)^{\ell+1} \left(\frac{\|k\|_{m,0}}{\|k\|_{0,m}}\right)^{\ell+1} (\|I - P_\nu\|)^{2\ell}.$$

The first factor tends to zero with ℓ, and in most cases the second one also. The third factor depends on the chosen quadrature and the interpolation estimation. The fourth one depends on the regularity properties of the kernel, and the last factor depends on the chosen projection P_ν.

In the case where $k \in C^{2m}([0,1] \times [0,1])$, we have

$$\kappa^{(k)} = c_I \left(\frac{1}{b_m}\right)^k \left(\frac{c_4}{c_6}\right)^k \left(\frac{a_q c_I}{c_q}\right)^k (\|I - P_\nu\|)^{k-1} \left(\frac{\|k\|_{m,0}}{\|k\|_{m,m}}\right)^k.$$

Here, the last term generally tends to zero with k.

Acknowledgement. This work was partially supported by the program ARCUS-INDE (Actions en Régions de Coopération Universitaire et Scientifique) 2005–2008 of the Rhône-Alpes region. The author is also indebted to Rekha Kulkarni, whose ideas motivated this chapter.

References

[AhLa01] Ahues, M., Largillier, A., Limaye, B.V.: *Spectral Computations for Bounded Operators,* Chapman & Hall, Boca Raton, FL (2001).

[At97] Atkinson, K.E.: *The Numerical Solution of Integral Equations of the Second Kind,* Cambridge University Press, London (1997).

[Li06] Limaye, B.V.: *Functional Analysis,* 2nd ed., New Age International Publishers, New Delhi, India (2006).

[ChLe84] Chatelin F., Lebbar, R.: Superconvergence results for the iterated projection method applied to a Fredholm integral equation of the second kind and the corresponding eigenvalue problem. *J. Integral Equations Appl.,* **6**, 71–91 (1984).

[Ku04] Kulkarni, R.P.: Approximate solution of multivariate integral equation of the second kind. *J. Integral Equations Appl.,* **16**, 343–374 (2004).

[Va05] Vainikko, G.: Fast solvers of integral equations of the second kind: quadrature methods. *J. Integral Equations Appl.,* **17**, 91–120 (2005).

[We02] Werschulz, A.G.: An overview of information-based complexity. Technical Report CUCS-022-02 (2002).

[We03] Werschulz, A.G.: Where does smoothness count the most for Fredholm equations of the second kind with noisy information? *J. Complexity,* **19**, 758–798 (2003).

17

Homogenized Models of Radiation Transfer in Multiphase Media

A.V. Gusarov and I. Smurov

École Nationale d'Ingénieurs de Saint-Étienne (ENISE), France;
gusarov@enise.fr, smurov@enise.fr

17.1 Introduction

The physical model of an absorbing scattering medium (ASM) and the corresponding mathematical model of the radiation transfer equation (RTE) were originally formulated to study dilute dispersed systems like fog [SiHo02]. Such media contain well-separated small particles (droplets), and so they are essentially heterogeneous. Therefore, even the derivation of the conventional RTE can be considered as a problem of homogenization. Nevertheless, homogenization of radiation transfer is often meant as a problem for the conventional RTE with oscillating coefficients [Pa05]. Oscillations with a period much smaller than the characteristic length scale of the problem can be averaged to obtain the effective coefficients of the homogenized RTE. Thus, from a physical point of view, a heterogeneous ASM with short-scale variations of the radiative properties is replaced by an equivalent homogeneous ASM with the effective radiative properties. Such a homogenization problem contains two small length scales of different size. The smallest scale is the size of the scattering inhomogeneity (particle) and the intermediate scale is the period of oscillations of the radiative properties.

The radiative properties of dilute dispersed media can be obtained from the scattering properties of a single particle [SiHo02], but considerable difficulties arise in dense dispersed systems where the volume fraction of the dispersed phases is comparable with the volume fraction of the matrix. The distances between the scatterers (particles) become comparable with their sizes in such systems, so that a mutual influence of the scatterers should be taken into account. This is referred to as dependent scattering. The current mathematical approach to dependent scattering is the RTE with modified radiative properties [BaSa00]. However, the applicability of the RTE to dense dispersed systems has never been rigorously proved.

A model of homogenization relative to the smallest scale of a uniform monophase domain was recently proposed for two-phase composite

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_17,

media [Gu08]. Each phase is considered as an absorbing but not a scattering medium in this model. The reflection and refraction on the boundaries between the phases are the only scattering events. This model has no restrictions on the phase composition and allows us to derive the conventional RTE for dilute dispersed systems as well as to describe the effects of dependent scattering when the volume fractions of the two phases are comparable.

This chapter aims to generalize the previous two-phase model [Gu08] to the case of an arbitrary number of phases and to analyze the derived equations.

17.2 Multiphase Model

A detailed description of radiation is given by its angular intensity $i(\mathbf{r}, \mathbf{\Omega})$ [SiHo02] defined at point $\mathbf{r}$ in the direction specified by its unit vector $\mathbf{\Omega}$. A detailed distribution of N phases in space can be characterized by their phase functions $\phi_\gamma(\mathbf{r})$ [To02] defined for $\gamma = 0, ..., N-1$ as

$$\phi_\gamma = \begin{cases} 1 & \text{in phase } \gamma, \\ 0 & \text{elsewhere.} \end{cases}$$

The detailed radiation intensity $i(\mathbf{r}, \mathbf{\Omega})$ could be calculated by transport equations

$$\mathbf{\Omega} \nabla i_\gamma = -\alpha_\gamma i_\gamma$$

in each monophase domain with the absorption coefficient α_γ related by the boundary conditions of reflection/refraction on the phase boundaries.

The structure of a multiphase medium is often characterized by averages as the volume fractions of phases and the specific surface of phase boundaries while the detailed phase distribution is unknown. In this case a detailed description of radiation transfer becomes impossible. Moreover, it will be excessive if the detailed radiation intensity is averaged at a physical measurement. The first question is: What average value can correctly represent the detailed radiation intensity in a domain containing a great number of morphological features (monophase domains)? The average of the intensity i itself is a bad choice because i can be considerably different even in neighbor monophase domains. The examples are given by an opaque phase with zero intensity and a transparent phase with nonzero intensity and two transparent phases of different refraction indices in thermal equilibrium where the ratio of the intensities is proportional to the ratio of the refraction indices squared [Gu08].

The most precise homogenized description is supposed to be given by N values obtained by averaging over each phase within the representative domain [ZeIaTa06, Gu08]. Note that the radiation intensity is defined by the energy flux through a surface [SiHo02]. Therefore, a representative surface is proposed rather than a representative volume domain [Gu08]. The partial averaged radiation intensities I_γ are defined on the representative surface S containing a great number of intersections with monophase domains [Gu08]:

$$I_\gamma = \frac{1}{Sf_\gamma} \int\limits_S i\phi_\gamma d\mathbf{s},$$

where the volume fraction of phase γ can be obtained on the same surface S with the surface element $d\mathbf{s}$:

$$f_\gamma = \frac{1}{S} \int\limits_S \phi_\gamma d\mathbf{s}.$$

The energy balance is given by a system of N integro-differential equations similar to the RTE but containing additional integral terms corresponding to radiation exchange between phases [ZeIaTa06, Gu08]. The two-phase model [Gu08] evaluates the coefficients of these equations based on the assumptions that the medium is statistically isotropic and that consecutive reflection/refraction events do not correlate. The same assumptions and a similar derivation procedure applied to a multiphase medium result in the following energy balance for phase γ:

$$\begin{aligned}\mathbf{\Omega}\nabla I_\gamma = -\big(\alpha_\gamma + \frac{A_\gamma}{4f_\gamma}\big)I_\gamma + \frac{\rho_\gamma A_\gamma}{4f_\gamma}\frac{1}{4\pi}\int\limits_{4\pi} I_\gamma(\mathbf{\Omega}')P_{\gamma\gamma}(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}' \\ + \sum_{\delta\neq\gamma}\frac{(1-\rho_{\delta\gamma})A_{\gamma\delta}}{4f_\gamma}\frac{1}{4\pi}\int\limits_{4\pi} I_\delta(\mathbf{\Omega}')P_{\delta\gamma}(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}', \quad (17.1)\end{aligned}$$

where $A_{\gamma\delta} = A_{\delta\gamma}$ is the specific surface of the boundary between phases γ and δ per unit volume of the multiphase medium,

$$A_\gamma = \sum_\delta A_{\delta\gamma} \quad (17.2)$$

is the specific surface of phase γ, $\rho_{\delta\gamma}$ is the hemispherical reflectivity of the interface γ/δ for the incidence from phase δ, and

$$\rho_\gamma = \frac{1}{A_\gamma}\sum_\delta \rho_{\gamma\delta}A_{\delta\gamma} \quad (17.3)$$

is the weighted average of the hemispherical reflectivity of the boundary of phase γ for the incidence from this phase. The hemispherical reflectivity is the optical property of a surface defined in [SiHo02]. Note that generally $\rho_{\gamma\delta} \neq \rho_{\delta\gamma}$. The relation

$$\frac{1-\rho_{\gamma\delta}}{1-\rho_{\delta\gamma}} = m_{\delta\gamma}^2 \quad (17.4)$$

follows from the optical reversibility [SiHo02] where $m_{\delta\gamma} = m_\delta/m_\gamma$ is the ratio of refraction indices m_δ and m_γ of phases δ and γ, respectively.

Symmetry relations for the scattering phase functions also follow from the optical reversibility:

$$P_{\gamma\delta}(\mathbf{\Omega}', \mathbf{\Omega}) = P_{\delta\gamma}(\mathbf{\Omega}, \mathbf{\Omega}'). \tag{17.5}$$

In an isotropic medium scattering phase functions depend on the scattering angle ψ between directions $\mathbf{\Omega}$ and $\mathbf{\Omega}'$; therefore, the two arguments can be exchanged:

$$P_{\gamma\delta}(\mathbf{\Omega}', \mathbf{\Omega}) = P_{\gamma\delta}(\psi) = P_{\gamma\delta}(\mathbf{\Omega}, \mathbf{\Omega}'). \tag{17.6}$$

Equations (17.5) and (17.6) show that the indices and the arguments can be exchanged in any combination in the isotropic medium. This proves the symmetry of the scattering matrix. In addition, the conventional normalizing condition is required:

$$\frac{1}{4\pi} \int\limits_{4\pi} P_{\delta\gamma}(\mathbf{\Omega}', \mathbf{\Omega}) d\mathbf{\Omega}' = 1. \tag{17.7}$$

The phase functions with $\gamma \neq \delta$ describe radiation exchange between phases γ and δ by refraction at the interface. These components are evaluated as [Gu08]:

$$P_{\gamma\delta}(\psi) = P_{\delta\gamma}(\psi) = 2 \frac{1 - \rho'_{\gamma\delta}(\chi)}{1 - \rho_{\gamma\delta}} \frac{d \cos^2 \chi}{d \cos(\chi - \chi')}, \tag{17.8}$$

where χ is the incidence, χ' the refraction, and $\psi = |\chi - \chi'|$ the scattering angles, and $\rho'_{\gamma\delta}(\chi)$ is the directional-hemispherical reflectivity for the incidence from phase γ defined in [SiHo02]. The symmetry of the right-hand side (17.8) against the exchange of the indices follows from the optical reversibility for the directional-hemispherical reflectivity, $\rho'_{\gamma\delta}(\chi) = \rho'_{\delta\gamma}(\chi')$, Snell's law of refraction, and relation (17.4). The hemispherical and the directional-hemispherical reflectivities are related as [SiHo02]

$$\rho_{\gamma\delta} = 2 \int\limits_0^1 \rho'_{\gamma\delta}(\chi) \cos \chi d \cos \chi. \tag{17.9}$$

The diagonal phase functions $P_{\gamma\gamma}$ describe back reflection of radiation by the boundary of phase γ. They are given by the weighted average

$$P_{\gamma\gamma}(\psi) = \frac{1}{\rho_\gamma A_\gamma} \sum_\delta \rho'_{\gamma\delta}(\chi) A_{\delta\gamma}, \tag{17.10}$$

with the scattering ψ and incidence χ angles related at specular reflection as $\psi + 2\chi = \pi$. Note that (17.8) and (17.10) satisfy normalizing condition (17.7). To prove this, one can choose the spherical coordinates with the axis parallel to the $\mathbf{\Omega}$ direction where the polar angle is the scattering angle ψ and $d\mathbf{\Omega}' = 2\pi d \cos \psi$ and then apply (17.9) along with definitions (17.2) and (17.3). Examples of scattering phase functions (17.8) and (17.10) derived from the Fresnel formulas for reflection and refraction are shown in [Gu08].

17.3 Dilute Dispersed Media

Suppose that the phase denoted by $\gamma = 0$ (the matrix phase) prevails in volume:

$$1 - f_0 \ll 1, \tag{17.11}$$

and that the grains of the other $N-1$ phases (the dispersed phases) do not touch each other:

$$A_{\gamma\delta} = 0 \quad \text{if } \gamma\delta \neq 0.$$

Equation (17.11) implies that the volume fractions of the dispersed phases are small. Because the volume fraction is in the denominators of the terms of the right-hand side of (17.1), the left-hand side of this equation for a dispersed phase can be neglected. This brings system (17.1) to the form

$$\begin{aligned}
\mathbf{\Omega}\nabla I_0 &= -\big(\alpha_0 + \frac{A_0}{4}\big)I_0 + \frac{\rho_0 A_0}{4}\frac{1}{4\pi}\int\limits_{4\pi} I_0(\mathbf{\Omega}')P_{00}(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}' \\
&\quad + \sum_{\delta=1}^{N-1}\frac{(1-\rho_{\delta 0})A_{0\delta}}{4}\frac{1}{4\pi}\int\limits_{4\pi} I_\delta(\mathbf{\Omega}')P_{\delta 0}(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}', \\
\mathbf{0} &= -\big(\alpha_\gamma f_\gamma + \frac{A_\gamma}{4}\big)I_\gamma + \frac{\rho_\gamma A_\gamma}{4}\frac{1}{4\pi}\int\limits_{4\pi} I_\gamma(\mathbf{\Omega}')P_{\gamma\gamma}(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}' \\
&\quad + \frac{(1-\rho_{0\gamma})A_{\gamma 0}}{4}\frac{1}{4\pi}\int\limits_{4\pi} I_0(\mathbf{\Omega}')P_{0\gamma}(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}', \ \gamma = 1,\ldots,N-1.
\end{aligned} \tag{17.12}$$

The second equation of system (17.12) is the Fredholm integral equation relative to I_γ. It indicates that radiation in dispersed phase $\gamma \neq 0$ is locally consistent with the radiation in the matrix phase. The general solution of this equation is given through the resolving kernel $K_\gamma(\mathbf{\Omega}',\mathbf{\Omega})$:

$$I_\gamma(\mathbf{\Omega}) = \frac{m_{\gamma 0}^2}{4\pi}\int\limits_{4\pi} I_0(\mathbf{\Omega}')K_\gamma(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}', \tag{17.13}$$

where factor $m_{\gamma 0}^2/4\pi$ is separated from the kernel. In a statistically isotropic medium the kernel depends on the angle between directions $\mathbf{\Omega}$ and $\mathbf{\Omega}'$ only and is, therefore, symmetric like the scattering phase functions (17.6). Substituting (17.13) into the second equation (17.12) gives the following equation for the kernel:

$$\begin{aligned}
\left(1 + \frac{4\alpha_\gamma f_\gamma}{A_\gamma}\right)K_\gamma(\mathbf{\Omega}',\mathbf{\Omega}) &= \frac{\rho_\gamma}{4\pi}\int\limits_{4\pi} K_\gamma(\mathbf{\Omega}',\mathbf{\Omega}'')P_{\gamma\gamma}(\mathbf{\Omega}'',\mathbf{\Omega})d\mathbf{\Omega}'' \\
&\quad + (1-\rho_{\gamma 0})P_{0\gamma}(\mathbf{\Omega}',\mathbf{\Omega}),
\end{aligned} \tag{17.14}$$

where relation (17.4) is taken into account. The normalizing condition for K_γ is obtained by integration of (17.14) over $\mathbf{\Omega}$:

$$\frac{1}{4\pi}\int\limits_{4\pi} K_\gamma(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega} = \frac{1-\rho_{\gamma 0}}{1-\rho_\gamma + 4\alpha_\gamma f_\gamma/A_\gamma}. \tag{17.15}$$

Note that in case of a transparent dispersed phase with $\alpha_\gamma = 0$, the right-hand side equals unity because $\rho_{\gamma 0} = \rho_\gamma$ for the considered medium. Thus, condition (17.15) for the kernel is similar to (17.7) for the scattering phase function.

Substituting (17.13) into the first equation (17.12) reduces the problem to a conventional RTE:

$$\mathbf{\Omega}\nabla I_0 = -\big(\alpha_0 + \frac{A_0}{4}\big)I_0 + \frac{1}{4\pi}\int\limits_{4\pi} I_0(\mathbf{\Omega}')Q(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}', \tag{17.16}$$

where the kernel of the integral transform is

$$\begin{aligned} Q(\mathbf{\Omega}',\mathbf{\Omega}) &= \frac{\rho_0 A_0}{4}P_{00}(\mathbf{\Omega}',\mathbf{\Omega}) \\ &\quad + \sum_{\delta=1}^{N-1}\frac{(1-\rho_{0\delta})A_{\delta 0}}{4}\frac{1}{4\pi}\int\limits_{4\pi} K_\delta(\mathbf{\Omega}',\mathbf{\Omega}'')P_{\delta 0}(\mathbf{\Omega}'',\mathbf{\Omega})d\mathbf{\Omega}''. \end{aligned} \tag{17.17}$$

17.3.1 Effective Radiative Properties

In the case of dilute dispersed systems, the multiphase model explained above is rigorously reduced to the conventional model of an absorbing scattering medium described by RTE (17.16). The effective radiative properties follow from this equation. The effective extinction coefficient is the absolute value of the factor before I_0 in the first term of the right-hand side:

$$\beta_e = \alpha_0 + A_0/4. \tag{17.18}$$

The effective scattering coefficient is evaluated from (17.17):

$$\sigma_e = \frac{1}{4\pi}\int\limits_{4\pi} Q(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega} = \frac{1}{4}\sum_{\delta=1}^{N-1} A_{\delta 0}\frac{1-\rho_{\delta 0}+4\rho_{0\delta}\alpha_\delta f_\delta/A_{\delta 0}}{1-\rho_{\delta 0}+4\alpha_\delta f_\delta/A_{\delta 0}}. \tag{17.19}$$

The effective scattering phase function is the normalized kernel (17.17):

$$P_e(\mathbf{\Omega}',\mathbf{\Omega}) = Q(\mathbf{\Omega}',\mathbf{\Omega})/\sigma_e. \tag{17.20}$$

In what follows, (17.18)–(17.20) are compared with the known results obtained by ray optics. Equation (17.18) presents extinction as a superposition of internal absorption in the matrix phase and shadowing by dispersed particles. The second term responsible for the shadowing is rigorous for randomly oriented particles of arbitrary convex shape [SiHo02]. The effective scattering is given by (17.19) as the result of independent scattering by the dispersed phases. The input of transparent phase δ with $\alpha_\delta = 0$ equals $A_{\delta 0}/4$.

The input of an opaque phase with $\alpha_\delta \to \infty$ equals $\rho_{0\delta} A_{\delta 0}/4$. These limits are rigorous for randomly oriented convex particles in the framework of ray optics [SiHo02, Gu08]. The intermediate case of semi-transparent spherical particles was studied in [Gu08] where differences in the effective scattering coefficient between the multiphase model and ray optics were found.

The effective scattering phase function (17.20) was validated by comparison with ray tracing for transparent spheres of the dispersed phase in transparent matrix [Gu08]. The multiphase model was found to smooth the angular distribution of the scattered radiation. The revealed discrepancies in the effective scattering coefficient and phase function were explained by strong correlations between the consecutive reflection/refraction events in transparent spheres neglected by the multiphase model [Gu08]. The mentioned correlations are less important for dispersed particles of irregular shape where the discrepancies are expected to decrease.

17.3.2 Opaque Dispersed Phases

The size of a dispersed particle is estimated as the ratio of its volume to the surface, f_γ/A_γ. Phase γ is referred to as the opaque phase if the size of its particles is much greater than the absorption length $1/\alpha_\gamma$:

$$\alpha_\gamma f_\gamma / A_\gamma \gg 1. \tag{17.21}$$

Let first N_p dispersed phases satisfy condition (17.21). According to the second equation of (17.12), $I_\gamma = 0$ for these phases. This means that radiation does not penetrate into opaque particles. This does not change the effective extinction coefficient (17.18) and simplifies the terms responsible for scattering by opaque phases in equations (17.19) and (17.17):

$$\sigma_e = \frac{1}{4}\sum_{\delta=1}^{N_p} \rho_{0\delta} A_{\delta 0} + \frac{1}{4}\sum_{\delta=N_p+1}^{N-1} A_{\delta 0} \frac{1-\rho_{\delta 0} + 4\rho_{0\delta}\alpha_\delta f_\delta / A_{\delta 0}}{1-\rho_{\delta 0} + 4\alpha_\delta f_\delta / A_{\delta 0}}, \tag{17.22}$$

$$Q(\psi) = \frac{1}{4}\sum_{\delta=1}^{N_p} \rho'_{0\delta}\big(\frac{\pi-\psi}{2}\big) A_{\delta 0} + \frac{1}{4}\sum_{\delta=N_p+1}^{N-1} \Big[\rho'_{0\delta}\big(\frac{\pi-\psi}{2}\big) A_{\delta 0} + (1-\rho_{0\delta}) A_{\delta 0} \frac{1}{4\pi}\int_{4\pi} K_\delta(\mathbf{\Omega}', \mathbf{\Omega}'') P_{\delta 0}(\mathbf{\Omega}'', \mathbf{\Omega}) d\mathbf{\Omega}''\Big]. \tag{17.23}$$

The first terms on the right-hand sides of (17.22) and (17.23) give exactly the same contribution of opaque particles to the effective scattering coefficient and phase function as evaluated by ray optics [SiHo02].

17.4 Dense Dispersed Media with Opaque Particles

Let radiation be transferred in a continuous transparent or partially absorbing matrix denoted by index $\gamma = 0$ filled with opaque inclusions of phases $1, ..., N-$

1 satisfying condition (17.21). The total volume fraction of the inclusions can be considerable, so that condition (17.11) is not required. The absorption term $-\alpha_\gamma I_\gamma$ dominates the right-hand side of (17.1) for an opaque phase. Thus, $I_\gamma = 0$ for $\gamma = 0, \ldots, N-1$. Equation (17.1) for the matrix becomes

$$\mathbf{\Omega}\nabla I_0 = -\big(\alpha_0 + \frac{A_0}{4f_0}\big)I_0 + \frac{\rho_0 A_0}{4f_0}\frac{1}{4\pi}\int\limits_{4\pi} I_0(\mathbf{\Omega}')P_{00}(\mathbf{\Omega}',\mathbf{\Omega})d\mathbf{\Omega}'. \tag{17.24}$$

Thus, the problem is again reduced to the conventional RTE (17.24) with the effective extinction and scattering coefficients

$$\beta_e = \alpha_0 + \frac{A_0}{4f_0} = \alpha_0 + \frac{1}{4f_0}\sum_{\gamma=1}^{N-1} A_{\gamma 0}, \tag{17.25}$$

$$\sigma_e = \frac{\rho_0 A_0}{4f_0} = \frac{1}{4f_0}\sum_{\gamma=1}^{N-1} \rho_{0\gamma} A_{\gamma 0}, \tag{17.26}$$

respectively, and the scattering phase function

$$P_e(\mathbf{\Omega}',\mathbf{\Omega}) = P_{00}(\mathbf{\Omega}',\mathbf{\Omega}) = \frac{1}{\rho_0 A_0}\sum_{\gamma=1}^{N-1} \rho'_{0\gamma} A_{\gamma 0}. \tag{17.27}$$

According to (17.25)–(17.27) each opaque phase contributes proportionally to its specific surface. Similar equations were obtained in [GuKr05] by physical reasoning and validated by Monte Carlo ray tracing simulation.

17.4.1 Dependent Scattering

The obtained radiative properties are generally in line with the theory of independent scattering [SiHo02]. The only difference is the factor of

$$g = 1/f_0 \tag{17.28}$$

before sums in (17.25) and (17.26). The physical meaning is that specific surfaces $A_{\gamma 0}$ and A_0 should be referred not to the total volume of the media but to the volume occupied by the matrix phase. This seems to be natural because the radiation is transferred in the matrix only and does not penetrate into the opaque phases.

Factor (17.28) takes into account dependent scattering in the considered medium. It was initially introduced as the result of the analysis of Monte Carlo ray tracing simulation [SiKa92] and referred to as the scaling factor. The points in Figure 17.1(a) show the values of g given by numerical experiments in the cited work while the line is the analytical formula (17.28) resulting from the general multiphase model. The model agrees with the simulation and confirms that the principal effect is that the radiation does not penetrate into

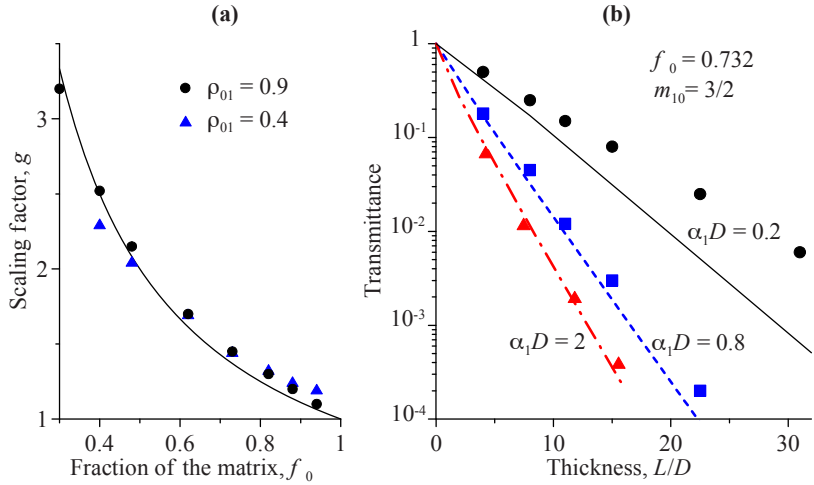


Fig. 17.1. Dependent scattering for opaque (a) and semi-transparent (b) spheres of diameter D in transparent matrix. The multiphase model (curves) and the Monte Carlo simulation (points [SiKa92]): (a), scaling factor g versus the volume fraction of the matrix phase f_0; (b), hemispherical transmittance versus the thickness L.

opaque phases. The theory is corrected for dependent scattering, so that the effective extinction and scattering coefficients are proportional to the scaling factor (17.28) and the effective scattering phase function does not change. The scaling factor is a strong function of the volume fraction of the matrix but is independent of the reflectivities of the opaque phases.

17.5 Dependent Scattering by (Semi)transparent Inclusions in a (Semi)transparent Matrix

In the general case when the matrix does not dominate in volume and the inclusions are not opaque, the full system of N equations (17.1) should be applied. Figure 17.1(b) compares the numerical solution of this system [Gu08] with the Monte Carlo simulation [SiKa92] for semi-transparent spheres of phase 1 of diameter D in transparent matrix 0. The hemispherical transmittance of a layer of this medium is plotted against its thickness L. A good agreement of the multiphase model with the Monte Carlo simulation is attained at the absorption parameters $\alpha_1 D$ of 0.8 and 2 while the multiphase model underestimates the hemispherical transmittance at $\alpha_1 D = 0.2$. The probable reason for this discrepancy could be the simplified boundary conditions used for the numerical solution of (17.1) [Gu08].

17.6 Conclusion

The multiphase model of radiation transfer (17.1) generalizes the equations obtained for a two-phase heterogeneous medium in [Gu08]. In the case of dilute dispersed systems, the multiphase model reduces to the conventional RTE. In the case of dense dispersed systems formed by opaque inclusions in a semi-transparent or transparent matrix, the multiphase model also reduces to the conventional RTE. It evaluates the effective radiative properties describing the dependent scattering. In the general case of transparent or semi-transparent inclusions in a transparent or semi-transparent matrix, the numerical solution of the model equations can explain the dependent scattering effects.

References

[SiHo02] Siegel, R., Howell, J.R.: *Thermal Radiation Heat Transfer,* Taylor & Francis, Washington, D.C. (2002).

[Pa05] Panasenko, G.: *Multi-scale Modelling for Structures and Composites,* Springer, Dordrecht (2005).

[BaSa00] Baillis, D., Sacadura, J.F.: Thermal radiation properties of dispersed media: theoretical prediction and experimental characterisation. *J. Quant. Spectrosc. Radiat. Transfer,* **67**, 327–363 (2000).

[Gu08] Gusarov, A.V.: Homogenization of radiation transfer in two-phase media with irregular phase boundaries. *Phys. Rev. B,* **77**, article 144201 (2008).

[To02] Torquato, S.: *Random Heterogeneous Materials,* Springer, New York (2002).

[ZeIaTa06] Zeghondy, B., Iacona, E., Taine, J.: Determination of the anisotropic radiative properties of a porous material by radiative distribution function identification (RDFI). *Internat. J. Heat Mass Transfer,* **49**, 2810–2819 (2006).

[GuKr05] Gusarov, A.V., Kruth, J.-P.: Modelling of radiation transfer in metallic powders at laser treatment. *Internat. J. Heat Mass Transfer,* **48**, 3423–3434 (2005).

[SiKa92] Singh, B.P., Kaviany, M.: Modelling radiative heat transfer in packed beds. *Internat. J. Heat Mass Transfer,* **35**, 1397–1405 (1992).

18

A Porous Finite Element Model of the Motion of the Spinal Cord

P.J. Harris[1] and C. Hardwidge[2]

[1] University of Brighton, UK; p.j.harris@brighton.ac.uk
[2] Hurstwood Park Neurological Unit, UK; carl.hardwidge@bsuh.nhs.uk

18.1 Introduction

The medical condition syringomyelia is characterized by the formation of large fluid-filled cavities in the spinal cord (called syrinxes in the medical literature). The exact mechanism by which these cavities form is not fully understood, although it has been theorized that changes in the pressure of the fluid surrounding the spinal cord could be responsible. There have been some studies carried out for the pressure levels in the cerebrospinal fluid [BW81], but none of these is over the time scales that are necessary to verify that the pressure changes are the cause of syringomyelia. Generally, detailed experimental data is needed over a period of months or even years in order to verify this hypothesis.

The alternative is to develop a mathematical model of the spinal cord and the surrounding liquid. A number of mathematical models of the spinal cord have been developed by applying various analytical and numerical methods to simulate the motion of the spinal cord and the surrounding liquid [CDB05, LEB06].

A more complete simulation of the whole cord was developed by Harris and Hardwidge [PJH07] by using a simple finite element model of the spinal cord based on the assumption that the cord had either linearly elastic or viscoelastic properties. However, this model did not take the permeable nature of the spinal cord into account, and so the internal loading of the pressure of the liquid inside the cord was missing. Further, as the cord was not permeable, it was not possible to include any changes to the pressure of the liquid in the central cavity or the consequential effects that such pressure changes would have had on the motion of the cord.

A more appropriate method is to treat the spinal cord as a porous medium saturated with the surrounding spinal fluid. This can be achieved by adapting existing mathematical models of porous media, developed in other areas of science and engineering (most notably soil mechanics) [RWL98], for use with biological and medical applications. The usual differential equations for

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_18,

the deformations of an elastic medium are modified to include a load term due to the pressure of the liquid inside the pores, and supplemented with an additional differential equation which can be solved for the pressure of this liquid. The resulting system of coupled partial differential equations can then be solved numerically using the finite element method.

18.2 Mathematical Model

Consider a liquid saturated porous medium, and let n denote the fluid-filled void fraction of the medium. Then the mass balance equation for the liquid phase is

$$\frac{\partial}{\partial t}\left[n\rho_l\right] + \nabla \cdot \left[n\rho_l \mathbf{v}_l\right] = 0, \tag{18.1}$$

where ρ_l is the density of the liquid and $\mathbf{v}_l$ is the velocity of the liquid phase. Assuming that the volume fraction of the liquid phase is constant, the liquid density does not vary in space, and the effects of the motion of the solid phase on the liquid phase can be neglected, then Darcy's law can be used to rewrite (18.1) as

$$\frac{\partial \rho_l}{\partial t} + \frac{\rho_l}{\nu_l}\nabla \cdot \left(-\kappa \nabla p\right) = 0, \tag{18.2}$$

where p is the excess pressure in the liquid phase, κ is the permeability, and ν_l is the viscosity of the liquid. If the excess pressure and the density are related by an equation of the form

$$p = \frac{h}{\rho_0}\left(\rho_l - \rho_0\right),$$

where ρ_0 is the density of the liquid phase when it is at rest and h is the bulk modulus of the liquid, then (18.2) leads to the linear diffusion equation

$$\frac{\partial p}{\partial t} = \mu \nabla^2 p, \tag{18.3}$$

where the diffusion constant μ is given by

$$\mu = \frac{h\kappa}{\nu_l}.$$

Now consider the stress in the porous medium. Using the same notation as is frequently used with the finite element method for stress analysis, we let σ denote the vector of the nonzero components of the symmetric stress tensor. This can be expressed as a linear combination of the contributions σ_l and σ_s from the liquid and solid phases, respectively, in the form

$$\sigma = (1-n)\sigma_s + n\sigma_l.$$

Assuming that the liquid phase is inviscid, the stress in the liquid phase is related to the pressure by

$$\sigma_l = -\mathbf{m}p,$$

where, for an axisymmetric problem, $\mathbf{m} = [1, 1, 1, 0]^T$. In the solid phase the usual linear stress–strain and strain–displacement relationships can be used to give the total stress as

$$\sigma = (1-n)DB\mathbf{u} + n\mathbf{m}p,$$

where D is the stress–strain matrix and B the strain–displacement matrix, as given in [OCZ91]. Assuming that there are no body forces acting on the solid phase, the equation of motion for the solid phase can be expressed as

$$-\nabla \cdot \sigma = [(1-n)\rho_s + n\rho_l]\,\frac{\partial^2 \mathbf{u}}{\partial t^2} + \mu_s \frac{\partial \mathbf{u}}{\partial t}, \tag{18.4}$$

where ρ_s is the density of the solid phase and μ_s is the damping coefficient for the solid phase.

Applying the finite element method to equations (18.3) and (18.4) yields the coupled matrix system of equations

$$\begin{aligned} &M\ddot{\mathbf{u}} + C\dot{\mathbf{u}} + (1-n)K\mathbf{u} = [(1-n)L + nQ_0]\,\mathbf{p}_0 + nQ\mathbf{p}, \\ &S\dot{\mathbf{p}} - H\mathbf{p} = -S_0\dot{\mathbf{p}}_0 + H_0\mathbf{p}_0, \end{aligned} \tag{18.5}$$

where M, C, and K are the mass, damping, and stiffness matrices given by

$$\begin{aligned} M &= [(1-n)\rho_s + n\rho_l]\int_V N^T N\,dv, \\ C &= \mu_s \int_V N^T N\,dv, \\ K &= \int_V B^T DB\,dv, \end{aligned}$$

respectively. The matrix N is the usual matrix constructed from the finite element basis functions $\{\phi_i\}$ (see [OCZ91] for further details). The fluid phase finite element matrices S and H are given by

$$\begin{aligned} S_{ij} &= \int_V \phi_i \phi_j\,dv, \\ H_{ij} &= \mu \int_V \nabla\phi_i \cdot \nabla\phi_j\,dv. \end{aligned}$$

The coupling matrices Q and L can be expressed as block matrices, where each appropriately sized block is given by

$$Q_{ij} = \int_V B_i^T \mathbf{m}\phi_j\,dv$$

and

$$L_{ij} = \int_S \phi_i \mathbf{n}\, dS,$$

respectively. Here S denotes the part of the surface of the porous medium where the pressure is known, and $\mathbf{n}$ is the unit normal vector to S directed into the porous medium. The subscript 0 is used to denote quantities that are at nodes where the pressure is known or is given by the boundary conditions.

The coupled finite element equations given by (18.5) can be expressed as

$$\begin{bmatrix} M & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & S \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}} \\ \dot{\mathbf{u}} \\ \dot{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} -C & -(1-n)K & nQ \\ I & 0 & 0 \\ 0 & 0 & H \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} + \begin{bmatrix} [(1-n)L + nQ_0]\,\mathbf{p}_0 \\ 0 \\ H_0\mathbf{p}_0 - S_0\dot{\mathbf{p}}_0 \end{bmatrix}, \tag{18.6}$$

where $\mathbf{v} = \dot{\mathbf{u}}$. Clearly, (18.6) forms a linear system of coupled differential equations in time which can be written in the form

$$A_0\dot{\mathbf{y}} = A_1\mathbf{y} + \mathbf{f}, \tag{18.7}$$

where

$$A_0 = \begin{bmatrix} M & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & S \end{bmatrix}, \; A_1 = \begin{bmatrix} -C & -(1-n)K & nQ \\ I & 0 & 0 \\ 0 & 0 & H \end{bmatrix},$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{v} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} [(1-n)L + nQ_0]\,\mathbf{p}_0 \\ 0 \\ H_0\mathbf{p}_0 - S_0\dot{\mathbf{p}}_0 \end{bmatrix}.$$

The system (18.7) can be integrated through time analytically. However, this is computationally expensive as it requires us to calculate all the eigenvalues and eigenvectors of the generalized eigenvalue problem $A_0\mathbf{y} = \lambda A_1\mathbf{y}$. In addition, we would also be required to calculate integrals involving exponential functions of the eigenvalues multiplied by the terms in the vector $\mathbf{f}$, which may only be possible with the use of numerical methods.

The alternative is to use a numerical method to integrate the system (18.7) through time. The method used here is based on the trapezium method. Let $\mathbf{y}_j$ denote the solution to (18.7) at the jth time step, and let δt denote the length of the time step. Then, at the time halfway between time steps j and $j+1$, we have

$$\dot{\mathbf{y}} \approx \frac{\mathbf{y}_{j+1} - \mathbf{y}_j}{\delta t}, \quad \mathbf{y} \approx \frac{\mathbf{y}_{j+1} + \mathbf{y}_j}{2}. \tag{18.8}$$

Substituting (18.8) into (18.7) and rearranging gives

$$(2A_0 - \delta t A_1)\,\mathbf{y}_{j+1} = (2A_0 + \delta t A_1)\,\mathbf{y}_j + 2\delta t \mathbf{f}_{j+1/2}, \tag{18.9}$$

which can be solved for $\mathbf{y}_{j+1}$. Note that the notation $\mathbf{f}_{j+1/2}$ is used to denote that $\mathbf{f}$ should be evaluated at the time halfway between the jth and $(j+1)$th time steps. Since $\mathbf{y}_0$ can be found from the initial conditions and $\mathbf{f}_{j+1/2}$ can be calculated from the boundary conditions, it is possible to use (18.9) to numerically integrate the system through time and calculate the approximate solution at a later time step.

For a solid cord, we can show that this method is stable by showing that the eigenvalues of the iteration matrix are always less than or equal to one in magnitude. Rewrite the system of equations (18.9) as

$$\mathbf{y}_{j+1} = \left(I - \frac{\delta t}{2} A_0^{-1} A_1\right)^{-1} \left[\left(I + \frac{\delta t}{2} A_0^{-1} A_1\right) \mathbf{y}_j + A_0^{-1} \mathbf{f}_{j+1/2}\right].$$

If λ is an eigenvalue of $A_0^{-1} A_1$, then

$$\frac{1 + \frac{\delta t}{2}\lambda}{1 - \frac{\delta t}{2}\lambda} \tag{18.10}$$

is an eigenvalue of the trapezium rule iteration matrix. It can be shown that the eigenvalues of $A_0^{-1} A_1$ are either

$$\frac{-\mu_s \pm \sqrt{\mu_s^2 - 4\omega^2}}{2},$$

where ω is a natural frequency of the structural phase, or an eigenvalue of the diffusion equation of the liquid phase (which are all real and negative or zero). In either case, the real part of each eigenvalue is either zero or negative, and so the magnitude of the numerator in (18.10) is less than or equal to the magnitude of the denominator. Hence, all the eigenvalues of the iteration matrix are less than or equal to one in magnitude, and the above trapezium method is stable.

18.3 Numerical Results

The results presented here are for a section of the spinal cord 5cm long and of radius 0.5cm. An axisymmetric finite element model is used to reduce the size of the computational model. The two cases being considered are for a solid section of cord, and a section of cord with an elliptical-shaped cavity in the centre which has vertical axis of length 2cm and horizontal axis of length 0.2cm. In each case two meshes (coarse and fine) of quadratically curved triangular elements [OCZ91] will be used in the calculations. The coarse meshes have 500 solid phase elements and 1111 nodes; and the fine meshes have 2000 solid phase elements and 4221 nodes. For the solid cord, the nodes are equally spaced both horizontally and vertically, but for the cord with the cavity, the mesh is graded so that there are smaller elements close to the top and bottom

of the cavity to take into account the known problems with computing the stresses at such points.

The material parameters of the model were chosen as follows. A value of $10^6 \mathrm{Nm}^2$ is used for Young's modulus as this is the value determined experimentally by Bilston and Thibault [LEB96]. Poisson's ratio is set to 0.49 as the cord is thought to be almost incompressible, and the density is taken to be $1100 \mathrm{kgm}^{-3}$, slightly denser than water. The cerebrospinal fluid is assumed to be essentially water with density $1000 \mathrm{kgm}^{-3}$. The void fraction is set to 0.18, and a range of different values are used for the diffusion parameter μ in (18.3). The different values of the diffusion parameter correspond to different values of the permeability of the spinal cord. The exterior pressure loading applied to the outside of the cord is constant in space, and its temporal variation is shown in Figure 18.1.

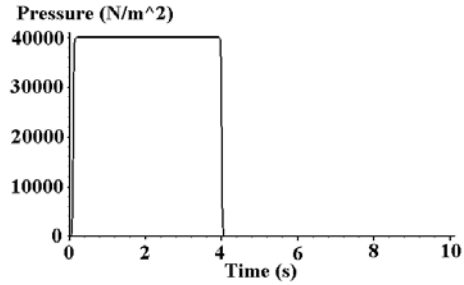


Fig. 18.1. The pressure loading applied to the outer surfaces of the spinal cord.

The main results presented here will be for computing the mean stress at a point of interest inside the spinal cord. The mean stress is simply the mean of the three components of the normal stress, and it can be shown that this is a stress invariant. That is, the mean stress does not depend on the coordinate system being used, or on the orientation of that system.

Figure 18.2 shows the results of computing the mean stress in the solid phase at the centre of the solid cord with $\mu = 5 \times 10^{-6}$ using both the coarse 500 element mesh and the fine 2000 element mesh. In this case the two curves on the graph are superimposed, indicating that the two meshes are giving almost identical results. The corresponding results for a cord with a cavity, where the mean stress has been calculated at one end of the cavity, are given in Figure 18.3. The results given in these figures show that the finite element method is yielding accurate results for this problem, since refining the mesh and time steps does not significantly change the calculated stress. Therefore, all the remaining calculations will only be carried out using the meshes with 2000 elements.

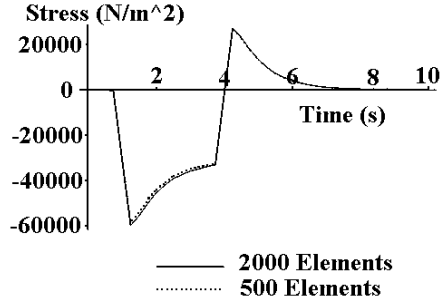


Fig. 18.2. A comparison of the calculated mean stress at the centre of the solid cord using both the 500 and 2000 element meshes.

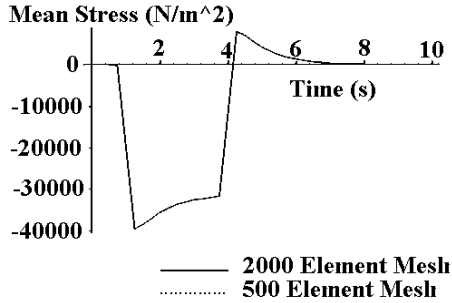


Fig. 18.3. A comparison of the calculated mean stress at one end of the cavity in the cord using both the 500 and 2000 element meshes.

Figure 18.4 shows the pressure in the liquid phase (top) and mean stress in the solid phase (bottom) for four different values of the diffusion parameter μ. As can be seen, if the diffusion constant is relatively large (5×10^{-5}), then the pressure changes travel through the cord almost instantaneously, whereas if the diffusion constant is relatively small (10^{-6}), then the peak pressure in the liquid phase is smaller than the peak applied pressure. This has a significant effect on the peak tensile stress. These results show that the tensile stress is maximized for the values of the diffusion constant such that the pressure in the liquid phase reaches its maximum value at the center of the cord just as the external pressure loading is removed. Figure 18.5 shows the corresponding results for the cord with a cavity, where the mean stress has been calculated at one end of the cavity, and we note that the peak tensile stress is much greater in this case.

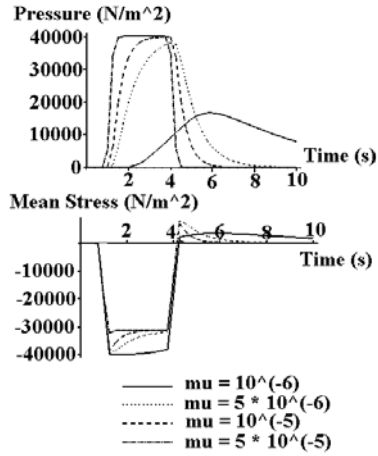


Fig. 18.4. A comparison of the calculated pressure in the liquid phase (top) and mean stress in the solid phase (bottom) at the centre of a solid cord for different values of the liquid phase diffusion parameter.

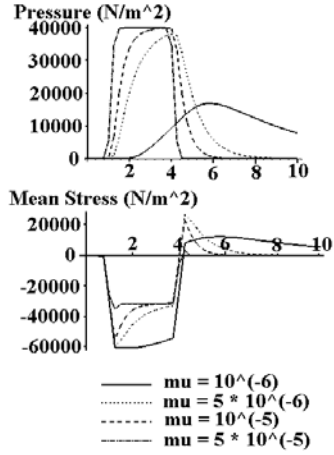


Fig. 18.5. A comparison of the calculated pressure in the liquid phase (top) and mean stress in the solid phase (bottom) at the bottom of the cavity for different values of the liquid phase diffusion parameter.

18.4 Conclusions

These results show that the finite element method can be used to accurately model the deformations of the spinal cord. Clearly, the value of the diffusion parameter μ appearing in (18.3) plays an important role in determining the behavior of the cord, as it controls how fast the external liquid pressure

changes are transmitted through the liquid phase of the cord and hence has effect on the mean stress in the centre of the cord. However, currently there is no information on what is an appropriate value for the diffusion parameter, and some further experimental work is needed to resolve this issue.

The higher mean stresses obtained in the case where the cord already has a cavity shows that once the cavities have formed in a patient they are likely to get bigger, and this has been observed in real patients.

The results presented in this chapter demonstrate that the hypothesis that syringomyelia is caused by physical fluid mechanics processes in the spine is credible. Further work is needed to modify the model to be able to simulate the actual formation and growth of the cavities in the spinal cord.

References

[CDB05] Bertram, C.D., Brodbelt, A.R, Stoodley, M.A.: The origins of syringomyelia: numerical models of fluid/structure interactions in the spinal cord. *J. Biomech. Engng.,* **127**, 1099–1109 (2005).

[LEB96] Bilston, L.E., Thibault, L.E.: The mechanical properties of the human cervical spinal cord in vitro. *Biomed. Engng. Soc.,* **24**, 67–74 (1996).

[LEB06] Bilston, L.E., Fletcher, D.F., Stoodley, M.A.: Focal spinal arachnoiditis increases subarachnoid space pressure: a computational study. *Clinical Biomech.,* **21**, 579–584 (2006).

[PJH07] Harris, P.J., Hardwidge, C.: The mathematical modelling of syringomyelia, in *Integral Methods in Science and Engineering. Techniques and Applications,* Constanda, C., Potapenko, S., eds., Birkhäuser, Boston, MA (2007).

[RWL98] Lewis, R.W., Schrefler, B.A.: *The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media,* 2nd ed., Wiley, Chichester (1998).

[BW81] Williams, B.: Simultaneous cerebral and spinal fluid recording: 2. Cerebrospinal dissociation with lesions at the foramen magnum. *Acta Neuro.,* **59**, 123–142 (1981).

[OCZ91] Zienkiewicz, O.C., Taylor, R.L.: *The Finite Element Method. Vols. 1 and 2,* McGraw-Hill, London (1991).

19

Boundary Hybrid Galerkin Method for Elliptic and Wave Propagation Problems in $\mathbb{R}^3$ over Planar Structures

C. Jerez-Hanckes[1] and J.-C. Nédélec[2]

[1] ETH Zürich, Switzerland; cjerez@math.ethz.ch
[2] École Polytechnique, Palaiseau, France; nedelec@cmapx.polytechnique.fr

19.1 Motivation

Consider a flat smooth manifold $\Gamma_m \subset \mathbb{R}^3$ of codimension one with Lipschitz boundary $\partial\Gamma_m$ and large aspect ratios such as the one depicted in Figure 19.1. Let the associated unbounded domain $\Omega := \mathbb{R}^3 \setminus \bar{\Gamma}_m$ be isotropic and homogeneous for the moment. We seek solutions $u \in H^1_{loc}(\Omega)$ of the Laplace and Helmholtz equations when a Dirichlet condition g_D is applied on Γ_m such that

$$\gamma_D^+ u\big|_{\Gamma_m} \;=\; \gamma_D^- u\big|_{\Gamma_m} \;=\; g_D \;\in\; H^{1/2}(\Gamma_m),$$

where $\gamma_D^{\pm}$ are the Dirichlet trace operators from either side of Γ_m. If $[\,\cdot\,]_{\Gamma_m}$ denotes the jump across Γ_m, clearly $[\gamma_D u]_{\Gamma_m} = 0$. Thus, solutions over Ω can be built [Mc00] via the *single-layer potential* $\boldsymbol{\Psi}^k_{SL}$, i.e.,

$$u(\mathbf{x}) \;=\; -\boldsymbol{\Psi}^k_{SL}\left([\gamma_N u]_{\Gamma_m}\right)(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Omega, \tag{19.1}$$

where

$$\boldsymbol{\Psi}^k_{SL}(\varphi)(\mathbf{x}) := \int_{\Gamma_m} G_k(\mathbf{x}-\mathbf{y})\,\varphi(\mathbf{y})d\mathbf{y} \qquad \text{for} \quad \mathbf{x} \in \Omega,$$

γ_N is the Neumann trace operator, and the integral kernel G_k takes the form

$$G_k(\mathbf{z}) \;=\; \frac{1}{4\pi}\frac{\exp\left(\imath k|\mathbf{z}|\right)}{|\mathbf{z}|} \quad \text{for } k \in \mathbb{R}, \tag{19.2}$$

being the associated fundamental solution of the differential equation.

Thus, we reduce the problem to that of finding the Neumann trace jump in (19.1), henceforth denoted $\sigma := [\gamma_N u]_{\Gamma_m}$, and which must lie in $\widetilde{H}^{-1/2}(\Gamma_m)$ by regularity of solutions over Ω. Upon taking the Dirichlet trace over Γ_m, one obtains a Fredholm integral equation of the first kind:

$$-\boldsymbol{V}_k(\sigma)(\mathbf{x}) \;=\; g_D(\mathbf{x}) \quad \text{for } \mathbf{x} \in \Gamma_m, \tag{19.3}$$

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_19,

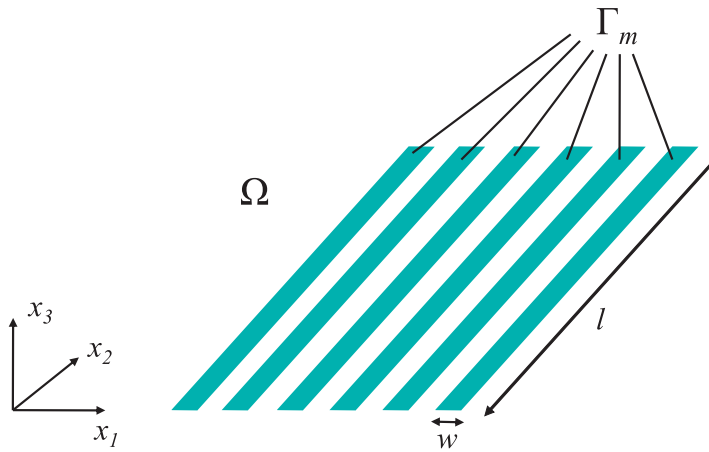


Fig. 19.1. Model geometry schematics and coordinates definition. Notice that Γ_m may not necessarily be connected. Furthermore, we assume $l \gg w$.

where

$$\boldsymbol{V}_k(\varphi)(\mathbf{x}) \;:=\; \gamma_D \circ \boldsymbol{\Psi}_{SL}^k\,(\varphi)\,(\mathbf{x}) \;=\; \int_{\Gamma_m} G_k(\mathbf{x}-\mathbf{y})\,\varphi(\mathbf{y})d\mathbf{y}. \tag{19.4}$$

This situation can be encountered under the categories of screen, crack, or interface—when the surface containing Γ_m lies between two different materials—problems for which solutions are known to possess singular behaviors (see [St87], [NiSa94], [CoDa02], and [Gr85]). Although several methods have been successfully proposed to handle such singularities, in our case the very elongated form of Γ_m renders them impractical. We overcome this by developing an augmented bases approach that takes into account the adequate boundary singularity. More precisely:

1. At edges, the boundary integral operator is turned into a compactly perturbed logarithmic singular integral operator for the transverse edge coordinate. We recall that weighted first-kind Chebyshev polynomials are shown to constitute an optimal discretization base.
2. At corners, the problem is reminiscent of that of finding the charge singularity of perfect conductor sectors [Ke99], [MoLe76]. In our case, we use first-order polynomials over anisotropically graded meshes that follow singular coefficients.

This scheme was already presented in [JLNL08], and it is the purpose here to provide a mathematical framework for its analysis.

19.2 Integral Operators with Logarithmic Kernels in $\mathbb{R}^2$

Let us first focus on the line segment $\Gamma_m = (a,b) \times \{0\} \in \mathbb{R}^2$ with bounded $a, b \in \mathbb{R}$. The associated Laplacian and Helmholtz Green's function becomes

$$G_k(\mathbf{z}) \;=\; \begin{cases} \dfrac{\imath}{4} H_0^{(1)}\left(k|\mathbf{z}|\right) & \text{for} \quad k \neq 0, \\ -\dfrac{1}{2\pi} \log|\mathbf{z}| & \text{for} \quad k = 0, \end{cases} \tag{19.5}$$

where $H_0^{(1)}$ is the Hankel function of the first kind and which behaves as G_0 for small arguments. Hence, the boundary integral operator (19.4) in $\mathbb{R}^2$ can be written as $\boldsymbol{V}_k = \boldsymbol{L} + \boldsymbol{K}_k$ where, in view of Γ_m,

$$\boldsymbol{L}(\varphi)(x_1, 0) \;:=\; \frac{1}{2\pi} \int_a^b \log \frac{1}{|x_1 - y_1|}\, \varphi(y_1)\, dy_1 \quad \text{for } x_1 \in (a,b) \tag{19.6}$$

and $\boldsymbol{K}_k$ is compact or null for k zero [SSC00].

Proposition 1. *The operator $\boldsymbol{L} : \widetilde{H}^{-1/2}(\Gamma_m) \to H^{1/2}(\Gamma_m)$ is a bounded Fredholm operator of index zero and the Gårding-type inequality*

$$\left((\boldsymbol{L} + \boldsymbol{K}_k)\, \varphi\,,\, \varphi\right)_{L^2(\Gamma_m)} \;\geq\; \gamma\, \|\varphi\|_{H^{-1/2}(\Gamma_m)} \tag{19.7}$$

holds. Moreover, if ϱ denotes the distance towards the endpoints of Γ_m, the solutions φ of (19.3) behave as $\mathcal{O}\left(\varrho^{-1/2}\right)$.

Remark 1. If Γ_m is in fact a Jordan curve, then a suitable parametrization can render a logarithmic integral operator plus compact perturbations and the above results still hold with different γ.

If we define

$$\widetilde{H}_0^{-1/2}(\Gamma_m) = \left\{ f \in \widetilde{H}^{-1/2}(\Gamma_m) \,:\, \langle f\,,\, 1\rangle = 0 \right\},$$
$$H_*^{1/2}(\Gamma_m) \;=\; \left\{ f \,\in\, H^{1/2}(\Gamma_m) \,:\, \int_{\Gamma_m} \frac{f(t)\, dt}{\varrho(t)} \;=\; 0 \right\},$$

a refinement of the above is stated as follows.

Proposition 2. *Operator $\boldsymbol{L}$ is bijective between $\widetilde{H}_0^{-1/2}(\Gamma_m)$ and $H_*^{1/2}(\Gamma_m)$.*

19.2.1 Logarithmic Operators in Weighted L^2-Spaces

Let $T_n(\zeta)$ and $U_n(\zeta)$ denote the Chebyshev polynomials of first and second kinds, respectively, defined by the relations

$$T_n(\zeta) \;=\; \cos n\theta, \quad U_n(\zeta) \;=\; \frac{\sin(n+1)\,\theta}{\sin\theta} \quad \text{with} \quad \zeta \;=\; \cos\theta$$

with real values over $[-1,1]$. For the same interval, we define the weight function $w(\zeta) := (1-\zeta^2)^{1/2}$. Then, the T_n are orthogonal with respect to w^{-1}:

$$\int_{-1}^{1} T_n(\zeta)\, T_m(\zeta)\, w^{-1}(\zeta)\, d\zeta \;=\; \begin{cases} 0 & n \neq m, \\ \pi/2 & n = m \neq 0, \\ \pi & n = m = 0, \end{cases} \tag{19.8}$$

while the U_n satisfy

$$\int_{-1}^{1} U_n(\zeta)\, U_m(\zeta)\, w(\zeta)\, d\zeta \;=\; \begin{cases} 0 & n \neq m, \\ \pi/2 & n = m \neq 0. \end{cases} \tag{19.9}$$

On the other hand, we introduce the weighted Hilbert spaces

$$\begin{aligned} L^2_{1/w} &:= \left\{ f \,:\, \|f\|^2_{1/w} := \int_{-1}^{1} |f(\zeta)|^2\, w^{-1}(\zeta)\, d\zeta \,<\, \infty \right\}, \\ L^2_{w} &:= \left\{ f \,:\, \|f\|^2_{w} := \int_{-1}^{1} |f(\zeta)|^2\, w(\zeta)\, d\zeta \,<\, \infty \right\}, \\ W &:= \left\{ f \,:\, f \in L^2_{1/w} \,:\, f' \in L^2_{w} \right\}, \end{aligned}$$

endowed with the obvious scalar products to induce the norms $\|\cdot\|_w$, $\|\cdot\|_{1/w}$ and $\|\cdot\|_W$, this last being the associated graph norm.

Proposition 3. *Let $\boldsymbol{W}\varphi := w\varphi$ and $\boldsymbol{W}^{-1}\varphi = w^{-1}\varphi$. Then we have the isometries*

$$\boldsymbol{W} \,:\, L^2_w \longrightarrow L^2_{1/w}, \quad \boldsymbol{W}^{-1} \,:\, L^2_{1/w} \longrightarrow L^2_w, \tag{19.10}$$

and there is a continuous inclusion $L^2_{1/w} \subset L^2_w$.

Proposition 4. *For a given $\zeta \in [-1,1]$, the logarithmic kernel admits the Chebyshev polynomial expansion*

$$\log \frac{1}{|\zeta-\eta|} \;=\; \log 2 \,+\, \sum_{n=1}^{\infty} \frac{2}{n} T_n(\zeta) T_n(\eta) \quad \forall\, \eta \in [-1,1] \tag{19.11}$$

as a function in $L^2_{1/w}$.

Now, without loss of generality (see Remark 1), consider the endpoints of Γ_m to lie at ± 1. Then, one can define the modified logarithmic integral operator $\boldsymbol{L}_{1/w} := \boldsymbol{L} \circ \boldsymbol{W}^{-1}$, i.e.,

$$\boldsymbol{L}_{1/w}(\varphi)(\zeta) \;=\; \frac{1}{2\pi} \int_{-1}^{1} \log \frac{1}{|\zeta-\eta|}\, \frac{\varphi(\eta)}{\sqrt{1-\eta^2}}\, d\eta \quad \text{for} \quad \zeta \in (-1,1).$$

Proposition 5. *The operator* $\boldsymbol{L}_{1/w} : L^2_{1/w} \to W$ *is bounded and continuously invertible. If* $f \in W$*, the unique solution of the integral equation with purely logarithmic kernel normalized on the interval* $[-1, 1]$ *is given by*

$$\varphi(x) = \frac{f_0}{\log 2} T_0(x) + 2 \sum_{n=1}^{\infty} n f_n T_n(x), \tag{19.12}$$

where the coefficients f_n *are given by*

$$f_n = \frac{2}{\pi} (f, T_n)_{1/w}, \quad n \in \mathbb{N}_0.$$

By using these two last propositions, we can easily derive the following result.

Corollary 1. *The original logarithmic singular operator* $\boldsymbol{L} : L^2_w \to W$ *is bounded and continuously invertible, i.e., it is a zero-index Fredholm operator.*

19.2.2 An Adapted Spectral Boundary Element Method

In what follows, we link the results over Sobolev and weighted spaces. Construct the approximation spaces:

$$\mathbb{Q}_N(\Gamma_m) = \operatorname{span}\left\{ w_m^{-1}(t) T_n(t_m) \right\}_{n=1}^{N}, \tag{19.13}$$

where, for $t \in [a, b]$, we have defined

$$t_m := \frac{2(t - t_m^c)}{b - a} \quad \text{with} \quad t_m^c := \frac{a + b}{2}$$

and $w_m(t) := \sqrt{(b - t)(t - a)}$. In the case $a = -1$, $b = 1$, this space belongs to L^2_w by Proposition 3. The idea is to describe φ through the truncated expansion

$$\varphi_N(t) = \sum_{n=0}^{N} \varphi_n \, w^{-1}(t) \, T_n(t). \tag{19.14}$$

We now show that $\mathbb{Q}_N(\Gamma_m)$ satisfies the approximation property over the associated Sobolev space.

Lemma 1. *The following properties hold:*

1. *The space* $\mathbb{Q}_N(\Gamma_m)$ *is a closed subspace of* $\widetilde{H}^{-1/2}(\Gamma_m)$*;*
2. $\mathbb{Q}_\infty(\Gamma_m) = \lim_{N \to \infty} \mathbb{Q}_N(\Gamma_m)$ *is dense in* $\widetilde{H}^{-1/2}(\Gamma_m)$*.*

Proposition 6. *Let* $g_D \in H_*^{1/2}(\Gamma_m)$*. With the Galerkin variational formulation for* $\varphi_N \in \mathbb{Q}_N(\Gamma_m)$*,*

$$\left\langle \boldsymbol{L}\varphi_N , \varphi_N^t \right\rangle_{\Gamma_m} = \left\langle g_D , \varphi_N^t \right\rangle_{\Gamma_m} \quad \forall \, \varphi_N^t \in \mathbb{Q}_N(\Gamma_m), \tag{19.15}$$

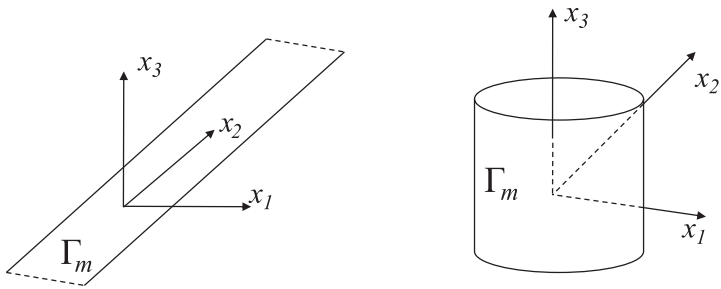


Fig. 19.2. Edge geometry (left) and cylinder geometry (right).

the following stability condition and error bound hold:

$$\|\varphi_N\|_{\widetilde{H}^{-1/2}(\Gamma_m)} \;\leq\; C\,\|g_D\|_{H^{1/2}_*(\Gamma_m)}\,,$$
$$\|\varphi - \varphi_N\|_{\widetilde{H}^{-1/2}_0(\Gamma_m)} \;\leq\; C \inf_{\psi_N \in \mathbb{Q}_N(\Gamma_m)} \|\varphi - \psi_N\|_{\widetilde{H}^{-1/2}_0(\Gamma_m)}\,.$$

Remark 2. Extension to compactly perturbed operators is achieved by Fredholmness. Thus, if besides the principal logarithmic term, continuous functions are introduced in the kernel, the solution scheme remains stable. This will prove to be the case in the upcoming applications.

19.3 Localization of Single-Layer Operators in $\mathbb{R}^3$

Before considering our initial problem (Section 19.1), we study two geometric configurations in $\mathbb{R}^3$ (Figure 19.2) for which the solution scheme requires the solution of logarithmic integral equations with continuous perturbations. Thus, sets of weighted Chebyshev polynomials of the first kind $\mathbb{Q}_N$ can provide suitable approximation spaces.

19.3.1 Flat Edge Problem

Let us consider an infinitely long, perfectly conducting strip with zero thickness and finite width inside an isotropic material in $\mathbb{R}^3$. Without loss of generality, we define the metallized domain

$$\Gamma_m \;=\; (-1,1)\times\mathbb{R}\times\{0\}.$$

From (19.4) and (19.2), we can also write $\boldsymbol{V}_k = \boldsymbol{T} + \boldsymbol{K}_k$, where

$$\boldsymbol{T}(\sigma)(\mathbf{x}) \;:=\; \frac{1}{4\pi}\int_{\Gamma_m} \frac{1}{|\mathbf{x}-\mathbf{y}|}\,\sigma(\mathbf{y})\,d\mathbf{y}, \tag{19.16a}$$

$$\boldsymbol{C}_k(\sigma)(\mathbf{x}) \;:=\; \frac{1}{4\pi}\int_{\Gamma_m} \frac{1-e^{\imath k|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|}\,\sigma(\mathbf{y})\,d\mathbf{y}, \tag{19.16b}$$

for $\mathbf{x} \in \Gamma_m$. The integral kernel in $\boldsymbol{T}$ is singular while the one in $\boldsymbol{C}_k$ is continuous for all $\mathbf{x}$ and $\mathbf{y}$. However, the domain is unbounded and, consequently, compactness arguments do not hold for the latter operators as in $\mathbb{R}^2$.

Proposition 7. *Let $g_D \in H^{1/2}(\Gamma_m)$ and consider the boundary integral equation*

$$\boldsymbol{T}(\sigma)(\mathbf{x}) \;=\; g_D(\mathbf{x}), \quad \mathbf{x} \in \Gamma_m. \tag{19.17}$$

Then the partial Fourier transform of the density σ along the edge direction x_2 is obtained as the solution of the logarithmic singular integral equation

$$\frac{1}{2}\int_{-1}^{1} K_0(|(x_1 - y_1)\xi|)\,\widehat{\sigma}(y_1,\xi)\,dy_1 \;=\; \widehat{g}(x_1,\xi) \quad \forall\, x_1 \in [-1,1], \tag{19.18}$$

where K_0 is the modified Bessel function of the second kind.

Proof. We start from the integral equation (19.16a). We take the partial Fourier transform along the edge axis, i.e., x_2. This yields

$$\frac{1}{4\pi\sqrt{2\pi}}\int_{\Gamma_m}\int_{\mathbb{R}} \frac{e^{\imath\xi x_2}}{\sqrt{(x_1-y_1)^2 \,+\, (x_2-y_2)^2}}\,\sigma(y_1,y_2)\,dx_2 dy_1 dy_2 \;=\; \widehat{g}(x_1,\xi),$$

where we have interchanged integrals formally. Now, applying the variable change $(x_1 - y_1)\zeta \;=\; x_3 \,-\, y_3$ and the identity [GrRy94]

$$K_0(z) \;=\; \frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{\imath z\zeta}}{\sqrt{1+\zeta^2}}\,d\zeta,$$

we finally obtain (19.18). Now, for a small argument, K_0 behaves as a logarithm while for large real z, it decreases exponentially. Thus, the integral operator is Fredholm as the kernel is continuous, weakly singular at the logarithm and exponentially decreasing along the unbounded direction.

Remark 3. Thus, we can take the variational form using $\mathbb{Q}_N([-1,1])$ as in Proposition 6 to solve equation (19.18).

19.3.2 Bounded Cylindrical Screen

Consider the following cylindrical screen of radius $\rho_0 \;>\; 0$, centered at the origin in an isotropic three-dimensional space:

$$\Gamma_m \;=\; \left\{\mathbf{x} \in \mathbb{R}^3 \;:\; \sqrt{x_1^2 \,+\, x_2^2} \;=\; \rho_0\,,\; |x_3| \;<\; 1\right\}.$$

We introduce cylindrical coordinates (ρ, φ, x_3) such that

$$x_1 \;=\; \rho\,\cos\varphi, \quad x_2 \;=\; \rho\,\sin\varphi, \quad x_3 \;=\; x_3,$$

with domains $\rho \in [0,\infty)$, $\varphi \in [0, 2\pi)$, and $x_3 \in \mathbb{R}$.

Proposition 8. *The solution of (19.17) is given by the Fourier expansion in the angular variable*

$$\sigma(\mathbf{y}) \;=\; \frac{1}{\sqrt{2\pi}} \sum_{m\,\in\,\mathbb{Z}} \sigma_m(\rho_0, y_3)\, e^{-\imath m\varphi}, \tag{19.19}$$

where each coefficient $\sigma_m(\rho_0, y_3)$ is the solution of the logarithmic singular integral equation

$$\frac{(-1)^m}{2\pi\rho_0} \int_{-1}^{1} \boldsymbol{Q}_{|m|-1/2} \left(\frac{1}{2} \left[\frac{x_3 - y_3}{\rho_0} \right]^2 + 1 \right) \sigma_m(\rho_0, y_3)\, dy_3 \;=\; g_m(\rho_0, x_3),$$

where $g_m(\rho_0, x_3)$ is the m-angular Fourier coefficient of g_D and $\boldsymbol{Q}_\nu$ denotes the Legendre function of second kind [AbSt72].

When the argument is close to one, $\boldsymbol{Q}_\nu$ non-integer index ν has a logarithmic singularity. In our case, for $y_3 \to x_3$, the argument in $\boldsymbol{Q}_{|m|-1/2}$ becomes

$$\boldsymbol{Q}_{|m|-1/2} \left(\frac{1}{2} \left[\frac{x_3 - y_3}{\rho_0} \right]^2 + 1 \right) = \log \frac{\rho_0}{|x_3 - y_3|} + C_{|m|} + \mathcal{O}\big((x_3 - y_3)^2\big)\,,$$

where $C_{|m|}$ is a bounded constant depending on $|m|$. Since $\boldsymbol{Q}_\nu$ is continuous elsewhere [AbSt72] and the integration domain $[-1, 1]$ is bounded, we conclude that the coefficients $\sigma_m(\rho_0, y_3)$ are solutions of compactly perturbed logarithmic integral equations depending on m and g_m.

Remark 4. Notice the logarithmic singular behavior as ρ_0 goes to zero. This is consistent with the knowledge that for filaments the solutions are singular as a logarithm along the radial direction.

19.4 Hybrid Element Description

We finally analyze surfaces possessing corners. Consider a domain similar to the infinite strip but with bounded length l along x_2 (Figure 19.3.) More precisely,

$$\Gamma_m \;=\; (-1, 1) \times (-l/2, l/2) \times \{0\},$$

with $l \gg 2$. Let d satisfy $0 < d < l/2$, and introduce subdomains

$$\Gamma_m^d \;:=\; (-1, 1) \times (-d/2, d/2) \times \{0\} \subsetneq \Gamma_m$$

and $\Gamma_m^f := \Gamma_m \setminus \bar{\Gamma}_m^d$. Define the cut-off function $\chi_d \in \mathcal{C}^\infty(\Gamma_m)$ equal to one inside Γ_m^d and zero elsewhere. We introduce the notation

$$\sigma_d \;:=\; \chi_d \sigma \quad \text{and} \quad \sigma_f \;:=\; (1 - \chi_d)\sigma \tag{19.20}$$

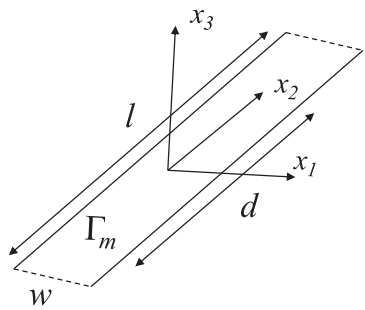


Fig. 19.3. Rectangle geometry and localization.

for the solution $\sigma \in \widetilde{H}^{-1/2}(\Gamma_m)$ of the boundary integral equation (19.3). Thus, the following system is built:

$$\begin{pmatrix} \chi_d \boldsymbol{V}_k & \chi_d \boldsymbol{V}_k \\ (1-\chi_d)\boldsymbol{V}_k & (1-\chi_d)\boldsymbol{V}_k \end{pmatrix} \begin{pmatrix} \sigma_d \\ \sigma_f \end{pmatrix} = \begin{pmatrix} \chi_d g_D \\ (1-\chi_d) g_D \end{pmatrix}. \tag{19.21}$$

Let ϱ be the distance function to $\partial\Gamma_m$. It is well known that, away from corners, solutions behave as $\varrho^{-1/2}$, while near them the singular behavior depends on the corner angle ν in a nonexplicit fashion. In Γ_m^d one can describe the function as tensor product $\sigma_d = \sigma_{1,d}\sigma_{3,d}$ so that

$$\boldsymbol{T}(\sigma_d)(x_1, x_3) \;=\; \boldsymbol{R}_k(\sigma_{3,d})(x_3)\boldsymbol{L}(\sigma_{1,d})(x_1) \;+\; \boldsymbol{K}_k(\sigma_d)(x_1, x_3),$$

where $\boldsymbol{L}$ is the logarithmic operator along x_1 and $\boldsymbol{R}_k$, $\boldsymbol{K}_k$ are compact operators. Thus, from a numerical point of view, a tensor product of weighted Chebyshev polynomials along x_1 and regular polynomials along x_3, i.e., $\sigma^{NM} \in \mathbb{Q}_N([-1,1]) \otimes \mathbb{P}_M([-d/2, d/2])$, can correctly describe σ_d but not σ_f. For the latter, we implement a triangular mesh $\mathcal{T}_h^f$ of Γ_m^f where an anisotropic refinement is carried out towards the boundaries according to the singularity order.

The associated variational formulation consists in finding σ_d^{NM} and σ_f^h in the corresponding approximation spaces such that

$$\left\langle \begin{pmatrix} \chi_d \boldsymbol{V}_k & \chi_d \boldsymbol{V}_k \\ (1-\chi_d)\boldsymbol{V}_k & (1-\chi_d)\boldsymbol{V}_k \end{pmatrix} \begin{pmatrix} \sigma_d^{NM} \\ \sigma_f^h \end{pmatrix}, \begin{pmatrix} \varphi_d^t \\ \varphi_f^t \end{pmatrix} \right\rangle = \begin{pmatrix} \langle g_D\,,\, \varphi_d^t \rangle \\ \langle g_D\,,\, \varphi_f^t \rangle \end{pmatrix}$$

for all $\varphi_d^t \in \mathbb{Q}_N([-1,1]) \otimes \mathbb{P}_M([-d/2, d/2])$ and $\varphi_f^t \in \mathbb{P}_0(\mathcal{T}_h^f)$.

19.5 Final Remarks

We have shown the appearance of a boundary operator with logarithmic singularities for Laplace and Helmholtz problems in $\mathbb{R}^2$ and especially in $\mathbb{R}^3$ when

objects are very elongated. We have proposed a scheme for the numerical approximation of Neumann jumps required to recover the entire solution using the single layer potential. This requires further study, as questions such as the precise functional space characterization of tensor schemes or the stability and error analysis remain open.

References

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

[St87] Stephan, E.: Boundary integral equations for screen problems in $\mathbb{R}^3$. *Integral Equations Operator Theory,* **10**, 236–257 (1987).

[CoDa02] Costabel, M., Dauge, M.: Crack singularities for general elliptic systems. *Math. Nachrichten,* **235**, 29–49 (2002).

[NiSa94] Nicaise, S., Sändig, A.-M.: General interface problems. I. *Math. Methods Appl. Sci.,* **17**, 395–429 (1994).

[Gr85] Grisvard, P.: *Elliptic Problems in Nonsmooth Domains,* Pitman, London (1985).

[Ke99] Keller, J.: Singularities at the tip of a plane angular sector. *J. Math. Phys.,* **40**, 1087–1092 (1999).

[MoLe76] Morrison, J., Lewis, J.: Charge singularity at the corner of a flat plate. *SIAM J. Appl. Math.,* **31**, 233–250 (1976).

[JLNL08] Jerez-Hanckes, C., Laude, V., Nédélec, J.-C., Lardat, R.: 3-D electrostatic hybrid elements model for SAW interdigital transducers. *IEEE Trans. Ultrason., Ferroelec., Freq. Control,* **55**, 686–695 (2008).

[SSC00] Shestopalov, Y., Smirnov, Y., Chernokozhin, E.: *Logarithmic Integral Equations in Electromagnetics,* VSP, Utrecht (2000).

[GrRy94] Gradshteyn, I., Ryzhik, I.: *Table of Integrals, Series, and Products,* Academic Press, London (1994).

[AbSt72] Abramowitz, M., Stegun, I.: *Handbook of Mathematical Functions,* Dover, New York (1972).

20

Boundary Integral Solution of the Time-Fractional Diffusion Equation

J. Kemppainen and K. Ruotsalainen

University of Oulu, Finland; jukemppa@paju.oulu.fi, keijo.ruotsalainen@ee.oulu.fi

20.1 Introduction

In this chapter we discuss the boundary integral solution of the fractional diffusion equation

$$\begin{aligned} \partial_t^\alpha \Phi - \Delta\Phi = 0, \text{ in } Q_T = \Omega \times (0,T), \\ \Phi = g, \text{ on } \Sigma_T = \Gamma \times (0,T), \\ \Phi(x,0) = 0, \ x \in \Omega, \end{aligned} \tag{20.1}$$

where $\Omega \subset \mathbb{R}^n$ is a smooth, bounded domain and ∂_t^α is the Caputo time derivative of the fractional order $0 < \alpha \leq 1$. For $\alpha = 1$ we get the ordinary diffusion equation and for $\alpha = 0$ we have the Helmholtz equation.

We present the fundamental solution by means of the Fox H-functions, and represent the solution of (1) as a single-layer potential. By the jump relations of the potential we derive the appropriate boundary integral operator. We give detailed mapping properties of the single-layer operator in anisotropic Sobolev spaces, which yields the unique solution of the boundary integral equation and thus the unique solution of the initial boundary value problem.

20.2 Function Spaces

Let $r, s \geq 0$. The anisotropic Sobolev space $H^{r,s}(\mathbb{R}^n \times \mathbb{R})$ consists of those distributions $u \in S'(\mathbb{R}^{n+1})$ for which the norm

$$\|u\|_{H^{r,s}} = (2\pi)^{-\frac{n+1}{2}} \left(\int_{\mathbb{R}^{n+1}} \{(1+|\xi|^2)^r + (1+|\eta|^2)^s\} |\widehat{u}(\xi,\eta)|^2 \mathrm{d}\xi \mathrm{d}\eta \right)^{\frac{1}{2}}$$

is finite. The spaces $H^{r,s}(Q_T)$ are defined by restrictions of elements in $H^{r,s}(\mathbb{R}^n \times \mathbb{R})$ to Q_T equipped with the norm

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_20,

$$\|u\|_{r,s;T} = \inf\{\|U\|_{H^{r,s}} : u = U|_{Q_T}\}.$$

Furthermore, the space $H_0^{r,s}(Q_T)$ is defined as the closure of $\mathcal{C}_0^\infty(Q_T)$ in $H^{r,s}(Q_T)$ and $H^{-r,-s}(Q_T)$ is defined by duality $H^{-r,-s}(Q_T) = (H_0^{r,s}(Q_T))'$. For $r, s \geq 0$ the space $H^{r,s}(\Gamma \times \mathbb{R})$ is defined by

$$H^{r,s}(\Gamma \times \mathbb{R}) = L^2(\mathbb{R}; H^r(\Gamma)) \cap H^r(\mathbb{R}; L^2(\Gamma)),$$

with the norm

$$\|u\|^2_{H^{r,s}(\Gamma\times\mathbb{R})} = \|u\|^2_{L^2(\mathbb{R};H^r(\Gamma))} + \|u\|^2_{H^s(\mathbb{R};L^2(\Gamma))}.$$

The spaces $H^{r,s}(\Sigma_T)$ and $H^{-r,-s}(\Sigma_T)$ are defined analogously with $H^{r,s}(Q_T)$ and $H^{-r,-s}(Q_T)$.

In what follows, we need the anisotropic Sobolev space $\widetilde{H}^{r,s}(\mathbb{R}^n \times \mathbb{R})$, which takes the vanishing initial condition at $t = 0$ into account and is defined by

$$\widetilde{H}^{r,s}(\mathbb{R}^n \times \mathbb{R}) = \{u \in H^{r,s}(\mathbb{R}^n \times \mathbb{R}) : \ \mathrm{supp}(u) \subset \mathbb{R}^n \times [0, \infty[\ \}.$$

For a finite time interval, we write $\mathbb{R}^{n+1}_T := \mathbb{R}^n \times (0, T)$ for $T > 0$ and define the space

$$\widetilde{H}^{r,s}(\mathbb{R}^{n+1}_T) = \{u = U|_{\mathbb{R}^n\times(-\infty,T)} : U \in \widetilde{H}^{r,s}(\mathbb{R}^n \times \mathbb{R})\},$$

equipped with the norm

$$\|u\|_{r,s;T} = \inf\{\|U\|_{H^{r,s}} : u = U|_{\mathbb{R}^n\times(-\infty,T)}\}.$$

The spaces $\widetilde{H}^{r,s}(\Sigma_T)$ are defined analogously.

20.3 Boundary Integral Formulation of the Problem

20.3.1 The Fundamental Solution

In order to formulate the boundary integral equation corresponding to (20.1), we need to calculate the fundamental solution $E(x, t)$. It is constructed by taking the Laplace transform of the time variable and the Fourier transform of the spatial variable in the fractional diffusion equation

$$(\partial_t^\alpha - \Delta)E(x, t) = \delta(x, t),$$

where $\delta(x, t)$ is the Dirac delta distribution. The transformed equation is then

$$(|\xi|^2 + s^\alpha)\widehat{\widetilde{E}}(\xi, s) = 1,$$

where the Fourier transform is defined by

$$\widehat{u}(\xi,t)=\int_{\mathbb{R}^n} e^{-i\langle x,\xi\rangle}u(x,t)\mathrm{d}x$$

and the Laplace transform by

$$\widetilde{u}(x,s)=\int_0^\infty e^{-st}u(x,t)\mathrm{d}t.$$

Hence, the Fourier–Laplace transform of the fundamental solution is

$$\widehat{\widetilde{E}}(\xi,s)=\frac{1}{|\xi|^2+s^\alpha}. \tag{20.2}$$

By taking the inverse Laplace and Fourier transforms, we notice that the fundamental solution is

$$E(x,t)=\begin{cases} \pi^{-n/2}t^{\alpha-1}|x|^{-n}H^{20}_{12}\big(\tfrac{1}{4}|x|^2t^{-\alpha}\big|^{(\alpha,\alpha)}_{(n/2,1),(1,1)}\big) & , \quad x\in\mathbb{R}^n,\ t>0,\\ 0 & , \quad x\in\mathbb{R}^n,\ t<0,\end{cases}$$

where H is the Fox H-function (see [KiS04], [Po99], and [PBM90]).

20.3.2 Mapping Properties of the Single-Layer Potential

Once the fundamental solution is known, we now define the single-layer potential

$$\Phi(x,t)=S\sigma(x,t)=\int_0^t\int_\Gamma \sigma(y,\tau)E(x-y,t-\tau)\mathrm{d}s_y\mathrm{d}\tau,\quad x\in\Omega,\ t\in(0,T),$$

for a given boundary distribution $\sigma\in C^\infty(\Sigma_T)$. The potential is the solution of the fractional diffusion equation both in the interior domain $\Omega\times(0,T)$ and on the exterior domain $[\mathbb{R}^n\setminus\overline{\Omega}]\times(0,T)$ with the zero initial condition. We denote the direct value of $S\sigma$ on the boundary by $V\sigma$.

The single-layer potential $S\sigma(x,t)$ is continuous up to the boundary due to the asymptotic properties of the fundamental solution. This leads us to the boundary relation

$$\gamma(S\sigma)(x,t)=\gamma(\Phi)(x,t)=V\sigma(x,t).$$

In other words, we have converted the initial boundary value problem of the fractional diffusion equation (20.1) to a boundary integral equation

$$V\sigma(x,t)=\gamma(\Phi)(x,t)=g(x,t),\ (x,t)\in\Sigma_T. \tag{20.3}$$

In our analysis we need the mapping properties of the single-layer potential in Sobolev spaces. The single-layer potential can be written as

$$S\phi=E*\gamma'(\phi). \tag{20.4}$$

By (20.2) we have

$$\mathcal{F}(E * f)(\xi, \eta) = \frac{1}{|\xi|^2 + (\mathrm{i}\eta)^\alpha} \hat{f}(\xi, \eta) \tag{20.5}$$

for smooth f with $\mathrm{supp} f \subset \mathbb{R}^n \times [0, \infty[$. It follows that the map

$$\psi \mapsto E * \psi : \widetilde{H}^{r, \frac{\alpha}{2} r}_{\mathrm{comp}}(\mathbb{R}^n \times (0, T)) \to \widetilde{H}^{r+2, \frac{\alpha}{2}(r+2)}_{\mathrm{loc}}(\mathbb{R}^n \times (0, T)) \tag{20.6}$$

is continuous for any $r \in \mathbb{R}$, where comp means compact support and loc local behavior in space variables. Since the trace map $\gamma : H^{r,s}(Q_T) \to H^{\lambda,\mu}(\Sigma_T)$ is continuous for every $\lambda = r - \frac{1}{2}$, $\mu = \frac{s}{r}\lambda$, $r > \frac{1}{2}$, and $s \geq 0$ ([LiMaI72], Theorem 4.2 of Chapter 1, and [LiMaII72], Theorem 2.1 of Chapter 4), by duality we have

$$\gamma' : H^{-\lambda, -\mu}(\Sigma_T) \to H^{-r, -s}_{\mathrm{comp}}(\mathbb{R}^n \times (0, T)). \tag{20.7}$$

Using the trace theorem once again, combining (20.6) and (20.7), and noting that the spaces $H^{r,s}(\Sigma_T)$ and $\widetilde{H}^{r,s}(\Sigma_T)$ coincide if and only if $|s| < \frac{1}{2}$, we may conclude the following result.

Theorem 1. *Let* $0 < s < 1$. *The operator*

$$V : \widetilde{H}^{-s, -\frac{\alpha}{2} s}(\Sigma_T) \to \widetilde{H}^{1-s, \frac{\alpha}{2}(1-s)}(\Sigma_T) \tag{20.8}$$

is continuous.

20.3.3 Jump Relations

As usual, we define the jump of the traces across the boundary as

$$[\gamma(u)] = \gamma(u_+) - \gamma(u_-),$$

where γ is the spatial trace operator and $u_+ = u|_{\overline{\Omega}^c}$ ($u_- = u|_\Omega$) is a function which is defined in the exterior (interior) of the domain Ω. Similarly, the jump of the normal derivative across the boundary is defined as

$$[\gamma_1(u)] = [\gamma(\partial_n u)] = \gamma(\partial_n u_+) - \gamma(\partial_n u_-).$$

For the proof of the jump relations we need some basic properties of fractional derivatives and Green's formula in the case of fractional time derivatives. Because the properties of fractional derivatives are crucial for the proof of Green's formula, we consider them first.

In what follows, $\partial_t^\alpha := {}^c D^\alpha_{0+}$ denotes the left Caputo derivative on time interval $(0, T)$, and the right Caputo derivative on the interval $(0, T)$ is denoted by ${}^c D^\alpha_{T-}$. They are defined by the formulas

$$
\begin{aligned}
{}^cD^\alpha_{0+}\varphi(t) &= \frac{1}{\Gamma(1-\alpha)}\int_0^t \frac{\varphi'(s)}{(t-s)^\alpha}\mathrm{d}s,\\
{}^cD^\alpha_{T-}\varphi(t) &= -\frac{1}{\Gamma(1-\alpha)}\int_t^T \frac{\varphi'(s)}{(s-t)^\alpha}\mathrm{d}s.
\end{aligned}
$$

The right and left Riemann–Liouville derivatives on the interval $(0,T)$ are defined by setting

$$
\begin{aligned}
D^\alpha_{0+}\varphi(t) &= \frac{1}{\Gamma(1-\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^t \frac{\varphi(s)}{(t-s)^\alpha}\mathrm{d}s,\\
D^\alpha_{T-}\varphi(t) &= -\frac{1}{\Gamma(1-\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\int_t^T \frac{\varphi(s)}{(s-t)^\alpha}\mathrm{d}s,
\end{aligned}
$$

respectively. Note that for sufficiently smooth functions φ for which $\varphi(0) = 0$ the left Caputo and Riemann–Liouville derivatives coincide (see [Po99], formula (2.165)), i.e.,

$$
{}^cD^\alpha_{0+}\varphi(t) = D^\alpha_{0+}\varphi(t). \tag{20.9}
$$

Integration by parts gives the following relation between the left Caputo and the right Riemann–Liouville derivative:

$$
\int_0^T \partial_t^\alpha\varphi(t)\psi(t)\mathrm{d}t = \int_0^T \varphi(t)D^\alpha_{T-}\psi(t)\mathrm{d}t \tag{20.10}
$$

for $\varphi \in C^1([0,T])$ with $\varphi(0) = 0$ and $\psi \in C^1([0,T])$.

The time reversal operator on the interval $(0,T)$ is defined by setting

$$
\kappa_T\varphi(t) = \varphi(T-t).
$$

Applying the time reversal operator to the left Riemann–Liouville derivative, we have

$$
D^\alpha_{T-}(\kappa_T\varphi)(t) = \kappa_T D^\alpha_{0+}\varphi(t). \tag{20.11}
$$

Let us next consider Green's formula for the fractional diffusion equation. Let $u, v \in C^1(\overline{Q}_T)$ with $v(\cdot,0) = 0$. Using the properties (20.9), (20.10), (20.11), and Green's formula with respect to the space variable, we obtain *Green's formula for the fractional diffusion equation:*

$$
\begin{aligned}
\int_{Q_T} \{(\partial_t^\alpha - \Delta)u\kappa_T v - \kappa_T u(\partial_t^\alpha - \Delta)v\}\mathrm{d}x\mathrm{d}t &= \int_{\Sigma_T} \{u\partial_n\kappa_T v - \partial_n u\kappa_T v\}\mathrm{d}s_\Gamma\mathrm{d}t\\
&= \langle\gamma u, \gamma_1\kappa_T v\rangle - \langle\gamma_1 u, \gamma\kappa_T v\rangle.
\end{aligned}
$$

By density arguments, Green's formula extends to functions $u, v \in \widetilde{H}^{1,\frac{\alpha}{2}}(Q_T)$ such that $(\partial_t^\alpha - \Delta)u, (\partial_t^\alpha - \Delta)v \in L^2(Q_T)$.

Now we are able to state and prove the jump relations for the single-layer potential of the fractional diffusion operator.

Theorem 2. *For every $\psi \in H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)$, the following jump relations hold:*

$$[\gamma(S\psi)] = 0, \ [\gamma_1(S\psi)] = -\psi.$$

Proof. For the function ψ let us denote $u = S\psi$. By the assumption on ψ and the properties of the trace map, we have $u \in \widetilde{H}^{1,\frac{\alpha}{2}}(B_R \times (0,T))$, where the radius of the ball B_R is so large that $\overline{\Omega} \subset B_R$. By the trace theorem ([LiMaI72], Theorem 4.2 of Chapter 1, and [LiMaII72], Theorem 2.1 of Chapter 4), $\gamma(u|_{Q_T}) = \gamma(u|_{Q_T^c})$, where $Q_T^c := (B_R \setminus \overline{\Omega}) \times (0,T)$. Hence, the continuity of the trace across the boundary is proved.

Using the representation formula (20.4), we have

$$(\partial_t^\alpha - \Delta)u = \gamma'(\psi)$$

in the distributional sense in $\mathbb{R}^n \times (0,T)$. Choosing $\phi \in C_0^\infty(B_R \times (0,T))$, we get

$$\langle \psi, \gamma(\phi)\rangle = \langle \gamma'(\psi), \phi\rangle = \langle (\partial_t^\alpha - \Delta)u, \phi\rangle = \langle u, (D_{T-}^\alpha - \Delta)\phi\rangle.$$

Making the time reversal and using the properties of the Caputo fractional derivatives, $D_{T-}^\alpha \kappa_T \phi = \kappa_T \partial_t^\alpha \phi$, from the previous equation we obtain

$$\langle \psi, \gamma(\kappa_T\phi)\rangle = \int_{B_R\times(0,T)} (\partial_t^\alpha - \Delta)\phi\, \kappa_T u \mathrm{d}x\mathrm{d}t. \tag{20.12}$$

Using Green's formula for the fractional diffusion operator with respect to the sets Q_T and Q_T^c, we get (recall that on $Q_T \cup Q_T^c$ we have $(\partial_t^\alpha - \Delta)u = 0$)

$$\int_{Q_T} (\partial_t^\alpha - \Delta)\phi\, \kappa_T u \mathrm{d}x\mathrm{d}t \quad = \langle \gamma_1(u), \gamma(\kappa_T\phi)\rangle - \langle \gamma(u), \gamma_1(\kappa_T\phi)\rangle, \tag{20.13}$$

$$\int_{Q_T^c} (\partial_t^\alpha - \Delta)\phi\, \kappa_T u \mathrm{d}x\mathrm{d}t \quad = -\langle \gamma_1(u), \gamma(\kappa_T\phi)\rangle + \langle \gamma(u), \gamma_1(\kappa_T\phi)\rangle. \tag{20.14}$$

The jump of the traces for u is $[\gamma(u)] = 0$ by the first part of the theorem. Since the test function ϕ is smooth, its traces are continuous across the boundary, i.e., $[\gamma(\kappa_T\phi)] = [\gamma_1(\kappa_T\phi)] = 0$.

Adding equations (20.13) and (20.14) together and using the previous trace properties of u and $\kappa_T\phi$, we obtain

$$\int_{B_R\times(0,T)} (\partial_t^\alpha - \Delta)\phi\, \kappa_T u \mathrm{d}x\mathrm{d}t = -\langle [\gamma_1(u)], \gamma(\kappa_T\phi)\rangle. \tag{20.15}$$

Combining equations (20.12) and (20.15), we finally obtain

$$\langle \psi, \gamma(\kappa_T\phi)\rangle = -\langle [\gamma_1(u)], \gamma(\kappa_T\phi)\rangle \ \forall \phi \in C_0^\infty(B_R \times (0,T)), \tag{20.16}$$

which proves the second statement.

20.3.4 Coerciveness of the Single-Layer Potential

For the proof of coercivity we use the standard technique by proving Gårding's inequality and positivity for the single-layer potential (see the proof of Theorem 3.11 in [Co92]). To begin with we apply Green's formula to the function $u = S\psi$, where $\psi \in H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)$. By the Gauss divergence formula,

$$\langle \nabla u, \nabla v\rangle_{Q_T} + \langle \partial_t^\alpha u, v\rangle_{Q_T} = \langle \gamma_1(u), \gamma(v)\rangle_{\Sigma_T} + \langle (\partial_t^\alpha - \Delta)u, v\rangle_{Q_T} \tag{20.17}$$

first for smooth u and v and then by the density argument and continuity for $v \in \widetilde{H}^{1,\frac{\alpha}{2}}(Q_T)$. Since $u = S\psi$ is the solution of the homogeneous fractional diffusion equation, we obtain from (20.17) again by the density argument and continuity:

$$\langle \nabla u, \nabla u\rangle_{Q_T} + \langle \partial_t^\alpha u, u\rangle_{Q_T} = \langle \gamma_1(u), \gamma(u)\rangle_{\Sigma_T}.$$

Since the Caputo derivative is positive semidefinite, the previous equality implies

$$\langle \gamma_1(u|_{Q_T}), \gamma(u|_{Q_T})\rangle_{\Sigma_T} \geq \int_{Q_T} |\nabla u|^2 \mathrm{d}x\mathrm{d}t.$$

On the domain Q_T^c we obtain

$$\langle \gamma_1(u|_{Q_T^c}), \gamma(u|_{Q_T^c})\rangle = \int_{\partial B_R\times(0,T)} u\partial_n u \mathrm{d}s_{\partial B_R}\mathrm{d}t - \int_{Q_T^c} |\nabla u|^2 \mathrm{d}x\mathrm{d}t - \langle \partial_t^\alpha u, u\rangle_{Q_T^c}.$$

By the jump relations we have

$$\langle \psi, V\psi\rangle = \langle \gamma_1(u|_{Q_T}), \gamma(u|_{Q_T})\rangle - \langle \gamma_1(u|_{Q_T^c}), \gamma(u|_{Q_T^c})\rangle.$$

Hence,

$$\langle \psi, V\psi\rangle \geq \int_{Q_T\cup Q_T^c} |\nabla u|^2 \mathrm{d}x\mathrm{d}t - \int_{\partial B_R\times(0,T)} u\partial_n u \mathrm{d}s_{\partial B_R}\mathrm{d}t.$$

Since the fundamental solution $E(x,t)$ is smooth on the boundary ∂B_R, the mapping $\psi \mapsto u|_{\partial B_R\times(0,T)} : H^{-\frac{1}{2},-\frac{\alpha}{2}}(\Sigma_T) \to H^{r,s}(\partial B_R \times (0,T))$ is continuous for any $r, s \in \mathbb{R}$, and the same is true for the mapping $\psi \mapsto \partial_n u|_{\partial B_R\times(0,T)}$. Hence, there exists a compact operator $T_1 : H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T) \to H^{\frac{1}{2},\frac{\alpha}{4}}(\Sigma_T)$ such that

$$\int_{\partial B_R\times(0,T)} u\partial_n u \mathrm{d}s_{\partial B_R}\mathrm{d}t = \langle \psi, T_1\psi\rangle_{\Sigma_T}.$$

On the other hand, since $\widetilde{H}^{1,\frac{\alpha}{2}}(Q_T) \hookrightarrow L^2(Q_T)$ is a compact embedding, there exists a compact operator $T_2 : H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T) \to H^{\frac{1}{2},\frac{\alpha}{4}}(\Sigma_T)$ such that

$$\int_{Q_T\cup Q_T^c} |\nabla u|^2 \mathrm{d}x\mathrm{d}t = \|u\|^2_{H^{1,0}(Q_T)} + \|u\|^2_{H^{1,0}(Q_T^c)} - \langle \psi, T_2\psi\rangle.$$

We need the following lemma.

Lemma 1. *The norms of $\widetilde{H}^{1,\frac{\alpha}{2}}(Q_T)$, $0<\alpha<1$, and $\widetilde{H}^{1,0}(Q_T)$ are equivalent on the subspace of functions satisfying the homogeneous fractional diffusion equation.*

The proof of this lemma is similar to that of Lemma 2.15 in [Co92] and is based on defining the space $\mathcal{V}(Q_T)$, which consist of those functions $u \in L^2((0,T);H^1(\Omega))$ such that $\partial_t^\alpha u \in L^2((0,T);H^{-1}(\Omega))$ and is equipped with the norm

$$\|u\|^2_{\mathcal{V}(Q_T)} = \|u\|^2_{H^{1,0}(Q_T)} + \|\partial_t^\alpha u\|^2_{H^{-1,0}(Q_T)}.$$

After that, we prove that the norms in spaces $\mathcal{V}(Q_T)$, $H^{1,0}(Q_T)$, and $H^{1,\frac{\alpha}{2}}(Q_T)$ are equivalent by using a proper interpolation result.

Utilizing the norm equivalence, we obtain

$$\langle \psi, (V+T_1+T_2)\psi\rangle \geq C(\|u\|^2_{H^{1,\frac{\alpha}{2}}(Q_T)} + \|u\|^2_{H^{1,\frac{\alpha}{2}}(Q_T^c)}). \tag{20.18}$$

By Theorem 2 we have

$$\|\psi\|_{H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)} = \|\gamma_1(u|_{Q_T}) - \gamma_1(u|_{Q_T^c})\|_{H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)}.$$

Combining this with inequality (20.18) and the trace theorem and denoting $T := T_1 + T_2$, we finally get *Gårding's inequality:*

$$\langle \psi, (V+T)\psi\rangle \geq C\|\psi\|^2_{H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)}.$$

We now consider the positiveness of the single-layer operator.

Lemma 2. *For all $\sigma \in \widetilde{H}^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)$ we have*

$$\mathrm{Re}(V\sigma,\sigma) > 0 \quad \text{if } \sigma \neq 0.$$

Proof. By the standard density argument it is enough to show the positivity for smooth functions $\sigma(x,t)$ for which the initial condition $\sigma(x,0)=0$ is valid.

Let us define the potential $\phi = S\sigma$ which is the solution of the homogeneous equation:

$$(\partial_t^\alpha - \Delta)\phi(x,t) = 0 \ \forall (x,t) \in Q_T \cup Q_T^c.$$

For a fixed $t>0$ we get, by the Gauss divergence formula,

$$0 = \int_\Omega (\partial_t^\alpha \phi - \Delta\phi)\cdot\phi\,\mathrm{d}x = \int_\Omega \partial_t^\alpha\phi\cdot\phi\,\mathrm{d}x + \int_\Omega |\nabla\phi|^2\mathrm{d}x - \int_\Gamma \partial_n\phi_-\cdot\phi\,\mathrm{d}s_\Gamma,$$
$$0 = \int_{\overline{\Omega}^c} (\partial_t^\alpha \phi - \Delta\phi)\cdot\phi\,\mathrm{d}x = \int_{\overline{\Omega}^c} \partial_t^\alpha\phi\cdot\phi\,\mathrm{d}x + \int_{\overline{\Omega}^c} |\nabla\phi|^2\mathrm{d}x + \int_\Gamma \partial_n\phi_+\cdot\phi\,\mathrm{d}s_\Gamma.$$

Adding these identities together, we obtain

$$\int_\Gamma -[\gamma_1\phi]\phi\,\mathrm{d}s_\Gamma = \int_{\Omega\cup\overline{\Omega}^c} \{\partial_t^\alpha\phi\cdot\phi + |\nabla\phi|^2\}\mathrm{d}x.$$

Note that on the right-hand side we have used the continuity of the traces of the single-layer potential proved in Theorem 2.

Integrating the previous identity with respect to the time variable over the interval $[0,T]$ yields

$$-\int_{\Sigma_T}[\gamma_1\phi]\phi\,\mathrm{d}s_\Gamma\mathrm{d}t=\int_{Q_T\cup Q_T^c}\{\partial_t^\alpha\phi\cdot\phi+|\nabla\phi|^2\}\mathrm{d}x\mathrm{d}t.$$

By Theorem 2 we have

$$\int_{\Sigma_T}\sigma V\sigma\,\mathrm{d}s_\Gamma\mathrm{d}t=\int_{Q_T\cup Q_T^c}\{\partial_t^\alpha S\sigma\cdot S\sigma+|\nabla S\sigma|^2\}\mathrm{d}x\mathrm{d}t.$$

Since the operators J^1 and D^1 commute when acting on functions with zero initial conditions, and they possess the semigroup property $J^\alpha J^\beta=J^{\alpha+\beta}$, we obtain, using the positive semidefiniteness of the operator J^α,

$$\begin{aligned}\int_0^T\partial_t^\alpha\phi\cdot\phi\,\mathrm{d}t &= \int_0^T J^{1-\alpha}D^1\phi\cdot\phi\,\mathrm{d}t=\int_0^T J^{1-\alpha}D^1\phi\cdot J^1D^1\phi\,\mathrm{d}t\\ &= \int_0^T J^{1-\alpha}D^1\phi\cdot J^\alpha J^{1-\alpha}D^1\phi\,\mathrm{d}t\\ &= \int_0^T \underbrace{J^{1-\alpha}D^1\phi}_{g}\cdot J^\alpha\underbrace{J^{1-\alpha}D^1\phi}_{g}\,\mathrm{d}t=\int_0^T J^\alpha g\cdot g\,\mathrm{d}t\geq 0.\end{aligned}$$

Hence, the single-layer operator is at least positive semidefinite; that is,

$$\int_{\Sigma_T}\sigma V\sigma\,\mathrm{d}s_\Gamma\mathrm{d}t\geq\int_{Q_T\cup Q_T^c}|\nabla S\sigma|^2\mathrm{d}x\mathrm{d}t\geq 0.$$

If there exists a boundary distribution such that

$$\int_{\Sigma_T}\sigma V\sigma\,\mathrm{d}s_\Gamma\mathrm{d}t=0,$$

then $\nabla S\sigma=0$ for all $x\in Q_T\cup Q_T^c$ and $0<t<T$. Thus, $\partial_t^\alpha\phi-\Delta\phi=\partial_t^\alpha\phi=0$. Now for every fixed $x\in Q_T\cup Q_T^c$ we have

$$\partial_t^\alpha\phi=J^{1-\alpha}D^1\phi=0.$$

Since the Abel integral equation $J^{1-\alpha}\varphi=\psi$ is uniquely solvable, we have $D^1\phi=0$, and thus $\phi(x,t)=C$. By the zero initial condition the constant $\phi(x,t)=C=0$, proving the positivity of the single-layer operator.

As in [HsSa89] or [Co92], we obtain the strong coerciveness of the single-layer operator, and thus we are able to state our main assertion.

Theorem 3. *The single-layer operator* $V : \widetilde{H}^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T) \to \widetilde{H}^{\frac{1}{2},\frac{\alpha}{4}}(\Sigma_T)$ *is an isomorphism. Furthermore, it is coercive, i.e., there exists a positive constant* c *such that*

$$\mathrm{Re}(V\sigma,\sigma) \geq c\|\sigma\|^2_{H^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)}$$

for all $\sigma \in \widetilde{H}^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)$.

Corollary 1. *For every* $g \in \widetilde{H}^{\frac{1}{2},\frac{\alpha}{4}}(\Sigma_T)$, *the fractional diffusion equation admits a unique solution* $\Phi(x,t) \in \widetilde{H}^{1,\frac{\alpha}{2}}(Q_T)$ *which is given by the single-layer potential*

$$\Phi(x,t) = S\sigma(x,t),$$

where $\sigma \in \widetilde{H}^{-\frac{1}{2},-\frac{\alpha}{4}}(\Sigma_T)$ *is the unique solution of the boundary integral equation*

$$V\sigma = g.$$

References

[AbSt71] Abramowitz, M., Stegun, I.A. (eds.): *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables,* U.S. Government Printing Office, Washington, D.C. (1971).

[Co92] Costabel, M.: Boundary integral operators for the heat equation. *Integral Equations Oper. Theory,* **13**, 498–552 (1992).

[Co04] Costabel, M.: Time-dependent problems with the boundary integral equation method, in *Encyclopedia of Computational Mechanics,* Stein, E., de Borst, R., Hughes, T.J.R. (eds.), Wiley (2004).

[CoSa01] Costabel, M., Saranen, J.: Parabolic boundary integral operators, symbolic representations and basic properties. *Integral Equations Oper. Theory,* **40**, 185–211 (2001).

[HsSa89] Hsiao, G.C., Saranen, J.: Coercivity of the single layer heat operator, Report 89-2, Center for Mathematics and Waves, University of Delaware, Newark, DE (1989).

[HsSa93] Hsiao, G.C., Saranen, J.: Boundary integral solution of the two-dimensional heat equation. *Math. Methods Appl. Sci.,* **16**, 87–114 (1993).

[KiS04] Kilbas, A.A., Saigo, M.: *H-transforms: Theory and Applications,* CRC Press, Boca Raton, FL (2004).

[LiMaI72] Lions, J.-L., Magenes, E.: *Non-homogeneous Boundary Value Problems and Applications. Vol. I,* Springer, Berlin (1972).

[LiMaII72] Lions, J.-L., Magenes, E.: *Non-homogeneous Boundary Value Problems and Applications. Vol. II,* Springer, Berlin (1972).

[Po99] Podlubny, I.: *Fractional Differential Equations,* Academic Press, San Diego, CA (1999).

[PBM90] Prudnikov, A.P., Brychkov, Y.A., Marichev, O.I.: *Integrals and Series. Vol. 3. More Special Functions,* Overseas Publishers Association, Amsterdam (1990).

21

Boundary Element Collocation Method for Time-Fractional Diffusion Equations

J. Kemppainen and K. Ruotsalainen

University of Oulu, Finland; jukemppa@paju.oulu.fi, keijo.ruotsalainen@ee.oulu.fi

21.1 Introduction

In this chapter, we discuss the numerical solution of the space-time boundary integral equation

$$S_\Gamma u_\Gamma(x,t) = \int_0^t \int_\Gamma u_\Gamma(y,\tau) E(x-y,t-\tau) \mathrm{d}s_y \mathrm{d}\tau = f(x,t), \qquad x \in \Gamma, \quad 0 < t < T,$$

where Γ is a smooth plane curve. The kernel of the integral operator,

$$E(x,t) = \frac{1}{\pi} t^{\alpha-1} |x|^{-2} H_{12}^{20}\left(\frac{1}{4}|x|^2 t^{-\alpha}\Big|_{(1,1),(1,1)}^{(\alpha,\alpha)}\right), \quad 0 < \alpha \le 1,$$

is the fundamental solution of the time-fractional diffusion equation (see [KiSa04] and [PBM90]). We consider the problem

$$\begin{aligned} \partial_t^\alpha \Phi - \Delta\Phi &= 0 \quad \text{in } Q_T = \Omega \times (0,T), \\ B(\Phi) &= g \quad \text{on } \Sigma_T = \Gamma \times (0,T), \\ \Phi(x,0) &= 0, \quad x \in \Omega, \end{aligned}$$

where the boundary operator $B(\Phi) = \Phi|_{\Sigma_T}$ and ∂_t^α is the Caputo time derivative of the fractional order $0 < \alpha \le 1$.

We shall consider the spline collocation method for the numerical approximation of the solution on quasi-uniform meshes with piecewise linear tensor product splines as the approximation space. We will show that the spline collocation method is stable in a suitable anisotropic Sobolev space, and it furnishes quasi-optimal error estimates.

In [KeRuI] we have considered the mapping properties of the single-layer operator S_Γ. We have shown that the single-layer operator defines an isomorphism on a scale of anisotropic Sobolev spaces, and that the operator

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_21,

is coercive in its natural energy norm. Also, the equivalence of the indirect boundary integral formulation and the time-fractional diffusion equation has been deduced.

The paper is organized as follows. In Section 21.2 we will briefly recall the basic definitions of smoothness spaces, spline spaces, and their approximation properties. In Section 21.3 we will present the basic mapping properties of the single-layer operator and present its Fourier representation, which will be used in the analysis of the spline collocation method. Finally, in Section 21.4 we will tackle the spline collocation method for the numerical approximation of the single-layer equation. We will show that the problem is equivalent to a modified Galerkin problem. The stability and error analysis is then based on the properties of the corresponding bilinear form of the modified Galerkin method.

21.2 Preliminaries

21.2.1 Smoothness Spaces

The space of continuous functions which are 1-periodic in the spatial variable is denoted by $C_1^{0,0}(\mathbb{R}^2)$. Moreover, the space $C_1^{k,l}(\mathbb{R}^2)$ contains the continuous functions u for which $\partial_\theta^k \partial_t^l u \in C_1^{0,0}(\mathbb{R}^2)$. Let $R_T^2 = \mathbb{R} \times (0,T)$. The space $C_1^{k,l}(\overline{\mathbb{R}}_T^2)$ consists of restrictions $u = U|_{\mathbb{R}\times[0,T]}$. Finally, $C_{00}^{k,l}(\overline{\mathbb{R}}_T^2)$ is defined to be the space of restrictions $u = U|_{\mathbb{R}\times[0,T]}$, where $U \in C_1^{k,l}(\mathbb{R}^2)$ is such that $U|_{\mathbb{R}\times(-\infty,0)} = 0$. The spaces are equipped with the natural maximum norm

$$\|u\|_{C_1^{k,l}(\mathbb{R}^2)} = \max_{\substack{0\le k_1\le k\\ 0\le l_1\le l}} \sup_{(\theta,t)\in\mathbb{R}^2} |\partial_\theta^{k_1}\partial_t^{l_1} u(\theta,t)|.$$

Let $r, s \in \mathbb{R}$. The anisotropic Sobolev space $H^{r,s}(\mathbb{R}^2)$ consists of distributions, which are 1-periodic with respect to the spatial variable, equipped with the norm

$$\|u\|_{r,s} = \Big(\sum_{k\in\mathbb{Z}} (1+|2\pi k|^2)^r \int_{\mathbb{R}} \big(1+|\eta|^2\big)^s |\widehat{u}(k,\eta)|^2 d\eta\Big)^{\frac{1}{2}}.$$

Here $\widehat{u}(k,\eta)$ for $(k,\eta) \in \mathbb{Z}\times\mathbb{R}$ is defined as the Fourier transform with respect to the spatial variable and the Fourier transform with respect to the time variable, i.e.,

$$\widehat{u}(k,\eta) = \int_0^1 \int_{\mathbb{R}} e^{-i(2\pi kx + t\eta)} u(\theta,t) d\theta\, dt.$$

Furthermore, we define $H^{r,s}(\mathbb{R}_T^2)$ as the space of restrictions $u = U|_{\mathbb{R}\times(0,T)}$, $U \in H^{r,s}(\mathbb{R}^2)$, equipped with the usual infimum norm

$$\|u\|_{r,s;T} = \inf\{\|U\|_{r,s} : u = U_{\mathbb{R}\times(0,T)}\}.$$

Finally, let us introduce the anisotropic Sobolev space $\widetilde{H}^{r,s}(\mathbb{R}^2)$, which takes the vanishing initial condition at $t = 0$ into account,

$$\widetilde{H}^{r,s}(\mathbb{R}^2) = \{u \in H^{r,s}(\mathbb{R}^2) : \ \mathrm{supp}(u) \subset \mathbb{R} \times [0, \infty)\}.$$

The spaces $\widetilde{H}^{r,s}(\mathbb{R}^2_T)$ are defined analogously.

For a more detailed study of the properties of the anisotropic Sobolev spaces, we refer to the book by Lions and Magenes [LiMaI72], [LiMaII72], and to the series of papers [CoSa01], [Co92], [HsSa89], [HsSa93], [Hä98], and to references therein.

The following embedding results can be found, for example, in [Hä98], Theorem 2.6.

Theorem 1. *Let $r > k + \frac{1}{2}$ and $s > l + \frac{1}{2}$, $k, l \in \mathbb{N}_0$. Then the embedding $\widetilde{H}^{r,s}(\mathbb{R}^2_T) \subset C_1^{k,l}(\overline{\mathbb{R}}^2_T)$ is continuous.*

21.2.2 The Approximation Spaces

For the approximation of the single-layer equation for the time-fractional diffusion we need suitable approximation spaces. They are defined with respect to the quasi-uniform meshes $\Delta_\theta = \{\theta_k : \theta_{k+N} = \theta_k + 1\}_{k \in \mathbb{Z}}$ and $\Delta_t = \{t_k\}_{k=0}^M$ given by

$$0 < \theta_1 < \cdots < \theta_N = 1, \ 0 = t_0 < t_1 < \cdots < t_M = T,$$

where the mesh parameters are

$$h_\theta = \max_{1 \le i \le N-1} |\theta_{i+1} - \theta_i|, \quad h_t = \max_{1 \le j \le M-1} |t_{j+1} - t_j|,$$

respectively.

Let $S^1(\Delta_\theta)$ be the space of 1-periodic piecewise linear continuous splines. The space $S_0^1(\Delta_t)$ consists of piecewise linear continuous splines such that $\phi(0) = 0$. Our approximation space is defined as the space of tensor product splines $M^1 = S^1(\Delta_\theta) \otimes S_0^1(\Delta_t)$:

$$M^1 = \mathrm{span}\{\psi_n \phi_m : \ \psi_n \in S^1(\Delta_\theta), \ \phi_m \in S_0^1(\Delta_t), \ 1 \le n \le N, \ 1 \le m \le M\}.$$

For the 1-periodic piecewise linear smoothest splines, the well-known approximation properties in periodic Sobolev spaces hold [ElSc85]:

$$\inf_{\psi \in S^1(\Delta_\theta)} \|u - \psi\|_{H^r} \le C h_\theta^{s-r} \|u\|_{H^s}, \ u \in H^s, \ r \le s \le 2, \ r < \frac{3}{2}.$$

Also, in $S_0^1(\Delta_t)$ the approximation property

$$\inf_{\phi \in S_0^1(\Delta_t)} \|u - \phi\|_{H^r(0,T)} \le C h_t^{s-r} \|u\|_{H^s(0,T)}$$

holds for any $u \in H^s(0,T)$, $\frac{1}{2} < s \le 2$, $u(0) = 0$ and $0 \le r \le 1, \ r \le s$.

21.3 The Single-Layer Operator

In this section we recall the main results from [KeRuI] and [KeRuII] concerning the mapping properties of the single-layer operator. We assume that the boundary Γ has a smooth, 1-periodic parametric representation $\theta \mapsto x(\theta)$ such that $|x'(\theta)| > 0$. We denote $u(\theta,t) = u_\Gamma(x(\theta),t)$ and $Vu(\theta,t) = (S_\Gamma u_\Gamma)(x(\theta),t)$. Then the single-layer operator can be written in the form

$$Vu(\theta,t) = \int_0^t \int_0^1 E(x(\theta)-x(\phi),t-\tau)u(\phi,\tau)|x'(\phi)|\mathrm{d}\phi\mathrm{d}\tau.$$

We notice that the single-layer operator is of Volterra type, i.e., if $u(\theta,\tau)=0$, when $\tau < t$, then $Vu(\theta,\tau)=0$. This is a consequence of the properties of the fundamental solution.

Note that the analysis presented here is valid for general smooth boundary curves when arc length parametrization is used. Thus, we may assume that the Jacobian of the parametric representation $|x'(\theta)| = \rho > 0$ is a constant.

The single-layer operator defines an anisotropic pseudodifferential operator which has the following representation [KeRuII]:

$$\begin{aligned} Vu(\theta,t) &= \frac{\rho}{2\pi}\sum_{m\in\mathbb{Z}}\int_{\mathbb{R}_\eta} a(m,\eta)\widehat{u}(m,\eta)e^{i2\pi m\theta+i\eta t}\mathrm{d}\eta + Bu(\theta,t) \\ &= V_0u(\theta,t)+Bu(\theta,t), \end{aligned}$$

where B is an operator of Volterra type which is a bounded operator between the anisotropic Sobolev spaces $\widetilde{H}^{s,\frac{\alpha}{2}s}(\Sigma_T) \to \widetilde{H}^{s+2,\frac{\alpha}{2}(s+2)}(\Sigma_T)$, and the principal part V_0 has the anisotropic symbol

$$a(\xi,\eta) = \frac{1}{2}([\frac{2\pi\xi}{\rho}]^2 + (i\eta)^\alpha)^{-\frac{1}{2}}.$$

In our analysis the following properties are crucial:

(i) The symbol is quasi-homogeneous of order $\beta=-1$; that is,

$$a(\theta,\lambda p,\lambda^{\frac{2}{\alpha}}\eta) = \lambda^{-1}a(\theta,p,\eta),\ \lambda\geq 1.$$

(ii) The mapping $\eta \to a(\theta,p,\eta)$ has polynomially bounded analytic continuation into the domain $\{z\in\mathbb{C}|\ z=\eta-i\sigma,\ \sigma>0\}$, and is continuous for $\sigma\geq 0$.

Note that our symbol a does not satisfy (ii), but we may define a proper approximation for it which satisfies (ii) and we may conclude the following theorem (see [CoSa00]).

Theorem 2. *The single-layer operator has the following properties:*

(i) $V : \widetilde{H}^{s,\frac{\alpha}{2}s}(\mathbb{R}^2_T) \to \widetilde{H}^{s+1,\frac{\alpha}{2}(s+1)}(\mathbb{R}^2_T)$ *is bounded for all* $s\in\mathbb{R}$.

(ii) There exists positive constant C_0 such that

$$\mathrm{Re}(Vu,u) \geq C_0 \|u\|^2_{-\frac{1}{2},-\frac{\alpha}{4}}$$

for all $u \in \widetilde{H}^{-\frac{1}{2},-\frac{\alpha}{4}}(\mathbb{R}^2_T)$.

(iii) $V : \widetilde{H}^{-\frac{1}{2},-\frac{\alpha}{4}}(\mathbb{R}^2_T) \to \widetilde{H}^{\frac{1}{2},\frac{\alpha}{4}}(\mathbb{R}^2_T)$ is an isomorphism.

In the analysis of the collocation scheme we will make use of the two following commutation relations. The first of them is the modification of the corresponding result for the single-layer operator for the heat equation [Hä98], Lemma 4.1.

Lemma 1. *For every $u \in \widetilde{H}^{1,1}(\mathbb{R}^2_T)$ there holds the commutation relation*

$$\partial_t V u = V \partial_t u.$$

Proof. Let us assume that u is a smooth function; then integrating by parts we will get

$$\begin{aligned} V(\partial_t u)(\theta,t) &= \rho \int_0^1 \int_0^t E(x(\theta)-x(\phi),t-\tau)\partial_\tau u(\phi,\tau)\mathrm{d}\tau\mathrm{d}\phi \\ &= \rho \int_0^1 \Big[\Big/_0^t E(x(\theta)-x(\phi),t-\tau)u(\phi,\tau)\mathrm{d}\tau \Big]\mathrm{d}\phi \\ &\qquad -\rho \int_0^1 \int_0^t \partial_\tau E(x(\theta)-x(\phi),t-\tau)u(\phi,\tau)\mathrm{d}\tau\mathrm{d}\phi. \end{aligned}$$

Now for the fundamental solution of the time-fractional diffusion equation there holds the asymptotic estimate [EiKo04]

$$|E(z,t-\tau)| \leq C|t-\tau|^{-1}\exp\{-\sigma z^{\frac{2}{2-\alpha}}|t-\tau|^{-\frac{\alpha}{2-\alpha}}\},$$

when $t-\tau \to 0^+$ and $z \neq 0$. Using this estimate and the initial condition $u(\phi,0)=0$, we obtain

$$V(\partial_t u)(\theta,t) = -\rho \int_0^1 \int_0^t \partial_\tau E(x(\theta)-x(\phi),t-\tau)u(\phi,\tau)\mathrm{d}\tau\mathrm{d}\phi.$$

The statement follows now from the equation

$$(\partial_t + \partial_\tau)E(x(\theta)-x(\phi),t-\tau) = 0.$$

Let us define the operator $\underline{\partial}_\theta$ by setting

$$\underline{\partial}_\theta u(\theta,t) = \partial_\theta u(\theta,t) + Ju(t),$$

where $Ju(t) = \int_0^1 u(\theta,t)\mathrm{d}\theta$. Clearly, this operator extends to an isomorphism from $\widetilde{H}^{r,s}(\mathbb{R}^2_T)$ onto $\widetilde{H}^{r-1,s}(\mathbb{R}^2_T)$ for all $r,s \in \mathbb{R}$. Denote its inverse by $\underline{\partial}_\theta^{-1}$.

Next, we introduce the operators

$$Ku = (V - V_0)u, \quad \underline{K} = \underline{\partial}_\theta K \underline{\partial}_\theta^{-1}, \quad \underline{V}u = V_0 u + \underline{K}u.$$

Here the operator V_0 is the single-layer operator on the circle with radius $\frac{\rho}{2\pi}$ having the parametric representation $x_0(\theta) = \frac{\rho}{2\pi}(\cos(2\pi\theta), \sin(2\pi\theta))$. Since the single-layer operator on a circle is the convolution operator in the spatial variable, it is clear that V_0 and $\underline{\partial}_\theta$ commute, i.e.,

$$\underline{\partial}_\theta V_0 = V_0 \underline{\partial}_\theta, \quad u \in C_{00}^{1,0}(\overline{\mathbb{R}}_T^2).$$

Hence, we obtain

$$\underline{V}u = \underline{\partial}_\theta V \underline{\partial}_\theta^{-1}.$$

Finally, we can prove the following assertion (see [Hä98] and [KeRuII]).

Theorem 3. *The operator $\underline{K}$ extends to a bounded operator of $\widetilde{H}^{s,\frac{\alpha}{2}s}(\Sigma_T)$ into $\widetilde{H}^{s+2,\frac{\alpha}{2}(s+2)}(\Sigma_T)$ if $s \geq -\frac{1}{\alpha}$.*

21.4 The Spline Collocation Method

21.4.1 Formulation of the Problem

Assuming that the right-hand side of the single-layer operator equation is continuous, then the collocation problem can be stated as: *Find $u_\Delta \in M^1$ such that*

$$Vu_\Delta(\theta_n, t_m) = Vu(\theta_n, t_m), \quad 1 \leq n \leq N, \ 1 \leq m \leq M.$$

The collocation problem is well defined provided both Vu_Δ and Vu are continuous functions.

Lemma 2. *Let $u \in \widetilde{H}^{1,1}(\mathbb{R}_T^2)$. Then*

(i) $Vu \in C_{00}^{1,0}(\overline{\mathbb{R}}_T^2)$;
(ii) $\partial_t Vu \in \widetilde{H}^{2,0}(\mathbb{R}_T^2)$;
(iii) $\partial_t \partial_\theta Vu \in \widetilde{H}^{1,0}(\mathbb{R}_T^2)$.

Proof. Using the commutation relation, we get

$$Vu = \partial_t^{-1} V \partial_t u, \quad u \in \widetilde{H}^{1,1}(\mathbb{R}_T^2),$$

where the operator ∂_t^{-1} is defined by setting

$$\partial_t^{-1} u(\theta, t) = \int_0^t u(\theta, \tau) \mathrm{d}\tau.$$

Obviously, the inverse operator has the following properties:

$$\partial_t^{-1}\partial_t u = 0,\ u \in \widetilde{H}^{0,1}(\mathbb{R}^2_T), \quad \partial_t^{-1}:\ \widetilde{H}^{r,k}(\mathbb{R}^2_T) \to \widetilde{H}^{r,k+1}(\mathbb{R}^2_T).$$

Hence, we have the following representation for V:

$$Vu = \partial_t^{-1}\underline{\partial}_\theta^{-1}[\underline{\partial}_\theta V \underline{\partial}_\theta^{-1}]\underline{\partial}_\theta \partial_t u, \quad u \in \widetilde{H}^{1,1}(\mathbb{R}^2_T).$$

Now it can be shown [KeRuI] that the operator $\underline{V} = \underline{\partial}_\theta^{-1} V \underline{\partial}_\theta$ is a continuous mapping from $\widetilde{H}^{r,\frac{\alpha}{2}r}(\mathbb{R}^2_T)$ into $\widetilde{H}^{r+1,\frac{\alpha}{2}(r+1)}(\mathbb{R}^2_T)$. Hence, for every $u \in \widetilde{H}^{1,1}(\mathbb{R}^2_T)$ we have $\underline{V}\partial_t\underline{\partial}_\theta u \in \widetilde{H}^{1,\frac{\alpha}{2}}(\mathbb{R}^2_T) \subset \widetilde{H}^{1,0}(\mathbb{R}^2_T)$, and therefore,

$$Vu = \partial_t^{-1}\underline{\partial}_\theta^{-1}\underline{V}\partial_t\underline{\partial}_\theta u \in \widetilde{H}^{2,1}(\mathbb{R}^2_T) \subset C^{1,0}_{00}(\overline{\mathbb{R}}^2_T).$$

The statements *(ii)* and *(iii)* follow from the relation $Vu \in \widetilde{H}^{2,1}(\mathbb{R}^2_T)$ and the mapping properties of the derivative operators.

Since the tensor product splines $M^1 \subset \widetilde{H}^{1,1}(\mathbb{R}^2_T)$, the collocation problem is well defined whenever the right-hand side of the equation $Vu = f$ is continuous.

21.4.2 Galerkin Formulation

For the unique solvability and stability of the collocation problem we will proceed as in [HaSa94], where for the spline collocation problem of the single-layer heat equation an equivalent Galerkin formulation was formulated by means of the integration by parts trick. For this we will define the operator

$$\underline{V}_\Delta u = \underline{\partial}_{\theta,\Delta} V \underline{\partial}_\theta^{-1},$$

where $\underline{\partial}_{\theta,\Delta}$ is the approximation of the operator $\underline{\partial}_\theta$ defined by

$$\underline{\partial}_{\theta,\Delta} u(\theta,t) = \partial_\theta u(\theta,t) + \sum_{n=0}^{N-1} \frac{\theta_{n+1}-\theta_{n-1}}{2} u(\theta_n,t).$$

In *the modified Galerkin method* we will find a function $u_\Delta \in M^1$ such that

$$\langle \underline{V}_\Delta \partial_t \underline{\partial}_\theta u_\Delta, \partial_t \underline{\partial}_\theta v\rangle = \langle \underline{V}_\Delta \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v\rangle \quad \forall v \in M^1,$$

where u is the solution of the single-layer operator equation.

The following theorem is a slight modification of Theorem 3.1 in [HaSa94], and we will omit its proof.

Theorem 4. *Let $u \in \widetilde{H}^{1,1}(\mathbb{R}^2_T)$ be the solution of the single-layer operator equation. Then solution of the collocation problem solves the modified Galerkin problem and vice versa.*

21.4.3 Stability and the Error Analysis

For the stability analysis we will need the following discrete norm in the approximation space M^1, which will be defined by setting

$$|\!|\!| u |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}} = \|\partial_t \underline{\partial}_\theta u\|_{-\frac{1}{2},-\frac{\alpha}{4};T}.$$

As usual, the stability of the Galerkin method relies on the Lax–Milgram lemma, which is valid if the corresponding bilinear form is coercive and continuous. Hence, we have the following.

Theorem 5. *Assume that $u \in \widetilde{H}^{1,1}(\mathbb{R}^2_T)$ be the solution of the single-layer equation $Vu = f$. Then for all $0 < h_\theta, h_t \le h_0$ there exists a unique solution $u_\Delta \in M^1$ to the collocation equations. Furthermore, we have the quasi-optimal error estimates*

$$|\!|\!| u - u_\Delta |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}} \le C \inf_{v \in M^1} |\!|\!| u - v |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}}.$$

Proof. Let us define the bilinear form $a(u,v) = \langle \underline{V} \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle$. By the mapping properties of the operator $\underline{V} = \underline{\partial}_\theta V \underline{\partial}_\theta^{-1}$, the bilinear form is continuous; that is,

$$|a(u,v)| \le C \|\partial_t \underline{\partial}_\theta u\|_{-\frac{1}{2},-\frac{\alpha}{4}} \|\partial_t \underline{\partial}_\theta v\|_{-\frac{1}{2},-\frac{\alpha}{4}}.$$

On the other hand, since the bilinear form can be decomposed as

$$\begin{aligned} a(u,v) &= \langle \underline{V} \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle \\ &= \langle V_0 \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle + \langle \underline{\partial}_\theta (V - V_0) \partial_t u, \partial_t \underline{\partial}_\theta v \rangle, \end{aligned}$$

where $V - V_0$ is a compact mapping and V_0 coercive, we get the inf-sup condition [BaAz72]

$$\inf_{0 \ne u \in M^1} \sup_{0 \ne v \in M^1} \frac{\langle \underline{V} \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle}{|\!|\!| u |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}} |\!|\!| v |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}}} \ge C > 0.$$

Let us now approximate the bilinear form $a(u,v) = \langle \underline{V} \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle$ with

$$a_\Delta(u,v) = \langle \underline{V}_\Delta \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle.$$

By the approximation properties of the trapezoidal rule, we have

$$\langle (\underline{V} - \underline{V}_\Delta) \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle| \le C h_\theta |\!|\!| u |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}} |\!|\!| v |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}}.$$

Hence, for sufficiently small h_0, the bilinear form $a_\Delta(u,v)$ is continuous in the discrete norm $|\!|\!| \cdot |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}}$ and satisfies the inf-sup condition

$$\inf_{0 \ne u \in M^1} \sup_{0 \ne v \in M^1} \frac{\langle \underline{V}_\Delta \partial_t \underline{\partial}_\theta u, \partial_t \underline{\partial}_\theta v \rangle}{|\!|\!| u |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}} |\!|\!| v |\!|\!|_{-\frac{1}{2},-\frac{\alpha}{4}}} \ge C > 0.$$

The unique solvability and quasi-optimality now follow by the standard techniques of Galerkin methods.

21.4.4 Error Analysis

From now on we will assume that the solution of the single-layer equation $u \in \widetilde{H}^{2,1}(\mathbb{R}^2_T) \cap \widetilde{H}^{1,2}(\mathbb{R}^2_T)$. By the approximation properties of the spline spaces, we will get the following lemma [KeRuII].

Lemma 3. *Let $u \in \widetilde{H}^{2,1}(\mathbb{R}^2_T) \cap \widetilde{H}^{1,2}(\mathbb{R}^2_T)$ and $P_{\theta,t} : L^2(\mathbb{R}^2_T) \to M^1$ be the orthogonal projection. Then we have the error estimate*

$$\|\!|u - P_{\theta,t}u|\!\|_{-\frac{1}{2},-\frac{\alpha}{4}} \le Ch_\theta^{\frac{3}{2}} \|u\|_{2,1} + Ch_t^{1+\frac{\alpha}{4}} \|u\|_{1,2}.$$

The quasi-optimality of the solution of the collocation problem and the previous lemma finally yield the error estimate.

Theorem 6. *Assume that $0 < h_\theta, h_t < h_0$. Let $u_\Delta \in M^1$ be the solution of the collocation equations and $u \in \widetilde{H}^{2,1}(\mathbb{R}^2_T) \cap \widetilde{H}^{1,2}(\mathbb{R}^2_T)$ be the solution of the single-layer equation $Vu = f$. Then*

$$\|\!|u - u_\Delta|\!\|_{-\frac{1}{2},-\frac{\alpha}{4}} \le C_1 h_\theta^{\frac{3}{2}} \|u\|_{2,1} + C_2 h_t^{1+\frac{\alpha}{4}} \|u\|_{1,2}.$$

References

[BaAz72] Babuška, I., Aziz, A.K.: Survey lectures on the mathematical foundations of the finite element method, in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations,* Aziz, A.K., ed., Academic Press (1972).

[Co92] Costabel, M.: Boundary integral operators for the heat equation. *Integral Equations Oper. Theory,* **13**, 498–552 (1992).

[CoSa00] Costabel, M., Saranen, J.: Spline collocation for convolutional parabolic boundary integral equations. *Numer. Math.,* **84**, 417–449 (2000).

[CoSa01] Costabel, M., Saranen, J.: Parabolic boundary integral operators, symbolic representations and basic properties. *Integral Equations Oper. Theory,* **40**, 185–211 (2001).

[EiKo04] Eidelman, S.D., Kochubei, A.N.: Cauchy problem for fractional diffusion equation. *J. Differential Equations,* **199**, 211–255 (2004).

[ElSc85] Elschner, J., Schmidt, G.: On spline interpolation in periodic Sobolev spaces. Preprint 01/83, Dept. Math. Akademie der Wissenschaften der DDR, Berlin (1985).

[HaSa94] Hamina, M., Saranen, J.: On the spline collocation method for the single-layer heat operator equation. *Math. Comp.,* **62**, 41–64 (1994).

[HsSa89] Hsiao, G.C., Saranen, J.: Coercivity of the single-layer heat operator. Report 89-2, Center for Mathematics and Waves, University of Delaware (1989).

[HsSa93] Hsiao, G.C., Saranen, J.: Boundary integral solution of the two-dimensional heat equation. *Math. Methods Appl. Sci.,* **16**, 87–114 (1993).

[Hä98] Hämäläinen J.: Spline collocation for the single-layer heat equation. *Ann. Acad. Sci. Fenn.,* Mathematica Dissertationes **113** (1998).

[KeRuI] Kemppainen, J., Ruotsalainen, K.: Boundary integral solution of the time-fractional diffusion equation. This volume, 213–222.

[KeRuII] Kemppainen, J., Ruotsalainen, K.: Numerical approximation of the boundary integral equations for two-dimensional fractional diffusion equations (in preparation).

[KiSa04] Kilbas, A.A., Saigo, M.: *H-transforms: Theory and Applications,* CRC Press, Boca Raton, FL (2004).

[LiMaI72] Lions, J.-L., Magenes, E.: *Non-homogeneous Boundary Value Problems and Applications. Vol. I,* Springer, Berlin (1972).

[LiMaII72] Lions, J.-L., Magenes, E.: *Non-homogeneous Boundary Value Problems and Applications. Vol. II,* Springer, Berlin (1972).

[PBM90] Prudnikov, A.P., Brychkov, Y.A., Marichev, O.I.: *Integrals and Series. Vol. 3. More Special Functions,* Overseas Publishers Association, Amsterdam (1990).

22

Wavelet-Based Hölder Regularity Analysis in Condition Monitoring

V. Kotila, S. Lahdelma, and K. Ruotsalainen

University of Oulu, Finland; vesa.kotila@oulu.fi, sulo.lahdelma@oulu.fi, keijo.ruotsalainen@ee.oulu.fi

22.1 Introduction

Condition monitoring is becoming more and more important in various areas of industry, due to the demands of efficiency and prolonged continuous running time of machinery. For example, in the Finnish pulp industry there have been demands for continuous running times of up to 18 months. To be cost efficient, maintenance operations should be carried out during scheduled downtime; hence, early and reliable fault detection is very important.

Vibration measurements have been the central tool in condition monitoring. Signals from displacement, velocity, and acceleration sensors have been used to estimate the condition of the machinery. For example, increased root-mean-square (RMS) values or changes in the frequency spectrum may indicate different types of faults, such as unbalance, misalignment, and bearing defects.

In rolling element bearings, a local fault on the raceways or on the rolling elements causes wideband bursts in the vibration signal measured from the bearing house. When the fault is on the inner race, the time interval between the bursts corresponds to the shaft frequency. If the shaft is rotating slowly, as in pulp washers, these bursts occur at long intervals and may be hard to detect from the frequency spectrum or the RMS value of the signal.

It has been reported that in some cases higher time derivatives of the displacement are more sensitive to certain faults than the velocity $\dot{x}$ or acceleration $\ddot{x}$. In a case study, a fault on a roller bearing inner race produced the largest relative peak value, compared to that of an intact bearing, when the fractional order of the time derivative was 4.75 [LaKo03]. It is then natural to assume that at least some faults result in reduced regularity of the vibration signal.

Another phenomenon where sudden bursts can be detected from the acceleration signal is cavitation in water turbines. Cavitation occurs when the local water pressure falls below the vaporization point and gas bubbles are formed. As the bubbles collapse, shock waves are created which detach metal

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_22,

from the turbine blades. It has been suggested that cavitation bursts have the form of a chirp.

The aim of our study is to find if wavelet methods can be applied to these problems in condition monitoring.

22.2 Hölder Regularity

First we summarize some definitions regarding Hölder regularity.

A function f is said to possess a Hölder exponent α at a point x_0, if there exists a polynomial P_n of degree $n \leq \alpha$, such that

$$|f(x) - P_n(x - x_0)| \leq C|x - x_0|^\alpha \tag{22.1}$$

when x is close to x_0. Polynomial P_n is typically the Taylor polynomial of f. The corresponding function class is denoted by $C^\alpha(x_0)$. The supremum of all values of α such that inequality (22.1) is valid is called the Hölder regularity of f at x_0.

If inequality (22.1) holds for all x and x_0 on an interval, α is called the uniform Hölder exponent of f.

For tempered distributions of finite order, if α is not an integer, f has a uniform Hölder exponent α if and only if the primitive of f is uniformly $\alpha+1$ on the same interval.

22.3 Wavelet Characterization of Hölder Regularity

The local Hölder regularity of a function f is characterized by the decay of its continuous wavelet transform.

A wavelet, as usual, is a function ψ satisfying an admissibility condition

$$\int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega = \int_{-\infty}^0 \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty.$$

The continuous wavelet transform is now the inner product of f against ψ translated by x and dilated by scale s, and normalized by a factor $\frac{1}{s}$,

$$Wf(s,x) = \frac{1}{s} \int_{-\infty}^\infty f(t)\psi(\frac{t-x}{s})dt.$$

A wavelet ψ is said to have N vanishing moments if the inner product with polynomials of degree at most $N-1$ is zero, that is,

$$\int_{-\infty}^\infty x^k \psi(x)dx = 0, \quad k = 0, 1, \ldots, N-1.$$

Let ψ be an admissible real-valued wavelet which is n-times continuously differentiable and has n vanishing moments and compact support.

The following theorem of Jaffard gives a necessary and sufficient condition for f to be of Hölder regularity α at a given point x_0 [Ja89].

Theorem 1. *Let $n \in \mathbb{N}$ and $f \in L^2(\mathbb{R})$. If $\alpha \leq n$ and $f \in C^\alpha(x_0)$, then for any scale s,*

$$|Wf(s,x)| \leq A(s^\alpha + |x - x_0|^\alpha).$$

On the other hand, if $\alpha < n$, $\alpha \notin \mathbb{Z}$, and if for any scale s and all x in a neighborhood of x_0,

$$\exists \epsilon > 0, A > 0 : |Wf(s,x)| \leq As^\epsilon$$
$$\exists B > 0 : |Wf(s,x)| \leq B(s^\alpha + \frac{|x - x_0|^\alpha}{|\log|x - x_0||}),$$

then $f \in C^\alpha(x_0)$.

In practice, a widely used method by Mallat and Hwang [MaHw92] relates local Hölder regularity only to the local extrema ridges of the wavelet transform. A ridge is a series of local maxima through the time-scale half-plane. In the case of an isolated singularity, the Hölder regularity can be estimated from the maxima values along a ridge pointing to x_0. Assume that the wavelet ψ has a compact support, n continuous derivatives, and itself is the nth derivative of a smoothing function, for example, a B-spline. The following theorem characterizes the Hölder regularity of an isolated singularity [MaHw92].

Theorem 2. *If there exists a scale s_0, an interval $]a,b[$, and a constant C such that for all $x \in]a,b[$ and $s < s_0$, all modulus maxima of $Wf(s,x)$ belong to the cone*

$$|x - x_0| < Cs, \tag{22.2}$$

then $f(x)$ has a uniform Hölder exponent n in a neighborhood of any $x_1 \in]a,b[$, $x_1 \neq x_0$. Function f belongs to $C^\alpha(x_0)$ if and only if the wavelet transform of f satisfies

$$|Wf(s,x)| \leq As^\alpha \tag{22.3}$$

on the ridges inside the cone (22.2).

We describe a procedure to estimate the location and the Hölder exponent of an isolated singularity.

1. Compute the wavelet transform.
2. Find the local maxima and minima at each scale.

3. Find and register the ridges by following the maxima and minima from fine scales to coarser scales.
4. Estimate the points where each ridge would end at the time axis by fitting a line at fine scales.
5. Of all ridges pointing to x_0, choose the one with the largest maximum value on the finest available scale.
6. Take the logarithm of the bounding inequality

$$\log(|Wf(s,x)|) \leq \log(A) + \alpha \log(s),$$

and estimate the regularity from the slope of a least-squares line.

An illustration of the procedure with a synthetic signal is plotted in Figure 22.1, with regularity $\alpha = 0.8$ at location $x = 1$ and regularity $\alpha = 1.6$ at location $x = 3$.

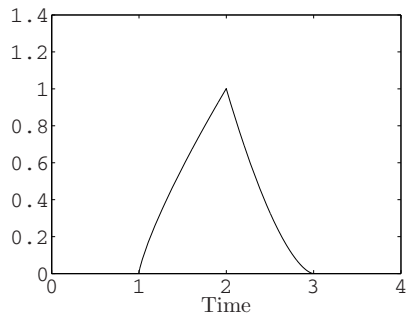


Fig. 22.1. Synthetic signal with regularity 0.8 at $x = 1$ and regularity 1.6 at $x = 3$.

The wavelet used is the second derivative of the Gaussian function, which is essentially supported on the interval $[-5, 5]$ and has two vanishing moments. The finest scale used is normalized as 1. In the wavelet transform, there are two ridges pointing to each singularity at the time axis, as can be seen in Figure 22.2. Let us study the ridges pointing to $x_0 = 1$.

From the doubly logarithmic plot of the modulus maxima values versus scale in Figure 22.3 it can be seen that the maxima values obey the predicted upper bound of Theorem 2. On finer scales, the effects of limited precision start to show. A linear least-squares fit gives slope 0.79, which is correct to one decimal. Similarly, at location 3 the estimate is 1.62.

22.4 Chirps

A class of singularities that are not isolated is that of chirps. In signal processing, a chirp means a short signal with either increasing or decreasing instantaneous frequency. We follow here the definition of Jaffard and Meyer [JaMe96].

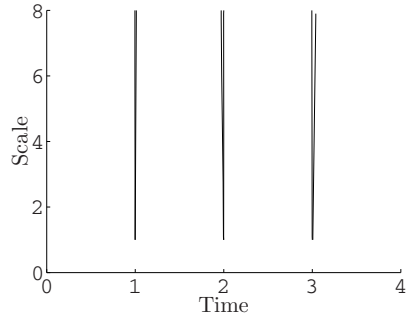


Fig. 22.2. The ridges of the wavelet transform of the signal in Figure 22.1.

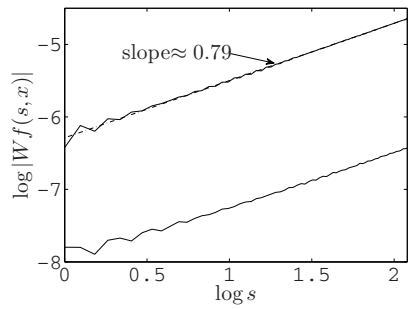


Fig. 22.3. Log-log plot of the modulus of the wavelet transform along the ridges pointing to $x_0 = 1$, as plotted in Figure 22.2.

Definition 1. *Let $\alpha > -1$ and $\beta > 0$. A function f, integrable in a neighborhood $[-\eta, \eta]$ of zero, is a generalized chirp of type (α, β), if*

$$f(x) = x^{\alpha} g_{+}(x^{-\beta}) \quad \text{if } 0 < x < \eta \tag{22.4}$$

$$f(x) = |x|^{\alpha} g_{-}(|x|^{-\beta}) \quad \text{if } -\eta < x < 0, \tag{22.5}$$

where g_- and g_+ are indefinitely oscillating functions on $[\eta^{-\beta}, \infty[$.

These chirps can be characterized by the decay of wavelet coefficients, modulo a smooth residual term [JaMe96]. Let ψ be a wavelet that belongs to the Schwartz class $S(\mathbb{R})$ of smooth and rapidly decreasing functions and has all vanishing moments. Further, assume that the dual wavelet $\tilde{\psi}$ exists, has N vanishing moments, and is smooth and compactly supported.

Theorem 3. *Let f be integrable in a neighborhood of the origin. Then*

$$f(x) = |x|^{\alpha} g_{\pm}(|x|^{-\beta}) + r(x), \quad r(x) \in C^{\infty}$$

with $g_{\pm} \in C^{r}([T, \infty[)$, if and only if there exists $\delta > 0$ such that

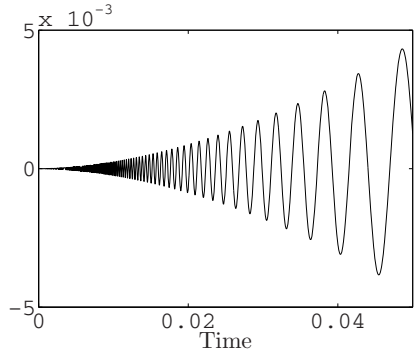


Fig. 22.4. A chirp with $\alpha = 1.8$ and $\beta = 1.2$.

$$|Wf(s,x)| \leq C|x|^{\alpha}\Big(\frac{s}{|x|^{1+\beta}}\Big)^{r} \quad \textit{if } 0 < s \leq x^{1+\beta} \leq \delta \tag{22.6}$$

$$|Wf(s,x)| \leq C_m|x|^{\alpha}\Big(\frac{|x|^{1+\beta}}{s}\Big)^{m} \quad \forall m, \textit{ if } x^{1+\beta} \leq s \leq |x| \leq \delta \tag{22.7}$$

$$|Wf(s,x)| \leq C_m s^m \quad \forall m, \textit{ if } |x| \leq s < \delta. \tag{22.8}$$

Inequalities (22.6) and (22.7) imply that upper bounds reach a maximum along the curve $s = |x|^{1+\beta}$. If such a ridge can be found from the wavelet transform, parameters α and β can be estimated:

- The ridge curve becomes a straight line in logarithmic coordinates, with slope $1 + \beta$.
- α is the slope of ridge height versus x in logarithmic coordinates.

We show an illustration with a function with $\alpha = 1.8$ and $\beta = 1.2$ plotted in Figure 22.4.

The wavelet used is the first derivative of the Mexican hat function, which does not exactly satisfy all of the given conditions, but which is compensated by not having to go to the finest scales. The modulus of the wavelet transform is plotted in Figure 22.5, with the maxima ridge (dark wavy curve) superimposed on it.

The parameter β can be approximated from the log-log data of time and scale along the ridge, plotted in Figure 22.6. The slope of a least-squares line is 2.19, giving the value 1.19 for β.

Similarly, the parameter α can be approximated from the log-log data of time and the modulus maxima along the ridge (Figure 22.7). A value of 1.79 is obtained for α.

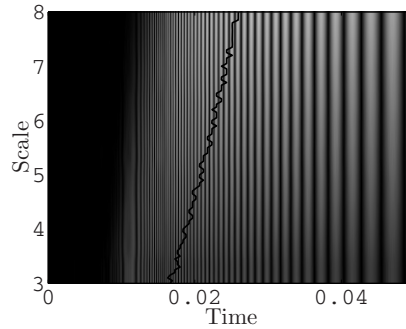


Fig. 22.5. Modulus maxima ridge of the wavelet transform of the signal in Figure 22.4.

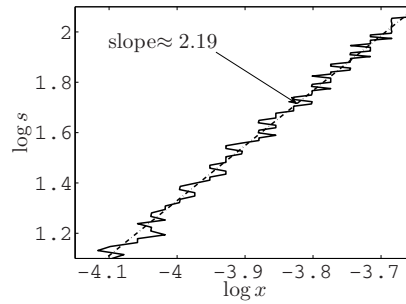


Fig. 22.6. Log-log plot of the ridge curve in Figure 22.5.

22.5 Hölder Regularity for Roller Bearings

Four double-row spherical roller bearings of type SKF 24124, typically found in felt guide rolls of a paper machine, were used in experimental equipment. Three of the bearings had inner race faults of variable degrees and one was intact. The acceleration signal (Figure 22.8) was measured from the bearing house of each bearing. Continuous wavelet transforms were computed for each signal using the second derivative of the Gaussian function (Figure 22.9). Only ridges above a threshold proportional to the L^2 norm of the signal were considered. The Hölder exponent was estimated from the slope of each ridge. From Figure 22.10, it can be seen that above the threshold, the negative exponents were distinctive to faulty bearings. Also, the locations of the negative exponents seem to correspond to where the rolling elements hit the fault.

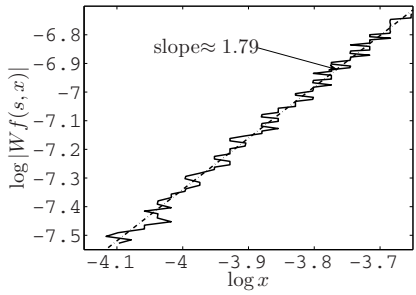


Fig. 22.7. Log-log plot of the ridge height versus time.

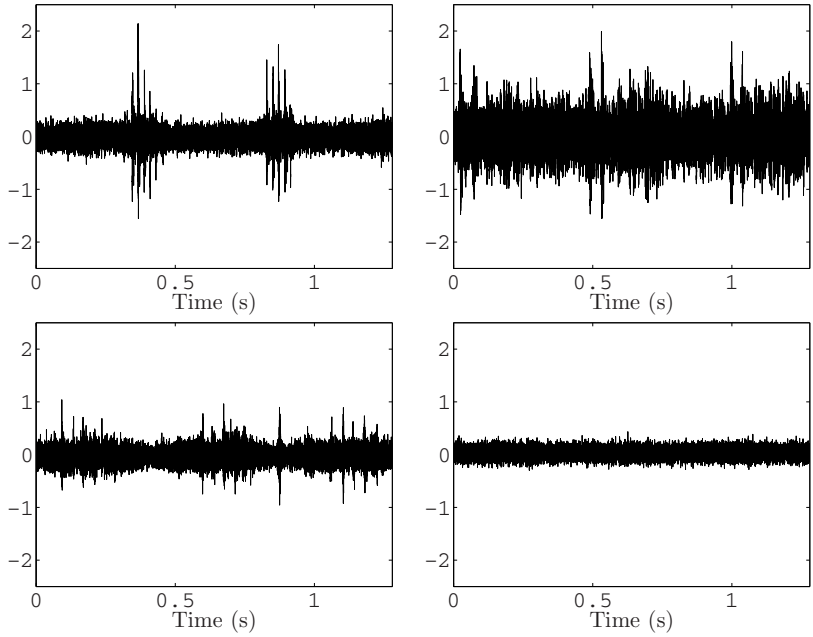


Fig. 22.8. Acceleration signals measured from the bearing house of four bearings. The signal in the lower right corner is from an intact bearing.

22.6 Discussion

The experiments clearly indicated that the local Hölder regularity of the vibration signal can be useful in condition monitoring of bearings. The slopes of the most prominent ridges of wavelet transform seem to reveal faults on the inner race, at least in the cases studied. Larger sets of test data with different kinds of bearing defects would give more precise results.

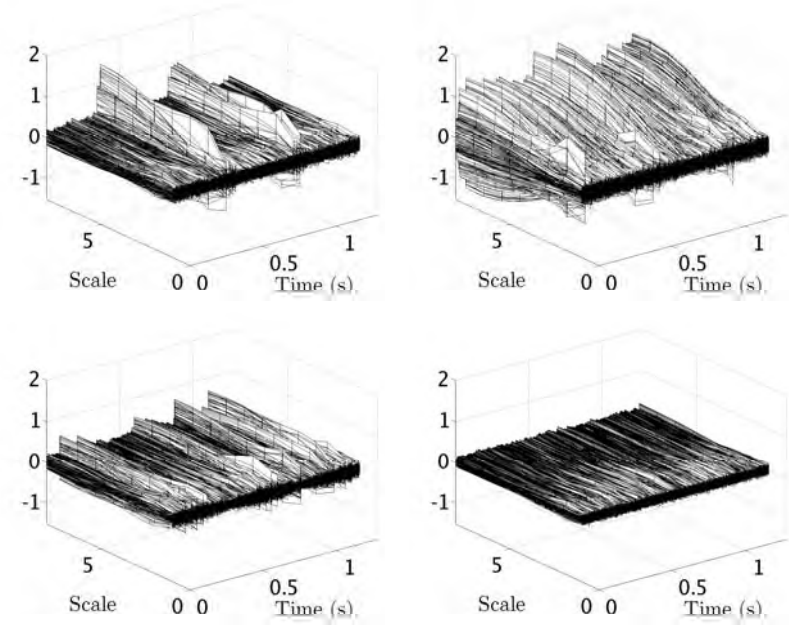


Fig. 22.9. Continuous wavelet transforms of the signals in Figure 22.8.

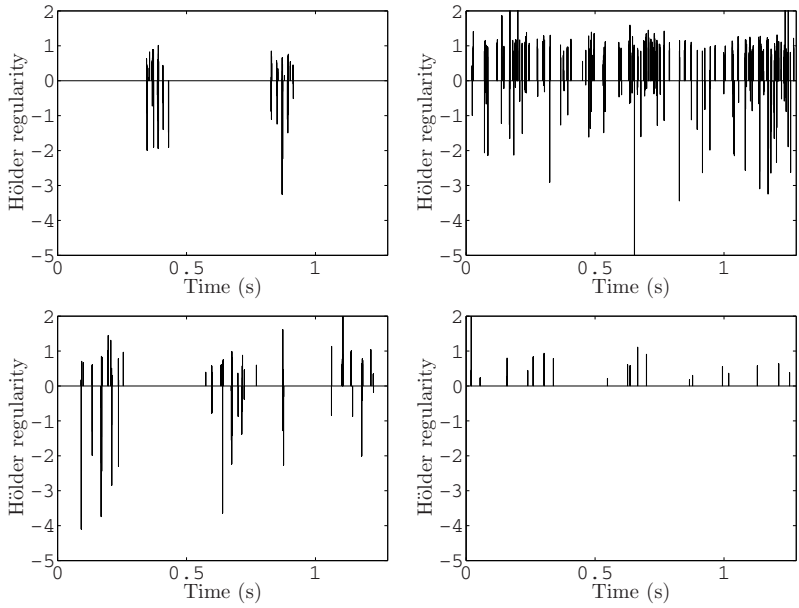


Fig. 22.10. Estimated local Hölder regularity of the signals in Figure 22.8.

References

[LaKo03] Lahdelma, S., Kotila, V.: Real order derivatives—new signal processing method. *Kunnossapito,* **17**, No. 8, 39–42 (2003) (Finnish).

[Ja89] Jaffard, S.: Exposants de Hölder en des points donnés et coefficients d'ondelettes. *C.R. Acad. Sci. Paris Sér. 1,* **308**, 79–81 (1989).

[MaHw92] Mallat, S., Hwang, W.L.: Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory,* **38**, 617–643 (1992).

[JaMe96] Jaffard, S., Meyer, Y.: Wavelet methods for pointwise regularity and local oscillations of functions. *Mem. Amer. Math. Soc.,* **123** (1996).

23

Integral Equation Technique for Finding the Current Distribution of Strip Antennas in a Gyrotropic Medium

A.V. Kudrin,[1] E.Yu. Petrov,[1] and T.M. Zaboronkova[2]

[1] University of Nizhny Novgorod, Russia; kud@rf.unn.ru, epetrov@rf.unn.ru
[2] Technical University of Nizhny Novgorod, Russia; zabr@nirfi.sci-nnov.ru

23.1 Introduction

Much previous work on the characteristics of wire antennas in gyrotropic media such as a magnetoplasma, for example, either applies to electrically small antennas for which the current distribution along the antenna wire can be assumed given [Ko99], or employs the transmission line theory for determining the current distribution (see [Ad77] and [Oh86]).

In this study, the problem of finding the current distribution of strip antennas in a homogeneous gyrotropic medium is attacked using an integral equation method. Although our approach is applicable to a general gyrotropic medium, our primary attention will be paid to the case of a resonant gyroelectric medium in which the refractive index of one of the characteristic waves tends to infinity when the angle between the wave normal direction and the gyrotropic axis approaches a certain value determined by the medium parameters. In this case, the classical thin-antenna theory cannot be employed readily since no matter how small the cross section of the antenna wire might be physically, it is always possible to find some wave normal direction for which one wavelength in the medium will become less than the wire cross-sectional extent and the antenna wire will appear to be "thick." We do not consider the antenna problem in its full generality, but focus on two particular strip geometries for which the problem is mathematically tractable.

23.2 Basic Formulation

Consider a homogeneous lossless gyrotropic medium described by the dielectric tensor

$$\boldsymbol{\varepsilon} = \epsilon_0 \begin{pmatrix} \varepsilon & -ig & 0 \\ ig & \varepsilon & 0 \\ 0 & 0 & \eta \end{pmatrix}, \tag{23.1}$$

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_23,

where ϵ_0 is the permittivity of free space. It is assumed that the gyrotropic axis is aligned with the z-axis. Note that in a gyroelectric medium, the direction of the gyrotropic axis coincides with that of a superimposed static magnetic field $\boldsymbol{B}_0$. The elements ε, g, and η of the tensor in (23.1) are functions of the medium parameters and the frequency ω of an electromagnetic field. Recall that a gyrotropic medium is resonant if $\operatorname{sgn}(\varepsilon) \neq \operatorname{sgn}(\eta)$, and is nonresonant otherwise [Ko99].

Two geometries of a perfectly conducting strip of width $2b$ excited by a time harmonic ($\sim \exp(i\omega t)$) voltage generator will be discussed. Firstly, we will consider a straight strip of infinite length, which is aligned with the x-axis and perpendicular to the y-axis. We assume that the current on such a strip antenna is excited by a given voltage V_0 that is applied over an interval $|x| \leq d$ and creates the field E_x^{ext} at $y = 0$ and $|z| \leq b$ (see Figure 23.1). Explicitly,

$$E_x^{\text{ext}}(x, 0, z) = \frac{V_0}{2d}\big[U(x+d) - U(x-d)\big]\big[U(z+b) - U(z-b)\big],$$

where U is the Heaviside step function. The current on the strip can be represented as

$$\boldsymbol{J} = \hat{x}_0 \delta(y) I(x, z),$$

where $\delta(y)$ is the Dirac delta and $I(x, z)$ is the surface current density.

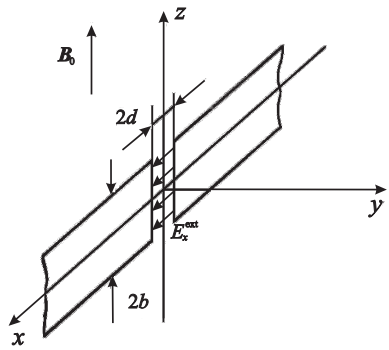


Fig. 23.1. Straight strip antenna.

Secondly, a strip coiled into a circular loop of radius a whose axis is parallel to the z-axis will be considered. Such an annular strip antenna is excited by a voltage V_0 which creates an electric field with the only nonzero azimuthal component E_ϕ^{ext} at $\rho = a$ and $|z| \leq b$ in the angular interval $|\phi - \phi_0| \leq \Delta \ll 2\pi$ (see Figure 23.2). Thus, for $\rho = a$,

$$E_\phi^{\text{ext}}(a, \phi, z) = \frac{V_0}{2a\Delta}\big[U(\phi - \phi_0 + \Delta) - U(\phi - \phi_0 - \Delta)\big]\big[U(z+b) - U(z-b)\big]. \tag{23.2}$$

Here, 2Δ is the angular opening of the interval to which the voltage is applied and ρ, ϕ, and z are cylindrical coordinates. The current of the annular strip antenna is written in the form

$$\boldsymbol{J} = \hat{\phi}_0 \delta(\rho - a) I(\phi, z),$$

where $I(\phi, z)$ is the surface current density.

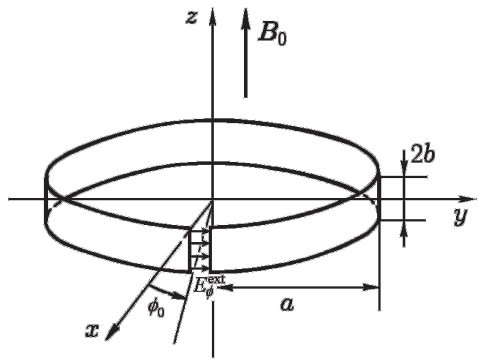


Fig. 23.2. Annular strip antenna.

To derive integral equations for the current distributions of the antennas, we should write the field excited by the unknown antenna current and ensure the required boundary conditions for the tangential components of the electric field on the surface of a perfectly conducting strip.

23.3 Field Representation

To find a formal representation of the electric field $\boldsymbol{E}$ due to the unknown current $\boldsymbol{J}$, we start from the Maxwell equations

$$\nabla \times \boldsymbol{E} = -i\omega \boldsymbol{B}, \quad \nabla \times \boldsymbol{B} = i\,\omega\mu_0\,\boldsymbol{\varepsilon} \cdot \boldsymbol{E} + \mu_0 \boldsymbol{J},$$

where μ_0 is the permeability of free space. Upon eliminating the magnetic field $\boldsymbol{B}$, one obtains

$$\nabla \times \nabla \times \boldsymbol{E} - \omega^2 \mu_0\,\boldsymbol{\varepsilon} \cdot \boldsymbol{E} = -i\,\omega\mu_0 \boldsymbol{J}. \tag{23.3}$$

We now go over to the Fourier-transformed version of (23.3). Our definition of the spatial Fourier transform of any function $F(\boldsymbol{r})$ is

$$F(\boldsymbol{n}) = \int F(\boldsymbol{r}) \exp(ik_0 \boldsymbol{n} \cdot \boldsymbol{r}) d\boldsymbol{r},$$

where $k_0 = \omega(\epsilon_0\mu_0)^{1/2}$ is the wave number in free space. The Fourier transform of (23.3) is written as

$$\mathbf{T} \cdot \boldsymbol{E}(\boldsymbol{n}) = -ik_0^{-1} Z_0 \, \boldsymbol{J}(\boldsymbol{n}).$$

Here, $Z_0 = (\mu_0/\epsilon_0)^{1/2}$ is the characteristic impedance of free space and $\mathbf{T}$ is a tensor whose elements T_{ij} $(i, j = x, y, z)$ are defined by

$$T_{ij} = n^2\delta_{ij} - n_i n_j - \epsilon_0^{-1}\varepsilon_{ij},$$

where δ_{ij} is the Kronecker delta. The Fourier-transformed equation can be solved formally for $\boldsymbol{E}(\boldsymbol{n})$ to yield

$$\boldsymbol{E}(\boldsymbol{n}) = -ik_0^{-1} Z_0 \mathbf{T}^{-1} \cdot \boldsymbol{J}(\boldsymbol{n}). \tag{23.4}$$

The elements of the inverse tensor $\mathbf{T}^{-1}$ are given by $T_{ij}^{-1} = C_{ij}/D$, where C_{ij} and D are the adjoint and determinant of the matrix T_{ij}, respectively. The determinant D is written as

$$D = -\eta \left[n_z^2 - n_{z,\mathrm{o}}^2(n_\perp)\right] \left[n_z^2 - n_{z,\mathrm{e}}^2(n_\perp)\right], \;\; n_\perp^2 = n_x^2 + n_y^2,$$

where

$$n_{z,\alpha}^2(n_\perp) = \varepsilon - \frac{1}{2}\Big(1 + \frac{\varepsilon}{\eta}\Big) n_\perp^2 + \chi_\alpha \left[\frac{1}{4}\Big(1 - \frac{\varepsilon}{\eta}\Big)^2 n_\perp^4 - \frac{g^2}{\eta} n_\perp^2 + g^2\right]^{1/2}. \tag{23.5}$$

Here, the subscript α stands for the "ordinary" ($\alpha = \mathrm{o}$) and "extraordinary" ($\alpha = \mathrm{e}$) characteristic waves of a gyrotropic medium, and $\chi_\mathrm{o} = -\chi_\mathrm{e} = -\mathrm{sgn}(1-\varepsilon/\eta)$. The square root in (23.5) is chosen to have the positive real part. The functions $n_{z,\alpha}(n_\perp)$ are defined to have the negative imaginary part. It is worth noting that the quantities $n_{z,\alpha}$ and $n_\perp$ are, respectively, the longitudinal and transverse components of the normalized (to k_0) propagation vector of the corresponding characteristic wave. It is easily verified that if $\mathrm{Re}\, n_\perp \to 0$, then $n_{z,\mathrm{e}}(n_\perp) \to \mathrm{sgn}(\varepsilon)\,(-\varepsilon/\eta)^{1/2} n_\perp$ and $n_{z,\mathrm{e}}(n_\perp) \to -i(\varepsilon/\eta)^{1/2} n_\perp$ in the resonant and nonresonant cases, respectively. For the "ordinary" wave in both cases, we have $n_{z,\mathrm{o}}(n_\perp) \to -in_\perp$ as $\mathrm{Re}\, n_\perp \to 0$.

23.4 Integral Equations for a Straight Strip Antenna

We start from a straight strip antenna whose surface current can be represented as

$$I(x, z) = \frac{k_0}{2\pi} \int_{-\infty}^{\infty} I(n_x, z) \exp(-ik_0 n_x x) dn_x. \tag{23.6}$$

To obtain integral equations for $I(n_x, z)$, we use the boundary conditions $E_x = -E_x^{\mathrm{ext}}$ and $E_z = 0$ on the strip surface. Allowing for (23.6), we find the Fourier-transformed electric field $\boldsymbol{E}(\boldsymbol{n})$ in (23.4) and take the inverse Fourier

transforms of the functions $E_x(\boldsymbol{n})$ and $E_z(\boldsymbol{n})$ in the form of integrals over n_x, n_y, and n_z. Observing that the corresponding integrands are singular at zeros of D, i.e., at $n_z = \pm n_{z,\mathrm{o}}(n_\perp)$ and $n_z = \pm n_{z,\mathrm{e}}(n_\perp)$, we perform integration over n_z using Cauchy's theorem of residues and write expressions for the Fourier-transformed (with respect to n_x) tangential components $E_x(n_x, y, z)$ and $E_z(n_x, y, z)$. Applying the boundary conditions for these quantities at $y = 0$ and $|z| < b$, we then obtain

$$\int_{-b}^{b} \mathcal{K}^{(x)}(n_x, z - z') I(n_x, z')\, dz' = -\frac{4\pi V_0}{Z_0 k_0} \frac{\sin(k_0 d n_x)}{k_0 d n_x}, \tag{23.7}$$

$$\int_{-b}^{b} \mathcal{K}^{(z)}(n_x, z - z')\, I(n_x, z') dz' = 0, \tag{23.8}$$

where $|z| < b$ and

$$\mathcal{K}^{(x)}(n_x, \zeta) = \sum_{\alpha=\mathrm{o}}^{\mathrm{e}} \chi_\alpha \frac{1}{\eta} \int_{-\infty}^{\infty} \frac{(n_\perp^2 - \eta)(n_\alpha^2 - \varepsilon)}{n_\perp^2 n_{z,\alpha}(n_{z,\mathrm{e}}^2 - n_{z,\mathrm{o}}^2)} \times \left[n_x^2 + \frac{g^2 n_y^2}{(n_\alpha^2 - \varepsilon)^2} \right] \exp(-ik_0 n_{z,\alpha} |\zeta|) dn_y, \tag{23.9}$$

$$\mathcal{K}^{(z)}(n_x, \zeta) = \mathrm{sgn}(\zeta) \sum_{\alpha=\mathrm{o}}^{\mathrm{e}} \chi_\alpha \int_{-\infty}^{\infty} \frac{n_x (n_\alpha^2 - \varepsilon)}{n_{z,\mathrm{e}}^2 - n_{z,\mathrm{o}}^2} \exp(-ik_0 n_{z,\alpha} |\zeta|) dn_y. \tag{23.10}$$

Hereafter, $n_\alpha^2 = n_\perp^2 + n_{z,\alpha}^2$. Thus, the problem of finding the current distribution has become one of solving integral equations (23.7) and (23.8) with kernels (23.9) and (23.10), respectively.

The behavior of the solutions of integral equations (23.7) and (23.8) is determined by the properties of their kernels $\mathcal{K}^{(x)}$ and $\mathcal{K}^{(z)}$. We will discuss in detail the case of a resonant medium and then give only the resulting formulas for the case of a nonresonant medium. It can be shown that kernels (23.9) and (23.10) can be divided into singular terms $K^{(x,z)}(n_x, \zeta)$ and nonsingular terms $F^{(x,z)}(n_x, \zeta)$:

$$\mathcal{K}^{(x,z)}(n_x, \zeta) = K^{(x,z)}(n_x, \zeta) + F^{(x,z)}(n_x, \zeta).$$

In the resonant case, the singular terms are represented as follows:

$$K^{(x)}(n_x, \zeta) = -2 \int_0^\infty \left[n_x^2 |\varepsilon\eta|^{-1/2} \cos\left(k_0 \sigma \mathrm{sgn}(\varepsilon) |\zeta| \sqrt{n_x^2 + n_y^2} \right) + i \exp\left(-k_0 |\zeta| \sqrt{n_x^2 + n_y^2} \right) \right] (n_x^2 + n_y^2)^{-1/2} dn_y,$$

$$K^{(z)}(n_x, \zeta) = 2 n_x \mathrm{sgn}(\zeta) \int_0^\infty \exp\left(-ik_0 \sigma \mathrm{sgn}(\varepsilon) |\zeta| \sqrt{n_x^2 + n_y^2} \right) dn_y,$$

where $\sigma = (-\varepsilon/\eta)^{1/2}$. The expressions for the nonsingular terms $F^{(x)}$ and $F^{(z)}$ are evident and are omitted for brevity.

If the strip is so narrow that the inequalities

$$b \ll 2d, \quad (k_0 b)^2 \max(|\varepsilon|, |g|, |\eta|) \ll 1$$

are satisfied, then the solutions of integral equations (23.7) and (23.8) can be found in analytical form. For the narrow strip, the kernels of these integral equations can be approximated, and one arrives at

$$\int_{-b}^{b} I(n_x, z') \ln \frac{k_0|z-z'|}{2} dz' = -\frac{2\pi V_0}{Z_0 k_0} \frac{\sin(k_0 d n_x)}{k_0 d n_x} \beta - S(n_x) \int_{-b}^{b} I(n_x, z') dz', \tag{23.11}$$

$$\int_{-b}^{b} \frac{I(n_x, z')}{z - z'} dz' = 0. \tag{23.12}$$

In the preceding equations,

$$S(n_x) = \beta \left[n_x^2 |\varepsilon\eta|^{-1/2} (\ln\sigma + \gamma + \ln|n_x|) + i \left(\gamma + \ln|n_x| + iF^{(x)}(n_x, 0)/2 \right) \right],$$

$$\beta = |\varepsilon\eta|^{-1/2} \left(n_x^2 + i\,|\varepsilon\eta|^{1/2} \right)^{-1},$$

where $\gamma = 0.5772\ldots$ is Euler's constant, and use is made of the fact that $F^{(z)}(n_x, 0) = 0$.

Equations (23.11) and (23.12) are approximate integral equations which can be solved exactly. Observe that the solution of equation (23.11) with the logarithmic kernel automatically satisfies equation (23.12) with a Cauchy singular kernel (see [Ga90] and [Vo74]). This circumstance allows us to consider only equation (23.11). Its solution, upon substituting into (23.6), yields

$$I(x, z) = -\frac{V_0}{Z_0 \pi \sqrt{b^2 - z^2}} \int_{-\infty}^{\infty} \frac{\sin(k_0 d n_x)}{k_0 d n_x} \frac{\beta \exp(-ik_0 n_x x)}{\ln(4/k_0 b) - S(n_x)} dn_x. \tag{23.13}$$

In the vicinity of the strip edges at $z = \pm b$, the surface current density exhibits the edge behavior consistent with the Meixner condition [Me72]. The total current $I_\Sigma(x)$ through the cross section $x = \text{const}$ of the strip is given by $I_\Sigma(x) = \int_{-b}^{b} I(x, z) dz$ and is evidently finite. Its closed-form expression can be obtained approximately if the strip is so narrow that the inequality $\ln(4/k_0 b) \gg S(n_x)$ is valid for $|n_x| < (k_0 d)^{-1}$. Then the term $S(n_x)$ can be neglected and the n_x integration in (23.13) is performed using the technique of contour integration. For $|x| > d$, we thus have approximately that

$$I_\Sigma(x) = \frac{\pi V_0 h}{Z_0 k_0 \ln(4/k_0 b)} \exp(-ih|x|), \tag{23.14}$$

where

$$h = k_0|\varepsilon\eta|^{1/4}(1-i)/\sqrt{2}. \tag{23.15}$$

Closed-form expression (23.14) for the current distribution refers to the case of a resonant medium. It can be shown that in the nonresonant case where $\operatorname{sgn}(\varepsilon) = \operatorname{sgn}(\eta)$, the expression for the current distribution is again given by (23.14) if the quantity h is taken in the form

$$h = \begin{cases} k_0(\varepsilon\eta)^{1/4} & \text{if } \varepsilon > 0 \text{ and } \eta > 0, \\ -ik_0(\varepsilon\eta)^{1/4} & \text{if } \varepsilon < 0 \text{ and } \eta < 0. \end{cases} \tag{23.16}$$

23.5 Integral Equations for an Annular Strip Antenna

We now turn to an annular strip antenna, as shown in Figure 23.2. To find the function $I(\phi, z)$, we should apply the boundary conditions $E_\phi = -E_\phi^{\text{ext}}$ and $E_z = 0$ on the antenna surface (for $\rho = a$ and $|z| < b$).

To derive representations for the E_ϕ and E_z components on the strip surface, we expand the unknown surface current density $I(\phi, z)$ into the Fourier series

$$I(\phi, z) = \sum_{m=-\infty}^{\infty} I_m(z)\exp(-im\phi). \tag{23.17}$$

Similarly, the quantity E_ϕ^{ext} given by (23.2) is expanded into the Fourier series

$$E_\phi^{\text{ext}} = \sum_{m=-\infty}^{\infty} A_m \exp(-im\phi),$$

where

$$A_m = \frac{V_0}{2\pi a}\frac{\sin(m\Delta)}{m\Delta}\exp(im\phi_0).$$

Then we calculate the quantities $J_x(\boldsymbol{n})$ and $J_y(\boldsymbol{n})$ corresponding to representation (23.17) and find $E_{x,y,z}(\boldsymbol{n})$ from (23.4). Evaluating the field components $E_\phi(\boldsymbol{r})$ and $E_z(\boldsymbol{r})$ and satisfying the boundary conditions for them on the strip surface, we get the integral equations

$$\int_{-b}^{b} \mathcal{K}_m^{(\phi)}(z-z')I_m(z')\,dz' = -\frac{2A_m}{Z_0 k_0^2 a}, \tag{23.18}$$

$$\int_{-b}^{b} \mathcal{K}_m^{(z)}(z-z')\,I_m(z')dz' = 0, \tag{23.19}$$

where $|z| < b$, $m = 0, \pm1, \pm2, \ldots$, and

$$\mathcal{K}_m^{(\phi)}(\zeta) = \sum_{\alpha=\text{o}}^{\text{e}} \frac{\chi_\alpha}{\eta}\int_0^\infty \left[\frac{m}{k_0 a n_\perp}J_m(k_0 a n_\perp) - \frac{g}{n_\alpha^2 - \varepsilon}J_m'(k_0 a n_\perp)\right]^2$$
$$\times \frac{n_\perp}{n_{z,\alpha}}\frac{(n_\perp^2-\eta)(n_\alpha^2-\varepsilon)}{n_{z,\text{e}}^2 - n_{z,\text{o}}^2}\exp(-ik_0 n_{z,\alpha}|\zeta|)\,dn_\perp, \tag{23.20}$$

$$\mathcal{K}_m^{(z)}(\zeta) = \operatorname{sgn}(\zeta) \sum_{\alpha=\mathrm{o}}^{\mathrm{e}} \frac{\chi_\alpha}{\eta} \int_0^\infty \left[\frac{m}{k_0 a n_\perp} J_m(k_0 a n_\perp) - \frac{g}{n_\alpha^2 - \varepsilon} J_m'(k_0 a n_\perp) \right]$$
$$\times J_m(k_0 a n_\perp) \frac{n_\perp^2 (n_\alpha^2 - \varepsilon)}{n_{z,\mathrm{e}}^2 - n_{z,\mathrm{o}}^2} \exp(-ik_0 n_{z,\alpha} |\zeta|)\, dn_\perp. \quad (23.21)$$

Here, J_m and J_m' stand for the mth-order Bessel function of the first kind and its first derivative with respect to the argument.

Equations (23.18) and (23.19), with the kernels given by (23.20) and (23.21), respectively, are singular integral equations for $I_m(z)$. As for a straight strip, the kernels can be represented as

$$\mathcal{K}_m^{(\phi,z)}(\zeta) = K_m^{(\phi,z)}(\zeta) + F_m^{(\phi,z)}(\zeta),$$

where the singular terms $K_m^{(\phi,z)}(\zeta)$ in the case of a resonant medium are written as

$$K_m^{(\phi)}(\zeta) = -\int_0^\infty \left[\frac{m^2}{(k_0 a)^2 |\varepsilon\eta|^{1/2}} J_m^2(k_0 a n_\perp) \exp(-ik_0 \sigma \operatorname{sgn}(\varepsilon) |\zeta| n_\perp) \right.$$
$$\left. + iJ_{m-1}^2(k_0 a n_\perp) \exp(-k_0 |\zeta| n_\perp) \right] dn_\perp, \quad (23.22)$$

$$K_m^{(z)}(\zeta) = \frac{m}{k_0 a \eta} \operatorname{sgn}(\zeta) \int_0^\infty n_\perp J_m^2(k_0 a n_\perp) \exp(-ik_0 \sigma \operatorname{sgn}(\varepsilon) |\zeta| n_\perp) dn_\perp, \quad (23.23)$$

while the regular terms $F_m^{(\phi,z)}$ are not present here in the interests of brevity.

The integrals in (23.22) and (23.23) can be expressed in terms of the Legendre functions of the first and second kinds, which can be approximated by simpler functions in the case of a narrow strip where the following conditions hold:

$$b \ll a, \quad b \ll a|\eta/\varepsilon|^{1/2}, \quad (k_0 b)^2 \max(|\varepsilon|, |g|, |\eta|) \ll 1. \quad (23.24)$$

As a result, we arrive at the approximate integral equations

$$\int_{-b}^{b} I_m(z') \ln \frac{|z - z'|}{2a} dz' = -\frac{2\pi A_m}{Z_0 k_0} \frac{(k_0 a)^2 |\varepsilon\eta|^{1/2}}{m^2 + i(k_0 a)^2 |\varepsilon\eta|^{1/2}} - S_m \int_{-b}^{b} I_m(z') dz', \quad (23.25)$$

$$\int_{-b}^{b} m \frac{I_m(z')}{z - z'} dz' = 0, \quad (23.26)$$

where $|z| < b$ and

$$S_m = \frac{1}{m^2 + i(k_0 a)^2 |\varepsilon\eta|^{1/2}} \left\{ m^2 \left[\ln \sigma + \gamma + \psi\left(m + \frac{1}{2}\right) + i\frac{\pi}{2} \operatorname{sgn}(\varepsilon) \right] \right.$$
$$\left. + i(k_0 a)^2 |\varepsilon\eta|^{1/2} \left[\gamma + \psi\left(m - \frac{1}{2}\right) - i\pi k_0 a\, F_m^{(\phi)}(0) \right] \right\}.$$

Here, $\psi(z) = d\ln\Gamma(z)/dz$ is the logarithmic derivative of the gamma function.

For each m, the solution of equation (23.25), which also satisfies equation (23.26), can be found by the methods presented in [Vo74], and is written as follows:

$$I_m(z) = \frac{2}{Z_0 k_0 \sqrt{b^2 - z^2}} \frac{(k_0 a)^2 |\varepsilon\eta|^{1/2}}{m^2 + i(k_0 a)^2 |\varepsilon\eta|^{1/2}} \frac{A_m}{\ln(4a/b) - S_m}.$$

The resultant expression for total current $I_\Sigma(\phi) = \int_{-b}^{b} I(\phi, z)dz$ can be written as

$$I_\Sigma(\phi) = a_0 + \sum_{m=1}^{\infty} \left\{2a_m \cos[m(\phi - \phi_0)] - (a_m - a_{-m}) \exp[im(\phi - \phi_0)]\right\}, \tag{23.27}$$

where

$$\begin{aligned} a_0 &= -\frac{iV_0}{Z_0 k_0 a} \frac{1}{\ln(4a/b) - 2 + 2\ln 2 + i\pi k_0 a\, F_m^{(\phi)}(0)}, \\ a_m &= \frac{iV_0}{Z_0 k_0 a} \frac{\sin(m\Delta)}{m\Delta} \frac{\alpha_m}{\ln(4a/b) - S_m}, \end{aligned}$$

in which

$$\alpha_m = -\frac{i(k_0 a)^2 |\varepsilon\eta|^{1/2}}{m^2 + i(k_0 a)^2 |\varepsilon\eta|^{1/2}}.$$

The last term in the braces in (23.27) appears due to the gyrotropy of a medium. It can be shown that in the case of a narrow strip, the last term gives the very small contribution to the total value of $I_\Sigma(\phi)$. A closed-form expression for the current distribution can be obtained if the strip is so narrow that $b \ll 2a\Delta \ll a$ and $\ln(4a/b) \gg S_m$ for $m < [\Lambda]$, where $\Lambda = \min(a/b, a/\sigma b)$ and the notation $[\Lambda]$ designates the integer part of Λ. Then the total current given by (23.27) can be evaluated approximately as

$$I_\Sigma(\phi) = -\frac{iV_0 \pi h}{Z_0 k_0 \ln(4a/b)} \frac{\cos[(\pi - \phi + \phi_0)ha]}{\sin(\pi h a)},$$

where $0 \le \phi - \phi_0 \le 2\pi$ and the quantity h coincides with that in (23.15). The above formula, obtained for the antenna current $I_\Sigma(\phi)$ in a resonant gyrotropic medium, remains valid in the case of a nonresonant medium if h is taken in the form given by (23.16).

23.6 Conclusions

We considered the problem of finding the current distribution on perfectly conducting strips in a gyrotropic medium described by the off-diagonal permittivity tensor. Primary attention has been focused on the case of a resonant medium in which the refractive index surface of one characteristic wave

extends to infinity at a certain wave normal direction. For two special geometries, the problem was reduced to a set of singular integral equations. Based on the results of the solution of these equations, closed-form expressions have been derived for the current distribution along the antenna surface. It should be noted that this work differs from that of other workers, who considered similar problems (e.g., [Ad77] and [Oh86]), in that we have started from a full-wave treatment of the problem and have obtained a solution by the use of an integral equation method. Finally, we note that the theory developed in this chapter makes it possible to establish the applicability conditions of the transmission line theory, which was used in earlier papers for describing the current distribution on narrow strips in a gyrotropic medium.

Acknowledgement. This work was supported by the Russian Foundation for Basic Research (project Nos. 08–02–97026-a and 09–02–00164-a), the Program for the State Support of the Leading Scientific Schools of the Russian Federation (project No. NSh–5180.2008.2), and the Dynasty Foundation (Russia).

References

[Ad77] Adachi, S., Ishizone, T., Mushiake, Y.: Transmission line theory of antenna impedance in magnetoplasma. *Radio Sci.,* **12**, 23–31 (1977).

[Ga90] Gakhov, F.D.: *Boundary Value Problems,* Dover, New York (1990).

[Ko99] Kondrat'ev, I.G., Kudrin, A.V., Zaboronkova, T.M.: *Electrodynamics of Density Ducts in Magnetized Plasmas,* Gordon & Breach, Amsterdam (1999).

[Me72] Meixner, J.: The behavior of electromagnetic fields at edges. *IEEE Trans. Antennas Propagat.,* **AP-20**, 442–446 (1972).

[Oh86] Ohnuki, S., Sawaya, K., Adachi, S.: Impedance of a large circular loop antenna in a magnetoplasma. *IEEE Trans. Antennas Propagat.*, **AP-34**, 1024–1029 (1986).

[Vo74] Vorovich, I.I., Aleksandrov, V.M., Babeshko, V.A.: *Nonclassical Mixed Problems in the Theory of Elasticity,* Nauka, Moscow (1974) (Russian).

24

A Two-Grid Method for a Second Kind Integral Equation with Green's Kernel

R.P. Kulkarni

Indian Institute of Technology, Powai, Mumbai, India; rpk@math.iitb.ac.in

24.1 Introduction

In this chapter we consider two-grid methods for the second kind integral equation with Green's kernel. Two-grid methods based on the Nyström operator have been studied in [At97], whereas methods based on a new approximating operator have been defined in [Ku04]. In the case of an integral operator with Green's kernel, the usual definition of the Nyström operator by replacing the integration with a numerical quadrature results in a loss of accuracy. Whereas in [AtSh07] a modified Nyström operator is defined explicitly, a similar treatment is considered in [Ku05]. The purpose of this chapter is to define two-grid methods based on the modified Nyström operator. We compare the performance of these methods by a concrete example.

24.2 Two-Grid Methods

Consider the following integral equation:

$$\lambda\, u(s) - \int_a^b k(s,t)u(t)\, dt = f(s), \quad s \in [a,b],$$

that is,

$$(\lambda I - \mathcal{K})u = f. \tag{24.1}$$

It is assumed that the kernel $k(s,t)$ is continuous in s and t and that $(\lambda I - \mathcal{K})$ is invertible.

Let $a = s_1 < s_2 < \cdots < s_{n+1} = b$ be a uniform partition of $[a,b]$ and let $h = s_{i+1} - s_i = \dfrac{b-a}{n}$ be the norm of the partition. Choose a basic quadrature formula

$$\int_{-1}^{1} g(t)dt \approx \sum_{j=1}^{r} w_j\, g(\xi_j).$$

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_24,

By considering the affine 1-1 map from $[-1,1]$ onto $[s_i, s_{i+1}]$, a corresponding quadrature formula on $[s_i, s_{i+1}]$ is given by

$$\int_{s_i}^{s_{i+1}} g(t)dt \approx \frac{h}{2}\sum_{j=1}^{r} w_j g(\eta_{i,j}), \text{ where } \eta_{i,j} = \frac{s_i + s_{i+1}}{2} + \frac{h}{2}\,\xi_j.$$

The composite quadrature formula is then given by

$$\int_a^b g(t)dt = \sum_{i=1}^{n}\int_{s_i}^{s_{i+1}} g(t)dt \approx \frac{h}{2}\sum_{i=1}^{n}\sum_{j=1}^{r} w_j g(\eta_{i,j}) = \sum_{j=1}^{nr} w_{n.j}\; g(\tau_{n,j}),$$

where

$$w_{n,(i-1)r+k} = \frac{h}{2}w_k, \quad \tau_{n,(i-1)r+k} = \eta_{i,k}, \quad i = 1,\ldots,n, \;\; k = 1,\ldots,r.$$

Using the above quadrature formula, the Nyström operator is defined as follows.

$$\mathcal{K}_n x(s) = \sum_{j=1}^{nr} w_{n,j} k(s, \tau_{n,j}) x(\tau_{n,j}).$$

It is well known that $\mathcal{K}_n$ converges to $\mathcal{K}$ in a collectively compact fashion. Hence, for all n large enough,

$$(\lambda I - \mathcal{K}_n)u_n = f \tag{24.2}$$

has a unique solution, which provides an approximation to the solution of (24.1). The solution of (24.2) is obtained by solving a system of equations of size nr. In order to achieve the required accuracy, one may have to choose n very large. Two-grid methods, which involve solution of a system of relatively small size, provide iterative approximations to u_n. A two-grid method based on the Nyström operator corresponding to a coarse grid is discussed in Atkinson [At97]. A method based on a New approximating operator is defined and analysed in Kulkarni [Ku04]. For the sake of completeness, we describe these two methods below.

24.2.1 Two-Grid Method: Nyström Operator

Let

$$a = t_1 < t_2 < \cdots < t_{m+1} = b, \quad m < n,$$

be a uniform partition of $[a,b]$ corresponding to a coarse grid and let $\mathcal{K}_m$ be the Nyström operator corresponding to the coarse grid.
Initial Guess: $u_n^{(0)}$

$$r^{(k)} = f - (\lambda I - \mathcal{K}_n)u_n^{(k)}$$
$$u_n^{(k+1)} = u_n^{(k)} + (\lambda I - \mathcal{K}_m)^{-1} r^{(k)}, \quad k = 0,1,2,\ldots.$$

Then

$$u_n^{(k+1)} = u_n^{(k)} + (\lambda I - \mathcal{K}_m)^{-1}(\lambda I - \mathcal{K}_n)(u_n - u_n^{(k)}).$$

Thus,

$$u_n - u_n^{(k+1)} \quad = \quad (\lambda I - \mathcal{K}_m)^{-1}(\mathcal{K}_n - \mathcal{K}_m)(u_n - u_n^{(k)})$$

and

$$u_n - u_n^{(k+1)} \quad = \quad [(\lambda I - \mathcal{K}_m)^{-1}(\mathcal{K}_n - \mathcal{K}_m)]^2(u_n - u_n^{(k-1)}).$$

Since $\mathcal{K}_m \to \mathcal{K}$ in a collectively compact fashion, for m large enough,

$$\beta_m = \sup_{n \geq m} \|[(\lambda I - \mathcal{K}_m)^{-1}(\mathcal{K}_n - \mathcal{K}_m)]^2\| < 1.$$

Hence, as $k \to \infty$,

$$\|u_n - u_n^{(k+1)}\|_\infty \leq \beta_m \|u_n - u_n^{(k-1)}\|_\infty \leq (\beta_m)^{\frac{k+1}{2}} \|u_n - u_n^{(0)}\|_\infty \to 0. \quad (24.3)$$

24.2.2 Two-Grid Method: New Operator

Let

$$a = t_1 < t_2 < \cdots < t_{m+1} = b, \quad m < n,$$

be a uniform partition of $[a, b]$ corresponding to a coarse grid. Let X_m be the space of piecewise polynomials of degree $\leq r - 1$ with respect to the above partition. Choose r distinct points $\nu_{i,j},\ j = 1, \ldots, r$ in each of the subintervals $[t_i, t_{i+1}]$. Let $P_m : C[a, b] \to X_m$ be defined by

$$(P_m x)(\nu_{i.j}) = x(\nu_{i,j}), \quad i = 1, \ldots, m, \ \ j = 1, \ldots, r.$$

If the end points of $[t_i, t_{i+1}]$ are included in the set of interpolation points, then $P_m x \in C[a, b]$ and $P_m^2 = P_m$. Otherwise, P_m can be extended to $L^\infty[a, b]$ and $P_m^2 = P_m$ (see [AtGa83]). In both the cases, $P_m \to I$ pointwise as $m \to \infty$.

Define

$$T_m \quad = \quad P_m \mathcal{K}_n + \mathcal{K}_n P_m - P_m \mathcal{K}_n P_m.$$

The method starts with an initial guess $v_n^{(0)}$ and proceeds according to the scheme

$$r^{(k)} = f - (\lambda I - \mathcal{K}_n) v_n^{(k)},$$
$$v_n^{(k+1)} = v_n^{(k)} + (\lambda I - T_m)^{-1} r^{(k)}, \quad k = 0, 1, 2, \ldots.$$

Then

$$v_n^{(k+1)} = v_n^{(k)} + (\lambda I - T_m)^{-1}(\lambda I - \mathcal{K}_n)(u_n - v_n^{(k)}).$$

Thus,

$$\begin{aligned} u_n - v_n^{(k+1)} &= (\lambda I - T_m)^{-1}(I - P_m)\mathcal{K}_n(I - P_m)(u_n - v_n^{(k)}) \\ &= M_{n,m}(u_n - v_n^{(k)}). \end{aligned}$$

Since $P_m \to I$ pointwise and $\mathcal{K}_n \to \mathcal{K}$ in a collectively compact fashion, it follows that, as $m \to \infty$,

$$\sup_{n \geq m} \|M_{n,m}\| \leq \|(\lambda I - T_m)^{-1}\|(I \ - P_m)\| \sup_{n \geq m} \|(I - P_m)\mathcal{K}_n\| \to 0.$$

Hence, for m large enough, $\delta_m = \sup_{n \geq m} \|M_{n,m}\| < 1$. Consequently, as $k \to \infty$,

$$\|u_n - v_n^{(k+1)}\|_\infty \leq \delta_m \|u_n - v_n^{(k)}\|_\infty \leq (\delta_m)^{k+1} \|u_n - v_n^{(0)}\|_\infty \to 0. \qquad (24.4)$$

A comparison of (24.3) and (24.4) suggests that the number of iterates needed to achieve a certain accuracy in the new method should be about half as compared to the Nyström method. It is validated by the numerical results in the last section.

24.3 Modified Nyström Method: Green's Kernel

In this section we consider two-grid methods for an integral operator with Green's kernel. Since the kernel lacks differentiability properties along the diagonal, it is necessary to modify the definition of the Nyström operator. We illustrate the basic principle by defining a modified Nyström operator based on the Simpson integration (see [AtSh07] and [Ku05]).
Let $0 = s_1 < s_2 < \cdots < s_{n+1} = 1$ be a uniform partition of $[0, 1]$ and let $h = s_{i+1} - s_i = \dfrac{1}{n}$ be the norm of the partition. The basic Simpson integration is defined by

$$\int_{s_i}^{s_{i+1}} g(t)dt \approx \frac{h}{6}\left(g(s_i) + 4g(\frac{s_i + s_{i+1}}{2}) + g(s_{i+1})\right) = S(g, s_i, s_{i+1})$$

and the corresponding composite quadrature formula is given by

$$\begin{aligned} \int_0^1 g(t)dt &= \sum_{i=1}^{n} \int_{s_i}^{s_{i+1}} g(t)dt \approx \frac{h}{6} \sum_{i=1}^{n} \left(g(s_i) + 4g(\frac{s_i + s_{i+1}}{2}) + g(s_{i+1})\right) \\ &= \sum_{j=1}^{2n+1} w_{n.j}\, g(\tau_{n,j}), \end{aligned}$$

where $w_{n,1} = w_{n,2n+1} = \dfrac{1}{6n}$, $w_{n,j} = \dfrac{2}{3n}$, j even, $w_{n,j} = \dfrac{1}{3n}$, j odd, and

$$\tau_{n,j} = \frac{(j-1)}{2n}, \quad j = 1, \ldots, 2n+1.$$

If g is four times differentiable, then

$$\left| \int_0^1 g(t)dt - \sum_{j=1}^{2n+1} w_{n.j}\ g(\tau_{n,j}) \right| \leq C\|g^{(4)}\|_\infty h^4.$$

Let $\mathcal{K}$ be an integral operator with a kernel that is four times differentiable in the second variable, and let

$$\mathcal{K}_n x(s) = \sum_{j=1}^{nr} w_{n,j} k(s, \tau_{n,j}) x(\tau_{n,j})$$

be associated with the composite Simpson integration. Then, for $x \in C^4[0,1]$,

$$\|(\mathcal{K} - \mathcal{K}_n)x\|_\infty \leq C\|x^{(4)}\|_\infty h^4.$$

As a consequence, if the right-hand side f in (24.1) is sufficiently differentiable, then

$$\|u - u_n\|_\infty \leq Ch^4.$$

Consider the integral operator $\mathcal{K}$ given by

$$(\mathcal{K}x)(s) = \int_0^1 k(s,t)\ x(t)\ dt, \quad s \in [0,1],$$

where

$$k(s,t) = \begin{cases} s(1-t) & \text{if} \quad 0 \leq s \leq t \leq 1, \\ t(1-s) & \text{if} \quad 0 \leq t \leq s \leq 1. \end{cases}$$

The Green's kernel described above is continuous on $[0,1] \times [0,1]$, but not differentiable along the diagonal. In this case, for $x \in C^2[0,1]$,

$$\|(\mathcal{K} - \mathcal{K}_n)x\|_\infty \leq C\|x^{(2)}\|_\infty h^2$$

and

$$\|u - u_n\|_\infty \leq Ch^2.$$

The order of convergence h^4 can be restored by modifying the definition of the Nyström operator in the following fashion.

For a fixed $s \in [0,1]$, let $k_s(t) = k(s,t)$, $t \in [0,1]$, and let $s \in [s_q, s_{q+1}]$. We write

$$\int_a^b k(s,t)x(t)dt = \sum_{\substack{i=1 \\ i \neq q}}^{n} \int_{s_i}^{s_{i+1}} (k_s x)(t)dt + \int_{s_q}^{s} (k_s x)(t)dt + \int_s^{s_{q+1}} (k_s x)(t)dt.$$

Let $\tilde{x}$ be the quadratic polynomial interpolating x at s_q, s_{q+1} and $\frac{s_q + s_{q+1}}{2}$. For $s \in [s_q, s_{q+1}]$, define

$$\tilde{\mathcal{K}}_n x(s) = \sum_{\substack{i=1 \\ i \neq q}}^{n} S(k_s x, s_i, s_{i+1}) + S(k_s \tilde{x}, t_q, s) + S(k_s \tilde{x}, s, t_{q+1}).$$

Thus, $\tilde{\mathcal{K}}_n x(s) = \sum_{j=1}^{2n+1} \tilde{w}_{n,j}(s) x(\tau_{n,j})$, so, for $x \in C^4[a,b]$,

$$\|(\mathcal{K} - \tilde{\mathcal{K}}_n)x\|_\infty \leq C \|x^{(4)}\|_\infty h^4.$$

If the integral equation (24.1) is approximated by $(\lambda I - \tilde{\mathcal{K}}_n)\tilde{u}_n = f$, then

$$\|u - \tilde{u}_n\|_\infty \leq C h^4.$$

Let X_m be the space of piecewise constant polynomials with respect to the coarse partition

$$0 = t_1 < t_2 < \cdots < t_{m+1} = 1, \quad m < n.$$

Let the interpolation points be the midpoints of $[t_i, t_{i+1}]$, and let $P_m : C[a,b] \to X_m$ be the map defined by

$$(P_m x)\left(\frac{t_i + t_{i+1}}{2}\right) = x\left(\frac{t_i + t_{i+1}}{2}\right), \quad i = 1, \ldots, m.$$

Define

$$\tilde{T}_m \quad = \quad P_m \tilde{\mathcal{K}}_n + \tilde{\mathcal{K}}_n P_m - P_m \tilde{\mathcal{K}}_n P_m.$$

The two-grid method starts with an initial guess $\tilde{v}_n^{(0)}$ and then proceeds according to the scheme

$$\tilde{r}^{(k)} = f - (\lambda I - \tilde{\mathcal{K}}_n)\tilde{v}_n^{(k)},$$
$$\tilde{v}_n^{(k+1)} = \tilde{v}_n^{(k)} + (\lambda I - \tilde{T}_m)^{-1}\tilde{r}^{(k)}, \quad k = 0, 1, 2, \ldots.$$

Then, in exactly the same manner as before, it can be shown that

$$\|u_n - \tilde{v}_n^k\|_\infty \to 0 \quad \text{as} \quad k \to \infty.$$

A two-grid method corresponding to the modified Nyström operator is defined in a similar manner.

24.4 Computational Cost

In each iteration we need to compute $y^{(k)} = (\lambda I - \tilde{T}_m)^{-1}\tilde{r}^{(k)}$, that is, we need to solve the following equation:

$$(\lambda I - \tilde{T}_m) y^{(k)} = \tilde{r}^{(k)}.$$

Since

$$\tilde{T}_m = P_m\tilde{\mathcal{K}}_n + \tilde{\mathcal{K}}_n P_m - P_m\tilde{\mathcal{K}}_n P_m,$$

we get

$$\lambda\, P_m\tilde{y}^{(k)} - P_m\tilde{\mathcal{K}}_n y^{(k)} = P_m\tilde{r}^{(k)}$$
$$\lambda\,(I - P_m)y^{(k)} - (I - P_m)\tilde{\mathcal{K}}_n P_m y^{(k)} = (I - P_m)\tilde{r}^{(k)}.$$

Thus

$$\lambda\, P_m y^{(k)} - P_m\tilde{\mathcal{K}}_n P_m y^{(k)} - \frac{1}{\lambda} P_m\tilde{\mathcal{K}}_n(I - P_m)\tilde{\mathcal{K}}_n P_m y^{(k)}$$
$$= P_m r^{(k)} + \frac{1}{\lambda}\, P_m\tilde{\mathcal{K}}_n(I - P_m)\tilde{r}^{(k)}.$$

This is a system of size m.

In the two-grid method corresponding to the modified Nyström operator, for each iteration we need to solve the following system of size $m + 1$:

$$(\lambda I - \tilde{\mathcal{K}}_m)y^{(k)} = \tilde{r}^{(k)}.$$

Thus, the costs in both iteration schemes are comparable. However, the numerical example in the next section shows that the number of iterates in the new method is about half of the number of iterates in the Nyström method.

24.5 Numerical Results

Consider the integral equation with Green's kernel

$$u(s) - \int_0^1 k(s,t)u(t)dt = \left(1 - \frac{1}{\pi^2}\right)\sin(\pi s), \quad s \in [0,1].$$

The exact solution is $u(s) = \sin(\pi s)$.

We choose $m = 4$, $n = 512$, and X_m is the space of piecewise constant functions with respect to the coarse partition. The interpolation points are chosen to be the midpoints of the subintervals, and the numerical quadrature is chosen to be the composite Simpson integration. The $\|.\|$ is calculated as the maximum value at 512 fine grid points. Tables 24.1 and 24.2 show a comparison between the new method and the Nyström and modified Nyström methods.

Remark 1. It is to be noted that in the case of a two-grid method defined using the new method with the modified Nyström operator, the error for the 7th iterate is of the order 10^{-14}, while the corresponding error in the two-grid method with the modified Nyström operator is of the order 10^{-8}. In the latter case, it was necessary to evaluate 12 iterates to achieve an accuracy of 10^{-14}.

Table 24.1. The new method versus the modified Nyström method.

k	New $\|u_n - \tilde{v}_n^k\|$	Modified Nyström $\|u_n - \tilde{u}_n^k\|$	New $\|u - \tilde{v}_n^k\|$	Modified Nyström $\|u - \tilde{u}_n^k\|$
1	1.4×10^{-2}	6.9×10^{-1}	1.4×10^{-2}	6.9×10^{-1}
2	9.3×10^{-5}	5.7×10^{-2}	9.3×10^{-5}	5.7×10^{-2}
3	5.9×10^{-7}	2.7×10^{-3}	5.9×10^{-7}	2.7×10^{-3}
4	3.7×10^{-9}	2.3×10^{-4}	3.7×10^{-9}	2.3×10^{-4}
5	2.3×10^{-11}	1.1×10^{-5}	4.3×10^{-11}	1.1×10^{-5}
6	1.5×10^{-13}	9.4×10^{-7}	4.3×10^{-11}	9.4×10^{-7}
7	2.0×10^{-14}	4.3×10^{-8}	4.3×10^{-11}	4.3×10^{-8}

Table 24.2. The new method versus the Nyström method.

k	New $\|u_n - v_n^k\|$	Nyström $\|u_n - u_n^k\|$	New $\|u - v_n^k\|$	Nyström $\|u - u_n^k\|$
1	1.4×10^{-2}	7.8×10^{-2}	1.3×10^{-2}	7.8×10^{-2}
2	9.3×10^{-5}	1.4×10^{-2}	9.3×10^{-5}	1.4×10^{-2}
3	5.9×10^{-7}	1.8×10^{-4}	7.7×10^{-6}	1.8×10^{-4}
4	3.7×10^{-9}	1.4×10^{-5}	7.7×10^{-6}	9.5×10^{-6}
5	2.4×10^{-11}	3.0×10^{-7}	7.7×10^{-6}	8.0×10^{-6}
6	1.5×10^{-13}	1.5×10^{-8}	7.7×10^{-6}	7.7×10^{-6}
7	2.0×10^{-14}	4.3×10^{-10}	7.7×10^{-6}	7.7×10^{-6}

Remark 2. If we compare the iterates with the exact solution, the error in the two-grid methods defined with the usual Nyström operator remains 10^{-6}, whereas the modification reduces this error to 10^{-11}.

Acknowledgement. This work was partially supported by the DST (SERC) grant SR/S4/MS:239/04.

References

[At97] Atkinson, K.E.: *The Numerical Solution of Integral Equations of the Second Kind,* Cambridge University Press, London (1997).

[AtGa83] Atkinson, K.E., Graham, I., Sloan, I.: Piecewise continuous collocation for integral equations. *SIAM J. Numer. Anal.,* **20**, 172–186 (1983).

[AtSh07] Atkinson, K.E., Shampine, L.F.: Solving Fredholm integral equations of the second kind in MATLAB. Research Report (2007).

[Ku04] Kulkarni, R.P.: Approximate solution of multivariable integral equations of the second kind. *J. Integral Equations,* **16**, 343–373 (2004).

[Ku05] Kulkarni, R.P.: On improvement of the iterated Galerkin solution of the second kind integral equations. *J. Numer. Math.,* **13**, 205–218 (2005).

25

A Brief Overview of Plate Finite Element Methods

C. Lovadina

Università di Pavia, Italy; carlo.lovadina@unipv.it

25.1 Introduction

In this chapter we present a brief account of possible finite element methods (FEMs) for the plate bending problem, when described by means of the *Reissner–Mindlin* model. We point out that the following overview is *far from being exhaustive:* we are perfectly aware that many important approaches are not even mentioned. Accordingly, also the references are very limited and lack completeness.

The choice of schemes that are going to be described is strongly biased by the author's experience, and it does not correspond to any efficiency or robustness criterion. We also remark that we are not going to detail any *rigorous* convergence and stability proof. Rather, we will try to heuristically explain

1. the main troubles arising from the FEM discretization of plate problems (Section 25.2);
2. why the methods under consideration succeed in the solution approximation (Section 25.3).

25.2 The Reissner–Mindlin Plate Model and Its FEM Discretization

25.2.1 The Reissner–Mindlin Plate Model

The Reissner–Mindlin equations for a clamped plate with a convex midplane domain Ω require us to find (θ, w) such that (see, for example, [Ba95] or [Hu87])

$$\begin{cases} -\operatorname{div}\mathbf{C}\varepsilon(\theta) - \lambda t^{-2}(\nabla w - \theta) = 0 & \text{in } \Omega, \\ -\operatorname{div}\left(\lambda t^{-2}(\nabla w - \theta)\right) = g & \text{in } \Omega, \\ \theta = 0,\ w = 0 & \text{on } \partial\Omega. \end{cases} \tag{25.1}$$

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_25,

In (25.1), t is the plate thickness, λ is the shear modulus, and $\mathbf{C}$ is the tensor of bending moduli, given (for isotropic materials) by

$$\mathbf{C}\tau := \frac{E}{12(1-\nu^2)}[(1-\nu)\tau + \nu\,\mathrm{tr}(\tau)\mathbf{I}], \tag{25.2}$$

where τ is a generic second-order symmetric tensor, $\mathrm{tr}(\tau)$ its trace, $\mathbf{I}$ is the second-order identity tensor, and E and ν are the Young modulus and Poisson ratio, respectively. Moreover, $\theta = (\theta_1, \theta_2)$ represents the (vector) rotation field and w is the transversal displacement (see Figure 25.1), while g is a given transversal load. Finally, ∇ is the usual symmetric gradient operator.

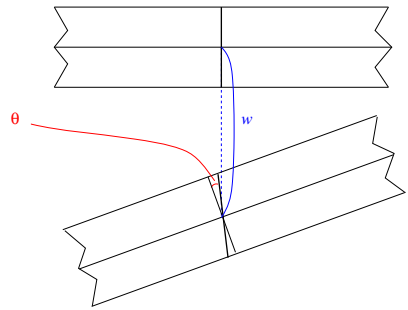


Fig. 25.1. Reissner–Mindlin kinematic variables.

Introducing the space $\Theta \times W = (H_0^1(\Omega))^2 \times H_0^1(\Omega)$, it is easily seen that problem (25.1) can be reformulated as the following *minimization* problem over $\Theta \times W$ for the elastic energy E_t:

$$\begin{cases} \text{Find } (\theta, w) \in \Theta \times W \text{ which minimizes} \\ E_t(\eta, v) := \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta) : \varepsilon(\eta) + \dfrac{\lambda t^{-2}}{2}\displaystyle\int_\Omega |\nabla v - \eta|^2 - \displaystyle\int_\Omega gv. \end{cases} \tag{25.3}$$

From a mechanical viewpoint, the term $\frac{1}{2}\int_\Omega \mathbf{C}\varepsilon(\eta) : \varepsilon(\eta)$ represents the bending energy, the term $\frac{\lambda t^{-2}}{2}\int_\Omega |\nabla v - \eta|^2$ gives the shear energy, and $\int_\Omega gv$ is the external load work. A standard computation leads to the Euler–Lagrange equations associated with problem (25.3):

$$\begin{cases} \text{Find } (\theta, w) \in \Theta \times W \text{ such that} \\ \displaystyle\int_\Omega \mathbf{C}\varepsilon(\theta) : \varepsilon(\eta) + \lambda t^{-2}\int_\Omega (\nabla w - \theta)\cdot(\nabla v - \eta) = \int_\Omega gv \end{cases} \tag{25.4}$$

for every $(\eta, v) \in \Theta \times W$. We remark that, for every fixed $t > 0$, the bilinear form

$$\int_\Omega \mathbf{C}\varepsilon(\theta) : \varepsilon(\eta) + \lambda t^{-2} \int_\Omega (\nabla w - \theta) \cdot (\nabla v - \eta)$$

is *continuous, symmetric, and coercive* over $\Theta \times W$. Moreover, for g smooth, $v \to \int_\Omega gv$ is a *linear and continuous functional* over W. Therefore, problem (25.4) is elliptic and the Lax–Milgram lemma implies the existence, uniqueness, and stability of its solution.

The ellipticity of the problem suggests to consider Galerkin discretization techniques for the solution approximation. Among them, the FEM is a very popular and flexible choice (see, for example, [Ci78]). We briefly recall that a *conforming* finite element procedure for our plate problem is based on the following steps.

- *Mesh generation.* Construct a decomposition $\mathcal{T}_h$ of Ω into triangular elements T. The mesh size h, defined as the maximum diameter of all the triangles in the decomposition, is an important geometric parameter. The mesh is typically required to fulfill some compatibility conditions. A typical mesh is displayed in Figure 25.2. We also remark that quadrilateral elements may be used as well.

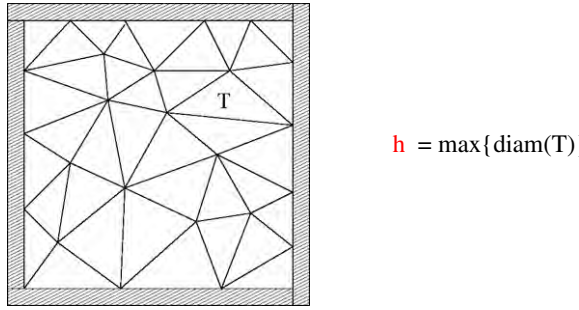


Fig. 25.2. A typical triangular mesh.

- *Local approximation.* For each T in the mesh $\mathcal{T}_h$, introduce $P(T)$, a polynomial space on T. Different choices for different elements may be made. However, the most common choice consists of selecting the same shape functions for every element.
- *Finite element space.* Form the *discrete* space

$$\Theta_h \times W_h = \left\{(\eta_h, v_h) \in \Theta \times W : (\eta_h, v_h)_{|T} \in P(T)\right\}.$$

  For instance, one could select piecewise linear and globally continuous functions for both rotations and vertical displacements. This choice is schematically depicted in Figure 25.3. Here the bullets mean that the relevant unknown is uniquely determined by assigning the values at the triangle vertices.

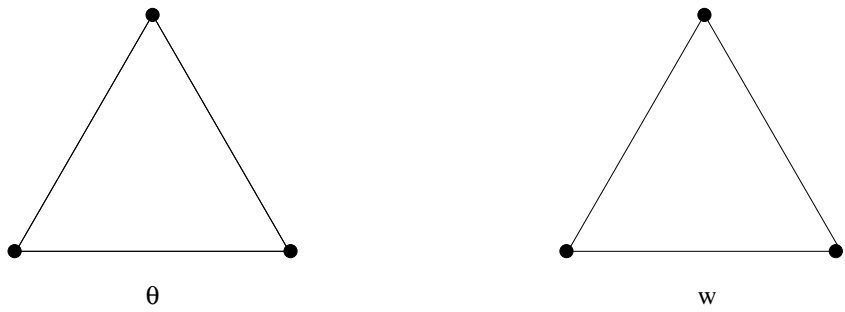


Fig. 25.3. The simplest FEM space.

- *Discrete problem.* Solve the problem

$$\begin{cases} \text{Find } (\theta_h, w_h) \in \Theta_h \times W_h \text{ s.t.} \\ \displaystyle\int_\Omega \mathbf{C}\varepsilon(\theta_h) : \varepsilon(\eta_h) + \lambda t^{-2} \int_\Omega (\nabla w_h - \theta_h) \cdot (\nabla v_h - \eta_h) = \int_\Omega g v_h \end{cases} \tag{25.5}$$

for every $(\eta_h, v_h) \in \Theta_h \times W_h$.

25.2.2 Locking Effects and Spurious Mode Occurrence

Since the problem is elliptic for each $t > 0$, the standard theory gives optimal error estimates for the discrete solution $(\theta_h, w_h) \in \Theta_h \times W_h$, as the mesh size tends to zero (see [Ci78] or [BrFo91], for instance). In practice, this means that reasonable outcomes are expected when using a mesh as in Figure 25.2, and the approximation spaces as in Figure 25.3. However, for a "small" thickness the discrete solution heavily underestimates the analytical solution (see Figure 25.4 for a pictorial representation of this occurrence, when the plate is clamped and subjected to a constant load).

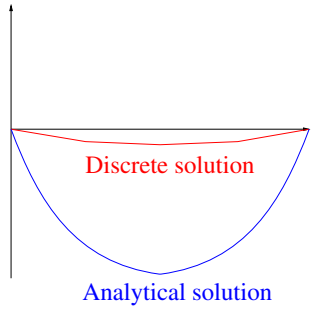


Fig. 25.4. Typical locking solution.

To understand this bad phenomenon, known as the *shear locking effect* (see, e.g., [BrFo91]), it is worth considering the asymptotic behavior of the problem as $t \to 0$. More precisely, it can be proved that problem (25.3) converges, in a suitable sense, to the *limit constrained problem* (see, for example, [SaPa92] or [ChPa94])

$$\begin{cases} \text{Find } (\theta^0, w^0) \in K \text{ which minimizes} \\ E_0(\eta, v) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta) : \varepsilon(\eta) - \int_\Omega gv, \quad (\eta, v) \in K, \end{cases} \tag{25.6}$$

where K is defined by

$$K = \{(\eta, v) \in \Theta \times W : \nabla v = \eta\}. \tag{25.7}$$

Even though we will not detail the convergence proof, we point out that the constraint $(\theta^0, w^0) \in K$ is very reasonable. Indeed, in minimizing the functional $E_t(\cdot,\cdot)$ in (25.3) *for very small* t, one should choose functions (η, v) such that

$$\int_\Omega |\nabla v - \eta|^2 \text{ is } \textit{“very small”} \quad (\text{which means } \nabla v - \eta \text{ is } \textit{“very small”}\ );$$

otherwise, one pays an enormous amount of shear energy in the term

$$\frac{\lambda t^{-2}}{2}\int_\Omega |\nabla v - \eta|^2.$$

We also remark that problem (25.6) is coercive and continuous on K. Furthermore, K is a *non-trivial* closed subspace of $\Theta \times W$. Indeed, given *any* compactly supported smooth function v, one may set $\eta := \nabla v$. By construction, $(\eta, v) \in K$. Therefore, the *continuous limit* problem (25.6) may be thought of as a standard well-posed (elliptic) problem.

We now turn our attention to the discrete problem. The discrete problem (25.5) is equivalent to a minimization problem for the *same* functional $E_t(\cdot,\cdot)$ (see (25.3)), but restricted to $\Theta_h \times W_h$, i.e.,

$$\begin{cases} \text{Find } (\theta_h, w_h) \in \Theta_h \times W_h \text{ which minimizes} \\ E_t(\eta_h, v_h) := \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) + \frac{\lambda t^{-2}}{2}\int_\Omega |\nabla v_h - \eta_h|^2 - \int_\Omega g v_h. \end{cases} \tag{25.8}$$

Therefore, the FEM problem converges, as $t \to 0$, to the discrete limit problem

$$\begin{cases} \text{Find } (\theta_h^0, w_h^0) \in K_h \text{ which minimizes} \\ E_0(\eta_h, v_h) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) - \int_\Omega g v_h, \quad (\eta_h, v_h) \in K_h, \end{cases} \tag{25.9}$$

where

$$K_h = \{(\eta_h, v_h) \in \Theta_h \times W_h \,:\, \nabla v_h = \eta_h\} = K \cap (\Theta_h \times W_h). \tag{25.10}$$

We remark that the discrete limit problem accounts for minimizing the same limit functional as for the continuous problem (see (25.6)), but this time on the discrete subspace K_h.

Let us analyze the structure of K_h, when using piecewise linear and globally continuous functions (see Figure 25.3). If $(\eta_h, v_h) \in K_h$, then $\nabla v_h = \eta_h$ (see (25.10)). Since $\eta_h \in C^0(\Omega)$, we deduce that $v_h \in C^1(\Omega)$. But v_h is also piecewise linear, so $v_h \in C^1(\Omega)$ implies that v_h is *a globally linear function* in Ω. If the plate is clamped on a part of the boundary of positive length, we then infer that $v_h = 0$. Recalling that $\eta_h = \nabla v_h$, we finally deduce that

$$(\eta_h, v_h) \in \Theta_h \times W_h \quad \Longrightarrow \quad (\eta_h, v_h) = (0, 0),$$

i.e., $K_h = \{(0,0)\}$. Therefore, the limit problem (25.9) is just a minimization problem for a "good" functional, but on a *trivial* space: the minimizing pair is surely $(\theta_h, w_h) = (0,0)$!

Of course, this is the limit "zero thickness" situation; however, for a "small thickness" (with respect to the mesh size h), the discrete problem is essentially so close to the limit case that the discrete solution is very small: *shear locking* has occurred.

Since the trouble stands in the shear energy term, a possible cure consists in reducing, somehow, its influence at the discrete level. This can be accomplished by considering the *modified* energy

$$E_{h,t}(\eta_h, v_h) = \frac{1}{2}\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) + \frac{\lambda\, t^{-2}}{2}\int_\Omega |R_h(\nabla v_h - \eta_h)|^2 - \int_\Omega g v_h,$$

where R_h is a suitable *reduction* operator. Therefore, the finite element scheme now reads, in its equivalent minimization formulation,

$$\begin{cases} \text{Find } (\theta_h, w_h) \in \Theta_h \times W_h \text{ which minimizes} \\ E_{h,t}(\eta_h, v_h) := \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) + \dfrac{\lambda t^{-2}}{2}\int_\Omega |R_h(\nabla v_h - \eta_h)|^2 - \int_\Omega g v_h. \end{cases} \tag{25.11}$$

As $t \to 0$, the problem will consequently converge to the problem

$$\begin{cases} \text{Find } (\theta_h^0, w_h^0) \in K_h \text{ which minimizes} \\ E_{h,0}(\eta_h, v_h) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) - \int_\Omega g v_h, \quad (\eta_h, v_h) \in K_h, \end{cases} \tag{25.12}$$

where K_h is now defined by

$$K_h = \{(\eta_h, v_h) \in \Theta_h \times W_h \,:\, R_h(\nabla v_h - \eta_h) = 0\}\,. \tag{25.13}$$

We point out that now the constraint has been relaxed to $R_h(\nabla v_h - \eta_h) = 0$, and we don't have $\nabla v_h - \eta_h = 0$ anymore. As a consequence, one may hope

that K is large enough to properly approximate the subspace K (see (25.7)), thus preventing shear locking effects. However, the reduction operator R_h must be carefully selected. Let us consider the following (not recommended) choice.

For the approximation space $\Theta_h \times W_h$, we select piecewise quadratic and globally continuous functions, as schematically depicted in Figure 25.5.

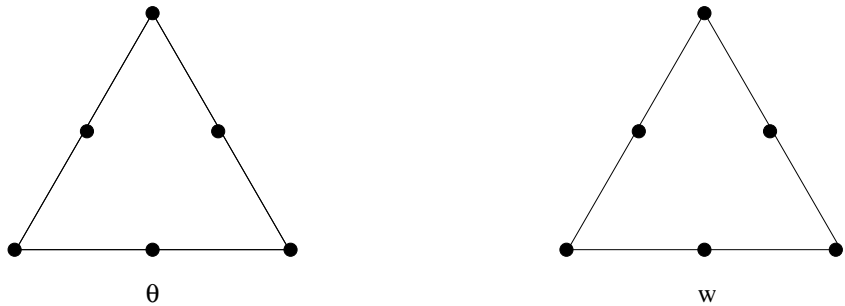


Fig. 25.5. Quadratic approximation.

We choose R_h as the L^2-projection operator onto the piecewise constant (vector-valued) functions. In can be proved that there is $v_h^* \in W_h$ for which $\nabla v_h^* \neq 0$ but $R_h(\nabla v_h^*) = 0$. Computing the energy $E_{h,t}(\cdot,\cdot)$ along the direction given by $(0, v_h^*) \in \Theta_h \times W_h$, we get (see (25.11))

$$E_h(\alpha(0, v_h^*)) = \frac{\lambda\, t^{-2}}{2} \int_\Omega |\alpha\, R_h(\nabla v_h^*)|^2 - \alpha \int_\Omega g v_h^* = -\alpha \int_\Omega g v_h^* \quad \forall \alpha \in \mathbb{R}.$$

Therefore, the functional $E_{h,t}(\cdot,\cdot)$ is *linear* along that particular direction. As a consequence, no minimizing pair $(\theta_h, w_h^*) \in \Theta_h \times W_h$ can be found, since ellipticity has been lost at the discrete level. From a practical point of view, one obtains a *singular* stiffness matrix. Of course, this is an extreme situation. However, even when the stiffness matrix is invertible, a naive choice of the reduction operator R_h may lead to a milder, though nasty, phenomenon: the occurrence of *spurious modes,* i.e., the discrete solution exhibits non-physical heavy oscillations.

Obviously, to avoid the existence of $v_h^* \in W_h$ such that $\nabla v_h^* \neq 0$ but $R_h(\nabla v_h^*) = 0$, one would like to choose $R_h = Id$, the identity operator. However, this choice will lead to trouble, again with the shear locking.

To summarize, we need to reduce the influence of the shear energy term, but

- If R_h reduces "too much," we risk *spurious modes* occurrence.
- If R_h does not reduce "enough," we risk *shear locking* effects.

Balancing R_h is not a trivial task. However, nowadays there are several efficient options in the literature. Some of them are briefly reviewed in the following section.

25.3 Some Efficient Finite Element Techniques

25.3.1 MITC Elements

We now describe one of the most popular and efficient strategies to approximate the solution of the Reissner–Mindlin plate equations: the so-called Mixed Interpolation of Tensorial Components (*MITC*) elements (see [BBF89], [BFS91], but also [TeHu85] and [Du92], for example). We here focus on a particular low-order element, but higher-order, as well as quadrilateral versions, are available. The scheme under consideration, known as the *MITC7* element, consists in making the following choice.

- To approximate each rotational component, we select piecewise quadratic and globally continuous functions. In addition, a local cubic bubble is inserted.
- To approximate the vertical displacements, we select piecewise quadratic and globally continuous functions.

To complete the element description, we need to specify the reduction operator R_h. To this end, for each triangle T we introduce the vectorial space

$$\Gamma(T) := (P_1(T))^2 + P_1(T)(y, -x)^T,$$

where $P_1(T)$ denotes the space of linear functions on T. It can be proved that a function in $\Gamma(T)$ is uniquely determined by assigning

- the moments up to the first order of its tangential component, for each edge of T (6 degrees of freedom);
- its mean value over T (2 degrees of freedom).

For a given smooth vectorial function $\delta = (\delta_1, \delta_2)$, we then define $R_h\delta$ by requiring that

$$\begin{cases} (R_h\delta)_{|T} \in \Gamma(T), \\ \displaystyle\int_T R_h\delta = \int_T \delta, \\ \displaystyle\int_e [(R_h\delta)\cdot \mathrm{t}]p_1(s)\mathrm{d}s = \int_e [\delta\cdot \mathrm{t}]p_1(s)\mathrm{d}s, \end{cases} \tag{25.14}$$

for every triangle $T \in \mathcal{T}_h$, and every edge e of T. Above, t is the tangent vector to the side e, while $p_1(s)$ is a linear polynomial with respect to a local

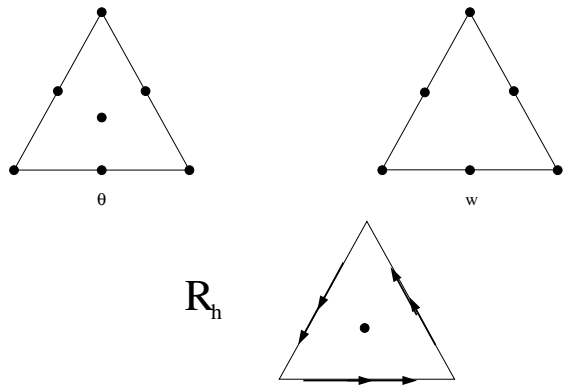


Fig. 25.6. The *MITC7* element.

coordinate s along e. All these choices and definitions are schematically shown in Figure 25.6.

The *MITC7* element, as well as all the other schemes based on the *MITC* philosophy, is carefully designed to fulfill the following crucial features.

- **P1.** R_h is the identity operator when applied to the gradients of *discrete* vertical displacements, i.e.,

$$R_h \nabla v_h = \nabla v_h \quad \forall v_h \in W_h.$$

- **P2.** If $\operatorname{curl} R_h \eta = 0$, then $R_h \eta$ is the gradient of a *discrete* vertical displacement, i.e.,

$$\{R_h \eta \,:\, \eta \in (H_0^1(\Omega))^2, \operatorname{curl} R_h \eta = 0\} = \nabla W_h.$$

  (Above and in what follows, the curl operator is defined as

$$\operatorname{curl} \varphi = \partial \varphi_2 / \partial x - \partial \varphi_1 / \partial y,$$

  for a generic vector-valued function $\varphi = (\varphi_1, \varphi_2)$.)

There exists an auxiliary space Q_h such that (for the *MITC7* element, this space consists of the locally linear functions, with no continuity requirements across the element interfaces):

- **P3.** The "commutative diagram property"

$$\operatorname{curl} R_h \eta = P_h \operatorname{curl} \eta, \quad \eta \in (H_0^1(\Omega))^2,$$

  holds, where $P_h : L^2 \to Q_h$ is the L^2 projection operator.
- **P4.** The pair of spaces (Θ_h, Q_h) is stable for Stokes-like problems.

We now comment on the properties above. We first notice that the very important property **P1** prevents the scheme from suffering from spurious modes.

To see which is the role played by the other properties, we focus on the *limit* problem, as $t \to 0$. As detailed in Section 25.2.2, this accounts for considering the problem (see (25.6)–(25.7)):

$$\begin{cases} \text{Find } (\theta^0, w^0) \in K \text{ which minimizes} \\ E_0(\eta, v) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta) : \varepsilon(\eta) - \int_\Omega gv, \quad (\eta, v) \in K, \end{cases} \tag{25.15}$$

where

$$K = \{(\eta, v) \in \Theta \times W \,:\, \nabla v = \eta\}\,. \tag{25.16}$$

The discrete counterpart reads

$$\begin{cases} \text{Find } (\theta_h^0, w_h^0) \in K_h \text{ which minimizes} \\ E_0(\eta_h, v_h) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) - \int_\Omega gv_h \quad (\eta_h, v_h) \in K_h, \end{cases} \tag{25.17}$$

where

$$K_h = \{(\eta_h, v_h) \in \Theta_h \times W_h : \nabla v_h = R_h\eta_h\}\,. \tag{25.18}$$

Roughly speaking, problem (25.17) has a chance to be a "good approximation" of problem (25.15), for all loads g, only if K_h is a "good approximation" of K. This means that, given $(\eta, v) \in K$, we need to find $(\eta_I, v_I) \in K_h$ such that

$$\eta_I \approx \eta, \quad v_I \approx v. \tag{25.19}$$

Above, $\eta \approx \eta_I$ means that some suitable norm of $\eta - \eta_I$ vanishes as the mesh size h tends to zero. The same remark applies to $v_I \approx v$, of course.

If $(\eta, v) \in K$ is sufficiently regular, the most natural choice would be to take η_I and v_I as the usual Lagrange interpolants of η and v, respectively (if $(\eta, v) \in K$ is less regular, one might think of the Clément's interpolants, see [Ci78]). Unfortunately, even though $\nabla v = \eta$, in general it is not true that the choice above leads to a (η_I, v_I) such that $\nabla v_I = R_h\eta_I$. Therefore, $(\eta_I, v_I) \notin K_h$, and a more sophisticated and subtle choice needs to be made, as sketched below.

Fix $(\eta, v) \in K$. We first consider the discretization of the Stokes-like problem:

$$\begin{cases} \text{Find } (\eta_I, p_h) \in \Theta_h \times Q_h \text{ such that} \\ \displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_I) : \varepsilon(\chi_h) + \int_\Omega p_h \operatorname{curl} \chi_h = \int_\Omega \mathbf{C}\varepsilon(\eta) : \varepsilon(\chi_h), \quad \chi_h \in \Theta_h, \\ \displaystyle\int_\Omega q_h \operatorname{curl} \eta_I = 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad q_h \in Q_h. \end{cases} \tag{25.20}$$

From property **P4**, we deduce that

$$\eta_I \approx \eta, \quad P_h \mathrm{curl}\, \eta_I = 0. \tag{25.21}$$

Such a $\eta_I \in \Theta$ will be our approximation of $\eta \in \Theta$. We now need to construct a suitable $v_I \in W_h$. From (25.21) and property **P3** we get

$$\mathrm{curl}\, R_h \eta_I = P_h \mathrm{curl}\, \eta_I = 0. \tag{25.22}$$

Property **P2** then implies that there exists $v_I \in W_h$ such that

$$\nabla v_I = R_h \eta_I. \tag{25.23}$$

Therefore, it holds for $(\eta_I, v_I) \in K_h$. In addition, since it holds for $\eta_I \approx \eta$ (see (25.21)), it follows that

$$R_h \eta_I \approx \eta = \nabla v. \tag{25.24}$$

Since $R_h \eta_I = \nabla v_I$ (see (25.23)), from (25.24) we deduce that

$$\nabla v_I \approx \nabla v, \tag{25.25}$$

which implies that $v_I \approx v$.

To summarize, using properties **P1**–**P4**, we have been able to find, for a given $(\eta, v) \in K$, a pair $(\eta_I, v_I) \in K_h$ such that $(\eta_I, v_I) \approx (\eta, v)$. This heuristically explains why the *MITC* elements are efficient.

Coming back to property **P4**, we point out that the connection between the Reissner–Mindlin problem and a Stokes-like problem is much deeper. Indeed, introducing the *Helmholtz decomposition* for $\lambda t^{-2}(\nabla w - \theta)$, that is,

$$\lambda t^{-2}(\nabla w - \theta) = \nabla \varphi + \mathrm{curl}\, p, \quad \varphi \in H_0^1(\Omega) = W, \quad p \in H_0^1(\Omega)/\mathbb{R}, \tag{25.26}$$

the plate problem (25.4) can be rewritten in the equivalent form (see [BrFo86])

$$\begin{cases} \text{Find } (\theta, w; \varphi, p) \in \Theta \times W \times W \times H_0^1(\Omega)/\mathbb{R} \text{ such that} \\ \displaystyle\int_\Omega \nabla\varphi \cdot \nabla v = \int_\Omega g v \quad \forall v \in W, \\ \begin{cases} \displaystyle\int_\Omega \mathbf{C}\varepsilon(\theta) : \varepsilon(\eta) - \int_\Omega p\, \mathrm{curl}\, \eta = \int_\Omega \nabla\varphi \cdot \eta \quad \forall \eta \in \Theta, \\ \displaystyle -\int_\Omega q\, \mathrm{curl}\, \theta - \lambda^{-1} t^2 \int_\Omega \mathrm{curl}\, p \cdot \mathrm{curl}\, q = 0 \quad \forall q \in H_0^1(\Omega)/\mathbb{R}, \end{cases} \\ \displaystyle\int_\Omega \nabla w \cdot \nabla \psi = \int_\Omega \theta \cdot \nabla\psi + \lambda^{-1} t^2 \int_\Omega \nabla\varphi \cdot \nabla\psi \quad \forall \psi \in W. \end{cases} \tag{25.27}$$

Above and in what follows, the curl operator is defined as

$$\mathrm{curl}\, q = (\partial q/\partial y, -\partial q/\partial x)^T,$$

for a generic scalar function q. Problem (25.27) reveals that the Reissner–Mindlin plate problem can be decomposed into

1. An elliptic problem for φ (the first variational equation);
2. A singularly perturbed *Stokes-like problem* for (θ, p) (the second and the third variational equation);
3. Another elliptic problem for w (the fifth variational equation).

In light of this reformulation, it should not be surprising that a Reissner–Mindlin element may contain ingredients peculiar to the mixed finite element machinery for the Stokes problem approximation.

25.3.2 The Arnold–Falk Element

We now present a triangular scheme which heavily exploits formulation (25.27) for the plate problem: the Arnold–Falk element (see [ArFa89]). This element is based on the following choices.

- Θ_h: each component of the rotation field is approximated by means of piecewise linear and globally continuous functions. In addition, a local cubic bubble is inserted per each triangle in the mesh.
- W_h: the vertical displacements are approximated by means of locally linear functions, which are continuous across adjacent elements *at the edge mid-points* (called the *non-conforming* P_1 *element*). It is easily seen that this approximation field is obtained by assigning its values at the edge mid-points.

The element is schematically shown in Figure 25.7.

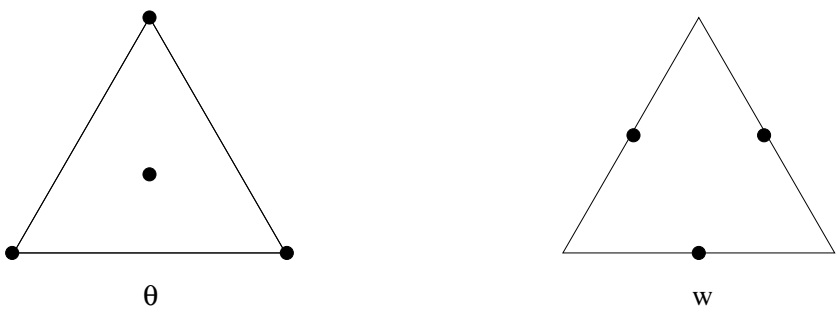


Fig. 25.7. The Arnold–Falk element.

Furthermore, we select R_h as P_0, the projection operator on the piecewise constant functions. Introducing the element-wise gradient operator ∇_h, and noting that $P_0 \nabla_h v_h = \nabla v_h$ for any $v_h \in W_h$, the discrete problem then reads

$$\begin{cases} \text{Find } (\theta_h, w_h) \in \Theta_h \times W_h \text{ which minimizes} \\ E_{h,t}(\eta_h, v_h) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) + \dfrac{\lambda\, t^{-2}}{2}\displaystyle\int_\Omega |\nabla_h v_h - P_0\eta_h|^2 - \displaystyle\int_\Omega g v_h. \end{cases} \tag{25.28}$$

The key point is to recognize that the *piecewise constant function* λt^{-2} $(\nabla_h w_h - P_0\theta)$ admits a *discrete Helmoholtz decomposition* (see (25.26)) as follows:

$$\lambda t^{-2}(\nabla_h w_h - R_h\theta) = \nabla_h\varphi_h + \operatorname{curl} p_h, \quad \varphi_h \in W_h, \quad p_h \in Q_h, \tag{25.29}$$

where Q_h is the space of piecewise linear and globally continuous functions. The discretization spaces for φ_h and p_h are depicted in Figure 25.8.

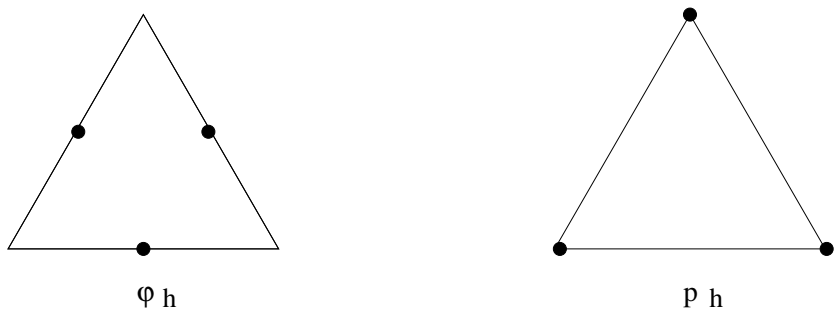


Fig. 25.8. Auxiliary spaces for the Arnold–Falk element.

Using that decomposition in the Euler–Lagrange equations emanated from problem (25.28), we get the variational system

$$\begin{cases} \text{Find } (\theta_h, w_h; \varphi_h, p_h) \in \Theta_h \times W_h \times W_h \times Q_h \text{ such that} \\ \displaystyle\int_\Omega \nabla_h\varphi_h \cdot \nabla_h v_h = \int_\Omega g v_h, \\ \begin{cases} \displaystyle\int_\Omega \mathbf{C}\varepsilon(\theta_h) : \varepsilon(\eta_h) - \int_\Omega p_h \operatorname{curl}\eta_h = \int_\Omega \nabla_h\varphi_h \cdot \eta_h, \\ \displaystyle -\int_\Omega q_h \operatorname{curl}\theta_h - \lambda^{-1}t^2\int_\Omega \operatorname{curl} p_h \cdot \operatorname{curl} q_h = 0, \end{cases} \\ \displaystyle\int_\Omega \nabla_h w_h \cdot \nabla_h\psi_h = \int_\Omega \theta_h \cdot \nabla\psi_h + \lambda^{-1}t^2\int_\Omega \nabla_h\varphi_h \cdot \nabla_h\psi_h, \end{cases} \tag{25.30}$$

for every $(\eta_h, v_h; \psi_h, q_h) \in \Theta_h \times W_h \times W_h \times Q_h$. Therefore, the Arnold–Falk scheme is equivalent to

1. Discretizing a Poisson problem by means of W_h, the space of non-conforming P_1 elements.

2. Discretizing a Stokes-like problem using the pair $\Theta \times Q_h$, which is the popular and stable *MINI element* (see [BrFo91], for instance).
3. Discretizing a further Poisson problem, still using W_h.

Since all the choices above are stable and convergent for the corresponding problems, the Arnold–Falk element results in a good approximation scheme for the Reissner–Mindlin plate problem.

25.3.3 Linked Interpolation Technique

We now describe a technique which has become quite popular, especially among the engineering community: the *linked interpolation technique* (see [AuTa94], [AuLo01], and [Lo98], for example). The main idea consists in improving the vertical displacements by using the rotational degrees of freedom. More precisely, the basic steps of this strategy are the following.

- Select finite element spaces Θ_h and W_h, as usual.
- Introduce a *suitable* linear operator (the *linking operator*)

$$L_h \,:\, \Theta_h \longrightarrow H_0^1(\Omega). \tag{25.31}$$

- Form the finite dimensional subspace of $\Theta \times W$:

$$X_h = \big\{(\eta_h, v_h^*) = (\eta_h, v_h + L_h\eta_h) : \eta_h \in \Theta_h\,,\ v_h \in W_h\big\}. \tag{25.32}$$

- Consider the discrete problem

$$\begin{cases} \text{Find } (\theta_h, w_h^*) \in X_h \text{ which minimizes} \\ E_{h,t}(\eta_h, v_h^*) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) + \dfrac{\lambda\, t^{-2}}{2}\displaystyle\int_\Omega |P_h(\nabla v_h^* - \eta_h)|^2 \\ \qquad\qquad\qquad\qquad\qquad\qquad - \displaystyle\int_\Omega g v_h^*, \end{cases} \tag{25.33}$$

  where P_h is typically a suitable L^2-projection operator.

The role of the *linking operator* L_h should be to help relax the constraint which causes locking effects. To give an example, we consider a triangular low-order element which corresponds to the following choices (see Figure 25.9).

- Θ_h: each component of the rotation field is approximated by means of piecewise linear and globally continuous functions. In addition, a local cubic bubble is inserted per each triangle in the mesh.
- W_h: the vertical displacements are approximated by means of piecewise linear and globally continuous linear functions.

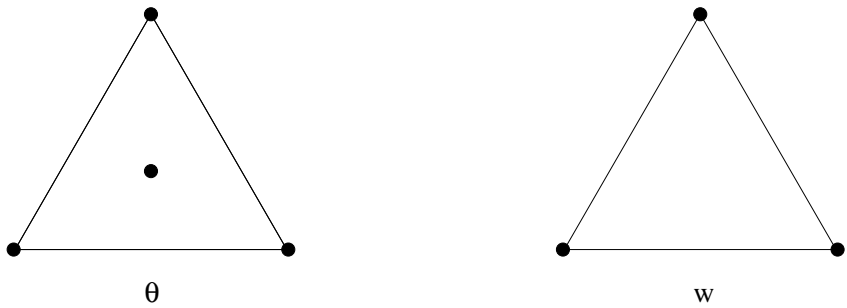


Fig. 25.9. Approximation of rotations and deflections for the lowest-order linked interpolation scheme.

The *linking operator* $L_h : \Theta_h \longrightarrow H_0^1(\Omega)$ is defined as follows. For each $T \in \mathcal{T}_h$, we set

$$\varphi_i = \lambda_j \lambda_k \quad \text{and} \quad EB_2(T) = \text{Span}\,\{\varphi_i\}_{1 \le i \le 3}\ , \tag{25.34}$$

where $\{\lambda_i\}_{1 \le i \le 3}$ are the barycentric coordinates of the triangle T and the indices (i, j, k) form a permutation of the set $(1, 2, 3)$. Then, the operator L_h is locally defined as

$$L_h \eta_{h|T} = \sum_{i=1}^{3} \alpha_i \varphi_i \in EB_2(T), \tag{25.35}$$

where the coefficients α_i are determined by requiring that

$$(\nabla L_h \eta_h - \eta_h) \cdot \mathrm{t} \quad \text{is constant on each } e. \tag{25.36}$$

Above, t denotes the tangential vector to the edge e. Therefore, a generic $v_h^* = v_h + L_h \eta_h$ (see (25.32)) is indeed a locally *quadratic* function. Finally, the operator P_h introduced in (25.33) is chosen as P_0, the L^2-projection operator over the piecewise constant functions.

The linked interpolation technique has some strong connections with the *MITC* elements described in Section 25.3.1. To see how this can occur, let us consider the term $\nabla v_h^* - \eta_h$ in (25.33). Recalling (25.32), we get

$$\nabla v_h^* - \eta_h = \nabla(v_h + L_h \eta_h) - \eta_h = \nabla v_h - [\eta_h - \nabla L_h \eta_h].$$

The vector $\eta_h - \nabla L_h \eta_h$ is often very similar (and sometimes even identical) to $R_h \eta_h$, where R_h is exactly the reduction operator of Section 25.3.1. For more details, the interested reader may see [Ly00].

25.3.4 PSRI Technique

The partial selective reduced integration (PSRI) technique is based on a suitable splitting of the shear energy term (see [ArBr93]). We illustrate the idea in

the easiest possible case: we first choose a real parameter α, with $0 < \alpha < t^{-2}$. We then introduce finite element spaces Θ_h and W_h, together with a reduction operator R_h. We finally consider the discrete problem:

$$\begin{cases} \text{Find } (\theta_h, w_h) \in \Theta_h \times W_h \text{ which minimizes} \\ E_{h,t}(\eta_h, v_h) = \dfrac{1}{2}\displaystyle\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) + \dfrac{\lambda\,\alpha}{2}\displaystyle\int_\Omega |(\nabla v_h - \eta_h)|^2 \\ \qquad\qquad + \dfrac{\lambda(t^{-2}-\alpha)}{2}\displaystyle\int_\Omega |R_h(\nabla v_h - \eta_h)|^2 - \displaystyle\int_\Omega g v_h. \end{cases} \tag{25.37}$$

Therefore, the shear energy term has been split into two parts, the first of which is *exactly* integrated, while the second one is *reduced* by means of R_h. The advantage of this formulation stands in the fact that the term

$$\frac{1}{2}\int_\Omega \mathbf{C}\varepsilon(\eta_h) : \varepsilon(\eta_h) + \frac{\lambda\,\alpha}{2}\int_\Omega |(\nabla v_h - \eta_h)|^2$$

is *always coercive over the whole space* $\Theta \times W$. As a consequence, spurious modes cannot occur independently of the chosen Θ_h and W_h spaces. With respect to the original discrete formulation (25.11), we now have much more flexibility in the choice of the approximation spaces. For instance, we could now reconsider the following element (see [Lo96]).

- Θ_h and W_h: both rotations and vertical displacements are discretized by means of piecewise quadratic and globally continuous functions (see Figure 25.10).
- $R_h = P_0$, the L^2-projection operator on the piecewise constant functions.

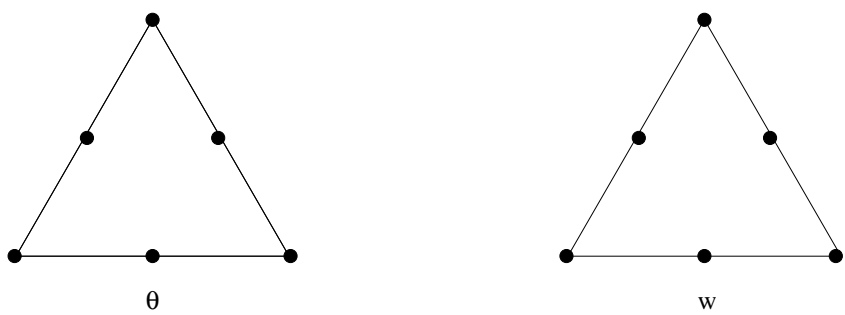


Fig. 25.10. Approximation of rotations and deflections for a low-order *PSRI* scheme.

This element shares the same degrees of freedom for all the kinematic variables, a feature which may be favorable for a possible extension to *shell* problems.

The main drawback of this approach is the presence of the parameter α to be chosen by the user. We point out that α cannot be arbitrarily selected. Indeed, looking at the *modified shear energy* (see (25.37))

$$\frac{\lambda\,\alpha}{2}\int_\Omega |(\nabla v_h - \eta_h)|^2 + \frac{\lambda(t^{-2}-\alpha)}{2}\int_\Omega |R_h(\nabla v_h - \eta_h)|^2,$$

one easily realizes the following.

- If α is "too small," we are essentially reducing the whole shear energy. Hence, spurious modes will likely occur.
- If α is "too large" (i.e., close to t^{-2}), we are essentially neglecting the effect of the reduction operator. Hence, the shear locking phenomenon will likely occur.

However, some numerical evidences reveal that the $PSRI$ technique is quite robust with respect to the parameter choice (see [ChLo95] and [AuLo99]). Therefore, one does not expect dramatic consequences even though one misses the "optimal" α (whatever "optimal" means in this context). We finally remark that one could set α varying from element to element, also selecting the *local* value $\alpha(T)$ as a function of the size of the current element T. This kind of choice sometimes leads to an improved performance of the scheme at hand. For more details on this point, as well as other similar techniques inspired by the *least-squares augmented formulations*, see [ChSt98] and [St95], for instance.

25.3.5 Non-Conforming Elements

Recently, the development of *discontinuous Galerkin (DG)* methods for elliptic problems have also suggested new approaches to the Reissner–Mindlin plate problems: non-conforming and DG-based elements have been designed and analyzed (see [ABM], [BrMa03], and [Mi01]). We however remark that a non-conforming element has already been presented in connection with the Arnold–Falk scheme (see Section 25.3.2), but only for the approximation of the vertical displacements.

We here focus on a *"fully" non-conforming* low-order triangular element stemming from the following choices, as proposed, analyzed, and numerically tested in [Lo05] and [CLM06].

- Θ_h and W_h: all the kinematic variables are approximated by means of locally linear functions, which are continuous across adjacent elements *at the edge mid-points* (*non-conforming P_1 element*, see Figure 25.11).
- $R_h = P_0$, the L^2-projection operator on the piecewise constant functions. Notice that $P_0(\nabla_h v_h) = \nabla_h v_h$ for every $v_h \in W_h$, which prevents the occurrence of spurious modes.

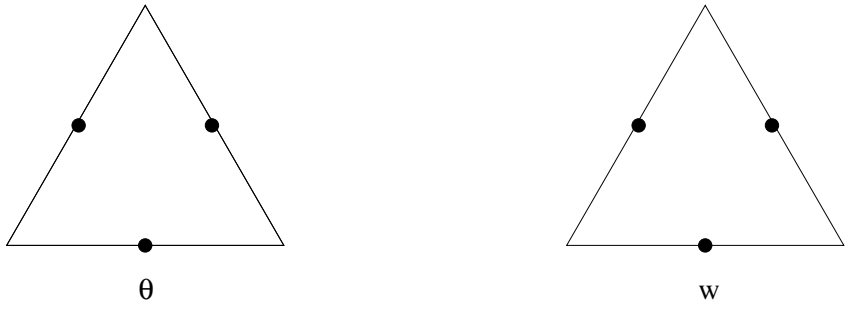


Fig. 25.11. Non-conforming element.

Then, the discrete problem reads

$$\begin{cases} \text{Find } (\theta_h, w_h) \in \Theta_h \times W_h \text{ which minimizes} \\ E_{h,t}(\eta_h, v_h) = \dfrac{1}{2} a_h(\eta_h, \eta_h) + \dfrac{\lambda\, t^{-2}}{2} \displaystyle\int_\Omega |\nabla_h v_h - P_0 \eta_h|^2 - \int_\Omega g v_h. \end{cases} \tag{25.38}$$

Above, the bilinear form $a_h(\cdot,\cdot)$ is defined by

$$a_h(\theta, \eta) := \int_\Omega \mathbf{C}\varepsilon_h(\theta) : \varepsilon_h(\eta) + \sum_{e \in \mathcal{E}_h} \frac{\kappa_e}{|e|} \int_e [\theta] : [\eta], \tag{25.39}$$

where

- ε_h denotes the element-by-element symmetric gradient operator,
- $[\cdot]$ is the jump operator,
- $\mathcal{E}_h$ is the set of edges e of $\mathcal{T}_h$,
- $|e|$ denotes the length of e,
- κ_e is a positive constant to be chosen.

We remark that κ_e must match the physical dimensions of $\mathbf{C}$. Therefore, a possible and highly reasonable choice is $\kappa_e = |\mathbf{C}|$, where $|\mathbf{C}|$ is some norm of $\mathbf{C}$. We point out that the jump term in (25.39), typical of the DG machinery, is necessary for stability: the term $\int_\Omega \mathbf{C}\varepsilon_h(\theta_h) : \varepsilon_h(\eta_h)$ alone is not positive definite on the non-conforming space Θ_h. In Figure 25.12 we display a rotation field $\eta_h \in \Theta_h$ such that $\int_\Omega \mathbf{C}\varepsilon_h(\eta_h) : \varepsilon_h(\eta_h) = 0$ but $\eta_h \neq 0$.

We also notice that the form $a_h(\cdot,\cdot)$ in (25.39) is a *strongly consistent* modification of the original form $\int_\Omega \mathbf{C}\varepsilon(\cdot) : \varepsilon(\cdot)$. In fact, computing $a_h(\cdot,\cdot)$ on smooth functions $\theta, \eta \in (H_0^1(\Omega))^2$, the jump term vanishes and one has $a_h(\theta, \eta) = \int_\Omega \mathbf{C}\varepsilon(\theta) : \varepsilon(\eta)$.

We now give a hint on why this approach gives rise to a locking-free scheme. For $t \to 0$, problem (25.38) becomes

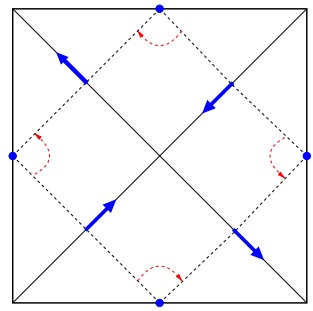


Fig. 25.12. Rotational spurious mode.

$$\begin{cases} \text{Find } (\theta_h^0, w_h^0) \in K_h \text{ which minimizes} \\ E_{h,0}(\eta_h, v_h) = \dfrac{1}{2} a_h(\eta_h, \eta_h) - \displaystyle\int_\Omega g v_h, \end{cases} \tag{25.40}$$

where

$$K_h = \{(\eta_h, v_h) \in \Theta_h \times W_h \,:\, \nabla_h v_h - P_0 \eta_h = 0\}\,. \tag{25.41}$$

Recalling the mid-point integration formula, one deduces from (25.41) that $(\eta_h, v_h) \in K_h$ means that the constraint $\nabla_h v_h = \eta_h$ is imposed *only* at the triangle barycenters, and *not everywhere*. Together with the continuity requirement only at the edge mid-points (see Figure 25.11), this makes the space K_h large enough to approximate the continuous space K defined in (25.7).

References

[ArBr93] Arnold, D.N., Brezzi, F.: Some new elements for the Reissner–Mindlin plate model, in *Boundary Value Problems for Partial Differential Equations and Applications,* Lions, J.-L., Baiocchi, C. (eds.), Masson, Paris (1993), 287–292.

[ABM] Arnold, D.N., Brezzi, F., Marini, L.D.: A family of discontinuous Galerkin finite elements for the Reissner–Mindlin plate. *J. Sci. Comp.* **22**, 25–45 (2005).

[ArFa89] Arnold, D.N., Falk, R.S.: A uniformly accurate finite element method for the Reissner–Mindlin plate. *SIAM J. Numer. Anal.,* **26**, 1276–1290 (1989).

[AuLo99] Auricchio, F., Lovadina, C.: Partial selective reduced integration schemes and kinematically linked interpolations for plate bending problems. *Math. Models Methods Appl. Sci.,* **9**, 693–722 (1999).

[AuLo01] Auricchio, F., Lovadina, C.: Analysis of kinematic linked interpolation methods for Reissner–Mindlin plate problems. *Comput. Methods Appl. Mech. Engrg.,* **190**, 18–19 (2001).

[AuTa94] Auricchio, F., Taylor, R.L.: A shear deformable plate element with an exact thin limit. *Comput. Methods Appl. Mech. Engrg.,* **118**, 393–412 (1994).

[Ba95] Bathe, K.J.: *Finite Element Procedures,* Englewood Cliffs, NJ (1995).

[BBF89] Brezzi, F., Bathe, K.J., Fortin, M.: Mixed-interpolated elements for Reissner–Mindlin plates. *Internat. J. Numer. Methods Engrg.,* **28**, 1787–1801 (1989).

[BrFo86] Brezzi, F., Fortin, M.: Numerical approximation of Mindlin–Reissner plates. *Math. Comput.,* **47**, 151–158 (1986).

[BrFo91] Brezzi, F., Fortin, M.: *Mixed and Hybrid Finite Element Methods,* Springer, New York (1991).

[BFS91] Brezzi, F., Fortin, M., Stenberg, R.: Error analysis of mixed-interpolated elements for Reissner–Mindlin plates. *Math. Models Methods Appl. Sci.,* **1**, 125–151 (1991).

[BrMa03] Brezzi, F., Marini, L.D.: A nonconforming element for the Reissner–Mindlin plate. *Computers and Structures,* **81**, 515–522 (2003).

[ChSt98] Chapelle, D., Stenberg, R.: An optimal low-order locking-free finite element method for Reissner–Mindlin plates. *Math. Models Methods Appl. Sci.,* **8**, 407–430 (1998).

[ChPa94] Chenais, D., Paumier, J.-C.: On the locking phenomenon for a class of elliptic problems. *Numer. Math.,* **67**, 427–440 (1994).

[ChLo95] Chinosi, C., Lovadina, C.: Numerical analysis of some mixed finite element methods for Reissner–Mindlin plates. *Comput. Mech.,* **16**, 36–44 (1995).

[CLM06] Chinosi, C., Lovadina, C., Marini, L.D.: Nonconforming locking-free finite elements for Reissner–Mindlin plates. *Comput. Methods Appl. Mech. Engrg.,* **195**, 3448–3460 (2006).

[Ci78] Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam (1978).

[Du92] Duran, R., Liberman, E.: On mixed finite element methods for the Reissner–Mindlin plate model. *Math. Comput.,* **58**, 561–573 (1992).

[Hu87] Hughes, T.J.R.: *The Finite Element Method,* Englewood Cliffs, NJ (1987).

[Lo96] Lovadina, C.: A new class of mixed finite element methods for Reissner–Mindlin plates. *SIAM J. Numer. Anal.,* **33**, 2457–2467 (1996).

[Lo98] Lovadina, C.: Analysis of a mixed finite element method for the Reissner–Mindlin plate problems. *Comput. Methods Appl. Mech. Engrg.,* **163**, 71–85 (1998).

[Lo05] Lovadina, C.: A low-order nonconforming finite element for Reissner–Mindlin plates. *SIAM J. Numer. Anal.,* **42**, 2688–2705 (2005).

[Ly00] Lyly, M.: On the connection between some linear triangular Reissner–Mindlin plate bending elements. *Numer. Math.,* **85**, 77–107 (2000).

[Mi01] Ming, P., Shi, Z.-C.: Nonconforming rotated Q_1 element for Reissner–Mindlin plate. *Math. Models Methods Appl. Sci.,* **11**, 1311–1342 (2001).

[SaPa92] Sanchez-Palencia, E.: Asymptotic and spectral properties of a class of singular-stiff problems. *J. Math. Pures Appl.,* **71**, 379–406 (1992).

[St95] Stenberg, R.: A new finite element formulation for the plate bending problem, in *Asymptotic Methods for Elastic Structures,* Ciarlet, P.G., Trabucho, L., Viaño, J. (eds.). de Gruyter (1995), 209–221.

[TeHu85] Tessler, A., Hughes, T.J.R.: A three-node Mindlin plate element with improved transverse shear. *Comput. Methods Appl. Mech. Engrg.,* **50**, 71–101 (1985).

26

Influence of a Weak Aerodynamics/Structure Interaction on the Aerodynamical Global Optimization of Shape

A. Nastase

RWTH-Aachen, Germany; nastase@lafaero.rwth-aachen.de

26.1 Introduction

The aim of this chapter is to perform a multidisciplinary aerodynamical global optimal design (OD) and to find the shape of a flying configuration (FC) that has high aerodynamical performance (a high ratio of L/D (lift to drag)) and also satisfies the requirements of the structure.

The global aerodynamical OD tries to determine the shapes of the external surface of an FC and of its planform in order to obtain high aerodynamical performances, and the structural OD tries to obtain a structure with minimal weight, inside this surface, with satisfactory stiffness. A certain degree of independence and also a certain interdependence occur between the global aerodynamical and structural ODs of the shape of the FC, which must be harmonized in order to obtain an FC with high aerodynamical performance which also satisfies all the requirements of the structure. In what follows we propose a weak aerodynamics/structure interaction via new and/or modified constraints that are required from a structural viewpoint.

26.2 The Three-Dimensional, Hyperbolic, Potential Solutions

In some previous papers [Na73], [Na86], and [Na07], the author has proposed some three-dimensional, hyperbolic, potential solutions for the computation of the axial disturbance velocities over several wings such as delta, rectangular, trapezoidal, and FCs such as wing–fuselage, wing–fuselage with leading edge flaps, etc. (all with arbitrary camber, twist, and thickness distributions), in inviscid supersonic flow. All these analytical solutions are written in closed (integrated) forms, have well-suited singularities, which are located only along the singular lines, and are easy to use for the computation of the inviscid aerodynamical characteristics and for performing the inviscid global OD of

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_26,

the shape of the FC. All these solutions are deduced by means of Carafoli's hydrodynamic analogy and van Dyke's minimum singularities principle, so they are matched, in the first approximation, with the Navier–Stokes layer (NSL) solutions.

The integrated wing–fuselage (IWF) is taken here as an example. The wing–fuselage (WF) is an FC whose wing has a central fuselage zone and is treated here as a wing alone, the surface of which is discontinuous along the junction lines between the wing and the fuselage. The junction lines are mathematically simulated as two artificial ridges. The IWFs are such WFs, with continuous mean surfaces; the thickness distributions on their wings and on their fuselage zones have the same tangent planes at each point of their junction lines but can be discontinuous in their higher derivatives. First, we introduce some dimensionless coordinates, namely

$$\tilde{x}_1 = \frac{x_1}{h_1}, \quad \tilde{x}_2 = \frac{x_2}{\ell_1}, \quad \tilde{x}_3 = \frac{x_3}{h_1},$$
$$\left(\tilde{y} = \frac{y}{\ell}, \quad \ell = \frac{\ell_1}{h_1}, \quad c = \frac{c'}{h_1}, \quad \nu = B\ell, \quad \bar{\nu} = B\,c, \quad B = \sqrt{M_\infty^2 - 1}\right).$$

Here, ℓ_1 and c' are the maximal half-span of the wing and of the fuselage zone, h_1 and ℓ are the maximal depth of the IWF and the dimensionless span of the wing, ν and $\bar{\nu}$ are the similarity parameters of the planform of the wing and of the fuselage zone, and the quotient $\bar{k} = \bar{\nu}/\nu$ depends on the purpose of the supersonic FC.

The downwashes w and w^*, w'^* on the thin and thick-symmetrical IWF components, respectively, are expressed as

$$w \equiv \tilde{w} = \sum_{m=1}^{N} \tilde{x}_1^{m-1} \sum_{k=0}^{m-1} \tilde{w}_{m-k-1,k}\, |\tilde{y}|^k$$

on the entire thin IWF (i.e., $-1 < \tilde{y} < 1$, $\bar{k} = \bar{\nu}/\nu$ with $\bar{\nu} = B\,c$), as

$$w^* \equiv \tilde{w}^* = \sum_{m=1}^{N} \tilde{x}_1^{m-1} \sum_{k=0}^{m-1} \tilde{w}^*_{m-k-1,k}\, |\tilde{y}|^k$$

on the parts of the thick-symmetrical component corresponding to the wing of the IWF (i.e., $-1 < \tilde{y} < -\bar{k}$ and $\bar{k} < \tilde{y} < 1$), and as

$$w'^* \equiv \bar{w}^* = \sum_{m=1}^{N} \tilde{x}_1^{m-1} \sum_{k=0}^{m-1} \bar{w}^*_{m-k-1,k}\, |\tilde{y}|^k$$

on the central part of the thick-symmetrical IWF component corresponding to the fuselage (i.e., $|\tilde{y}| < \bar{k}$).

The corresponding axial disturbance velocities u and u^* on the thin and, respectively, thick-symmetrical components of the thick, lifting IWF with subsonic leading edges (i.e., $\nu < 1$), fitted with two lateral artificial ridges (that

simulate the jump of the higher derivatives of the axial velocity along the junction lines WF) and also fitted with a central ridge, as in [Na86] and [Na07], are

$$u = \ell \sum_{n=1}^{N} \tilde{x}_1^{n-1} \left\{ \sum_{q=0}^{E\left(\frac{n}{2}\right)} \frac{\tilde{A}_{n,2q}\, \tilde{y}^{2q}}{\sqrt{1-\tilde{y}^2}} + \sum_{q=1}^{E\left(\frac{n-1}{2}\right)} \tilde{C}_{n,2q}\, \tilde{y}^{2q} \cosh^{-1} \sqrt{\frac{1}{\tilde{y}^2}} \right\},$$

$$\begin{aligned} u^* = \ell \sum_{n=1}^{N} \tilde{x}_1^{n-1} \Bigg[& \sum_{q=0}^{n-1} \tilde{H}^*_{nq}\, \tilde{y}^q \left(\cosh^{-1} M_1 + (-1)^q \cosh^{-1} M_2\right) \\ & + \sum_{q=0}^{E\left(\frac{n-2}{2}\right)} \tilde{D}^*_{n,2q}\, \tilde{y}^{2q} \sqrt{1-\nu^2\tilde{y}^2} + \sum_{q=1}^{E\left(\frac{n-1}{2}\right)} \tilde{C}^*_{n,2q}\, \tilde{y}^{2q} \cosh^{-1} \sqrt{\frac{1}{\nu^2\tilde{y}^2}} \\ & + \sum_{q=0}^{n-1} \tilde{G}^*_{nq}\, \tilde{y}^q \left(\cosh^{-1} S_1 + (-1)^q \cosh^{-1} S_2\right) \Bigg], \end{aligned}$$

where

$$M_1 = \sqrt{\frac{(1+\nu)(1-\nu\tilde{y})}{2\nu(1-\tilde{y})}}, \qquad M_2 = \sqrt{\frac{(1+\nu)(1+\nu\tilde{y})}{2\nu(1+\tilde{y})}},$$

$$S_1 = \sqrt{\frac{(1+\bar{\nu})(1-\nu\tilde{y})}{2(\bar{\nu}-\nu\tilde{y})}}, \qquad S_2 = \sqrt{\frac{(1+\bar{\nu})(1+\nu\tilde{y})}{2(\bar{\nu}+\nu\tilde{y})}}.$$

The coefficients of the axial and vertical disturbance velocities are related through Germain's compatibility conditions, which are linear and homogeneous relations with respect to the coefficients of the downwashes. On the upper side of the IWF, we have $u_e = -u + u^*$.

In a modern concept, these solutions are used twice: as outer solutions, at the edge of the NSL, and to reinforce the numerical solutions, inside the NSL.

26.3 NSL Spectral Forms of the Physical Entities

We introduce the spectral coordinate

$$\eta = (x_3 - Z(x_1, x_2))/\delta(x_1, x_2).$$

Here, $Z(x_1, x_2)$ is the equation of the upper surface of the FC and δ is the thickness of the NSL. Inside the NSL, the range of η is $0 < \eta < 1$, as desired.

We propose the following spectral forms for the velocity components, the density function R (defined here as $R = \ln \rho$), and the absolute temperature T (see [Na02, Na04], and [Na07]):

$$u_\delta = u_e \sum_{i=1}^{N} u_i \eta^i, \quad v_\delta = v_e \sum_{i=1}^{N} v_i \eta^i, \quad w_\delta = w_e \sum_{i=1}^{N} w_i \eta^i,$$

$$R = R_w + (R_e - R_w) \sum_{i=1}^{N} r_i \eta^i, \quad T = T_w + (T_e - T_w) \sum_{i=1}^{N} t_i \eta^i, \tag{26.1}$$

where R_w and T_w are the values of R and T at the wall and u_e, v_e, w_e, R_e, and T_e are the edge values of u_δ, v_δ, w_δ, and R, T, obtained from the outer inviscid flow at the edge of the NSL. The spectral coefficients u_i, v_i, and w_i of the velocity components u_δ, v_δ, and w_δ, and the spectral coefficients r_i and t_i of the density function R and the absolute temperature T are used to satisfy exactly the partial differential equations (PDEs) for the NSL at a finite number of points. By using the physical gas equation (i.e., the equation of a perfect gas), we find that the pressure p inside the NSL is expressed as a function only of the absolute temperature T and of the density function R; specifically,

$$p = R_g \rho T = R_g e^R T \quad . \tag{26.2}$$

The viscosity μ depends only on the temperature T. An exponential law is accepted, namely,

$$\mu = \mu_\infty \left(\frac{T}{T_\infty} \right)^{n_1}, \tag{26.3}$$

where R_g is the universal gas constant, μ_∞ and T_∞ are the values of viscosity and absolute temperature of the undisturbed flow and n_1 is the exponent of the exponential law ($n_1 = 0.76$ for air).

26.4 Dependence of the Density Function and Absolute Temperature on the Spectral Coefficients of the Velocity Components

We now consider the PDE of continuity; that is,

$$\frac{\partial(\rho u_\delta)}{\partial x_1} + \frac{\partial(\rho v_\delta)}{\partial x_2} + \frac{\partial(\rho w_\delta)}{\partial x_3} = 0.$$

If the density function $R = \ln \rho$ is introduced instead of the density ρ, the the PDE of continuity assumes the form

$$u_\delta \frac{\partial R}{\partial x_1} + v_\delta \frac{\partial R}{\partial x_2} + w_\delta \frac{\partial R}{\partial x_3} = -\left(\frac{\partial u_\delta}{\partial x_1} + \frac{\partial v_\delta}{\partial x_2} + \frac{\partial w_\delta}{\partial x_3} \right). \tag{26.4}$$

If the spectral form of R given in the fourth equation (26.1) is now introduced in the continuity equation (26.4), which is linear in the spectral

coefficients r_i of R, and if the collocation method is used, we arrive at the linear algebraic system

$$\sum_{i=1}^{N} g_{i,j} r_i = \gamma_j, \quad j = 1, 2, ..., N. \tag{26.5}$$

Solving this linear algebraic system, we obtain explicit expressions for the spectral coefficients r_i, which are functions only of the spectral coefficients of the velocity components.

Next, we consider the PDE for the absolute temperature T:

$$u_\delta \frac{\partial T}{\partial x_1} + v_\delta \frac{\partial T}{\partial x_2} + w_\delta \frac{\partial T}{\partial x_3} = \frac{1}{\rho C_p}\left[u_\delta \frac{\partial p}{\partial x_1} + v_\delta \frac{\partial p}{\partial x_2} + w_\delta \frac{\partial p}{\partial x_3} + \lambda \Delta_3 T + \mu \phi_d \right], \tag{26.6}$$

where p, ρ, and μ are the local pressure, density, and viscosity, λ is the coefficient of the thermal conductivity of the gas, C_p is the coefficient of specific heat at constant pressure, and ϕ_d is the dissipation function

$$\phi_d = 2\left[\left(\frac{\partial u_\delta}{\partial x_1}\right)^2 + \left(\frac{\partial v_\delta}{\partial x_2}\right)^2 + \left(\frac{\partial w_\delta}{\partial x_3}\right)^2 \right] + \left(\frac{\partial v_\delta}{\partial x_1} + \frac{\partial u_\delta}{\partial x_2}\right)^2 + \left(\frac{\partial w_\delta}{\partial x_2} + \frac{\partial v_\delta}{\partial x_3}\right)^2 + \left(\frac{\partial u_\delta}{\partial x_3} + \frac{\partial w_\delta}{\partial x_1}\right)^2 - \frac{2}{3}(\mathrm{div}\vec{V})^2.$$

The PDE (26.6) is used to compute the spectral coefficients t_i of the absolute temperature T as functions of the spectral coefficients u_i, v_i, w_i of the velocity components. If the spectral form of temperature, given by the fifth equation (26.1), the pressure p, given by (26.2), the density function R, expressed as a function only of the velocity components and obtained by solving the linear algebraic system (26.5), and the exponential law, given by (26.3) for the viscosity μ in terms of temperature, are used, then all these entities are eliminated from the temperature equation (26.6). The coefficients t_i depend only on the spectral coefficients u_i, v_i, w_i of the velocity components. If the collocation method is also applied, we obtain a transcendental algebraic system in the spectral coefficients t_i of the absolute temperature T:

$$\sum_{i=1}^{N} h_{ip} t_i + h_{0p}(T^{n_1})_p = \theta_p, \quad p = 1, 2, \ldots, N.$$

The coefficients h_{ip}, h_{0p}, and θ_p depend only on the spectral coefficients u_i, v_i, and w_i of the velocity components. The PDEs of the NSL are split, and all entities are expressed as functions only of the spectral coefficients of the velocity components, which are determined by solving the impulse PDEs of the NSL written in spectral forms, as in [Na07].

26.5 The Iterative Optimum-Optimorum Theory

The strategy of the aerodynamical global OD of the shape of the FC used here is a two-times enlarged variational method developed by the author.

The *first enlargement* allows the performing of the inviscid global OD of the shape of the FC (namely, the optimization of its distributions of camber, twist, and thickness, and *also* of its similarity parameters of the planform) and leads to an enlarged variational problem with free boundaries. The author developed an optimum-optimorum (OO) theory for the solution of this enlarged variational problem. The global optimized shape of the FC is chosen within *a class of admissible FCs* defined by some suitable properties. A lower-limit hypersurface of the drag functional as a function of the similarity parameters ν_i is defined, namely,

$$C_d^{(i)} = f(\nu_i).$$

Each point of this hypersurface is obtained by solving a classical variational problem with given boundaries (i.e., a given set of similarity parameters). The position of the minimum of this hypersurface, which is numerically determined, gives the best set of the similarity parameters, and the optimal shape of the FC, which corresponds to this set, is also the global optimized shape of the FC in the class. This OO theory was used by the author for the aerodynamical inviscid global optimization of the shapes of three models with respect to minimum inviscid drag, namely the delta wing model ADELA and the integrated WF models FADET I and FADET II, at cruising Mach numbers $M_\infty = 2; 2.2; 3$, respectively. All these three global optimized models have high aerodynamical performances.

The *second enlargement* of the variational method consists in the development of an iterative OO theory, in order to also introduce the influence of friction in the total drag functional and in the aerodynamical OD of the shape of the FC. The previous inviscid global optimized shape of the FC now represents the first step in the iterative viscous shape optimization process. An intermediate computational checking of the inviscid global optimized shape of the FC is made with the author's zonal, reinforced spectral viscous solutions for the three-dimensional NSL, which use the author's analytical hyperbolic potential solutions at the edge of the NSL edge and reinforce the numerical solutions of the NSL. These numerical solutions, with analytical properties, have correct jumps along the singular lines of the FCs and a correct last behavior. The friction drag coefficient $C_d^{(f)}$ of the FC is determined. The inviscid global optimized FC shape is also checked from a structural point of view. A weak aerodynamics/structure interaction via additional or modified constraints can produce important changes in the final global aerodynamical optimized shapes of FCs. The flow chart of the iterative OO theory with weak interaction is presented in Figure 26.1.

An intermediate computational checking of the inviscid global optimized shape of the FC is made with the author's zonal, spectral viscous solvers for

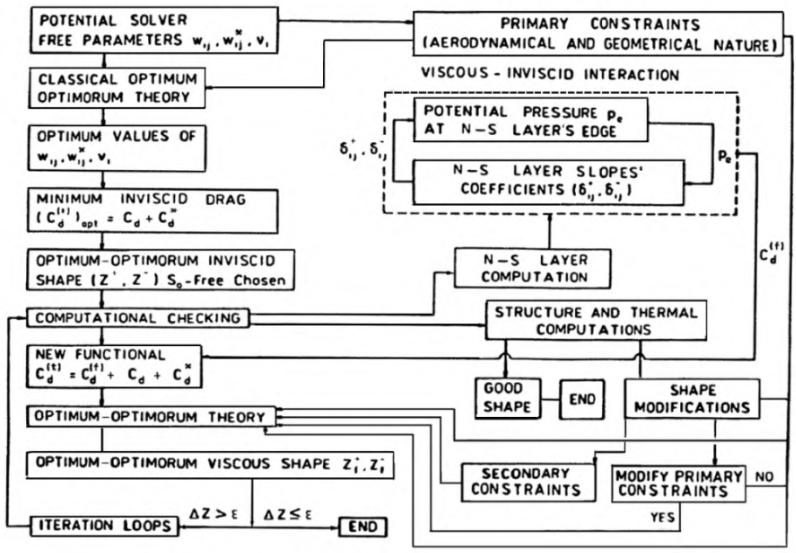


Fig. 26.1. The iterative optimum-optimorum theory, with weak interaction.

the three-dimensional PDEs of the NSL. The friction drag coefficient $C_d^{(f)}$ of the FC is determined. The inviscid global optimized shape of the FC is also checked from the structural point of view. Additional or modified constraints, introduced in order to control the camber, twist, and thickness distributions of the aerodynamical, global optimized shape of the FC for structural reasons, are proposed here. In the second step of optimization the predicted inviscid optimized shape of the FC is corrected by including these supplementary constraints in the variational problem and the friction drag coefficient in the drag functional. The iterative optimization process is repeated until the maximal local modification of the shape in two consecutive optimization steps presents no significant change.

26.6 Weak Aerodynamics/Structure Interaction

We propose a weak aerodynamics/structure interaction via new and modified constraints introduced for structural reasons in the global aerodynamical optimization problem, in order to obtain a final shape that is good from the aerodynamical point of view and also satisfies the stiffness requirements of the structure.

Reductions of the magnitudes of the aerodynamical optimized FC camber and twist distributions may be necessary, especially when the FC is optimized

at higher supersonic cruising Mach numbers. These reductions can be obtained if the Kutta condition on the leading edges of the FC is satisfied at a supersonic Mach number lower than the cruising Mach number (modified constraints).

The control of the magnitude of the aerodynamical optimized thickness of the FC along its central longitudinal section is useful because aerodynamical optimization has the tendency to push the maximal thickness away from the central section, especially in the rear part of the FC. This control can be realized by the introduction of a central fuselage zone on the wing, as treated in the previous section, in order to create enough place for a structure.

The augmentation of the aerodynamical optimized thickness distribution may be necessary, especially at the rear part of the FC, because the aerodynamical global optimized shape of the FC looks, in its longitudinal cuts in the vicinity of the trailing edge of the FC, like Joukowsky profiles. An augmentation of the thickness in the vicinity of the trailing edge with a small loss in drag is obtained by means of the following procedure. First, a small extension of the leading edges is made, and the zero thickness line is moved behind its initial position (on the trailing edge). The thickness distribution of the FC with an artificial augmented area of the planform is optimized at cruising Mach number, and the condition of null-thickness is now satisfied along the new artificial trailing edge, which is parallel to the initial trailing edge. After this optimization, the artificial augmented part of the FC, located behind the initial position of the trailing edge, is cut along the initial trailing edge and is eliminated. The global optimized thickness distribution is augmented in the rear part of the FC, as required by the structure stiffness. Also, the optimal distributions of thickness and of the angles of aperture of the FC along its initial trailing edge are small and asymptotically cancel along the artificial position of the trailing edge, as desired from the aerodynamical point of view.

There are two possibilities for designing a supersonic transport aircraft (STA).

• If the first, classical solution of the supersonic FC with one central integrated fuselage is chosen, the augmentation of the thickness of the FC in its central section can be obtained partially by increasing the relative thickness of the wing or, more efficiently, by introducing a central fuselage zone with augmented relative thickness, as used by the author for the design of the fully optimized and fully integrated models FADET I and FADET II. This classical solution is far from the tendency of aerodynamical thickness optimization.

• If the second, nonclassical solution of the FC with two twin integrated fuselages located in the vicinity of the central zone of the wing is adopted, the shape of this supersonic FC, proposed by the author in the form of a fully optimized and fully integrated Catamaran supersonic transport aircraft (CATA-STA) with fuselages almost completely embedded in the wing and shown in Figure 26.2, is more adequate for the requirement of aerodynamical thickness optimization: it has more lateral stability; it has more stiffness because, for the same number of passengers and the same transversal section, the fuselages are half as long as those obtained from the solution with one central

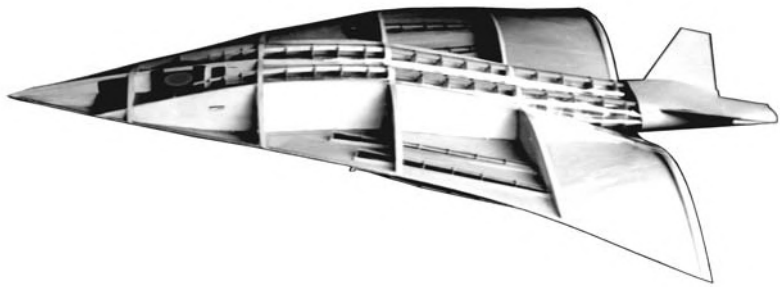


Fig. 26.2. Fully optimized and fully integrated catamaran STA.

fuselage; and, since it can fly with one characteristic surface (instead of two shock waves produced by a classical nonintegrated central fuselage), it cannot produce sonic boom interference.

26.7 Conclusions

The iterative optimum-optimorum theory is stable, robust, and rapidly convergent. It has almost all the attributes of genetic algorithms like evolutions, mutations, crossovers, and migrations, it easily allows multidisciplinary design, and it also takes care of friction. The theory is used for the optimal design of the proposed CATA-STA.

The analytical three-dimensional hyperbolic potential solutions previously given are successfully used as the outer flow at the edge of the NSL and to reinforce the Navier–Stokes zonal spectral solutions proposed here. The solution can be easily used by other researchers to reinforce and stabilize their own numerical solutions.

The weak aerodynamics/structure interaction via new or modified constraints introduced for structure reasons in the aerodynamical global optimal design, leads to a reshape of the FC in order to obtain an aerodynamical, global optimized shape that is also good for structural purposes.

The fully optimized and fully integrated shapes of the models ADELA, FADET I, FADET II, and CATA-STA, designed by the author using the OO theory, look like birds; namely, they are flattened, convex shaped in the frontal part, and have a wave shape in the rear part.

References

[Na73] Nastase, A.: *Use of Computers in the Optimization of Aerodynamic Shapes,* Editura Academii, Bucharest (1973) (Romanian).

[Na86] Nastase, A.: Optimum-optimorum integrated wing-fuselage configuration for supersonic transport aircraft of second generation, in *Proceedings 15th ICAS Congress,* London (1986).

[Na07] Nastase, A.: *Computation of Supersonic Flow over Flying Configurations,* Elsevier, Oxford (2007).

[Na02] Nastase, A.: Design of aerodynamical optimal shape of an integrated STA via spectral Navier–Stokes layer. AIAA-2002-5552, technical paper, 9th AIAA/ISSMO MAO Symposium, Atlanta, GA (2002).

[Na04] Nastase, A.: Zonal, spectral solutions for Navier–Stokes layer and applications, in *Proceedings Fourth ECCOMAS-2004,* Zienkiewicz, O.C. et al. (eds.), Wiley, Chichester (2004).

27

Multiscale Investigation of Solutions of the Wave Equation

M. Perel,[1] M. Sidorenko,[2] and E. Gorodnitskiy[2]

[1] Ioffe Physical-Technical Institute, St. Petersburg, Russia; perel@mph.phys.spbu.ru, eugy@yandex.ru
[2] St. Petersburg University, Russia; m-sidorenko@yandex.ru

27.1 Initial Value Problem for the Wave Equation

We consider here an initial value problem for the homogeneous wave equation with constant coefficients in three spatial dimensions, that is,

$$\begin{cases} u_{tt} - c^2(u_{xx} + u_{yy} + u_{zz}) = 0, \\ u|_{t=0} = w(\boldsymbol{r}), \quad \left.\dfrac{\partial u}{\partial t}\right|_{t=0} = v(\boldsymbol{r}). \end{cases} \tag{27.1}$$

The number of dimensions is not essential, and the method proposed can be generalized with minor changes to the case of an arbitrary number of spatial dimensions. We suppose that the initial data for the problem (27.1) has a complicated multiscale structure, i.e., the initial data possesses rapid changes of local frequency, a high degree of localization, singularities, discontinuities, and sharp edges. An example of such data is presented in Figure 27.1. We also note that this image is represented in discrete, not analytic, form. The most convenient mathematical apparatus for describing initial data of this kind is a continuous wavelet transform [AnMu04]. Not only does the wavelet transform contain complete information about the local structure of the data, i.e., it has an inverse, but it is also known to be the most adequate transform for qualitative analysis of the data.

When the initial data has a multiscale structure, the wave field is also multiscale at any time. This means that different spatial scales of a wave field at a fixed time may have localization in different spatial areas. Then it is useful to know the time evolution of the wavelet transform taken with respect to the spatial coordinates. We offer an analytic formula for the time dependency of the wavelet transform, which does not require the calculation of the wave field itself.

We define a class of solutions denoted by $\mathcal{H}$ of functions $u(\boldsymbol{r}, t) \in L_2(\mathbb{R}^3)$ that satisfy the wave equation as a distribution [GeSh67]:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\langle u(\boldsymbol{r}, t),\, \alpha(\boldsymbol{r})\rangle = c^2\,\langle u(\boldsymbol{r}, t),\, \Delta\alpha(\boldsymbol{r})\rangle \quad \forall \alpha \in S(\mathbb{R}^3), \tag{27.2}$$

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_27,

Fig. 27.1. An image of the Olga pond in Peterhof, St. Petersburg.

where

$$\langle u(\boldsymbol{r},t), \alpha(\boldsymbol{r})\rangle = \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{r}\, u(\boldsymbol{r},t)\, \overline{\alpha(\boldsymbol{r})}.$$

The initial conditions are taken in the form

$$u(\boldsymbol{r},0) = w(\boldsymbol{r}),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle u(\boldsymbol{r},t), \alpha(\boldsymbol{r})\rangle\Big|_{t=0} = \langle v(\boldsymbol{r}), \alpha(\boldsymbol{r})\rangle \quad \forall \alpha \in S(\mathbb{R}^3). \tag{27.3}$$

We require that $w(\boldsymbol{r}) \in L_2(\mathbb{R}^3)$ and $\widehat{v}(\boldsymbol{k})/|\boldsymbol{k}| \in L_2(\mathbb{R}^3)$, where $\widehat{v}(\boldsymbol{k})$ is the Fourier transform of $v(\boldsymbol{r})$. These conditions ensure that $u(\boldsymbol{r},t) \in L_2(\mathbb{R}^3)$.

This chapter presents a development of our results in [PeSi08].

27.2 A Continuous Wavelet Transform and Its Main Properties

To make the paper self-contained, we include necessary facts on the continuous wavelet transform. Numerous books on wavelet analysis are now available; for example, see [Da92], [AnMu04]. A wavelet transform F of a given function $f \in L_2(\mathbb{R}^3)$

$$F(a,\beta,\gamma,\boldsymbol{b}) = \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{r}\, f(\boldsymbol{r})\, \overline{\phi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r})}. \tag{27.4}$$

depends on the choice of an analyzing function ϕ referred to as a mother wavelet. The mother wavelet $\phi(\boldsymbol{r}) \in L_2(\mathbb{R}^3)$ is an arbitrary function with symmetry about the OZ axis. Usually the mother wavelet is assumed to satisfy an additional condition, which is necessary to use the inverse transform. We will discuss this condition below. The family of wavelets $\phi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r})$ is derived from the chosen mother wavelet $\phi(\boldsymbol{r})$ by the formula

$$\phi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r}) = \frac{1}{a^{3/2}}\phi\left(\mathrm{M}_{\beta\gamma}^{-1}\frac{\boldsymbol{r}-\boldsymbol{b}}{a}\right), \tag{27.5}$$

where $\mathrm{M}_{\beta\gamma}$ is the rotation matrix through angles β and γ,

$$\mathrm{M}_{\beta\gamma} = \mathrm{M}_\beta\,\mathrm{M}_\gamma = \begin{pmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{pmatrix},$$

$\boldsymbol{b} \in \mathbb{R}^3$ characterizes translation, and $a \in (0,\infty)$ is a dilation parameter. The coefficient $a^{-3/2}$ is introduced to retain the $L_2(\mathbb{R}^3)$ norm of wavelets independent of parameters.

The wavelet transform depends on a set of parameters $a,\beta,\gamma,\boldsymbol{b}$. With a simple example we show what kind of information can be extracted from the wavelet transform for different values of parameters. The choice of a mother wavelet determines the possibilities of analyzing the data with the help of the wavelet transform. The Morlet wavelet [Da92], [AnMu04] reads

$$\varphi(\boldsymbol{r}) = \mathrm{e}^{-|\boldsymbol{r}|^2}\mathrm{e}^{\mathrm{i}\boldsymbol{r}\cdot\boldsymbol{l}}, \quad |\boldsymbol{l}| = \kappa. \tag{27.6}$$

The vector $\boldsymbol{l}$ is the direction of the Morlet wavelet. The wavelet transform (27.4) with the Morlet wavelet can be interpreted as a windowed Fourier transform

$$F(a,\beta,\gamma,\boldsymbol{b}) = a^{-3/2}\int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{r}\, f(\boldsymbol{r})\, \mathrm{e}^{-|\boldsymbol{r}-\boldsymbol{b}|^2/a^2}\mathrm{e}^{-\mathrm{i}(\boldsymbol{r}-\boldsymbol{b})\cdot\mathrm{M}_{\beta\gamma}\boldsymbol{l}/a}.$$

The vector $\mathrm{M}_{\beta\gamma}\boldsymbol{l}/a$ has the sense of a wave vector $\boldsymbol{k}$. The exponent $\exp(-|\boldsymbol{r}-\boldsymbol{b}|^2/a^2)$ is a window that cuts part of the function f in the vicinity of the point $\boldsymbol{b}$. The modulus of $\boldsymbol{k}$ is κa^{-1}, and the angles β,γ determine its spatial direction. Small values of a, i.e., large values of spatial frequencies, are responsible for discontinuities near the point $\boldsymbol{b}$ in space. It is known that any wavelet also has this property if $\widehat{\phi}(0) = 0$ and $\phi(\boldsymbol{r}) \in L_1(\mathbb{R}^3)$. The Morlet wavelet (27.6) satisfies this condition approximately for large κ. The Morlet wavelet extracts discontinuities near the point $\boldsymbol{b}$ in the direction defined by the angles β and γ. If the direction is not important, the integral

Fig. 27.2. The modulus of the direction-independent wavelet transform of the image in Figure 27.1 for small a plotted against b_x, b_y.

$$\int_0^{2\pi} \mathrm{d}\beta \int_0^{\pi} \mathrm{d}\gamma \, \sin\gamma \, |F(a,\beta,\gamma,\boldsymbol{b})|^2$$

as a function of $\boldsymbol{b}$ for a small fixed a characterizes the distribution of discontinuities. The same analysis of the data can be obtained by means of a spherically symmetric wavelet, for example, by the Mexican hat [AnMu04].

A similar analysis can be carried out in the two-dimensional case. We demonstrate the possibilities of the wavelet transform with two-dimensional examples. The formula for the wavelet transform is similar to (27.4), but the family of solutions reads

$$\phi^{a,\beta,\boldsymbol{b}}(\boldsymbol{r}) = \frac{1}{a}\phi\left(\mathrm{M}_\beta^{-1}\frac{\boldsymbol{r}-\boldsymbol{b}}{a}\right), \quad \mathrm{M}_\beta = \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix}.$$

The modulus of the wavelet transform of the image in Figure 27.1 for small fixed $a = 0.003$ is plotted against b_x and b_y in Figure 27.2. It is calculated with a mother wavelet [KiPe00], [PeSi07] for the parameters $p = 0.4$ and $\varepsilon = \gamma = 4$. The wavelet is almost spherically symmetric. The level of brightness is indicative of the value of $|F(a,\boldsymbol{b})|$.

To show the directional potentialities of the wavelet transform, we consider a simpler example plotted in Figure 27.3.

In this case, the function f is a characteristic function of a rectangle. We use a wavelet [KiPe00], [PeSi07] that is numerically close to the Morlet wavelet

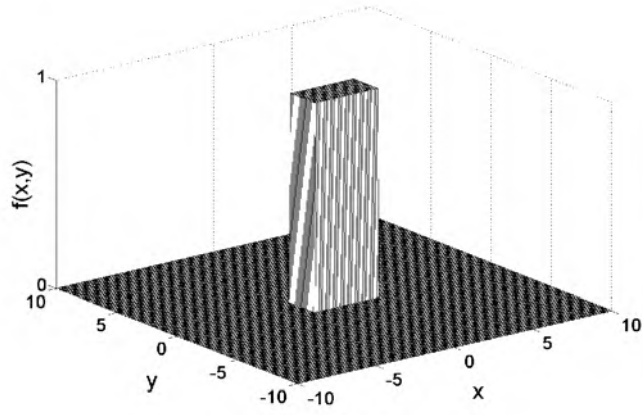


Fig. 27.3. A simple example of a function with discontinuities.

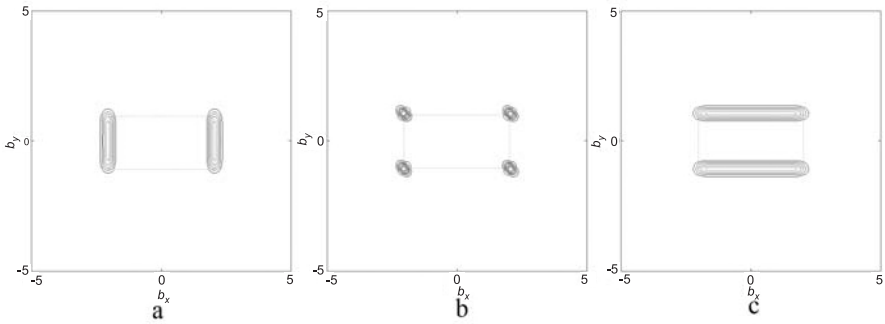


Fig. 27.4. Levels of the modulus of the wavelet transform of the function in Figure 27.3 for small a plotted against b_x, b_y for (a) $\beta = 0^o$, (b) $\beta = 45^o$, and (c) $\beta = 90^o$. The contour of the original rectangle is plotted with dotted lines.

if we choose the parameters $p = 4$, $\varepsilon = 16$, and $\gamma = 0.5$. The modulus of the wavelet transform of f is plotted against b_x and b_y in the subfigures of Figure 27.4. All these subfigures are built for one and the same small fixed value $a = 0.1$ but for different fixed angles β. The angle is measured counterclockwise from the x-axis. In the pictures, we see the lines perpendicular to the direction of the axis, $\boldsymbol{l}$, of the wavelet. If the direction $\boldsymbol{l}$ does not coincide with the direction of the sides of the rectangle, we observe only the corners.

The global directional properties of the image are characterized by the integral function [AnMu04]

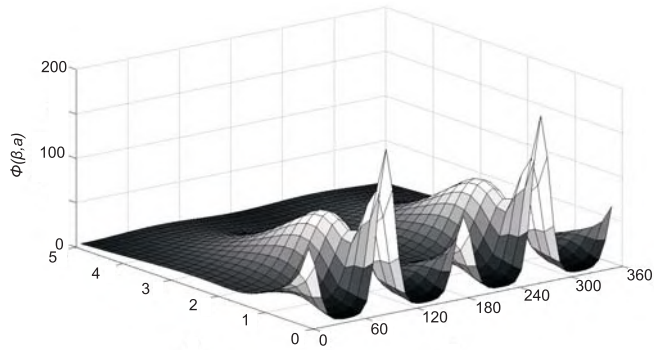


Fig. 27.5. Scale-angle diagram Φ of the function in Figure 27.3 plotted against β and a, $a \in (0,5)$, $\beta \in (0,360)$.

$$\Phi(\beta, a) = \int\limits_{\mathbb{R}^2} \mathrm{d}^2\boldsymbol{b}\, |F(a,\beta,\boldsymbol{b})|^2$$

plotted against a and β in Figure 27.5.

We see that Φ has local maxima in the direction perpendicular to the sharp sides of the rectangle, which are $\beta = 0^o$ and $\beta = 180^o$, $\beta = 90^o$ and $\beta = 270^o$ for small values of a. For each of these two angles we have a maximum for $a = 0$ and one more maximum that characterizes the width and the length of the rectangle.

The wavelet transform can be inverted:

$$f(\boldsymbol{r}) = \frac{1}{C_{\phi\chi}} \int\limits_0^{2\pi} \mathrm{d}\beta \int\limits_0^{\pi} \mathrm{d}\gamma\, \sin\gamma \int\limits_0^{\infty} \frac{\mathrm{d}a}{a^4} \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{b}\, F(a,\beta,\gamma,\boldsymbol{b})\, \chi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r}), \qquad (27.7)$$

where $\chi(\boldsymbol{r}) \in L_2(\mathbb{R}^3)$ is another axisymmetric mother wavelet with the OZ axis, $\chi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r})$ is the family of wavelets constructed according to (27.5), and

$$C_{\phi\chi} \equiv \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k} \frac{\widehat{\phi}(\boldsymbol{k})\overline{\widehat{\chi}(\boldsymbol{k})}}{|\boldsymbol{k}|^3}, \qquad \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k} \frac{|\widehat{\phi}(\boldsymbol{k})\widehat{\chi}(\boldsymbol{k})|}{|\boldsymbol{k}|^3} < \infty.$$

The formula (27.7) allows us to represent the function $f(\boldsymbol{r})$ as a superposition of the functions $\chi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r})$, which form an overcomplete set. The wavelets ϕ and χ may coincide.

A simplified reconstruction formula also exists [AnMu04]. It allows one to reconstruct the function f from its wavelet transform directly, without the wavelet χ:

$$f(\boldsymbol{r}) = \frac{1}{\widetilde{C}} \int\limits_0^\infty \frac{\mathrm{d}a}{a^{5/2}} \int\limits_0^{2\pi} \mathrm{d}\beta \int\limits_0^\pi \mathrm{d}\gamma\, \sin\gamma\, F(a,\beta,\gamma,\boldsymbol{r}),$$

where

$$\widetilde{C} \equiv \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k}\, \frac{\overline{\widehat{\phi}(\boldsymbol{k})}}{|\boldsymbol{k}|^3}, \quad \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k}\, \frac{|\widehat{\phi}(\boldsymbol{k})|}{|\boldsymbol{k}|^3} < \infty. \tag{27.8}$$

27.3 Time-Dependent Wavelet Transform

We study here wavelet transform (27.4) of a solution with respect to $\boldsymbol{r}$, the time t being a parameter:

$$U(a,\beta,\gamma,\boldsymbol{b};t) \equiv \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{r}\, u(\boldsymbol{r},t)\, \overline{\phi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r})}. \tag{27.9}$$

To state the result we choose a solution $\varphi(\boldsymbol{r},t)$ of the wave equation (27.2) that belongs to $L_2(\mathbb{R}^3)$ with respect to r and has only positive frequencies, i.e.,

$$\varphi(\boldsymbol{r},t) = \frac{1}{(2\pi)^3} \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k}\, \widehat{\varphi}(\boldsymbol{k},t) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}, \quad \widehat{\varphi}(\boldsymbol{k},t) = \widehat{\phi}(\boldsymbol{k}) \mathrm{e}^{-\mathrm{i}|\boldsymbol{k}|ct}.$$

We construct a family of solutions by the formula

$$\varphi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r},t) = \frac{1}{a^{3/2}} \varphi\left(\mathrm{M}_{\beta\gamma}^{-1} \frac{\boldsymbol{r}-\boldsymbol{b}}{a}, \frac{t}{a}\right).$$

When $t=0$, this formula determines a family of wavelets (27.5), where $\phi(\boldsymbol{r}) = \varphi(\boldsymbol{r},0)$. We introduce the second solution $\psi(\boldsymbol{r},t)$ as

$$\psi(\boldsymbol{r},t) = \int\limits_{-\infty}^{t} \mathrm{d}\tau \varphi(\boldsymbol{r},\tau).$$

We require that the solution $\psi(\boldsymbol{r},t)$ belong to $L_2(\mathbb{R}^3)$. If the solution $\varphi(\boldsymbol{r},t)$ additionally belongs to $L_1(\mathbb{R}^3)$, then $\widehat{\varphi}(\boldsymbol{k})$ is continuous and bounded in the vicinity of $\boldsymbol{k}=0$ and then $\widehat{\varphi}(\boldsymbol{k})/|\boldsymbol{k}| \in L_2(\mathbb{R}^3)$ and $\psi(\boldsymbol{r},t) \in L_2(\mathbb{R}^3)$ for fixed t.

Proposition 1. *The time-dependent wavelet transform (27.9) of the solution of the initial value problem (27.3) is expressed as*

$$\begin{aligned} &U(a,\beta,\gamma,\boldsymbol{b},t) \\ &= \frac{1}{2} \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{r}\, w(\boldsymbol{r}) \left[\overline{\varphi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r},t)} + \overline{\varphi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r},-t)}\right] \\ &+ \frac{a}{2} \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{r}\, v(\boldsymbol{r}) \left[\overline{\psi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r},t)} - \overline{\psi^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{r},-t)}\right]. \end{aligned} \tag{27.10}$$

Proof. We rewrite the expression (27.9) using the Plancherel formula

$$U(a,\beta,\gamma,\boldsymbol{b};t) = \frac{1}{(2\pi)^3}\int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k}\,\widehat{u}(\boldsymbol{k},t)\,\overline{\widehat{\phi}^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{k})}. \tag{27.11}$$

The Fourier transform of each solution of the wave equation (27.2) can be split into a sum of positive- and negative-frequency components

$$\widehat{u}(\boldsymbol{k},t) = \widehat{u}_+(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}|\boldsymbol{k}|ct} + \widehat{u}_-(\boldsymbol{k})\mathrm{e}^{\mathrm{i}|\boldsymbol{k}|ct}. \tag{27.12}$$

The initial conditions yield

$$\widehat{u}_+(\boldsymbol{k}) = \frac{1}{2}\left[\widehat{w}(\boldsymbol{k}) - \frac{1}{\mathrm{i}c|\boldsymbol{k}|}\widehat{v}(\boldsymbol{k})\right], \quad \widehat{u}_-(\boldsymbol{k}) = \frac{1}{2}\left[\widehat{w}(\boldsymbol{k}) + \frac{1}{\mathrm{i}c|\boldsymbol{k}|}\widehat{v}(\boldsymbol{k})\right]. \tag{27.13}$$

We substitute (27.13) into (27.12) and then into (27.11). Upon combining the terms, we obtain

$$\begin{aligned} U(a,\beta,\gamma,\boldsymbol{b};t) = \frac{1}{2(2\pi)^3}\int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k}\,\Bigg\{\widehat{w}(\boldsymbol{k})\left[\overline{\widehat{\phi}^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{k})\mathrm{e}^{\mathrm{i}|\boldsymbol{k}|ct}} + \overline{\widehat{\phi}^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}|\boldsymbol{k}|ct}}\right] \\ + \widehat{v}(\boldsymbol{k})\left[-\frac{1}{\mathrm{i}c|\boldsymbol{k}|}\,\overline{\widehat{\phi}^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{k})\mathrm{e}^{\mathrm{i}|\boldsymbol{k}|ct}} + \frac{1}{\mathrm{i}c|\boldsymbol{k}|}\,\overline{\widehat{\phi}^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}|\boldsymbol{k}|ct}}\right]\Bigg\}. \end{aligned} \tag{27.14}$$

We note that

$$\widehat{\phi}(\boldsymbol{k})\mathrm{e}^{\mp\mathrm{i}|\boldsymbol{k}|ct} = \widehat{\varphi}(\boldsymbol{k},\pm t), \qquad \frac{1}{\mathrm{i}c|\boldsymbol{k}|}\widehat{\phi}(\boldsymbol{k})\mathrm{e}^{\mp\mathrm{i}|\boldsymbol{k}|ct} = -\widehat{\psi}(\boldsymbol{k},\pm t). \tag{27.15}$$

The representation of $\widehat{\psi}^{a,\beta,\gamma,\boldsymbol{b}}$ in terms of $\widehat{\varphi}^{a,\beta,\gamma,\boldsymbol{b}}$ contains an additional factor a:

$$\frac{1}{\mathrm{i}ca|\boldsymbol{k}|}\widehat{\phi}^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{k})\mathrm{e}^{\mp\mathrm{i}|\boldsymbol{k}|ct} = -\widehat{\psi}^{a,\beta,\gamma,\boldsymbol{b}}(\boldsymbol{k},\pm t). \tag{27.16}$$

We obtain (27.10) from (27.14) by means of the Plancherel formula and (27.15) and (27.16).

27.4 Numerical Examples of Wavelet Transform at Different Moments in Time

We consider the functions in Figure 27.1 and 27.3 as the initial data $w(\boldsymbol{r})$ and obtain a time-dependent wavelet transform by (27.10), taking into account only the positive-frequency part. The wavelet transform of the function in Figure 27.3 at a fixed time $t_1(ct_1 = 5)$ for small $a = 0.1$ and a fixed set of angles $\beta = 0^o$, 45°, 90° is plotted against b_x and b_y in Figure 27.6. The directional-independent wavelet transform (Figure 27.2) of the image in Figure 27.1 calculated at a time $t_*(ct_* = 100)$ for small values of $a = 0.003$ is plotted against b_x and b_y in Figure 27.7.

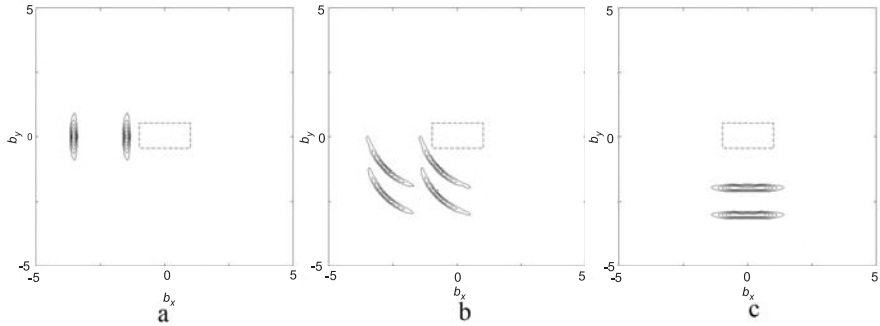


Fig. 27.6. Levels of the modulus of the wavelet transform of the function in Figure 27.3 at time t_1 for small a plotted against b_x, b_y for (a) $\beta = 0^o$, (b) $\beta = 45^o$, and (c) $\beta = 90^o$. The contour of the original rectangle is plotted with dotted lines.

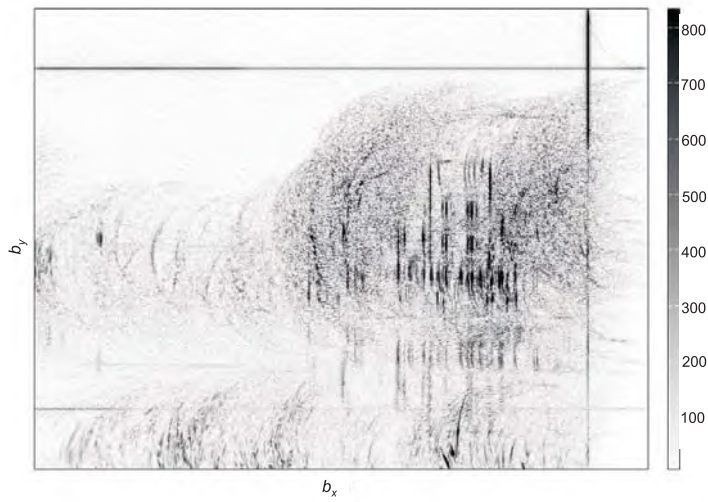


Fig. 27.7. The modulus of the direction-independent wavelet transform of the image in Figure 27.1 at time t_* for a small scale a plotted against b_x, b_y.

27.5 Reconstruction Formula

Proposition 2. *The solution of the initial value problem (27.3) can be reconstructed from the time-dependent wavelet transform (27.9) by the formula*

$$u(\boldsymbol{r},t) = \frac{1}{\widetilde{C}} \int\limits_0^\infty \frac{\mathrm{d}a}{a^{5/2}} \int\limits_0^{2\pi} \mathrm{d}\beta \int\limits_0^\pi \mathrm{d}\gamma\, \sin\gamma\, U(a,\beta,\gamma,\boldsymbol{r};t),$$

where $\widetilde{C}$ is defined by (27.8). (This is understood as an equality of distributions. The formula can be useful for partial reconstruction of the field if we are interested only in propagation in a given interval of directions for given scales.)

Proof. We consider the expression

$$\frac{1}{\widetilde{C}} \int\limits_0^\infty \frac{\mathrm{d}a}{a^{5/2}} \int\limits_0^{2\pi} \mathrm{d}\beta \int\limits_0^\pi \mathrm{d}\gamma\, \sin\gamma\, \langle U(a,\beta,\gamma,\boldsymbol{r};t),\, \alpha(\boldsymbol{r})\rangle \quad \forall \alpha \in S(\mathbb{R}^3). \tag{27.17}$$

The Plancherel formula gives

$$\langle U(a,\beta,\gamma,\boldsymbol{r};t),\, \alpha(\boldsymbol{r})\rangle = \frac{1}{(2\pi)^3} \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k}\; \widehat{u}(\boldsymbol{k},t)\, a^{3/2}\, \overline{\widehat{\phi}\left(a\mathrm{M}_{\beta\gamma}^{-1}\boldsymbol{k}\right)\, \widehat{\alpha}(\boldsymbol{k})}. \tag{27.18}$$

Substituting (27.18) into (27.17) and changing the order of integration yield

$$\frac{1}{\widetilde{C}} \frac{1}{(2\pi)^3} \int\limits_{\mathbb{R}^3} \mathrm{d}^3\boldsymbol{k}\, \widehat{u}(\boldsymbol{k},t)\, \overline{\widehat{\alpha}(\boldsymbol{k})} \int\limits_0^\infty \frac{\mathrm{d}a}{a} \int\limits_0^{2\pi} \mathrm{d}\beta \int\limits_0^\pi \mathrm{d}\gamma\, \sin\gamma\, \overline{\widehat{\phi}\left(a\mathrm{M}_{\beta\gamma}^{-1}\boldsymbol{k}\right)}.$$

Taking a new variable of integration $\boldsymbol{\zeta} = a\mathrm{M}_{\beta\gamma}^{-1}\boldsymbol{k}$ in the inner integral and using the fact that, in the case of an axially symmetric wavelet, the domain of integration is $\mathbb{R}^3$, we obtain $\widetilde{C}$. Plancherel's formula implies that the expression is equal to $\langle u, \alpha\rangle$.

References

[AnMu04] Antoine, J.P., Murenzi, R., Vandergheynst, P., Ali, S.T.: *Two-Dimensional Wavelets and Their Relatives,* Cambridge University Press, London (2004).

[GeSh67] Gelfand, I.M., Shilov, G.E.: *Generalized Functions. Vol. 3: Theory of Differential Equations,* Academic Press, New York (1967).

[Da92] Daubechies, I.: *Ten Lectures on Wavelets,* SIAM, Philadelphia, PA (1992).

[KiPe00] Kiselev, A.P., Perel, M.V.: Highly localized solutions of the wave equation. *J. Math. Phys.,* **41**, 1934–1955 (2000).

[PeSi07] Perel, M.V., Sidorenko, M.S.: New physical wavelet "Gaussian wave packet." *J. Phys. A,* **40**, 3441–3461 (2007).

[PeSi08] Perel, M.V., Sidorenko, M.S.: Wavelet-based integral representation for solutions of the wave equation. Preprint arXiv:0809.2211 [math-ph] (2008).

28

The Laplace Transform Method for the Albedo Boundary Conditions in Neutron Diffusion Eigenvalue Problems

C.Z. Petersen,[1] M.T. Vilhena,[1] D. Moreira,[2] and R.C. Barros[3]

[1] Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil; claudiopetersen@yahoo.com.br, vilhena@pq.cnpq.br
[2] Universidade Federal do Pampa, Bagé, RS, Brazil; davidson@pq.cnpq.br
[3] Universidade do Estado do Rio de Janeiro, Nova Friburgo, RJ, Brazil; rcbarros@pq.cnpq.br

28.1 Introduction

Some heavy nuclei are fissile after having absorbed a neutron, i.e., they violently split into two unequal fragments, while at the same time ejecting two or three neutrons on average. This phenomenon is called fission. Neutrons ejected during one fission can, in turn, be absorbed by other neighboring fissile nuclei, thus creating a chain reaction. If this reaction is controlled and stabilized, one gets an energy source—this is what happens in a nuclear reactor [WF07]. Nuclear power is a proven technology and has the potential to generate virtually limitless energy with no significant greenhouse gas emissions. From a physical understanding of criticality, it appears that any system containing fissile material could be made critical by arbitrarily varying the number of neutrons emitted in fission. It is well known that criticality calculations can often be best approached by solving eigenvalue problems. In elementary nuclear reactor theory, the dominant eigenvalue, i.e., the effective multiplication factor *(keff)*, is thought of as the ratio between the numbers of neutrons generated in successive fission reactions. The eigenfunction corresponding to the dominant eigenvalue is proportional to the neutron flux within the reactor core. Furthermore, in most realistic reactor global calculations, it is necessary to consider an approximation of the energy-dependent eigenvalue problem in which the energy variable is discretized. The most common energy discretization method is the conventional multigroup approximation, in which the neutron energy range is divided into contiguous energy groups. In practice, multigroup diffusion theory has been applied extensively to nuclear reactor analyses and generally found to perform better than it theoretically has any right to, because it does not include the direction-of-motion variable [AlOd86]. Neutron fission events do not take place in the non-multiplying regions of nuclear re-

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_28,

actors, e.g., moderator, reflector, and structural core; therefore, we claim that we can improve the efficiency of nuclear reactor global calculations by eliminating the explicit numerical calculations within the non-multiplying regions around the active domain. In this chapter, we describe the application of the Laplace transform method in order to determine the energy-dependent albedo matrix that we use in the boundary conditions of multigroup neutron diffusion eigenvalue problems in slab geometry to substitute the explicit numerical calculations within the baffle–reflector system around a thermal nuclear reactor core. Albedo, the Latin word for "whiteness," was defined by Lambert (1760) as the fraction of incident light reflected diffusely by a surface [Pa61]. This word has remained the usual scientific term in astronomy. Here, we extend it to the reflection of neutrons. At this point, an outline of the remainder of this chapter follows. In Section 28.2 we present the mathematical formulation. In Section 28.3 we present numerical results, while concluding remarks with suggestions for future work are given in Section 28.4.

28.2 Mathematical Formulation

Let us consider Figure 28.1, which illustrates a slab where regions F stand for the fuel regions, region B is the baffle of width $l_b = x_b - x_a$, and region R is the reflector of width $l_r = x_c - x_b$. Our goal is to determine an albedo matrix that we can use in the boundary conditions at $x = x_a$ of multigroup neutron diffusion eigenvalue problems in slab geometry to substitute the explicit numerical calculations within the baffle–reflector system ($x_a \leq x \leq x_c$), which does not generate power.

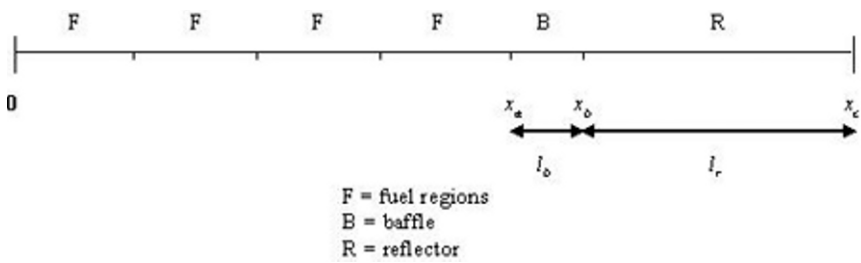


Fig. 28.1. Slab domain.

We now write the two-energy group slab-geometry neutron diffusion equations for the two non-multiplying regions of Figure 28.1:

$$\frac{dJ_{1,q}(x)}{dx} + \Sigma_{R1,q}(x)\phi_{1,q}(x) = 0, \tag{28.1}$$

$$J_{1,q}(x) = -D_{1,q}\frac{d\phi_{1,q}(x)}{dx}, \tag{28.2}$$

$$\frac{dJ_{2,q}(x)}{dx} + \Sigma_{a2,q}(x) = \Sigma_{s1\to 2,q}\phi_{1,q}(x), \tag{28.3}$$

$$J_{2,q}(x) = -D_{2,q}\frac{d\phi_{2,q}(x)}{dx}. \tag{28.4}$$

Equations (28.1) and (28.2) hold for the fast energy group ($g = 1$) and (28.3) and (28.4) are valid for the thermal energy group ($g = 2$). Here, $x_a \le x \le x_b$ for $q = b$ (baffle) and $x_b \le x \le x_c$ for $q = r$ (reflector). Moreover, we define

$J(x)$: neutron current;
$\phi(x)$: neutron scalar flux;
D: diffusion coefficient;
Σ_R: removal macroscopic cross section;
Σ_a: absorption macroscopic cross section;
$\Sigma_{s1\to 2}$: downscattering macroscopic cross section.

Furthermore, we consider the boundary conditions

$$\Phi_{1,r}(x_c) = 0, \tag{28.5}$$

$$\Phi_{2,r}(x_c) = 0. \tag{28.6}$$

At this point, we apply the Laplace transformation in space to (28.1) and (28.2) with $q = b$, making use of the fact that we can move the origin $x = 0$ to $x = x_a$. Therefore, by solving the resulting linear system and applying the inverse Laplace transform, we obtain

$$\begin{aligned}\phi_{1,b}(x) = {} & \frac{\phi_{1,b}(x_a)\sqrt{D_{1,b}\Sigma_{R1,b}} - J_{1,b}(x_a)}{2\sqrt{D_{1,b}\Sigma_{R1,b}}}\exp^{k_{1,b}x} \\ & + \frac{\phi_{1,b}(x_a)\sqrt{D_{1,b}\Sigma_{R1,b}} - J_{1,b}(x_a)}{2\sqrt{D_{1,b}\Sigma_{R1,b}}}\exp^{-k_{1,b}x},\end{aligned} \tag{28.7}$$

$$\begin{aligned}J_{1,b}(x) = {} & \frac{J_{1,b}(x_a) - \sqrt{D_{1,b}\Sigma_{R1,b}}\,\phi_{1,b}(x_a)}{2}\exp^{k_{1,b}x} \\ & + \frac{J_{1,b}(x_a) + \sqrt{D_{1,b}\Sigma_{R1,b}}\,\phi_{1,b}(x_a)}{2}\exp^{-k_{1,b}x}.\end{aligned} \tag{28.8}$$

Next, we substitute (28.7) in (28.3) with $q = b$, apply the Laplace transformation in space, solve the resulting linear system, and apply the inverse Laplace transformation to obtain

$$
\begin{aligned}
\phi_{2,b}(x) = & \frac{J_{2,b}(x_a)}{2\sqrt{D_{2,b}\Sigma_{a2,b}}}\left(-e^{k_{2,b}l_b} + e^{-k_{2,b}l_b}\right) \\
& + \frac{\phi_{2,b}(x_a)}{2}\left(e^{k_{2,b}l_b} + e^{-k_{2,b}l_b}\right) \\
& + \frac{\Sigma_{s1\to 2,b}}{2\sqrt{D_{1,b}\Sigma_{R1,b}\Sigma_{a2,b}}}\left(\phi_{1,b}(x_a)\sqrt{D_{1,b}\Sigma_{R1,b}} - J_{1,b}(x_a)\right)e^{k_{1,b}l_b} \\
& + \frac{\Sigma_{s1\to 2,b}}{2\sqrt{D_{1,b}\Sigma_{R1,b}\Sigma_{a2,b}}}\left(\phi_{1,b}(x_a)\sqrt{D_{1,b}\Sigma_{R1,b}} + J_{1,b}(x_a)\right)e^{-k_{1,b}l_b} \\
& + \frac{\Sigma_{s1\to 2,b}}{2\sqrt{D_{1,b}\Sigma_{R1,b}}(\sqrt{\Sigma_{a2,b}D_{2,b}})k_{1,b}}\left(2J_{1,b}(-e^{k_{2,b}l_b} + e^{-k_{2,b}l_b})\right), \qquad (28.9)
\end{aligned}
$$

$$
\begin{aligned}
J_{2,b}(x) = & \frac{J_{2,b}(x_a)}{2}\left(e^{k_{2,b}l_b} - e^{-k_{2,b}l_b}\right) \\
& + \frac{\phi_{2,b}(x_a)\sqrt{D_{2,b}\Sigma_{a2,b}}}{2}\left(e^{k_{2,b}l_b} + e^{-k_{2,b}l_b}\right) \\
& + \frac{\Sigma_{s1\to 2,b}\phi_{1,b}(x_a)}{2(k_{2,b}+k_{1,b})}\left(e^{k_{2,b}l_b} + e^{-k_{2,b}l_b}\right) \\
& + \frac{\Sigma_{s1\to 2,b}\phi_{1,b}(x_a)}{2(k_{2,b}-k_{1,b})}\left(e^{k_{2,b}l_b} + e^{-k_{2,b}l_b}\right) \\
& + \frac{\Sigma_{s1\to 2,b}J_{1,b}(x_a)}{2\sqrt{D_{1,b}\Sigma_{R1,b}}(k_{2,b}+k_{1,b})}\left(e^{k_{2,b}l_b} - e^{-k_{2,b}l_b}\right) \\
& + \frac{\Sigma_{s1\to 2,b}J_{1,b}(x_a)}{2\sqrt{D_{1,b}\Sigma_{R1,b}}(k_{2,b}-k_{1,b})}\left(e^{k_{2,b}l_b} - e^{-k_{2,b}l_b}\right), \qquad (28.10)
\end{aligned}
$$

where we have defined

$$
k_{1,q} = \sqrt{\frac{\Sigma_{R1,q}}{D_{1,q}}}, \quad k_{2,q} = \sqrt{\frac{\Sigma_{a2,q}}{D_{2,q}}},
$$
$$
q = b \text{ (baffle) or } r \text{ (reflector)}.
$$

To proceed further, we follow similar steps for the reflector region, i.e., $q = r$ ($x_b \le x \le x_c$), and use the boundary conditions (28.5–28.6). Moreover, we write the two-energy group albedo equation as

$$
\begin{pmatrix} J_1(x_a) \\ J_2(x_a) \end{pmatrix} = \begin{pmatrix} \alpha_{1,1} & 0 \\ -\alpha_{2,1} & \alpha_{2,2} \end{pmatrix} \begin{pmatrix} \phi_1(x_a) \\ \phi_2(x_a) \end{pmatrix},
$$

where the 2 x 2 albedo matrix is to substitute the baffle–reflector system ($x_a \le x \le x_c$) in Figure 28.1. At this point, we remark that the albedo entry α_{12} is set equal to zero, because we have neglected the upscattering events in this approach, as we see in equation (28.1), where $\Sigma_{s2\to 1,q} = 0$. By using the Laplace transform technique, as described previously, the entries of the albedo matrix, after tedious algebra, appear as

$$\alpha_{11} = \frac{\sqrt{D_{1,b}\Sigma_{R1,b}}\sqrt{D_{1,b}\Sigma_{R1,b}}\sinh(k_{1,b}l_b)}{\sqrt{D_{1,b}\Sigma_{R1,b}}\cosh(k_{1,b}l_b) + \sqrt{D_{1,R}\Sigma_{R1,R}}\coth(k_{1,R}l_R)\sinh(k_{1,b}l_b)} + \frac{\sqrt{D_{1,b}\Sigma_{R1,b}}\sqrt{D_{1,R}\Sigma_{R1,R}}\coth(k_{1,R}l_R)\cosh(k_{1,b}l_b)}{\sqrt{D_{1,b}\Sigma_{R1,b}}\cosh(k_{1,b}l_b) + \sqrt{D_{1,R}\Sigma_{R1,R}}\coth(k_{1,R}l_R)\sinh(k_{1,b}l_b)}$$

$$\begin{aligned}\alpha_{12} = {} & \frac{(\Sigma_{S1\to 2,b})^2\sqrt{D_{2,b}\Sigma_{a2,b}}}{(k_{2,b})^2 - (k_{1,b})^2\sqrt{D_{2,b}\Sigma_{a2,b}}\cosh(k_{2,b}l_b) + \sqrt{D_{2,R}\Sigma_{a2,R}}\sinh(k_{2,b}l_b)} \\ & - \frac{\left(\sqrt{D_{1,b}\Sigma_{R1,b}}\sinh(k_{1,b}l_b) + \sqrt{D_{1,R}\Sigma_{R1,R}}\coth(k_{1,R}l_R)\cosh(k_{1,b}l_b)\right)}{\left(\sqrt{D_{1,b}\Sigma_{R1,b}}\cosh(k_{1,b}l_b) + \sqrt{D_{1,R}\Sigma_{R1,R}}\coth(k_{1,R}l_R)\sinh(k_{1,b}l_b)\right)} \\ & \times \frac{(\Sigma_{S1\to 2,b})^2}{(k_{2,b})^2 - (k_{1,b})^2} \\ & \times \left(\sqrt{D_{2,b}\Sigma_{R1,b}} - \frac{\sqrt{D_{2,b}\Sigma_{a2,b}}}{\sqrt{D_{2,b}\Sigma_{a2,b}}\cosh(k_{2,b}l_b) + \sqrt{D_{2,r}\Sigma_{a2,r}}\sinh(k_{1,b}l_b)}\right) \\ & + \frac{(\Sigma_{S1\to 2,b})^2\sqrt{D_{1,b}\Sigma_{R1,b}}}{(k_{2,R})^2 - (k_{1,R})^2\sqrt{D_{1,b}\Sigma_{R1,b}}\cosh(k_{1,b}l_b) + \sqrt{D_{1,R}\Sigma_{R1,R}}\sinh(k_{1,b}l_b)} \\ & + \frac{\sqrt{D_{2,b}\Sigma_{a2,b}}D_{2,b}(k_{2,R} - k_{1,R})}{\sqrt{D_{2,b}\Sigma_{a2,b}}\cosh(k_{2,b}l_b) + \sqrt{D_{2,R}\Sigma_{a2,R}}\sinh(k_{2,b}l_b)},\end{aligned}$$

$$\alpha_{22} = \frac{\sqrt{D_{2,b}\Sigma_{a2,b}}\sqrt{D_{2,b}\Sigma_{a2,b}}\sinh(k_{2,b}l_b)}{\sqrt{D_{2,b}\Sigma_{a2,b}}\cosh(k_{2,b}l_b) + \sqrt{D_{2,R}\Sigma_{a2,R}}\coth(k_{2,R}l_R)\sinh(k_{2,b}l_b)} + \frac{\sqrt{D_{2,b}\Sigma_{a2,b}}\sqrt{D_{2,R}\Sigma_{a2,R}}\coth(k_{2,R}l_R)\cosh(k_{2,b}l_b)}{\sqrt{D_{2,b}\Sigma_{a2,b}}\cosh(k_{2,b}l_b) + \sqrt{D_{2,R}\Sigma_{a2,R}}\coth(k_{2,R}l_R)\sinh(k_{2,b}l_b)}.$$

28.3 Illustrative Example

In this section we examine the numerical results for one model problem. This two-energy group model problem consists of the heterogeneous slab illustrated in Figure 28.2. The material parameters for each zone are given in Table 28.1.

Moreover, vacuum boundary conditions apply on the outer boundary of the reflector region, i.e., at $x = 174.857$ cm and reflective boundary conditions apply at x = 0. We assume that this half domain generates 200 MW per cm^2 cross-sectional area and the energy release is equal to 200 MeV for each fission reaction. To solve this problem, the convergence criterion for the effective multiplication factor *(keff)* is

$$\left|\frac{k^{(l)} - k^{(l-1)}}{k^{(l)}}\right| \leq \epsilon_k \tag{28.11}$$

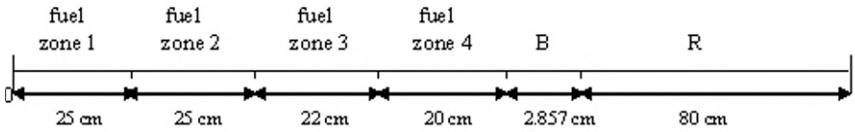


Fig. 28.2. Model problem.

Table 28.1. Material parameters for the Model Problem

Material zones	D_1	D_2	Σ_{R1}	Σ_{a2}	$\Sigma_{s1\to 2}$	$\nu\Sigma_{f1}$	$\nu\Sigma_{f2}$
fuel zone 1	1.438000	0.397600	0.029350	0.104900	0.015630	0.008350	0.155618
fuel zone 2	1.438000	0.382500	0.026600	0.078620	0.017380	0.008350	0.155618
fuel zone 3	1.409000	0.405100	0.026610	0.106300	0.015830	0.010500	0.222180
fuel zone 4	1.466000	0.385800	0.026150	0.111000	0.015590	0.008064	0.199381
baffle	1.049000	0.333493	0.004634	0.151881	0.001012	0	0
reflector	1.871400	0.283409	0.035411	0.031579	0.043400	0	0

and the convergence criterion for the group scalar flux is

$$\max_{\substack{J=1:J+1\\ g=1:2}} \left| \frac{\phi^{(l)}_{g,j-1/2} - \phi^{(l-1)}_{g,j-1/2}}{\phi^{(l)}_{g,j-1/2}} \right| \leq \epsilon_\phi, \tag{28.12}$$

where J is the total number of discretization cells in the spatial grid set up on the domain. In (28.11), we have defined $k^{(l)}$ as the lth estimate of the dominant eigenvalue k in the power iterative scheme. In (28.12), $\phi^{(l)}_{g,j-1/2}$ has been defined as the lth estimate of the group cell-edge scalar flux. For this model problem, we assumed $\epsilon_k = 10^{-5}$ in (28.11) and $\epsilon_\Phi = 10^{-4}$ in (28.12). The fast-group scalar flux and the thermal-group scalar flux displayed by Tables 28.2 and 28.3, respectively, show that the albedo boundary conditions, as described in this chapter, for one non-multiplying region and for two non-multiplying regions are very accurate at substituting the reflector region and the baffle–reflector system around the active domain, when compared with the results generated by the finite difference code explicitly.

In addition, Table 28.4 shows the power distribution per unit cross-sectional area and the effective multiplication factor as generated by the finite difference code explicitly and using the albedo boundary conditions. As we see, the results are very accurate with respect to the results generated explicitly. In addition, the efficiency of the computational finite difference code increased significantly by the use of the present albedo boundary conditions. That is, the CPU running time for convergence of the model problem decreased 39% by the use of the one-region albedo boundary condition and 59% by the use of the two-region albedo boundary condition. Moreover, the use of

Table 28.2. Neutron scalar flux for the fast energy group ($g = 1$).

($g = 1$)	$X = 0$ cm	$X = 25$ cm	$X = 50$ cm	$X = 72$ cm	$X = 92$ cm
Neutron scalar flux for the fast energy group Finite difference	1.043374x10^{14}	8.444018x10^{16}	1.635675x10^{18}	9.573301x10^{17}	1.750701x10^{16}
Neutron scalar flux for the fast energy group One-region albedo	1.043131x10^{14}	8.44212x10^{16}	1.635346x10^{18}	9.572398x10^{17}	1.764890x10^{16}
Relative deviation with respect to the explicit calculation (%)	0.023290	0.022430	0.020114	0.009432	0.810475
Neutron scalar flux for the fast energy group Two-region albedo	1.042678x10^{14}	8.438297x10^{16}	1.634609x10^{18}	9.570281x10^{17}	1.773688x10^{16}
Relative deviation with respect to the explicit calculation (%)	0.066707	0.067752	0.065172	0.031546	1.313017

the Chebyshev acceleration scheme for the convergence of the power iterative method [Fl72], decreased the number of iterations by 19%; that is, 113 power iterations for unaccelerated convergence of the model problem, and 92 power iterations with the Chebyshev acceleration scheme.

28.4 Concluding Remarks

We described in this chapter the use of the Laplace transform method in space for the calculation of the albedo boundary conditions in energy-dependent neutron diffusion eigenvalue problems in slab geometry. Besides generating very accurate results, the albedo boundary conditions, as described in this chapter, improved the efficiency of the fine-grid running code, as the execution time has shortened considerably. The extension of the present albedo boundary conditions for multigroup neutron diffusion eigenvalue problems for nuclear reactor global calculations with more than two energy groups, say with four energy groups, is straightforward except for the fact that the matrix algebra involved to obtain the 4 x 4 albedo matrix will be much more

Table 28.3. Neutron scalar flux for the thermal energy group ($g = 2$).

($g = 2$)	$X = 0$ cm	$X = 25$ cm	$X = 50$ cm	$X = 72$ cm	$X = 92$ cm
Neutron scalar flux for the thermal energy group Finite difference	4.771184x10^{12}	4.781230x10^{15}	8.733574x10^{16}	4.159283x10^{16}	4.252688x10^{14}
Neutron scalar flux for the thermal energy group One-region albedo	4.770100x10^{12}	4.780186x10^{15}	8.731865x10^{16}	4.158912x10^{16}	4.262688x10^{14}
Relative deviation with respect to the explicit calculation (%)	0.022720	0.021835	0.019568	0.008920	0.235145
Neutron scalar flux for the thermal energy group Two-region albedo	4.767740x10^{12}	4.777731x10^{15}	8.727408x10^{16}	4.157735x10^{16}	4.282688x10^{14}
Relative deviation with respect to the explicit calculation (%)	0.072183	0.073182	0.070601	0.037218	0.705436

tedious. Although the present albedo boundary conditions do not directly apply to multidimensional diffusion eigenvalue problems, we can use the idea in an approximate way by neglecting the transverse leakage terms in order to derive the albedo expressions. We expect that the efficiency shall be even more pronounced in fine-grid multigroup multidimensional methods, such as the conventional finite difference. In addition, it is well known that the convergence rate of the power method [Fl72] depends highly on the dominance ratio in the eigenvalue spectrum and may be very slow. As we can see in the previous section, the acceleration of the power method using a technique based on the two-parameter Chebyshev extrapolation of the fission source improved the convergence rate of the power method. In slab geometry, this may not represent a significant contribution; however, in multidimensional multigroup diffusion eigenvalue problems for nuclear reactor global calculations, the use of multigroup albedo boundary conditions together with the Chebyshev acceleration scheme might improve the efficiency of the computer code considerably.

Table 28.4. Power distribution (MW/cm^2) and effective multiplication factor (*keff*).

	$R1^a$	$R2^b$	$R3^c$	$R4^d$	$R5^e$	$R6^f$	*keff*
Power (Finite difference)	0.759797	41.299690	145.096800	12.84142	0	0	9.991050x10^{-1}
Power (one-region albedo)	0.759625	41.291030	145.071100	12.87562	0	0	9.991060x10^{-1}
Relative deviation with respect to the explicit calculation %	0.022519	0.020969	0.017712	0.266326	-	-	0.000100
Power (two-region albedo)	0.759283	41.272220	145.012400	12.95612	0	0	9.991080x10^{-1}
Relative deviation with respect to the explicit calculation %	0.067663	0.066514	0.058168	0.893203	-	-	0.000300

a = Fuel zone 1 in Fig. 28.2
b = Fuel zone 2 in Fig. 28.2
c = Fuel zone 3 in Fig. 28.2
d = Fuel zone 4 in Fig. 28.2
e = Baffle
f = Reflector

Acknowledgement. This work was sponsored by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq - Brazil) and is related to the research project of the National Institute of Science and Technology on Innovative Nuclear Reactors.

References

[AlOd86] Alcouffe, R.E., O'Dell, R.D.: Transport calculations for nuclear reactors, in *CRC Handbook of Nuclear Reactor Calculations. Vol. 1,* CRC Press, Boca Raton, FL (1986).

[Bu05] Burden, R.L.: *Numerical Analysis,* 8th ed., Thomson-Brooks/Cole, Belmont, CA (2005).

[Fl72] Fladmark, G.E.: Numerical reactor calculations. IAEA-SM-154/20, 497–507 (1972).

[Pa61] Pannekoek, A.: *A History of Astronomy.* Dover, New York (1961).

[Pe08] Petersen, C.Z.: Aplicação da Transformada de Laplace para Determinação de Condições de Contorno tipo Albedo para Cálculos Neutrônicos. M.Sc. dissertation, PROMEC/UFRGS, Brazil (2008) (Portuguese).

[WF07] World Federation of Engineering Organizations (WFEO). Nuclear Power Feasibility, Energy Committee (2007).

29

Solution of the Fokker–Planck Pencil Beam Equation for Electrons by the Laplace Transform Technique

B. Rodriguez and M.T. Vilhena

Universidade Federal do Rio Grande, RS, Brazil; barbara.orodriguez@gmail.com, vilhena@pq.cnpq.br

29.1 Introduction

While many medical physicists understand the basic principles underlying Monte Carlo codes such as EGS [Ka00], Geant [Wr01], and MCNP [Br93], there is less appreciation of the capabilities of deterministic methods which in principle can provide comparable accuracies to Monte Carlo. Only within the last years have serious studies been made on the appliance of deterministic calculations to medical physics applications. The most versatile and widely used deterministic methods are the P_N approximation [Da57]; [SeViPa00], the S_N method (discrete ordinates method) [ViBa95]; [ViSeBa95], and their variants [SeVi94]; [RoViVo06]. The method of discrete ordinates has been used successfully in neutral particle applications [D096]; [Da92] and gamma ray transport calculations for many years. The calculations for these two types of radiation are done very similarly, since they are both neutral particles. On the other hand, to our knowledge, the P_N approximation has not yet been applied in the solution of the charged particle pencil beam transport equation. Pencil beam equations are used to model, e.g., problems of collimated electron and photon particles penetrating piecewise homogeneous regions. The collisions between the beam particles and particles from beams with different directions cause deposit of some part of the energy carried by the beams at the collision sites. To obtain a desired "amounts of energy deposited at certain parts of the target region" (dose) is of crucial interest in radiative cancer therapy.

In this chapter, we present a closed-form solution for the two-dimensional Fokker–Planck pencil beam equation for electron transport [BoLa96]; [BoLa95] in a homogeneous rectangular domain. This solution can be considered an alternative approach for the Boltzmann transport equation for charged particles. The Fokker–Planck (FP) approximation represents the impact of soft reactions as continuously slowing down the electrons, while also continuously changing their direction; e.g., a monodirectional beam will be dispersed into a finite beam width. This approximation can be derived from a Taylor series

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_29,

expansion of the integrand in the scatter source term appearing in the Boltzmann equation, with the assumption that only small changes in energy and direction are significant. The main idea described in this chapter relies on applying the P_N approximation, in the angular variable, to the two-dimensional Fokker–Planck equation and then applying the Laplace transform in the spatial x-variable. As a result, a first order linear differential equation in the spatial y-variable is attained, for which the solution is straightforward. The P_N approximation consists in expanding the angular variable of the angular flux in terms of the Legendre polynomials. In Section 29.2 we describe in detail the two-dimensional FP pencil beam equation solution. We conclude the chapter with Section 29.3, where we give some illustrative examples.

29.2 Mathematical Formulation

In order to determine the angular flux of electrons in a rectangular domain, let us consider the following two-dimensional, time-independent electron transport equation

$$\mu\frac{\partial\psi(x,y,\overline{\Omega},E)}{\partial x}+\eta\frac{\partial\psi(x,y,\overline{\Omega},E)}{\partial y}+\sigma_t(E)\psi(x,y,\overline{\Omega},E)$$
$$=\int_{4\pi}d\overline{\Omega}'\sigma_s(E'\to E,\overline{\Omega}'\cdot\overline{\Omega})\psi(x,y,\overline{\Omega}',E'), \tag{29.1}$$

in a rectangle $0\le x\le a$ and $0\le y\le b$, subject to vacuum boundary conditions. Here the angular flux, denoted as $\psi(x,y,E,\overline{\Omega})$, represents the flux of particles at position (x,y), with energy E travelling in direction $\overline{\Omega}=(\mu,\eta)$. The quantity σ_s in (29.1) is the differential scattering cross section and is written as

$$\sigma_s(E,\mu_0)=\sum_{l=0}^{L\le N}\frac{2l+1}{2}\sigma_{sl}(E)P_l(\mu_0),\ N \text{ odd},$$

where $\mu_0=\overline{\Omega}'\cdot\overline{\Omega}$ is the cosine of the scattering angle and σ_{sl} are the Legendre moments of the scattering cross section. In this chapter we focus on screened Rutherford scattering, which can be written as

$$\sigma_s(E,\mu_0)=\frac{\sigma_t(E)\eta^*(\eta^*+1)}{\pi(1+2\eta^*-\mu_0^2)},$$

where $\eta^*>0$ is a typically small constant called the screening parameter. Screened Rutherford scattering is one of the simplest models of elastic scattering of electrons from nuclei taking into account the screening of the nuclei by atomic electrons. It is obtained from the Schrödinger equation in the first Born approximation, using an exponential factor in the potential to model the screening effect [Re85]. An approximate formula for the screening parameter is written as

$$\eta^* = \frac{h^2 Z^{\frac{2}{3}}}{4(a_H)^2(m_e v)^2}, \tag{29.2}$$

where Z denotes the atomic number of the nucleus, mv is the (relativist) momentum of the electron that is being scattered, and C is a constant. In terms of the Planck constant h and the Bohr radius a_H,

$$C = \frac{h^2}{4{a_H}^2}.$$

The FP equation can take many different forms depending on the order of approximation employed and the characteristics of the scattering cross section. In all cases, the integral Boltzmann scattering operator is approximated with a differential operator obtained using Taylor expansion techniques. This equation represents an approximation to the Boltzmann transport equation that is valid whenever small-angle scattering is predominant [Ta67].

We now assume that the scattering process is sufficiently peaked in the forward direction so that the FP scattering description [Po83] is appropriate. Thus, the FP approximation [BoLa96] to transport problem (29.1) is given by

$$\mu \frac{\partial \psi^{FP}(x, y, \overline{\Omega}, E)}{\partial x} + \eta \frac{\partial \psi^{FP}(x, y, \overline{\Omega}, E)}{\partial y} = \frac{\sigma_{tr}}{2} \frac{\partial}{\partial \mu} \Big[(1 - \mu^2) \frac{\partial}{\partial \mu} \Big] \psi^{FP}(x, y, \overline{\Omega}, E), \tag{29.3}$$

where $\psi^{FP}(x, y, \overline{\Omega}, E)$ represents the FP angular flux of particles at position (x, y), with energy E travelling in direction $\overline{\Omega} = (\mu, \eta)$, and the coefficient σ_{tr} is called the transport cross section and is defined as

$$\sigma_{tr} = 2\pi \int_{-1}^{1} \int_{0}^{1} \sigma_s(E, \mu_0)(1 - \mu_0) d\mu_0 d\eta. \tag{29.4}$$

The differential term on the right-hand side of (29.3) can be replaced by

$$\frac{\partial}{\partial \mu} \Big[(1 - \mu^2) \frac{\partial}{\partial \mu} \Big] \psi^{FP}(x, y, \overline{\Omega}, E) = \Big[(1 - \mu^2) \frac{\partial^2}{\partial \mu^2} - 2\mu \frac{\partial}{\partial \mu} \Big] \psi^{FP}(x, y, \overline{\Omega}, E). \tag{29.5}$$

Substituting (29.5) into (29.3), we obtain

$$\mu \frac{\partial \psi^{FP}(x, y, \overline{\Omega}, E)}{\partial x} + \eta \frac{\partial \psi^{FP}(x, y, \overline{\Omega}, E)}{\partial y} = \frac{\sigma_{tr}}{2} \Big[(1 - \mu^2) \frac{\partial^2}{\partial \mu^2} - 2\mu \frac{\partial}{\partial \mu} \Big] \psi^{FP}(x, y, \overline{\Omega}, E). \tag{29.6}$$

Applying in (29.6) the operator

$$\int_{-1}^{1}\int_{0}^{1}(\ )P_m(\mu)d\mu d\eta, \text{ com } m=0,\ldots,N,$$

and using the recursion formula [Kr66]

$$\mu P_n(\mu)=\frac{n+1}{2n+1}P_{n+1}(\mu)+\frac{n}{2n+1}P_{n-1}(\mu),$$

as well as the Legendre polynomial properties, we arrive at the following P_N equations:

$$\frac{n+1}{2n+1}\frac{\partial}{\partial x}\psi_{n+1}^{FP}(x,y,E)+\frac{n}{2n+1}\frac{\partial}{\partial x}\psi_{n-1}^{FP}(x,y,E)$$
$$+\frac{2n+1}{2}\frac{\partial}{\partial y}\psi_{n}^{FP}(x,y,E)T_n=\frac{\sigma_{tr}}{2}[-n(n+1)]\psi_{n}^{FP}(x,y,E), \tag{29.7}$$

with the angular flux moments in discrete ordinates approximated by a quadrature formula as follows:

$$\psi^{FP}(x,y,\overline{\Omega},E)=\sum_{l=0}^{L}\frac{2n+1}{2}\psi_{n}^{FP}(x,y,E)P_n(\mu),$$

for $n=0,...,N$, with $\psi_{N+1}^{FP}(x,y,E)=0$ in the P_N approximation and T_n represented by an integral term, which can be analytically solved, written as

$$T_n=\int_{-1}^{1}\sqrt{(1-\mu^2)}P_n(\mu)P_{n+1}(\mu)d\mu. \tag{29.8}$$

Applying the Laplace transformation in (29.7) in the spatial variable x, we obtain the linear algebraic system

$$\frac{n+1}{2n+1}\Big[s\overline{\psi_{n+1}^{FP}}(s,y,E)-\overline{\psi_{n+1}^{FP}}(0,y,E)\Big]$$
$$+\frac{n}{2n+1}\Big[\overline{\psi_{n-1}^{FP}}(s,y,E)-\overline{\psi_{n-1}^{FP}}(0,y,E)\Big]+\frac{2n+1}{2}\frac{\partial}{\partial y}\overline{\psi_{n}^{FP}}(s,y,E)T_n$$
$$=\frac{\sigma_{tr}}{2}\Big[-n(n+1)\Big]\overline{\psi_{n}^{FP}}(s,y,E) \tag{29.9}$$

for $n=0,\ldots,N$, and $\overline{\psi_{n-1}^{FP}}(s,y,E)$, $\overline{\psi_{n}^{FP}}(s,y,E)$, and $\overline{\psi_{n+1}^{FP}}(s,y,E)$ are the transformed angular fluxes in the spatial x variable. The linear algebraic system (29.9) can be recast in the matrix form

$$A_n\overline{\psi_{n}^{FP}}'(s,y,E)+B_n(s)\overline{\psi_{n}^{FP}}(s,y,E)-C_n\psi_{n}^{FP}(0,y,E)=0. \tag{29.10}$$

Here, $\overline{\psi_{n}^{FP'}}(s,y,E)$ is the N components vector of the derivative of the angular flux Laplace-transformed in the x variable with respect to y and is written as

$$\overline{\psi_n^{FP}}'(s,y,E) = \left[\overline{\psi_0^{FP}}'(s,y,E) \quad \overline{\psi_1^{FP}}'(s,y,E) \quad ... \quad \overline{\psi_N^{FP}}'(s,y,E)\right]^T.$$

Here, the column vector $\overline{\psi_n^{FP}}(s,y,E)$ is the N components of the angular flux Laplace-transformed vector in the x-variable and $\psi_n^{FP}(0,y,E)$ is the N components of the angular flux vector in the x-variable at $x=0$. They have the form

$$\overline{\psi_n^{FP}}(s,y,E) = \left[\overline{\psi_0^{FP}}(s,y,E) \ \overline{\psi_1^{FP}}(s,y,E) \ \ldots \ \overline{\psi_N^{FP}}(s,y,E)\right]^T,$$
$$\psi_n^{FP}(0,y,E) = \left[\psi_0^{FP}(0,y,E) \ \psi_1^{FP}(0,y,E) \ ... \ \psi_N^{FP}(0,y,E)\right]^T.$$

On the other hand, the components of matrices A_n, $B_n(s)$, and C_n are given, respectively, by

$$A_n = \begin{bmatrix} 1T_0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 9T_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 25T_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & (2N+1)^2T_N \end{bmatrix},$$

$$B_n(s) = \begin{bmatrix} 0 & 2s & 0 & 0 & \cdots & 0 \\ 2s & 6\sigma_{tr} & 4s & 0 & \cdots & 0 \\ 0 & 4s & 30\sigma_{tr} & 6s & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2Ns & N(N+1)(2N+1)\sigma_{tr} \end{bmatrix},$$

$$C_n = \begin{bmatrix} 0 & 2 & 0 & 0 & \cdots & 0 \\ 2 & 0 & 4 & 0 & \cdots & 0 \\ 0 & 4 & 0 & 6 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2N-2 & 0 & 2N \\ 0 & 0 & 0 & \cdots & 2N & 0 \end{bmatrix},$$

where σ_{tr} and T_n are defined by (29.4) and (29.8), respectively.

The solution of (29.10) is

$$\overline{\psi_n^{FP}}(s,y,E) = c1(s) \cdot e^{-\left[B_n(s)\cdot A_n^{-1}\right]} + C_n \cdot [B_n(s)]^{-1} \cdot \psi_n^{FP}(0,y,E),$$

where $c_1(s)$ is an arbitrary constant. In this problem we determine the $c_1(s)$ value by applying the boundary and interface conditions. Due to the linear character of the inverse Laplace transform operator, taking the Laplace inversion of the above ansatz, we get

$$\psi_n^{FP}(x,y,E) = \mathcal{L}^{-1}\left\{c1(s)\cdot e^{-\left[B_n(s)\cdot A_n^{-1}\right]}\right\} + C_n\cdot\mathcal{L}^{-1}\left\{[B_n(s)]^{-1}\right\}\cdot\psi_n^{FP}(0,y,E). \quad (29.11)$$

Once we have obtained the inverse matrices A_n^{-1}, $B_n^{-1}(s)$, and C_n^{-1}, we calculate the inverse of the first term in (29.11) by using the Laplace convolution property. Here, it is important to mention that the inverse matrix $B_n^{-1}(s)$ was not obtained analytically, due to the existence of the s parameter, a non-numeric parameter. Therefore, we opt to calculate the inverse Laplace transform numerically—here we apply the Gauss quadrature inversion method [StSe86]; [DaMa79].

29.3 Illustrative Examples

In order to illustrate the aptness of the discussed methodology to solve the two-dimensional FP pencil beam transport equation, in what follows we present numerical simulation examples for the absorbed energy in rectangular domains with different dimensions and compositions. The illustrative examples are presented under absorbed energy form, i.e., the deposited energy in several points of interest.

We considered a homogeneous rectangular domain composed of water, tissue, or bone. We also assume a monoenergetic (E = 1.25 MeV) and monodirectional photon beam incoming on the edge of a rectangle. The incoming photons will be tracked until their entire energy is deposited and/or they leave the domain of interest. In this study, the energy deposited by the secondary electrons, generated by the Compton effect, will be considered. The remaining effects will not be taken into account. The numerical results encountered for absorbed energy are compared with the ones obtained by the program Geant4 v8, using the Monte Carlo technique for low energy data [Ho07]; [Ro07]. The package includes detailed simulations of the interactions of particles with energies from about 250 eV to 250 GeV.

Geant4 [Ag03] is a toolkit for simulating the passage of particles through matter. It includes a complete range of functionality including tracking, geometry, physics models, and hits. The physics processes offered cover a comprehensive range, including electromagnetic, hadronic, and optical processes, a large set of long-lived particles, materials and elements, over a wide energy range starting, in some cases, from 250 eV and extending in others to the TeV energy range. It has been designed and constructed to expose the physics models utilized, to handle complex geometries, and to enable its easy adaptation for optimal use in different sets of applications. It has been used in applications in particle physics, nuclear physics, accelerator design, space engineering, and medical physics.

In what follows, we present numerical results for three problems.

Problem 1. Let us consider a homogeneous rectangular domain, constituted by liquid water ($Z/A = 0.55508$, $\rho = 1$ g/cm^3) and with the vacuum boundary condition.

In Table 29.1 we present the P_N approximation numerical simulations for the absorbed energy and comparisons with the Geant4 results [Wr01]. Bearing

Table 29.1. Absorbed energy in a rectangular domain composed by water.

	Water, liquid		
Domain dimensions	P_9	Geant4	absolute relative error
10 cm x 10 cm	0.02149855	0.02289237	6.0885%
10 cm x 20 cm	0.02100552	0.02240557	6.2487%
20 cm x 10 cm	0.01845323	0.01971250	6.3882%
20 cm x 20 cm	0.03378688	0.03609384	6.3916%
30 cm x 40 cm	0.04580598	0.04893386	6.3921%

in mind that the Geant4 program applies the Monte Carlo technique, giving a closer look at the results in Table 29.1, we promptly realize a good coincidence. In Table 29.2 we display the numerical convergence of the P_N approximation results in a rectangular domain composed of water for increasing N. In fact, observing the results for $N = 7$ and $N = 9$ we notice a coincidence of four significant digits.

Problem 2. To check the influence of the material density in the absorbed energy calculation, let us consider a rectangular domain composed of cortical bone ($Z/A = 0.51478$, $\rho = 1.92$ g/cm^3) and with the vacuum boundary condition.

Problem 3. Let us consider a homogeneous rectangular geometry constituted by soft tissue ($Z/A = 0.54996$, $\rho = 1.06$ g/cm^3) and with the vacuum boundary condition.

In Tables 29.3 and 29.4, we present the P_N approximation numerical simulations for the absorbed energy in a rectangle composed, respectively, of cortical bone and soft tissue, and comparisons with the Geant4 program results, where the maximum discrepancy found is lower than 9%. From the analysis of the results encountered for the above problem, we promptly realize a good agreement between the proposed methodology and the Monte Carlo technique results. Our numerical results demonstrate that, for higher density materials,

Table 29.2. P_N numerical convergence for Problem 1.

N	20 cm x 20 cm
1	0.02590432
3	0.03199219
5	0.03252043
7	0.03370622
9	0.03378688

Table 29.3. Absorbed energy in a rectangular domain composed by cortical bone [IC89].

	Bone, cortical (ICRU44)		
Domain dimensions	P_9	Geant4	absolute relative error
20 cm x 10 cm	0.83789957	0.91244397	8.1697%
20 cm x 20 cm	0.79284239	0.86380422	8.2150%
30 cm x 40 cm	0.89218297	0.97249263	8.2581%

other effects must be taken into account, because when the density increases, the number of interactions increases as well as the possibility of other processes involving the production of secondary electrons. We must also mention that we have done all the calculations using an AMD Athlon 1700 (1.4 GHz) microcomputer. Furthermore, the maximum computational time observed to generate all the results in each table was 30 minutes while the computational time to generate the Geant4 results was approximately one day.

In this chapter we obtained a closed-form solution for the Fokker–Planck pencil beam equation for rectangular geometries. This procedure allows us to calculate the energy deposited by secondary electrons generated by the Compton effect. We must recall that, to our knowledge, the P_N approximation of the two-dimensional Fokker–Planck equation has not been analytically solved yet. We must emphasize that the P_N solution of the Fokker–Planck pencil beam equation reported keeps the analytical feature, in the sense that no approximation is made along its derivation from the P_N equations, ex-

Table 29.4. Absorbed energy in a rectangular domain composed by soft tissue [IC89].

	Tissue, soft (ICRU44)		
Domain dimensions	P_9	Geant4	absolute relative error
20 cm x 10 cm	0.02288590	0.02440210	6.2134%
20 cm x 20 cm	0.03317010	0.03542490	6.3650%
30 cm x 40 cm	0.04951665	0.05288919	6.3766%

cept for the round-off error. Bearing in mind, besides the analytical feature of solution, the good agreement between the results attained by the proposed methodology with the ones of Geant4 with a small computational effort, we are confident in stressing that this technique is quite robust and promising, either under a mathematical or a computational point of view, to solve the two-dimensional Fokker–Planck pencil beam equation.

References

[Ag03] Agostinelli, S., et al.: Geant4—a simulation toolkit. *Nuclear Instruments Methods Phys. Research A,* **506**, 250–303 (2003).

[BoLa95] Börges, C., Larsen, E.W.: The transversely integrated scalar flux of a narrowly focused particle beam. *SIAM J. Appl. Math.,* **55**, 1–22 (1995).

[BoLa96] Börges, C., Larsen, E.W.: On the accuracy of the Fokker–Planck and Fermi pencil beam equations for charged particle transport. *Medical Phys.,* **23**, 1749–1759 (1996).

[Br93] Briesmeister, J.F.: MCNP—a general Monte Carlo N-particle transport code. Los Alamos National Laboratory Report LA-12625-M (1993).

[Da92] DANTSYS 3.0: *One-, Two-, and Three-Dimensional Multigroup, Discrete Ordinates Transport Code System,* RSICC Computer Code Collection CCC-547, Oak Ridge National Laboratory (1992).

[DaMa79] Davies, B., Martin, B.: Numerical inversion of the Laplace transform: a survey and comparison of methods. *J. Comput. Phys.,* **33**, 1–32 (1979).

[Da57] Davison, B.: *Neutron Transport Theory,* Oxford University Press, London (1957).

[D096] DOORS 3.1: *One-, Two-, and Three-Dimensional Discrete Ordinates Neutron/Photon Transport Code System,* RSICC Code Package CCC-650, Oak Ridge National Laboratory (1996).

[Ho07] Hoff, G.: Personal communication. Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil (2007).

[IC89] *International Commission on Radiation Units and Measurements, Tissue Substitutes in Radiation Dosimetry and Measurement,* ICRU 44, Bethesda, Maryland (1989).

[Ka00] Kawrakow, I.: Accurate condensed history Monte Carlo simulation of electron transport. I: EGSnrc, the new EGS4 version. *Medical Phys.,* **27**, 485–498 (2000).

[Kr66] Kreider, D.L.: *An Introduction to Linear Analysis,* Addison-Wesley, Reading, MA (1966).

[LMFB97] Larsen, E.W., Miften, M.M., Fraass, B.A., Brinvis, I.D.: Electron dose calculations using the method of moments. *Medical Phys.,* **24**, 111–125 (1997).

[Po83] Pomraning, G.C.: Flux-limited diffusion and Fokker–Planck equations. *Nuclear Sci. Engrg.,* **85**, 116–126 (1983).

[Re85] Reimer, L.: *Scanning Electron Microscopy,* Springer, Berlin (1985).

[Ro07] Rodriguez, B.D.A.: Methodology for obtaining a solution for the Boltzmann transport equation considering Compton scattering simulated by Klein–Nishina. Doctoral dissertation, Universidade Federal do Rio Grande do Sul (2007).

[RoViVo06] Rodriguez, B.D.A., Vilhena, M.T., Borges, V.: The determination of the exposure buildup factor formulation in a slab using the LTS_N method. *Kerntechnik,* **71**, 182–184 (2006).

[SeVi94] Segatto, C.F., Vilhena, M.T.: Extension of the LTS_N formulation for discrete ordinates problems without azimuthal symmetry. *Ann. Nuclear Energy,* **21**, 701–710 (1994).

[SeViPa00] Segatto, C.F., Vilhena, M.T., Pazos, R.P.: On the convergence of the spherical harmonics approximations. *Nuclear Sci. Engrg,.* **134**, 114–119 (2000).

[StSe86] Stroud, A.H., Secrest, E.: *Gaussian Quadrature Formulas,* Prentice-Hall, Englewood Cliffs, NJ (1986).

[Ta67] Tannenbaum, B.S.: *Plasma Physics,* McGraw-Hill, New York (1967).

[ViBa95] Vilhena, M.T., Barichello, L.B.: An analytical solution for the multigroup slab geometry discrete ordinates problems. *Transport Theory Statistical Phys.,* **24**, 1029–1037 (1995).

[ViSeBa95] Vilhena, M.T., Segatto, C.F., Barichello, L.B.: A particular solution for the Sn radiative transfer problems. *J. Quant. Spectrosc. Radiat. Transfer,* **53**, 467–469 (1995).

[Wr01] Wright, D.H.: *Physics Reference Manual,* `http://cern.ch/geant4` (2001). See User Documents at the Geant4 Web page `http://cern.ch/geant4` under Documentation.

30

Nonlinear Functional Parabolic Equations

L. Simon

L. Eötvös University of Budapest, Hungary; simonl@ludens.elte.hu

30.1 Introduction

This work was motivated by works where nonlinear parabolic functional differential equations were considered which arise in certain applications. (See the references in [SiJa08].) In [SiJa08], existence theorems and some qualitative properties were proved on solutions to initial value problems for the functional equations (connected with the above applications)

$$D_t u - \sum_{i=1}^{n} D_i[a_i(t,x,u,Du;u)] + a_0(t,x,u,Du;u) = f. \tag{30.1}$$

The aim of this chapter is to formulate existence theorems if certain modified (in some sense more general) assumptions are fulfilled and to show several examples satisfying these conditions such that the assumptions of [SiJa08] are not fulfilled. Some qualitative properties of the solutions are proved in [Si09].

30.2 Existence of Solutions

Denote by $\Omega \subset \mathbb{R}^n$ a bounded domain having the uniform C^1 regularity property (see [Ad75]), $Q_T = (0,T) \times \Omega$, and let $p \geq 2$ be a real number. Let $V \subset W^{1,p}(\Omega)$ be a closed linear subspace of the usual Sobolev space $W^{1,p}(\Omega)$ (of real-valued functions). Denote by $L^p(0,T;V)$ the Banach space of the set of measurable functions $u : (0,T) \to V$ with the norm

$$\| u \|^p_{L^p(0,T;V)} = \int_0^T \| u(t) \|^p_V \, dt.$$

The dual space of $L^p(0,T;V)$ is $L^q(0,T;V^\star)$ where $1/p + 1/q = 1$ and $V^\star$ is the dual space of V (see, e.g., [Ze90]).

About the functions a_i, we make the following assumptions:

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_30,

(A_1) The functions $a_i : Q_T \times \mathbb{R}^{n+1} \times L^p(0,T;V) \to \mathbb{R}$ satisfy the Carathéodory conditions for arbitrary fixed $u \in L^p(0,T;V)$ $(i = 0,1,...,n)$.

(A_2) There exist bounded (nonlinear) operators $g_1 : L^p(0,T;V) \to \mathbb{R}^+$ and $k_1 : L^p(0,T;V) \to L^q(\Omega)$ such that

$$|a_i(t,x,\zeta_0,\zeta;u)| \le g_1(u)[|\zeta_0|^{p-1} + |\zeta|^{p-1}] + [k_1(u)](x)$$

for a.a. $(t,x) \in Q_T$, each $(\zeta_0,\zeta) \in \mathbb{R}^{n+1}$, and $u \in L^p(0,T;V)$.

(A_3) There holds the inequality

$$\sum_{i=1}^{n}[a_i(t,x,\zeta_0,\zeta;u) - a_i(t,x,\zeta_0,\zeta^\star;u)](\zeta_i - \zeta_i^\star) \ge [g_2(u)](t)|\zeta - \zeta^\star|^p, \quad (30.2)$$

where

$$[g_2(u)](t) \ge c^\star \left[1 + \| u \|_{L^p(0,t;V)}\right]^{-\sigma^\star}, \quad (30.3)$$

$c^\star$ is some positive constant, and $0 \le \sigma^\star < p-1$.

(A_4) There holds the inequality

$$\sum_{i=0}^{n} a_i(t,x,\zeta_0,\zeta;u)\zeta_i \ge [g_2(u)](t)[|\zeta_0|^p + |\zeta|^p] - [k_2(u)](t,x),$$

where $k_2(u) \in L^1(Q_T)$ satisfies (for some positive $\sigma < p - \sigma^\star$)

$$\| k_2(u) \|_{L^1(Q_t)} \le \text{const}\left[1 + \| u \|_{L^p(0,t;V)}\right]^{\sigma}.$$

(A_5) There exists $\delta > 0$ such that if $(u_k) \to u$ weakly in $L^p(0,T;V)$, strongly in $L^p(0,T;W^{1-\delta,p}(\Omega))$, $(\zeta_0^k) \to \zeta_0$ in $\mathbb{R}$, and $(\zeta^k) \to \zeta$ in $\mathbb{R}^n$, then for a.a. $(t,x) \in Q_T$,

$$\lim_{k\to\infty} a_i(t,x,\zeta_0^k,\zeta^k;u_k) = a_i(t,x,\zeta_0,\zeta;u).$$

Remark 1. Assumption (A_5) is weaker than the corresponding assumption in [SiJa08], thus equation (30.1) may contain more general "nonlocal" terms in this chapter. (See the examples in Section 30.3.)

Definition 1. *Assuming that properties (A_1)–(A_5) hold, we define an operator $A : L^p(0,T;V) \to L^q(0,T;V^\star)$ by*

$$[A(u),v] = \int_{Q_T}\left\{\sum_{i=1}^{n} a_i(t,x,u,Du;u)D_i v + a_0(t,x,u,Du;u)v\right\}dtdx, \quad (30.4)$$

where the brackets $[\cdot,\cdot]$ mean the dualities in $L^q(0,T;V^\star)$ and $L^p(0,T;V)$.

Theorem 1. *Assume (A_1)–(A_5). Then for any $f \in L^q(0,T;V^\star)$ and $u_0 \in L^2(\Omega)$ there exists $u \in L^p(0,T;V)$ such that $D_t u \in L^q(0,T;V^\star)$,*

$$D_t u + A(u) = f, \quad u(0) = u_0. \quad (30.5)$$

Proof. Clearly, (A_1), (A_2) imply that A is bounded, (i.e., it maps bounded sets of $L^p(0,T;V)$ into bounded sets of $L^q(0,T;V^\star)$) and demicontinuous:

$$(u_j) \to u \text{ in } L^p(0,T;V) \text{ implies } (A(u_j)) \to A(u) \text{ weakly in } L^q(0,T;V^\star).$$

(See, e.g., [Si08], [Ze90].) Further, (A_4) implies that A is coercive:

$$[A(u_j), u_j] \to +\infty \text{ if } \| u_j \|_{L^p(0,T;V)} \to \infty$$

because, by (A_4) (for $\| u_j \| > 1$),

$$\begin{aligned}[A(u_j), u_j] &\geq \frac{c^\star}{[1+ \| u_j \|_{L^p(0,T;V)}]^{\sigma^\star}} \| u_j \|^p_{L^p(0,T;V)} \\ &\qquad - \text{const}[1+ \| u_j \|_{L^p(0,T;V)}]^\sigma \\ &\geq (c^\star/2) \| u_j \|_{L^p(0,T;V)}]^{p-\sigma^\star} - \text{const}[1+ \| u_j \|_{L^p(0,T;V)}]^\sigma \to \infty\end{aligned}$$

as $\| u_j \|_{L^p(0,T;V)} \to \infty$ since $p - \sigma^\star > \sigma$.

Now we show that A is pseudomonotone with respect to

$$D(L) = \{u \in L^p(0,T;V) : D_t u \in L^q(0,T;V^\star), u(0) = 0\}$$

in the sense of [BeMu92]: defining the operator L by $Lu = D_t u$ for $u \in D(L)$, if

$$u_j, u \in D(L), \quad (u_j) \to u \text{ weakly in } L^p(0,T;V), \tag{30.6}$$

$$(Lu_j) \to Lu \text{ weakly in } L^q(0,T;V^\star), \tag{30.7}$$

$$\limsup_{j\to\infty} [A(u_j), u_j - u] \leq 0, \tag{30.8}$$

we then have

$$\lim_{j\to\infty} [A(u_j), u_j - u] = 0 \text{ and } A(u_j) \to A(u) \text{ weakly in } L^q(0,T;V^\star) \tag{30.9}$$

because, by the well-known compact embedding theorem (see, e.g., [Li69], [Si08]) (30.6) and (30.7) imply that there is a subsequence $(\tilde{u}_j)$ of (u_j) such that

$$(\tilde{u}_j) \to u \text{ in } L^p(0,T;W^{1-\delta,p}(\Omega)) \text{ and a.e. in } Q_T. \tag{30.10}$$

Since $(D_i u_j)$ is bounded in $L^p(Q_T)$, we may assume on the subsequence $(\tilde{u}_j)$

$$(D_i \tilde{u}_j) \to D_i u \text{ weakly in } L^p(Q_T), \quad i = 1, ..., n. \tag{30.11}$$

Next,

$$\begin{aligned}[A(\tilde{u}_j), \tilde{u}_j - u] &= \int_{Q_T} a_0(t,x,\tilde{u}_j, D\tilde{u}_j; \tilde{u}_j)(\tilde{u}_j - u)dtdx \\ &\quad + \sum_{i=1}^n \int_{Q_T} [a_i(t,x,\tilde{u}_j, D\tilde{u}_j; \tilde{u}_j) - a_i(t,x,\tilde{u}_j, Du; \tilde{u}_j)]dtdx \\ &\qquad + \sum_{i=1}^n \int_{Q_T} a_i(t,x,\tilde{u}_j, Du; \tilde{u}_j)(D_i\tilde{u}_j - D_i u)dtdx. \end{aligned} \tag{30.12}$$

The first term on the right-hand side of (30.12) converges to 0 since the $L^p(Q_T)$ norm of $\tilde{u}_j - u$ tends to 0 by (30.10) and its multiplier is bounded in $L^q(Q_T)$ by (A_2). Further, the third term on the right-hand side tends to 0, too, by (30.11), because, by (30.6), (30.10), (A_1), (A_2), (A_5), and Vitali's theorem,

$$a_i(t, x, \tilde{u}_j, Du; \tilde{u}_j) \to a_i(t, x, u, Du; u) \text{ in } L^q(Q_T).$$

Consequently, from (30.8) and (30.12) we obtain

$$\limsup_{j\to\infty} \sum_{i=1}^{n} \int_{Q_T} [a_i(t, x, \tilde{u}_j, D\tilde{u}_j; \tilde{u}_j) - a_i(t, x, \tilde{u}_j, Du; \tilde{u}_j) \times (D_i\tilde{u}_j - D_i u)]dtdx \le 0. \quad (30.13)$$

Since $(\tilde{u}_j)$ is bounded in $L^p(0, T; V)$, (A_3) and (30.13) imply

$$\lim_{j\to\infty} \int_{Q_T} |D\tilde{u}_j - Du|^p dtdx = 0 \text{ and } (D\tilde{u}_j) \to Du \quad \text{a.e. in } Q_T, \quad (30.14)$$

for a subsequence (denoted again, for simplicity, by $(\tilde{u}_j)$). Therefore, by (A_1), (A_2), (A_5), (30.6), (30.10), (30.14), and Vitali's theorem,

$$a_i(t, x, \tilde{u}_j, D\tilde{u}_j; \tilde{u}_j) \to a_i(t, x, u, Du; u) \quad \text{in } L^q(Q_T), \quad i = 0, 1, \ldots, n,$$

which implies (30.9) for the subsequence $(\tilde{u}_j)$ by (30.10) and (30.14). Consequently, (30.9) holds for (u_j), too (see, e.g., [BeMu92], [Ze90]).

Since A is bounded, demicontinuous, coercive, and pseudomonotone with respect to $D(L)$, we obtain the assertion. (See, e.g., [BeMu92] and [Si08].)

We now formulate an existence theorem in $(0, \infty)$. Denote by $L^p_{loc}(0, \infty; V)$ the set of functions $u : (0, \infty) \to V$ such that for each fixed finite $T > 0$, $u|_{(0,T)} \in L^p(0, T; V)$ and let $Q_\infty = (0, \infty) \times \Omega$, $L^\alpha_{loc}(Q_\infty)$ be the set of functions $u : Q_\infty \to \mathbb{R}$ such that $u|_{Q_T} \in L^\alpha(Q_T)$ for any finite T.

Theorem 2. *Assume that the functions*

$$a_i : Q_\infty \times \mathbb{R}^{n+1} \times L^p_{loc}(0, \infty; V) \to \mathbb{R}$$

satisfy (A_1)–(A_5) for any finite T and that the $a_i(t, x, \zeta_0, \zeta; u)|_{Q_T}$ depend only on $u|_{(0,T)}$ (Volterra property). Then for any $f \in L^q_{loc}(0, \infty; V^\star)$, there exists $u \in L^p_{loc}(0, \infty; V)$ which is a solution of (30.5) for any finite T.

Theorem 2 follows from Theorem 1 if we use a diagonal process and the Volterra property (see, e.g., [Si00]).

30.3 Examples

In [SiJa08] examples of the following type were considered:

$$a_i(t,x,\zeta_0,\zeta;u) = b([H(u)](t,x))\zeta_i|\zeta|^{p-2}, \quad i=1,\dots,n,$$
$$a_0(t,x,\zeta_0,\zeta;u) = b_0([H_0(u)](t,x))\zeta_0|\zeta_0|^{p-2} + \hat{b}_0([F_0(u)](t,x))\hat{\alpha}_0(t,x,\zeta_0,\zeta),$$

where $b, b_0, \hat{b}_0$ are continuous, and $\hat{\alpha}_0$ is measurable in t, x, continuous in the other variables, and they satisfy

$$b(\theta) \geq \frac{c_2}{1+|\theta|^{\sigma^\star}}, \quad b_0(\theta) \geq \frac{c_2}{1+|\theta|^{\sigma^\star}}$$

with some positive constants c_2 and $\sigma^\star < p-1$,

$$|\hat{b}_0(\theta)| \leq 1 + |\theta|^{p-1-\varrho^\star}$$

with $\varrho^\star < p-1$, and

$$|\hat{\alpha}_0(t,x,\zeta_0,\zeta)| \leq c_1(|\zeta_0|^{\hat{\varrho}} + |\zeta|^{\hat{\varrho}}), \quad 0 \leq \hat{\varrho}, \quad \sigma^\star + \hat{\varrho} < \varrho^\star.$$

Finally,

$$H, H_0 : L^p(0,T;V) \to C(\overline{Q_T}), \quad F_0 : L^p(0,T;V) \to L^p(Q_T)$$

are linear continuous operators of Volterra type. Thus, $[H(u)](t,x)$ and $[H_0(u)](t,x)$ may have one of the forms

$$\int_{Q_t} d(t,x,\tau,\xi)u(\tau,\xi)d\tau d\xi, \qquad \sup_{(t,x)\in Q_T} \int_{Q_T} |d(t,x,\tau,\xi)|^q d\tau d\xi < \infty,$$
$$\int_{\Gamma_t} d(t,x,\tau,\xi)u(\tau,\xi)d\tau d\sigma_\xi, \qquad \sup_{(t,x)\in Q_T} \int_{\Gamma_T} |d(t,x,\tau,\xi)|^q d\tau d\sigma_\xi < \infty,$$

$d(t,x,\tau,\xi)$ is continuous in (t,x), $\Gamma_t = (0,t)\times\partial\Omega$ or

$$\sum_{j=1}^n d_j(t,x)\int_{Q_t} \tilde{d}_j(\tau,\xi)D_j u(\tau,\xi)d\tau d\xi, \quad d_j \in C(\overline{Q_T}), \quad \tilde{d}_j \in L^q(Q_T).$$

One can show that the examples of the above type satisfy the conditions of Theorems 1 and 2 in the case when

$$H, H_0 : L^p(Q_T) \to L^p(Q_T)$$

are continuous linear operators (for a fixed $T > 0$ or arbitrary finite $T > 0$, respectively) and b, b_0 are bounded. Thus (for bounded b, b_0), $[H(u)](t,x)$ and $[H_0(u)](t,x)$ may also have the forms given in [SiJa08] for F_0; that is,

$$\int_0^t d(t,x,\tau)u(\tau,x)d\tau, \qquad \int_\Omega d(t,x,\xi)u(t,\xi)d\xi,$$

where

$$\int_0^T \sup_{x\in\Omega}\left[\int_0^T |d(t,x,\tau)|^q d\tau\right]^{p/q} dt < \infty,$$
$$\int_\Omega \sup_{t\in[0,T]}\left[\int_\Omega |d(t,x,\xi)|^q d\xi\right]^{p/q} dx < \infty,$$

respectively.

Acknowledgement. This work was supported by the Hungarian National Foundation for Scientific Research under grant OTKA T 049819.

References

[Ad75] Adams, R.A.: *Sobolev Spaces,* Academic Press, New York (1975).

[BeMu92] Berkovits, J., Mustonen, V.: Topological degree for perturbations of linear maximal monotone mappings and applications to a class of parabolic problems. *Rend. Mat. Ser. VII Roma,* **12**, 597–621 (1992).

[Li69] Lions, J.-L.: *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires,* Dunod, Paris (1969).

[Si00] Simon, L.: On nonlinear hyperbolic functional differential equations. *Math. Nachr.,* **217**, 175–186 (2000).

[SiJa08] Simon, L., Jäger, W.: On non-uniformly parabolic functional differential equations. *Studia Sci. Math. Hungar.,* **45**, 285–300 (2008).

[Si08] Simon, L.: Application of monotone type operators to parabolic and functional parabolic PDE's. *Handbook of Differential Equations, Evolutionary Equations,* **4**, 267–321 (2008).

[Si09] Simon, L.: On some properties of nonlinear functional parabolic equations. *Internat. J. Qualitative Theory Differential Equations Appl.* (to appear).

[Ze90] Zeidler, E.: *Nonlinear Functional Analysis and its Applications. II A, II B,* Springer, Berlin (1990).

31

Grid Computing for Multi-Spectral Tomographic Reconstruction of Chlorophyll Concentration in Ocean Water

R.P. Souto,[1] H.F. de Campos Velho,[1] F.F. Paes,[1] S. Stephany,[1] P.O.A. Navaux,[2] A.S. Charão,[3] and J.K. Vizzotto[4]

[1] National Institute for Space Research, São José dos Campos, Brazil; rpsouto@gmail.com, haroldo@lac.inpe.br, fabiana.paes@lac.inpe.br, stephan@lac.inpe.br
[2] Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil; navaux@inf.ufrgs.br
[3] Universidade Federal de Santa Maria, Brazil; andrea@inf.ufsm.br
[4] Centro Universitário Franciscano, Santa Maria, Brazil; juvizzotto@gmail.com

31.1 Introduction

In the last decades, the development of inversion methodologies for radiative transfer problems has been an important research topic in many branches of science and engineering [Go02, Mc92]. The direct or forward radiative transfer problem in hydrologic optics, in the steady state, involves the determination of the radiance distribution in a body of water, given the boundary conditions, source term, inherent optical properties (IOPs), such as the absorption and scattering coefficients, and the phase function. The inverse radiative transfer problem arises when physical properties, internal light sources, and/or boundary conditions must be estimated from radiometric measurements of the underwater light field. A challenge in the inverse hydrological optics problem is to determine the IOPs, considering only the water-leaving radiance.

The inverse problem is formulated as an optimization problem and iteratively solved using a recent intrinsic regularization scheme [PrEtAl04, SoEtAl04b] coupled to an ant colony optimization (ACO). The regularization scheme pre-selects candidate solutions based on their smoothness, quantified by a Tikhonov norm [PrEtAl04]. Profiles generated with the wrong curvature are filtered out using a second derivative criterion [SoEtAl09, SoEtAl07]. An objective function is given by the square difference between computed and experimental radiances at every iteration. Each candidate solution corresponds to a discrete chlorophyll profile.

The chlorophyll profile is reconstructed from multi-spectral water-leaving radiances of ocean surface, following Chalhoub and Campos Velho [ChCV03].

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_31,

Vertical values of the absorption and scattering coefficients are estimated from the chlorophyll profile by means of bio-optical models [Mo94], for each pixel in the satellite image. For an ocean surface, the image contains many pixels. However, each pixel inversion is independent from the inversion of other pixels. This constitutes a good challenge to be addressed by the grid computing approach. The OurGrid middleware [CiEtAl05] was used to manage these jobs (pixel inversion). A grid infrastructure was built to perform the inversion for each pixel, managing a queue of independent jobs submitted to three clusters spread over Brazil.

31.2 Light Transmission in Natural Water

The radiative transfer equation (RTE) models the transport of photons through a medium [Go02]. Light intensity is given by a directional quantity, the radiance L, measuring the rate of energy being transported at a given point and in a given direction. This direction is defined by a polar angle θ (relative to the normal of the plane) and an azimuthal angle φ (a possible direction in that plane). At any point of the medium, light can be absorbed, scattered or transmitted, according to the absorption (a) and scattering (b) coefficients and to a scattering phase function that models how light is scattered in any direction. An attenuation coefficient c is defined as $c = a + b$, and the geometrical depth is mapped to an optical depth τ. Assuming a plane-parallel geometry, for the case of azimuthal symmetry (no dependence on j), isotropic medium, and absence of a source term, and making the radiance $L_\lambda(\tau, \mu, \varphi) = L_\lambda$, the one-dimensional integro-differential RTE can be written as

$$\mu \frac{d}{d\tau} L_\lambda(\tau, \mu) + L_\lambda(\tau, \mu) = \frac{\varpi_0(\tau, \lambda)}{2} \int_{-1}^{1} L_\lambda(\tau, \mu) d\mu',$$

subjected to the boundary conditions

$$L_\lambda(0, \mu) = F\delta(\mu - \mu_0), \quad L_\lambda(\zeta, -\mu) = 0.$$

A heterogeneous medium can be modeled as a set of R homogeneous finite layers. Optical variable τ is discretized in $R + 1$ values, varying from $\tau_0 = 0$ up to $\tau_R = \zeta$, where ζ is the medium *optical depth*. Then, for $r = 1, 2, \ldots, R$ and $\mu \in (0, 1]$, the problem in this multi-region geometry can be given by

$$\mu \frac{d}{d\tau} L_{r,\lambda}(\tau, \mu) + L_{r,\lambda}(\tau, \mu) = \frac{\varpi_r(\lambda)}{2} \int_{-1}^{1} L_{r,\lambda}(\tau, \mu) d\mu',$$

with

$$\varpi_0(\tau, \lambda) = \varpi_r(\lambda) = \frac{b_r(\lambda)}{c_r(\lambda)} = \frac{b_r(\lambda)}{a_r(\lambda) + b_r(\lambda)}$$

constant in the region r, for any value of τ, where $c_r(\lambda)$, $a_r(\lambda)$, and $b_r(\lambda)$ are, respectively, the attenuation, the absorption, and the scattering coefficients, for a given wavelength λ.

Bio-optical models are employed to correlate the absorption and scattering coefficients to the chlorophyll concentration. These coefficients are assumed to be constant in each region. Therefore, discrete values a_r and b_r can be estimated for each region from the discrete values C_r. Chlorophyll profiles can be represented according to Gaussian distributions [Mo94]:

$$C(z) = C_{bg} + \frac{h}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - z_{max}}{\sigma}\right)^2 \right], \tag{31.1}$$

where z is the depth in meters and $C(z)$ is given in mg/m^3. This profile can be seen in the results section of this work, termed *exact*. A bio-optical model was formulated by Morel [Go84] for the absorption coefficient, and for the scattering coefficient by Gordon and Morel [Ch60],

$$a_r(\lambda) = [a^w + 0.06a^c C_r^{0.65}][1 + 0.2e^{-0.014(\lambda-440)}],$$
$$b_r(\lambda) = (550/\lambda)\, 0.30 C_r^{0.62},$$

where a^w is the pure water absorption and a^c is a nondimensional, statistically derived chlorophyll-specific absorption coefficient, and λ is the considered wavelength. The values of a^w and a^c depend on the wavelength and can be found in tables [Mo94].

31.3 Inversion Scheme

The inverse problem is formulated according to an implicit approach, leading to an optimization problem. The algorithm is expressed as a constrained nonlinear optimization problem, in which the direct problem is iteratively solved for successive approximations of the unknown parameters. Iteration proceeds until an objective function, representing the least-squares fit of the model results and experimental data added to a regularization term, converges to a specified small value.

The set of parameters to be estimated is $R + 1$ discrete values of the chlorophyll concentration C_r, for $r = 0, 1, \ldots, R$ at optical depths τ_r taken at the upper interface of the regions. Experimental data are the discrete radiances $L^{\exp}(\tau_0, \mu_i, \lambda_j)$ for $i = 1, 2, \ldots, N_\mu/2$ and $j = 1, 2, \ldots, N_\lambda$. Therefore, the $R+1$ discrete values of the concentration are estimated from $N_\mu/(2\,N_\lambda)$ spectral radiance values right above the sea surface. The objective function $J(C)$ is given by the square difference between the experimental and model radiances plus a regularization term,

$$J(C) = \sum_{i=1}^{N_\mu/2} \sum_{j=1}^{N_\lambda} [L^{\exp}(\tau_0, -\mu_i, \lambda_j) - L_C(\tau_0, -\mu_i, \lambda_j)]^2 + \gamma\Gamma(C), \tag{31.2}$$

where $\Gamma(C)$ is the regularization function, which is weighted by a regularization parameter γ. For instance, the second order Tikhonov regularization [TiAr77] is defined by

$$\Gamma(C) = \sum_{r=0}^{R-2} (C_r - 2C_{r+1} + C_{r+2})^2.$$

31.3.1 Ant Colony Optimization

The ant colony optimization (ACO) is a method based on the collective behavior of ants choosing the shortest path between the nest and a food source [DoEtAl96]. Each ant marks its path with an amount of pheromone and the marked path is further employed by other ants as a reference. Several generations of ants are produced. For each generation, a fixed amount of ants (na) is evaluated. Each ant is associated to a feasible path and this path represents a candidate solution, being composed of a particular set of edges of the graph that contains all possible solutions. Each ant is generated by choosing these edges on a probabilistic basis. A solution is composed of linking ns nodes and in order to connect each pair of nodes, np discrete values can be chosen. This approach was used to deal with a continuous domain. Therefore, there are $ns \times np$ possible paths $[i, j]$ available. Denoting by ρ the pheromone decay rate, the amount of pheromone T_{jj} at generation t is given by

$$T_{ij}(t) = (1 - \rho)T_{ij}(t - 1) \quad t = 1, 2, \ldots, mit,$$

where mit is the maximum number of iterations.

This approach was successfully used for many graph-like problems [DoEtAl96]. The best ant of each generation is then chosen, and it is allowed to mark its path with pheromone. This will influence the creation of ants in future generations. The pheromone decays due to an evaporation rate. Finally, at the end of all generations, the best solution is assumed to be achieved.

A parallel implementation [SoEtAl04b] of the ACO-IR (ACO with intrinsic regularization) was executed in a distributed memory machine. Parallelization is important since this problem is very computationally intensive.

31.3.2 Intrinsic Regularization and the Concavity Criterion

In this chapter, an ACO-based inverse solver with an intrinsic regularization scheme [PrEtAl04, SoEtAl04b] is employed without the regularization term ($\gamma = 0$) shown in equation (31.2).

As a smooth profile is required, this is known information about the inverse solution. Such knowledge is included in the generation of the candidate solutions by means of pre-selecting the smoother ants according to the second order Tikhonov norm. Actually, a kind of pre-regularization is performed. Therefore, the usual regularization term is not required.

Besides the smoothness, additional information is also used to compute the inverse solution: the concavity of the chlorophyll profile, which is verified by means of its second derivative. Since only curves with negative concavity are expected, a penalty is assigned to profiles with positive concavity. For each of these profiles, an overhead value is added to the evaluated objective function (equation (31.2)).

31.4 Chlorophyll Concentration: 3D Reconstruction

We simulate a specific case with dimensions of 60 km × 60 km, and 40 meters of depth for the ocean spatial domain. The horizontal domain is uniformly divided into 36 smaller regions of 10 km × 10 km. There are three profiles to be recovered in the whole domain. Each profile was generated employing the Gaussian model given by (31.1). The parameters used to construct each profile are shown in Table 31.1.

Table 31.1. Parameters of Gaussian chlorophyll profiles for equation (31.1).

Profile	C_{bg}	h	σ	z_{max}
1	0.2	144.0	9.0	17.0
2	0.2	144.0	9.0	25.0
3	0.2	144.0	12.0	17.0

The concentration profiles are shown in Figure 31.1, and the profile distribution is shown in Figure 31.2. As one can note, there are 20 sub-regions with profile 1, 12 with profile 3, and 4 sub-regions which correspond to profile 2. For each profile there is a set of radiance multi-spectral values which come from the ocean surface. A random noise of 1% was added in the radiance values of all regions. Each region with the same profile has a different initial random sequence of noise values, i.e., a different seed sequence.

It is supposed that a good estimation was obtained up to the peak of the curve (average profile) with poorer agreement to the lower part of the profile (depth below the peak). In order to improve the inverse solution, a two-step strategy is used: in *step 1,* the estimation has already been performed for the whole profile, and then, in *step 2,* the reconstruction is carried out only for the lower part of the curve. In *step 2,* each ant is still related to the whole profile, but the values obtained in *step 1* for the upper part of the water layer are frozen. In other words, *step 2* is a new inverse problem, but simpler than the original problem, because the *step 2* problem has a lower dimension and a good first guess (obtained in *step 1*).

In each region, the inverse problem of recovering the chlorophyll concentration profile, based on the water-leaving radiances, must be solved. As these profile reconstructions are independent of each other, the set of inversions is

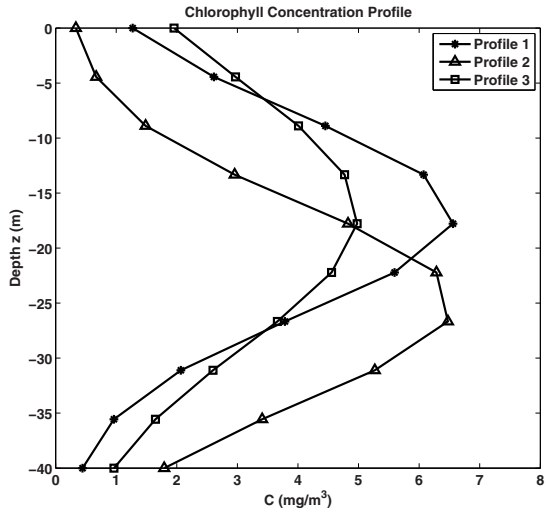


Fig. 31.1. Chlorophyll concentration profiles for the parameters in Table 31.1.

P1	P1	P1	P1	P1	P1
P1	P3	P3	P3	P3	P1
P1	P3	P2	P2	P3	P1
P1	P3	P2	P2	P3	P1
P1	P3	P3	P3	P3	P1
P1	P1	P1	P1	P1	P1

Fig. 31.2. Profile distribution in a spatial domain split into 36 regions of 10 km × 10 km each.

treated here as a "Bag-of-Tasks" application, i.e., a fully independent set of tasks. Therefore, the use of a grid environment to perform the inversion in the complete spatial domain, in feasible time, was a natural choice. It was performed with a total of 56 jobs in the grid, where 36 are in regard to *step 1* profile recovering, and 20 jobs concern the *step 2* reconstruction for only profile 1. The jobs on the grid are assigned each pixel on the ocean surface (a sub-domain), *vectorizing* the radiances associated to each pixel, and preparing them for the inversion procedure. The process starts for the execution of the Bag-of-Tasks procedure, where a script provides the jobs to the clusters.

A grid with three clusters was geographically spread over Brazil, and it was configured on three clusters, as described in Table 31.2. The OurGrid [CiEtAl05] middleware was employed in our tomographic reconstruction. This middleware is targeted to grid computing with Bag-of-Tasks applications. Three parts of this middleware can be identified:

Table 31.2. Hardware equipment for the "3D Ocean Color" grid.

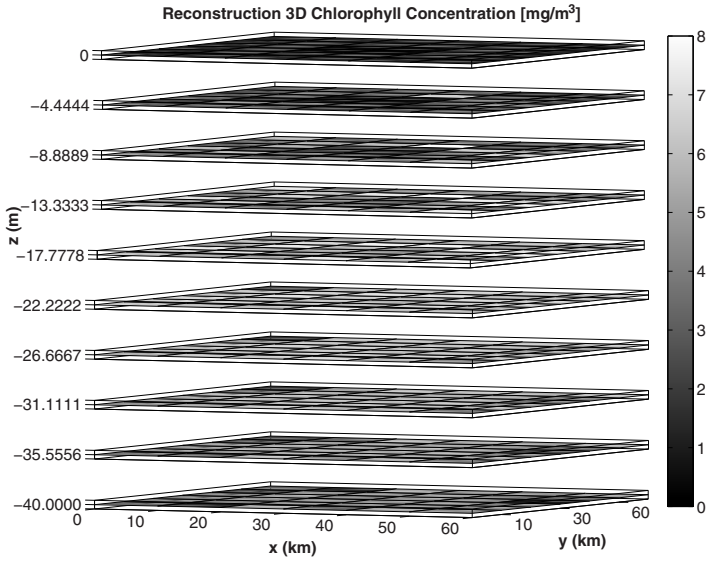
Institution	Equipment
Computing/UFSM	SGI Altix XE server: 8 cores (2 Intel Xeon quad-core 2.0 GHz)
II/UFRGS	Cluster Cray XD1: 4 processors AMD Opteron, 2.8 GHz
INPE/LAC	Cluster Cray XD1: 8 processors AMD Opteron, 2.8 GHz

- **mygrid**: a user interface for job submission and execution from the *home machine*;
- **peer**: provide the computers linked in a home machine (it is the *peer machine*): the component is installed on different machines;
- **useragent**: run in each grid machine, these are the machines that run the tasks on OurGrid.

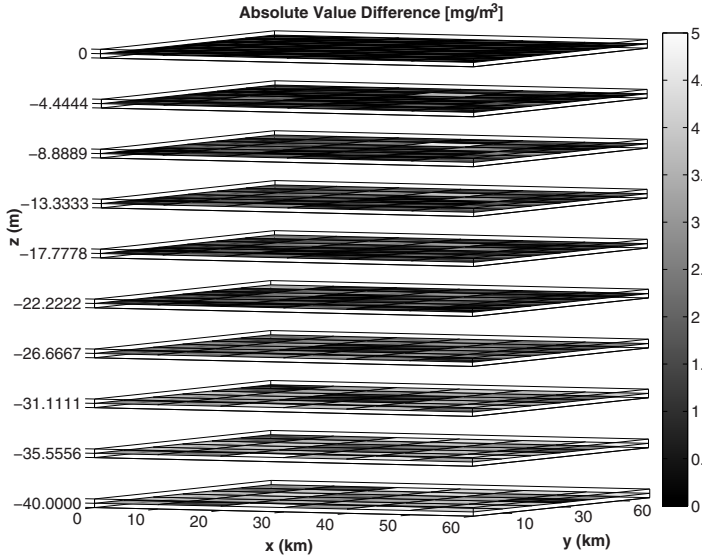
The inverse solver was tested for a multi-region ($R = 9$) offshore ocean water radiative transfer problem with azimuthal symmetry, using multi-spectral radiance data, with a 1% random noise. This data is related to the emerging radiances at the water surface and includes $N_\mu/2 = 10$ polar directions for each of the $N_\lambda = 10$ wavelengths. In the considered test cases, synthetic data was used to simulate the experimental values.

The tuning of the parameters in the ACO may have a big influence on the results. The ACO implementation required adjustment of parameters such as the pheromone decay rate (ρ) and q_0, used in the roulette scheme. Here $\rho = 0.03$ and $q_0 = 0$ were used. A faster convergence is obtained for higher values of ρ, but the inverse solution is not good. There are other parameters in the process, such as the number of possible paths (np) between each pair of the ns nodes, the number of ants na, or the maximum number of iterations mit. These values are shown in Table 31.3, and they are used for the test cases. The ACO was executed using $na = 360$ and pre-selecting 1/30 of these ants ($na_p = 12$) according to their smoothness. The range of search for the chlorophyll concentration varied from 0.0003 (1.0/3000.0) up to 10.0 (3000.0/300.0).

Figure 31.3(a) shows the inversion for the 3D chlorophyll concentration to the ocean surface containing 36×36 pixels, and Figure 31.3 (b) is the modulus of the difference between the true and estimated chlorophyll concentration. It is possible to realize that a good reconstruction is obtained up to 22 m.



(a)



(b)

Fig. 31.3. Chlorophyll concentration: (a) estimated, (b) difference from the real value.

Table 31.3. Seeds and ACO parameters.

Seeds (10)	3, 15, 21, 31, 45, 63, 77, 81, 95, 99						
ACO parameters	ns	np	na	na_p	mit	ρ	$q0$
	10	3000	360	12	400	0.03	0.0

The accumulated middleware cluster time (T_p) is defined here as the overall runtime of jobs submitted to grid nodes by OurGrid middleware. Such accumulated time corresponds to a sequential execution of a job set. Table 31.4 presents T_p measures for each cluster during our experiment. This experiment consisted of 56 jobs and required 21:34 hours to complete, as shown in Table 31.4. All time measures presented here were extracted from the OurGrid log file, which registers the start/end time for each job.

Table 31.4. Accumulated cluster time (hours).

Cluster	Jobs	T_p (hh:mm)	T_p/Jobs (hh:mm:ss)
Computing-UFSM	30	07:20	00:14:40
II-UFRGS	9	06:48	00:45:20
INPE/LAC	17	07:26	00:26:11
	56	21:34	00:23:06

Table 31.4 also presents the average duration of a single job on each cluster and for the whole experiment. This analysis shows that each cluster has a different job execution rate. The first cluster (Computing-UFSM) presents the shorter runtimes, which can be explained by its shared-memory multicore architecture, which reduces communication costs for the parallel program implementing the ACO. Since computations are eventually overlapped in time (up to the number of available grid nodes), and the time of grid usage demanded to perform the whole set of jobs is defined here as grid elapsed time (T_g), which is naturally shorter than T_p. For the experiment we consider here, the grid elapsed time was **07:43 hours**. The ratio T_p/T_g gives the speedup for the grid execution, which is **2.79** for this experiment.

31.5 Conclusion

In this work, we tested the feasibility of recovering different kinds of chlorophyll profiles, in a given spatial domain in the ocean, from basically two points of view: the accuracy of the reconstruction, and also as an application that can intensively use a grid environment.

By distributing the jobs over the grid, we have effectively reduced the time needed to obtain the results. The grid was easily deployed using the OurGrid middleware, which scheduled the job set onto the available grid nodes. The cluster usage statistics can be calculated from the ratio between T_p (for each cluster) and T_g. These statistics confirm that each cluster was busy most of the time, although there is still room to improve the grid usage. The performance gain could be even better, considering that not all the clusters were simultaneously used all the time.

References

[ChCV03] Chalhoub, E.S., Campos Velho, H.F.: Multispectral reconstruction of bioluminescence term in natural waters. *Appl. Numer. Math.*, **47**, 365–376 (2003).

[Ch60] Chandrasekhar, S.: *Radiative Transfer,* Dover, New York (1960).

[CiEtAl05] Cirne, W., Brasileiro, F., Paranhos, D., Costa, L., Santos-Neto, E., Osthoff, C.: *Building a User-Level Grid for Bag-of-Tasks Applications in the HPC: Paradigm and Infrastructure,* Wiley, New York (2005).

[DoEtAl96] Dorigo, M., Maniezzo, V., Colorni, A.: The ant optimization: optimization by a colony of cooperating agents. *IEEE Trans. Syst. Man Cybernet. Part B,* **26**, 29–41 (1996).

[Go84] Gordon, H.R.: Remote sensing marine bioluminescence: the role of the in-water scalar irradiance. *Appl. Optimization,* **24**, 1694–1696 (1984).

[Go02] Gordon, H.R.: Inverse methods in hydrologic optics. *Oceanologia,* **44**, 9–58 (2002).

[Mc92] McCormick, N.: Inverse radiative transfer problems: a review. *Nuclear Sci. Engrg.,* **112**, 185–198 (1992).

[Mo94] Mobley, C.: *Light and Water: Radiative Transfer in Natural Waters,* Academic Press, New York (1994).

[Mor91] Morel, A.: Light and marine photosynthesis: a spectral model with geochemical and climatological implications. *Progress Oceanography,* **26**, 263–306 (1991).

[PrEtAl04] Preto, A.J., Campos Velho, H.F., Becceneri, J., Arai, N. Souto, R.P., Stephany, S.: A new regularization technique for an ant-colony-based inverse solver applied to a crystal growth problem, in *13th Inverse Problem in Engineering Seminar* (2004), 147–153.

[SoEtAl07] Souto, R.P., Barbosa, V.C., Campos Velho, H.F., Stephany, S.: Determining chlorophyll concentration in off-shore sea water from multispectral radiances by using second derivative criterion and ant colony meta-heuristic, in *Inverse Problems, Design and Optimization Symposium. Vol. I* (2007), 341–348.

[SoEtAl09] Souto, R.P., Campos Velho, H.F., Stephany, S., Barbosa, V.C.: Multispectral inversion for chlorophyll concentration in offshore sea water by using the ant colony optimization and the second derivative criterion. *J. Quant. Spectrosc. Radiat. Transfer* (submitted).

[SoEtAl04a] Souto, R.P., Campos Velho, H.F., Stephany, S., Chaulhoub, E.: Performance analysis of radiative transfer algorithms in a parallel environment. *Transport Theory Stat. Phys.,* **33**, 449–468 (2004).

[SoEtAl04b] Souto, R.P., Campos Velho, H.F., Stephany, S., Sandri, S.A.: Reconstruction of chlorophyll concentration profile in offshore ocean water using a parallel ant colony code, in *Hybrid Metaheuristics (Proceedings)* (2004), 19–24.

[TiAr77] Tikhonov, A., Arsenin, V.: *Solutions of Ill-Posed Problems,* Winston & Sons, Washington, D.C. (1977).

32

Long-Time Solution of the Wave Equation Using Nonlinear Dissipative Structures

J. Steinhoff and S. Chitta

University of Tennessee Space Institute, Tullahoma, TN; jsteinho@utsi.edu, subha@flowanalysis.com

32.1 Introduction

A new method, "wave confinement" (WC), is developed to efficiently solve the scalar wave equation on a discretized domain. This method is similar to the originally developed method, "vorticity confinement," which is used to solve a vast range of fluid dynamics problems [StWePu95]. WC involves modifying the discretized wave equation by adding a nonlinear term to generate traveling "dissipative solitary" waves that are stable to perturbations due to numerical effects, such as dissipation and dispersion. As the present study involves treating thin waves propagating long distances, on feasible computational grids, the propagating functions cannot be more than 2–3 cells wide. In these cases, since the accuracy of conventional higher-order schemes increases only as the number of points across the pulse becomes relatively large, they are not useful. Often, for these cases, the main quantities of interest in the far field are the integrated amplitude and the motion of the centroid surfaces (which we use to represent wave fronts), rather than the details of the internal structure of the pulse. For realistic problems, these pulse surfaces can have multiple sources and scattering surfaces, propagate through regions with a varying refraction index, and have complex topology. Accordingly, we only consider Eulerian methods, where such general surface topologies can automatically be treated with no need for complex "surface fitting" or adaptive grids.

WC has the potential to greatly extend the range of application of existing computational methods for certain problems. The new method has many of the advantages of Green's function-based integral equation methods for long-distance propagation, since the propagation distance can be indefinitely long. However, unlike Green's function schemes, which are most useful for a uniform index of refraction in simple domains, WC allows short pulses to automatically and efficiently propagate through regions with a varying index of refraction and undergo multiple scattering in complex domains, since it is an Eulerian finite difference technique.

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_32,

32.2 Approach

WC involves treating a thin feature, such as a pulse, as a type of weak solution of the governing partial differential equation (PDE). Within the feature, a discretized nonlinear PDE is solved, whose solution can be as thin as 2–3 grid cells, so that it does not necessarily represent an accurate Taylor expansion discretization of the PDE, yet retains the essential physical features. The approach is similar to shock capturing [La58], where conservation laws are satisfied, so that integral quantities such as total amplitude and centroid motion are accurately computed for the feature.

32.2.1 WC as a PDE

For simplicity, we first consider a scalar, ϕ, advecting at a constant speed c:

$$\partial_t \phi = -c\partial_x \phi. \tag{32.1}$$

Our basic point is that there will be errors when we discretize equation (32.1) using conventional schemes based on Taylor expansions. When we confine the pulse solution to $\sim$ 2–3 grid cells, which is our goal, the derivatives of ϕ and hence these "errors" will be large. Also, corresponding to the small number of grid points within the pulse, there will be only a small number of quantities that we can conserve. Adding a term $E = \partial_x^2 F(\{\phi\})$ to (32.1) that vanishes at the boundaries, along with derivatives, will not affect the conservation of these quantities, which include the total amplitude

$$A = \int \phi dx, \tag{32.2}$$

and the speed of the centroid

$$\frac{d\langle x\rangle}{dt} = \frac{\int \phi c(x) dx}{A},$$

where the centroid is $\langle x\rangle = \frac{\int x\phi dx}{A}$. To preserve these essential physical characteristics of (32.1) for a short convecting pulse, we want E to satisfy a set of conditions described above. In addition, it should be homogenous of degree one, so like the original PDE, the dynamics of propagation does not depend on the magnitude of ϕ. This is an important distinction of the WC equation. Many nonlinear equations use nonhomogenous terms for the nonlinear term [RoHySt07]. In fact, Cahn and Hilliard in 1958 used such a nonlinear term under a second derivative as in our equation, but one that was not homogenous [CaHi83]. The PDE in (32.1) with the confinement term is

$$\partial_t \phi = -c\partial_x \phi + \partial_x^2 F. \tag{32.3}$$

One example of F that proves to be stable is

$$F = \frac{\alpha}{\psi^2}\left[\partial_x^2\psi - \lambda\psi\right],$$

where $\psi = \phi^{-1}$ and λ is a constant that defines the width of the pulse. Using the chain rule, we have

$$\partial_t\phi = -c\partial_x\phi - \alpha\lambda\partial_x^2\phi - \alpha\partial_x^2\left(\partial_x^2\phi - 2\frac{(\partial_x\phi)^2}{\phi}\right). \tag{32.4}$$

Then, we define the three "confinement" terms, $F = F_0 + F_1 + F_2$, where $F_0 = -\alpha\lambda\partial_x^2\phi$ and $F_1 = -\alpha\partial_x^4\phi$ are linear, and $F_2 = 2\alpha\partial_x^2\left(\frac{(\partial_x\phi)^2}{\phi}\right)$ is nonlinear. It is interesting that the second-order term, F_0, in (32.4) behaves in a different way from most popular nonlinear PDEs, such as KdV, that harbor solitary wave. In these, the linear term is the "expansion" term, and the "contraction" or "steepener" term is the nonlinear Burgers-like convection: $(\partial_x\phi^2/2)$. In WC, the linear second-order term, F_0, acts to *contract* the pulse, and the nonlinear term, F_2, prevents ϕ from changing sign and transfers the amplitude from large wavelengths to small. The higher-order term, F_1, acts as diffusion for short wavelengths and prevents the pulse from diverging. In the convecting frame of the pulse, $\xi = x - ct$, the PDE becomes the heat equation

$$\partial_t\phi = \partial_\xi^2 F. \tag{32.5}$$

When (32.5) converges, the pulse then relaxes to the form

$$\phi \to \phi_0 \sec h\beta(\xi - \xi_0),$$

where $\beta = \sqrt{\lambda}$, and ϕ_0 and ξ_0 are arbitrary constants. An important point is that wavelengths created by perturbations (such as numerical errors) that are longer than the thin features that are to be confined must have a negative diffusive behavior, so that the features remain confined, and are stable to perturbations against spreading. This means that F_2 must be nonlinear. It is easy to show by von Neumann analysis that a linear combination of terms, with a negative lower-order dissipation, cannot lead to a stable confinement for any finite range of coefficients: any wavelength that exhibits negative diffusion would eventually diverge.

The appearance of ϕ in the denominator of (32.4) makes F_2 diverge as $\phi \to 0$. This prevents ϕ from changing sign. Since A in (32.2) is conserved, the integral of ϕ over any finite region cannot then diverge. In the discretized version defined below, none of the grid values can diverge. This ensures realizability if ϕ is a physical quantity. Smolarkiewicz [Sm83] also has rearranged the discretized convection equation so that there is such a term in the denominator for this reason.

32.2.2 Discretized Representation: 1D Scalar Advection

One discretized formulation of the PDE given in (32.3) can be written in the form

$$\phi_j^{n+1} = \phi_j^n - \frac{\nu}{2}\left(\phi_{j+1}^n - \phi_{j-1}^n\right) + a\delta_j^2 F_j^n, \tag{32.6}$$

where $\delta_j^2 f_j = f_{j-1} - 2f_j + f_{j-1}$, $\nu = \frac{c\Delta t}{h}$, $a = \frac{\Delta t}{h^2}$, Δt is the time step, and h is the grid cell size. Many conventional schemes can be put in this form, where F adds a (typically linear) stabilizing dissipation. However, the role of F is very different here. The confinement term, F, is defined as

$$F_j^n = \mu\phi_j^n - \varepsilon\Phi_j^n,$$

where Φ is a nonlinear function of ϕ (given below) and μ is a diffusion coefficient that can include numerical discretization effects in a conventional convection or wave equation solution (we assume physical diffusion is much smaller). ε is a numerical coefficient that, together with μ, controls the size and time scales of the confined features. For this reason, we refer to the two terms as "confinement terms."

There are many possibilities for Φ on the grid. A simple class is

$$\Phi_j^n = \left[\frac{\sum_{l=-1}^{+1}(\phi_{j+l}^n)^{-1}}{N}\right]^{-1}. \tag{32.7}$$

The above sum is over a set of $N-1$ neighboring grid nodes and the node where Φ is computed. Upon Taylor expansion, we wish to recover the PDE given by (32.4) in the fine grid limit. The two (positive) parameters, ε and μ, are determined by the two small scales of the computation, h and Δt, since we want the small features to relax to their solitary wave shape in a small number of time steps and to have an effective support of a small number of grid cells. Thus, even though h may be small, the Laplacian will be large and the total effect also large. At convergence, $\mu\phi - \varepsilon\Phi \approx 0$ (not exactly zero because the convection term is continually adding a perturbation). The solution to the above equation that vanishes in the far field is then

$$\phi \to \phi_0 \mathrm{sech}\left[\gamma(j - j_0 - \nu n)\right],$$

where j_0 is the approximate initial position of the centroid and ϕ_0 is an arbitrary constant. The pulse width coefficient, γ, is a function of $\dfrac{\varepsilon}{\mu}$ and is given as

$$\cosh(\gamma) = \frac{1}{2}\left(\frac{3\varepsilon}{\mu} - 1\right).$$

A pulse, which is (in this example) initially a (Kronecker) delta, is solved using (32.6) with periodic boundary conditions, and compared to the solution of a higher-order, conventional method. For this computation, the parameters used are $\nu = 0$, $\mu = 0.2$, and $\varepsilon = 0.3$. As can be seen in Figure 32.1, the higher-order method quickly spreads the pulse to many grid cells while WC keeps the pulse (effectively) compact.

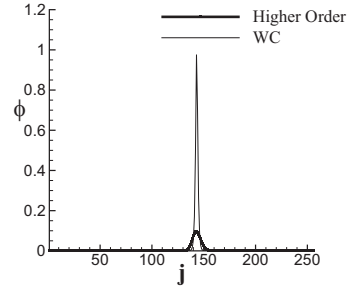


Fig. 32.1. Pulse with WC compared to higher-order method.

32.2.3 Wave Equation

We start with the one-dimensional (1D) scalar wave equation with constant wave speed, c, for simplicity. As in scalar convection, we add an additional term to control the shape of a short pulse. We take the wave equation analog of the dissipative form used in the advection equation:

$$\partial_t^2 \phi = c^2 \partial_x^2 \phi + \partial_t \partial_x^2 F,$$

or, using a simple time discretization,

$$\delta_n^2 \phi = \nu^2 \delta_j^2 \phi + a' \delta_n, \delta_j^2 F,$$

where $\delta_n f^n = f^n - f^{n-1}, \delta_n^2 f^n = f^n - 2f^{n-1} + f^{n-2}, a' = \frac{\Delta t^2}{h^2}$. It was seen above and in other works [StDiHa], that the addition of WC terms in the form of second derivatives of a function that has short range do not change the propagation speed (or the total amplitude) of an advecting, confined pulse. The same is true for the wave equation, if an additional time derivative is applied. The main constraint on the confinement term, F, as in advection, is that it force an initial isolated, propagating short range pulse with a single maximum to remain short range and also not develop any additional maxima. We again use $F^n = \mu\phi^n - \varepsilon\Phi^n$, where Φ has the form given by (32.7) in terms of its argument.

Results for the 1D wave equation are shown in Figure 32.2. When $\mu = 0$ and $\varepsilon = 0$, the solution is dispersive. Adding a small quantity of positive diffusion, $\mu = 0.2$, will smooth the solution but it will be highly dissipative. To overcome anomalous dissipation, the confinement term is added, which will stabilize the solution. An important feature of the method is that the waves do not suffer a "phase shift" when they pass through each other, like most soliton solutions. This is an obvious requirement for the equation we want to simulate—the linear wave equation. However, the confinement term is nonlinear. Such a phase shift would ordinarily show up as a kink in two waves in two or three dimensions that are passing through each other. Results for the centroid trajectories for two wave equation pulses passing through each

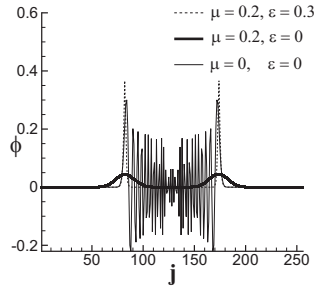


Fig. 32.2. Wave equation solution.

other in 1D are presented in Figure 32.3. It can be seen that there is no phase shift to plottable accuracy in spite of the nonlinearity. The pulses seem to be effectively transparent to each other (after a short relaxation time).

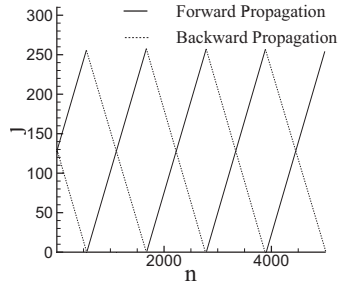


Fig. 32.3. Centroid positions of forward and backward propagating pulses.

The extension to multidimensions is almost trivial. We simply substitute a multidimensional Laplacian in the original wave equation and use a multidimensional harmonic mean, where we sum inverse values of ϕ over the central point and the $N-1$ neighboring grid points on the multidimensional grid. If we consider a straight, 2D propagating pulse aligned at an angle, θ, where θ is arbitrary and propagating in the normal direction (for isotropic media), the solution at convergence is

$$\phi_{i,j} = \phi_0 \text{sech}\left[\gamma\left(r_{i,j} - r_0\right)\right],$$

where $r - r_0$ is the distance from a point (grid point in the discrete case) to the centroid along θ. As before, ϕ_0 is arbitrary and γ is given for a 2-D planar pulse at an angle θ, on a grid by

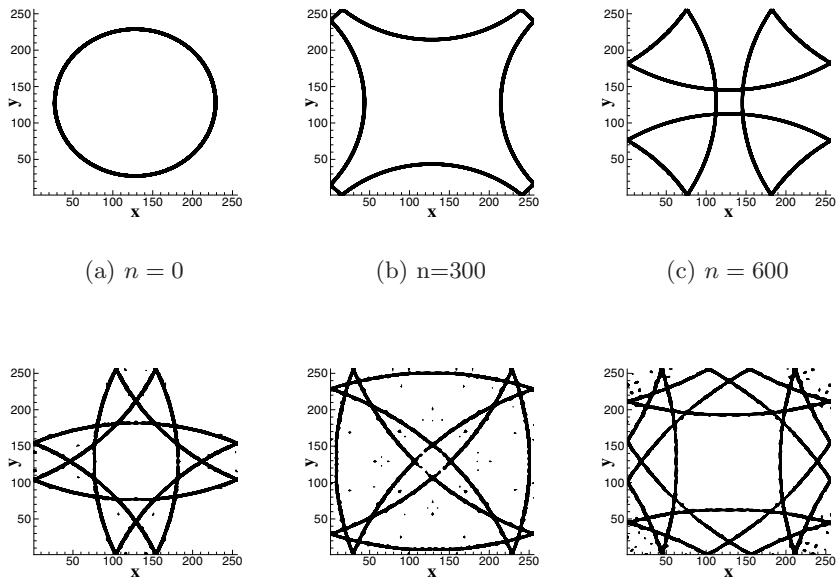
$$\frac{\epsilon}{\mu} = \frac{\left[1 + 2\cosh\left(\gamma h \cos\theta\right) + 2\cosh\left(\alpha h \sin\theta\right)\right]}{N},$$

where we take the number of grid points in the sum, N (central and nearest neighbors) to be 5.

32.3 Results

32.3.1 Propagation/Planar Reflection

A circular wave propagating inside a 2D square domain with reflecting boundaries is shown in Figure 32.4. It is obvious that the wave does not deteriorate



(a) $n = 0$ (b) n=300 (c) $n = 600$

(d) $n = 900$ (e) $n = 1200$ (f) $n = 1500$

Fig. 32.4. 2D wave equation solution.

even after many reflections. Most discretization effects should appear as a deviation from circular symmetry, since the grid is Cartesian. No such effects appear, to plottable accuracy. Ray tracing techniques also suffer from numerical dissipation as interpolation techniques have to be used to add and fit markers to approximate a continuous wave as the curve lengthens. The method we use is very different from converging methods, as we are converging the local function of the grid to 2–3 grid cells. So, in the fine grid limit,

it remains spread over the same number of grid cells and requires no logic to treat or "fit in" marker points. It can be seen in Figure 32.4 that there are also no curvature effects to plottable accuracy. Also, as in 1D, there is no discernable interaction between intersecting waves. The waves retain their form and orientation in spite of multiple head-on collisions. Long-distance propagation in 3D is simulated for an expanding, initially spherical wave and is shown in Figure 32.5. The computation is done on a coarse, 64^3 cell grid

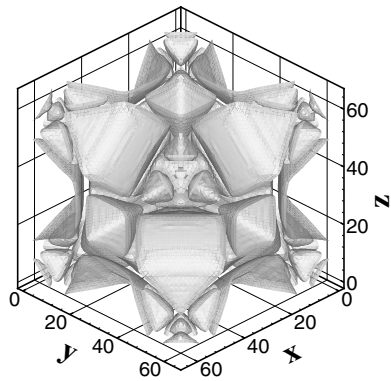


Fig. 32.5. 3D wave equation.

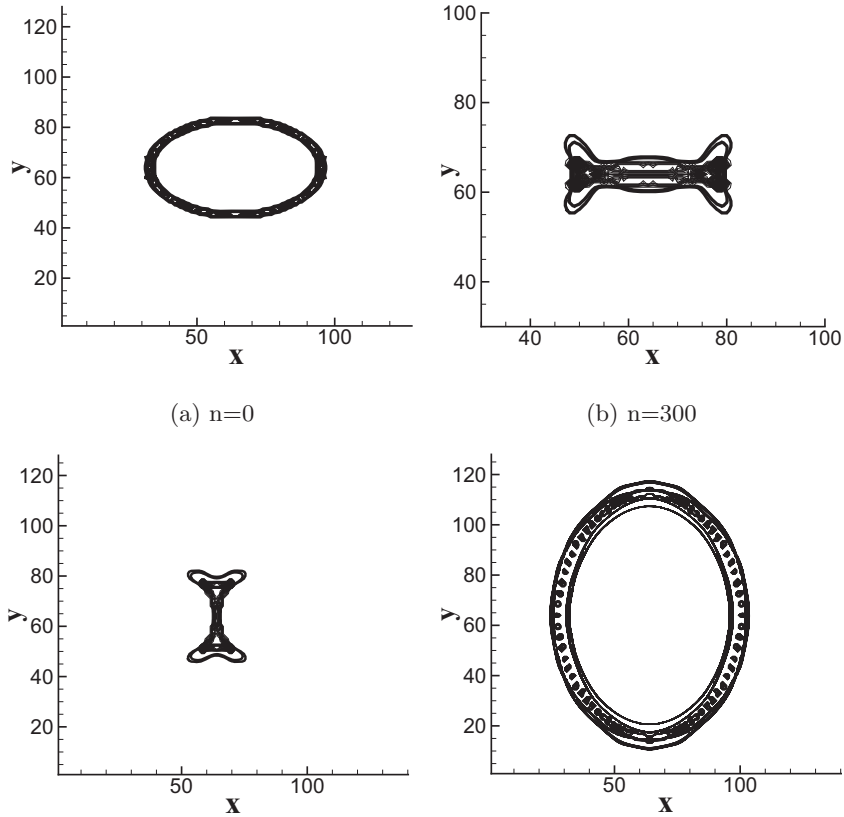
with periodic boundary conditions. The initial diameter for this computation is 16 grid cells. Ray tracing techniques become highly complicated in 3D when there are multiple intersecting surfaces. Robust interpolation methods then have to be used to fit the surfaces.

32.3.2 Focusing Waves

WC is also applied to converging/focusing waves (also in 2D). Here, there are a number of conserved variables which allow the propagation of waves through the focusing regions and automatically reconstruct the waves after focusing. A focusing elliptical wave front is computed and is displayed as amplitude contours. The elliptical centroid line is initially

$$\left(\frac{i_0}{a}\right)^2 + \left(\frac{j_0}{b}\right)^2 = 1,$$

where (i_0, j_0) is the center of the domain, $a = 32$ and $b = 20$. For an accuracy check, the results are compared to the results of a ray tracing, using Lagrangian markers. In Figure 32.6, amplitude contours are compared to Lagrangian markers (these are exact solutions in the high frequency approximation for the specific marker location). It can be seen that the basic information



(a) n=0

(b) n=300

(c) n=600

(d) n=900

Fig. 32.6. Focusing elliptical wave.

defined by the initial conditions is not lost, even though only the simple Eulerian algorithm was used, with no additional logic, and the grid was not fine enough to resolve the focusing. Since the interest is in the long-distance propagation (after the focusing), the detailed resolution at the focusing itself is not an issue.

32.3.3 Varying Index of Refraction

Another important study involves the pulse speed in nonuniform index of refraction fields. An initially straight pulse (again 2D), propagating through a region with refractive index defined as

$$\nu_j = 0.5/(1 + e^{-0.001(j-j_0)^2})$$

is simulated by means of the equation

$$\phi_{i,j}^{n+1} = 2\phi_{i,j}^{n} - \phi_{i,j}^{n-1} + \bigtriangledown^2 \left(\nu^2 \phi_{i,j}^{n}\right) + \delta_n \bigtriangledown^2 F_{i,j}^{n},$$

where $\bigtriangledown^2 = \delta_i^2 + \delta_j^2$. It is observed that the isolated pulse trajectory is correct with no diffusion or dispersion when compared to accurate ray tracing computations. It is also seen that information is not lost in spite of a limited density of grid points across the focusing regions. It can be further seen that in the far field, ray tracing techniques cannot continue to describe the wave as a smooth surface. Also, unlike ray tracing schemes, which suffer from scarcity of grid nodes in the far field, WC can still capture waves as smooth surfaces without complex logic involving allocation of new markers and interpolation. A comparison is shown in Figure 32.7, in which the smooth contours are calculated by the confinement method and compared with ray tracing (depicted as "blobs").

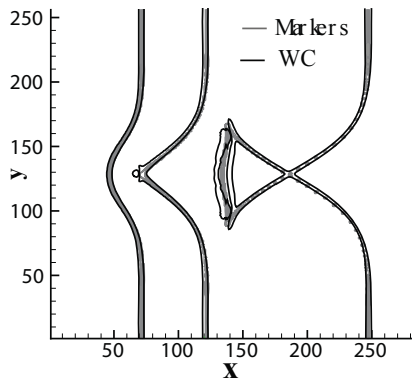


Fig. 32.7. Propagation of plane wave through regions of varying index of refraction.

32.4 Conclusion

A method, "wave confinement," has been described which propagates thin wave equation pulses. The pulses are solutions to the scalar wave equation with an added nonlinear term. When discretized on an Eulerian grid, the pulse solutions are represented as thin "shells" in 3D and wavefront "curves" in 2D, which remain only 2–3 grid cells thick, and propagate indefinitely as nonlinear solitary waves with no numerical spreading. As such, they serve to accurately transport total amplitude, integrated along a normal, and compute arrival times at each grid point. These pulses can accurately propagate through focusing and varying index of refraction regions and reflect from complex boundaries.

Acknowledgement. This work was mainly supported by the AFOSR under Dr. Arje Nachman.

References

[CaHi83] Cahn, J.W., Hilliard, J.E.: Free energy of a nonuniform system. *J. Chemical Phys.,* **28**, 258–267 (1983).

[La58] Lax, P.: Hyperbolic systems of conservation laws. II. *Comm. Pure Appl. Math.,* **10**, 537–566 (1958).

[RoHySt07] Rosenau, P., Hyman, M.J., Stanley, M.: Multidimensional compactons. *Phys. Review Lett.,* **98**, article 24101 (2007).

[Sm83] Smolarkiewicz, P.A.: A simple positive definite advection scheme with small implicit diffusion. *Monthly Weather Revue,* **111**, 479–486 (1983).

[StWePu95] Steinhoff, J., Wenren, Y., Puskas, E.: Computation of short acoustic pulses, in *Proceedings Sixth Internat. Symp. CFD,* (1995).

[StDiHa] Steinhoff, J., Dietz, W., Haas, S., Xiao, M., Lynn, N., Fan, M.: Simulating small scale features in fluid dynamics and acoustics as nonlinear solitary waves. *AIAA,* 2003–0078 (2003).

33

High-Performance Computing for Spectral Approximations

P.B. Vasconcelos,[1] O. Marques,[2] and J.E. Roman[3]

[1] Universidade do Porto, Portugal; pjv@fep.up.pt
[2] Lawrence Berkeley National Laboratory, Berkeley, CA, USA; oamarques@lbl.gov
[3] Universidad Politécnica de Valencia, Spain; jroman@dsic.upv.es

33.1 Introduction

In this chapter, we focus on the numerical solution of large eigenvalue problems arising in finite-rank discretizations of integral operators.

Let X be a Banach space over $\mathbb{C}$ and T a compact linear operator defined on X. We aim to solve numerically the eigenvalue problem

$$T\varphi = \lambda\varphi,$$

with λ nonzero and φ defined in X. Approximations λ_m and φ_m for the spectral elements of the integral operator can be obtained by solving

$$T_m\varphi_m = \theta_m\varphi_m,$$

where (T_m) is a sequence of finite-rank operators converging to T [AhLa01]. By evaluating the projected problem on a specific basis function, it is reduced to a matrix spectral problem

$$A_m x_m = \theta_m x_m \tag{33.1}$$

for a finite matrix A_m [AhLa06].

In what follows, we review the numerical computation of solutions of (33.1) using state-of-the-art numerical methods implemented in publicly available software packages, assembling previous results from [VaMa08] and [VaRo08]. Emphasis is given to parallel strategies provided by ScaLAPACK [BlCh97] and SLEPc [HeRo05]. Numerical experiments are performed on a weakly singular integral operator, where A_m is large and banded. Direct methods for the computation of the whole spectrum and iterative methods to compute a (small) set of eigenpairs will be presented.

The chapter is structured as follows. Section 33.2 presents the computer architecture specifications used for the performance results. Section 33.3 includes

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_33,

a brief discussion of the software strategies to tackle this kind of problem. An illustrative example is then given in Section 33.4. Section 33.5 presents some numerical results along with some conclusions and insights on how to solve similar problems.

33.2 Hardware: Parallel Machines

The numerical tests used for analyzing the performance of the codes have been carried out on four computing platforms.

Bassi is an IBM p575 POWER 5 system located at NERSC, the US National Energy Research Scientific Computing Center. It is a distributed memory computer with 888 processors. The processors are distributed among 111 compute nodes with 8 processors at 1.9 GHz per node. Processors on each node have a shared memory pool of 32 GBytes. The compute nodes are connected to each other with a high-bandwidth, low-latency switching network. Each node runs its own full instance of the standard AIX operating system. Jacquard, also located at NERSC, is an AMD Opteron cluster with 356 dual-processor nodes, 2.2 GHz processors, 6 GB of memory per node, interconnected with a high-speed InfiniBand network.

GridUP is an AMD Opteron 250 cluster with 24 dual-processor nodes, with a total of 48 processors at 2.4 GHz. The nodes have 4 GB of memory each, and they are interconnected via Gigabit Ethernet network. This system belongs to Universidade do Porto and is part of the Portuguese national grid infrastructure.

The Odin cluster, located at Universidad Politécnica de Valencia, has 55 nodes with dual Pentium Xeon processor at 2 GHz with 1 GB of memory per node, with a total of 110 processors. The nodes are interconnected with a high-speed SCI network with 2D torus topology.

33.3 Software: Numerical Methods and Libraries

In order to compute the solution of large scale eigenvalue problems on parallel computers, we can develop a parallel program using MPI, the message passing interface standard for programming distributed-memory parallel computers [GrLu99], and make use of parallel libraries that are based on that paradigm. Two types of numerical strategies are available: direct and iterative methods. Here we employ direct methods implemented in the ScaLAPACK library (Scalable Linear Algebra PACKage [BlCh97]), and iterative methods implemented in SLEPc, the Scalable Library for Eigenvalue Problem Computations [HeRo05], to compute a few eigenpairs. These are open source software packages available in the ACTS Collection of the US Department of Energy (DOE) [DrMa05].

ScaLAPACK does not currently provide any expert driver for the eigensolution of band matrices. This means that the problem must be treated as a general one. In the symmetric case, one can call the expert driver subroutine *pdsyevx* that handles the reduction to tridiagonal form followed by the computation of the eigendecomposition of the generated tridiagonal. In the nonsymmetric case, the user has to explicitly call subroutines for the two steps: first reduction of the matrix to Hessenberg form by calling the subroutine *pdgehrd* (which applies orthogonal similarity transformations) and then computing the Schur decomposition of the Hessenberg form by calling subroutine *pdlahqr* (which uses the QR algorithm).

Concerning SLEPc, it provides a number of eigensolvers that are appropriate for large sparse eigenproblems in which only part of the spectrum is required. We used a version of the Krylov–Schur method [St01], which is a faster variant of the Arnoldi algorithm. Additionally, SLEPc allows the transparent use of other eigensolver libraries such as ARPACK [LeSo98] and PRIMME [St07]. In order to enhance convergence of the iterative eigensolvers, SLEPc provides a built-in implementation of the shift-and-invert spectral transformation technique. For the solution of the linear systems involved, the user can employ different solvers such as GMRES and different preconditioners, including those provided by external libraries (e.g., hypre [FaYa02]).

Some platforms have specifically tuned versions of some libraries. On Bassi we used the BLAS available in PESSL (Parallel Engineering and Scientific Subroutine Library), a mathematical subroutine library from IBM designed to provide high performance for numerically intensive computing jobs running on IBM systems. It is IBM's parallel analogue of its serial library ESSL. Although PESSL contains a subset of ScaLAPACK, we have used our own full installation of ScaLAPACK. On other computer architectures, optimized libraries exist such as MKL for Intel processors and ACML for AMD processors (as is the case for Jacquard).

33.4 Illustrative Example

We consider an eigenvalue problem, issued from a real application [Ru04] and [AhAl02], where the integral operator $T : X \to X$, $X = L^1([0, \tau^\star])$, is defined by

$$(T\varphi)(\tau) = \frac{\varpi}{2} \int_0^{\tau^\star} E_1\left(|\tau - \tau^\star|\right) \varphi(\tau') d\tau, \quad \tau \in [0, \tau^\star] \,.$$

The kernel of the integral operator, which is weakly singular, is defined through the first exponential integral function

$$E_1(\tau) = \int_1^\infty \frac{\exp(-\tau\mu)}{\mu} d\mu, \quad 0 < \tau \leq \tau',$$

and depends on the albedo, $\varpi \in [0,1]$.

Defining a grid of m points, $0 = \tau_{m,0} < \tau_{m,1} < \cdots < \tau_{m,m} = \tau^\star$, we build a finite-dimensional subspace given by $X_m = \text{span}\{e_{m,j} : j = 1, \ldots, m\}$, where $e_{m,j} = 1$ if $\tau \in]\tau_{m,j-1}, \tau_{m,j}[$ and 0 otherwise. Defining the projection

$$\pi_m \varphi = \sum_{j=1}^{m} \langle \varphi, e^*_{m,j} \rangle e_{m,j},$$

where

$$\langle \varphi, e^*_{m,j} \rangle = \frac{1}{\tau_{m,j} - \tau_{m,j-1}} \int_{\tau_{m,j-1}}^{\tau_{m,j}} \varphi(\tau') d\tau',$$

we compute

$$T_m \varphi = \pi_m T \varphi.$$

Finally, the spectral problem for the finite-rank operator is reduced to a spectral problem for an $m \times m$ matrix by considering $A_m(i,j) = \langle T e_{m,j}, e^*_{m,i} \rangle$, [AhLa06].

For the tests, we consider $\varpi = 0.75$, $\tau^\star = 8000$, and m ranging from 8000 to $64,000$. The larger values of m are only possible using iterative methods on parallel machines and with appropriate linear algebra kernels to deal with sparse matrices.

The coefficients of A_m decay in magnitude significantly from the diagonal, and for practical purposes A_m can be considered a band matrix. For a fixed $\tau^\star$, the bandwidth increases with larger values of m. For very large values of m such as $32,000$ or $64,000$, the storage and computational effort required for building the matrix and solving the problem is high. For practical purposes, this implies the use of an increasing number of processors.

33.5 Numerical Results

In this section, we present some results for the different approaches described above. We used a relative error $\varepsilon \leq 10^{-12}$ for the iterative method in order to obtain solutions as "accurate" as the direct method and therefore perform a more realistic comparison of the algorithms' computational performance.

Let us begin by considering nonuniform grids on the interval $[0, \tau^\star]$, leading to nonsymmetric matrices.

In Table 33.1, we present the times for the computation of the full spectrum of matrix A_{8000} with the direct methods provided by ScaLAPACK in two machines, Bassi and GridUP. The generation of A_{8000} on the GridUP machine was much faster than on Bassi, as a direct consequence of the faster processor. In contrast, the timings for ScaLAPACK were better in Bassi, and with greater speedup (the ratio of the computing time with one processor over

Table 33.1. ScaLAPACK timings (seconds) on Bassi and on GridUP for nonsymmetric A_{8000}: number of processors (p), generation of the matrix (GEN), reduction to Hessenberg form (*pdgehrd*), and Schur decomposition (*pdlahqr*), on up to 16 processors.

	Bassi			GridUP		
p	GEN	*pdgehrd*	*pdlahqr*	GEN	*pdgehrd*	*pdlahqr*
1	308.4	1504.4	2389.9	196.25	5192.1	7315.5
2	154.1	875.7	1801.2	97.90	4955.1	5569.5
4	76.9	466.7	1235.4	48.71	3614.8	4609.7
8	38.6	235.8	908.8	23.96	2072.6	2567.6
16	19.3	108.3	834.8	11.88	1453.9	5659.8

the computing time with p processors). The reason for this is that on Bassi we used the optimized BLAS from the PESSL library.

For larger values of m, it is very computationally expensive to obtain the whole spectrum. To have access to a subset of the eigenvalues, we can use iterative methods, such as those provided by SLEPc.

Table 33.2 shows execution times for different matrix sizes on Odin. The

Table 33.2. SLEPc timings (seconds) on Odin for nonsymmetric A_m: number of processors (p), generation of the matrix (GEN), 5 largest eigenpairs using Krylov–Schur method, for several values of m on up to 32 processors.

m	p	GEN	Krylov–Schur
8000	1	522.21	110.66
	2	214.57	72.40
	4	91.41	38.44
	8	35.23	21.60
	16	17.61	16.25
	32	7.08	29.77
32,000	1	3964.10	1907.42
	2	2020.84	936.54
	4	1070.76	415.12
	8	542.73	237.47
	16	277.57	158.03
	32	143.40	123.44
64,000	2	8474.29	–
	4	4565.43	1355.83
	8	2397.99	650.20
	16	1225.71	312.64
	32	665.30	201.70

results show good speedup for all values of m, both for matrix generation and eigencomputation, except for the smallest value of m with 32 processors (not enough computational work per processor compared to communication time).

One can infer that the code is reasonably scalable because a good speedup is maintained when the number of processors grows together with an increase of the problem size. The missing values correspond to cases where there is not enough physical memory to store the matrix and other auxiliary data ($m = 64,000$ with one and two processors).

We also compared the performance of Krylov–Schur to ARPACK. For this problem, the latter requires a larger Krylov subspace to converge, and this has a direct impact on the computation time. For instance, for $m = 8000$ using 16 processors, ARPACK took 78.57 seconds to compute the largest five eigenvalues.

The larger the value of m, the smaller the separation of the eigenvalues, so Krylov eigensolvers will have more difficulties. To cope with this, it is necessary to increase the dimension of the Krylov subspace, which means an increase in memory requirements.

Convergence can be improved if a reference value that is close to the wanted eigenvalues is known. Table 33.3 presents results for the shift-and-invert technique. These runs need much fewer vectors for the basis. In this case, the

Table 33.3. SLEPc timings (seconds) on Odin for nonsymmetric A_m: number of processors (p), 5 largest eigenpairs using Krylov–Schur method with shift-and-invert for several values of m on up to 32 processors. Linear systems are solved with GMRES and different preconditioners (*its* is the accumulated number of iterations of the linear systems).

m	p	Block Jacobi		AMG	
		time	*its*	time	*its*
32,000	1	23.03	46	136.50	201
	2	23.73	160	67.36	182
	4	83.89	1279	44.59	207
	8	211.16	7073	23.20	207
	16	n/c	n/c	18.90	244
	32	n/c	n/c	10.88	215
64,000	2	109.48	161	277.69	206
	4	316.73	1343	180.88	224
	8	765.75	7034	105.23	226
	16	n/c	n/c	65.91	258
	32	n/c	n/c	35.30	229

eigensolver is applied to the operator $(A_m - \varpi I)^{-1}$ to compute eigenvalues closest to ϖ. The inverse of $A_m - \varpi I$ is handled implicitly by solving linear systems within the eigensolver iterations. We have chosen to solve these linear systems with GMRES combined with two parallel preconditioners, as follows. Block Jacobi consists in computing an incomplete LU factorization without fill-in for each diagonal block (in our case, one block per processor). This preconditioner is easy to implement in parallel, but it loses efficiency when

the number of blocks is increased. This is the reason why in the case of 16 and 32 processors some linear solvers have reached the maximum number of allowed iterations, resulting in an insufficient accuracy in the computed eigenpairs (in the table, "n/c" indicates this circumstance). A preconditioner with a better scaling is the algebraic multigrid (AMG) preconditioner [HeYa02], as shown in the table.

We now consider the case where a regular grid induces a symmetric matrix. Symmetric eigenproblems are more common in practice, and more methods and software are available for this case.

Table 33.4 shows timings on Bassi for symmetric A_{8000}, for the computation of the full spectrum (eigenvalues only). As before, the generation of A

Table 33.4. ScaLAPACK timings (seconds) on Bassi for symmetric A_{8000}: generation of the matrix (GEN), and eigencomputation (*pdsyevx*), on up to 16 processors.

p	GEN	*pdsyevx*
1	310.0	418.7
2	154.5	478.8
4	77.4	322.1
8	38.8	165.2
16	19.4	79.8

scales well with an increasing number of processors. The scaling of the eigensolution phase is satisfactory for $p > 4$. Interestingly, the eigensolution phase takes more time on two processors than on one processor, which suggests a poor load balancing given the dimension of the matrix.

Table 33.5 shows timings on Jacquard for symmetric A of different sizes, for the computation of the five largest eigenvalues and corresponding eigenvectors. The scaling of the eigensolution phase is similar to the one observed on Bassi. A 2D block cyclic distribution for A might lead to a better performance, but we anticipate that for much larger matrices a direct method would be impractical. This is mainly because of the costs required for the reduction of A to tridiagonal form as part of the eigensolution strategy. If required, the computation of all eigenvectors of A from the eigenvectors of the tridiagonal would also add to the costs.

Table 33.6 shows results for iterative methods for the symmetric case, where in addition to Krylov–Schur we have considered a Davidson-type eigensolver implemented in PRIMME [St07]. For the generation of the matrix, we exploit symmetry and have to compute only half of the matrix elements. Furthermore, we optimize the generation by making use of a small software cache mechanism that stores recently computed integral values. In this way, we can avoid up to 75% computation time in some cases. This mechanism was also present for the nonsymmetric case, yet the percentage of cache hits was far smaller.

Table 33.5. ScaLAPACK timings (seconds) on Jacquard for symmetric A_m: generation of the matrix (GEN), and eigencomputation (*pdsyevx*), on up to 64 processors.

m	p	GEN	*pdsyevx*
4000	1	41.6	137.4
	2	21.2	271.4
	4	10.4	206.7
	8	5.2	128.4
	16	2.5	76.6
8000	2	84.5	2271.9
	4	41.7	1459.1
	8	20.8	881.8
	16	10.1	499.1
	32	5.1	285.2
16000	8	83.7	5511.1
	16	40.4	2879.4
	32	20.3	1656.9
	64	10.2	945.0
32000	32	81.4	11210.0
	64	40.7	6580.1

Table 33.6. SLEPc timings (seconds) on Odin for symmetric A_m: number of processors (p), generation of the matrix (GEN), 5 largest eigenpairs using Krylov–Schur method, and 5 largest eigenpairs using PRIMME, for several values of m on up to 32 processors.

m	p	GEN	Krylov–Schur	PRIMME
8000	1	39.61	135.68	40.79
	2	29.92	66.80	23.88
	4	16.86	34.17	13.02
	8	9.08	16.14	9.30
	16	4.68	27.40	20.38
	32	2.41	29.71	22.09
32000	1	1885.17	1446.70	799.78
	2	1403.00	550.65	351.01
	4	815.04	249.54	214.65
	8	435.45	160.43	102.33
	16	224.75	74.11	72.85
	32	114.57	69.91	48.13
64000	1	7913.90	–	2655.99
	2	6054.33	1784.94	1503.41
	4	3543.23	943.70	975.45
	8	1894.70	447.77	408.31
	16	975.14	236.56	355.60
	32	500.20	145.92	139.37

In these tests, most of the time PRIMME was faster in computing the eigenpairs than Krylov–Schur, particularly when the number of processors is small. As the number of processors grows, both solvers show similar performance. We must emphasize that these results were obtained with a smaller subspace dimension than in the nonsymmetric case.

Table 33.7 shows results for shift-and-invert with a symmetric matrix. Similarly to the nonsymmetric case, the AMG preconditioner is very effective and results in very good speedup, because the number of iterations required by the linear systems is moderate and almost constant for different numbers of processors. Block Jacobi still has difficulties for large numbers of processors, although in this case it was able to compute the solution to the required precision.

Table 33.7. SLEPc timings (seconds) on Odin for symmetric A_m: number of processors (p), 5 largest eigenpairs using Krylov–Schur method with shift-and-invert for several values of m on up to 32 processors. Linear systems are solved with GMRES and different preconditioners (*its* is the accumulated number of iterations of the linear systems).

m	p	Block Jacobi		AMG	
		time	*its*	time	*its*
32000	1	10.18	41	75.09	133
	2	12.43	153	34.52	136
	4	35.26	977	20.23	161
	8	91.48	4641	11.80	178
	16	77.20	8162	9.53	170
	32	73.99	12717	7.73	184
64000	1	63.96	46	231.18	122
	2	56.58	160	118.97	132
	4	120.67	974	62.58	139
	8	241.09	3963	33.86	151
	16	238.66	7455	24.76	175
	32	213.47	12831	15.82	185

References

[AhAl02] Ahues, M., d'Almeida, F.D., Largillier, A., Titaud, O., Vasconcelos, P.B.: An L^1 refined projection approximate solution of the radiation transfer equation in stellar atmospheres. *J. Comput. Appl. Math.,* **140**, 13–26 (2002).

[AhLa06] Ahues, M., Largillier, A., d'Almeida, F.D., Vasconcelos, P.B.: Defect correction for spectral computations for a singular integral operator. *Comm. Pure Appl. Anal.,* **5**, 241–250 (2006).

[AhLa01] Ahues, M., Largillier, A., Limaye, B.V.: *Spectral Computations with Bounded Operators,* CRC, Boca Raton, FL (2001).

[BlCh97] Blackford, L.S. et al.: *ScaLAPACK Users' Guide,* Society for Industrial and Applied Mathematics, Philadelphia, PA (1997).

[DrMa05] Drummond, L.A., Marques, O.A.: An overview of the advanced computational software (ACTS) collection. *ACM Trans. Math. Software,* **31**, 282–301 (2005).

[FaYa02] Falgout, R.D., Yang, U.M.: hypre: a library of high performance preconditioners, in *Lecture Notes in Computer Science,* **2331**, 632–641 (2002).

[GrLu99] Gropp, W., Lusk, E., Skjellum, A.: *Using MPI: Portable Parallel Programming with the Message-Passing Interface,* MIT Press, Cambridge, MA (1999).

[HeYa02] Henson, V.E., Yang, U.M.: BoomerAMG: a parallel algebraic multigrid solver and preconditioner. *Appl. Numer. Math. Trans. IMACS,* **41**, 155–177 (2002).

[HeRo05] Hernandez, V., Roman, J.E., Vidal, V.M.: SLEPc: a scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Trans. Math. Software,* **31**, 351–362 (2005).

[LeSo98] Lehoucq, R.B., Sorensen, D.C., Yang, C.: *ARPACK Users' Guide. Solution of Large-Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods,* Society for Industrial and Applied Mathematics, Philadelphia, PA (1998).

[Ru04] Rutily, B.: Multiple scattering theoretical and integral equations, in *Integral Methods in Science and Engineering: Analytic and Numerical Techniques,* Constanda, C., Ahues, M., Largillier, A. (eds.), Birkhäuser, Boston, 211–231 (2004).

[St07] Stathopoulos, A.: Nearly optimal preconditioned methods for Hermitian eigenproblems under limited memory. Part I: Seeking one eigenvalue. *SIAM J. Scientific Comput.,* **29**, 481–514 (2007).

[St01] Stewart, G.W.: A Krylov–Schur algorithm for large eigenproblems. *SIAM J. Matrix Anal. Appl.,* **23**, 601–614 (2001).

[VaMa08] Vasconcelos, P.B., Marques, O.: Comparison of parallel eigensolvers for a discretized radiative transfer problem, *Proc. Appl. Math. Mech.,* **7**, 1022805–1022806 (2008).

[VaRo08] Vasconcelos, P.B., Marques, O., Roman, J.E.: Parallel eigensolvers for a discretized radiative transfer problem, in *Lecture Notes in Computer Science,* **5336**, 336–348 (2008).

34

An Analytical Solution for the General Perturbed Diffusion Equation by an Integral Transform Technique

M.T. Vilhena, B.E.J. Bodmann, and I.R. Heinen

Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil;
vilhena@pq.cnpq.br, bardo.bodmann@ufrgs.br, heire@bol.com.br

34.1 Introduction

In the developments of nuclear energy, new reactor concepts are being proposed and explored, where innovative ideas need to be tested by means of simulations. Although the original neutron calculations start from a transport equation, many approaches reduce the calculation to diffusion equations, since the Boltzmann equation for neutron transport is still considered a challenge (see, for example, [Le05], [Se07], and the references therein). A detailed sequence, starting from a neutron transport equation (Boltzmann equation) until the reduction to a diffusion phenomenon using Fick's hypothesis, is given, for instance, in [Se07]. Our principal concern here is an effective analytical method for the general perturbed neutron diffusion equation by an integral transform technique. To this end, we present a procedure that allows us to construct an analytical solution of the multi-group neutron diffusion equation in Cartesian geometry using well-established integral transform procedures [He05]. Once the general structure of the solution is determined, we may directly calculate the neutron flux (which is an analytical expression), and the only quantity which is determined numerically at the end of the calculation is criticality. In what follows we present the procedure, considering a generic multi-group calculation for an arbitrary number of energy intervals. Due to the fact that the geometric extension of the reactor core is typically very much larger in one dimension compared to the other two length scales, we may cast the calculation into a two-dimensional (2D) setting.

34.2 The Multi-Group Diffusion Equation with Constraints

The general multi-group diffusion problem in two dimensions is given by the equation system

C. Constanda and M.E. Pérez (eds.), *Integral Methods in Science and Engineering, Volume 2: Computational Methods*, DOI 10.1007/978-0-8176-4897-8_34,

$$\mathbb{L}\mathbf{\Phi} = \mathbf{S}, \tag{34.1}$$

where $\mathbb{L}$ represents the local nonhomogeneous diffusion operator and includes particle multiplication from fission, $\mathbf{\Phi} = (\phi_1(x,y), \ldots, \phi_G(x,y))^T$ signifies the local multi-group neutron flux (in vector representation), and $\mathbf{S} = (S_1(x,y), \ldots, S_G(x,y))^T$ is a local multi-group neutron source, for energy groups $g \in \{1, \ldots, G\}$. The diffusion operator may be decomposed further into group-preserving and group-mixing terms: $\mathbb{L} = \mathbb{L}_P + \mathbb{L}_M$. The diagonal elements contain a local diffusion operator, with absorption and fission of the same energy group:

$$\mathbb{L}_P = \mathrm{diag}\left(\nabla D_1 \nabla + \Sigma_{a1} - \frac{\chi_1}{k_{eff}}\nu\Sigma_{f1}, \ldots, \nabla D_G \nabla + \Sigma_{aG} - \frac{\chi_G}{k_{eff}}\nu\Sigma_{fG}\right).$$

The nondiagonal elements of $\mathbb{L}_M$ contain fission and scattering terms:

$$(\mathbb{L}_M)_{gg'} = -\frac{\chi_{g'}}{k_{eff}}\nu\Sigma_{fg'} + \Sigma_{gg'}.$$

Here, $D_g = D_g(x,y)$ represents the local diffusion coefficient for energy group g, and $\Sigma_{ag}(x,y)$, $\Sigma_{fg'}(x,y)$, and $\Sigma_{gg'}(x,y)$ are the macroscopic position-dependent absorption, fission, and scattering cross sections, respectively. The weight factor ν is due to neutron multiplication from the fission process, χ_g is the integrated neutron spectrum from fission of group g, and k_{eff} is the effective multiplication factor measuring criticality.

The solutions obey the piecewise open surface boundary conditions defined by the neutron current density and scalar flux at the contours of the sheet. If Γ denotes the 2D volume and $\partial\Gamma$ the boundary, and if $\partial\Gamma_{xy}$ are the boundary pieces shown in Figure 34.1, then these conditions are

$$\left.\frac{\partial\phi_g}{\partial x}\right|_{\partial\Gamma_{0y}} = \left.\frac{\partial\phi_g}{\partial y}\right|_{\partial\Gamma_{x0}} = 0,$$
$$\phi_g|_{\partial\Gamma_{\bar{x}y}} = \phi_g|_{\partial\Gamma_{x\bar{y}}} = 0.$$

Since the problem has mirror symmetry with respect to the coordinate axes with either $x = 0$ or $y = 0$, it is sufficient to determine only the solution in the section with $x, y \geq 0$; the rest of it may be completed using the mentioned reflection symmetries. A further constraint breaks the scale invariance of the nonhomogeneous diffusion equation upon introduction of the energy release (E_R) per unit time of the sheet, which implicitly correlates the multiplication factor k_{eff} to the power through the total neutron flux:

$$P = E_R \int_\Gamma \sum_g \Sigma_{fg}\phi_g d\Gamma.$$

As it stands, (34.1) is unlikely to be solved in closed analytical form. In order to introduce a simplification which nevertheless permits us to control

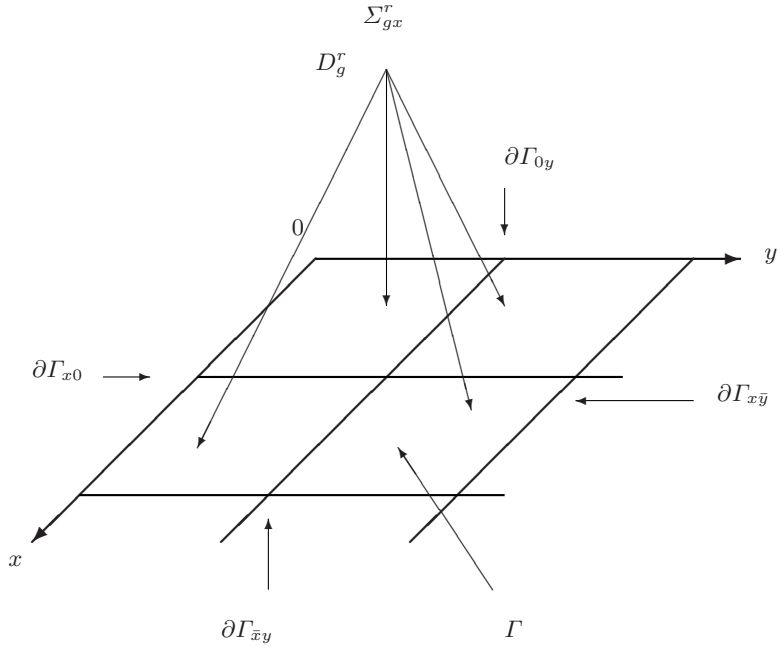


Fig. 34.1. The 2D sheet with boundaries, internal interfaces, and locally homogeneous physical coefficients.

convergence in a strict mathematical sense, we may make use of the physical resolution scale set by the inverse of the largest macroscopic cross-sectional value of the problem under consideration, and segment the sheet into several regions $r \in \{1, \ldots, R\}$, with linear dimension smaller than the mean free path. In Figure 34.1, we showed for simplicity an example with four patches disregarding this criterion, since our goal is to introduce the procedure and show how it works. A further comment with respect to convergence is in order here; mathematical convergence signifies that one may derive convergence criteria which exactly evaluate the quality of the solution which differs from the heuristic criteria usually used in numerical or stochastic methods.

We assume that the only neutron source is fission and, consequently, ignore the source term ($S(x, y) \to 0$). Besides the dependence on the specific energy group, the physical coefficients are now "locally" homogeneous, i.e., constant in a specific region r:

$$D_g(x, y) \to D^r_g, \quad \Sigma_{ag}(x, y) \to \Sigma^r_{ag},$$
$$\Sigma_{fg'}(x, y) \to \Sigma^r_{fg'}, \quad \Sigma_{gg'}(x, y) \to \Sigma^r_{gg'}.$$

The only quantity that preserves its original dependence on the coordinates is the scalar neutron flux, which is determined in its analytical form for each region ($\Phi(x, y) \to \Phi^r(x, y)$). Taking into account the modifications from above

and on multiplication of the operator $\mathbb{L}$ from the left by the matrix $\mathbb{D}_r^{-1}$ that contains the inverse multi-group diffusion constants of the respective region, we can rewrite the simplified equation (34.1) in a more compact form for each homogeneous region, namely,

$$(\Delta - \mathbb{W}_P^r + \mathbb{W}_M^r)\,\Phi^r = 0. \tag{34.2}$$

In addition to the boundary conditions, we have the piecewise open interface conditions, which combine the solutions of adjacent regions (r and r', respectively) into one unique solution for the whole problem. These conditions are

$$\mathbb{D}^r \nabla \Phi^r|_{\partial\Gamma_{xy}} = \mathbb{D}^{r'} \nabla \Phi^{r'}\Big|_{\partial\Gamma_{xy}},$$
$$\Phi^r|_{\partial\Gamma_{xy}} = \Phi^{r'}\Big|_{\partial\Gamma_{xy}}.$$

Except for the interface conditions, one may consider the total problem divided into smaller similar rescaled problems, each of them having the same solution structure but with different coefficients. Equation (34.2) together with the boundary and interface conditions define the problem to be solved analytically.

34.3 An Analytical Solution

The constant approximation for the physical parameters of each region together with a combination of a limited Laplace transform and a method called the generalized integral transform technique (GITT) [Co93, CoMi97], which splits the differential operator into eigenvalues and polynomials, allow us to apply standard methods of linear algebra and determine the analytical structure of the solution. Equation (34.2) is symmetric under the swap $x \leftrightarrow y$, so we may apply the GITT to the x degree of freedom and convert the remaining degree (y) by means of the Laplace transformation. As a first step towards the decomposition of (34.2), the scalar flux may be replaced by an expansion of the form

$$\phi_g^r = \sum_{i=0}^{\infty} \xi_{gi}^r(x)\eta_{gi}^r(y).$$

If there were only one energy group and the problem were one dimensional, then (34.2) would assume the form of a Sturm–Liouville problem. Hence, we may think of the terms $\xi_{gi}^r(x)$ as representing a linearly independent functional basis which, because of similarity of the structure of the equations, may be determined from the auxiliary problem, i.e., the Sturm–Liouville problem. The principal idea of GITT is then to substitute differential operators by eigenvalues of that auxiliary problem with known analytical solutions. This auxiliary problem satisfies the same boundary conditions as the original problem in

order to minimize the dimension of the functional basis (the eigenfunctions) it supplies for each eigenvalue. More specifically, the solutions of the Sturm–Liouville problem with nonzero eigenvalues $\lambda_i = \frac{(2i-1)\pi}{2(b_r-a_r)} \neq 0$ satisfy the same boundary conditions as the total problem, that is,

$$(\partial_x)^2 \xi_i^r + \lambda_i^2 \xi_i^r = 0, \quad \partial_x \xi_i^r|_{a_r} = 0, \quad \xi_i^r|_{b_r} = 0.$$

In order to adjust the solutions at the interface and take into account deviations of the interface conditions from the boundary conditions of the total problem, a nonorthogonal but linearly independent solution of the Sturm–Liouville problem with zero eigenvalue ($\lambda_0 = 0$) is added. Thus, the structure of the solution is the same as the one for a totally homogenized problem except for an additional linear function with coefficients to be determined from the boundary or interface conditions. Note that, by this procedure, interfaces and boundaries are determined with the same technique. The orthogonality property of the basis of the subspace (with nonzero eigenvalue) offers a way to decouple the equation into a set of independent equations. Next, the orthogonal basis is the same for all energy groups, so that the coefficients that differentiate the solutions for each energy group are absorbed in the η functions.

The differential operator with respect to y may be eliminated by the use of the limited Laplace transform $\mathcal{L}^r[\eta(y)] = \tilde{\eta}^r(s)$, defined within the limits of each region. Then the derivative term is

$$\mathcal{L}^r[(\partial_y)^2 \eta(y)] = s^2 \tilde{\eta}^r + s\Upsilon_1^r + \Upsilon_0^r,$$

which substitutes all terms containing degrees of freedom along y. Here, the Υs play a role analogous to that of the linear functions of the Sturm–Liouville problem and take care of the matching of the solutions at the boundaries and interfaces. Upon insertion of the expansion and application of the Laplace transformation, we arrive at an equation that, in component notation for $\mathbb{W}_P$ and $\mathbb{W}_M$, is of the form

$$\sum_{i=0}^{\infty} \left(\left(s^2 - (\lambda_i^r)^2 - (\mathbb{W}_P^r)_g \right) \tilde{\eta}_{gi}^r + \sum_{g'} (\mathbb{W}_M^r)_{gg'} \tilde{\eta}_{g'i}^r + s\Upsilon_{g1i}^r + \Upsilon_{g0i}^r \right) \xi_i^r = 0.$$

We may now use the projection operator $\int_{a_r}^{b_r} dx[\xi_i^r]$ with $i \neq 0$ to decompose equation (34.2) into a set of separate equations, which depend only on the y-dual variable s. For convenience, we introduce the notation

$$s\Upsilon^r_{g1j} + \Upsilon^r_{g0j} = Y_{jg},$$

$$\int_{a_r}^{b_r} \xi^r_j \xi^r_i dx = \delta_{ij} N^r_j \quad \text{for } i, j \neq 0,$$

$$\int_{a_r}^{b_r} \xi^r_j \xi^r_0 dx = \delta_{ij} A^r_j \quad \text{for } j \neq 0,$$

$$\int_{a_r}^{b_r} \xi^r_0 \xi^r_0 dx = B^r,$$

which modifies the equation set into a system for each j with generic structure (rows and columns refer to the energy groups)

$$\begin{pmatrix} \bullet & \bullet & \bullet & \cdots & \bullet \\ \bullet & \bullet & 0 & 0 & 0 \\ \bullet & 0 & \bullet & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ \bullet & 0 & 0 & 0 & \bullet \end{pmatrix} \begin{pmatrix} \bullet \\ \bullet \\ \bullet \\ \vdots \\ \bullet \end{pmatrix} + \begin{pmatrix} \bullet \\ \bullet \\ \bullet \\ \vdots \\ \bullet \end{pmatrix},$$

where the blobs indicate nonzero elements and all other components are zero. In terms of the specific expressions, the resulting equation system is

$$A^r_j \left(\left(s^2 - (\mathbb{W}^r_P)_g\right) \tilde{\eta}^r_{g0} + \sum_{g'} (\mathbb{W}^r_M)_{gg'} \tilde{\eta}^r_{g'0} + Y^r_{g0} \right)$$
$$+ N^r_j \left(\left(s^2 - \left(\lambda^r_j\right)^2 - (\mathbb{W}^r_P)_g\right) \tilde{\eta}^r_{gj} + \sum_{g'} (\mathbb{W}^r_M)_{gg'} \tilde{\eta}^r_{g'j} + Y^r_{gj} \right) = 0.$$

Let $\mathbb{W}^r_i$ be the matrix in the equation above, containing $s^2, \lambda^2_i, \mathbb{W}^r_P$, and $\mathbb{W}^r_M$. Using the linear independence, the solutions $\tilde{\eta}^r_{gi}$ may be determined simply by Cramer's rule; that is,

$$\tilde{\eta}^r_{gi} = -\frac{\det(\mathbb{W}^r_{ig})}{\det(\mathbb{W}^r_i)},$$

where $\mathbb{W}^r_{ig}$ signifies the modified matrix with the gth column replaced by the inhomogeneity $\mathbf{Y}^r_i = (Y^r_{1i}, \ldots, Y^r_{Gi})^T$ of the matrix equation. The Laplace-transformed factors are rational functions in s, so that the solution may be obtained from the Heaviside expansion, where the sum runs over the roots of $\det(\mathbb{W}^r_i)$ in s,

$$\eta^r_{gi}(y) = \sum_{j=1}^{p} \left. \frac{\det(\mathbb{W}^r_{ig})}{\frac{\partial}{\partial s}\det(\mathbb{W}^r_i)} \right|_{s=s_i} e^{s_j y} .$$

Recalling that the solutions $\eta^r_{gi}(y)$ still contain terms due to $\Upsilon^r_{(0,1)}$ from the limited Laplace transform, we eliminate these unknowns using the boundary and interface conditions in y and integrating out the second dimension in x:

$$\sum_i \left(D_g^r \int \xi_i^r dx \, \frac{\partial \eta_{gi}^r}{\partial y}(b_r) - D_g^{r'} \int \xi_i^{r'} dx \, \frac{\partial \eta_{gi}^{r'}}{\partial y}(a_{r'}) \right) = 0,$$

$$\sum_i \left(\int \xi_i^r dx \, \eta_{gi}^r(b_r) - \int \xi_i^{r'} dx \, \eta_{gi}^{r'}(a_{r'}) \right) = 0.$$

The remaining missing pieces are the solutions ξ_0^r of the Sturm–Liouville problem with zero eigenvalue. For the solutions constituting an orthogonal basis per region, the boundary conditions are the same as for the total problem with boundary $\partial\Gamma_{x0} \cup \partial\Gamma_{\bar{x}y} \cup \partial\Gamma_{x\bar{y}} \cup \partial\Gamma_{0y}$. The only terms that are needed to match the local solutions at the interfaces are the linear terms. Thus,

$$\sum_i \left(D_g^r \frac{\partial \xi_i^r}{\partial x}(b_r) \int \eta_{gi}^r dy \right) - D_g^{r'} \frac{\partial \xi_0^{r'}}{\partial x}(a_{r'}) \int \eta_{g0}^{r'} dy = 0,$$

$$\xi_0^r(b_r) \int \eta_{g0}^r dy - \sum_i \left(\xi_i^{r'}(a_{r'}) \int \eta_{gi}^{r'} dy \right) = 0.$$

Except for k_{eff}, the solution is known in closed analytical form. Since the multiplication factor enters the solution ϕ in a way that in general does not permit inversion, the integral that relates the power of the sheet to the neutron flux may not be cast into a form that allows us to solve for k_{eff} explicitly, so that one has to resort to a numerical procedure, which takes place at the end of the solving procedure. Thus the only nonanalytical step is the determination of the numerical value of k_{eff}.

34.4 Conclusion

This chapter presented a new method, which generates analytical solutions for the globally heterogeneous problem of neutron diffusion in two dimensions. The principal steps employed are the Laplace transform and the GITT. Motivated by recent developments in reactor concepts, we developed an effective procedure which permits to analyze in an analytical way what changes in the reactor core geometry or composition occur and can lead to an optimized setup. Since the only quantity determined by numerical means is the effective multiplication factor, the quality of the solution may be controlled by mathematical convergence criteria. A detailed analysis compares the error from numerical methods such as the finite difference method, which typically scales with the step size, to our procedure with an error that depends on the truncation of the expansion and the region size determined by a scale which may be determined from the largest macroscopic cross section present in the problem.

Although algebraic manipulations are typically slower in execution than numerical procedures, in the present approach, because the homogenized

global problem has the same solution as the rescaled smaller problem, restricted to a specific region except for the differences imposed by the interface conditions, which are taken care of by a linear correction of the solution, the structure of the solution is the same for all regions and can also be applied to the outer regions which are limited partially by the outer boundary $\partial\Gamma$. Once the number of energy groups and regions is defined, and further the truncation of the expansion is determined, then one may prepare a library of solutions using the proposed method. The only task to be executed then is to determine numerically the GITT eigenvalues and coefficients from the Laplace transform and solve the power integral for the effective multiplication factor. We believe that in order to get a comparable precision with numerical or stochastic procedures will be more time consuming, especially if modifications in geometry and material composition are to be examined.

References

[Co93] Cotta, R.M.: *Integral Transforms in Computational Heat and Flow,* CRC Press, Boca Raton, FL (1993).

[CoMi97] Cotta, R.M., Mikhailov, M.D.: *Heat Conduction: Lumped Analysis, Integral Transforms, Symbolic Computation,* McGraw-Hill, Chichester, UK (1997).

[He05] Heinen, I.R.: Master thesis. Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil (2005) (Portuguese).

[Le05] Leppänen, J.: A new assembly-level Monte Carlo neutron transport code for reactor physics calculations, in *Mathematics and Computation, Supercomputing, Reactor Physics and Nuclear and Biological Applications,* American Nuclear Society, LaGrange Park, IL (2005) (on CD-ROM).

[Se07] Sekimoto, H.: *Nuclear Reactor Theory. Part II,* COE-INES Tokyo Institute of Technology (2007).

Index