MONOGRAFÍAS MATEMÁTICAS
"GARCÍA DE GALDEANO"

2010

N.º 35

Tenth International Conference Zaragoza-Pau on Applied Mathematics and Statistics

M. C. López de Silanes
M. Palacios
G. Sanz
J. J. Torrens
M. Madaune-Tort
C. Paroissin
D. Trujillo
(Editors)

MONOGRAFÍAS MATEMÁTICAS
GARCÍA DE GALDEANO

Número **35**, 2010

Tenth International Conference Zaragoza-Pau on Applied Mathematics and Statistics

Jaca (Spain), September 15–17, 2008

Editors

M. C. López de Silanes
M. Palacios
G. Sanz
Universidad de Zaragoza, Spain

J. J. Torrens
Universidad Pública de Navarra, Spain

M. Madaune-Tort
C. Paroissin
D. Trujillo
Université de Pau et des Pays de l'Adour, France

JORNADAS ZARAGOZA–PAU DE MATEMÁTICA APLICADA Y ESTADÍSTICA (10ª. Jaca 2008. Jaca)

Tenth International Conference Zaragoza–Pau on Applied Mathematics and Statistics : Jaca, Spain, September 15 – 17, 2008 / editors M. C. López de Silanes… [et al.]. — Zaragoza : Prensas Universitarias de Zaragoza : Instituto Universitario de Matemáticas y Aplicaciones, Universidad de Zaragoza, 2010

XXX, 302 p. ; 25 cm. — (Monografías Matemáticas García de Galdeano ; 35)
ISBN 978-84-15031-53-6
Matemáticas-Congresos y asambleas
LÓPEZ DE SILANES, M. C.
51(063)

Monografías Matemáticas García de Galdeano nº 35
Septiembre 2010
Universidad de Zaragoza

Imprime: Servicio de Publicaciones. Universidad de Zaragoza

D.L.: Z-3228-2010
ISBN: 978-84-15031-53-6

The edition of this volume has been partially subsidized by the Vicerrectorado de Investigación de la Universidad de Zaragoza

X Jornadas ZARAGOZA - PAU X Journées

Tenth International Conference Zaragoza-Pau on Applied Mathematics and Statistics

Residencia Universitaria de Jaca, September 15 – 17, 2008

http://pcmap.unizar.es/~jaca2008

Departamentos de Matemática Aplicada
y de Métodos Estadísticos
Universidad de Zaragoza

Información:
M.C. López de Silanes
Tel.: 976 76 19 86
mcruz@unizar.es

M. Palacios
Tel.: 976 76 19 81
mpala@unizar.es
Departamento de Matemática Aplicada

G. Sanz
Tel.: 976 76 28 60
gerardo@unizar.es
Departamento de Métodos Estadísticos

Laboratoire de Mathématiques et leurs Applications
UMR CNRS 5142
Université de Pau et des Pays de l'Adour

Renseignements:
M. Madaune-Tort
Tel.: 559 40 75 19
monique.madaune-tort@univ-pau.fr

C. Paroissin
Tel.: 559 40 75 69
cristian.paroissin@univ-pau.fr

D. Trujillo
Tel.: 559 40 75 60
david.trujillo@univ-pau.fr
Laboratoire de Mathématiques et leurs Applications de Pau

Contents

Applied Mathematics 1

Statistics 253

Preface

The *International Conference Zaragoza-Pau on Applied Mathematics and Statistics* is organized every two years by the *Departamento de Matemática Aplicada*, the *Departamento de Métodos Estadísticos*, both from the *Universidad de Zaragoza* (Spain), and the *Laboratoire de Mathématiques Appliquées*, from the *Université de Pau et des Pays de l'Adour* (France). The aim of this conference is to present recent works in Applied Mathematics and Statistics, putting special emphasis on subjects linked to petroleum engineering and environmental problems.

The Tenth Conference took place in Jaca (Spain) from 15th to 17st September 2005. Breaking for the first time the two-year periodicity of the Conference, its tenth edition did not take place in 2007, since two related events filled the calendar in this year. The official opening ceremony was graced by the presence of the Chancellor of the University of Zaragoza, Excmo. Sr. Rector Mgfco. D. Manuel J. López Pérez, and the Chancellor of the University of Pau, M. le Président Jean-Louis Gout. During those three days, 111 mathematicians, coming from different universities, research institutes or the industrial sector, attended 10 plenary lectures, 52 contributed talks and a poster session with a total of 11 posters. The principal topics were: theoretical and numerical analysis of deterministic models described by differential equations, statistics and stochastic processes, surface approximation and image analysis. At the same time, there was also a session devoted to Algebra and Geometry. These proceedings contain 2 papers based on the corresponding invited lectures along with 30 full length refereed research papers.

We would like to thank the following institutions for their regular financial and material support in our cooperation programs: Université de Pau et des Pays de l'Adour, Universidad de Zaragoza, Conseil Régional d'Aquitaine, Gobierno de Aragón, Conseil Régional de Midi-Pyrénées, Gobierno de Navarra, and Pyrenean Work Community. Thanks are also due to the Centre National de la Recherche Scientifique (CNRS), Common Funds Aquitaine-Aragón and European Social Fund (ESF), for the grants specially allotted at the time of the Tenth Conference.

We wish to express our gratitude to Mohamed Amara (U. Pau), Laurent Bordes (U. Pau), Mira Bozzini (U. Milano-Bicocca), Jacqueline Fleckinger (U. Toulouse I), Raúl Gouet (U. Chile), Miguel Pasadas (U. Granada) and Tomas Sauer (U. Justus-Leibig-Geissen), who, together with us, formed the Scientific Committee, and Diego Izquierdo, Javier López and Pedro Mateo, from the University of Zaragoza, who shared with us the tasks of the Organizing Committee. We are also indebted to all others who helped in the organization of the conference, in particular, Beatriz Palacios and José Manuel Palacios.

We finally acknowledge the assistance provided for the realization of the proceedings by the Instituto Universitario de Matemáticas y Aplicaciones, inside the Monografías Matemáticas García de Galdeano, and the Servicio de Publicaciones of the University of Zaragoza, as well as the kind cooperation of the referees. Special thanks are also due to Diego Izquierdo for his help in the preparation of the proceedings.

The next edition of the Conference Zaragoza-Pau will be held in Jaca from 15th to 17th September 2010. All of you are cordially invited to participate in this event.

Pau and Zaragoza, August 2010
The Editors

María Cruz López de Silanes
Manuel Palacios
Departamento de Matemática Aplicada
Universidad de Zaragoza

Monique Madaune-Tort
C. Paroissin
David Trujillo
Laboratoire de Mathématiques Appliquées
Université de Pau et des Pays de l'Adour

Gerardo Sanz
Departamento de Métodos Estadísticos
Universidad de Zaragoza

Juan José Torrens
Departamento de Matemática e Informática
Universidad Pública de Navarra

Contributors

List of participants

ADIMY, Mostafa
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
mostafa.adimy@univ-pau.fr

AGUT, Cyril
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour et INRIA MAGIQUE-3D,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
mostafa.adimy@univ-pau.fr

AKHTAR, Waseem
Abdus Salam School of Mathematical Sciences,
GC University,
Lahore, 68-B New Muslim Town, Lahore, Pakistan.
wasakh75@hotmail.com

AMARA, Mohamed
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
mohamed.amara@univ-pau.fr

AMROUCHE, Chérif
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
cherif.amrouche@univ-pau.fr

ARTAL, Enrique
Departamento de Matemáticas,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
artal@unizar.es

BADRA, Mehdi
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
mehdi.badra@univ-pau.fr

BALDASSARI, Carolina
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
cbaldass3@hotmail.fr

BARBET, Luc
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
luc.barbet@univ-pau.fr

BARUCQ, Hélène
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour et INRIA MAGIQUE-3D,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
helene.barucq@univ-pau.fr

BEGOUT, Pascal
Université de Toulouse I,
Résidence le Clos du Village 3,
Avenue Marc Laurent,
78370 Plaisir, France.
begout@ann.jussieu.fr

BERRADE, Lola
Departamento de Métodos Estadísticos,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
berrade@unizar.es

BIRITXINAGA, Edurne
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
edurne.biritxinaga@univ-pau.fr

BLANC, Jérémy
Université de Grenoble I,
Institut Fourier,
BP 74, 38402 Saint-Martin d'Hères, France.
Jeremy.Blanc@fourier.ujf-grenoble.fr

BOAL, Natalia
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
nboal@unizar.es

BONZOM, Florian
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
florian.bonzom@univ-pau.fr

BORDES, Laurent
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
laurent.bordes@univ-pau.fr

BOZZINI, Mira
Dipartimento di Matematica ed Applicazioni,
Università di Milano-Bicocca,
Via Cozzi 53, 20125 Milano, Italy.
mira.bozzini@unimib.it

CARNICER, Jesús Miguel
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
carnicer@unizar.es

CASADO, José Luis
Centro Meteorológico Territorial en Canarias
c/ Historiador Fernando de Armas 12,
Urbanización El Zurbarán,
35017 Las Palmas de Gran Canaria Oriental,
Spain.
jl_casado@yahoo.es

CLAVERO, Carmelo
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
clavero@unizar.es

COGOLLUDO, José Ignacio
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
jicogo@unizar.es

CORTÉS, Vanesa
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
vcortes@unizar.es

CRESSON, Jacky
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
jacky.cresson@univ-pau.fr

DAHOUMANE, Fabien
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
fa.da64@tele2.fr

DAMBRINE, Marc
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
`marc.dambrine@utc.fr`

DIAZ, Julien
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
`julien.diaz@inria.fr`

DOSSOU-GBÉTÉ, Simplice
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
`simplice.dossou-gbete@univ-pau.fr`

DUPRAT, Véronique
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
`veronique.duprat@inria.fr`

FAENZI, Daniele
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
`daniele.faenzi@univ-pau.fr`

FERREIRA, Chelo
Departamento de Matemática Aplicada,
Facultad de Veterinaria,
Universidad de Zaragoza,
C/ Miguel Servet 117,
Zaragoza, Spain.
`cferrei@unizar.es`

FLECKINGER, Jacqueline
CEREMATH - Université Toulouse I,
Pl. du Doyen G. Marty,
31042 Toulouse Cedex, France.
`jfleckinger@gmail.com`

FLORENS, Vincent
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
`vincent.florens@univ-pau.fr`

GARCÍA, Pilar
Departamento de Mécanica de Fluidos,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
`pigar@unizar.es`

GASCA, Mariano
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
`gasca@unizar.es`

GASPAR, Francisco
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
`fjgaspar@unizar.es`

GIACOMONI, Jacques
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
`jacques.giacomoni@univ-pau.fr`

GÓMEZ, Inmaculada
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
`igomez@unizar.es`

GRACIA, José Luis
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
jlgracia@unizar.es

GREFF, Isabelle
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
isabelle.greff@univ-pau.fr

GRIGOROSCUTA, Magdalena
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
magdalena.grigoroscuta@etud.univ-pau.fr

GUTIÉRREZ, José M.
Departamento de Matemáticas y Computación,
Universidad de La Rioja,
Edificio Vives,
c/ Luis de Ulloa s/n,
26004, Logroño, Spain.
jmguti@unirioja.es

HANNA, Hanen
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
hhanna@uc.edu.ve

IBÁÑEZ, M. José
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Granada,
Campus de Fuentenueva s/n,
18071 Granada. Spain.
mibanez@ugr.es

IZQUIERDO, Diego
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
dizquier@unizar.es

JAMES, François
Université d'Orléans,
UMR CNRS 6628,
BP 6759, 45067 Orléans Cedex 2,France.
francois.james@univ-orleans.fr

JIMENEZ, Julien
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
julien.jimenez@univ-pau.fr

JODRA, Pedro
Departamento de Métodos Estadísticos,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
pjodra@unizar.es

JOIE, Julie
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
julie.joie@etud.univ-pau.fr

KAMLA, Vivient
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
vivientcorneille.kamla@univ-pau.fr

KOMATITSCH, Dimitri
Laboratoire de Modélisation et d'Imagerie en
Géosciences de Pau (MIGP),
Université de Pau et des Pays de l'Adour et
INRIA MAGIQUE 3D,
IPRA MIGP, Avenue de l'Université,
BP 1155, 64013 Pau Cedex, France.
dimitri.komatitsch@univ-pau.fr

LABURTA, M. Pilar
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
laburta@unizar.es

LAHOZ, David
Departamento de Métodos Estadísticos,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
davidla@unizar.es

LANCHARES, Víctor
Departamento de Matemáticas y Computación,
Universidad de La Rioja,
Edificio Vives,
c/ Luis de Ulloa s/n,
26004 Logroño, Spain.
vlancha@unirioja.es

LATORRE, Borja
Departamento de Mécanica de Fluidos,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
borja.latorre@unizar.es

LECUREUX, Marie Hélène
CEREMATH - Université Toulouse I,
Pl. du Doyen G. Marty,
31042 Toulouse Cedex, France.
mh.lecureux@free.fr

LÓBON, Javier
Museo Nacional de Ciencias Naturales (CSIC),
c/ José Gutierrez Abascal, 2,
28006 Madrid. Spain.
MCNL178@mncn.csic.es

LÓPEZ DE SILANES, María Cruz
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
mcruz@unizar.es

LOULY, Mohamed Salem
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
mohamed.salem@etud.univ-pau.fr

LUCE, Robert
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
robert.luce@univ-pau.fr

MADAUNE-TORT, Monique
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
monique.madaune-tort@univ-pau.fr

MADEC, Ronan
Laboratoire de Modélisation et d'Imagerie en
Géosciences de Pau (MIGP),
Université de Pau et des Pays de l'Adour et
INRIA MAGIQUE 3D,
IPRA MIGP, Avenue de l'Université,
BP 1155, 64013 Pau Cedex, France.
ronan.madec@etud.univ-pau.fr

MALLOR, Fermín
Departamento de Estadística e Investigación
Operativa,
Universidad Pública de Navarra,
Campus de Arrosadía, 31006 Pamplona, Spain.
mallor@unavarra.es

MARQUET, Catherine
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
catherine.marquet@univ-pau.fr

MARTÍN, Jorge
Departamento de Matemáticas,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
jorge@unizar.es

MARTÍN, Raúl
Departamento de Matemáticas
Universidad de Castilla-La Mancha,
Escuela Superior de Informática,
Paseo de la Universidad 4, 13071 Ciudad Real,
Spain.
Raul.MMartin@uclm.es

MARTIN, Roland
Laboratoire de Modélisation et d'Imagerie en
Géosciences de Pau (MIGP),
Université de Pau et des Pays de l'Adour et
INRIA MAGIQUE 3D,
IPRA MIGP, Avenue de l'Université,
BP 1155, 64013 Pau Cedex, France.
roland.martin@univ-pau.fr

MIANA, Pedro J.
Departamento de Matemáticas,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
pjmiana@unizar.es

MOGUEN, Yann
Laboratoire de Thermique, Énergétique et
Procédés,
Université de Pau et des Pays de l'Adour
IUT-GTE,
Avenue de l'Université, 64000 Pau, France.
yann.moguen@etud.univ-pau.fr

MOKRANI, Amar
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
amar.mokrani@etud.univ-pau.fr

MOLER, José Antonio
Departamento de Estadística e Investigación
Operativa,
Universidad Pública de Navarra,
Campus de Arrosadía, 31006 Pamplona, Spain.
jmoler@unavarra.es

NAFIDI, Ahmed
Departamento de Estadística e Investigación
Operativa,
Facultad de Ciencias,
Universidad de Granada,
Campus de Fuentenueva,
18071 Granada, Spain.
nafidiah@ugr.es

NAZAR, Mudassar
Abdus Salam School of Mathematical Sciences,
GC University,
Lahore, 68-B New Muslim Town, Lahore,
Pakistan.
mudassar_666@yahoo.com

NGUYEN, Huy Hoang
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
huy_hoang.nguyen@yahoo.com

NUALART, Eulalia
Departamento de Estadística e Investigación
Operativa,
Universidad Pública de Navarra,
Campus de Arrosadía, 31006 Pamplona, Spain.
eulalia@nualart.es

PALACIOS, Manuel
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
mpala@unizar.es

PARENT, Eric
AgroParisTech,
ENGREF/MORSE,
19 Av. du Maine, 75732 Paris Cedex 15, France.
eric.parent@agroparistech.fr

PAROISSIN, Christian
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
christian.paroissin@univ-pau.fr

PASCUAL, Ana Isabel
Departamento de Matemáticas y Computación,
Universidad de La Rioja,
c/ Luis de Ulloa s/n,
26004 Logroño, Spain.
aipasc@unirioja.es

PEÑA, Juan Manuel
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
jmpena@unizar.es

PÉREZ, Ester
Departamento de Ingeniería Matemática e Informática,
Universidad Pública de Navarra,
Campus de Arrosadía, 31006 Pamplona, Spain.
ester.perez@unavarra.es

PETRAU, Agnes
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
agnes.petrau@etud.univ-pau.fr

PIERRE, Charles
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
charles.pierre@univ-pau.fr

PLO, Fernando
Departamento de Métodos Estadísticos,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
fplo@unizar.es

PORTERO, Laura
Departamento de Ingeniería Matemática e Informática,
Universidad Pública de Navarra,
Campus de Arrosadía, 31006 Pamplona, Spain.
laura.portero@unavarra.es

PUIG, Bénédicte
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
benedicte.puig@univ-pau.fr

PUISEUX, Pierre
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
pierre.puiseux@univ-pau.fr

RODRIGO, Carmen
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
carmenr@unizar.es

RODRIGUEZ, Licesio
Departamento de Matemáticas,
Universidad de Castilla-La Mancha,
E.T.S. Ingenieros Industriales,
Avda. Camilo José Cela s/n,
13071, Ciudad Real, Spain.
L.RodriguezAragon@uclm.es

RODRÍGUEZ, Miguel Luis
Departamento de Matemática Aplicada,
E.T.S.I. Caminos, Canales y Puertos,
Universidad de Granada,
Campus de Fuentenueva,
18071 Granada, Spain.
miguelrg@ugr.es

ROMERO, Natalia
Departamento de Matemáticas y Computación,
Universidad de la Rioja,
Edificio Vives,
c/ Luis de Ulloa, s/n,
26004 Logroño, Spain.
natalia.romero@unirioja.es

SAINT-GUIRONS, Anne-Gaëlle
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
annegaelle_saintguirons@yahoo.fr

SANGÜESA, Carmen
Departamento de Métodos Estadísticos,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
csangues@unizar.es

SANZ, Gerardo
Departamento de Métodos Estadísticos,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
gerardo.sanz@unizar.es

SAOUDI, Kamel
CEREMATH,
Université Toulouse I,
Manufacture des Tabacs,
21 Allées de Brienne,
31000 Toulouse. France.
kamel.saoudi@univ-tlse1.fr

SAUER, Tomas
Lehrstuhl für Numerische Mathematik,
Justus-Leibig-Universität-Giessen,
Heinrich-Buff-Ring 44, 35392 Giessen,
Germany.
tomas.sauer@math.uni-giessen.de

SAWADOGO, Amadou
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
amadou.sawadogo@etud.univ-pau.fr

SCHINDLER, Ian
Université Toulouse I,
Manufacture des Tabacs,
21 Allées de Brienne,
31000 Toulouse. France.
ian.schindler@univ-tlse1.fr

SEBASTIÁN, María Victoria
Departamento de Matemática Aplicada,
C.P.S., Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
msebasti@unizar.es

SELOULA, Nour
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
nourelhouda.seloula@etud.univ-pau.fr

TAAKILI, Abdelaziz
Laboratoire de Mathématiques et leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
abdelaziz.taakili@etud.univ-pau.fr

TAKÁČ, Peter
Department of Mathematics,
University of Rostock,
18055 Rostock, Germany.
peter.takac@uni-rostock.ge

TCHOUANMO, Stephane
IRSN - DSU SSIAD,
BERIS BP 17 92262,
Fontenay-aux-Roses Cedex, France.
s.tchouanmo@etud.univ-pau.fr

TORRENS, Juan José
Departamento de Ingeniería Matemática e Informática,
Universidad Pública de Navarra,
Campus de Arrosadía, 31006 Pamplona, Spain.
jjtorrens@unavarra.es

TOVIO, Víctor
Departamento de Matemáticas,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
tovio@unizar.es

TRUJILLO, David
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
david.trujillo@univ-pau.fr

TURLOT, Jean Christophe
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
jean-christophe.turlot@univ-pau.fr

URMENETA, Henar
Departamento de Estadística e Investigación Operativa,
Universidad Pública de Navarra,
Campus de Arrosadía, 31006 Pamplona, Spain.
henar@unavarra.es

VALLÈS, Jean
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
jean.valles@univ-pau.fr

VALLET, Guy
Laboratoire de Mathématiques et leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
guy.vallet@univ-pau.fr

WITTBOLD, Petra
Institut für Mathematik,
Technische Universät Berlin,
Straße des 17 Juni 136,
10623 Berlin, Germany.
wittbold@math.tu-berlin

Other communications

The following contributions are the ones which were presented but not included in this book. Some will appear in other publications.

Groundstat positivity or negativity for some elliptic systems involving Schrödinger operators
B. Alziary, J. Fleckinger and M. H. Lécureux

A stabilized discontinuous Galerkin formulation for Helmholtz problems
M. Amara, R. Djellouli and M. Grigoroscuta

A 3D finite volume scheme for a water-gas flow model in porous media
B. Amaziane, S. Tchouanmo and M. Dymitrowska

Stokes problems in unbounded domains: an approach in weighted Sobolev spaces
C. Amrouche and F. Bonzom

The stationary Navier-Stokes Equations with a non-zero constant velocity at infinity in three-dimensional unbounded domains
C. Amrouche and H. H. Nguyen

Discontinuous Galerkin method for the Reverse Time Migration
C. Baldassari, H. Barucq, H. Calandra, B. Denel and J. Diaz

Some remarks on second-order optimality conditions for the class of $C^{1,1}$-functions
L. Barbet and A. Daniilidis

New perfectly matched layers for short water waves
H. Barucq, J. Diaz and M. Tlemcani

Performance assesment of new approximate local *DtN* boundary conditions for prolate spheroidal-shaped boundaries
H. Barucq, R. Djellouli and A-G. Saint-Guirons

Probabilistic evaluation of the delay elapsed between an event and its discovery during a maintenance operation
E. Biritxinaga-Etxart, L. Bordes, C. Paroissin, B. Puig, W. Tinsson, S. Baysset, J. L. Vérit and J. M. Bosc

On the complement of curves in the projective plane
J. Blanc

A globally convergent interior point algorithm for the posynomial model
N. Boal

Hitting time distribution of a given length in some growth models
N. Bru and C. Paroissin

Curve complements, combinatorics and pencils
J. I. Cogolludo

Parameter estimation for seasonal fractional ARIMA with stable innovations
S. Dossou-Gbété, M. Ndongo and A. Diop

Maximum likelihood estimation for Mallows-Bradley-Terry models for the analysis of ranking data
S. Dossou-Gbété and A. Sawadogo

Matemáticas en la Naturaleza
M. Gasca

Analyzing biomechanical curves as functional data
M. Gastón, T. León and F. Mallor

A class of explicit parallel Runge-Kutta-Nyström methods of high order
I. Gómez and J. M. Franco

Mathematical analysis of a class of coupling hyperbolic-parabolic problems
J. Jimenez and L. Lévi

A new integral representation of the polylogarithm function from a probabilistic perspective
P. Jodrá

An individual-base model of the spread of HIV in a heterosexual population that includes sex workers and their clients
V. C. Kamla and M. Artzrouni

A 21 billion degrees of freedom, 2.5 terabyte simulation of seismic wave propagation in the Earth in parallel on 2166 processors
D. Komatitsch, J. Labarta and D. Michéa

Forecasting wind data with a MLP NN trained by means of a combination of EA and LR
D. Lahoz and P. M. Mateo

Animal population regulation theory: A field test based on the long-term monitoring of two contrasting populations of stream-living brown trout (Salmo trutta)
J. Lobón-Cerviá

Implementation of incident plane waves in the Spectral-Element Method and comparison with the Method of Fundamental Solutions in 2D
R. Madec, D. Komatitsch and F. J. Sánchez-Sesma

Applied Mathematics

Monografías Matemáticas *García de Galdeano* **35**, 3–10 (2010)

Construction of quasi-interpolants on uniform partitions

A. Abbadi, D. Barrera, M. J. Ibáñez and D. Sbibih

Abstract. We propose a new general method for constructing standard quasi-interpolation operators into the space spanned by the integer translates of a B-spline defined on a uniform partition of $\mathbb{R}^s$. The key tool is an appropriate error estimate with a leader term that contains and expression measuring the quality of the approximation. It is a function on the sequence defining the quasi-interpolating operator, and therefore, we define and solve a minimization problem in such a way that their solutions are characterized in terms of some splines that do not depend on the linear form defining the operator.

Keywords: B-splines, box splines, discrete quasi-interpolants, differential quasi-interpolants, integral quasi-interpolants, approximation power, error estimates.

AMS classification: 41A05, 41A15, 65D05, 65D07.

§1. Introduction

We propose a new general method for constructing quasi-interpolation operators based on B-splines defined on uniform partitions τ of $\mathbb{R}^s$, $s \geq 1$. Let ϕ be such a B-spline on τ, normalized by $\sum_{i\in\mathbb{Z}^s} \phi(\cdot - i) = 1$. Let $\mathcal{S} := \operatorname{span}(\phi(\cdot - i))_{i\in\mathbb{Z}^s}$ be the cardinal spline space spanned by the shifts of ϕ.

The classical structure for a quasi-interpolant is given by the expression

$$Q(f) := \sum_{i\in\mathbb{Z}^s} \lambda f(\cdot + i)\, \phi(\cdot - i),$$

λ being a linear functional (see e.g. [3], [2], [5]). Usually, λf is a linear combination of values of f and some of its derivatives at some points in some open set containing the support of ϕ; or a linear combination of values of f at some points in this set; or a linear combination of weighted mean values of the function to be approximated, i.e. λf is given by

$$\sum_{j\in J} c_j f(-j), \ \sum_{|i|\leq\ell} \sum_{j\in J_i} c_{i,j} D^{(i)} f(-j), \ \text{or} \sum_{j\in J} c_j \langle f, \psi(\cdot - j)\rangle,$$

J and J_i, $|i| \leq \ell$ for $0 \leq \ell \leq \deg\psi$, being finite subsets of $\mathbb{Z}^s$, and $\langle\cdot,\cdot\rangle$ and ψ standing for the usual inner product and another B-spline.

That linear functional is defined to produce a quasi-interpolant Q exact on a polynomial space included in $\mathcal{S}$. We will restrict our attention to these cases. More precisely, we will demand the exactness of Q on $\mathbb{P}_n$, with n such that $\mathbb{P}_n \subset \mathcal{S}$ and $\mathbb{P}_{n+1} \not\subseteq \mathcal{S}$, i.e. Q realizes the approximation power of $\mathcal{S}$.

§2. Estimating the quasi-interpolation error

For the scaled quasi-interpolant

$$Q_h f := \sum_{i\in\mathbb{Z}^s} \lambda f\,(h\,(\cdot + i))\,\phi\left(\frac{\cdot}{h} - i\right)$$

considered here, we have the following result concerning the error $E_h f := f - Q_h f$. The notation m_α is used for the normalized monomial of order α: $m_\alpha(x) = x^\alpha/\alpha!$.

Proposition 1. *Let $f \in C^{n+2}(\mathbb{R}^s)$. For every triangle T in $h\tau$, there exist both a neighborhood $V = V(T)$, independent of f, and a constant $C > 0$, independent of h and T, such that*

$$\|E_h f\|_{\infty,T} \le T_{n,Q} h^{n+1} |f|_{\infty,n+1,V} + C\,h^{n+2}\,|f|_{\infty,n+2,V}\,,$$

where

$$T_{n,Q} := \max_{\alpha\in\mathbb{N}_0^s,\ |\alpha|=n+1} \|Qm_\alpha - m_\alpha\|_{\infty,[0,1]^s}\,.$$

Proof. Suppose that Q is an integral quasi-interpolation operator. Then, we have

$$Qf = \sum_{i\in\mathbb{Z}^s} \lambda f\,(\cdot + i)\,\phi\,(\cdot - i)$$

with

$$\lambda f\,(\cdot + i) = \sum_{j\in J} c_j \left\langle f\,(\cdot + i)\,, \psi\,(\cdot - j)\right\rangle = \int_{\mathbb{R}^s} f\,(t)\,H\,(t - i)\,dt,$$

where

$$H := \sum_{j\in J} c_j \psi\,(\cdot - j)\,.$$

Thus,

$$Qf = \int_{\mathbb{R}^s} f\,(t)\,K\,(t,\cdot)\,dt,$$

with

$$K\,(t,\cdot) := \sum_{i\in\mathbb{Z}^s} H\,(t - i)\,\phi\,(\cdot - i)\,.$$

Taking into account that the scaled quasi-interpolant Q_h in is equal to $\sigma_h Q \sigma_{1/h}$ where the scaling operator σ_h is defined as

$$\sigma_h f = f\left(\frac{\cdot}{h}\right),$$

we get

$$Q_h f = \int_{\mathbb{R}^s} f\,(t)\,\frac{1}{h^s} K\left(\frac{t}{h}, \frac{\cdot}{h}\right) dt.$$

The kernel in this integral representation of Q_h is $\mathbb{P}_{n-1}$-reproducing and shift-invariant, and has sufficient decay. Then, the next error estimate for the integral quasi-interpolation operator considered here follows from [4].

The proof for discrete quasi-interpolants is given in [1]. A similar method can be used to prove the result in the differential case. □

The constant $T_{n,Q}$ in the leading term of $E_h f$ is determined by how well Q_h approximates the monomials of order $n + 1$.

§3. Achieving the required exactness

The operator Q is exact on $\mathbb{P}_n$ if for all $\alpha \in \mathbb{N}_0^s$ such that $|\alpha| \leq n$ one gets

$$\lambda\,(m_\alpha) = g_\alpha\,(0)\,,$$

where the polynomials g_α can be recursively computed as follows (see e.g. [3]):

$$g_0 = m_0, \quad g_\alpha = m_\alpha - \sum_{j\in\mathbb{Z}^s} \phi\,(j) \sum_{\beta\lneqq\alpha} m_{\alpha-\beta}\,(-j)\, g_\beta, \;\; |\alpha| > 0.$$

They are only sufficient conditions to guarantee the exactness of Q on $\mathbb{P}_n$.

§4. A minimization problem

It is natural to construct Q by solving this minimization problem:

Problem 1. Minimize $T_{n,Q}$ subject to the exactness conditions $\lambda\,(m_\alpha) = g_\alpha\,(0)\,,\; |\alpha| \leq n$.

The solutions of this problem (and the corresponding quasi-interpolants Q) can be easily characterized using the well-known Schoenberg operator

$$S f := \textstyle\sum_{i\in\mathbb{Z}^s} f\,(i)\,\phi\,(\cdot - i)\,.$$

Proposition 2. *Let Q be one of the quasi-interpolants considered here defined from the linear functional λ. Let us suppose that Q is exact on $\mathbb{P}_n$. If*

$$\lambda m_\alpha = g_\alpha\,(0) + \frac{1}{2}\left(\max_{[0,1]^s} G_\alpha + \min_{[0,1]^s} G_\alpha\right)$$

for all $\alpha \in \mathbb{N}_0^s$ such that $|\alpha| = n+1$, where

$$G_\alpha := m_\alpha - S\,g_\alpha,$$

then $T_{n,Q}$ attains its minimum value.

Note that G_α does not depend on λ.

§5. A differential example

Let ϕ be the quadratic box-spline on the criss-cross triangulation τ_2, centered at the origin (see e.g. [3]). Then $n = 2$, i.e. we can construct differential quasi-interpolants exact on $\mathbb{P}_2$ by minimizing the errors associated with the cubic monomials. We will restrict our attention to the case $\ell = 1$, i.e. we will suppose that the values of f and its first order partial derivatives at the grid points are known.

We have

$$\begin{aligned}\lambda_\mu f = f\,(0) &+ \frac{1}{16}\left(D^{(1,0)} f\,(1,0) - D^{(1,0)} f\,(-1,0)\right) - \mu\left(D^{(1,0)} f\,(0,1) - D^{(1,0)} f\,(0,-1)\right)\\ &- \mu\left(D^{(0,1)} f\,(1,0) - D^{(1,0)} f\,(-1,0)\right) + \frac{1}{16}\left(D^{(0,1)} f\,(0,-1) - D^{(0,1)} f\,(0,1)\right).\end{aligned}$$

The exactness of Q on $\mathbb{P}_2$ is guaranteed by the conditions

$$\lambda m_\alpha = g_\alpha(0), \ |\alpha| \le 2.$$

Since $\max_{[0,1]^2} G_\alpha = -\min_{[0,1]^2} G_\alpha$ when $|\alpha| = 3$, the new linear equations yielding the minimum of $T_{2,Q}$ are given by

$$\lambda m_\alpha = g_\alpha(0), \ |\alpha| = 3.$$

When $J_{0,0} = \{(0,0)\}$ and $J_{1,0} = J_{0,1} = \{(0,0), (\pm 1, 0), (0, \pm 1)\}$, the solution of this linear system depends on a parameter μ, and provides the linear functional

$$\begin{aligned} \lambda_\mu f = f(0) &+ \frac{1}{16}\left(D^{(1,0)} f(1,0) - D^{(1,0)} f(-1,0)\right) - \mu\left(D^{(1,0)} f(0,1) - D^{(1,0)} f(0,-1)\right) \\ &- \mu\left(D^{(0,1)} f(1,0) - D^{(1,0)} f(-1,0)\right) + \frac{1}{16}\left(D^{(0,1)} f(0,-1) - D^{(0,1)} f(0,1)\right). \end{aligned}$$

The value $\mu = 0$ gives a differential quasi-interpolant Q^* having minimally supported fundamental functions. We have the following result concerning its associated error.

Proposition 3. *Let $f \in C^3\left(\mathbb{R}^2\right)$. For every triangle T in $h\tau_2$, there exist both a neighborhood V_T, independent of f, and constants $C_\alpha > 0$, independent of h and T, such that*

$$\left\| D^\alpha\left(Q_h^* f - f\right)\right\|_{\infty,T} \le C_\alpha h^{3-|\alpha|} \left\| D^3 f\right\|_{\infty,V_T}.$$

Moreover,

$$\begin{aligned} C_{0,0} &= \frac{153 + 15\sqrt{10} + 13\sqrt{13}}{648} \simeq 0.381646, \\ C_{1,0} = C_{0,1} &= \frac{198 + 10\sqrt{10} + 13\sqrt{13}}{324} \simeq 0.853379. \end{aligned}$$

We consider the test function, whose graphic is given in Figure 1.

$$f(x,y) = 3(1-x)^2 e^{-x^2-(y+1)^2} - 10\left(\frac{x}{5} - x^3 - y^5\right) e^{-x^2-y^2} - \frac{1}{3} e^{-(x+1)^2-y^2}.$$

Figure 2 shows the errors associated with the new differential quasi-interpolation operator Q_h^* for some different values of the steplength h.

In order to show the performance of Q_h^*, we also give in Figure 3 the plots of the errors associated with the classical differential quasi-interpolant $\widetilde{Q}_h$ that uses the partial derivatives up to the order two, for the same values of h:

$$\widetilde{Q}_h f = \sum_{i \in \mathbb{Z}^2} \left(f(ih) - \frac{h^2}{8}\left(D^{(2,0)} f(ih) + D^{(0,2)} f(ih)\right)\right) \phi\left(\frac{\cdot}{h} - i\right).$$

The operator $Q_h^* f$ obtained solving the minimization problem gives good results when compared with $\widetilde{Q}_h f$, although the latter uses second order partial derivatives.

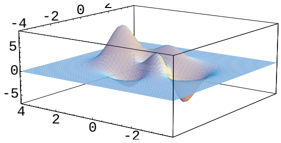


Figure 1: The test function f.

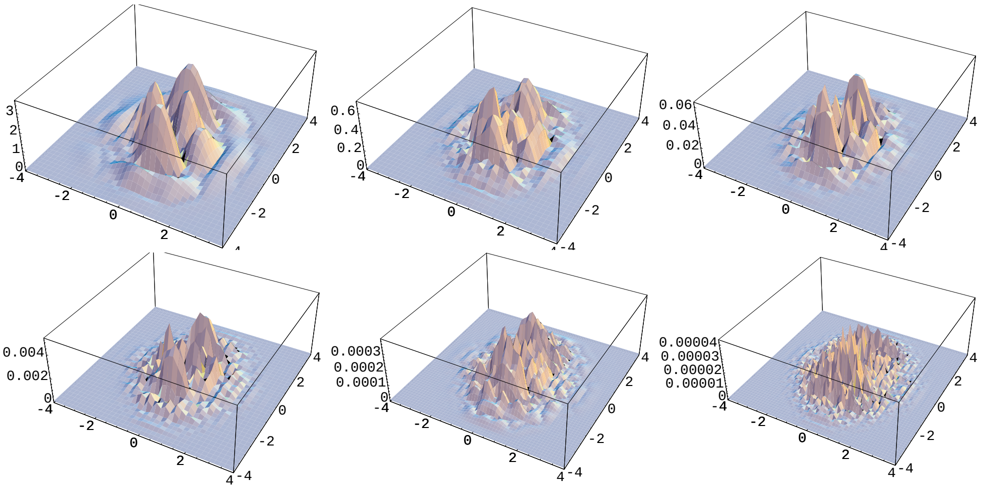


Figure 2: Quasi-interpolation errors $Q_h^* f$ for the test functions for $h = \frac{1}{2^n}$, $0 \leq n \leq 5$.

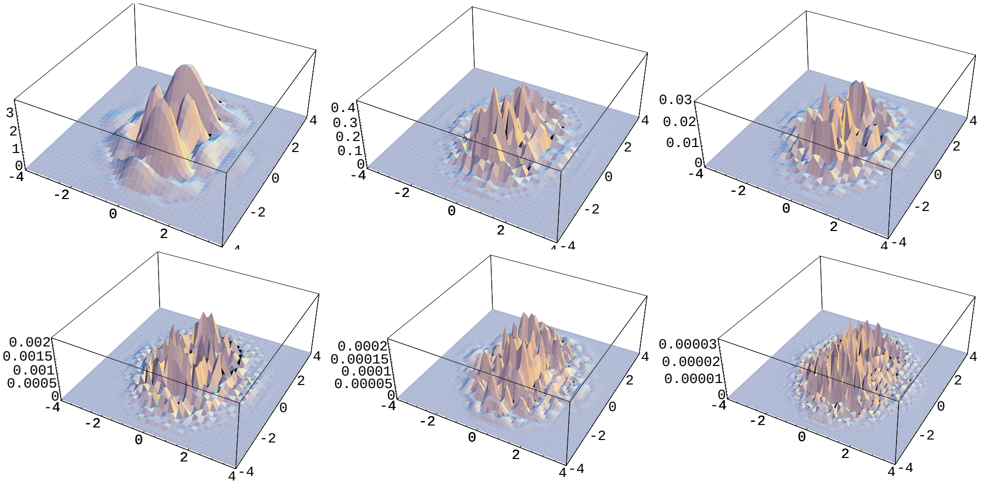


Figure 3: Quasi-interpolation errors $\widetilde{Q}_h f$ for the test functions for $h = \frac{1}{2^n}$, $0 \leq n \leq 5$.

§6. An integral example

Let τ be the uniform mesh of the plane generated by the directions $d_1 := (1,0)$, $d_2 := (0,1)$, $d_3 := d_1 + d_2$ and $d_4 := -d_1 + d_2$. Let ϕ be the box spline associated to the direction set $X = \{d_1, d_1, d_2, d_2, d_3, d_4\}$, centered at the origin (cf. [3]). It is one of the two box splines in $\mathbb{P}_4^2(\tau_2)$. It is well known (cf. [2]) that $\mathbb{P}_3$ is the space of maximal total degree included in $\mathcal{S}(\phi)$, that is the construction we have given runs with $n = 3$. It can be easily verified that the unique nonzero values of ϕ at the integers are

$$\begin{aligned}
&\phi(0,0) = \frac{5}{12},\\
&\phi(1,0) = \phi(-1,0) = \phi(0,1) = \phi(0,-1) = \frac{1}{8},\\
&\phi(1,1) = \phi(-1,1) = \phi(-1,-1) = \phi(1,-1) = \frac{1}{48}.
\end{aligned}$$

From these values we obtain the following expressions for the polynomials in the Appell sequence associated to ϕ:

$$\begin{aligned}
&g_{0,0} = 1,\ g_{1,0} = m_{1,0},\ g_{0,1} = m_{0,1},\ g_{2,0} = m_{2,0} - \frac{1}{6},\ g_{1,1} = m_{1,1},\ g_{0,2} = m_{0,2} - \frac{1}{6},\\
&g_{3,0} = m_{3,0} - \frac{1}{6}m_{1,0},\ g_{2,1} = m_{2,1} - \frac{1}{6}m_{0,1},\ g_{1,2} = m_{1,2} - \frac{1}{6}m_{1,0},\ g_{0,3} = m_{0,3} - \frac{1}{6}m_{0,1},\\
&g_{4,0} = m_{4,0} - \tfrac{1}{6}m_{2,0} + \frac{1}{72},\ g_{3,1} = m_{3,1} - \frac{1}{6}m_{1,1},\ g_{2,2} = m_{2,2} - \frac{1}{6}m_{2,0} - \tfrac{1}{6}m_{0,2} + \frac{5}{144},\\
&g_{1,3} = m_{1,3} - \frac{1}{6}m_{1,1},\ g_{0,4} = m_{0,4} - \frac{1}{6}m_{0,2} + \frac{1}{72}.
\end{aligned}$$

After some computations, we get $G_{3,1} = G_{1,3} = 0$, and

$$\begin{aligned}
&\max_{[0,1]^2} G_{4,0} = G_{4,0}\left(\frac{1}{2},0\right) = \frac{1}{384}, && \min_{[0,1]^2} G_{4,0} = G_{4,0}(0,0) = 0,\\
&\max_{[0,1]^2} G_{2,2} = G_{2,2}(0,0) = 0, && \min_{[0,1]^2} G_{2,2} = G_{2,2}\left(\frac{1}{2},\frac{1}{2}\right) = -\frac{1}{192},\\
&\max_{[0,1]^2} G_{0,4} = G_{0,4}\left(0,\frac{1}{2}\right) = \frac{1}{384}, && \min_{[0,1]^2} G_{0,4} = G_{0,0}(0,0) = 0.
\end{aligned}$$

Thus, given a discrete, differential or integral linear form λ, we obtain the following equations that characterize the solutions of the minimization problem:

$$\begin{aligned}
&\lambda m_{0,0} = 1,\ \lambda m_{1,0} = \lambda m_{0,1} = 0,\\
&\lambda m_{2,0} = \lambda m_{0,2} = -\frac{1}{6},\ \lambda m_{1,1} = 0,\ \lambda m_{3,0} = \lambda m_{2,1} = \lambda m_{1,2} = \lambda m_{0,3} = 0,\\
&\lambda m_{4,0} = \lambda m_{0,4} = \frac{35}{2304},\ \lambda m_{2,2} = \frac{37}{1159},\ \lambda m_{3,1} = \lambda m_{1,3} = 0.
\end{aligned}$$

As a integral linear functional uses a B-spline ψ as weight function in the inner products, we choose $\psi = \phi$. Moreover, let J be the set of the integer $i = (i_1, i_2)$ such that $|i_1|+|i_2| \leq 2$.Taking into account that the nonzero moments of ψ are

$$\mu_{0,0} = 1,\ \mu_{2,0} = \mu_{0,2} = \frac{1}{3},\ \mu_{4,0} = \mu_{0,4} = \frac{3}{10},\ \mu_{2,2} = \frac{17}{180},$$

the expansion of λm_α, $|a| \leq 4$, results in a linear system on $c = (c_j)_{|j_1|+|j_2|\leq 2}$ whose unique solution is

$$\begin{aligned} c_{0,0} &= \frac{11071}{2880}, c_{1,0} = c_{0,1} = c_{-1,0} = c_{0,-1} = -\frac{11}{12}, \\ c_{2,0} &= c_{0,2} = c_{-2,0} = c_{0,-2} = \frac{991}{11520}, \\ c_{1,1} &= c_{-1,1} = c_{-1,-1} = c_{1,-1} = \frac{689}{5760}. \end{aligned}$$

Note that c is a lozenge sequence and so the fundamental function of the associated quasi-interpolant has the same symmetries than the box spline ϕ.

Acknowledgements

This work has been supported by the Junta de Andalucía (Research group FQM/191) and by Spain's Ministerio de Educación y Ciencia (Research project MTM2008-00671).

References

[1] Barrera, D., and Ibáñez, M. J. Minimizing the quasi-interpolation error for bivariate discrete quasi-interpolants. *Comput. Appl. Math. 224* (2009), 250–268.

[2] Chui, C. K. *Multivariate Splines*, vol. 40 of *Regional Conference Series in Applied Mathematics*. Siam, Philadelphia, 1988.

[3] de Boor, C., and Riemenschneider, S. *Box splines*, vol. 98 of *Applied Mathematical Sciencies*. Springer, New York, 1993.

[4] Dekel, S., and Leviatan, D. On measuring the efficiency of kernel operators $l_p\left(\mathbb{R}^d\right)$. *Adv. Comput. Math. 20* (2004), 53–65.

[5] Lai, M.-J., and Schumaker, L. L. *Spline Functions on Triangulations*, vol. 110 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, Cambridge, 2007.

A. Abbadi
Département de Mathématiques et Informatique
Université Mohammed 1er
60000 Oujda, Maroc
abzabbadi@yahoo.fr

D. Barrera
Departamento de Matemática Aplicada
E.T.S. Ingenieros de Caminos, Canales y Puertos
Campus de Fuentenueva s/n
18071 Granada, España
dbarrera@ugr.es

M. J. Ibáñez
Departamento de Matemática Aplicada
Facultad de Ciencias
Campus de Fuentenueva s/n
18071 Granada, España
mibanez@ugr.es

D.Sbibih
Département de Mathématiques et Informatique
Université Mohammed 1er
60000 Oujda, Maroc
sbibih@sciences.univ-oujda.ac.ma

Monografías Matemáticas *García de Galdeano* **35**, 11–18 (2010)

A LEAST SQUARES APPROACH FOR AN INVERSE TRANSMISSION PROBLEM

Lekbir Afraites, Marc Dambrine and Djalil Kateb

Abstract. We consider the question of recovering the shape of an unknown inclusion ω inside a body Ω from a single boundary measurement. This inverse problem —known as electrical impedance tomography— is seen through the minimization of some Least Squares criteria. We provide the first and second order derivatives with respect of perturbations of the shape of the interface $\partial\omega$ of the state functions and of the objectives. We study the stability of the optimization and prove that the shape Hessian at an optimal inclusion is not coercive but compact explaining the ill-posedness of the proposed approach.

Keywords: Inverse conductivity problem, shape optimization, second order method.

AMS classification: 49Q10, 49Q12, 65N21.

§1. Introduction

Consider a body constant conductivity σ_1 occupying a bounded domain Ω in $\mathbb{R}^N$ with $N \geq 3$. Inside Ω, there is an unknown inclusion ω whose conductivity σ_2 differs from the background conductivity σ_1 ($\sigma_1, \sigma_2 > 0$). The electrical potential u solves the partial differential equation

$$-\text{div}\,(\sigma_\omega(x)\nabla u) = 0 \text{ in } \Omega, \tag{1}$$

with $\sigma_\omega = \sigma_1\chi_{\Omega\setminus\overline{\omega}} + \sigma_2\chi_\omega$. The notation χ_E denotes the characteristic function of a measurable subset E of Ω. By measuring the input voltage and the corresponding output current on $\partial\Omega$, we gain access to a Cauchy pair (f, g) for (1). In others words, both Dirichlet boundary condition $u = f$ and Neumann boundary condition $\sigma_1\partial_{\mathbf{n}}u = g$ are known on $\partial\Omega$. We consider the question of a practical reconstruction of ω by these redundant informations on $\partial\Omega$.

This problem is a particular case of the inverse conductivity problem of Calderón that concerns the determination of the conductivity distribution σ from boundary measurements ([11, 9, 4]). The identification problem of an inclusion by boundary measurements is usually written from a numerical point a view as the minimization of a cost function: typically a Least Squares matching criterion. Many authors have investigate the steepest descent method for this problem [7, 6, 2] with the methods of shape optimization.

We address in this manuscript the stability of the optimization problems obtained with different Least Square cost function. By introducing second order methods, we analyze the wellposedness of the optimization method. We explain the instability in the continuous settings in terms of shape optimization: the shape Hessian is not coercive —in fact its Riesz operator turns out to be compact— and hence the criterion to minimize does not have necessarily a local strict minimum. A Kohn-Vogelius type objective is studied in [3] and simplified models can be found in [5, 1]. In this note, we present a Least Squares approach for this inverse problem and obtain similar results. This fact is surprising since a Kohn-Vogelius criteria is expected to lead to more stable optimization schemes.

The present manuscript is organized as follows. In Section 2, we reformulate the identification problem as shape optimization problems, tracking with a Least Squares formulation the Dirichlet and Neumann boundary conditions. We precise the first and second derivative of the state and the corresponding expressions for the criteria by introducing an adjoint state. Finally, we present our main result: a compactness result for the shape hessian at a critical point. In Section 3, we justify some shape derivatives and explain the main steps of the proof for the compactness theorem that explains the ill-posedness of the underlying identification problem.

§2. The results

Let us fix the geometrical setting under consideration and the notations. We consider a bounded domain $\Omega \subset \mathbb{R}^N$ ($N \geq 3$) with a C^2 boundary. It is fulfilled with a material whose conductivity is σ_1, an unknown inclusion ω in Ω of conductivity $\sigma_2 \neq \sigma_1$. In the sequel, we fix $d_0 > 0$ and consider inclusions ω such that $\omega \subset\subset \Omega_{d_0} = \{x \in \Omega,\ d(x, \partial\Omega) > d_0\}$. We also assume that the boundary $\partial\omega$ is of class $C^{4,\alpha}$.

In the sequel, a bold character denotes a vector. If $\mathbf{h}$ denotes a deformation field, it can be written as $\mathbf{h} = \mathbf{h}_\tau + h_n\mathbf{n}$ on $\partial\omega$. Note also that in the following lines, $\mathbf{n}$ denotes the outer normal field to $\partial\omega$ pointing into $\Omega \setminus \overline{\omega}$. Hence, for $x \in \partial\omega$, we define, when the limit exists, $u^\pm(x)$ (resp. $(\partial_n u)^\pm(x)$) as the limit of $u(x \pm t\mathbf{n}(x))$ (resp. $\langle \nabla u(x \pm t\mathbf{n}(x), \mathbf{n}(x)) \rangle$) when $t > 0$ tends to 0. Note that $\mathbf{h}_\tau$ is a vector while h_n is a scalar quantity. Admissible deformation fields have to preserve $\partial\Omega$ and the regularity of the boundaries. Therefore, we consider the space of admissible fields

$$\mathcal{H} = \left\{ \mathbf{h} \in C^{4,\alpha}(\mathbb{R}^N, \mathbb{R}^N), \operatorname{supp}(\mathbf{h}) \subset \Omega_{d_0} \right\}.$$

2.1. The shape optimization problem

In order to recover the shape of the inclusion ω, an possible strategy is to minimize a cost function. Many choices are possible, in particular a Least Squares type objective. In this paper, we study two different Least Square cost functions. We now define these criteria. Fixing the Neumann boundary data, we can track Dirichlet boundary conditions:

$$J_{LS}(\omega) = \frac{1}{2}\int_{\partial\Omega} |u_n - f|^2,$$

where f is the disturbed boundary measurements and the u_n is solution of the Neumann boundary value problem:

$$\begin{cases} -\operatorname{div}(\sigma_\omega \nabla u_n) = 0 & \text{in } \Omega, \\ \sigma_1 \partial_n u_n = g & \text{on } \partial\Omega. \end{cases} \tag{2}$$

To obtain uniqueness of the solution of (2), we add the normalization condition

$$\int_{\partial\Omega} u_n = \int_{\partial\Omega} f. \tag{3}$$

Another possible choice is to fix Dirichlet boundary condition and track the outgoing flux:

$$J_{DLS}(\omega) = \frac{1}{2}\int_{\partial\Omega} |\sigma_1 \partial_n u_d - g|^2,$$

where u_d is solution of the Dirichlet boundary value problem:

$$\begin{cases} -\mathrm{div}\,(\sigma_\omega \nabla u_d) = 0 & \text{in } \Omega, \\ u_d = f & \text{on } \partial\Omega. \end{cases} \tag{4}$$

To ensure that the cost function J_{DLS} is well defined, we assume that the Dirichlet data $f \in H^{3/2}(\partial\Omega)$. To avoid this assumption, one usually prefers to consider J_{LS} than J_{DLS}.

2.2. Differentiability results for the state u_n and u_d

We quote from [6, 10, 2] the first order derivative of the state u_n and u_d.

Theorem 1. *Let Ω be a open subset of $\mathbb{R}^N$ with a C^2 boundary and ω a subdomain in Ω_{d_0} with a $C^{4,\alpha}$ boundary. The state functions u_n and u_d are shape differentiable and their shape derivative u_n' and u_d' belong to* $\mathrm{H}^1(\Omega \setminus \overline{\omega}) \cup \mathrm{H}^1(\omega)$ *and satisfy*

$$\begin{cases} \Delta u_n' = 0 \;\; \textit{in } \Omega \setminus \overline{\omega} \textit{ and in } \omega, \\ [u_n'] = h_n \dfrac{[\sigma]}{\sigma_1}\, \partial_{\mathbf{n}} u_n^- \;\; \textit{on } \partial\omega, \\ [\sigma \partial_n u_n'] = [\sigma]\mathrm{div}_\tau\,(h_n \nabla_\tau u_n) \;\; \textit{on } \partial\omega, \\ \sigma_1 \partial_n u_n' = 0 \;\; \textit{on } \partial\Omega, \end{cases} \quad \textit{and} \quad \begin{cases} \Delta u_d' = 0 \;\; \textit{in } \Omega \setminus \overline{\omega} \textit{ and in } \omega, \\ \left[u_d'\right] = h_n \dfrac{[\sigma]}{\sigma_1}\, \partial_{\mathbf{n}} u_d^- \;\; \textit{on } \partial\omega, \\ \left[\sigma \partial_n u_d'\right] = [\sigma]\mathrm{div}_\tau\,(h_n \nabla_\tau u_d) \;\; \textit{on } \partial\omega, \\ u_d' = 0 \;\; \textit{on } \partial\Omega. \end{cases} \tag{5}$$

The second order derivative of the state functions u_n is computed in [3].

Theorem 2. *Let Ω be a open subset of $\mathbb{R}^N$ with a C^2 boundary and ω a element of Ω_{d_0} with a $C^{4,\alpha}$ boundary. Let $\mathbf{h}_1$ and $\mathbf{h}_2$ be two deformation fields in $\mathcal{H}$. The state u_n is has a second order shape derivative $u_n'' \in$* $\mathrm{H}^1(\Omega \setminus \overline{\omega}) \cup \mathrm{H}^1(\omega)$ *that solves*

$$\begin{cases} \Delta u_n'' = 0 \textit{ in } \Omega \setminus \overline{\omega} \textit{ and in } \omega, \\ [u_n''] = (h_{1,n} h_{2,n} H - \mathbf{h_1}_\tau.(D\mathbf{n}\,\mathbf{h_2}_\tau))\,[\partial_{\mathbf{n}} u_n] - (h_{1,n}[\partial_{\mathbf{n}}(u_n)_2'] + h_{2,n}[\partial_{\mathbf{n}}(u_n)_1']) \\ \qquad + (\mathbf{h_1}_\tau.\nabla h_{2,n} + \mathbf{h_2}_\tau.\nabla h_{1,n})\,[\partial_{\mathbf{n}} u_n] \textit{ on } \partial\omega, \\ [\sigma \partial_n u_n''] = \mathrm{div}_\tau\,(h_{2,n}\,[\sigma \nabla_\tau (u_n)_1'] + h_{1,n}\,[\sigma \nabla_\tau (u_n)_2'] + \mathbf{h_1}_\tau.(D\mathbf{n}\,\mathbf{h_2}_\tau)[\sigma \nabla_\tau u_n]) \\ \qquad - \mathrm{div}_\tau\,((\mathbf{h_1}_\tau.\nabla_\tau h_{2,n} + \nabla_\tau h_{1,n}.\mathbf{h_2}_\tau)\,[\sigma \nabla_\tau u_n]) \\ \qquad + \mathrm{div}_\tau\,(h_{2,n} h_{1,n}(2D\mathbf{n} - HI)\,[\sigma \nabla_\tau u_n]) \textit{ on } \partial\omega, \\ \sigma_1 \partial u_n'' = 0 \textit{ on } \partial\Omega. \end{cases} \tag{6}$$

Here, $(u_n)_i'$ denotes the first order derivative of u in the direction of h_i as given in (5)*, $D\mathbf{n}$ stands for the second fundamental form of the manifold $\partial\omega$ and H stands for the mean curvature of $\partial\omega$. Note that H is then the sum of the main curvatures and not the scaled version (divided by $n-1$) in dimension n.*

The result concerning u_d is an easy adaption of Theorem 2. Once the differentiability of the state function has been established, the chain rule provides the differentiability with respect to the shape of criterion.

2.3. Differentiability of the objective

As usual for Least Squares objective, this derivative can be simplified thanks to an adjoint state denoted by w_{LS} for J_{LS} and w_{DLS} for J_{DLS}.

Theorem 3. *Let Ω be a open subset of $\mathbb{R}^N$ ($N \geq 3$) with a C^2 boundary and ω a element of Ω_{d_0} with a $C^{4,\alpha}$ boundary. The Least-Square objective J_{LS} and J_{DLS} are differentiable with respect to the shape and their derivatives in the direction of a deformation field $\mathbf{h}$ in $\mathcal{H}$ are given by*

$$DJ_{LS}(\omega).h = \frac{\sigma_1 - \sigma_2}{\sigma_2} \int_{\partial\omega} (\sigma_1 \partial_n w_{LS}^+ \partial_n u_n^+ + \nabla_\tau u_n . \nabla_\tau w_{LS})\, h_n,$$
$$DJ_{DLS}(\omega).h = -\frac{\sigma_1 - \sigma_2}{\sigma_2} \int_{\partial\omega} \Big(\sigma_1 \partial_n w_{DLS}^+ \partial_n u_d^+ + \nabla_\tau u_d . \nabla_\tau w_{DLS}\Big)\, h_n,$$

where the adjoint functions w_{LS} and w_{DLS} solve the boundary value problem

$$\begin{cases} -\mathrm{div}\,(\sigma \nabla w_{LS}) = 0 & in\ \Omega, \\ \sigma_1 \partial_n w_{LS} = u_n - f & on\ \partial\Omega, \end{cases} \quad and \quad \begin{cases} -\mathrm{div}\,(\sigma \nabla w_{DLS}) = 0 & in\ \Omega, \\ w_{DLS} = \sigma_1 \partial u_d - g & on\ \partial\Omega, \end{cases} \qquad (7)$$

The compatibility condition is satisfied thanks to the normalization (3). The adjoint has to be normalized for example as in (3).

In this work, we are interested by the second order shape derivative of the cost functions objectives and the study of the stability of these criteria. For this, will need the shape derivatives of the adjoint states w_{LS} and w_{DLS} obtained as a consequence of Theorem 1. The state functions w_{LS} and w_{DLS} are shape differentiable and their shape derivatives w'_{LS} and w'_{DLS} belong to $\mathrm{H}^1(\Omega \setminus \overline{\omega}) \cup \mathrm{H}^1(\omega)$ and satisfy

$$\begin{cases} \Delta w'_{LS} = 0 \ \text{ in } \Omega \setminus \overline{\omega} \text{ and in } \omega, \\ [w'_{LS}] = h_n \dfrac{[\sigma]}{\sigma_1} \partial_{\mathbf{n}} w_{LS}^- \ \text{ on } \partial\omega, \\ [\sigma \partial_n w'_{LS}] = [\sigma] \mathrm{div}_\tau (h_n \nabla_\tau w_{LS}) \ \text{ on } \partial\omega, \\ \sigma_1 \partial_n w'_{LS} = u'_n \ \text{ on } \partial\Omega, \end{cases}$$

and

$$\begin{cases} \Delta w'_{DLS} = 0 \ \text{ in } \Omega \setminus \overline{\omega} \text{ and in } \omega, \\ [w'_{DLS}] = h_n \dfrac{[\sigma]}{\sigma_1} \partial_{\mathbf{n}} w_{DLS}^- \ \text{ on } \partial\omega, \\ [\sigma \partial_n w'_{DLS}] = [\sigma] \mathrm{div}_\tau (h_n \nabla_\tau w_{DLS}) \ \text{ on } \partial\omega, \\ w'_{DLS} = \sigma_1 \partial_n u'_d \ \text{ on } \partial\Omega. \end{cases}$$

We now give the second order derivatives of the Least Square criterions J_{LS} and J_{DLS}:

Theorem 4. *Let Ω be a open subset of $\mathbb{R}^N$ with a C^2 boundary and ω a element of Ω_{d_0} with a $C^{4,\alpha}$ boundary. Let $\mathbf{h}_1$ and $\mathbf{h}_2$ be two deformation fields in $\mathcal{H}$. The Least Square objectives J_{LS} and J_{DLS} are twice differentiable with respect to the shape and their second derivatives in the direction* $\mathbf{h}$ *are given by*

$$\begin{aligned} D^2 J_{LS}(\omega)(\mathbf{h},\mathbf{h}) = & \int_{\partial\omega} \sigma_1 \partial_n w_{LS}'^{+} \, [(u_n')] + [\sigma \partial_n w_{LS}'] \, (u_n')^- - [\sigma \partial_n (u_n')] \, w_{LS}'^{-} \\ & + \int_{\partial\omega} \sigma_2 \partial_{\mathbf{n}} w_{LS}^{-} \, [(u_n)''] - \sigma_1 \partial_n (u_n')^+ \, [w_{LS}'] - w_{LS} \, [\sigma \partial_{\mathbf{n}} (u_n)''] \, , \\ D^2 J_{DLS}(\omega)(\mathbf{h},\mathbf{h}) = & \int_{\partial\omega} \left[\sigma \partial_n (u_d')\right] w_{DLS}'^{-} + \sigma_1 \partial_n (u_d')_1^+ \, [w_{DLS}'] - \sigma_1 \partial_n w_{DLS}'^{-} \left[(u_d')\right] \\ & + \int_{\partial\omega} [(u_d)''] + [\sigma \partial_n w_{DLS}'] \, (u_d')^- - w_{DLS} \, [\sigma \partial_{\mathbf{n}} (u_d)''] - \sigma_2 \partial_{\mathbf{n}} w_{DLS}^{-} \, . \end{aligned} \tag{8}$$

Let us investigate the properties of stability of this cost functions. We focus the study J_{LS} cost function but we can use the same techniques for J_{DLS}. We *assume* that there exists an admissible inclusion ω^* such that $J_{LS}(\omega^*) = 0$. It realizes the absolute minimum of the criterion J_{LS}. This is satisfied by solution of the inverse problem.Then, Euler's equation $DJ_{LS}(\omega^*)(\mathbf{h}) = 0$ holds and that we prove that

$$D^2 J_{LS}(\omega^*)(\mathbf{h},\mathbf{h}) = \int_{\Omega} (u_n')^2. \tag{9}$$

Moreover, if $h_n \neq 0$, then $D^2 J_{LS}(\omega^*)(\mathbf{h},\mathbf{h}) > 0$ holds. Nevertheless, (9) does not means that the minimization problem is well posed. In fact, the following theorem explains the instability of standard minimization algorithms.

Theorem 5. *Assume that ω^* is a critical shape of J_{LS} for which the additional condition $u_n = f$ holds, then the Riesz operator corresponding to $D^2 J_{LS}(\omega^*)$ defined from* $\mathrm{H}^{1/2}(\partial\omega^*)$ *with values in* $\mathrm{H}^{-1/2}(\partial\omega^*)$ *is compact.*

Theorem 5 has two main consequences. First, the shape Hessian at the global minimizer is not coercive. This means that this minimizer may be no local strict minimum of the criterion. Moreover, J_{LS} is not locally convex (at least uniformly in the directions of deformations) around the minimizer ω^*: the criterion provide no control of the distance between the parameter ω and the target ω^*. The second consequence concerns any numerical scheme used to obtain this optimal domain ω^*. One has to face this difficulty. This explains why frozen Newton schemes or Levenberg-Marquard schemes are used to numerically solve this problem [6, 2].

§3. Ideas of the proofs

3.1. Proof of Theorem 4

The differentiability of the objective is a direct application of Theorem 2. The computation we make here is based on the relation

$$D^2 J_{LS}(\omega)(\mathbf{h}_1,\mathbf{h}_2) = D\left(DJ_{LS}(\omega)\mathbf{h}_1\right)(\omega)\mathbf{h}_2 - DJ_{LS}(\omega)D\mathbf{h}_1\mathbf{h}_2). \tag{10}$$

To obtain (8), we first compute the shape gradient in the direction $\mathbf{h}_1$, then differentiate it in the direction of $\mathbf{h}_2$ to get

$$DJ_{LS}(\omega)\mathbf{h}_1 = \int_{\partial\Omega} (u_n - f)\,(u_n)'_1.$$

Then,

$$D\,(DJ_{LS}(\omega)\mathbf{h}_1)\,\mathbf{h}_2 = \int_{\partial\Omega} (u_n)'_1(u_n)'_2 + ((u_n)'_1)'_2(u_n - f).$$

Thanks to formula (10), we obtain

$$D^2 J_{LS}(\omega)(\mathbf{h}_1, \mathbf{h}_2) = \int_{\partial\Omega} (u_n)'_1(u_n)'_2 + (u_n)''_{1,2}(u_n - f). \tag{11}$$

Introducing the adjoint state function w_{LS} and the first derivative adjoint state w'_{LS}, we transform the integral on $\partial\Omega$ at integral on $\partial\omega$ thanks to Green's formulas:

$$\begin{aligned}\int_{\partial\Omega} (u_n)'_1(u_n)'_2 &= \int_{\partial\Omega} \sigma_1\partial_n w'_{LS}\,(u_n)'_1\\ &= \int_{\partial\omega} \sigma_1\partial_n(w'_{LS})^+\,[(u_n)'_1] + (u_n^-)'_1\,[\sigma_1\partial_n w'_{LS}] - [\sigma\partial_n(u_n)'_1]\,(w'_{LS})^- - [w'_{LS}]\,\sigma_1\partial_n(u_n^+)'_1,\end{aligned}$$

$$\int_{\partial\Omega} (u_n)''_{1,2}(u_n - f) = \int_{\partial\Omega} \sigma_1\partial_n w_{LS}\,(u_n)''_{1,2} = \int_{\partial\omega} \sigma_2\partial_n w^-_{LS}\left[(u_n)''_{1,2}\right] - w_{LS}\left[\sigma\partial_n(u_n)''_{1,2}\right].$$

We gather these formulae to obtain the result (8).

3.2. Sketch of proof of Theorem 5

We follow the strategy of analysis of [5, 3]. We specify the domain ω that is assumed to be a critical shape for J_{LS}. Moreover, we assume that the additional condition $u_n = f$ on $\partial\Omega$ holds, then the adjoint state $w_{LS} = 0$ in the Ω and the first derivative adjoint state w'_{LS} becomes :

$$\begin{cases} -\mathrm{div}\,(\sigma_\omega \nabla w'_{LS}) = 0 \text{ in } \Omega,\\ \qquad \sigma_1\partial_n w'_{LS} = u'_n \text{ on } \partial\Omega.\end{cases}$$

To emphasize that we deal with such a special domain, we will denote it by ω^*. The assumptions mean that the measurements are compatible and that ω^* is a global minimum of the criterion. From the necessary condition of order two at a minimum, the shape Hessian is positive at such a point.

Let us notice that only the normal component of $\mathbf{h}$ appears. Let us also emphasize that there is no hope to get $\mathbf{h} = 0$ from the structure theorem for second order shape derivative. The deformation field $\mathbf{h}$ appears in $D^2 J_{LS}(\omega^*)(\mathbf{h}, \mathbf{h})$ only thought its normal component h_n since ω^* is a critical point for J_{LS}. This remark explains why we consider in the statement of Theorem 5 the scalar Sobolev space corresponding to the normal components of the deformation field.

We now prove Theorem 5. From (8), we deduce

$$D^2 J_{LS}(\omega^*)(\mathbf{h}, \mathbf{h}) = \int_{\partial\omega^*} \sigma_1\partial_n w'^+_{LS}\,[(u'_n)] - \int_{\partial\omega^*} [\sigma\partial_n(u'_n)]\,w'^-_{LS}$$

Substituting their values to the quantities $[(u'_n)]$ and $[\sigma\partial_n(u'_n)]$, we get

$$D^2 J_{LS}(\omega^*)(\mathbf{h},\mathbf{h}) = [\sigma]\left(\left\langle \sigma_1 h_n \partial_n u_n^-, \partial_n w'^+_{LS}\right\rangle - \left\langle \mathrm{div}_\tau\left(h_n \nabla_\tau u_n\right), w'^-_{LS}\right\rangle\right),$$

where $\langle\cdot,\cdot\rangle$ denotes the duality between $\mathrm{H}^{1/2}(\partial\omega^*) \times \mathrm{H}^{-1/2}(\partial\omega^*)$. Let us introduce the operators

$$\begin{aligned} T_1 : \mathrm{H}^{1/2}(\partial\omega^*) &\to \mathrm{H}^{-1/2}(\partial\omega^*) & \qquad M_1 : \mathrm{H}^{1/2}(\partial\omega^*) &\to \mathrm{H}^{1/2}(\partial\omega^*) \\ \mathbf{h} &\mapsto \mathrm{div}_\tau\left(h_n \nabla_\tau u_n\right) & \mathbf{h} &\mapsto w'^-_{LS} \\ T_2 : \mathrm{H}^{1/2}(\partial\omega^*) &\to \mathrm{H}^{1/2}(\partial\omega^*) & M_2 : \mathrm{H}^{1/2}(\partial\omega^*) &\to \mathrm{H}^{-1/2}(\partial\omega^*) \\ \mathbf{h} &\mapsto h_n \partial_{\mathbf{n}} u_n^- & \mathbf{h} &\mapsto \partial_{\mathbf{n}} w'^+_{LS} \end{aligned}$$

The Hessian can then be written under the form

$$D^2 J_{LS}(\omega^*)(\mathbf{h},\mathbf{h}) = [\sigma]\left(\left\langle M_2(\mathbf{h}), T_2(\mathbf{h})\right\rangle - \sigma_1\left\langle T_1(\mathbf{h}), M_1(\mathbf{h})\right\rangle\right).$$

From the classical results of Maz'ya and Shaposhnikova on multipliers ([8]), we get easily that T_1 and T_2 are continuous operators. Operator M_1 is the composition of the operators

$$\begin{aligned} R_1 : \mathrm{H}^{1/2}(\partial\omega^*) &\to \mathrm{H}^{1/2}_\diamond(\partial\Omega) & \quad \text{and} \quad R_2 : \mathrm{H}^{1/2}_\diamond(\partial\Omega) &\to \mathrm{H}^{1/2}(\partial\omega^*) \\ \mathbf{h} &\mapsto u'_n & \phi &\mapsto \psi \end{aligned}$$

where ψ is the trace on $\partial\omega^*$ of Ψ solution of

$$\begin{cases} -\mathrm{div}\left(\sigma_{\omega^*}\nabla\Psi\right) = 0 & \text{in } \Omega, \\ \sigma_1 \partial_n \Psi = \phi & \text{on } \partial\Omega, \end{cases} \tag{12}$$

and $\mathrm{H}^{1/2}_\diamond(\partial\Omega)$ is the Sobolev space

$$\mathrm{H}^{1/2}_\diamond(\partial\Omega) = \left\{\phi \in \mathrm{H}^{1/2}(\partial\Omega) \ : \ \int_{\partial\Omega} \phi = 0\right\}.$$

While R_1 is a continuous operator, R_2 is compact. To prove this claim, let us express $u_{|\partial\omega^*} = \psi$. We use the integral representation formula and classical notation for the layers operators: we use the convention that the letter S is used for single layer potentials while K is used for double layer potentials. All the justifications of next claims are standart in the theory of integral equations. If u solves the boundary value problem (12), then it also solves the following system of integral equation

$$\begin{bmatrix} \frac{1}{2}I + \mu K_{\omega^*} & \kappa K_{\partial\Omega\partial\omega^*} \\ \mu K_{\partial\omega^*\partial\Omega} & \kappa\left(-\frac{1}{2}I + K_\Omega\right) \end{bmatrix} \begin{bmatrix} (u)_{|\partial\omega^*} \\ (u)_{|\partial\Omega} \end{bmatrix} = \kappa \begin{bmatrix} S_{\partial\Omega\partial\omega^*}\phi \\ S_\Omega \phi \end{bmatrix},$$

where $\kappa = -\sigma_1/(\sigma_1+\sigma_2)$ and $\mu = [\sigma]\,/\,(\sigma_1 + \sigma 2)$. The matricial operator arising in this equation has a continuous inverse. A straightforward computation gives that $u|_{\partial\omega^*} = \psi$ solves

$$\left[\left(\tfrac{1}{2}I + \mu K_{\omega^*}\right) + \mu K_{\partial\Omega\partial\omega^*}\left(-\tfrac{1}{2}I + K_\Omega\right)^{-1} K_{\partial\omega^*\partial\Omega}\right]\psi = \kappa\left[S_{\partial\Omega\partial\omega^*} - K_{\partial\Omega\partial\omega^*}\left(-\tfrac{1}{2}I + K_\Omega\right)^{-1} S_\Omega\right]\phi.$$

Since the operators $K_{\partial\Omega\partial\omega^*}$ and $S_{\partial\Omega\partial\omega^*}$ are compact, the operator R_2 is compact, hence M_1 is compact. The proof of compactness of M_2 is similar and therefore the Hessian is compact.

References

[1] Afraites, L., Dambrine, M., Eppler, K., and Kateb, D. Detecting perfectly insulated obstacles by shape optimization techniques of order two. *Discrete Contin. Dyn. Syst. Ser. B 8*, 2 (2007), 389–416 (electronic).

[2] Afraites, L., Dambrine, M., and Kateb, D. Shape methods for the transmission problem with a single measurement. *Numer. Funct. Anal. Optim. 28*, 5-6 (2007), 519–551.

[3] Afraites, L., Dambrine, M., and Kateb, D. On second order shape optimization methods for electrical impedance tomography. *SIAM J. Control Optim. 47*, 3 (2008), 1556–1590.

[4] Astala, K., and Päivärinta, L. Calderón's inverse conductivity problem in the plane. *Ann. of Math. (2) 163*, 1 (2006), 265–299.

[5] Eppler, K., and Harbrecht, H. A regularized Newton method in electrical impedance tomography using shape Hessian information. *Control Cybernet. 34*, 1 (2005), 203–225.

[6] Hettlich, F., and Rundell, W. The determination of a discontinuity in a conductivity from a single boundary measurement. *Inverse Problems 14*, 1 (1998), 67–82.

[7] Kirsch, A. The domain derivative and two applications in inverse scattering theory. *Inverse Problems 9*, 1 (1993), 81–96.

[8] Maz′ya, V. G., and Shaposhnikova, T. O. *Theory of multipliers in spaces of differentiable functions*, vol. 23 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[9] Nachman, A. I. Reconstructions from boundary measurements. *Ann. of Math. (2) 128*, 3 (1988), 531–576.

[10] Pantz, O. Sensibilité de l'équation de la chaleur aux sauts de conductivité. *C. R. Math. Acad. Sci. Paris 341*, 5 (2005), 333–337.

[11] Sylvester, J., and Uhlmann, G. A global uniqueness theorem for an inverse boundary value problem. *Ann. of Math. (2) 125*, 1 (1987), 153–169.

Lekbir Afraites
École Nationale des Sciences Appliquées, Safi
Université Cadi Ayyad, Maroc
Lekbir.Afraites@cea.fr

Marc Dambrine
Laboratoire de Mathématiques et de leurs Applications
Université de Pau et des Pays de l'Adour
Marc.Dambrine@univ-pau.fr

Djalil Kateb
Laboratoire de Mathématiques Appliquées de Compiègne
Université de Technologie de Compiègne
Djalil.Kateb@utc.fr

Monografías Matemáticas *García de Galdeano* **35**, 19–26 (2010)

A NEW MODIFIED EQUATION APPROACH FOR SOLVING THE WAVE EQUATION

C. Agut, J. Diaz and A. Ezziani

Abstract. The main topic of this work is to provide a fast and accurate solution of the wave equation. We will present new numerical schemes based on the modified equation technique using a switch between the space discretization and the time one. Numerical results illustrate the performances of these methods with respect to the accuracy and the computational burden.

Keywords: High order schemes, discontinuous Galerkin method, acoustic wave equation.

AMS classification: 65M12, 65M60, 35L05.

§1. Introduction

The solution of the full wave equation implies very high computational burdens to get high accurate results. Indeed, to improve the accuracy of the numerical solution, one must considerably reduce the space step, which is the distance between two points of the mesh representing the computational domain. Obviously this results in increasing the number of unknowns of the discrete problem. Besides, the time step, whose value fixes the number of required iterations for solving the evolution problem, is linked to the space step through the CFL (Courant-Friedrichs-Lewy) condition. The CFL number defines an upper bound for the time step in such a way that the smaller the space step is, the higher the numbers of iterations and of discrete unknowns will be. In the three-dimensional case, the problem can have more than ten millions of unknowns which must be evaluated at each time-iteration. However, high-order numerical methods can be used for computing accurate solutions with larger space and time steps. Recently, Joly and Gilbert (cf. [1]), have optimized the modified equation technique, which was proposed by Shubin and Bell (cf. [3]) for solving the wave equation and it seems to be very promising providing some improvements. In this work, we apply this technique in a original way. Indeed, most of the works devoted to the solution of the wave equation consider first the space discretization of the system before addressing the question of the time discretization. We intends here to invert the discretization process by applying first the time discretization thanks to the modified equation and after to consider the space discretization. After the time discretization an additional bilaplacian operator appears and we have therefore to consider C^1 finite elements (such as the Hermite ones) or Discontinuous Galerkin finite elements whose C^1 continuity is enforced through an appropriate penalty term. We provide a numerical comparison of the performance of the new method in order to illustrate the gains of accuracy and computational burden.

§2. Modified Equation technique

In this section, we describe the classical modified equation technique and we recall its main properties.

We consider the wave equation in a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$. We impose here a Neumann boundary condition of Ω but this study can be extended to other type of boundary condition without difficulty. Similarly, for a sake of simplicity, we do not consider any source term:

$$\begin{cases} \text{Find } u : (0,T) \times \Omega \to \mathbb{R} \text{ such that} & \\ \dfrac{\partial^2 u}{\partial t^2} - c^2 \Delta u = 0 & \text{in } (0,T) \times \Omega, \\ u(0,x) = u_0(x), \dfrac{\partial u}{\partial t}(0,x) = u_1(x) & \text{in } \Omega, \\ \nabla u \cdot \mathbf{n} = 0 & \text{on } \Gamma = \partial\Omega, \end{cases} \tag{1}$$

where T is the final time, c the velocity of the waves and u_0, u_1 are initial data. We assume here that the velocity is piecewise constant.

After a space discretisation (Finite Elements, Discontinous Galerkin, Finite Differences, etc.), the system can be rewritten as a linear system:

$$M \frac{d^2 U}{dt^2} + KU = 0, \tag{2}$$

where U is a vector whose components represent an approximation of u in a suitable basis of function, M is the mass matrix which is invertible and K is the stiffness matrix. To discretize (2) in time, we use Taylor expansions to obtain

$$\frac{U(t+\Delta t) - 2U(t) + U(t-\Delta t)}{\Delta t^2} = \frac{d^2 U(t)}{dt^2} + \frac{\Delta t^2}{12} \frac{d^4 U(t)}{dt^4} + O\left(\Delta t^4\right).$$

where Δt is the time step. Then, applying (2), we have that

$$\frac{d^4 U(t)}{dt^4} = M^{-1} K M^{-1} K U(t).$$

Consequently, we obtain an explicit fourth-order scheme:

$$U^{n+1} = 2U^n - U^{n-1} - \Delta t^2 \left[M^{-1} K \left(U^n - \frac{\Delta t^2}{12} \left(M^{-1} K U^n \right) \right) \right], \tag{3}$$

where U^n denotes the approximation of U at time $t = n\Delta t$.

This technique is the so called modified equation technique and was introduced by Shubin and Bell ([3]). We precise that it can be applied to obtain a scheme of arbitrary even order.

This scheme is stable under the following CFL condition [1]:

$$\frac{\Delta t}{h} \leq \alpha_{LF} \sqrt{3},$$

where h is the typical space step of the mesh and α_{LF} denotes the CFL condition we would have obtained with a classical leapfrog scheme:

$$U^{n+1} = 2U^n - U^{n-1} - \Delta t^2 M^{-1} K U^n. \tag{4}$$

We remark that this scheme requires one more multiplication by $M^{-1}K$ than the classical second order leapfrog scheme, but its CFL condition is multiplied by $\sqrt{3} \simeq 1.73$, so that it increases the order of convergence by two orders, without penalizing too much the computational burden.

§3. Scheme with the bilaplacian operator

We present here the construction of a new scheme using the modified equation technique by first applying the time discretization before the space one.

3.1. Construction of the semi-discrete scheme

Using Taylor expansions on the continuous unknown, we have

$$\frac{u(t+\Delta t) - 2u(t) + u(t-\Delta t)}{\Delta t^2} = \frac{d^2u(t)}{dt^2} + \frac{\Delta t^2}{12}\frac{d^4u(t)}{dt^4} + O\left(\Delta t^4\right).$$

Then, applying the wave equation (1) to the second and the fourth derivative of $u(t)$ with respect to the time, we easily obtain

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} = c^2 \Delta u^n + \frac{\Delta t^2}{12} c^4 \Delta^2 u^n. \tag{5}$$

In the following, this scheme will be called “scheme with bilaplacian operator”. To discretize the bilaplacian operator, we have to consider a space discretization which is able to take into account some H^2 quantities. Consequently, in this work, we have to consider C^1 finite elements (such as the Hermite ones) or Discontinuous Galerkin elements whose C^1 continuity is enforced through an appropriate penalty term.

3.2. Hermite finite elements

We first present the space discretization of (5) by Hermite elements. We restrict ourselves to the 1D-case since these elements are difficult to adapt to the higher dimensions.

Because of the bilaplacian operator, we need an additional boundary condition. Deriving two times the equation $\nabla u \cdot \mathbf{n} = 0$ with respect to the time and using the wave equation (1), we obtain

$$\frac{\partial^2 \nabla u}{\partial t^2} \cdot \mathbf{n} = \nabla \frac{\partial^2 u}{\partial t^2} \cdot \mathbf{n} = c^2 \nabla\left(\Delta u\right) \cdot \mathbf{n} = 0.$$

Consequently, we have to impose $\nabla u \cdot \mathbf{n} = 0$ and $\nabla\left(\Delta u\right) \cdot \mathbf{n} = 0$ on Γ. Similarly, for Dirichlet boundary conditions we would have $u = 0$ and $\Delta u = 0$ on Γ.

We multiply (5) by a test function $v \in H^2(\Omega)$, we integrate this equation over Ω, we apply Green formula and we use the two boundary conditions to obtain

$$\int_\Omega \left(\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} \right) v = a_1 \left(u^n, v\right) + \frac{\Delta t^2}{12} a_2 \left(u^n, v\right),$$

where

$$\begin{aligned} a_1 \left(u^n, v\right) &= -c^2 \int_\Omega \nabla u^n \cdot \nabla v, \\ a_2 \left(u^n, v\right) &= c^4 \int_\Omega \Delta u^n \Delta v - c^4 \int_\Gamma \Delta u^n \left(\nabla v \cdot \mathbf{n}\right) - c^4 \int_\Gamma \Delta v \left(\nabla u^n \cdot \mathbf{n}\right). \end{aligned}$$

The last term of a_2 which vanishes on Γ is artificially introduced to symmetrize the bilinear form.

We consider $\Omega = [a, b] \subset \mathbb{R}$ and we introduce the following space of discretization:

$$V_h = \{v \in C^1 \left(\Omega\right) : v_{|K} \in P^3([x_j, x_{j+1}]),\ \forall j = 1 \ldots n-1\}.$$

where $\{x_j\}_{j=1\ldots n}$ are defined by

$$\forall j = 1 \ldots n-1,\ x_j \in [a, b] \text{ and } x_j < x_{j+1}.$$

The basis functions of Hermite's element method are defined by

$$\forall 1 \leq i, j \leq n-1, \begin{cases} \varphi_{2i-1}(x_j) = \delta_{2i-1,j}, & \varphi_{2i}(x_j) = 0, \\ \varphi^{'}_{2i-1}(x_j) = 0, & \varphi^{'}_{2i}(x_j) = \delta_{2i,j}. \end{cases}$$

We finally obtain the following linear system:

$$\frac{U^{n+1} - 2U^n + U^{n-1}}{\Delta t^2} = M^{-1} K U^n, \tag{6}$$

with $M_{i,j} = \int_\Omega \varphi_i \varphi_j$, $K_{i,j} = \frac{\Delta t^2}{12} a_2(\varphi_i, \varphi_j) - a_1(\varphi_i, \varphi_j)$ and $U_i^n = u^n(x_i)$, if i is odd, or $(u^n)^{'}(x_i)$, if i is even.

The CFL condition of this scheme is given by the following result:

Theorem 1. *A necessary and sufficient L^2-stability condition is given by*

$$c \frac{\Delta t}{h} \leq \frac{1}{\sqrt{5}},$$

where $h = \underset{j=1\ldots n-1}{min} (x_{j+1} - x_j)$

Proof. We give the main ideas of the proof. The necessary condition is proved by a classical discrete Fourier analysis. Likewise, for the sufficient condition, we use an energy estimate to obtain

$$\lambda_{\min} \geq 0 \quad \text{and} \quad \frac{\Delta t^2}{4} \lambda_{\max} \leq 1,$$

where $\lambda_{\min} = \min\{\lambda \in \mathrm{Sp}(-M^{-1/2} K M^{-1/2})\}$ and $\lambda_{\max} = \max\{\lambda \in \mathrm{Sp}(-M^{-1/2} K M^{-1/2})\}$. □

Remark 1. The stability condition of this scheme is approximatively ${}^{\Delta t}/_{h} \leq 0.447$ and we have only one multiplication by $M^{-1}K$, whereas the stability condition of a P^3-Lagrange discretization combined with the classical modified equation technique is approximatively ${}^{\Delta t}/_{h} \leq 0.266$ and the scheme requires two multiplications by $M^{-1}K$. So the new scheme is 3.4 times faster.

In a strongly heterogeneous media, the solution is no longer C^1 because of the discontinuities of the physical parameters and Hermite elements are not adapted to this problem. Consequently, we introduce another method based on Discontinuous Galerkin method in the next section.

3.3. Discontinuous Galerkin Method

In this part, we use a Discontinuous Galerkin Method (DGM) which takes into account the discontinuities between each elements of the mesh $\mathcal{T}_h$ of Ω. More precisely, we use the Interior Penalty Discontinuous Galerkin Method [2]. First, we multiply (5) by a test function v, we integer it over each element K and we sum it over all elements of the mesh $\mathcal{T}_h$:

$$\sum_{K\in\mathcal{T}_h}\int_K \frac{u^{n+1}-2u^n-u^{n-1}}{\Delta t^2} v\,dx - \sum_{K\in\mathcal{T}_h}\int_K c^2\Delta u^n v\,dx - \frac{\Delta t^2}{12}\sum_{K\in\mathcal{T}_h}\int_K c^4\Delta^2 u^n v\,dx = 0.$$

Now, we have to introduce various notations. The set of the mesh faces are denoted $\mathcal{F}_h$ which is partitionned into two subsets $\mathcal{F}_h^i$ and $\mathcal{F}_h^b$ corresponding respectively to the interior faces and those located on the boundary. For $F \in \mathcal{F}_h^i$, we note arbitrarily K^+ and K^- the two elements sharing F and we define $\boldsymbol{\nu}$ as the unit outward normal vector pointing from K^+ to K^-.

Using a classical IPDG method, the second term of the formulation is replaced by the bilinear form a_1 defined by

$$\begin{aligned} a_1(u,v) = &\sum_{K\in\mathcal{T}_h}\int_K c^2\nabla u^n \nabla v\,dx - \sum_{F\in\mathcal{F}_h}\int_F [\![v]\!]\{\!\{c^2\nabla u^n\}\!\}\cdot\boldsymbol{\nu}\,d\sigma \\ &- \sum_{F\in\mathcal{F}_h}\int_F [\![u^n]\!]\{\!\{c^2\nabla v\}\!\}\cdot\boldsymbol{\nu}\,d\sigma + \sum_{F\in\mathcal{F}_h}\int_F \alpha_1 [\![u^n]\!]\,[\![v]\!]\,d\sigma, \end{aligned}$$

where α_1 is a well chosen penalization coefficient and $[\![\cdot]\!]$ and $\{\!\{\cdot\}\!\}$ correspond respectively to the jump and the average of a piecewise smooth function v, on an interior edge such that :

$$[\![v]\!] := v^+ - v^-, \quad \{\!\{v\}\!\} := \frac{v^+ + v^-}{2}.$$

We denote also by $v^\pm$ the restriction of v to the element $K^\pm$.

Now, we consider the third term of the formulation, denoted by Q. Using two times a Green formula, we obtain

$$Q = -\sum_{K\in\mathcal{T}_h}\int_K c^4\Delta u^n \Delta v\,dx + \sum_{K\in\mathcal{T}_h}\int_{\partial K} c^4\Delta u^n(\nabla v\cdot\mathbf{n})\,d\sigma - \sum_{K\in\mathcal{T}_h}\int_{\partial K} c^4(\nabla(\Delta u^n)\cdot\mathbf{n})\,v\,d\sigma.$$

Then, we can rewrite the second term and the third one:

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K} c^4\Delta u^n(\nabla v\cdot\mathbf{n})\,d\sigma = \sum_{F\in\mathcal{F}_h}\int_F [\![\nabla v]\!]\cdot\boldsymbol{\nu}\{\!\{c^4\Delta u^n\}\!\} + [\![c^4\Delta u^n]\!]\{\!\{\nabla v\}\!\}\cdot\boldsymbol{\nu}\,d\sigma,$$

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K} c^4(\nabla(\Delta u^n)\cdot\mathbf{n})v\,d\sigma = \sum_{F\in\mathcal{F}_h}\int_F [\![v]\!]\{\!\{c^4\nabla(\Delta u^n)\}\!\}\cdot\boldsymbol{\nu} + [\![c^4\nabla(\Delta u^n)]\!]\cdot\boldsymbol{\nu}\{\!\{v\}\!\}\,d\sigma.$$

Combining the continuity of u and $\nabla u\cdot\mathbf{n}$ across the interfaces with the wave equation (1), we deduce the continuity of Δu and $\nabla(\Delta u)\cdot\mathbf{n}$ so that

$$Q = -\sum_{K\in\mathcal{T}_h}\int_K c^4\Delta u^n\Delta v\,dx + \sum_{F\in\mathcal{F}_h}\int_F [\![\nabla v]\!]\cdot\boldsymbol{\nu}\{\!\{c^4\Delta u^n\}\!\} - \sum_{F\in\mathcal{F}_h}\int_F [\![v]\!]\{\!\{c^4\nabla(\Delta u^n)\}\!\}\cdot\boldsymbol{\nu}.$$

Since the form is not symmetric, we add the corresponding symmetric terms which vanish because of the continuity of u and $\nabla u\cdot\mathbf{n}$, and to enforce the coercivity of the form we add a suitable penalization term $\alpha_2\in\mathbb{R}$ to obtain the bilinear form

$$\begin{aligned} a_2(u,v) = Q_2 &+ \sum_{F\in\mathcal{F}_h}\int_F [\![\nabla u^n]\!]\cdot\boldsymbol{\nu}\{\!\{c^4\Delta v\}\!\}\\ &- \sum_{F\in\mathcal{F}_h}\int_F [\![u^n]\!]\{\!\{c^4\nabla(\Delta v)\}\!\}\cdot\boldsymbol{\nu} + \sum_{F\in\mathcal{F}_h}\int_F \alpha_2[\![c\nabla u\cdot\boldsymbol{\nu}]\!]\,[\![c\nabla v\cdot\boldsymbol{\nu}]\!].\end{aligned}$$

Then, we introduce the space of discretization $V_h = \{v\in L^2(\Omega) : v_{|K}\in P^3(K), \forall K\in\mathcal{T}_h\}$ and we consider $\{\varphi_j\}_{j=1\ldots n}$, the classical discontinuous basis functions P^3 of V_h to obtain the scheme

$$U^{n+1} = 2U^n - U^{n-1} + \Delta t^2 M^{-1}\left(\frac{\Delta t^2}{12}K_2 - K_1\right),$$

where $(M)_{i,j} = \sum_{K\in\mathcal{T}_h}\int_K \varphi_i\varphi_j$, $(K_1)_{i,j} = a_1(\varphi_i,\varphi_j)$ and $(K_2)_{i,j} = a_2(\varphi_i,\varphi_j)$.

Numerical results will illustrate the fact that this scheme has the same stability condition as the classical IPDG method combined with a leapfrog scheme.

§4. Numerical Results

In this part, we present some results in the one-dimensional case. Experiments in higher dimensions are in progress and preliminary results confirms the 1D results. In all the experiments, we consider a domain $\Omega = [0,10]$, a final time $T = 100$ and a velocity $c = 1$. We consider periodic boundary conditions, to ensure that the boundary conditions do not deteriorate the performances of the scheme. The initial conditions are

$$\begin{cases} U^0(x) = \sin(\pi x),\\ U^1(x) = \sin(\pi(x-\Delta t)),\end{cases}$$

so that the exact solution is $U(x,t) = \sin(\pi(x-t))$.

First, we compare the scheme with the bilaplacian operator to the classical P^3 FEM with the

Ndof	150	300	600	1200
P_3 FE	$\Delta x = 0.200$ $\Delta t = 0.0531$ Err = 3.39E−03	$\Delta x = 0.100$ $\Delta t = 0.0266$ Err = 2.66E−04	$\Delta x = 0.050$ $\Delta t = 0.0133$ Err = 1.75E−05	$\Delta x = 0.025$ $\Delta t = 0.0066$ Err = 1.11E−06
Δ^2 Hermite FE	$\Delta x = 0.133$ $\Delta t = 0.0584$ Err = 6.63E−03	$\Delta x = 0.067$ $\Delta t = 0.0294$ Err = 4.2E−04	$\Delta x = 0.033$ $\Delta t = 0.0147$ Err = 2.56E−05	$\Delta x = 0.017$ $\Delta t = 0.0073$ Err = 1.58E−06

Table 1: Comparison between P^3 FE and Δ^2 Hermite FE

Ndof	150	300	600	1200
DGP_3	$\Delta x = 0.256$ $\Delta t = 0.0681$ Err = 3.421E−03	$\Delta x = 0.132$ $\Delta t = 0.0349$ Err = 2.7006E−04	$\Delta x = 0.066$ $\Delta t = 0.0176$ Err = 1.809E−05	$\Delta x = 0.033$ $\Delta t = 0.0088$ Err = 1.158E−06
Δ^2 DGP_3	$\Delta x = 0.256$ $\Delta t = 0.0467$ Err = 3.297E−03	$\Delta x = 0.132$ $\Delta t = 0.0240$ Err = 1.717E−04	$\Delta x = 0.066$ $\Delta t = 0.0121$ Err = 8.088E−06	$\Delta x = 0.033$ $\Delta t = 0.0060$ Err = 4.337E−07

Table 2: Comparison between classical IPDG and Δ^2 IPDG

modified equation scheme. Table 1 presents the $L^2\,(]0,T[\,,\Omega)$-error with various choices of the number of degree of freedom (*Ndof*), the space step (Δx) and the time step (Δt).

We can easily remark that, with each method, the ratio between two consecutive errors is almost 16 that is to say the two methods are indeed fourth order methods. Furthermore, we note that the error is smaller with "P_3 FE" than with "Δ^2 Hermite FE" for a given number of degrees of freedom (i.e. for an equivalent computational burden at each time step). However, the same level of error as P^3 FE can be reached by decreasing the time step by 25%. Keeping in mind that the Δ^2 scheme requires only one multiplication by $M^{-1}K$, it is still less expensive than the classical one.

Now we present the results using a DGM with the same parameters as previously and $\alpha_1 = 8$ and $\alpha_2 = -10$ (cf. Table 2).

Once again, these results confirms that the methods are fourth order methods and we remark that the results with the scheme with the bilaplacian operator provides smaller error than the classical IPDG. Moreover, we notice that, with the bilaplacian operator, the time step is smaller than IPDG method but this problem is in balance with the fact that we have only one multiplication by $M^{-1}K$.

We now investigate the influence of the boundary conditions on the stability of the schemes. Table 3 represents the CFL conditions (numerically computed) for periodic, Neu-

	Periodic	Dirichlet	Neumann
Leapfrog scheme P_3	0.15333	0.15333	0.15333
FE P_3	0.26558	0.26558	0.26558
DG P_3	0.2655	0.2655	0.2655
Δ^2 Hermite FE	0.4471	0.4471	**0.1995**
Δ^2 DGP_3	0.1821	0.1821	0.1821

Table 3: Comparison CFL conditions

mann and Dirichlet conditions for the various schemes we have presented.

The boundary conditions do not modify the stability of the Δ^2 IPDG scheme, whereas the Neumann condition deteriorate the stability ot the Δ^2 Hermite scheme. Besides, since the IPDG scheme can be more easily extended to multidimensional cases and is more adapted to deal with heterogeneous media, we will focus on this method in future works.

§5. Conclusion

In this work, we have constructed a new scheme based on the modified equation technique and a switch between the time discretization and the space discretization. This new scheme allows to reduce the computational time and improve the accuracy of the classical methods. We are now considering the two dimensional case and heterogeneous media. Next step will be the implementation of absorbing boundary conditions.

References

[1] Gilbert, J.-C., and Joly, P. Higher order time stepping for second order hyperbolic problems and optimal CFL conditions. *Numerical Analysis and Scientific Computing for PDE's and their Challenging Applications* (2006).

[2] Grote, M. J., Schneebeli, A., and Schötzau, D. Dicontinuous galerkin finite element method for the wave equation. *SIAM J. Numer. Anal. 44* (2006), 2408–2431.

[3] Shubin, G. R., and Bell, J. B. A modified equation approach to constructing fourth-order methods for acoustic wave propagation. *SIAM J. Sci. Statist. Comput. 8* (1987), 135–151.

Cyril Agut and Julien Diaz
INRIA Research Center Bordeaux-Sud Ouest
Team-project Magique 3D
cyril.agut@inria.fr
julien.diaz@inria.fr

Abdelaaziz Ezziani
Laboratoire de Mathématiques Appliquées -
UMR CNRS 5142
Université de Pau et des Pays de l'Adour - Bt.
IPRA - BP 1155 64013 PAU Cedex
abdelaaziz.ezziani@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **35**, 27–32 (2010)

On the helical flow of Newtonian fluids induced by time dependent shear

W. Akhtar and M. Nazar

Abstract. The velocity field and the shear stresses corresponding to the unsteady flow of Newtonian fluids in an infinite circular cylinder are determined by means of the Hankel and Laplace transforms. The motion is produced by the infinite cylinder that at the initial moment is subject to both longitudinal and rotational time dependent shear stresses.

Keywords: Newtonian fluids, velocity field, tangential stress, cylindrical domains.
AMS classification: 53B25, 53C15.

§1. Introduction

The study on the flow of a viscous fluid in a circular cylinder is not only of fundamental theoretical interest but it also occurs in many applied problems. The starting solutions for the motion of the second grade fluids due to longitudinal and torsional oscillations of a circular cylinder have been studied by Fetecau in [3]. Vieru et al [6], by means of the Laplace transform and Cauchy's residue theorem, have determined the starting solutions for the oscillating motion of a Maxwell fluid. Akhtar and Nazar [1] have studied the rotational flow of generalized Maxwell fluids in a circular cylinder which rotates around its axis.

The aim of this paper is to study the flow of a Newtonian fluid in an infinite circular cylinder of radius R. The motion is produced by the cylinder that at the initial moment is subjected to longitudinal and torsional time dependent shear stresses. The exact solutions of the problems with initial and boundary conditions are determined by means of the finite Hankel and Laplace transforms. The solutions obtained in this paper can be used to make a comparison between flows of Newtonian and non-Newtonian fluids.

§2. Governing equations

The Cauchy stress in an incompressible Newtonian fluid is characterized by the next constitutive equation [5]:

$$\mathbf{T} = -p\mathbf{I} + \mu\mathbf{A}, \tag{1}$$

where $-p\mathbf{I}$ denotes the indeterminate spherical stress, $\mathbf{A} = \mathbf{L} + \mathbf{L}^T$ is the first Rivlin Ericksen tensor, $\mathbf{L}$ is the velocity gradient, μ is the dynamic viscosity, the superscript T denotes the transpose operator.

In cylindrical coordinates (r, θ, z), the velocity of the flow is given by

$$\mathbf{v} = \mathbf{v}(r,t) = w(r,t)\mathbf{e}_\theta + v(r,t)\mathbf{e}_z, \tag{2}$$

where $\mathbf{e}_\theta$ and $\mathbf{e}_z$ are the unit vectors in the θ and z directions respectively. For such flows the constraint of incompressibility is automatically satisfied.

Introducing (2) into constitutive equation (1), we find

$$\tau_1(r,t) = \mu \frac{\partial v(r,t)}{\partial r}, \tag{3}$$

$$\tau_2(r,t) = \mu\Big(\frac{\partial w(r,t)}{\partial r} - \frac{1}{r} w(r,t)\Big), \tag{4}$$

where $\tau_1(r,t) = S_{rz}(r,t)$ and $\tau_2(r,t) = S_{r\theta}(r,t)$ are the shear stress which is different of zero. The last equations together with the equations of motion leads to the governing equations [4]

$$\frac{\partial v(r,t)}{\partial t} = \nu\Big(\frac{\partial^2 v(r,t)}{\partial r^2} + \frac{1}{r}\frac{\partial v(r,t)}{\partial r}\Big), \quad r \in (0,R), \quad t > 0, \tag{5}$$

$$\frac{\partial w(r,t)}{\partial t} = \nu\Big(\frac{\partial^2 w(r,t)}{\partial r^2} + \frac{1}{r}\frac{\partial w(r,t)}{\partial r} - \frac{1}{r^2} w(r,t)\Big), \quad r \in (0,R), \quad t > 0, \tag{6}$$

where $\nu = \mu/\rho$ is the kinematic viscosity and ρ is the constant density of the fluid.

§3. Helical flow through an infinite circular cylinder

Let us consider an incompressible Newtonian fluid at rest in an infinite circular cylinder of radius R. At time zero, the cylinder suddenly begins to rotate and move along its axis due to time dependent shear stress. Owing to the shear, the fluid is gradually moved, its velocity being given by Eq.(2) and the governing equations are (5) and (6). The appropriate initial and boundary conditions are

$$v(r,0) = 0, \quad w(r,0) = 0; \quad r \in [0,R], \tag{7}$$

$$\tau_1(R,t) = \mu \frac{\partial v(R,t)}{\partial r} = f.t; \quad t \geq 0, \tag{8}$$

$$\tau_2(R,t) = \mu\Big(\frac{\partial w(R,t)}{\partial r} - \frac{1}{R} w(R,t)\Big) = f.t; \quad t \geq 0. \tag{9}$$

To solve this problem we shall use as in [1, 2] the Laplace and Hankel transforms.

3.1. Calculation of the velocity field

Applying the Laplace transform to Eqs. (5), (6), (8) and (9) and using Eq. (7) we obtain the following problems with boundary conditions

$$q\bar{v}(r,q) = \nu\Big(\frac{\partial^2 \bar{v}(r,q)}{\partial r^2} + \frac{1}{r}\frac{\partial \bar{v}(r,q)}{\partial r}\Big), \tag{10}$$

$$\frac{\partial \bar{v}(R,q)}{\partial r} = \frac{f}{\mu q^2}, \tag{11}$$

$$q\overline{w}(r,q) = \nu\Big(\frac{\partial^2 \overline{w}(r,q)}{\partial r^2} + \frac{1}{r}\frac{\partial \overline{w}(r,q)}{\partial r} - \frac{1}{r^2}\overline{w}(r,q)\Big), \tag{12}$$

$$\frac{\partial \overline{w}(R,q)}{\partial r} - \frac{1}{R}\overline{w}(R,q) = \frac{f}{\mu q^2}, \tag{13}$$

where

$$\overline{v}(r,q)=\int_0^\infty v(r,t)e^{-qt}dt,\quad \overline{w}(r,q)=\int_0^\infty w(r,t)e^{-qt}dt$$

are the Laplace transforms of $v(r,t)$ and $w(r,t)$ respectively. In the following we denote by

$$\overline{v}_H(r_{0n},q)=\int_0^R r\overline{v}(r,q)J_0(rr_{0n})\,dr,\quad \overline{w}_H(r_{1n},q)=\int_0^R r\overline{w}(r,q)J_1(rr_{1n})\,dr, \tag{14}$$

the finite Hankel transforms of $\overline{v}(r,q)$ and $\overline{w}(r,q)$ respectively, where $J_0(\cdot)$ and $J_1(\cdot)$ are the Bessel functions of first kind of order zero and one and r_{0n} and r_{1n}, for $n=1,2,3,\ldots$, are the positive roots of the transcendental equations $J_1(Rr)=0$ and $J_2(Rr)=0$ respectively.

Multiplying both sides of Eq. (10) by $rJ_0(rr_{0n})$, integrating with respect to r from 0 to R and taking into account the condition (11) and the equality

$$\int_0^R r\Big[\frac{\partial^2\overline{v}(r,q)}{\partial r^2}+\frac{1}{r}\frac{\partial\overline{v}(r,q)}{\partial r}\Big]J_0(rr_{0n})\,dr=\frac{RfJ_0(Rr_{0n})}{\mu q^2}-r_{0n}^2\overline{v}_H(r_{0n},q), \tag{15}$$

we find that

$$\overline{v}_H(r_{0n},q)=\frac{Rf}{\rho}J_0(Rr_{0n})\frac{1}{q^2(q+\nu r_{0n}^2)}. \tag{16}$$

Multiplying both sides of Eq. (12) by $rJ_1(rr_{1n})$, integrating with respect to r from 0 to R and taking into account the condition (13) and the equality

$$\int_0^R r\Big[\frac{\partial^2\overline{w}(r,q)}{\partial r^2}+\frac{1}{r}\frac{\partial\overline{w}(r,q)}{\partial r}-\frac{1}{r^2}\overline{w}(r,q)\Big]J_1(rr_{1n})\,dr=\frac{RfJ_1(Rr_n)}{\mu q^2}-r_{1n}^2\overline{w}_H(r_{1n},q), \tag{17}$$

we find that

$$\overline{w}_H(r_{1n},q)=\frac{Rf}{\rho}J_1(Rr_{1n})\frac{1}{q^2(q+\nu r_{1n}^2)}. \tag{18}$$

Now, for a more suitable presentation of the final results, we rewrite Eqs. (16) and (18) in the following equivalent forms

$$\overline{v}_H(r_{0n},q)=\overline{v}_{1H}(r_{0n},q)+\overline{v}_{2H}(r_{0n},q), \tag{19}$$

where

$$\overline{v}_{1H}(r_{0n},q)=\frac{RfJ_0(Rr_{0n})}{r_{0n}^2}\frac{1}{\mu q^2},\quad \overline{v}_{2H}(r_{0n},q)=-\frac{RfJ_0(Rr_{0n})}{\mu r_{0n}^2}\frac{1}{q(q+\nu r_{0n}^2)} \tag{20}$$

and

$$\overline{w}_H(r_{1n},q)=\overline{w}_{1H}(r_{1n},q)+\overline{w}_{2H}(r_{1n},q), \tag{21}$$

where

$$\overline{w}_{1H}(r_{1n},q)=\frac{RfJ_1(Rr_{1n})}{r_{1n}^2}\frac{1}{\mu q^2},\quad \overline{w}_{2H}(r_{1n},q)=-\frac{RfJ_1(Rr_{1n})}{\mu r_{1n}^2}\frac{1}{q(q+\nu r_{1n}^2)} \tag{22}$$

A straightforward calculus deals to the following function-Hankel transform pairs

$$f(r)=\frac{fr^2}{2R},\quad f_H(r_{0n})=\frac{RfJ_0(Rr_{0n})}{r_{0n}^2},\quad g(r)=\frac{fr^3}{2R^2},\quad g_H(r_{1n})=\frac{RfJ_1(Rr_{1n})}{r_{1n}^2}. \tag{23}$$

The inverse Hankel transforms of the functions $\bar{v}_{2H}(r_{0n}, q)$ and $\overline{w}_{2H}(r_{1n}, q)$ are [2]

$$\bar{v}_2(r,q) = \frac{2}{R^2}\sum_{n=1}^{\infty}\frac{J_0(rr_{0n})}{J_0^2(Rr_{0n})}\bar{v}_{2H}(r_{0n},q),$$
$$\overline{w}_2(r,q) = -2\sum_{n=1}^{\infty}\frac{r_{1n}^2 J_1(rr_{1n})}{[(r_{1n}^2+h^2)R^2-1]J_1^2(Rr_{1n})}\overline{w}_{2H}(r_{1n},q), \tag{24}$$

where $h = -\frac{1}{R}$.

Applying the inverse Hankel transform to Eqs. (19)-(22) and using (23) and (24) we obtain the following form of the Laplace transforms of the functions $v(r,t)$ and $w(r,t)$

$$\bar{v}(r,q) = \frac{fr^2}{2R}\frac{1}{\mu q^2} - \frac{2f}{\mu R}\sum_{n=1}^{\infty}\frac{J_0(rr_{0n})}{r_{0n}^2 J_0(Rr_{0n})}\frac{1}{q(q+\nu r_{0n}^2)}, \tag{25}$$

$$\overline{w}(r,q) = \frac{fr^3}{2R}\frac{1}{\mu q^2} - \frac{2f}{\mu R}\sum_{n=1}^{\infty}\frac{J_1(rr_{1n})}{r_{1n}^2 J_1(Rr_{1n})}\frac{1}{q(q+\nu r_{1n}^2)}. \tag{26}$$

We denote by

$$h_i(r_{in},q) = \frac{1}{q+\nu r_{in}^2},\ i = 0,1,$$

and have [2]

$$L^{-1}\{h_i(r_{in},q)\} = \exp(-\nu r_{in}^2).$$

The inverse Laplace transform of the function $g_i(r_{in},q) = \frac{1}{q}h_i(r_{in},q)$ is

$$g_i(r_{in},t) = \int_0^t h_i(r_{in},u) = \frac{1}{\nu r_{in}^2}[1-\exp(-\nu r_{in}^2 t)]. \tag{27}$$

Applying inverse Laplace transform to Eqs. (25) and (26) and using (27) we find the following forms of the velocity fields:

$$v(r,t) = \frac{fr^2}{2\mu R}t - \frac{2f}{\nu\mu R}\sum_{n=1}^{\infty}\frac{J_0(rr_{0n})}{r_{0n}^4 J_0(Rr_{0n})}[1-\exp(-\nu r_{0n}^2 t)], \tag{28}$$

and

$$w(r,t) = \frac{fr^3}{2\mu R^2}t - \frac{2f}{\nu\mu R}\sum_{n=1}^{\infty}\frac{J_1(rr_{1n})}{r_{1n}^4 J_1(Rr_{1n})}[1-\exp(-\nu r_{1n}^2 t)]. \tag{29}$$

3.2. Calculation of the shear stresses

Applying the Laplace transform to Eqs. (3) and (4) we find that

$$\overline{\tau}_1(r,q) = \mu\frac{\partial\bar{v}(r,q)}{\partial r}, \tag{30}$$

$$\overline{\tau}_2(r,q) = \mu\Big(\frac{\partial\overline{w}(r,q)}{\partial r} - \frac{1}{r}\overline{w}(r,q)\Big). \tag{31}$$

Differentiating Eqs. (25) and (26) with respect to r we get

$$\frac{\partial \overline{v}(r,q)}{\partial r} = \frac{rf}{R}\frac{1}{\mu q^2} + \frac{2f}{\mu R}\sum_{n=1}^{\infty}\frac{J_1(rr_{0n})}{r_{0n}J_0(Rr_{0n})}\frac{1}{q(q+\nu r_{0n}^2)}, \tag{32}$$

respectively

$$\frac{\partial \overline{w}(r,q)}{\partial r} - \frac{1}{r}\overline{w}(r,q) = \frac{fr^2}{R^2}\frac{1}{\mu q^2} + \frac{2f}{\mu R}\sum_{n=1}^{\infty}\frac{J_2(rr_{1n})}{r_{1n}J_1(Rr_{1n})}\frac{1}{q(q+\nu r_{1n}^2)}. \tag{33}$$

Introducing (32) into (30) and (33) into (31) we find that

$$\overline{\tau}_1(r,q) = \frac{rf}{R}\frac{1}{q^2} + \frac{2f}{R}\sum_{n=1}^{\infty}\frac{J_1(rr_{0n})}{r_{0n}J_0(Rr_{0n})}\frac{1}{q(q+\nu r_{0n}^2)}, \tag{34}$$

$$\overline{\tau}_2(r,q) = \frac{r^2 f}{R^2}\frac{1}{q^2} + \frac{2f}{R}\sum_{n=1}^{\infty}\frac{J_2(rr_{1n})}{r_{1n}J_1(Rr_{1n})}\frac{1}{q(q+\nu r_{1n}^2)}. \tag{35}$$

Inverting Eqs. (34) and (35) and using (27), we find the following forms of the shear stresses

$$\tau_1(r,t) = \frac{rft}{R} + \frac{2f}{\nu R}\sum_{n=1}^{\infty}\frac{J_1(rr_{0n})}{r_{0n}^3 J_0(Rr_{0n})}[1-\exp(-\nu r_{0n}^2 t)], \tag{36}$$

$$\tau_2(r,t) = \frac{r^2 ft}{R^2} + \frac{2f}{\nu R}\sum_{n=1}^{\infty}\frac{J_2(rr_{1n})}{r_{1n}^3 J_1(Rr_{1n})}[1-\exp(-\nu r_{1n}^2 t)]. \tag{37}$$

From (36) and (37) it is easy to verify that $\tau_1(R,t) = ft$ and $\tau_2(R,t) = ft, \quad t \geq 0$.

§4. Conclusion

In this note, the velocity field and the resulting shear stresses corresponding to the helical flow induced by an infinite circular cylinder in an incompressible Newtonian fluid have been determined using the finite Hankel and Laplace transforms. The motion is produced by the cylinder that at the initial moment is subjected to both rotation and translation by time dependent shear. The solutions that have been obtained satisfy all imposed initial and boundary conditions and can be used to make a comparison between flows of Newtonian and non-Newtonian fluids. For $t \to \infty$, the solutions (28) and (29) reduce to the steady-state solutions

$$v(r,t) = \frac{fr^2}{2\mu R}t - \frac{2f}{\nu\mu R}\sum_{n=1}^{\infty}\frac{J_0(rr_{0n})}{r_{0n}^4 J_0(Rr_{0n})},$$

and

$$w(r,t) = \frac{fr^3}{2\mu R^2}t - \frac{2f}{\nu\mu R}\sum_{n=1}^{\infty}\frac{J_1(rr_{1n})}{r_{1n}^4 J_1(Rr_{1n})}.$$

Acknowledgements

The authors would like to thank Prof. Constantin Fetecau for his suggestions and constructive comments.

References

[1] Akhtar, W., and Nazar, M. Exact solutions for the rotational flow of generalized Maxwell fluids in a circular cylinder. *Bull. Math. Soc. Sci. Math., Romanie, 51(99)*, 2 (2008).

[2] Debnath, L. and Bhatta, D. *Integral Transforms and Their Applications* (second ed.). Chapman and Hall/CRC Press, Boca Raton, London, New York, 2007.

[3] Fetecau, C., and Corina, F.. Starting solutions for the motion of a second grade fluid due to longitudinal and torsional oscillations of a circular cylinder. *Int. J. Eng. Sci. 44*, 11-12 (2006), 788–796.

[4] Fetecau, C., Corina, F., and Vieru, D. On some helical flows of Oldroyd-B fluids. *Acta Mech. 189* (2007), 53–63.

[5] Spurk, J. H., and Aksel, N. *Fluid Mechanics*. Springer-Verlag, Berlin-Heidelberg, Germany, 2008.

[6] Vieru, D., Akhtar, W., Corina, F., and Fetecau, C. Starting solutions for the oscillating motion of a Maxwell fluid in cylindrical domains. *Meccanica 42* (2006), 573–583.

Waseem Akhtar and Mudassar Nazar
Abdus Salam School of Mathematical Sciences GC University
68-B New Muslim Town Lahore
PAKISTAN
wasakh75@yahoo.com

Monografías Matemáticas *García de Galdeano* **35**, 33–40 (2010)

ASYMPTOTIC KINETIC ENERGY CONSERVATION FOR LOW-MACH NUMBER FLOW COMPUTATIONS

Mohamed Amara, Yann Moguen and Eric Schall

Abstract. Numerical dissipation, often used in collocated mesh schemes to enforce stability or to avoid odd-even decoupling problem, may be undesirable, for example to compute turbulent fluid flows in *DNS* or *LES*. Unfortunately, on the other hand, central discretization suffers from loss of stability problems, in particular when the Reynolds number increases. Therefore, an important attention has been devoted to find criteria that could ensure the stability without any addition of non-physical dissipation into the numerical schemes.

It appears experimentally that, for incompressible flow, the discrete kinetic energy conservation is one of these criteria. The present study deals with *(1)* the asymptotic signification of this conservation property in the incompressible limit of the compressible flow model; *(2)* the conditions under which the discrete kinetic energy is conserved in the sense previously evidenced; *(3)* the benefits that can be expected from this conservation property in the computations and the numerical problems that it does not prevent.

Keywords: Low-Mach number, kinetic energy conservation, Mach-uniformity, pressure correction, numerical dissipation.

AMS classification: 65N22, 76G25, 76M45, 76N15.

§1. Introduction

In the context of finite volume method, it has been proved recently in Georges *et al.* [1] that central interpolations ensure the discrete kinetic energy conservation in the “incompressible limit”, which is in fact reduced to the condition $\text{div}(v) = 0$ in this reference. Here, we are aiming to extend this result in the more general case of the incompressible limit in the *asymptotic* sense of this expression, that is, the limit of the compressible flow model when the reference Mach number of the flow goes to zero. Physically, this can be interpreted as the non-convertion of the kinetic energy into the elastic one allowing the propagation of sound waves.

First, following Nicoud [4], a single scale continuous asymptotic analysis is employed to precise the conditions under which the kinetic energy is conserved when the reference Mach number of the flow goes to zero. Secondly, this conservation property is investigated at the discretized level. Finally, few numerical experiments based on an “all-Mach” algorithm described in Ref. [3] demonstrate that the check board decoupling problem is unfortunately not avoided when the Mach number becomes sufficiently small before unity.

§2. Continuous asymptotics

We claim that the "incompressible limit" of the compressible flow model should be understood rather as the model obtained when the characteristic Mach number of the flow goes to zero in the general compressible flow model, than the divergence-free velocity flow model. In fact this latter is only a particular case of the asymptotic model. In this section, we recall shortly the derivation of the convective space and time scale continuous asymptotics, and few basic properties of it (see *e.g.* Ref. [2] for further details).

The Euler equations are written in dimensional form as

$$\begin{aligned}
\partial_{\hat{t}}\hat{\varrho} + \hat{\mathrm{div}}(\hat{\varrho}\hat{v}) &= 0,\\
\partial_{\hat{t}}(\hat{\varrho}\hat{v}) + \hat{\mathrm{div}}(\hat{\varrho}\hat{v}\otimes\hat{v}) + \hat{\nabla}\hat{p} &= 0,\\
\partial_{\hat{t}}(\hat{\varrho}\hat{E}) + \hat{\mathrm{div}}(\hat{\varrho}\hat{v}\hat{H}) &= 0,\\
\hat{E} &= \hat{e} + \hat{K},\\
\hat{\varrho}\hat{H} &= \hat{\varrho}\hat{E} + \hat{p},\\
\hat{\varrho}\hat{e} &= \frac{\hat{p}}{\gamma - 1},
\end{aligned}$$

where $\hat{\varrho}$, $\hat{v}$, $\hat{p}$, $\hat{e}$, $\hat{E}$, $\hat{H}$ and γ denote the density, velocity, pressure, internal energy, total energy, total enthalpy and the ratio of the specific heats at constant pressure and constant volume, respectively. The kinetic energy is $\hat{\varrho}\hat{K}$ with

$$\hat{K} = \frac{\hat{v}\cdot\hat{v}}{2}.$$

Let us suppose that the following reference quantities are given: length $\hat{l}_r$, density $\hat{\varrho}_r$, pressure $\hat{p}_r$ and norm velocity $\hat{v}_r$. Then, non-dimensional quantities are defined, $x = \hat{x}/\hat{l}_r$, $v = \hat{v}/\hat{v}_r$, $p = \hat{p}/\hat{p}_r$, $\varrho = \hat{\varrho}/\hat{\varrho}_r$, $t = \hat{t}/(\hat{l}_r/\hat{v}_r)$, $E = \hat{E}/(\hat{p}_r/\hat{\varrho}_r)$, $e = \hat{e}/(\hat{p}_r/\hat{\varrho}_r)$ and $H = \hat{H}/(\hat{p}_r/\hat{\varrho}_r)$. In the following, ∇ and div denote the gradient and the divergence operators with respect to the non-dimensionalized spatial variable x. The non-dimensional Euler equations read:

$$\begin{aligned}
\partial_t\varrho + \mathrm{div}(\varrho v) &= 0,\\
\partial_t(\varrho v) + \mathrm{div}(\varrho v\otimes v) + \frac{1}{\mathrm{M}^2}\nabla p &= 0, && (1)\\
\partial_t(\varrho E) + \mathrm{div}(\varrho v H) &= 0, && (2)\\
E &= e + \mathrm{M}^2 K, && (3)\\
\varrho H &= \varrho E + p, && (4)\\
\varrho e &= \frac{p}{\gamma - 1}, && (5)
\end{aligned}$$

where we set

$$\mathrm{M} = \sqrt{\gamma}\,\frac{\hat{v}_r}{\sqrt{\gamma\hat{p}_r/\hat{\varrho}_r}},\quad K = \frac{v\cdot v}{2}.$$

By taking the scalar product of the velocity with the momentum equation (1), one obtains the transport equation of the kinetic energy

$$\partial_t(\varrho K) + \operatorname{div}(v\varrho K) = \frac{1}{\mathrm{M}^2}(p \operatorname{div}(v) - \operatorname{div}(pv)). \tag{6}$$

Next, let us suppose that

$$p(x, t, \mathrm{M}) = \sum_{n=0}^{N} \mathrm{M}^n p^{(n)}(x, t) + o(\mathrm{M}^N), \quad N = 0, 1, 2, \quad \mathrm{M} \to 0,$$

with similar expansions for ϱ and v. Then, these expansions are substituted into the non-dimensional Euler equations. From the momentum equation, collecting coefficients of powers -2 and -1 in the characteristic Mach number M,

$$p^{(0)} = p^{(0)}(t), \quad p^{(1)} = p^{(1)}(t). \tag{7}$$

This means that, at convective space and time scale, the spatial pressure variations are taken into account by the second-order pressure $p^{(2)}$, called the hydrodynamic pressure. Now, from Eqs. (3), (5) and (7),

$$\partial_t(\varrho E) = \frac{1}{\gamma - 1}\mathrm{d}_t p^{(0)} + o(\mathrm{M}), \quad \mathrm{M} \to 0. \tag{8}$$

Consequently, energy equation (2) leads to

$$\mathrm{d}_t p^{(0)} + \gamma p^{(0)} \operatorname{div}(v^{(0)}) = 0. \tag{9}$$

On the other hand, from Eq. (6), the zeroth-order transport equation of the kinetic energy reads, after integration over the computational domain Ω,

$$-\int_{\partial\Omega} p^{(2)} v^{(0)} \cdot n = \partial_t \int_{\Omega} \varrho^{(0)} K^{(0)} + \int_{\partial\Omega} \varrho^{(0)} K^{(0)} v^{(0)} \cdot n - \int_{\Omega} p^{(2)} \operatorname{div}(v^{(0)}). \tag{10}$$

Consequently, a sufficient condition for the kinetic energy conservation in the incompressible limit (asymptotically) is that the zeroth-order velocity field is divergence-free. Let us notice that, from Eqs. (8) and (9), this condition is equivalent to

$$\mathrm{d}_t(\varrho E)^{(0)} = 0 \quad \text{or} \quad \mathrm{d}_t p^{(0)} = 0.$$

In this case the zeroth-order pressure of the flow is constant in time and space. In fact, $p^{(0)}$ (called the thermodynamical pressure) is related to the adiabatic compression of the gas flow, because it verifies:

$$\mathrm{D}_t \log \varrho^{(0)} = \mathrm{d}_t \log(p^{(0)})^{1/\gamma}, \quad \mathrm{D}_t \equiv \partial_t + v^{(0)} \cdot \nabla.$$

If $p^{(0)}$ is constant, then the power of the hydrodynamic pressure forces on the boundary of the computational domain equals the time variation of the the zeroth-order kinetic energy, plus its injection or evacuation by the boundary. This is the significance of the kinetic energy conservation in the incompressible limit, given by Eq. (10) through the analysis of the continuous flow model.

§3. Semi-discrete asymptotics

Let us now consider the discretized counterpart of the continuous property of kinetic energy conservation in the incompressible limit. We are aiming to conserve this property after applying the discretization procedure, when a first-order cell centered finite-volume method is used. We follow the presentation of Georges *et al.* [1], but here we adopt an asymptotic point of view. As in the continuous case, the convective time and space scale is considered.

Let $V^h \subset \Omega$ a polygonal bounded domain in $\mathbb{R}^d$ ($d = 1, 2$ or 3), which consists of cells V_i such that

$$V^h = \bigcup_i V_i \,; \quad \left|V_i \cap V_j\right| = 0,\ i \neq j.$$

S_{ij} is the surface (if $d = 3$), the edge (if $d = 2$) or the point (if $d = 1$) between two adjacent cells V_i and V_j.

Let us first introduce the asymptotic semi-discrete continuity operator. On each cell V_i, the semi-discrete zeroth-order mass equation reads:

$$|V_i|\, \mathrm{d}_t \varrho^{(0)} + \sum_{S_{ij}} (\varrho v)^{(0)}_{ij} \cdot n_{ij} \left|S_{ij}\right| = 0,$$

where n_{ij} denotes the V_i unit outer normal on S_{ij}. Here, mass fluxes, pressures and velocities are centrally interpolated. Thus, the asymptotic semi-discrete Continuity operator is defined on the cell V_i by

$$\mathrm{C}_i = |V_i|\, \mathrm{d}_t \varrho^{(0)}_i + \sum_{S_{ij}} F^{\mathrm{C}}_{ij}, \quad F^{\mathrm{C}}_{ij} = \frac{(\varrho v)^{(0)}_i + (\varrho v)^{(0)}_j}{2} \cdot n_{ij} \left|S_{ij}\right|.$$

Let us now consider the zeroth-order semi-discrete momentum equation, the k^{th} component of which reads

$$|V_i|\, \mathrm{d}_t(\varrho^{(0)}_i u^{(0)}_i) + \sum_{S_{ij}} (\varrho v)^{(0)}_{ij} \cdot n_{ij} u^{(0)}_{ij} \left|S_{ij}\right| + \sum_{S_{ij}} p^{(2)}_{ij} n^k_{ij} \left|S_{ij}\right| = 0, \tag{11}$$

where we note $u \equiv v^k$ for convenience. As a generalization of the two first terms of the left-hand side of Eq. (11), we set

$$\mathrm{D}(\phi)_i = |V_i|\, \mathrm{d}_t(\varrho^{(0)}_i \phi^{(0)}_i) + \sum_{S_{ij}} F^{\mathrm{D}}_{ij}(\phi), \quad F^{\mathrm{D}}_{ij}(\phi) = \frac{(\varrho v)^{(0)}_i + (\varrho v)^{(0)}_j}{2} \cdot n_{ij} \frac{\phi^{(0)}_i + \phi^{(0)}_j}{2} \left|S_{ij}\right|,$$

where ϕ is a scalar field on V^h, constant on each cell V_i. It is called the Divergence operator. Let us also introduce the Advection operator,

$$\mathrm{A}(\phi)_i = |V_i|\, \varrho^{(0)}_i \mathrm{d}_t \phi^{(0)}_i + \sum_{S_{ij}} F^{\mathrm{A}}_{ij}(\phi), \quad F^{\mathrm{A}}_{ij}(\phi) = \frac{(\varrho v)^{(0)}_i + \varrho v)^{(0)}_j}{2} \cdot n_{ij} \frac{\phi^{(0)}_j - \phi^{(0)}_i}{2} |S_{ij}|,$$

and the Skew-symmetric operator,

$$\mathrm{S}(\phi)_i = \frac{1}{2} \left(\mathrm{D}(\phi)_i + \mathrm{A}(\phi)_i\right).$$

A simple calculus leads to

$$\mathrm{S}(\phi)_i = \frac{|V_i|}{2}\left(\mathrm{d}_t(\varrho_i^{(0)}\phi_i^{(0)}) + \varrho_i^{(0)}\mathrm{d}_t\phi_i^{(0)}\right) + \sum_{S_{ij}} F^{\mathrm{S}}_{ij}(\phi), \quad F^{\mathrm{S}}_{ij}(\phi) = \frac{(\varrho v)_i^{(0)} + (\varrho v)_j^{(0)}}{2} \cdot n_{ij}\frac{\phi_j^{(0)}}{2}\left|S_{ij}\right|.$$

Moreover, as

$$\mathrm{D}(\phi)_i = \mathrm{S}(\phi)_i + \frac{\phi_i^{(0)}}{2}\mathrm{C}_i,$$

Eq. (11) yields

$$\sum_i \left\{ u_i^{(0)}\mathrm{S}(u^{(0)})_i + \frac{u_i^{(0)}}{2}\mathrm{C}_i + u_i^{(0)}\sum_{S_{ij}} F^p_{ij} \right\} = 0, \qquad (12)$$

where

$$F^p_{ij} = \frac{p_i^{(2)} + p_j^{(2)}}{2} n^k_{ij}\ \left|S_{ij}\right|.$$

It is worthwhile to notice that

$$\phi_i^{(0)}\mathrm{S}(\phi)_i = |V_i|\ \mathrm{d}_t\left(\frac{\varrho_i^{(0)}(\phi_i^{(0)})^2}{2}\right) + \sum_{S_{ij}} F^{\mathrm{KS}}_{ij}(\phi),$$

where

$$F^{\mathrm{KS}}_{ij}(\phi) = \frac{(\varrho v)_i^{(0)} + (\varrho v)_j^{(0)}}{2} \cdot n_{ij}\frac{\phi_i^{(0)}\phi_j^{(0)}}{2}\ \left|S_{ij}\right|.$$

Thus, Eq. (12) becomes

$$\sum_i \left\{ |V_i|\ \mathrm{d}_t\left(\frac{\varrho_i^{(0)}(u_i^{(0)})^2}{2}\right) + \sum_{S_{ij}} F^{\mathrm{KS}}_{ij}(u) + \frac{u_i^{(0)}}{2}\mathrm{C}_i + u_i^{(0)}\sum_{S_{ij}} F^p_{ij} \right\} = 0. \qquad (13)$$

Focusing on the last terms in the brackets, let us mention that

$$u_i^{(0)}\sum_{S_{ij}} F^p_{ij} = \sum_{S_{ij}} F^{\mathrm{K}p}_{ij}(u) - p_i^{(2)}\sum_{S_{ij}} \frac{u_i^{(0)} + u_j^{(0)}}{2} n^k_{ij}\ |S_{ij}| + p_i^{(2)}u_i^{(0)}\sum_{S_{ij}} n^k_{ij}\ |S_{ij}|, \qquad (14)$$

where

$$F^{\mathrm{K}p}_{ij}(u) = \frac{u_i^{(0)}p_j^{(2)} + u_j^{(0)}p_i^{(2)}}{2}\ n^k_{ij}\ \left|S_{ij}\right|.$$

As $F^{\mathrm{K}p}_{ji}(u) = -F^{\mathrm{K}p}_{ij}(u)$, Eq. (13) leads to

$$\sum_i \left\{ |V_i|\ \mathrm{d}_t\left(\frac{\varrho_i^{(0)}\|v_i^{(0)}\|^2}{2}\right) - p_i^{(2)}\sum_{S_{ij}} \frac{v_i^{(0)} + v_j^{(0)}}{2} \cdot n_{ij}\ \left|S_{ij}\right| \right\} = 0, \qquad (15)$$

where $\|\cdot\|$ denotes the euclidean norm in $\mathbb{R}^d$. Finally, one obtains the following result:

Theorem 1. *Consider the Euler equations discretized using a first-order cell-centered finite-volume method. Let us assume that the thermodynamical pressure $p^{(0)}$ is constant in time, or equivalently, the divergence of the zeroth-order velocity is zero at any time. Then, the discrete kinetic energy is conserved on the whole computational domain when the Mach number goes to zero, provided that:*

1. *convective terms are spatially discretized in skew-symmetric form,*
2. *mass fluxes, velocities and pressures are centrally interpolated at the cell interfaces.*

One observes that Eq. (15) is in accordance with the original continuous form (10) of the kinetic energy conservation equation in the incompressible limit. The numerical benefits that can be expected from this property are carried out by the non-growth of the sum of the square of the velocities. This contributes to ensure the stability of the time scheme without any explicit numerical dissipation to introduce.

Unfortunately, a glance at Fig. 1 suffices to realize that check board decoupling is carried on by central discretizations as the Mach number goes to zero. For this computation, an algorithm suggested by Nerinckx *et al.* [3] is used. In predictor/corrector form, it is based on the energy equation at the correction step, and enables one the proper handling of the pressure field. This one plays a specific role in the progressive decoupling between the flow equations when the Mach number goes to zero (see *e.g.* [2]).

A one-dimensional inviscid steady flow of perfect gas is considered in a nozzle with a variable section. The throat Mach number is about 10^{-6}. In Fig. 1, the asymmetry in the check board distribution along the nozzle is due to the boundary conditions. The flow is oriented from the left to the right. At the inlet, density and velocity are prescribed and the pressure gradient is zero. At the outlet, the pressure is prescribed, while velocity and density are allowing to float.

Semi-discrete asymptotic analysis enables one to explain the origin of the pressure numerical oscillations. Returning to the momentum equation, one has at orders -2 ($l = 0$) and -1 ($l = 1$):

$$\sum_{S_{ij}} \frac{p_i^{(l)} + p_i^{(l)}}{2} n_{ij}^k \left|S_{ij}\right| = 0, \quad l = 0, 1.$$

In fact, at convective time and space scale, $p^{(1)}$ disappears from the flow model in the incompressible limit (see *e.g.* [2]). Since at this scale the check board effect is due only to the hydrodynamic pressure $p^{(2)}$, it can be removed by the addition of an explicit numerical dissipation in the following form:

$$(\varrho v)_{ij} = \frac{(\varrho v)_i + (\varrho v)_j}{2} + \alpha_{ij}(\mathrm{M})\,(p_i - p_j), \quad \alpha_{ij}(\mathrm{M}) = O(1/\mathrm{M}^2),\ \mathrm{M} \to 0. \qquad (16)$$

The efficiency of this technique can be viewed in Fig. 2. Let us emphasize that, in this figure, a steady flow with constant boundary conditions is considered. In contrast, for example when acoustic pressure fluctuations are imposed at the outlet of the nozzle, the non-physical coupling between pressure and velocity influences the kinetic energy. This is due to the fact that $p^{(1)}$, which is identified as the acoustic pressure in the flow, is taken into account when the numerical dissipation (16) is applied.

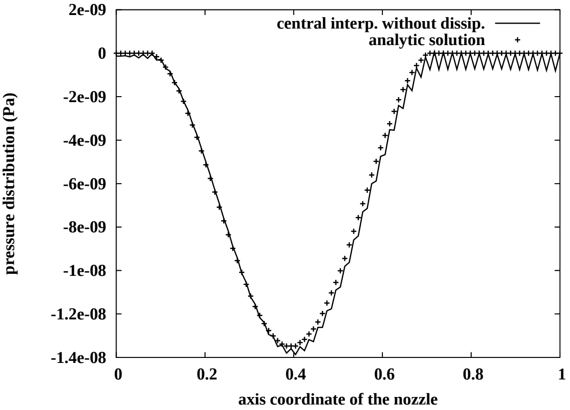


Figure 1: Pressure distribution (Pa) along the nozzle. Throat Mach number: 10^{-6}. Central discretization without numerical dissipation.

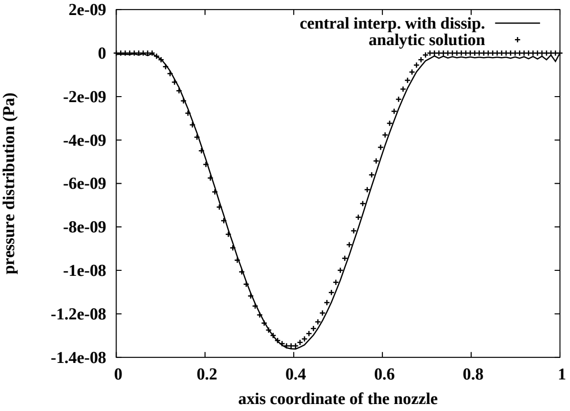


Figure 2: Pressure distribution (Pa) along the nozzle. Throat Mach number: 10^{-6}. Central discretizations with numerical dissipation.

§4. Conclusion

Kinetic energy conservation property in the incompressible limit was investigated at the continuous and discrete levels, through asymptotic analysis in convective time and space scale. Central interpolation of the pressures, mass fluxes and velocities allows one to conserve the continuous asymptotic property after applying the discretization procedure, provided that the skew-symmetric form is adopted for the convective terms. Unfortunately, check board effect is carried on by central interpolations. In the case of steady flows computations, a numerical dissipation enables one to avoid it.

References

[1] Georges, L., Winckelmans, G., and Geuzaine, P. Improving shock-free compressible RANS solvers for LES on unstructured meshes. *J. Comput. Appl. Math. 215* (2008), 419–428.

[2] Müller, B. Low Mach Number Asymptotics of the Navier-Stokes Equations and Numerical Implications. In *Lecture Series 1999-03* (March 1999), von Karman Institute for Fluid Dynamics.

[3] Nerinckx, K., Vierendeels, J., and Dick, E. Mach-uniformity through the coupled pressure and temperature correction algorithm. *J. Comput. Phys. 206* (2005), 597–623.

[4] Nicoud, F. Conservative High-Order Finite-Difference Schemes for Low-Mach Number Flows. *J. Comput. Phys. 158* (2000), 71–97.

Mohamed Amara
Laboratoire de Mathématiques et de leurs Applications - Bâtiment IPRA
Université de Pau et des Pays de l'Adour, Avenue de l'Université - F-64 013 Pau Cedex, France
mohamed.amara@univ-pau.fr

Yann Moguen and Eric Schall
Laboratoire de Thermique, Energétique et Procédés - IUT GTE
Université de Pau et des Pays de l'Adour, Avenue de l'Université - F-64 000 Pau, France
yann.moguen@free.fr and eric.schall@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **35**, 41–47 (2010)

NUMERICAL MODELLING OF POLLUTANT TRANSPORT

A. Balaguer, E. D. Fernández-Nieto, B. Latorre and V. Martínez

Abstract. In this work we study an artificial compression technique to treat discontinuities associated to linearly degenerated fields, with application to pollutant transport. The basic idea is to introduce a new flux in order to solve a new equation where the contact discontinuity is now a shock, travelling to the same velocity. We propose a flux-limiter method that combines the artificial compression technique and two second order methods. This method allows to apply the artificial compression technique in all the domain, without detecting the discontinuity jumps. We present a 2D test where the improvement of the presented technique can be observed.

Keywords: Artificial compression, pollutant transport, flux-limiters methods.

AMS classification: 65M05, 65M10.

§1. Introduction

A scalar conservation law, under certain regularity hypotheses, reduces to the partial differential equation

$$\begin{cases} u(x,t)_t + f(u(x,t))_x = 0, & (x,t) \in \mathbb{R}\times\mathbb{R}^+, \\ u(x,0) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{1}$$

where $u : \mathbb{R}\times\mathbb{R}^+ \to \mathbb{R}$ is the conserved variable and $f : \mathbb{R} \to \mathbb{R}$ is the flux function.

The pollutant transport is associated to linearly degenerated fields. If a is the velocity of the fluid, the pollutant concentration is the solution of the an advection equation

$$u_t + au_x = 0. \tag{2}$$

In the case that a is contant, the profile of the pollutant concentration can present contact discontinuities. The problem of the pressence of contact discontinuities is that they have a numerical diffusion more marked that shocks present in equations with non linear flux.

Harten presents a technique to treat this type of discontinuities in the pioneering work [3], dated in 1977 and in which some modifications of standard finite differences methods are discussed. Latter in 1989, Harten [4] again introduced the novel concept of subcell resolution.

Recently, the artificial compression method has been employed to improve the numerical solution of a great number of problems by using a great variety of techniques. For example, Yang [2], Lie and Noelle [5]. A brief description of the state of the art of this subject can be found in [1].

When the initial data of the problem has two constant states

$$u_0(x) = \begin{cases} u_-, & x < x_d, \\ u_+, & x > x_d. \end{cases}. \tag{3}$$

Martínez and Fernández-Nieto (see [6] and [1]) propose a procedure of artificial compression based on a modification of the flux to obtain a better numerical approach in the jumps of the solution. The idea is to detect the jump and to replace in this zone the linear flux in equation (2) by a nonlinear flux, so that the analytical solution in the original equation is conserved [6].

The objective of this work is to apply this technique to a second order scheme and to avoid the step of detection of discontinuity jumps. In Section 2 we propose a flux-limiter method based in the use of the compression technique and the combination of two second order methods. This method allows us to avoid the step of detection of discontinuity jumps and to improve the results of the second order methods. Finally, in Section 3 we present two numerical tests.

§2. Flux-limiter upwind method with artificial compression

In this section we first present a flux-limiter method that combines the first order upwind method with a second order one. For the second order method we present two possibilities: the classical Lax-Wendroff scheme (LW in what follows) and the second order upwind scheme (UP2 in what follows). After, we propose another scheme that uses a random combination of these two second order methods.

Numerical schemes using flux limiters can be defined by

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x}(\phi_{j+1/2}^n - \phi_{j-1/2}^n), \tag{4}$$

where

$$\phi_{j+1/2}^n = \phi_{j+1/2}^{1st} + \varphi(r_{j+1/2}^{2nd})(\phi_{j+1/2}^{2nd} - \phi_{j+1/2}^{1st}). \tag{5}$$

By $\phi_{j+1/2}^{1st}$ and $\phi_{j+1/2}^{2nd}$ we denote the numerical flux functions of first and second order respectively at time $t = t^n$. By $\varphi(r^{2nd})$ we denote a flux limiter function, which is defined in function of a non-dimensional quantity: r^{2nd}. The definition of r^{2nd} depends on the choice of the second order method. For the first order method we consider the upwind scheme:

$$\phi_{j+1/2}^{1st} = \frac{f(u_j) + f(u_{j+1})}{2} - \frac{1}{2}\left|a_{j+1/2}\right|(u_{j+1} - u_j), \tag{6}$$

where $a = \partial_u f$. And the compressed first order method reads

$$\phi_{j+1/2}^{1st,comp} = \frac{\tilde{f}(u_j) + \tilde{f}(u_{j+1})}{2} - \frac{1}{2}\left|\partial_u \tilde{f}_{j+1/2}\right|(u_{j+1} - u_j), \tag{7}$$

where $\tilde{f}(u)$ is defined by using an artificial compresion technique as follows. Following [1] we consider the flux

$$\tilde{f}(u) = au + g(u), \tag{8}$$

where

$$g(u) = \rho(u - u_-)(u - u_+), \tag{9}$$

where ρ is a parameter, which is chosen to assure the dynamical consistency of the jump (see [6] and [1]). It must verify:

$$\rho > 0, \text{ if } u_- > u_+ \quad \text{and} \quad \rho < 0, \text{ if } u_- < u_+. \tag{10}$$

In [6] it is proved that, if we consider a numerical scheme stable under a CFL condition λ_0 and if

$$|\rho| \leq \frac{\lambda_0 - |a| \frac{\Delta t}{\Delta x}}{\frac{\Delta t}{\Delta x} |u_- - u_+|}, \tag{11}$$

then the numerical scheme is also stable for the modified flux under the same CFL condition. For the numerical schemes that we consider in this work we have $\lambda_0 = 1$.

By $\partial_u \tilde{f}_{j+1/2}$ we denote the Roe average, that verifies

$$\tilde{f}(u_{j+1}) - \tilde{f}(u_j) = (\partial_u \tilde{f}_{j+1/2})(u_{j+1} - u_j). \tag{12}$$

As we mentioned previously, we consider two different possibilities for the second order method:

- Lax-Wendroff (LW):

$$\phi^{LW}_{j+1/2} = \frac{f(u_j) + f(u_{j+1})}{2} - \frac{1}{2}\frac{\Delta t}{\Delta x} a_{j+1/2}(f(u_{j+1}) - f(u_j)). \tag{13}$$

For Lax-Wendroff method $r^{2nd} = r^{LW}$ is defined by

$$r^{LW} = \begin{cases} (u_j - u_{j-1})/(u_{j+1} - u_j), & \text{if } a_{j+1/2} > 0, \\ (u_{j+1} - u_j)/(u_j - u_{j+1}), & \text{if } a_{j+1/2} < 0. \end{cases} \tag{14}$$

- Upwind second order ($UP2$):

$$\begin{aligned} \phi^{UP2}_{j+1/2} = \frac{1}{2}\Bigg(& f(u_j) + f(u_{j+1}) - \left|a_{j+1/2}\right| (u_{j+1} - u_j) \\ & + (1 - \lambda a^+_{j-1/2}) \frac{1 + \operatorname{sgn}(a_{j-1/2})}{2} (f(u_j) - f(u_{j-1})) \\ & - (1 + \lambda a^-_{j+3/2}) \frac{1 - \operatorname{sgn}(a_{j+3/2})}{2} (f(u_{j+2}) - f(u_{j+1})) \Bigg), \end{aligned}$$

where $a^\pm = (a \pm |a|)/2$. In this case $r^{2nd} = r^{UP2}$ is defined by

$$r^{UP2} = \begin{cases} (u_j - u_{j+1})/(u_{j-1} - u_j), & \text{if } a_{j+1/2} > 0, \\ (u_{j+1} - u_j)/(u_{j+2} - u_{j+1}) & \text{if } a_{j+1/2} < 0. \end{cases} \tag{15}$$

Finally, we present another numerical scheme based in a combination of previous one and the compression technique. The objective is to propose a new numerical scheme that improves previous second order methods, by using the artificial compression proposed technique, and to omit the detection of the discontinuity jump, that is, to apply the compression in all the domain without a conditionally jump detection.

The numerical flux function is

$$\phi_{j+1/2} = \phi^{1st,comp}_{j+1/2} + \varphi(r^{2nd}_{j+1/2})(\phi^{2nd}_{j+1/2} - \phi^{1st,comp}_{j+1/2}). \tag{16}$$

By $\phi^{1st,comp}_{j+1/2}$ we denote the numerical flux function of the first order upwind method applying the artificial compression technique (7). By $\phi^{2nd}_{j+1/2}$ we denote a second order method, for example LW or UP2. And by $\varphi(r)$ a flux-limiter function. For the numerical tests we have considered the minmod flux-limiter function.

Observe that the purpose to use flux-limiters functions is to combine two methods by applying the first order one near discontinuities and the second order one outside discontinuities. So, by applying the compression technique for the first order method only, we can omit the detection of discontinuity jumps.

Another improvement is the choice of the second order method. Instead of consider LW or UP2, we propose a combination of them. One possibility is to define $\phi^{2nd}_{j+1/2}$ as the mean average of LW and UP2. But in this case we must compute both fluxes. Another possibility is to choice in each intercell $j + 1/2$, one of them, for example we can use a random function to select of them. We have compared both possibilities for tests 4 and 5 and the final results are nearly the same. Then we only show the results corresponding to the cheaper possibility, the random choice.

The motivation to use a combination of LW and UP2 methods is illustrated in tests 1 and 2. We can observe that the results obtained by using the flux-limiter version with the LW and the UP2 method present a symmetrical and opposite behavior near discontinuities (see for example Figures 1(b) and 1(c)).

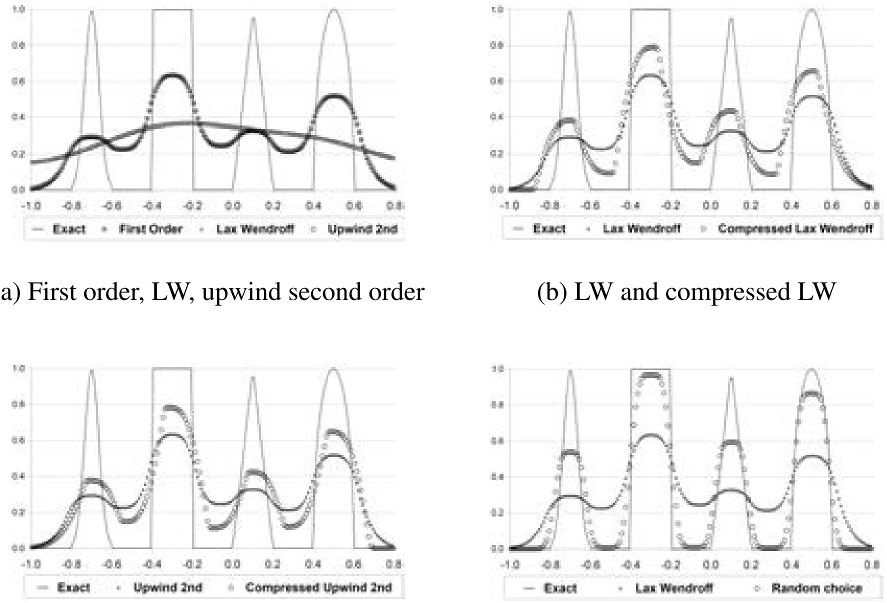
The artificial compression technique presented in the paper can be easily extended to 2D domains (see [1]). Basically, the finite volume method for 2D equations is based into apply a 1D flux at each edge of the 2D control volume. In test 2, we consider the same proposed combination using flux function (16), by combining the compressed first order upwind method, the 2D LW method and the 2D UP2 method.

§3. Numerical tests

3.1. Test 1: four profiles

We consider the following problem:

$$\begin{cases} u_t + u_x = 0, \quad -1 \leq x \leq 1, \\ u_0(x) = \begin{cases} e^{(\ln 2)\,(x+0.7)^2/0.0009}, & -0.8 \leq x \leq -0.6, \\ 1, & -0.4 \leq x \leq -0.2, \\ 1 - |10x - 1|, & 0 \leq x \leq 0.2, \\ \sqrt{1 - 100(x - 0.5)^2}, & 0.4 \leq x \leq 0.6, \\ 0, & \text{otherwise.} \end{cases} \end{cases} \tag{17}$$



(a) First order, LW, upwind second order

(b) LW and compressed LW

(c) Upwind second order and compressed upwind second order

(d) LW and compressed random choice

Figure 1: Test 1, comparison of first order, compressed Lax-Wendoff, compressed upwind second order and compressed random choice method.

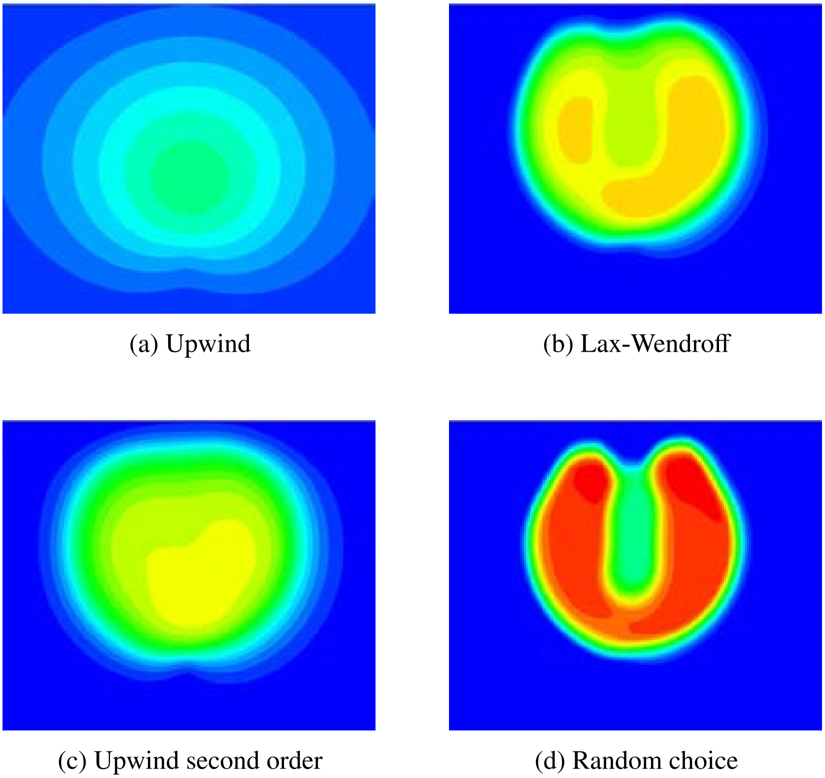
We consider $NX = 200$ points in $[-1, 1]$ and periodic boundary conditions. The final time is $t = 20$, and by the CFL condition we set $(\Delta t/\Delta x) = 0.5$. In Figure 1 we compare the results obtained with the first order upwind method (Figure 1(a)), the second order flux-limiter version with LW scheme (Figure 1(b)), the UP2 version (Figure 1(c)) and the proposed scheme (16), by using a random combination of LW and UP2 (Figure 1(d)). We observe that the less diffusive method is the proposed compressed random choice method. It improves the results for all the profiles.

3.2. Test 2: 2D test

In this subsection we consider a 2D problem, where the domain is $[0, 1]\times[0, 1]$. We discretize the domain in quadrangular cells, with $NX = NY = 200$. We consider the following problem:

$$\begin{cases} u_t + a(x, y)\, u_x + b(x, y)\, u_y = 0, \quad 0 \le x \le 1, \quad 0 \le y \le 1, \\ u_0(x, y) = \begin{cases} 1, & \text{if } (x, y) \in \Omega^1, \\ 0, & \text{otherwise,} \end{cases} \end{cases} \tag{18}$$

where Ω^1 is defined by the points (x, y) of the circle of ratio $r = 0.2$ and center $(0.5, 0.75)$, which are external to the rectangle $[0.475, 0.525] \times [0.65, 1]$.


(a) Upwind

(b) Lax-Wendroff

(c) Upwind second order

(d) Random choice

Figure 2: Test 2, t=4. (a) Upwind, (b) Flux limiter method with Lax-Wendroff, (c) Flux limiter method with Upwding second order (d) Random choice.

The velocity field is defined by a circular champ centered in $(0.5, 0.5)$:

$$a(x, y) = -2\pi\,(y - 0.5), \quad b(x, y) = 2\pi\,(x - 0.5). \tag{19}$$

Then, the test consist in a profile that is transported circularly around the center of the domain, $(0.5, 0.5)$. The time necessary to give a compleat turn is a period $T = 1$. By the CFL condition we set $(\Delta t/\Delta x) = \sqrt{2}/(4\pi)$.

In Figure 2 we present the level curves corresponding to the calculated profile at $t = 4T$. Figure 2(a) corresponds to the numerical result with the first order upwind method. Figure 2(b) corresponds to the LW with flux limiter scheme. Figure 2(c) corresponds to the UP2 method. And Figure 2(d) corresponds to the proposed scheme (16) in 2D. We observe that the proposed scheme present less diffusion in the four times presented.

Acknowledgements

The work of A. Balaguer has been supported by the Spanish Ministry of Education and Science and the FEDER in the framework of the projects CTM2006-11767 and CLG2006-

11242-C03/BTE. Fernández-Nieto has been partially supported by the Spanish Government Research project MTM2006-01275. B. Latorre has been partially supported by the Spanish Ministry of Science and Education under research project CGL2005-07059-C02-02.

References

[1] Fernández-Nieto, E., and Martínez, V. A treatment of discontinuities for nonlinear systems with linearly degenerate fields. *Computers & Fluids 36, n. 5* (2007), 987–1003.

[2] H., Y. A local extrapolation method for hyperbolic conservation laws. i. the eno underlying schemes. *J. Sci. Computing 15* (2000), 231–264.

[3] Harten, A. The artificial compression method for computation of shock and contact discontinuities I. Single conservation laws. *Commun. Pure Appl. Math. 30* (1977), 611–638.

[4] Harten, A. Eno schemes with subcell resolution. *J. Comp. Phys. 83* (1989), 148–184.

[5] Lie K.-A., N. S. On the artificial compression method for second-order nonoscillatory central difference schemes for systems of conservation laws. *SIAM J. Sci. Comput. 24* (2003), 1157–75.

[6] Martínez, V. An artifial compression procedure via flux correction, in book on "godunov methods: Theory and applications". *Kluwer Acad./Plenum Publ.* (2001), 595–602.

A. Balaguer
Departamento de Matemática Aplicada,
Universidad Politécnica de Valencia,
E.T.S.I. Geodésica, Cartográfica y Topográfica,
Camino de Vera s/n, 46022 Valencia, Spain
abalague@mat.upv.es

E. D. Fernández-Nieto
Departamento de Matemática Aplicada I,
Universidad de Sevilla. E.T.S. Arquitectura.
Avda, Reina Mercedes, s/n. 41012 Sevilla, Spain
edofer@us.es

B. Latorre
Departamento de Mecánica de Fluidos,
Universidad de Zaragoza
Centro Politécnico Superior
María de Luna 3. 50015 Zaragoza, Spain
borja.latorre@unizar.es

V. Martínez
Departamento de Matemáticas
Universitat Jaume I,
Campus de Riu Sec, 12071 Castelló, Spain
martinez@mat.uji.es

Monografías Matemáticas *García de Galdeano* **35**, 49–56 (2010)

Stability analysis of the Interior Penalty Discontinuous Galerkin method for solving the wave equation coupled with high-order absorbing boundary conditions

Hélène Barucq, Julien Diaz and Véronique Duprat

Abstract. In this paper, we study high-order absorbing boundary conditions (ABCs) for the acoustic wave equation the Higdon's one, which only take into account the propagative waves and Hagstrom-Warburton's one, which considers both the evanescent and proagative ones. We discretize the problem by a Discontinuous Galerkin (DG) method. Numerical results illustrate the instability of the method in particular cases.

Keywords: Absorbing boundary conditions, discontinuous Galerkin method, acoustic wave equation.

AMS classification: 65M12,65M60,35L05,35L20.

§1. Introduction

The numerical simulation of wave propagation generally involves boundary conditions which both represent the behavior at infinity and provide a mathematical tool to define a bounded computational domain in which a finite element method (FEM) can be applied. Most of these conditions are derived from the approximation of the Dirichlet-to-Neumann operator and when they both preserve the sparsity of the finite element matrix and enforce dissipation into the system, they are called absorbing boundary conditions. Most of the approximation procedures are justified into the hyperbolic region which implies that only the propagative waves are absorbed. If the exterior boundary is localized far enough from the source field, the approximation is accurate and the absorbing boundary condition is efficient. However, the objective is to use a computational domain whose size is optimized since the solution of wave problems requires to invert matrices whose order is very large and is proportional to the distance between the source field and the exterior boundary. Hence, it is a big deal to derive absorbing boundary conditions which are efficient when the exterior boundary is close to the source field and it is necessary to construct conditions which are efficient not only for propagative waves but both for evanescent and glancing waves. Recently, a new condition has been derived from an approximation of the Dirichlet-to-Neumann operator which is valid both for propagative and evanescent waves and extends the condition which was formerly proposed by Higdon [8]. By using a classical finite element scheme, Hagstrom et al. [7] have shown the improvements induced by the new condition. In this work, we intend to investigate whether the new condition can be introduced into a Interior Penalty Discontinuous Galerkin method [4] which is more accurate to reproduce the propagation of waves into heterogeneous media than standard FEMs. To analyze the impact of the new condition on the accuracy of the numerical solution, we also consider the Higdon condition and we compare the efficiency of the two conditions.

§2. Statement of the problem

In this section, we consider a model problem for the time-dependent wave equation in a two-dimensional domain Ω with a general ABC and we focus on the description of the Interior Penalty Discontinuous Galerkin (IPDG) method ([4]). We have:

$$(\mathcal{S})\begin{cases}\partial_t^2 u - \operatorname{div}\left(c^2\nabla u\right) = f, \quad \text{in } (0,T)\times\Omega,\\ u(0,x) = 0\ ;\ \partial_t u(0,x) = 0, \quad \text{in } \Omega,\\ \partial_{\mathbf{n}} u = 0, \quad \text{on } \Gamma_N,\\ \partial_{\mathbf{n}} u = B(\partial_t, \nabla_\Gamma)u, \quad \text{on } \Gamma_{\text{abs}},\end{cases}$$

where f is the source function, c the velocity of the wave, u the unknown field, T the final time, $\mathbf{n}$ the unit outward normal vector, Γ_N and Γ_{abs} respectively the boundary with the Neumann condition and the ABC which is represented by the operator B. The operator B is differential, for instance, it reads $\frac{1}{c}\partial_t$ which corresponds to the simplest ABC. We refer to [1], where the well-posedness of problem $(\mathcal{S})$ has been established for $f \in L^2(0,T;L^2(\Omega))$ by applying the semi-group theory. More precisely, if $\mathcal{U} = \{u \in H^1(\Omega), \partial_n u \in L^2(\Gamma_{\text{abs}})\}$, $u \in C^0(0,T;\mathcal{U}) \cap C^1(0,T;L^2(\Omega))$.

We consider a partition $\mathcal{T}_h$ of Ω composed of triangles K, we denote by Ω_h the set of triangles, by Σ_{abs} the set of the edges on the absorbing boundary, by Σ_{N} the set of the edges on the Neumann boundary and by Σ_{i} the set of the edges in the domain such that $\Sigma_{\text{i}} \cap (\Sigma_{\text{N}} \cup \Sigma_{\text{abs}}) = \emptyset$. For each $\Sigma \in \Sigma_{\text{i}}$, we have to distinguish the two triangles that share Σ: we note them arbitrarily K^+ and K^- . We introduce useful notations to define the jump and the average over edges:

$$[\![v]\!] := v^+\boldsymbol{\nu}^+ + v^-\boldsymbol{\nu}^- \quad \text{and} \quad \{\!\{v\}\!\} := \frac{v^+ + v^-}{2},$$

where v^+ and v^- respectively refers to the restriction of v in K^+ and K^- and $\boldsymbol{\nu}^\pm$ stands for the unit outward normal vector to $K^\pm$.

It is well-known the IPDG formulation of $(\mathcal{S})$ reads as ([4]):

$$\begin{cases}\text{Find } u \in \mathcal{U} \text{ such that } \forall v \in \mathrm{H}^1,\\ \displaystyle\sum_K \int_K \partial_t^2 uv + a(u,v) - \sum_{\Sigma\in\Sigma_{\text{abs}}} \int_\Sigma c^2 \partial_{\mathbf{n}} uv = \sum_K \int_K fv,\end{cases}$$

with

$$a(u,v) = \sum_K \int_K c^2\nabla u\nabla v - \sum_{\Sigma\in\Sigma_{\text{i}}} \int_\Sigma \Big(\{\!\{v\}\!\}[\![c\nabla u]\!] + \{\!\{u\}\!\}[\![c\nabla v]\!] + \alpha[\![u]\!][\![v]\!]\Big).$$

We seek an approximation of the solution in the finite element space V_h^k defined as follows:

$$V_h^k = \left\{ v \in L^2(\Omega); v_{|K} \in P^k, \forall K \right\}, k \in \mathbb{N}$$

where P^k is the set of polynomials of degree at most k on K.

§3. The Higdon's Condition

Here, we are going to study ABCs derived from a transparent boundary condition which only take the propagative waves into account. We will also discuss the implementation of those high-order conditions in the IPDG scheme.

We recall the Higdon's condition of order p, ($p \in \mathbb{N}$) (cf. [8]):

$$\prod_{j=1}^{P} (\cos a_j \, \partial_t + c \, \partial_{\mathbf{n}}) \, u = 0, \quad \text{on } \Gamma_{abs}. \tag{1}$$

Remark 1. The Engquist-Majda's condition (cf. [2]), which was one of the first ABCs to be designed, is a particular case of the Higdon one. Indeed, it is obtained by choosing all a_j equal to zero in (1).

To implement this condition in a numerical scheme, we define auxiliary functions u_j, for $1 \leq j \leq P$ on the absorbing boundary (cf. [3]):

$$\begin{cases} (\cos a_1 \partial_t + c\partial_{\mathbf{n}}) u = \partial_t u_1, & \\ (\cos a_j \partial_t + c\partial_{\mathbf{n}}) u_{j-1} = (\cos a_j \partial_t - c\partial_{\mathbf{n}}) u_j, & j = 2, \dots, P, \\ u_j(0, .) = 0, & j = 1, \dots, P. \end{cases}$$

By this way, we avoid to use high-order differential operators into the variational formulation. Indeed, it has been shown in [6] that

$$\prod_{j=1}^{P} (\cos a_j \, \partial_t + c \, \partial_{\mathbf{n}}) \, u = 0 \iff u_P = 0$$

and

$$(\partial_t^2 - \Delta) u_j = 0, \ \forall j = 1, \dots, P.$$

Then, thanks to these two properties, we can rewrite the problem including now P differential equations on the boundary which can be easily included and we obtain the following system:

$$\begin{cases} \partial_t^2 u - c^2 \triangle u = f, \quad \text{in } \Omega, \\ \partial_{\mathbf{n}} u = 0, \quad \text{on } \Gamma_N, \\ (\cos a_1 \partial_t + c\partial_{\mathbf{n}}) = \partial_t u_1, \quad \text{on } \Gamma_{\text{abs}}, \\ 2\cos a_2 (1 - \cos^2 a_1) \partial_t^2 u + l_{1,1} \partial_t^2 u_1 + (1 - \cos^2 a_2) \partial_t^2 u_2 \\ \qquad\qquad = c^2 (2 \cos a_2 \partial_\tau^2 u + \partial_\tau^2 u_1 + \partial_\tau^2 u_2), \quad \text{on } \Gamma_{\text{abs}}, \\ l_{j,j-1} \partial_t^2 u_{j-1} + l_{j,j} \partial_t^2 u_j + l_{j,j+1} \partial_t^2 u_{j+1} \\ \qquad\qquad = c^2 (m_{j,j-1} \partial_\tau^2 u_{j-1} + m_{j,j} \partial_\tau^2 u_j + m_{j,j+1} \partial_\tau^2 u_{j+1}), \quad \text{for } j = 2, \dots, P-1, \ \text{on } \Gamma_{\text{abs}}, \\ u_P = 0, \quad \text{on } \Gamma_{\text{abs}}, \end{cases}$$

where τ is the tangential component such that $(\mathbf{n},\tau)$ is a direct basis and

$$\begin{cases} l_{1,1} = 1 + 2\cos a_2 \cos a_1 + \cos^2 a_2, \\ l_{j,j-1} = \cos a_{j+1}(1-\cos^2 a_j), \\ l_{j,j} = \cos a_{j+1}(1+\cos^2 a_j) + \cos a_j(1+\cos^2 a_{j+1}), \\ l_{j,j+1} = \cos a_j(1-\cos^2 a_{j+1}), \end{cases}$$

and

$$\begin{cases} m_{j,j-1} = \cos a_{j+1}, \\ m_{j,j} = \cos a_{j+1} + \cos a_j, \\ m_{j,j+1} = \cos a_j. \end{cases}$$

Now, let us introduce the approximation space to discretize the ABC. Let W_h^k be defined as

$$W_h^k = \left\{ w \in L^2(\Gamma_{\text{abs}});\ w_{|\Sigma} \in P^k(\Sigma), \forall \Sigma \in \Sigma_{\text{abs}} \right\}.$$

The equations on Γ_{abs} are discretized by a 1D IPDG approximation and we define similar notations to the 2D case. N_{abs} is the set of the vertices of the edges of Σ_{abs}; for each point p in N_{abs}, we arbitrarily denote by Σ^+ and Σ^- the two edges sharing p, and by $\boldsymbol{\nu}^\pm$ the unit tangent vector to $\Sigma^\pm$ in p. The definition of the jumps and the averages are the same as in Section 2.

For a given j, consider the equation

$$l_{j,j-1}\partial_t^2 u_{j-1} + l_{j,j}\partial_t^2 u_j + l_{j,j+1}\partial_t^2 u_{j+1} = c^2(m_{j,j-1}\partial_\tau^2 u_{j-1} + m_{j,j}\partial_\tau^2 u_j + m_{j,j+1}\partial_\tau^2 u_{j+1}),$$

whose variational formulation reads as

$$\forall w \in H^1(\Gamma_{\text{abs}}),\ \sum_{\Sigma\in\Sigma_{\text{abs}}} \int_\Sigma \left(l_{j,j-1}\partial_t^2 u_{j-1} + l_{j,j}\partial_t^2 u_j + l_{j,j+1}\partial_t^2 u_{j+1} \right) w$$
$$= -m_{j,j-1}a_{j,j-1}(u_{j-1},w) - m_{j,j}a_{j,j}(u_j,w) - m_{j,j+1}a_{j,j+1}(u_{j+1},w),$$

where

$$a_{i,j}(u,w) = \sum_{\Sigma\in\Sigma_{\text{abs}}} \int_\Sigma c^2 \partial_\tau u \partial_\tau w - \sum_{z\in N_{\text{abs}}} \left(\{\!\{w\}\!\}[\![u]\!] + \{\!\{u\}\!\}[\![w]\!] - \alpha_{i,j}[\![u]\!][\![w]\!] \right)$$

and $\alpha_{i,j}$ is the penalization term depending on $\cos a_i$ and $\cos a_j$.

We obtain then,

$$\begin{cases} M\dfrac{d^2U}{dt^2} + C\dfrac{dU}{dt} + KU = F + G\dfrac{dU^1}{dt}, \quad \text{in } \Omega, \\ B_1\dfrac{d^2U}{dt^2} + l_{1,1}B_2\dfrac{d^2U^1}{dt^2} + (1-\cos^2 a_2)B_2\dfrac{d^2U^2}{dt^2} + EU + DU^1 + DU^2 = 0, \quad \text{on } \Gamma_{\text{abs}}, \\ l_{j,j-1}B_2\dfrac{d^2U^{j-1}}{dt^2} + l_{j,j}B_2\dfrac{d^2U^j}{dt^2} + l_{j,j+1}B_2\dfrac{d^2U^{j+1}}{dt^2} \\ \quad + m_{j,j-1}DU^{j-1} + m_{j,j}DU^j + m_{j,j+1}DU^{j+1} = 0, \quad \text{for } j = 2,\dots,P-1, \text{ on } \Gamma_{\text{abs}}, \\ U^P = 0, \quad \text{on } \Gamma_{\text{abs}}, \end{cases}$$

where U is the solution vector, U^j the auxiliary functions, M the mass matrix, K the stiffness matrix, F the source vector and all the other matrices come from the ABC.

To simplify, we rewrite this system. We have:

$$R\frac{d^2X}{dt^2} + S\frac{dX}{dt} + TX = \begin{pmatrix} F \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where X is the vector of all the unknowns (U and U^j).

Next, we apply a time-discretization using a second-order finite difference scheme:

$$\left(R + \frac{\Delta t}{2}S\right)X^{n+1} = \Delta t^2 \begin{pmatrix} F(n\Delta t, .) \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \Delta t^2 TX^n + 2RX^n - RX^{n-1} + \frac{\Delta t}{2}SX^{n-1},$$

with $X^n = X(n\Delta t)$ and Δt is the time step. Note that, since M, B_1, B_2, C and G are block-diagonal matrices, $\left(R + \frac{\Delta t}{2}S\right)$ is easily invertible.

§4. The Hagstrom-Warburton's condition

In this section, we study a new condition proposed by T. Hagstrom and T. Warburton [7] which takes into account not only propagative waves but also evanescent waves. More accuracy is then expected.

The Hagstrom-Warburton's ABC (H-W ABC) of order $P + Q$, ($P, Q \in \mathbb{N}$) is given by

$$\left[\prod_{j=1}^{Q}(\sigma_j + \partial_{\mathbf{n}})\right]\left[\prod_{j=1}^{P}(\cos a_j\, \partial_t + c\partial_{\mathbf{n}})\right]u = 0. \tag{2}$$

For the same reasons as for the Higdon's ABC, we introduce auxiliary functions defined on the absorbing boundary:

$$\begin{cases} (\cos a_1\partial_t + c\partial_{\mathbf{n}})u = \cos a_1\partial_t u_1, \\ (\cos a_j\partial_t + c\partial_{\mathbf{n}})u_{j-1} = (\cos a_j\partial_t - c\partial_{\mathbf{n}})u_j, \quad \text{for } 2 \le j \le P, \\ (\sigma_j + \partial_{\mathbf{n}})u_{P+j-1} = (\sigma_j - \partial_{\mathbf{n}})u_{P+j}, \quad \text{for } 1 \le j \le Q, \\ u_j((x,y),0) = 0, \quad \text{for } 1 \le j \le P + Q. \end{cases}$$

As for the Higdon's ABC, we have (cf. [6]):

$$\left[\prod_{j=1}^{Q}(\sigma_j + \partial_{\mathbf{n}})\right]\left[\prod_{j=1}^{P}(\cos a_j\, \partial_t + c\partial_{\mathbf{n}})\right]u = 0 \iff u_{P+Q} = 0$$

and

$$\forall j \in 1, \dots, P + Q,\ (\partial_t^2 - \Delta)u_j = 0.$$

Hence, the system can be rewritten in a more convenient way (cf. [5]). The approach is the same as before except when j is equal to P. For $j < P$ or $j > P$, we get:

$$\begin{cases} 2\cos a_2(1-\cos^2 a_1)\partial_t^2 u + l_{1,1}\cos a_1 \partial_t^2 u_1 + \cos a_1(1-\cos^2 a_2)\partial_t^2 u_2 \\ \qquad\qquad = 2c^2\cos a_2 \partial_\tau^2 u + c^2(\cos a_1 \partial_\tau^2 u_1 + \cos a_1 \partial_\tau^2 u_2), \quad \text{on } \Gamma_{\text{abs}}, \\ l_{j,j-1}\partial_t^2 u_{j-1} + l_{j,j}\partial_t^2 u_j + l_{j,j+1}\partial_t^2 u_{j+1} \\ \qquad\qquad = c^2(m_{j,j-1}\partial_\tau^2 u_{j-1} + m_{j,j}\partial_\tau^2 u_j + m_{j,j+1}\partial_\tau^2 u_{j+1}), \quad \text{for } j = 2,\dots,P-1, \text{ on } \Gamma_{\text{abs}}, \\ \bar{l}_{j,j-1}\partial_t^2 u_{P+j-1} + \bar{l}_{j,j}\partial_t^2 u_{P+j} + \bar{l}_{j,j+1}\partial_t^2 u_{P+j+1} \\ \qquad\qquad = c^2(\bar{m}_{j,j-1}\partial_\tau^2 u_{P+j-1} + \bar{m}_{j,j}\partial_\tau^2 u_{P+j} + \bar{m}_{j,j+1}\partial_\tau^2 u_{P+j+1}) \\ \qquad\qquad\quad + c^2(\bar{s}_{j,j-1}u_{P+j-1} + \bar{s}_{j,j}u_{P+j} + \bar{s}_{j,j+1}u_{P+j+1}), \quad \text{for } j = 2,\dots,P-1, \text{on } \Gamma_{\text{abs}}, \end{cases}$$

where l, m are the coefficients defined in Section 3 and $\bar{l}$, $\bar{m}$ and $\bar{s}$ are given by:

$$\begin{cases} \bar{l}_{j,j-1} = \bar{m}_{j,j-1} = \dfrac{1}{\sigma_j}, \\ \bar{l}_{j,j} = \bar{m}_{j,j} = \dfrac{1}{\sigma_j} + \dfrac{1}{\sigma_{j+1}}, \\ \bar{l}_{j,j+1} = \bar{m}_{j,j+1} = \dfrac{1}{\sigma_{j+1}}, \end{cases} \qquad \text{and} \qquad \begin{cases} \bar{s}_{j,j-1} = \sigma_j, \\ \bar{s}_{j,j} = -(\sigma_j + \sigma_{j+1}), \\ \bar{s}_{j,j+1} = \sigma_{j+1}. \end{cases}$$

When $j = P$, we have to introduce a seam function ψ which makes the link between the two kinds of auxiliary functions: those defined for the propagative waves (using cos) and those for the evanescent ones (using σ). Hence, we get two equations for $j = P$ which are:

$$\begin{cases} (1-\cos^2 a_P)\partial_t^2 u_{P-1} + (\cos^2 a_P + 1)\partial_t^2 u_P + \cos^2 a_P \partial_t^2 \psi = c^2(\partial_\tau^2 u_{P-1} + \partial_\tau^2 u_P), \quad \text{on } \Gamma_{\text{abs}}, \\ \partial_t^2 u_P + \partial_t^2 u_{P+1} - \cos a_P \sigma_1 c \partial_t \psi = \sigma_1^2 c^2(u_P + u_{P+1}) + c^2(\partial_\tau^2 u_P + \partial_\tau^2 u_{P+1}), \quad \text{on } \Gamma_{\text{abs}}. \end{cases}$$

For the space-discretization, we use a similar method to the one described in Section 3 and we finally get:

$$\left(R_2 + \frac{\Delta t}{2}S_2\right)X^{n+1} = \Delta t^2 \begin{pmatrix} F(n\Delta t, .) \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \Delta t^2 T_2 X^n + 2R_2 X^n - R_2 X^{n-1} + \frac{\Delta t}{2}S_2 X^{n-1},$$

where X is the vector of all the unknowns: u, u_j and ψ.

§5. Numerical results

We have considered the square $[-2; 2] \times [-2; 2]$ and the following Ricker-type source:

$$f(x, y, t) = \delta(x - x_0, y - y_0)2\lambda(2\lambda(t - t_0)^2 - 1)e^{-\lambda(t-t_0)^2},$$

(x_r, y_r)	Higdon				H-W
	$a_1 = 0$	$a_1 = \frac{\pi}{6}$	$a_1 = a_2 = 0$	$a_1 = 0, a_2 = \frac{\pi}{6}$	$a_1 = a_2 = 0$ $\sigma_1 = 10$
$(0, -1.8)$	1.98	9.14	0.54	0.54	0.54
$(0.7, -1.8)$	7.3	4.06	0.61	0.60	0.60
$(1.8, 1.8)$	18.0	12.0	1.02	0.82	0.90

Table 1: Relative L^2 error for Higdon's and H-W conditions

where $\lambda = (5\pi)^2$, $t_0 = 0.2$, $(x_0, y_0) = (0, -1)$ and δ denotes the Dirac distribution. The penalization coefficient in the domain is $\alpha = 8$. We have computed the solution U^{app} near the absorbing boundary at three different points (x_r, y_r) equal to $(0, -1.8)$, $(0.7, -1.8)$ and $(1.3, -1.8)$ for different values of the coefficients a_j and σ_j and we have compared it to the exact solution U (i.e. the solution of the wave equation in $\mathbb{R}^2$). On Tab.1, we represent the relative $L^2([0, T])$ error, $err = \frac{\|U^{\text{app}} - U\|_{L^2([0,T])}}{\|U\|_{L^2([0,T])}} * 100$ for four Higdon's conditions ($a_1 = 0$, $a_1 = \pi/6$, $a_1 = 0$ and $a_2 = 0$,$a_1 = 0$ and $a_2 = \pi/6$) and one H-W condition ($a_1 = a_2 = 0$ and $\sigma = 10$).

For the first two tests, we have no auxiliary functions since we consider first-order conditions. For the three other tests we have imposed the same penalization coefficient $\alpha_j = 16$ for all the auxiliary equations on Γ. We remark that, as expected, the second-order Higdon's condition performs better than the first-order one. However, the third-order H-W condition does not improve the error as compared to the second-order Higdon's condition. This is due to the discretization method, since the accuracy of the ABC can be improved by decreasing the penalization coefficient but if this coefficient is too small the scheme becomes unstable. We have observed the same problem with thethird-order Higdon's condition and for higher-order conditions too. Moreover, for some particular coefficients, for instance $a_1 = 0$, $a_2 = \pi/6$ and $a_3 = \pi/4$, the scheme is unconditionnaly unstable (i.e. there is no penalization parameters that stabilize the scheme). In Fig. 1, we have represented the solution U^{app} for these coefficients at point $(0, -1.8)$.

Therefore, the method of auxiliary functions proposed in [3] to implement Higdon's and H-W conditions is not adapted to an IPDG approximation and we are now considering other type of ABC compatible with the IPDG method. In the same time, we are looking for an enriched IPDG scheme which is able to use the ABCs we consider in this work.

.

References

[1] Barucq, H., Delaurens, F., and Hanouzet, B. Method of absorbing boundary conditions: Phenomena of error stabilization. *SIAM J. Num. Anal. 35* (1998), 1113–1129.

[2] Engquist, B., and Majda, A. Absorbing boundary conditions for the numerical simulation of wave. *Math. Comp. 31* (1977), 629–651.

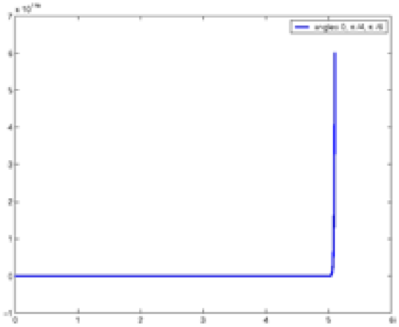


Figure 1: The solution U^{app} at point (0,-1.8) for $a_1 = 0$, $a_2 = \pi/6$ and $a_3 = \pi/4$

[3] Givoli, D., Hagstrom, T., and Patlashenko, I. Finite element formulation with high-order absorbing boundary conditions for time-dependent waves. *Comput. Meth. Appl. Mech. Engrg. 195* (2006), 3666–3690.

[4] Grote, M., Schneebeli, A., and Schötzau, D. Discontinuous Galerkin finite element method for the wave equation. *SIAM J. Num. Anal. 44* (2006), 2408–2431.

[5] Hagstrom, T., Mar-Or, A., and Givoli, D. High-order local absorbing conditions for the wave equation: extensions and improvements. *J. Comput. Phys. 227* (2008), 3322–3357.

[6] Hagstrom, T., and Warburton, T. A new auxiliary variable formulation of high-order local radiation boundary conditions: corner compatibility conditions and extensions to first-order systems. *Wave Motion 39* (2004), 327–338.

[7] Hagstrom, T., Warburton, T., and Givoli, D. Radiation boundary conditions for time-dependent waves based on complete plane wave expansions. *Submitted*.

[8] Higdon, R. Numerical absorbing boundary conditions for the wave equation. *Math. Comp. 49* (1987), 65–90.

INRIA Research Center Bordeaux Sud-Ouest, Team-project Magique-3D
Laboratoire de Mathématiques Appliquéee UMR CNRS 5142
Université de Pau et des Pays de l'Adour - Bt IPRA
BP 1155
64013 PAU Cedex
helene.barucq@inria.fr, julien.diaz@inria.fr, veronique.duprat@inria.fr

Monografías Matemáticas *García de Galdeano* **35**, 57–64 (2010)

NONCONFORMING FINITE ELEMENT DISCRETIZATION FOR THE NUMERICAL SIMULATION OF POLYMER FLOWS

Roland Becker, Daniela Capatina, Didier Graebling
and Julie Joie

Abstract. We present our first numerical results for the simulation of the Giesekus model, obtained by combining a $\boldsymbol{P}_1 \times P_0$ nonconforming finite element approximation for the velocity and the pressure with a $\underline{P}_0$ discontinuous Galerkin method for the stress tensor. For this purpose, we first consider a three-fields formulation of the Stokes equations and analyze its $\boldsymbol{P}_k \times P_{k-1} \times \underline{P}_{k-1}$ nonconforming approximation.

Keywords: Polymer liquid, Giesekus model, mixed formulation, nonconforming method.

AMS classification: 35Q30, 76A05, 65N30, 65N12.

§1. Introduction

Polymeric liquids are, from a rheological point of view, viscoelastic non- Newtonian fluids. Their non-Newtonian behavior can be observed in a variety of physical phenomena which are unseen with Newtonian liquids and which cannot be predicted by the Navier-Stokes equations. The rheological behavior of polymers is so complex that many different constitutive equations have been proposed in the literature in order to describe these phenomena (see for example [9]). We choose here to study the nonlinear differential model of Giesekus introduced in [5], which presents two main advantages. First, it yields a realistic behavior for shear flows, elongational flows and mixed flows. Second, only two material parameters are needed to describe the model: the viscosity η and the relaxation time λ.

In this paper, we employ for the discretization of the Giesekus model $\boldsymbol{P}_1 \times P_0$ nonconforming finite elements for the velocity and the pressure and $\underline{P}_0$ discontinuous elements for the stress tensor. We begin by considering a three-fields formulation of the Stokes problem and its $\boldsymbol{P}_k \times P_{k-1} \times \underline{P}_{k-1}$ nonconforming approximation for $k = 1, 2, 3$. A brief mathematical analysis of the discrete problem is presented before considering the nonlinear Giesekus model and its numerical approximation. Finally we present our first numerical results.

In what follows, we agree to write the vectors in bold letters and the tensors in underlined letters, $\underline{\tau} = (\tau_{ij})_{1\leq i,j\leq 2}$; the product of two tensors will be denoted by $\underline{\tau} : \underline{\sigma} = \sum_{i,j=1}^{2} \tau_{ij}\sigma_{ij}$. The letter c denotes a positive constant independent of the discretization.

§2. The Stokes problem

We first consider the Stokes equations, which describe the steady flow of an incompressible, Newtonian fluid at low Reynolds numbers. The governing equations are the momentum and the mass conservation laws and the constitutive equation of a Newtonian fluid. In order to compute non-Newtonian liquids (which is our further goal), one has to consider formulations with at least three unknows, since the stress tensor cannot be eliminated from the corresponding constitutive equation. Therefore, we write the Stokes equations as follows:

$$\begin{aligned} -\operatorname{div}\underline{\tau} + \nabla p &= \boldsymbol{f} && \text{in } \Omega, && (1)\\ \nabla\cdot\boldsymbol{u} &= 0 && \text{in } \Omega, && (2)\\ \underline{\tau} &= 2\eta\underline{D}(\boldsymbol{v}) && \text{in } \Omega, && (3)\end{aligned}$$

where $\underline{\tau}$ denotes the viscous stress tensor, p the pressure, η the fluid's viscosity and $\underline{D}(\boldsymbol{u}) = \frac{1}{2}\left(\underline{\nabla}\boldsymbol{u} + (\underline{\nabla}\boldsymbol{u})^t\right)$ the rate of strain tensor, with $\boldsymbol{u}$ the velocity. In view of the numerical approximation, we take Ω a polygonal domain of $\mathbb{R}^2$. We take the data $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$. For the sake of simplicity, we only consider here homogeneous Dirichlet boundary conditions.

2.1. Two equivalent variational formulations

Let $(\mathcal{T}_h)_{h>0}$ be a family of triangulations of Ω consisting of triangles: $\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h} T$. Then the three-fields formulation of the Stokes problem can be written as follows:

$$\begin{cases} (\boldsymbol{u}, p, \underline{\tau}) \in \boldsymbol{H}_0^1(\Omega)\times L_0^2(\Omega)\times \underline{X} & & & \\ \qquad\qquad b(p,\boldsymbol{v}) + c_0(\underline{\tau},\boldsymbol{v}) & = l(\boldsymbol{v}) & \forall \boldsymbol{v}\in \boldsymbol{H}_0^1(\Omega) \\ b(q,\boldsymbol{u}) & = 0 & \forall q \in L_0^2(\Omega) \\ c_0(\underline{\sigma},\boldsymbol{u}) \qquad - d_0(\underline{\sigma},\underline{\tau}) & = 0 & \forall \underline{\sigma}\in\underline{X}, \end{cases} \qquad (4)$$

where

$$c_0(\underline{\tau},\boldsymbol{v}) = \sum_{T\in\mathcal{T}_h}\int_T \underline{\tau} : \underline{D}(\boldsymbol{v})dx, \qquad d_0(\underline{\sigma},\underline{\tau}) = \frac{1}{2\eta}\int_\Omega \underline{\sigma} : \underline{\tau}dx,$$

$$b(p,\boldsymbol{v}) = -\sum_{T\in\mathcal{T}_h}\int_T p\nabla\cdot\boldsymbol{v}dx, \qquad l(\boldsymbol{v}) = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}dx,$$

and where

$$\underline{X} = \left\{\underline{\tau} = (\tau_{ij})_{1\le i,j\le 2}; \tau_{ij} = \tau_{ji},\ \tau_{ij}\in L_2(\Omega),\ i,j = 1,2\right\}.$$

The symmetry of the stress tensor is strongly imposed in the definition of the space $\underline{X}$. Note that the last equation of (4) implies $\underline{\tau} = 2\eta\underline{D}(\boldsymbol{u})$ and so, by substituting $\underline{\tau}$ in the first equation, one deduces the equivalence between (4) and the following two fields formulation:

$$\begin{cases} (\boldsymbol{u}, p) \in \boldsymbol{H}_0^1(\Omega)\times L_0^2(\Omega) & & \\ a_0(\boldsymbol{u},\boldsymbol{v}) + b(p,\boldsymbol{v}) & = l(\boldsymbol{v}) & \forall \boldsymbol{v}\in\boldsymbol{H}_0^1(\Omega) \\ b(q,\boldsymbol{u}) & = 0 & \forall q\in L_0^2(\Omega), \end{cases} \qquad (5)$$

where

$$a_0(\boldsymbol{u},\boldsymbol{v}) = 2\eta\sum_{T\in\mathcal{T}_h}\int_T \underline{D}(\boldsymbol{u}) : \underline{D}(\boldsymbol{v})dx.$$

The proof of the well-posedness of this formulation is well known, see for instance [6].

2.2. Discretization by means of nonconforming finite elements

We agree to denote by ε_h^{int} the set of internal edges of $\mathcal{T}_h$, by ε_h^{∂} the set of edges situated on the boundary $\partial\Omega$ and we put $\varepsilon_h = \varepsilon_h^{int} \cup \varepsilon_h^{\partial}$. As usually, let h_T be the diameter of the triangle T and let $h = \max_{T\in\mathcal{T}_h} h_T$. On every edge e belonging to ε_h^{int}, such that $\{e\} = \partial T^i \cap \partial T^j$, we define once for all the unit normal $\boldsymbol{n}_e$ oriented from T^i towards T^j. Then, for a given function φ, we define the jump across the edge e by $[\varphi] = \varphi_{/T^i} - \varphi_{/T^j}$. If $e \in \varepsilon_h^{\partial}$, we take for $\boldsymbol{n}_e$ the outward unit normal $\boldsymbol{n}$ and for $[\varphi]$ the trace of φ on e. We agree to denote the $L^2(e)$-orthogonal projection of a given function $\varphi \in L^2(e)$ on the polynomial space P_k ($k \in \mathbb{N}$) by $\pi_k v$.

In what follows, we take $k = 1$, 2 or 3 and we introduce the following discrete spaces:

$$\begin{aligned}
\boldsymbol{V}_h &= \left\{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega);\ (\boldsymbol{v})_{/T} \in \boldsymbol{P}_k,\ \forall T \in \mathcal{T}_h \text{ and } [\pi_{k-1}\boldsymbol{v}]_{/e} = 0,\ \forall e \in \varepsilon_h\right\},\\
Q_h &= \left\{q \in L_0^2(\Omega);\ (q)_{/T} \in P_{k-1},\ \forall T \in \mathcal{T}_h\right\},\\
\underline{X}_h &= \left\{\underline{\sigma} \in \underline{X};\ (\underline{\sigma})_{/T} \in \underline{P}_{k-1},\ \forall T \in \mathcal{T}_h\right\}.
\end{aligned}$$

For odd k, we recognize the Crouzeix-Raviart finite elements introduced in [3] (see also [2]); $k = 2$ corresponds to the elements of Fortin-Soulie introduced in [4]. It is known that there exist two interpolation operators $\boldsymbol{I}_h \in \mathcal{L}(\boldsymbol{H}^1(\Omega); \boldsymbol{V}_h)$ and $i_h \in \mathcal{L}(L_0^2; Q_h)$ satisfying usual interpolation estimates and moreover,

$$\begin{aligned}
&\int_T r\nabla\cdot(\boldsymbol{I}_h\boldsymbol{v} - \boldsymbol{v})dx = 0, \qquad \int_e \boldsymbol{r}\cdot[\boldsymbol{I}_h\boldsymbol{v}]ds = 0, \qquad \forall r \in P_{k-1},\ \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega),\\
&\int_T r(i_h q - q)dx = 0, \qquad \forall r \in P_{k-1},\ \forall q \in L_0^2(\Omega).
\end{aligned}$$

We can now write the nonconforming approximation of (5) as follows:

$$\begin{cases}
(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h & & \\
a(\boldsymbol{u}_h, \boldsymbol{v}_h) + b(p_h, \boldsymbol{v}_h) & = l(\boldsymbol{v}_h) & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h\\
b(q_h, \boldsymbol{u}_h) & = 0 & \forall q_h \in Q_h
\end{cases} \tag{6}$$

where now

$$a(\cdot,\cdot) = a_0(\cdot,\cdot) + \gamma a_1(\cdot,\cdot), \qquad a_1(\boldsymbol{u}_h, \boldsymbol{v}_h) = \eta \sum_{e\in\varepsilon_h^{\partial\mathrm{int}}} \frac{1}{|e|}\int_e [\pi_1(\boldsymbol{u}_h\cdot\boldsymbol{n}_e)][\pi_1(\boldsymbol{v}_h\cdot\boldsymbol{n}_e)]ds,$$

with γ a stabilization parameter which can be chosen independent of h. The stabilization term $a_1(\cdot,\cdot)$ is added only for $k = 1$ in order to retrieve the coercivity of the bilinear form $a(\cdot,\cdot)$, thanks to a discrete Korn inequality. Note that a different choice for $a_1(\cdot,\cdot)$ was introduced by Brenner in [1], enhancing the continuity of the whole vector $\pi_1\boldsymbol{u}_h$ across the internal edges. For $k = 2$ or 3, one obviously has $a_1(\boldsymbol{u}_h, \boldsymbol{v}_h) = 0$, for all $\boldsymbol{u}_h, \boldsymbol{v}_h \in \boldsymbol{V}_h$.

We now consider the following nonconforming approximation of (4):

$$\begin{cases}
(\boldsymbol{U}_h, P_h, \underline{\tau}_h) \in \boldsymbol{V}_h \times Q_h \times \underline{X}_h & & & \\
\gamma a_1(\boldsymbol{U}_h, \boldsymbol{v}_h) + b(P_h, \boldsymbol{v}_h) + c_0(\underline{\tau}_h, \boldsymbol{v}_h) & & = l(\boldsymbol{v}_h) & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h\\
b(q_h, \boldsymbol{U}_h) & & = 0 & \forall q_h \in Q_h\\
c_0(\underline{\sigma}_h, \boldsymbol{U}_h) & - d_0(\underline{\sigma}_h, \underline{\tau}_h) & = 0 & \forall \underline{\sigma}_h \in \underline{X}_h.
\end{cases} \tag{7}$$

2.3. Well-posedness of the approximated problems and error estimates

In order to prove the well-posedness of (6), we apply the Babǔska-Brezzi theorem. It is useful to introduce the following semi- norms on $\boldsymbol{H}^1(\Omega) + \boldsymbol{V}_h$:

$$|\boldsymbol{v}|^2_{1,h} = \sum_{T\in\mathcal{T}_h} |\boldsymbol{v}|^2_{1,T}, \qquad \|\underline{D}(\boldsymbol{v})\|^2_{0,h} = \sum_{T\in\mathcal{T}_h} \|\underline{D}(\boldsymbol{v})\|^2_{0,T},$$

$$[\![\boldsymbol{v}]\!] = \left(2\eta\|\underline{D}(\boldsymbol{v})\|^2_{0,h} + \gamma a_1(\boldsymbol{v},\boldsymbol{v})\right)^{1/2} = a(\boldsymbol{v},\boldsymbol{v})^{1/2},$$

and to recall the following result for piecewise H^1 functions, established by Brenner [1] in a stronger form and then improved by Mardal and Winther [8]:

$$|\boldsymbol{v}|_{1,h} \le c\left(\|\underline{D}(\boldsymbol{v})\|^2_{0,h} + \frac{1}{\eta}a_1(\boldsymbol{v},\boldsymbol{v}) + \phi(\boldsymbol{v})\right)^{1/2},$$

where $\phi : \boldsymbol{H}^1(\Omega) \to \mathbb{R}$ is a continuous semi-norm such that if $\phi(\boldsymbol{v}) = 0$ for a rigid motion $\boldsymbol{v}$, then $\boldsymbol{v}$ is a constant vector. With the choice

$$\phi(\boldsymbol{v}) = \sum_{e\in\varepsilon_h^\partial} \|\boldsymbol{\pi}_0\boldsymbol{v}\|^2_{0,e},$$

we can now deduce the following Korn inequality on $\boldsymbol{V}_h$:

$$|\boldsymbol{v}|_{1,h} \le c\left(\|\underline{D}(\boldsymbol{v})\|^2_{0,h} + \frac{1}{\eta}a_1(\boldsymbol{v},\boldsymbol{v})\right)^{1/2}. \tag{8}$$

Theorem 1. *Problem (6) has a unique solution.*

Proof. We check the hypotheses of the Babǔska-Brezzi theorem. The coercivity of the form $a(\cdot,\cdot)$ on $\boldsymbol{V}_h \times \boldsymbol{V}_h$ is obvious in view of (8), so one has only to prove the discrete inf-sup condition. For this purpose, we make use of the continuous inf-sup condition for the Stokes problem (cf [6]), and with any $q \in Q_h \subset L^2_0(\Omega)$ we associate $z \in \boldsymbol{H}^1_0(\Omega)$ such that $\nabla\cdot z = q$ and $\|z\|_{1,\Omega} \le c\|q\|_{0,\Omega}$. By putting $\boldsymbol{w} = \boldsymbol{I}_h z \in \boldsymbol{V}_h$, we immediately obtain $b(q,\boldsymbol{w}) = \|q\|^2_{0,\Omega}$ and $|\boldsymbol{w}|_{1,h} \le c\|q\|_{0,\Omega}$. For $k = 2$ or 3, one has $a_1(\boldsymbol{w},\boldsymbol{w}) = 0$ so the result is obvious. For $k = 1$, we still have to bound

$$a_1(\boldsymbol{w},\boldsymbol{w}) = \eta\sum_{e\in\varepsilon_h}\frac{1}{|e|}\|[\boldsymbol{w}\cdot\boldsymbol{n}_e]\|^2_{0,e}.$$

By combining the fact that $[z\cdot\boldsymbol{n}_e] = 0$, the next trace inequality on $\{e\} = \partial T_1 \cap \partial T_2$:

$$\begin{aligned}
&\frac{1}{\sqrt{|e|}}\|[(\boldsymbol{I}_h z - z)\cdot\boldsymbol{n}_e]\|_{0,e}\\
&\qquad \le c\left(\frac{1}{h_{T_1}}\|\boldsymbol{I}_h z - z\|_{0,T_1} + \frac{1}{h_{T_2}}\|\boldsymbol{I}_h z - z\|_{0,T_2} + |\boldsymbol{I}_h z - z|_{1,T_1} + |\boldsymbol{I}_h z - z|_{1,T_2}\right)\\
&\qquad \le c|z|_{1,T_1\cup T_2}
\end{aligned}$$

and the interpolation properties of the Crouzeix-Raviart operator $\boldsymbol{I}_h$, we get

$$a(\boldsymbol{w},\boldsymbol{w}) \le c\eta\,|z|^2_{1,h},$$

which allows us to conclude. □

The well-posedeness of (7) is now immediate, thanks to its equivalence with (6).

Theorem 2. *Problem (7) has a unique solution given by $(\boldsymbol{u}_h, p_h, 2\eta\underline{D}(\boldsymbol{u}_h))$, where $(\boldsymbol{u}_h, p_h)$ is the solution of (6).*

We have established, by using standard techniques in stabilized mixed formulations:

Theorem 3. *Let $(\boldsymbol{u}, p) \in \boldsymbol{H}^{k+1}(\Omega) \times H^k(\Omega)$ be the solution of the continuous Stokes problem. Then the solution $(\boldsymbol{u}_h, p_h)$ of (6) satisfies the following a priori error bound*

$$[\![\boldsymbol{u} - \boldsymbol{u}_h]\!] + \frac{1}{\sqrt{\eta}}\|p - p_h\|_{0,\Omega} \le ch^k\Big(\sqrt{\eta}|\boldsymbol{u}|_{k+1,\Omega} + \frac{1}{\sqrt{\eta}}|p|_{k,\Omega}\Big),$$

with c a constant independent of h and of η. If, moreover, Ω is convex, then

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega} \le ch^{k+1}\Big(|\boldsymbol{u}|_{k+1,\Omega} + \frac{1}{\eta}|p|_{k,\Omega}\Big).$$

§3. The Giesekus model

Giesekus introduced in [5] the following constitutive law, describing the behavior of a viscoelastic non-Newtonian liquid:

$$\lambda\overset{\triangledown}{\underline{\tau}} + \frac{\alpha}{G}\underline{\tau}\,\underline{\tau} + \underline{\tau} = 2\eta\underline{D}(\boldsymbol{u}), \tag{9}$$

whith G the elastic modulus and η the viscosity. The relaxation time λ is defined by the formula $\lambda = \eta/G$ and $\alpha \in [0; 1[$ is a parameter. An appropriate choice seems to be $\alpha = 1/2$. Note that $\alpha = 0$ yields the upper convected Maxwell model. Here above, $\overset{\triangledown}{\underline{\tau}}$ is the upper convective derivative defined by

$$\overset{\triangledown}{\underline{\tau}} = \frac{\partial}{\partial t}\underline{\tau} + (\boldsymbol{u}\cdot\nabla)\underline{\tau} - \Big(\underline{\tau}\,(\nabla\boldsymbol{u})^t + \nabla\boldsymbol{u}\,\underline{\tau}\Big).$$

We note that this constitutive law is strongly nonlinear since it involves a quadratic term in the stress tensor, which is difficult to handle and certainly explains the lack of mathematical and numerical studies of this model. The complete Giesekus model is obtained by adding the two conservation equations (1) and (2), as well as boundary conditions $\boldsymbol{u} = \boldsymbol{g}$ on $\partial\Omega$ and $\underline{\tau} = \underline{\tau}^-$ on the inflow boundary $\partial\Omega^-$, and for instationary flows, initial conditions on $\boldsymbol{u}$ and $\underline{\tau}$.

We are now interested in the discretization of the above Giesekus model. For this purpose we consider the previous spaces $\boldsymbol{V}_h$, Q_h and $\underline{X}_h$ for $k = 1$ and we write the discrete problem as follows:

$$\begin{cases} (\boldsymbol{u}_h^*,\ p_h^*, \underline{\tau}_h^*) \in \boldsymbol{V}_h \times Q_h \times \underline{X}_h & & & \\ \gamma a_1(\boldsymbol{u}_h^*, \boldsymbol{v}_h) + \quad b(p_h^*, \boldsymbol{v}_h) + \quad c_0(\boldsymbol{v}_h, \underline{\tau}_h^*) & = l(\boldsymbol{v}_h) & \forall \boldsymbol{v_h} \in \boldsymbol{V}_h \\ b(q_h, \boldsymbol{u_h^*}) & = 0 & \forall q_h \in Q_h \\ c(\boldsymbol{u}_h^*, \underline{\tau}_h^*; \underline{\sigma}_h) + \qquad d(\underline{\tau}_h^*, \underline{\tau}_h^*; \underline{\sigma}_h) & = 0 & \forall \underline{\sigma}_h \in \underline{X}_h. \end{cases}$$

The nonlinear forms $c(\cdot,\cdot;\cdot)$ and $d(\cdot,\cdot;\cdot)$ are defined by

$$\begin{aligned} c(\cdot,\cdot;\cdot) &= c_0(\cdot,\cdot) + c_1(\cdot,\cdot;\cdot) + c_2(\cdot,\cdot;\cdot), \\ d(\cdot,\cdot;\cdot) &= d_0(\cdot,\cdot) + d_1(\cdot,\cdot;\cdot). \end{aligned}$$

$c_1(\cdot,\cdot;\cdot)$ is the convective term that we treat by means of the Lesaint-Raviart upwind scheme (cf [7]). More precisely, we take

$$c_1(\boldsymbol{u}_h, \underline{\tau}_h; \underline{\sigma}_h) = \sum_{T\in\mathcal{T}_h} \int_{\partial T^-} \boldsymbol{u}_h \cdot \boldsymbol{n}\, \underline{\tau}_h^{\mathrm{ext}} : (\underline{\sigma}_h^{\mathrm{int}} - \underline{\sigma}_h^{\mathrm{ext}}) ds,$$

where $\partial T^- = \{e \subset \partial T; \pi_0(\boldsymbol{u}_h \cdot \boldsymbol{n}) < 0 \text{ on } e\}$. $c_2(\cdot,\cdot;\cdot)$ comes from the objective derivative

$$c_2(\boldsymbol{u}_h, \underline{\tau}_h; \underline{\sigma}_h) = -\lambda \sum_{T\in\mathcal{T}_h} \int_T \underline{\tau}_h (\underline{\nabla \boldsymbol{u}}_h)^t : \underline{\sigma}_h dx - \lambda \sum_{T\in\mathcal{T}_h} \int_T \underline{\nabla \boldsymbol{u}}_h\, \underline{\tau}_h : \underline{\sigma}_h dx,$$

whereas $d_1(\cdot,\cdot;\cdot)$ takes into account the quadratic term of (9) as follows:

$$d_1(\underline{\tau}_h, \underline{\tau}_h; \underline{\sigma}_h) = \sum_{T\in\mathcal{T}_h} \frac{1}{G} \int_T \underline{\tau}_h\, \underline{\tau}_h : \underline{\sigma}_h dx.$$

For the moment, the nonlinear problem is solved by means of Newton's method, which necessitates the computation of the jacobian matrix

$$J = \begin{pmatrix} \gamma A_1 & B & C_0 \\ B^T & 0 & 0 \\ C_0^T + C_{11} + C_{21} & 0 & D_0 + C_{12} + C_{22} + \alpha D_{11} \end{pmatrix}.$$

By denoting τ^i, respectively $\boldsymbol{v}^i$, the values of the stress tensor, respectively the velocity, at the previous Newton iteration, one has:

$$c_{11}(\boldsymbol{u}_h, \underline{\sigma}_h) = \sum_{T\in\mathcal{T}_h} \int_{\partial T^-} \boldsymbol{u}_h \cdot \boldsymbol{n}\, \underline{\tau}^{i\mathrm{ext}} : (\underline{\sigma}_h^{\mathrm{int}} - \underline{\sigma}_h^{\mathrm{ext}}) ds,$$

$$c_{12}(\underline{\tau}_h, \underline{\sigma}_h) = \sum_{T\in\mathcal{T}_h} \int_{\partial T^-} \boldsymbol{u}^i \cdot \boldsymbol{n}\, \underline{\tau}_h^{\mathrm{ext}} : (\underline{\sigma}_h^{\mathrm{int}} - \underline{\sigma}_h^{\mathrm{ext}}) ds,$$

$$c_{21}(\boldsymbol{u}_h, \underline{\sigma}_h) = -\lambda \sum_{T\in\mathcal{T}_h} \int_T \underline{\tau}^i (\underline{\nabla \boldsymbol{u}}_h)^t : \underline{\sigma}_h dx - \lambda \sum_{T\in\mathcal{T}_h} \int_T \underline{\nabla \boldsymbol{u}}_h\, \underline{\tau}^i : \underline{\sigma}_h dx,$$

$$c_{22}(\underline{\tau}_h, \underline{\sigma}_h) = -\lambda \sum_{T\in\mathcal{T}_h} \int_T \underline{\tau}_h (\underline{\nabla \boldsymbol{u}}^i)^t : \underline{\sigma}_h dx - \lambda \sum_{T\in\mathcal{T}_h} \int_T \underline{\nabla \boldsymbol{u}}^i\, \underline{\tau}_h : \underline{\sigma}_h dx,$$

$$d_{11}(\underline{\tau}_h, \underline{\sigma}_h) = \sum_{T\in\mathcal{T}_h} \frac{1}{G} \int_T \left(\underline{\tau}_h\, \underline{\tau}^i : \underline{\sigma}_h + \underline{\tau}^i\, \underline{\tau}_h : \underline{\sigma}_h \right) dx.$$

The problem has a rather large number of unknowns (six in 2D and ten in 3D which is the realistic framework for polymer flows). Therefore, it is important to reduce the computational cost by improving the Newton method. One may note that for $k = 1$, $c_{21}(\cdot,\cdot)$, $c_{22}(\cdot,\cdot)$ and $d_{11}(\cdot,\cdot)$ are defined locally on each triangle while the stencils of $c_{11}(\cdot,\cdot)$ and $c_{12}(\cdot,\cdot)$ are reduced to the element itself and its neighbours. The development of a specially designed Newton algorithm, allowing for the use of the Stokes matrix and taking advantage of the small stencils of the previous matrices, is undergoing work.

In perspective, we also intend to study higher order approximations. The design of a monotone DG method of degree ≥ 1 for the transport equation is an active research domain.

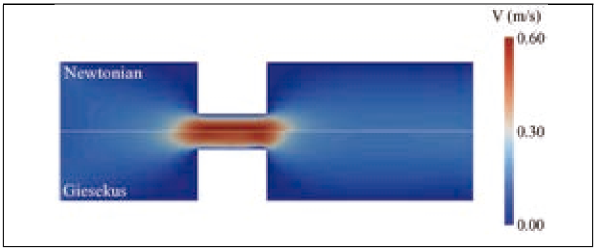


Figure 1: Velocity magnitude for a Newtonian and a Giesekus liquid.

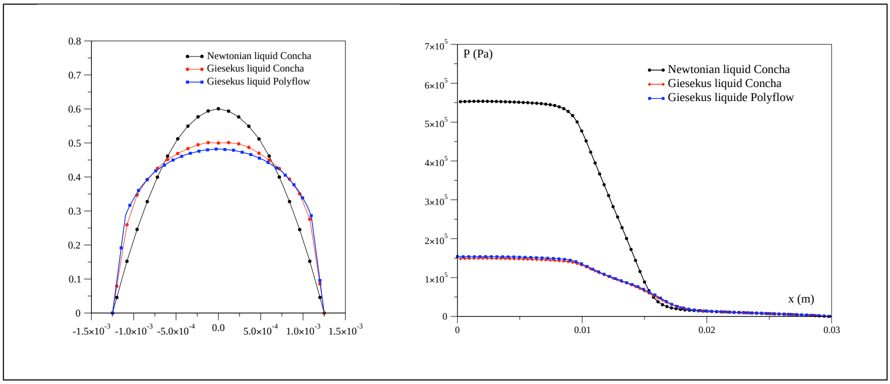


Figure 2: Velocities and pressures with Concha and Polyflow.

§4. Numerical results

There exist only relatively few numerical codes for the simulation of polymer flows. A major issue is the treatment of the internal coupling between the viscoelasticity of the liquid and the flow. This coupling is quantified by the Weissenberg number We, defined by $W_e = \lambda\dot{\gamma}$ where $\dot{\gamma}$ is the shear rate. The CFD codes for polymer liquids are generally only able to deal with W_e up to 10. The most popular code for the simulation of polymer flows is Polyflow (*http://www.ansys.com/products/polyflow*) developped by the Cesame team. We have computed our method in the C++ library Concha developped by the INRIA team Concha.

For all considered tests, we chose $\eta = 100$ Pa.s. For each value of λ, we compute the largest Weissenberg number for a corresponding Newtonian fluid as follows:

$$\dot{\gamma}_{max} = 6\bar{\boldsymbol{u}}/l \;\Rightarrow\; W_{e_{max}} = \lambda\, 6\bar{\boldsymbol{u}}/l,$$

where $\bar{\boldsymbol{u}}$ is the mean velocity on the channel and l its thickness. The geometry studied is a 4:1 planar contraction. We show the results for $W_e = 7.68$, and we consider a stationary flow.

In Fig. 1, we compare the norm of the velocity for a Newtonian and a Giesekus fluid. In Fig. 2, we compare our results with the ones obtained with Polyflow. We have used a mesh consisting of 25 794 triangles with Concha, respectively 14 866 with Polyflow. For the inflow and for the outflow, we impose 0.1 m/s as a normal velocity and a Neumann condition, respectively. With Concha, these results are obtained after 1500 s and with Polyflow after 3110 s. We observe a good agreement between the two approaches. The modification of the velocity profiles and the shut down of the pressure are typical behaviors of non- Newtonian fluids.

References

[1] Brenner, S. C. Korn's inequalities for piecewise H^1 vector fields. *Math. Comp. 73*, 247 (2004), 1067–1087.

[2] Crouzeix, M., and Falk, R. S. Nonconforming finite elements for the Stokes problem. *Math. Comp. 52*, 186 (1989), 437–456.

[3] Crouzeix, M., and Raviart, P.-A. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge 7*, R-3 (1973), 33–75.

[4] Fortin, M., and Soulie, M. A nonconforming piecewise quadratic finite element on triangles. *Internat. J. Numer. Methods Engrg. 19*, 4 (1983), 505–520.

[5] Giesekus, H. Constitutive equations for polymer fluids based on the concept of configuration-dependent molecular mobility : A generalized mean configuration model. *J. Non-Newtonian Fluids Mech. 17* (1985), 349.

[6] Girault, V., and Raviart, P.-A. *Finite element methods for Navier-Stokes equations*, vol. 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986.

[7] Lesaint, P., and Raviart, P.-A. On a finite element method for solving the neutron transport equation. In *Mathematical aspects of finite elements in partial differential equations*. 1974, pp. 89–123.

[8] Mardal, K.-A., and Winther, R. An observation on Korn's inequality for nonconforming finite element methods. *Math. Comp. 75*, 253 (2006), 1–6.

[9] Owens, R. G., and Phillips, T. N. *Computational rheology*. Imperial College Press, London, 2002.

Roland Becker, Daniela Capatina and Julie Joie
INRIA Bordeaux-Sud-Ouest - Equipe Projet Concha
Laboratoire de Mathématiques et Applications, UMR CNRS 5142
Université de Pau et des Pays de l'Adour
IPRA, BP 1155, 64013 Pau Cedex (France)
roland.becker@univ-pau.fr , daniela.capatina@univ-pau.fr and
julie.joie@univ-pau.fr

Didier Graebling
INRIA Bordeaux-Sud-Ouest - Equipe Projet Concha
IPREM/EPCP, UMR CNRS 5154
Université de Pau et des Pays de l'Adour
2 avenue Angot, 64053 Pau Cedex 9(France)
didier.graebling@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **35**, 65–71 (2010)

NUMERICAL SIMULATION OF LIQUID CRYSTALS

Roland Becker and Nour El Houda Seloula

Abstract. We consider the numerical simulation of nematic liquid crystal flows, modeled by a simplified version of [2] the Ericksen-Leslie model, imposing a nonconvex constraint on the director field. Computational experiments are used to compare the two approaches.

Keywords: Nematic liquid crystals, Ericksen-Lislie model, harmonic map heat flow, finite element method, discrete scheme.

AMS classification: 65M12, 65M60, 35K55, 35Q35.

§1. Introduction

In this paper, we consider a simplified version of the Ericksen- Leslie model, see for example Lin and Liu [2]. This model is a modified Navier- Stokes system that takes into account the liquid crystallinity, coupled with the Ginzburg-Landau equations.

$$
\begin{aligned}
\mathbf{v}_t - \nu\,\Delta\mathbf{v} + (\mathbf{v}.\nabla)\mathbf{v} + \nabla p + \lambda\,\nabla\cdot(\nabla\mathbf{d}\odot\nabla\mathbf{d}) &= 0 && \text{in } \Omega_T := (0,T)\times\Omega, && (1)\\
\mathbf{d}_t + (\mathbf{v}.\nabla)\mathbf{d} - \gamma\Delta\mathbf{d} &= \gamma|\Delta\mathbf{d}|^2\mathbf{d} && \text{in } \Omega_T, && (2)\\
\nabla\cdot\mathbf{v} &= 0 && \text{in } \Omega_T, && (3)
\end{aligned}
$$

and the nonconvex constraint

$$|\mathbf{d}(t,x)| = 1, \qquad (4)$$

and with the initial and boundary conditions

$$
\begin{aligned}
&\mathbf{v}(0,\mathbf{x}) = \mathbf{v}_0(\mathbf{x}), && \mathbf{d}(0,\mathbf{x}) = \mathbf{d}_0(\mathbf{x}), && \forall\mathbf{x}\in\Omega. && (5)\\
&\mathbf{v}(t,\mathbf{x}) = 0, && \partial_{\mathbf{n}}\mathbf{d}(t,\mathbf{x}) = 0, && \forall(t,\mathbf{x})\in\partial\Omega_T. && (6)
\end{aligned}
$$

The unknowns are the time-dependent divergence-free velocity field $\mathbf{v}(t,\mathbf{x})$, the pressure $p(t,\mathbf{x})$ of the fluid and the director field $\mathbf{d}(t,\mathbf{x})$ representing the orientation of the liquid crystal molecules. The fluid is confined in an open bounded domain Ω of $\mathbb{R}^3$ with a lipschitz boundary $\partial\Omega$. In the above, the vector $\mathbf{n}$ denotes the outward pointing unit normal and the matrix product is defined as

$$(\nabla\mathbf{d}\odot\nabla\mathbf{d})_{ij} = \sum_{k=1}^{2}\frac{\partial\mathbf{d}_k}{\partial\mathbf{x}_i}\frac{\partial\mathbf{d}_k}{\partial\mathbf{x}_j}.$$

The constraint (4) causes difficulties from both analytical and numerical points of view. A widely used approach is to approximate this constraint by a penalty function such as the

Ginzburg-Landau approximation $\mathbf{f}_\epsilon(\mathbf{d}) = \epsilon^{-2}(|\mathbf{d}|^2 - 1)\mathbf{d}$, for $0 < \epsilon \ll 1$. This penalisation function exhibits a potential structure, i.e., there exists a potential function $\mathbf{F}_\epsilon(\mathbf{d}) = \frac{\epsilon^{-2}}{4}(|\mathbf{d}|^2 - 1)^2$ such that $\mathbf{f}_\epsilon(\mathbf{d}) = \nabla_{\mathbf{d}}(\mathbf{F}_\epsilon(\mathbf{d}))$.
Accordingly, the penalised model reads as

$$\mathbf{v}_t - \nu\,\Delta\mathbf{v} + (\mathbf{v}.\nabla)\mathbf{v} + \nabla p + \lambda\,\nabla\cdot\,(\nabla\mathbf{d}\odot\nabla\mathbf{d}) = 0 \qquad \text{in } \Omega_T := (0,T)\times\Omega, \tag{7}$$

$$\mathbf{d}_t + (\mathbf{v}.\nabla)\mathbf{d} - \gamma(\Delta\mathbf{d} - \mathbf{f}_\epsilon(\mathbf{d})) = 0 \qquad \text{in } \Omega_T, \tag{8}$$

$$\nabla\cdot\mathbf{v} = 0 \qquad \text{in } \Omega_T, \tag{9}$$

subject to the initial and the boundary conditions (5) and (6).

Two fully discrete finite element methods for the system (1)-(3) and (7)-(9) have been recently studied by R. Becker, X. Feng, and A. Prohl [1], where the convergence of finite element approximations is established but the schemes do not satisfy the constraint (4). In this note, we are interested in a modification satisfying this contraint.

The paper is organized as follows. In the next section, we recall the energy estimates proven by Lin and Liu [2]. In section 3, we develop our modified scheme and in section 4, we prove that this scheme satisfies the constraint (4). Computational examples are given to prove the efficiency of the method.

§2. Energy estimates

It was observed in [2] that by using the differential identity $\nabla\cdot(\nabla\mathbf{z}\odot\nabla\mathbf{z}) = (\nabla\mathbf{z})^T\Delta\mathbf{z} + \frac{1}{2}\nabla(|\nabla\mathbf{z}|^2)$, the equation (7) can be rewritten as follows:

$$\mathbf{v}_t - \nu\,\Delta\mathbf{v} + (\mathbf{v}.\nabla)\mathbf{v} + \nabla p + \frac{\lambda}{2}\,\nabla(|\nabla\mathbf{d}|^2) + \lambda(\nabla\mathbf{d})^T\Delta\mathbf{d} = 0. \tag{10}$$

Notice that the term $\frac{\lambda}{2}\,\nabla(|\nabla\mathbf{d}|^2)$ can be absorbed into the definition of the pressure. Hence the system (7)–(9) satisfies the following basic energy law:

$$\frac{dE}{dt} = -\left(\nu\|\nabla\mathbf{v}\|^2_{L^2(\Omega)} + \lambda\gamma\|\Delta\mathbf{d} - \mathbf{f}_\epsilon(\mathbf{d})\|^2_{L^2(\Omega)}\right), \tag{11}$$

where

$$E = \frac{1}{2}\|\mathbf{v}\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|\nabla\mathbf{d}\|^2_{L^2(\Omega)} + \lambda\int_\Omega \mathbf{F}_\epsilon(\mathbf{d}).$$

This estimate was used by Lin and Liu [2] to establish existence, uniqueness and regularity of solutions to the coupled liquid crystal problem. The energy law (11) is obtained by multiplying the equation (10) by $\mathbf{v}$ and the director equation (8) by $-(\Delta\mathbf{d} - \mathbf{f}_\epsilon(\mathbf{d})$ and adding the two. The crucial observation is that the main term from the momentum equation $\nabla\mathbf{d}^T(\Delta\mathbf{d})\cdot\mathbf{v}$, cancels with the convective term $(\mathbf{v}\cdot\nabla)\mathbf{d}\cdot(-\Delta\mathbf{d}) = -\nabla\mathbf{d}^T(\Delta\mathbf{d})\mathbf{v}$ in the director equation. We have also, by using the facts $\operatorname{div}\mathbf{v} = 0$ and $\mathbf{v} = 0$ on $\partial\Omega$, that

$$\int_\Omega(\mathbf{v}\cdot\nabla)\mathbf{v}\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} = \int_\Omega \mathbf{v}\cdot\nabla p\,\mathrm{d}\mathbf{x} = \int_\Omega(\mathbf{v}\cdot\nabla)\mathbf{d}\cdot\mathbf{f}_\epsilon(\mathbf{d})\,\mathrm{d}\mathbf{x} = \int_\Omega \mathbf{v}\cdot\nabla\Big(\frac{|\mathbf{d}|^2}{2}\Big)\,\mathrm{d}\mathbf{x} = 0.$$

§3. Fully discrete finite element methods for the Ericksen-Leslie model

We assume that $\mathcal{T}_h$ is a quasi-uniform triangulation of a bounded polygonal domain $\Omega \subset \mathbb{R}^2$ into triangles of diameter $h > 0$, i.e., $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} \overline{K}$. Let $\mathcal{N}_h$ denote the set of all nodes of $\mathcal{T}_h$. We introduce the space

$$\mathbf{Y}_h = \left\{\mathbf{a}_h \in C(\overline{\Omega}, \mathbb{R}^2) : \mathbf{a}_h|_K \in P_1(K, \mathbb{R}^2)\right\},$$

and $\mathcal{I}_h : C(\overline{\Omega}, \mathbb{R}^2) \longrightarrow \mathbf{Y}_h$: the nodal interpolation operator such that $\mathcal{I}_h\Phi = \sum_{\mathbf{z}\in\mathcal{N}_h} \Phi(\mathbf{z})\varphi_{\mathbf{z}}$, where $\{\varphi_{\mathbf{z}} : \mathbf{z} \in \mathcal{N}_h\} \subset \mathbf{Y}_h$. Choose

$$\begin{aligned} \mathbf{X}_h &= \left\{\mathbf{v}_h \in C^0(\overline{\Omega}, \mathbb{R}^2) \cap H^1_0(\Omega, \mathbb{R}^2); \mathbf{v}_h/K \in P_2(K, \mathbb{R}^2)\right\}, \\ M_h &= \left\{q_h \in \mathbf{L}^2_0(\Omega);\ q_h/K \in P_0(K)\right\}, \end{aligned}$$

and

$$\mathbf{V}_h = \{\mathbf{v}_h \in \mathbf{X}_h :\ (\operatorname{div} \mathbf{v}_h, q_h) = 0\ \forall q_h \in M_h\}.$$

In the following, we use the L^2-orthogonal projections $Q_{\mathbf{Y}_h} : \mathbf{L}^2(\Omega, \mathbb{R}^2) \longrightarrow \mathbf{Y}_h$, $Q_{\mathbf{V}_h} : \mathbf{L}^2(\Omega, \mathbb{R}^3) \longrightarrow \mathbf{V}_h$ and the H^1-orthogonal projection $R_h : H^1(\Omega, \mathbb{R}^2) \longrightarrow \mathbf{Y}_h$.

In [1], the authors study a first fully discrete finite element approximation for the regularized problem (7)-(9), which uses the couple $(\mathbf{X}_h, M_h)$ of finite dimensional spaces for the velocity and for a new pressure $\widetilde{p} = \hat{p} + \lambda\mathbf{F}_\epsilon(\mathbf{d})$, where $\hat{p} = p + \frac{\lambda}{2}|\nabla\mathbf{d}|^2$.

Algorithm 1.

(1) Set $\mathbf{v}_h^0 := Q_{\mathbf{V}_h}\mathbf{v}_0{}^\epsilon$ and $\mathbf{d}_h^0 := R_{\mathbf{Y}_h}\mathbf{d}_0^\epsilon$.

(2) For $m = 1, ..., M$, let $\mathbf{f}_h^m := \left|\mathbf{d}_h^m\right|^2 \mathbf{d}_h^m - \mathbf{d}_h^{m-1}$. *Find* $(\mathbf{v}_h^m, \mathbf{d}_h^m, \tilde{p}_h^m, \mathbf{w}_h^m) \in \mathbf{X}_h \times \mathbf{Y}_h \times M_h \times \mathbf{Y}_h$ *such that, for all* $(\mathbf{u}_h, \mathbf{a}_h, q_h, \mathbf{b}_h) \in \mathbf{X}_h \times \mathbf{Y}_h \times M_h \times \mathbf{Y}_h$,

$$\begin{aligned} &(d_t\, \mathbf{v}_h^m, \mathbf{u}_h) + \nu(\nabla\mathbf{v}_h^m, \nabla\mathbf{u}_h) + \left((\mathbf{v}_h^{m-1}.\nabla)\, \mathbf{v}_h^m, \mathbf{u}_h\right) + \frac{1}{2}\left((\operatorname{div} \mathbf{v}_h^{m-1})\, \mathbf{v}_h^m, \mathbf{u}_h\right) \\ &\qquad\qquad + (\tilde{p}_h^m, \operatorname{div} \mathbf{u}_h) - \lambda\left((\nabla\mathbf{d}_h^{m-1})^T\, \mathbf{w}_h^m, \mathbf{u}_h\right) = \langle \mathbf{g}(t_m, .), \mathbf{u}_h\rangle, \\ &(d_t\, \mathbf{d}_h^m, \mathbf{a}_h) + \left((\mathbf{v}_h^m.\nabla)\, \mathbf{d}_h^{m-1}, \mathbf{a}_h\right) + \gamma(\mathbf{w}_h^m, \mathbf{a}_h) = 0, \\ &(\operatorname{div} \mathbf{v}_h^m, q_h) = 0, \\ &(\nabla\mathbf{d}_h^m, \nabla\mathbf{b}_h) + (\mathbf{f}_\epsilon^m, \mathbf{b}_h)_h - (\mathbf{w}_h^m, \mathbf{b}_h) = 0. \end{aligned}$$

Moreover, they have proved that the solution of Algorithm 1 verifies a discrete energy law, see [1] for more details. Next they study the following discrete scheme for the system (1)-(3), where an implicit treatment of the coupling terms is used in contrast to the semi-implicit discretization in Algorithm1. The discrete Laplacian $\Delta_h : W^{1,2}(\Omega) \to \mathbf{Y}_h$ and a temporal discretisation using the implicit midpoint rule are used.

Algorithm 2.

(1) Let $\mathbf{v}_h^0 := Q_{\mathbf{V}_h}\mathbf{v}^0$ and $\mathbf{d}_h^0 := \mathcal{I}_h\, \mathbf{d}^0$.

(2) Let $m = 1, ..., M$. *Find* $(\mathbf{v}_h^m, \mathbf{d}_h^m, \hat{p}_h^m) \in \mathbf{X}_h \times \mathbf{Y}_h \times M_h$ *such that, for all* $(\mathbf{u}_h, \mathbf{a}_h, q_h) \in \mathbf{X}_h \times \mathbf{Y}_h \times M_h$, *there holds*

$$\begin{aligned}&(d_t\,\mathbf{v}_h^m, \mathbf{u}_h) + \nu(\nabla\mathbf{v}_h^m, \nabla\mathbf{u}_h) + \Big((\mathbf{v}_h^{m-1}.\nabla)\,\mathbf{v}_h^m, \mathbf{u}_h\Big) + \frac{1}{2}\,\Big((\operatorname{div}\mathbf{v}_h^{m-1})\,\mathbf{v}_h^m, \mathbf{u}_h\Big)\\ &\qquad - (\hat{p}_h^m, \operatorname{div}\mathbf{u}_h) - \lambda\Big((\nabla\mathbf{d}_h^{m-1})^T\,\Delta_h\mathbf{d}^{m-\frac{1}{2}}, \mathbf{u}_h\Big) = \langle\mathbf{g}(t_m,.), \mathbf{u}_h\rangle,\\ &(\operatorname{div}\mathbf{v}_h^m, q_h) = 0,\\ &(d_t\,\mathbf{d}_h^m, \mathbf{a}_h) + \Big((\mathbf{v}_h^m.\nabla)\,\mathbf{d}_h^{m-1}, \mathbf{a}_h\Big) + \gamma\Big(\mathbf{d}_h^{m-\frac{1}{2}} \times (\mathbf{d}_h^{m-\frac{1}{2}} \times \Delta_h\mathbf{d}_h^{m-\frac{1}{2}}),\ \mathbf{a}_h\Big) = 0,\end{aligned}$$

where $\mathbf{d}_h^{m-\frac{1}{2}} = \frac{1}{2}(\mathbf{d}_h^{m-1} + \mathbf{d}_h^m)$.

A discrete energy law is also proved for the solutions of this scheme. The main contribution in this note is to change the term $\Big((\mathbf{v}_h^m.\nabla)\,\mathbf{d}_h^{m-1}, \mathbf{a}_h\Big)$ in Algorithm 2 by $\frac{1}{2}\Big((\mathbf{v}_h^m.\nabla)\,\mathbf{d}_h^{m-\frac{1}{2}}, \mathbf{a}_h\Big) - \frac{1}{2}(\mathbf{d}^{m-\frac{1}{2}}, \mathbf{v}_h^m.\nabla\mathbf{a}_h)$.

Then, the new Algorithm reads as follows:

Algorithm 3.

(1) Let $\mathbf{v}_h^0 := Q_{\mathbf{V}_h}\mathbf{v}^0$ and $\mathbf{d}_h^0 := \mathcal{I}_h\,\mathbf{d}^0$.

(2) Let $m = 1, ..., M$. *Find* $(\mathbf{v}_h^m, \mathbf{d}_h^m, \hat{p}_h^m) \in \mathbf{X}_h \times \mathbf{Y}_h \times M_h$ *such that, for all* $(\mathbf{u}_h, \mathbf{a}_h, q_h) \in \mathbf{X}_h \times \mathbf{Y}_h \times M_h$, *there holds*

$$\begin{aligned}&(d_t\,\mathbf{v}_h^m, \mathbf{u}_h) + \nu(\nabla\mathbf{v}_h^m, \nabla\mathbf{u}_h) + \Big((\mathbf{v}_h^{m-1}.\nabla)\,\mathbf{v}_h^m, \mathbf{u}_h\Big) + \frac{1}{2}\,\Big((\operatorname{div}\mathbf{v}_h^{m-1})\,\mathbf{v}_h^m, \mathbf{u}_h\Big)\\ &\qquad - (\hat{p}_h^m, \operatorname{div}\mathbf{u}_h) - \lambda\Big((\nabla\mathbf{d}_h^{m-1})^T\,\Delta_h\mathbf{d}^{m-\frac{1}{2}}, \mathbf{u}_h\Big) = \langle\mathbf{g}(t_m,.), \mathbf{u}_h\rangle,\\ &(\operatorname{div}\mathbf{v}_h^m, q_h) = 0,\\ &(d_t\,\mathbf{d}_h^m, \mathbf{a}_h) + \frac{1}{2}\Big((\mathbf{v}_h^m.\nabla)\,\mathbf{d}_h^{m-\frac{1}{2}}, \mathbf{a}_h\Big) - \frac{1}{2}(\mathbf{d}^{m-\frac{1}{2}}, \mathbf{v}_h^m.\nabla\mathbf{a}_h)\\ &\qquad + \gamma\Big(\mathbf{d}_h^{m-\frac{1}{2}} \times (\mathbf{d}_h^{m-\frac{1}{2}} \times \Delta_h\mathbf{d}_h^{m-\frac{1}{2}}),\ \mathbf{a}_h\Big) = 0,\end{aligned}$$

and with a judicious choice of the test function $\mathbf{a}_h = \mathbf{d}_h^{m-\frac{1}{2}}$, and by supposing that $\mathbf{d}_h^0 \in \mathbf{Y}_h$ satisfies $|\mathbf{d}_h^0| = 1$, the director field $\mathbf{d}$ satisfies the constraint (4).

§4. Numerical examples

In this section, we present and we compare numerical results using the algorithms 2 and 3. We use Newton's method for the solution of the nonlinear system at each time step. For this purpose, three Newton iterations are sufficient in our computations.

The following example is taken from [1] to approximate smooth solutions of (1)–(3).

Example 1. We consider $\Omega = (-1, 1)^2$, and $\mathbf{v}_0 \equiv 0$, $\mathbf{d}_0 = (\sin(a), \cos(a))^\top$, where $a = 2.0\,\pi(\cos(x) - \sin(y))$. The parameters are taken as follows: $\lambda = \gamma = 1$, $\nu = 0.1$. The initial condition $\mathbf{d}_0$ and the final state are shown in Figure 1.

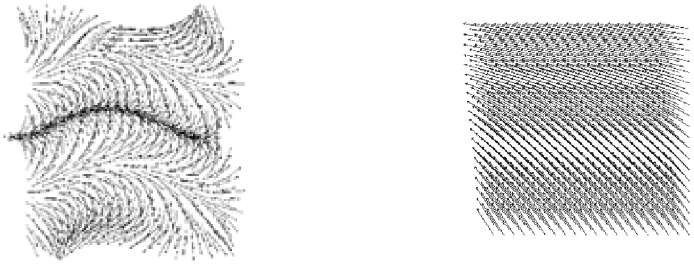


Figure 1: Algorithm 2: Initial (left) and final (right) director fields.

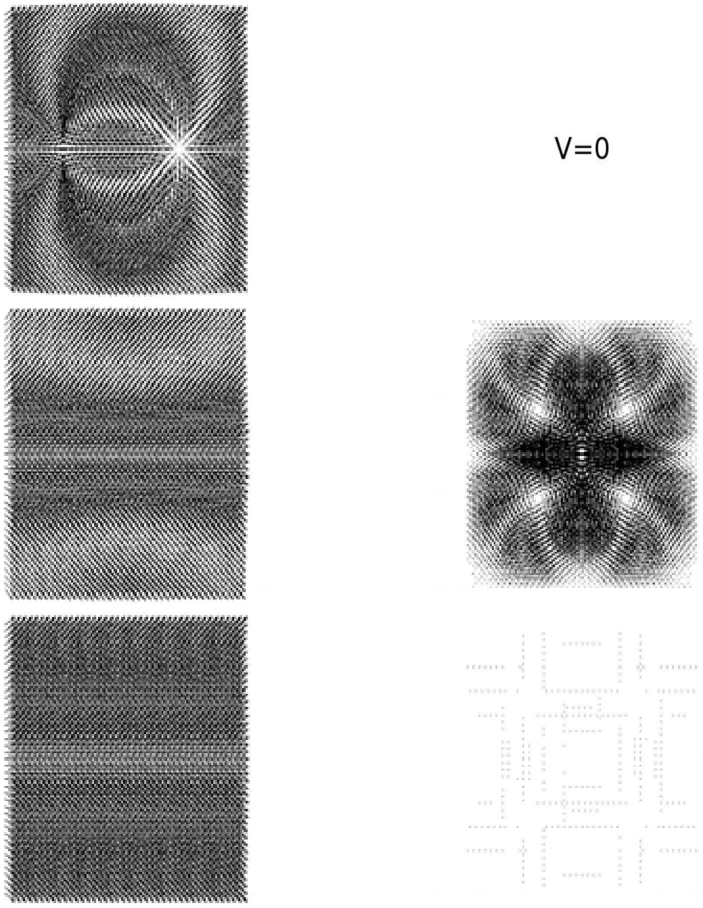


Figure 2: (Example 2) Using Algorithm 1. Snapshots at times $t = 0, 0.3, 0.9$ of $\{d_h^m\}$ (left) and $\{u_h^m\}$ (right).

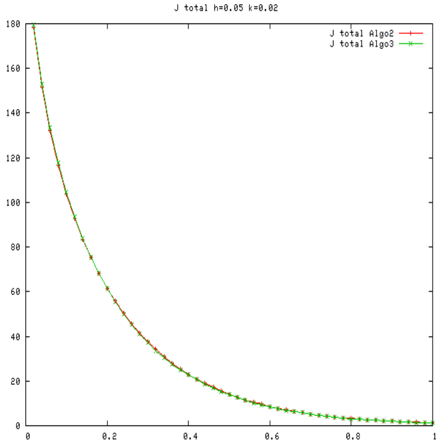


Figure 3: (Example 2) $\mathcal{J}_{total}$ with Algorithm 2 and Algorithm 3 for $k = 0.02$, $h = 0.05$ and $\eta = 0.1$.

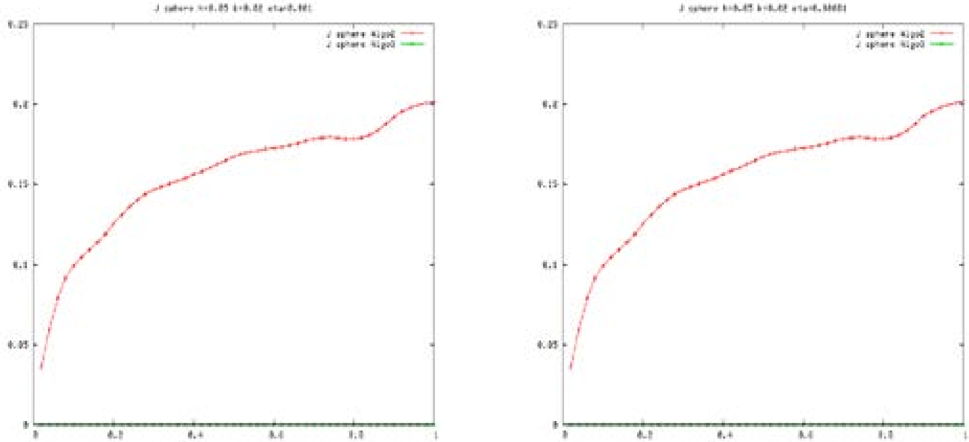


Figure 4: Comparison of Algorithm 2 and Algorithm 3 with Example 2 for $k = 0.02$, $h = 0.05$ and $\eta = 0.001$ (left), for $k = 0.02$, $h = 0.05$ and $\eta = 0.00001$ (right).

A uniform crisscross triangulation of Ω is used with uniform mesh size $h = 1/20$, $1/40$ and $1/80$. Next, we present the results for Algorithm 1.

Example 2. We consider $\Omega = (-1,1)^2$, $\mathbf{v}_0 \equiv 0$ and $\mathbf{d}_0 = \hat{\mathbf{d}} \,/\, (\left|\hat{\mathbf{d}}\right|^2 + \eta^2)^{1/2}$, with $\hat{\mathbf{d}}(x,y) = (x^2 + y^2 - 0.25, y)^T$. The parameters are taken as follows: $\lambda = \gamma = 1$, $\nu = 0.1$, $\eta = 0.05$, $\epsilon = 0.05$, $k = 0.01$ and $h = 0.1$. The evolution of this solution is shown in Figure 2.

In order to compare the solutions of Algorithm 2 with those of Algorithm 3, we give some notations. Let $\mathcal{J}_{kin}(\mathbf{v}_h) = \frac{1}{2}\int_\Omega |\mathbf{v}_h|^2$ be the kinetic energy, $\mathcal{J}_{ela}(\mathbf{d}_h) = \frac{\lambda}{2}\int_\Omega |\nabla \mathbf{d}_h|^2$, the elastic energy and finally $\mathcal{J}_{total}(\mathbf{v}_h, \mathbf{d}_h) = \mathcal{J}_{kin}(\mathbf{v}_h) + \mathcal{J}_{ela}(\mathbf{d}_h)$, the total energy.

We then compare the total energy for the two algorithms on the same mesh with the same time step. As can be seen from Figure 3, we have exactly the same solutions with the two algorithms.

We define now, the sphere energy $\mathcal{J}_{sphere}(\mathbf{d}_h) = 1 - |\mathbf{d}_h|$. The results presented in Figure 4 show that the sphere energy for the Algorithm 3 is zero and the constraint (4) is satisfied in contrast to Algorithm 2.

References

[1] Becker, R., Feng, X., and Prohl, A. Finite element approximation of the Ericksen-Leslie model for nematic liquid crystal flow. *SIAM J. Numer. Anal. 46*, 4 (2008), 1704–1731.

[2] Lin, F., and Liu, C. Nonparabolic dissipative systems modeling the flow of liquid crystals. *Communications on Pure and Applied Mathematics 48* (1995), 501–537.

Roland Becker
EPI Concha-Inria Bordeaux Sud-Ouest
Departement de Mathematiques Appliquées
Université de Pau et des Pays de l'Adour
I.P.R.A, B.P 1155
64013 Pau Cedex, France
roland.becker@univ-pau.fr

Nour El Houda Seloula
EPI Concha-Inria Bordeaux Sud-Ouest
Departement de Mathematiques Appliquées
Université de Pau et des Pays de l'Adour
I.P.R.A, B.P 1155
64013 Pau Cedex, France
nourelhouda.seloula@etud.univ-pau.fr

Monografías Matemáticas *García de Galdeano* **35**, 73–82 (2010)

STABLE MULTIQUADRIC APPROXIMATION BY LOCAL THINNING

Mira Bozzini and Licia Lenarduzzi

Abstract. In this paper our concern is the recovery of a highly regular function by a discrete set X of data with arbitrary distribution. We consider the case of a nonstationary multiquadric interpolant that presents numerical breakdown. Therefore we propose a global least squares multiquadric approximant with a center set T of maximal size and obtained by a new thinning technique. The new thinning scheme removes the local bad conditions in order to obtain $A_{X,T}$ well conditioned. The choice of working on local subsets of the data set X provides an effective solution. Some numerical examples to validate the goodness of our proposal are given.

Keywords: Scattered data, arbitrary distribution, thinning, least squares, non stationary multiquadric approximant.

AMS classification: 65D10, 41A05.

§1. Introduction

In this article we address the problem of recovering a function with high regularity by using a set of data with arbitrary distribution.

It is well known that the radial basis functions (RBF) are a powerful tool for the multivariate approximation from scattered data. Nevertheless the arbitrary distribution of the data can lead to an ill conditioned problem. In fact the standard methods involve the solution of linear systems whose matrices can be ill conditioned also for moderate size.

In the literature we find various approaches to solve the problem of the ill conditioning when RBFs are considered. A wide list of papers can be found in the recent article [2]. Anyway, the techniques developed take into account only the case of samples from quasi uniform distributions.

In the present paper, the reconstruction of the unknown function is provided by a multiquadric (MQ) least squares approximant with the basis functions located at centers, determined such that the collocation matrix is well conditioned in the sense that Matlab does not display a warning that it is close to singular. The procedure to select the centers is studied in order to provide a solution with very good accuracy.

The note is organized as follows: the main result is presented in §4 and it concerns the determination of the set of the centers; and before we give some notations in §2 and we present the least squares approximant by radial basis functions in §3. Finally in §5 we provide some numerical examples that simulate real applicative cases to show the effectiveness of the proposed technique.

§2. Notations

Given $\{(x_j, f_j),\ j = 1, \ldots, N\}$ with data sites $x_j \in D \subset \mathbb{R}^2$ and values $f_j = f(x_j) \in \mathbb{R}$ measured from some unknown function $f \in C^\alpha(D)$, $\alpha > 2$, we indicate with X the set of the data sites x_j, $j = 1, \ldots, N$, and with $q(X)$ the minimal distance among the X sites. Similarly we name $T := \{t_1, \ldots, t_M\}$ a set of distinct points $t_k \in D$ and $q(T)$ the minimal distance among themselves.

We consider the multiquadric function with fixed parameter δ

$$\phi((\cdot - y), \delta) := (\|\cdot - y\|_2^2 + \delta^2)^{1/2} = \phi_\delta(\cdot)$$

and we denote with $A_{X,X}$ the interpolation matrix with entries $a_{ij} = \phi((x_i - x_j), \delta)$, $x_i, x_j \in X$ and with $A_{X,T}$ a matrix whose entries are $b_{ij} = \phi((x_i - t_j), \delta)$, $x_i \in X$, $t_j \in T$. Let $\mathcal{K}_2(A_{X,X})$ be the spectral condition number.

Given two sets X and T, we define the covering radius according to the l_2 measure of the set T on X

$$r_{TX} = \max_{x_j \in X} d_T(x_j),$$

where

$$d_T(x_j) = \min_{t_k \in T} \|t_k - x_j\|_{l_2}. \tag{1}$$

Another important parameter is the fill distance

$$h_D(X) = \max_{x \in D} d_X(x),$$

where

$$d_X(x) = \min_{x_j \in X} \|x - x_j\|_{l_2}.$$

We observe that, in the case where we consider the generic point $x \in D$ in (1), the fill distance is synonymous of covering radius.

§3. About least squares approximation

We are interested in providing a solution of the least squares problem when we sample a function f on the set $X = \{x_1, \ldots x_N\}$ of data sites and consider a second set $T := \{t_1, \ldots, t_M\}$, at which we center the multiquadric bases with fixed parameter δ. Let it be $M < N$.

Let the approximant be of the form

$$Q_f(x) = \sum_{j=1}^{M} c_j \phi((x - t_j), \delta), \quad x \in \mathbb{R}^2.$$

The coefficients $\{c_j\}$ are obtained as solution of the least squares problem

$$A_{XT}\mathbf{c} = \mathbf{f},$$

where $\mathbf{f} = \{f_1, \ldots, f_N\}$. The system has a unique solution if the matrix $A_{X,T}$ of entries $\{A_{ij} := \phi((x_i - t_j), \delta)\}$, $i = 1, \ldots N$, $j = 1, \ldots M$, has full rank.

Unfortunately it is not clear how to choose the set T; in fact there is not much mathematical theory to guarantee that this approach is well posed. However, often the least squares method is a valid tool to obtain a global approximation which takes into account the whole information given by the problem. In this connection we want to show by a simple example that it is very important not to discard any datum.

We consider the set X shown in Fig. 6. In this set the presence of two couples, where the points are very close to each other, leads to an unstable interpolation matrix. A possible choice is that of considering for each couple only one point and then to interpolate the data of the new set $\tilde{X}$ whose matrix $A_{\tilde{X},\tilde{X}}$ is stable. In this way, considering Franke's test function, we obtain a maximum error

$$e_\infty = 7.56\ e(-2),$$

computed on a grid 61 × 61. Otherwise, considering the least squares method with full rank matrix $A_{X,T}$, we obtain a maximum error

$$e_\infty = 3.80\ e(-2).$$

We observe that to consider all the given functional values leads to an accuracy of one order greater than that obtained by stable interpolation.

Therefore, having considered what we have said above, we have developed a wide experimentation in order to find some information about the construction of a set T associated to a full rank matrix $A_{X,T}$, less hard than the theoretical properties given by Quak, Sivakumar and Ward in [9]. On the basis of our experimentation, one can make the conjecture: " the rank of the matrix $A_{X,T}$ mainly depends on the parameter $q(T)$". Such a statement was indicated also by Buhmann in [1].

§4. Determination of the set T

We suppose that the interpolation matrix $A_{X,X}$ presents a numerical breakdown. Therefore, as it is not possible to consider the interpolant, we want to individuate a set T of centers with a size as large as possible such that the matrix $A_{X,T}$ has full rank to provide a least squares approximant.

The first step in the construction of the set T deals with the determination of an upper bound M_0 for the size of T. For this aim, we recall that the numerical stability of the RBF-ϕ interpolation depends on $G_\phi(q(X))$, where $G_\phi : [0,\infty) \to [0,\infty)$ is a monotonically increasing function. It follows that as the size N of the sample increases and $q(X)$ decreases, the spectral condition number $\mathcal{K}_2(A_{X,X})$ grows.

But we observe that, for a given distribution of the sample, the value of N for which the matrix $A_{X,X}$ presents numerical breakdown depends on the function ϕ or on the value assigned to the parameter of those radial bases such as the Gaussian or the multiquadric for example.

For instance, considering the vertices of a regular hexagonal grid as point locations, we note that a fit by Gaussians, with parameter $\epsilon = 1$, quickly leads Matlab to display a warning of matrix close to be singular, as soon as $N > 46$; whereas with the multiquadric basis with parameter $\delta = 1$ the warning is displayed with $N \geq 85$.

The polyharmonic basis functions are more stable. Sizes of the sample of the order of one thousand can be used to interpolate without instability.

Having fixed the multiquadric basis, we must still define which set of points to choose from which to determine the value M_0 that bounds the size of T with the chosen basis. It is natural to require the stability of the solution for the set T, but also an optimal accuracy. So we recall that all the least squares estimates are based on interpolation error estimates.

In the interpolation problem the pointwise error bound depends on the fill distance of the set of the data sites in D. It follows that, to determine the maximum bound of the size for the set T, we have to search a set T_0 of discrete points in D for which on one hand the value $q(T_0)$ is as large as possible and on the other hand the value $h_D(T_0)$ is as small as possible in order to minimize the mesh ratio $\rho_D(T_0) = h_D(T_0)/q(T_0)$. It is known that the optimal set in $\mathbb{R}^2$ is given by the vertices of a regular hexagonal grid, [4].

4.1. Determination of the bound M_0

Our goal is to consider a set T with parameter $h_D(T)$ as small as possible; this is equivalent to look for a set T of cardinality as large as possible, constrained so that A_{TT} is numerically stable.

Having fixed the multiquadric ϕ_δ, we consider a regular hexagonal grid V_l and compute the spectral condition number $\mathcal{K}_2(A_{V_l,V_l})$. Then we decrease the step of the grid V_l and we obtain a new set of vertices V_{l+1} and a new value $\mathcal{K}_2(A_{V_{l+1},V_{l+1}})$. The process goes on, for a value $l = L$ of the index, until we meet the matrix A_{V_L,V_L} numerically unstable.

The set V_L is the one that corresponds to the value of minimal mesh ratio among all the sets of same cardinality: $\rho_D(V_L) = \min \rho_D(T)$, $|T| = |V_L|$. The value L is the upper bound M_0; it corresponds to the optimal distribution of centers, but it is not always suitable to take them, when considering a set X of scattered points with arbitrary distribution. The value of the size M of the set T will be the closer to M_0 the more the distribution of the sample is almost uniform.

4.2. Determination of T: sketch of the procedure

Once the basis ϕ_δ, for which the value M_0 is known, is fixed, the individuation of the set T is worked in two steps. At first we determine a proper subset of X and then we improve its covering radius on X. To construct the proper subset of X we take into account that:

- Coalescent points determine instability and hence matrices of moderate size can be unstable.
- A warning of ill conditioning depends on the value of $q(X)$ but also on the geometry of the points. With respect to this we show the following example to validate the statement. We consider two different small configurations X, both of size 5 that are shown in Fig. 1 and Fig. 2 respectively; we provide the values of $q(X)$ and $\mathcal{K}_2(A_{X,X})$, when interpolating with the multiquadric with parameter ($\delta = 10$).

  In the first case it is $\mathcal{K}_2(A_{X,X}) = 8.62\ e(15)$ and $q_1(X) = 10\ e(-4)$; in the second case it is $\mathcal{K}_2(A_{X,X}) = 1.77\ e(16)$ even if the minimal distance $q_2(X) = 10\ e(-3)$ is larger than $q_1(X) = 10\ e(-4)$.
- A bad local condition involves a bad global condition

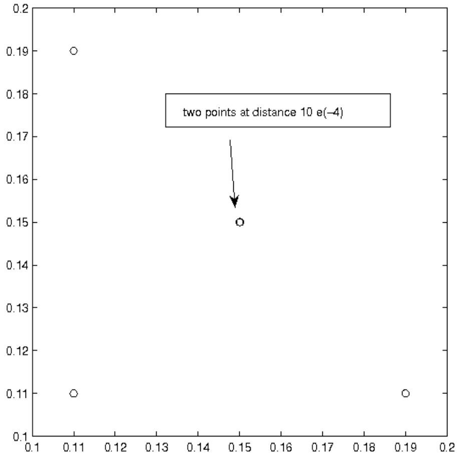


Figure 1:

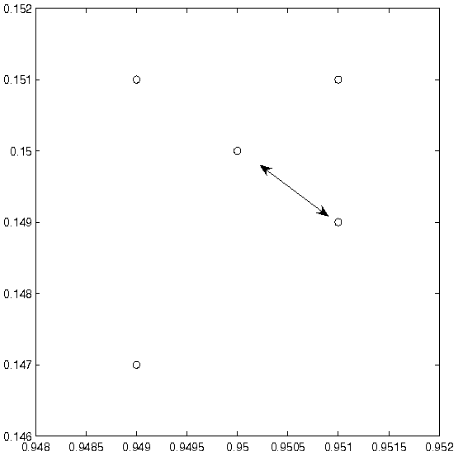


Figure 2:

In the following we sketch the procedure to determine the centers and we refer to [6] for the whole description of the algorithm. For clearness of exposition, we assume that we are in the presence of a sample of size N sampled from uniform distribution.

Using the Delaunay triangulation $\mathcal{D}(X)$, we sort the data sites in a vector $Y = \{y_1, \dots, y_N\}$ according to the increasing distance from the contiguous points within $\mathcal{D}(X)$. It follows that the last M_0 components of the vector Y correspond to points whose interpoint distances are larger.

We indicate with $Y_0 := \{y_{N-M_0+1}, \dots, y_N\}$ such a vector and we consider for each component $y_k \in Y_0$ its Voronoi cell $V(y_k)$. We calculate the set $Y_0^\star = \{y^\star_{N-M_0+1}, \dots, y^\star_N\}$ whose components correspond to the barycenters of the sets $X \cap V(y_k)$, $k = N - M_0 + 1, \dots, N$. The vector $Y_0^\star$ has covering radius $r_{Y_0^\star X}$ less than $r_{Y_0 X}$.

By this operation the value of $q(Y_0^\star)$ is larger than $q(X)$; nevertheless $A_{Y_0^\star, Y_0^\star}$ can be unstable due to particular geometries of the data sites. In this last case we consider a subdivision of $Y_0^\star$ in subsets $\{S_j\}$ worked in the following way.

We construct the Delaunay triangulation $\mathcal{D}(Y_0^\star)$ on $Y_0^\star$ and, for each $y_k^\star \in Y_0^\star$, we calculate the average distance

$$\tau_k = 1/n_k \sum_1^{n_k} \mathrm{dist}_2(y_k^\star, y_j^\star)$$

from its neighbouring centers. We construct the vector Z whose components $z_k \in Y_0^\star$ are sorted by increasing values of τ_k. The first components of Z correspond to regions of D with largest density of $Y_0^\star$ points.

Let m be fixed and let us start with the first component z_1 of Z to determine the $(m-1)$ points $y_i^\star \in Y_0^\star$ closest to z_1 according to dist_∞. Let us indicate with S_1 such a set. Successively we determine the subsets S_j in the same way by considering the component $z_j \in Z \setminus \cup_{k=1}^{j-1} S_k$ and by individuating the $(m-1)$ points $y_j^\star \in Y_0^\star \setminus \cup_{k=1}^{j-1} S_k$ closest to z_j. The process ends in a finite number of steps.

For each subset S_j we evaluate the condition $\mathcal{K}_2(A_{S_j,S_j})$. In the case of a bad condition we discard those points that determine instability. We indicate with $Y_1^\star$ the set of the points $y_k^\star \in Y_0^\star$ not discarded.

The set $Y_1^\star$ has been determined on the basis of local interpolation matrices well conditioned, but this does not ensure that the global matrix $A_{Y_1^\star,Y_1^\star}$ is numerically stable. In fact it could happen that, when subdividing the set $Y_0^\star$ into subsets, some geometry of points of the global set have been split. When the matrix $A_{Y_1^\star,Y_1^\star}$ is unstable, the step of the subdivision is repeated on $Y_1^\star$. The recursive process gets a set of centers well separated by few iterations. Let M_F be the cardinality of $Y^\star$. As before, we improve the covering radius by considering for each point $y_j \in Y^\star$ its Voronoi cell $V_j^\star$ and we construct the set T of the barycenters of the points $\{x_k \in X \cap V_j^\star\}$, $j = 1, \dots M_F$.

The procedure here described has low computational cost because it works on sets of small sizes. The bigger cost is due to the thinning scheme described in [3] to determine the initial set Y_0 needed when $N > M_0$. Such a cost is of the order of $N \log N$. When $N \leq M_0$ the procedure takes $Y_0 \equiv X$.

The procedure, here described in short, can be suitably adapted to the case of arbitrary distributions. In the already cited report [6] the cases of uniform distribution, clusters of data and distributions dependent on the phenomenon are considered.

§5. Examples

We shall show three examples relevant to three different distributions that can be met in different applicative problems. For each one of the examples quoted, we provide: the maximum error e_∞ computed on a grid 61×61, the cardinality of the set T and the value of the index of spectral condition $\mathcal{K}_2(A_{X,T})$ provided by Matlab as well as the size N of the set X. In all the examples we take the unitary square $[0,1]^2$ as D and we take the shifts of the multiquadric with parameter $\delta = 0.35$ as basis functions. For this basis the value of M_0 is 314.

Finally the results have been compared with some known methods in the literature, in particular with techniques of knot removal to construct the set of the centers, [4], and with the approximated interpolation, [10].

The last method has a solution given by a linear combination of shifts of a RBF$-\phi$

$$P_f(x) = \sum_1^N \hat{c}_j \phi(x - x_j)$$

whose coefficients $\{\hat{c}_j\}_1^N$ solve the system

$$(A_{XX} + \lambda I)\mathbf{c} = \mathbf{f},$$

where the parameter λ is chosen in a theoretical way based on the smoothness of the unknown function f.

5.1. Example 1

Let us consider a sample of size $N = 198$ from a distribution with variable densities depending on the behaviour of Franke's function. This way of collecting the information, relevant to

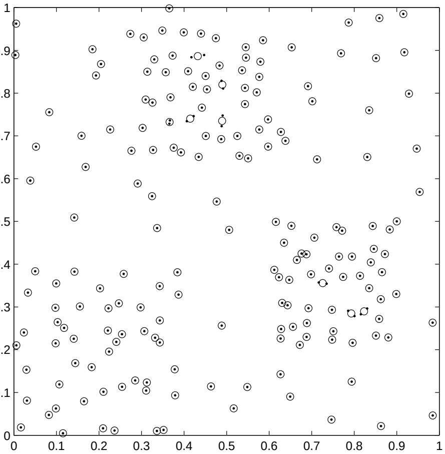


Figure 3: X dotted, T circled

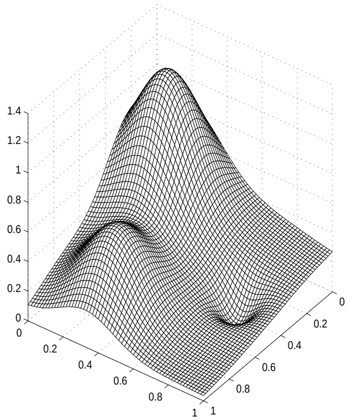


Figure 4: The approximant $s_{X,T}$

the phenomenon to be recovered at hand, is used, for example, in problems of clinical survey or in geophysical problems. In this case it is $N < M_0$ and it is $\mathcal{K}_2(A_{X,X}) = 2.38\ e(16)$ with a warning of not full rank from Matlab:

- According to the current sketched procedure the set T selected, by just one iteration, is the one shown in Fig. 3 of size 189 and with $\mathcal{K}_2(A_{X,T}) = 1.60\ e(13)$. The graphic is shown in Fig. 4 and the error is

  $$e_\infty(X,T) = 1.94\ e(-3).$$

  The running time was 0.88 sec., excluding the computation of Y_0, on a AMD 64 working as a monoprocessor.

- By using the thinning technique presented in [3] to construct the set of the centers, a set $\tilde{X}$ of 183 centers is selected.

  The set of the centers $\tilde{X}^\star$, obtained by improving the covering radius, determines a stable matrix $A_{X,\tilde{X}^\star}$ with error

  $$e_\infty(X,T) = 2.30\ e(-3),$$

  and the total running time was 1.20 sec, excluding the computation of Y_0.

- When constructing the approximated interpolation with the good value $\lambda = 10^{-10}$ we obtain

  $$e_\infty(X,X) = 2.17\ e(-3).$$

5.2. Example 2

We assign a sample of mildly scattered data of size $N = 1600$ from the valley test function, [7]. The sample is oversized to simulate the case of laser measures of a smooth feature

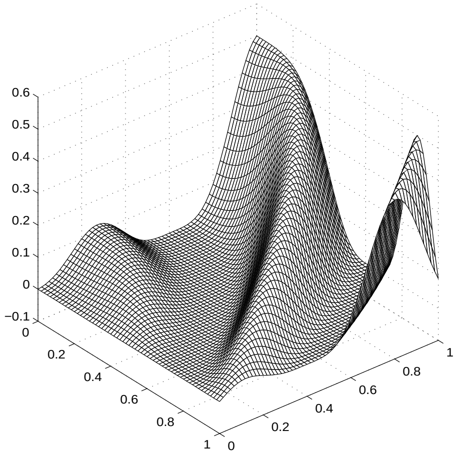


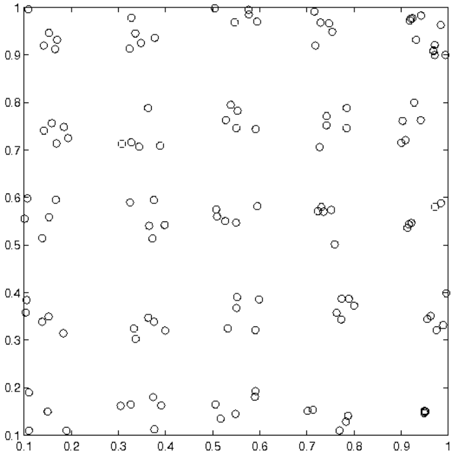
Figure 5: The approximant $s_{X,T}$

as in industrial applications. Moreover the coefficients and the centers of $s_{X,T}$ provide a compressed information for the data set.

- Our algorithm selects T of size $M = M_0 = 314$. The index of condition is $\mathcal{K}_2(A_{X,T}) = 2.92\ e(12)$ and the number of iterations is 3 . The graphical output is shown in Fig. 5 and the error is
$$e_\infty(X,T) = 1.88\ e(-3).$$
- By approximated interpolation with $\lambda = 10^{-9}$ the corresponding error is
$$e_\infty(X,X) = 8.59\ e(-3).$$
Besides, for $N \gg M_0$, the approximated interpolation is expensive, because it is necessary to work with a full matrix $N \times N$ for each value of λ.
- By using the modified Shepard's method run with our radial basis and with the same parameters of locality as described in [5], we obtain an error
$$e_\infty(X,X) = 3.20\ e(-3).$$
If we want a maximum error as small as the one obtained with our procedure, we have to consider a sample of size $N > 3000$.

5.3. Example 3

Here we consider the case of a configuration of clusters of points. Cluster sampling has many analogies to real-world sampling. It is a set of densely sampled areas with large gaps where no samples are taken, [8]. $N = 131$ data from Franke's test function are taken with $X = \cup_{r=1}^{n=25} S_r$ where with $\{S_r\}$ we have named the clusters, as shown in Fig. 6. The value of



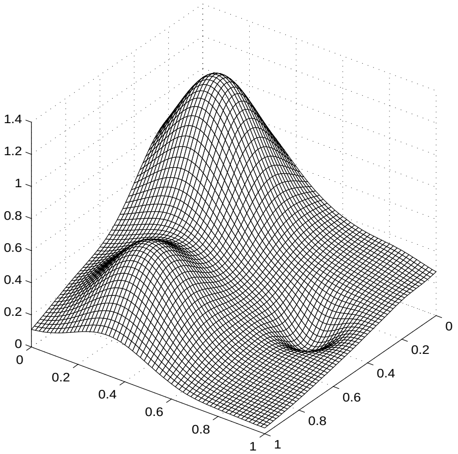


Figure 6: X locations

Figure 7: The approximant $s_{X,T}$

M_0 for this configuration is 260; so we put $Y_0 \equiv X$. The algorithm, working locally, by one iteration, determines the set T by discarding one point only belonging to the 25-th cluster located in the top right hand corner. It turns out that $M = 130$ and $\mathcal{K}_2(A_{X,T}) = 4.11\ e(13)$.

The graphical output of $s_{X,T}$ is shown in Fig. 7 and the error is

$$e_\infty(X, T) = 3.80\ e(-2).$$

There are cases where there are some points very near each other, as happens in the case of real-world sampling. We could use an adaptive technique that exchanges the data of locations, that are very near each other, with the average of their functional values, placed at their barycenter. By using such a technique in our case, the associated interpolation matrix $A_{\tilde{X},\tilde{X}}$ presents $\mathcal{K}_2(A_{\tilde{X},\tilde{X}}) = 4.04\ e(13)$ and the errors of the interpolant are

$$e_\infty(\tilde{X}, \tilde{X}) = 7.56\ e(-2).$$

Acknowledgements

We are grateful to Armin Iske who encouraged us during our work and gave us advice for the draft of the paper.

References

[1] Buhmann, M. *Radial Basis Functions*, vol. 12 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2003.

[2] Fasshauer, G., and Zhang, J. Preconditioning of radial basis function interpolation systems via accelerated iterated approximate moving least squares approximation. *Progress on Meshless Methods* (2008). A. J. M. Ferreira et al. (eds.).

[3] Floater, M., and Iske, A. Thinning and approximation of large sets of scattered data. *in Advanced Topics in Multivariate Approximation* (1996). F. Fontanella, K. Jetter, and P.J. Laurent (eds.), World Scientific, Singapore.

[4] Iske, A. *Multiresolution Methods in Scattered Data Modelling*, vol. 37 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin, 2004.

[5] Lazzaro, D., and Montefusco, L. B. Radial basis functions for the multivariate interpolation of large scattered data sets. *J. Comput. Appl. Math. 140* (2002), 521–536.

[6] Lenarduzzi, L. Stable multiquadric approximation of scattered data by local thinning: schemes of the algorithms. *IMATI report 3* (2009).

[7] Nielson, G. A first-order blending method for triangles based upon cubic interpolation. *Internat. J. Numer. Meth. Engr. 15* (1978), 308–318.

[8] Nielson, G. Scattered data modelling. *IEEE Comp. Graph. Appl.* (1993), 60–70.

[9] Quak, E., Sivakumar, N., and Ward, J. D. Least squares approximation by radial functions. *SIAM J. Math. Anal. 24* (1993), 1043–1066.

[10] Wendland, H., and Rieger, C. Approximate interpolation with applications to selecting smoothing parameters. *Num. Math. 101* (2005), 643–662.

Mira Bozzini
Università di Milano Bicocca
Dipartimento di Matematica ed Applicazioni
via Cozzi 53
20125 Milano, Italy
`mira.bozzini@unimib.it`

Licia Lenarduzzi
Istituto di Matematica Applicata e Tecnologie
Informatiche del C. N. R. Sez. Milano
via Bassini 15
20133 Milano
`licia@mi.imati.cnr.it`

Monografías Matemáticas *García de Galdeano* **35**, 83–90 (2010)

ERROR GROWTH IN THE NUMERICAL INTEGRATION OF PERIODIC ORBITS WITH PROJECTION METHODS

M. Calvo, M. P. Laburta, J. I. Montijano and L. Rández

Abstract. The aim of this work is to show that, when projection techniques are used in connection with Runge–Kutta (RK) methods to preserve first integrals of some periodic differential systems, the global error of the numerical solution presents a linear growth, even though the integration advances with a variable stepsize strategy.

Keywords: initial value problems, periodic solutions, Runge–Kutta methods, projections methods, error growth, preservation of first integrals.

AMS classification: 65L05, 65L06.

§1. Introduction

In this paper, autonomous initial value problems of the form

$$x'(t) = f(x(t)), \tag{1}$$

$$x(t_0) = x_0 \in \mathbb{R}^m, \tag{2}$$

which possess a unique periodic solution $x = \varphi(t; x_0)$ with period $T_0 > 0$, are considered. The function f is supposed as smooth as necessary.

Let φ_t be the t-flow map of (1), and ψ_h the function that defines a smooth one-step method to solve numerically (1)–(2). Thus, we obtain approximations x_n to the exact solution of this initial value problem at the gridpoints $t_n = t_{n-1} + h_{n-1}$:

$$x_n = \psi_{h_{n-1}}(x_{n-1}) = \psi_{h_{n-1}} \circ \cdots \circ \psi_{h_0}(x_0) \simeq \varphi_{t_n}(x_0), \quad n = 1, 2, 3, \dots,$$

where $h_0, h_1, h_2, \dots$ is a sequence of positive stepsizes.

B. Cano and J. M. Sanz-Serna have considered in [5] one-step methods ψ_h for the numerical integration of (1)–(2) satisfying the following conditions:

(i) ψ_h is defined in $\mathbb{R}^m$ for $|h| \le |h_0|$ for some $h_0 > 0$.

(ii) $\psi_h(x)$ depends smoothly on h and x.

(iii) The method ψ_h is consistent of order $r \ge 1$, r integer:

$$\psi_h(x) - \varphi_h(x) = O(h^{r+1}),\ h \to 0.$$

(iv) The Jacobian matrices satisfy

$$\psi'_h(x) - \varphi'_h(x) = O(h^{r+1}),\ h \to 0.$$

(v) The stepsizes are determined by

$$h_n = h\, s(x_n, h), \quad h > 0, \quad n = 0, 1, 2, \ldots,$$

where $s(x, h)$ is a smooth real-valued function such that

$$s_{\min} \leq s(x, h) \leq s_{\max},$$

for suitable positive constants $s_{\min}$ and $s_{\max}$.

Under these five assumptions, the authors prove in [5] the following asymptotic expansion in powers of h for the global error:

$$x_n - \varphi(t_n; x_0) = h^r e_r(t_n) + \cdots + h^{2r-1} e_{2r-1}(t_n) + h^{2r-1} R(t_n, h), \ h \to 0,$$

where the error functions $e_k(t)$ satisfy non-homogeneous variational equations of (1) with respect to $\varphi(t; x_0)$, and $R(t, h) \to 0$ as $h \to 0$ in bounded time intervals. The authors also prove in [5] that these error functions at integer multiples of the period, $e_k^{(N)} = e_k(NT_0)$, $r \leq k \leq 2r - 1$, satisfy:

$$e_k^{(N)} = \left(\sum_{i=0}^{N-1} M_{t_0}^i \right) e_k^{(1)}, \quad N = 1, 2, \ldots,$$

where $M_{t_0} = M(t_0 + T_0, t_0)$ is the monodromy matrix associated to the T_0-periodic solution $\varphi(t; x_0)$.

In particular, M. P. Calvo and J. M. Sanz-Serna had shown in [4] that integrating elliptic orbits in the two-body problem with a symplectic method using a constant stepsize policy, the global error grows linearly with the number of periods. They point out that such study is extensible to periodic Hamiltonian problems whose period depends only on the energy.

In this article we present a study of the growth of the global error integrating periodic differential systems (not necessarily Hamiltonian) so that the periodic orbit is embedded into a family of periodic orbits. We present some numerical experiments over this kind of problems with projection Runge–Kutta (RK) methods.

§2. Error behaviour

We assume the following hypothesis (H) for the differential system (1):

> *For all $\widetilde{x}_0$ in some neighbourhood of x_0, the solution of (1) with initial value $\widetilde{x}_0$ at time t_0, $\varphi(t; \widetilde{x}_0)$, is periodic with period $T = T(\widetilde{x}_0)$ where the function T is as smooth as required* (H)

To integrate this kind of differential problems we consider one step methods satisfying the above assumptions (i)–(v). In [3], we show that the matrix M_0 can be written as

$$M_0 = I - f(x_0) \nabla T(x_0)^T.$$

This expression for M_0 allows to simplify the powers of this matrix, and we obtain that the global error coefficients after N periods satisfy

$$e_k^{(N)} = N e_k^{(1)} - \frac{N(N-1)}{2} f(x_0) \nabla T(x_0)^T e_k^{(1)}, \quad r \leq k \leq 2r - 1$$

Furthermore, in [3] we show that the global error of the numerical method after N periods can be written as:

$$\begin{aligned} ge^{(N)} &= Nge^{(1)} - \frac{N(N-1)}{2} f(x_0)\nabla T(x_0)^T ge^{(1)} - h^{2r-1} NR(T_0,h) \\ &\quad + h^{2r-1}\frac{N(N-1)}{2} f(x_0)\nabla T(x_0)^T R(T_0,h) + h^{2r-1} R(NT_0,h). \end{aligned} \tag{3}$$

As a consequence, we prove the following result:

Theorem 1. *Let us consider a differential system (1), (2) satisfying the hypothesis* (H)*, and an one-step method satisfying the conditions (i)–(v). Then, if the method preserves the period T up to order $O(h^{2r})$, the global error grows linearly on t, provided that Nh^r is small.*

Let us see now how the error of first integrals of the differential system behaves. Let $G(x)$ be an scalar first integral of (1). We denote by

$$\Delta^{(N)}G = G(x_0 + ge^{(N)}) - G(x_0), \quad N = 1, 2, \ldots,$$

the error in the invariant G for the method ψ_h after N periods. Since

$$\Delta^{(N)}G = \nabla G(x_0)^T ge^{(N)} + O\left(\|ge^{(N)}\|^2\right),$$

taking into account (3) we obtain:

Theorem 2. $\Delta^{(N)}G = N\,\Delta^{(1)}G + O(Nh^{2r-1}) + h^{2r-1}\nabla G(x_0)^T R(NT_0,h) + O\left(\|ge^{(N)}\|^2\right).$

This asymptotic relation implies a linear error growth in the invariant with the number of periods provided that $\|ge^{(N)}\|$ is not too large.

A Runge–Kutta method with a projection technique applied to (1) with $x(t_0) = u$, provides a numerical approximation $\widehat{\psi}_h(u)$ given by

$$\widehat{\psi}_h(u) = \psi_h(u) + \lambda(u,h)w(u,h),$$

where:

- ψ_h is an s-stage RK method of order r:

$$\begin{cases} \psi_h(u) = u + h\displaystyle\sum_{j=1}^{s} b_j\, f(U_j), \\ U_j = u + h\displaystyle\sum_{k=1}^{s} a_{jk} f(U_k), \quad (j = 1, \ldots, s), \end{cases}$$

- The coefficient $\lambda(u,h) \in \mathbb{R}$ is computed so that $\widehat{\psi}_h(u)$ preserves the invariant G of (1): $G(\widehat{\psi}_h(u)) = G(u)$.
- The vector $w(u,h) \in \mathbb{R}^m$ depends on the type of projection used. In this article, we will use directional projection [2], which takes $w = \psi_h - \widetilde{\psi}_h$, where $\widetilde{\psi}_h$ is a RK method of order $q < r$ embedded in ψ_h.

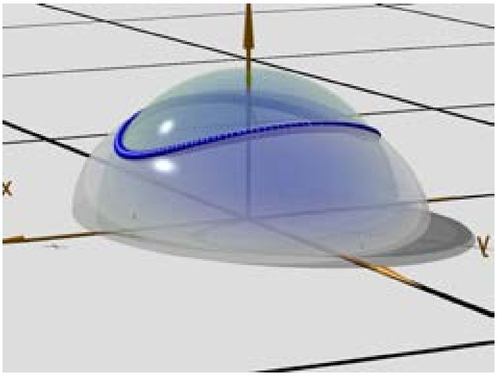


Figure 1: Solution of Euler's equations obtained with the projection dopri54 method

The results of the previous section on the growth of the errors are stated on the assumption that the one-step method ψ_h commutes with respect to differentiations with respect to the initial conditions, i.e. assumption (*iv*) is satisfied. It is known [1] that this relation holds for all the RK methods. We have proved in [3] that it is also satisfied for projection methods.

Next, we present some numerical experiments to corroborate the theoretical results presented in this article. The numerical method considered is the Runge–Kutta embedded pair of order 5(4) constructed by Dormand and Prince (see e.g. [7, p. 178]). We denote this pair "standard dopri54", whereas "projection dopri54" refers to that pair combined with the directional projection technique (see [2]) which makes the resulting method to preserve certain first integrals of the problem. Integrations have been carried out with a local error tolerance of 10^{-6}.

Our first test problem describes the motion of a free rigid body represented by the Euler's equations (see e.g. [6, p. 95]):

$$\begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} 0 & c_3y_3 & -c_2y_2 \\ -c_3y_3 & 0 & c_1y_1 \\ c_2y_2 & -c_1y_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

The vector $y = (y_1, y_2, y_3)^T$ is the angular momentum, and $c_j^{-1} > 0$, $j = 1, 2, 3$, are the principal momenta of inertia. This Poisson differential system has the two first integrals:

$$2H = c_1y_1^2 + c_2y_2^2 + c_3y_3^2,$$
$$L^2 = y_1^2 + y_2^2 + y_3^2,$$

where H and L represent the kinetic energy and the modulus of the angular momentum, respectively.

We have taken $c_1 = 1$, $c_2 = 1 - 0.51/\sqrt{1.51}$, $c_3 = 1 + 1/\sqrt{1.51}$, with initial conditions:

$$y_1(0) = 0, \; y_2(0) = y_3(0) = 1.$$

The solution of this initial value problem is periodic, and its period $T = T(H, L)$ depends only on those two quadratic invariants. In Figure 1 we have plotted the numerical solution of this problem obtained with the projection dopri54 method. It lies on the intersection of the sphere $L^2(y_1, y_2, y_3) = 2$ with the ellipsoid $2H(y_1, y_2, y_3) = c_2 + c_3$.

In Figure 2, the Euclidean norm of the global error against the number of periods is shown in a log-log scale for the Euler's equations. The integration has been carried out up to 8000 periods. The projection dopri54 method has been designed so that it preserves the two invariants of the problem (see [2]) and, in consequence, it preserves its period. As it can be seen, the global error grows linearly with the number of periods for this projection method which is in agreement with Theorem 1. As expected, this growth is quadratic for the standard dopri54. Dashed straight lines with slopes $m = 1$ and $m = 2$ have been drawed in order to show up clearly the type of growth. In Figure 3, the preservation of the invariants is clear for the projected RK method, whereas the error of the invariant grows linearly for the standard one, which agrees with Theorem 2. Here the dashed reference line has slope 1.

The second test problem is the well known planar two body problem, also called Kepler's problem, given by the equations:

$$p_i' = -\frac{q_i}{(q_1^2 + q_2^2)^{3/2}}, \; q_i' = p_i, \; i = 1, 2.$$

This is a Hamiltonian system with Hamiltonian function given by

$$H = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}}.$$

We have considered the initial conditions

$$p_1 = 0, \; p_2 = \sqrt{\frac{1+e}{1-e}}, \; q_1 = 1 - e, \; q_2 = 0,$$

which correspond to a 2π-periodic elliptic orbit with eccentricity e, $0 \leq e < 1$. For these numerical experiments we have taken $e = 0.3$, and we have integrated along 8000 periods.

In Figure 4, the evolution of the global error against the number of periods is shown. In this case, the projection has been made so that the resulting projection method preserves the Hamiltonian H. Therefore, it also preserves the period since the period only depends on the energy H. According to Theorem 1, the growth of the global error must be linear in this case, and it is just what happens for the projection dopri54. Once again, the global error grows quadratically for the standard method. Figure 5 shows up the preservation of the first integral H for the projection method and the linear growth of H with the number of periods for the standard one.

Acknowledgements

This work was supported by D.G.I. project MTM2007-67530-C02-01.

References

[1] Bochev, P. B., and Scovel, C. On quadratic invariants and symplectic structure. *BIT 34* (1994), 337–345.

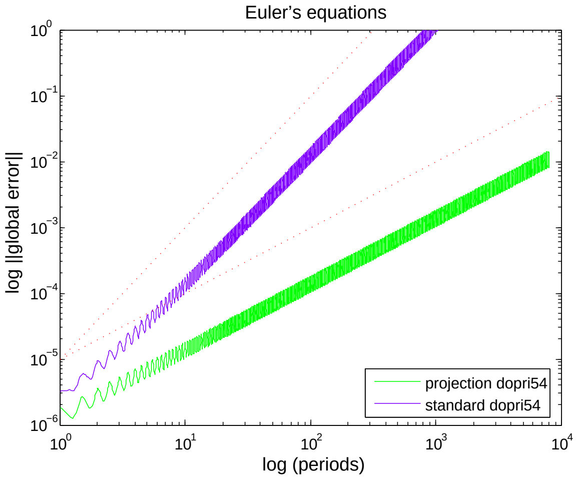


Figure 2: Euler's equations: global error vs. periods, tol=10^{-6}

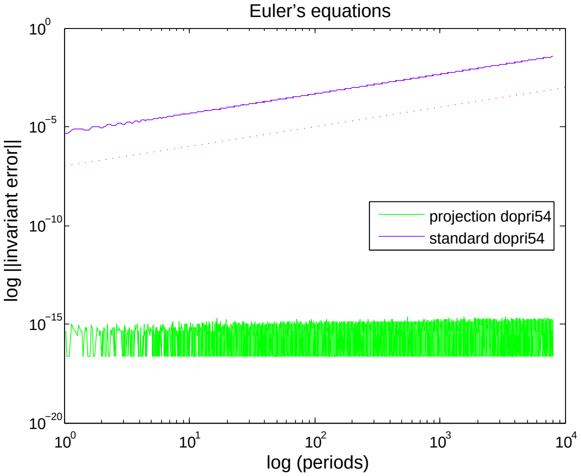


Figure 3: Euler's equations: invariants' error vs. periods, tol=10^{-6}

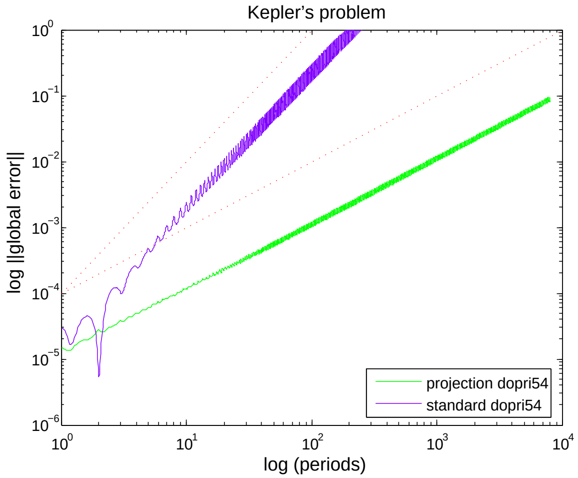


Figure 4: Kepler's problem: global error vs. periods, $e = 0.3$, tol=10^{-6}

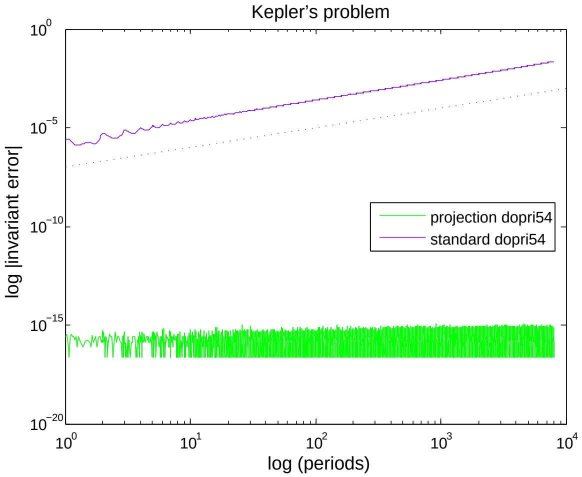


Figure 5: Kepler's problem: error in H vs. periods, $e = 0.3$, tol=10^{-6}

[2] Calvo, M., Hernández-Abreu, D., Montijano, J. I., and Rández, L. On the preservation of invariants by explicit runge-kutta methods. *SIAM J. Sci. Comput. 28* (2006), 868–885.

[3] Calvo, M., Laburta, M. P., Montijano, J. I., and Rández, L. Error growth in the numerical integration of periodic orbits. *Technical Report, Dpto. de Matemática Aplicada, Universidad de Zaragoza* (2008).

[4] Calvo, M. P., and Sanz-Serna, J. M. The development of variable step symplectic integrators with applications to the two-body problem. *SIAM J. Sci. Comput. 14* (1993), 936–952.

[5] Cano, B., and Sanz-Serna, J. M. Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible systems. *SIAM J. Numer. Anal. 34* (1997), 1391–1417.

[6] Hairer, E., Lubich, C., and Wanner, G. *Geometric Numerical Integration: Structure Preserving algorithms for Ordinary Differential Equations.* Springer–Verlag, Berlin, 2002.

[7] Hairer, E., Nørsett, S. P., and Wanner, G. *Solving Ordinary Differential Equations I, Nonstiff Problems.* Springer–Verlag, Berlin, 1993.

M. Calvo, M. P. Laburta, J. I. Montijano and L. Rández
Departamento de Matemática Aplicada
Universidad de Zaragoza
50009-Zaragoza (Spain)
`calvo, laburta, monti` and `randez@unizar.es`

Monografías Matemáticas *García de Galdeano* **35**, 91–98 (2010)

Optimal bases of spaces with trigonometric functions

J. M. Carnicer, E. Mainar and J. M. Peña

Abstract. The normalized B-basis of a space has optimal shape preserving properties. We present a procedure to construct the normalized B-basis. We illustrate this construction in the space $\bar{T}_{1/2}$ generated by 1, t, $\cos t$, $\sin t$, $\cos(t/2)$, $\sin(t/2)$. This space can be used to represent exactly the following remarkable curves: complete cycloidal arcs, cardioids, deltoids and Descartes' trifolium.

Keywords: Normalized B-basis, trigonometric functions, shape preserving representations.

AMS classification: 65D17, 42A10.

§1. Introduction

The Bernstein basis is optimal among all other shape preserving bases of the space of polynomials of degree not greater than n on a given compact interval [1]. Roughly speaking, this means that the curve represented by this basis is closer to its control polygon than with other kinds of representations for polynomial curves. In [2] it was proved that each space of functions admitting shape preserving representations (in the sense of [5]) always has an optimal shape preserving basis called *the normalized B-basis*. For polynomial curves, the shape preserving representations exist on intervals of any length. However, for more general spaces interesting in curve design, it is not clear whether there exist shape preserving representations on a given interval (see [6]). In the last years, a growing interest in the design of curves in spaces mixing algebraic, trigonometric and hyperbolic functions has arisen.

It is desirable to represent motions of objects with its natural velocity, which eliminates the freedom in the parameterization. In particular, in order to obtain uniform circular motions, it is necessary to represent a circle with its arc length parameterization. It is also convenient that all types of curves which have to be used in the design process can be obtained with the same kind of representation. This may imply to use a representation of curves in a computer graphics system working simultaneously with algebraic and transcendent curves.

In [4] we have found all six dimensional spaces invariant under translations and reflections containing the first degree polynomials and the trigonometric functions $\cos t$, $\sin t$ and admitting shape preserving representations on the interval $[0, 2\pi]$ (this implies that a complete circular arc can be represented using a single control polygon). Among these spaces, we find spaces mixing trigonometric functions with two angular frequencies

$$\bar{T}_w := \langle 1, t, \cos t, \sin t, \cos(wt), \sin(wt)\rangle, \quad 0 < w < 1.$$

We have chosen the space $\bar{T}_{1/2}$ to show how to obtain the optimal bases. This space allows us to represent exactly not only circles but also complete cycloidal arcs, cardioids, deltoids and Descartes' trifolium among other remarkable curves.

Section 2 presents the procedure to obtain the normalized B-basis of a space. Section 3 is devoted to the construction of the normalized B-basis of the space $\bar{T}_{1/2}$ and the obtention of control polygons for the representation of remarkable curves.

§2. Construction of the optimal shape preserving basis

Let us recall that an *extended Chebyshev space* of functions F defined on an interval I is a space such that each nonzero function of F has at most $\dim F - 1$ zeros (counting multiplicities) in I. In order to check the existence of normalized B-bases, we shall use Theorem 4.1 of [3] restated below.

Theorem 1. *Let F be an $(n+1)$-dimensional subspace of $C^n[a,b]$ such that $1 \in F$. Then F is an extended Chebyshev space with a normalized B-basis on $[a,b]$ if and only if the space of the derivatives*

$$F' := \{f' \mid f \in F\}$$

is an extended Chebyshev space.

Assuming that we have shown the existence of a normalized B-basis, we may proceed to its construction. First we construct a B-basis following the method suggested in Remark 2.3 and Theorem 2.4 of [3] and then, we shall normalize it following Remark 4.1 of [3]. Let us describe the steps of this construction.

Step 1. We start with a basis such $(u_0, \dots, u_n)$ such that the wronskian matrix at the left end of the interval

$$W(u_0, \dots, u_n)(a) = \begin{pmatrix} u_0(a) & u_1(a) & \cdots & u_n(a) \\ u_0'(a) & u_1'(a) & \cdots & u_n'(a) \\ \vdots & \vdots & \ddots & \vdots \\ u_0^{(n)}(a) & u_1^{(n)}(a) & \cdots & u_n^{(n)}(a) \end{pmatrix}$$

is a lower triangular matrix with nonzero diagonal entries.

Step 2. We compute $W(u_n, \dots, u_0)(b)$, the wronskian matrix of the basis $(u_n, \dots, u_0)$, where the ordering of the functions has been reversed, at the right end of the interval and obtain its LU factorization with L a lower triangular matrix with unit diagonal and U a nonsingular upper triangular matrix.

Step 3. We construct the basis $(b_0, \dots, b_n)$ defined by

$$(b_n, -b_{n-1}, \dots, (-1)^n b_0) := (u_n, u_{n-1}, \dots, u_0)U^{-1}.$$

This basis is a Bernstein-like basis in the sense that $W(b_0, \dots, b_n)(a)$ and $W(b_n, \dots, b_0)(b)$ are lower triangular matrices.

Step 4. In order to normalize the obtained basis, we solve the linear system

$$L(c_n, c_{n-1}, \dots, c_0)^T = (1, 0, \dots, 0)^T,$$

and then the normalized B-basis is

$$(B_0, \dots, B_n) := (c_0 b_0, \dots, c_n b_n).$$

Remark 1. Since the space is invariant under reflections we have for the functions of the normalized B-basis

$$B_i(t) = B_{n-i}(a+b-t), \quad t \in [a,b], \quad i = 0, \ldots, n.$$

So we only need to compute half of the basis functions B_i, $0 \le i \le n/2$.

§3. Designing with two angular frequencies

This section is devoted to the design of curves in the space

$$\bar{T}_{1/2} = \langle 1, t, \cos t, \sin t, \cos(t/2), \sin(t/2) \rangle,$$

on the interval $t \in [0, 2\pi]$. First we shall find the optimal basis (normalized B-basis) of $\bar{T}_{1/2}$ on $[0, 2\pi]$ and later we shall use it for the design of some remarkable curves.

It is a well-known fact that the space $\langle 1, \cos s, \sin s, \cos(2s), \sin(2s) \rangle$ of trigonometric polynomials of degree 2 is an extended Chebyshev space on $[0, 2\pi)$, that is, any nonzero function of the space has at most $\dim \bar{T}'_{1/2} - 1 = 4$ zeros (counting multiplicities).

By Theorem 1, there exists a normalized B-basis on the space $\bar{T}_{1/2}$ on $[0, 2\pi]$ if and only if the space of the derivatives

$$\bar{T}'_{1/2} = \langle 1, \cos t, \sin t, \cos(t/2), \sin(t/2) \rangle$$

is an extended Chebyshev space. Taking $s = t/2$, the space is transformed into the space of trigonometric polynomials of degree 2 on the interval $[0, \pi]$. As mentioned above, this space is extended Chebyshev on each interval contained in $[0, 2\pi)$.

Once we have shown the existence of a normalized B-basis, we proceed to its construction following the steps described in Section 2.

Step 1. We start with the basis $(u_0, \ldots, u_5)$ given by

$$\left(1, t, 1 - \cos t, t - \sin t, 4 - \frac{16}{3}\cos(t/2) + \frac{4}{3}\cos t, 3t - 8\sin(t/2) + \sin t\right),$$

$t \in [0, 2\pi]$, whose wronskian matrix at $t = 0$

$$W(u_0, \ldots, u_5)(0) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 3/4 \end{pmatrix}$$

is lower triangular with positive diagonal entries.

Let us observe that $u_4(t)$ can be factorized as $u_4(t) = \frac{8}{3}\Big(1 - \cos(t/2)\Big)^2$.

Step 2. We evaluate the wronskian matrix at $t = 2\pi$

$$W(u_0,\dots,u_5)(2\pi) = \begin{pmatrix} 1 & 2\pi & 0 & 2\pi & 32/3 & 6\pi \\ 0 & 1 & 0 & 0 & 0 & 8 \\ 0 & 0 & 1 & 0 & -8/3 & 0 \\ 0 & 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & -1 & 0 & 5/3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 5/4 \end{pmatrix},$$

reverse the columns and compute the factorization, $W(u_5,\dots,u_0)(2\pi) = LU$ obtaining

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 4\pi^{-1}/3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 3\pi/16 & 1 & 0 & 0 & 0 \\ -\pi^{-1}/3 & -1/4 & 2\pi^{-1} & 1 & 0 & 0 \\ 0 & -15\pi/128 & -5/8 & 3\pi/16 & 1 & 0 \\ 5\pi^{-1}/24 & 5/32 & -2\pi^{-1} & -1 & 4\pi^{-1}/3 & 1 \end{pmatrix},$$

$$U = \begin{pmatrix} 6\pi & 32/3 & 2\pi & 0 & 2\pi & 1 \\ 0 & -128\pi^{-1}/9 & -8/3 & 0 & -5/3 & -4\pi^{-1}/3 \\ 0 & 0 & \pi/2 & 1 & 5\pi/16 & 1/4 \\ 0 & 0 & 0 & -2\pi^{-1} & -3/8 & -\pi^{-1}/2 \\ 0 & 0 & 0 & 0 & 9\pi/128 & 3/32 \\ 0 & 0 & 0 & 0 & 0 & -\pi^{-1}/8 \end{pmatrix}.$$

Step 3. In order to construct the basis $(b_0,\dots,b_5)$ defined by

$$(b_5, -b_4, b_3, -b_2, b_1, -b_0) = (u_5, u_4, u_3, u_2, u_1, u_0)U^{-1},$$

we compute

$$U^{-1} = \begin{pmatrix} \pi^{-1}/6 & 1/8 & 0 & 0 & -16\pi^{-1}/9 & -4/3 \\ 0 & -9\pi/128 & -3/8 & -3\pi/16 & -1 & 0 \\ 0 & 0 & 2\pi^{-1} & 1 & -32\pi^{-1}/9 & -8/3 \\ 0 & 0 & 0 & -\pi/2 & -8/3 & 0 \\ 0 & 0 & 0 & 0 & 128\pi^{-1}/9 & 32/3 \\ 0 & 0 & 0 & 0 & 0 & -8\pi \end{pmatrix}.$$

Then we have

$$\begin{aligned} b_5(t) &:= \frac{1}{6\pi}(3t - 8\sin(t/2) + \sin t), \\ b_4(t) &:= \frac{1}{8}(3t - 8\sin(t/2) + \sin t) - \frac{3\pi}{16}(1 - \cos(t/2))^2, \\ b_3(t) &:= \frac{2}{\pi}(t - \sin t) - (1 - \cos(t/2))^2. \end{aligned}$$

Step 4. Solving the system

$$L(c_5, c_4, c_3, c_2, c_1, c_0)^T = (1, 0, 0, 0, 0, 0)^T,$$

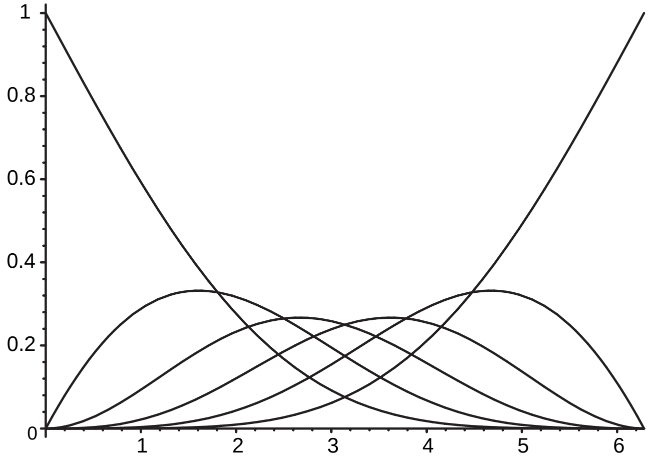


Figure 1: Normalized B-basis of $\bar{T}_{1/2}$ on $[0, 2\pi]$.

we obtain $c_5 = 1$, $c_4 = -4\pi^{-1}/3$, $c_3 = 1/4$ and then the normalized B-basis $(B_0, \dots, B_5)$ is given by

$$\begin{aligned} B_5(t) &:= \frac{1}{6\pi}\,(3t - 8\sin(t/2) + \sin t)\,, \\ B_4(t) &:= \frac{-1}{6\pi}\,(3t - 8\sin(t/2) + \sin t) + \frac{1}{4}\,(1 - \cos(t/2))^2\,, \\ B_3(t) &:= \frac{1}{2\pi}\,(t - \sin t) - \frac{1}{4}\,(1 - \cos(t/2))^2\,, \end{aligned}$$

and by Remark 1, we obtain the remaining basis functions

$$B_2(t) := B_4(2\pi - t),\ B_1(t) := B_4(2\pi - t),\ B_0(t) := B_5(2\pi - t).$$

Figure 1 shows the graphs of the functions of the normalized B-basis of $\bar{T}_{1/2}$ on $[0, 2\pi]$.

Now we are going to obtain control polygons of different curves. For this purpose, we need the coefficients of some functions with respect to the normalized B-basis. The coefficients of the function t, shown in Table 1, are called the Greville abscissae and are used for obtaining the control polygon

$$\begin{pmatrix}0\\c_0\end{pmatrix}\begin{pmatrix}3\pi/4\\c_1\end{pmatrix}\begin{pmatrix}3\pi/4\\c_2\end{pmatrix}\begin{pmatrix}5\pi/4\\c_3\end{pmatrix}\begin{pmatrix}5\pi/4\\c_4\end{pmatrix}\begin{pmatrix}2\pi\\c_5\end{pmatrix}$$

of the graph of $f(t) = \sum_{i=0}^{5} c_i B_i(t)$.

Table 1 contains the coefficients of the usual trigonometric functions $\cos t, \sin t$, the cycloidal cosine, $1 - \cos t$, and the cycloidal sine, $t - \sin t$, with respect to the normalized B-basis.

Figure 2 (left) shows a circle $(\sin t, 1 - \cos t)$, $t \in [0, 2\pi]$, and its control polygon

$$\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}3\pi/4\\0\end{pmatrix}\begin{pmatrix}3\pi/4\\4\end{pmatrix}\begin{pmatrix}-3\pi/4\\4\end{pmatrix}\begin{pmatrix}-3\pi/4\\0\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}.$$

Function	c_0	c_1	c_2	c_3	c_4	c_5
1	1	1	1	1	1	1
t	0	$3\pi/4$	$3\pi/4$	$5\pi/4$	$5\pi/4$	2π
$\cos t$	1	1	-3	-3	1	1
$\sin t$	0	$3\pi/4$	$3\pi/4$	$-3\pi/4$	$-3\pi/4$	0
$1-\cos t$	0	0	4	4	0	0
$t-\sin t$	0	0	0	2π	2π	2π
$\sin(t/2)$	0	$3\pi/8$	$3\pi/8$	$3\pi/8$	$3\pi/8$	0
$\cos(t/2)$	1	1	0	0	-1	-1
$(1+\cos t)/2$	1	1	-1	-1	1	1

Table 1: Coefficients of relevant functions in $\bar{T}_{1/2}$

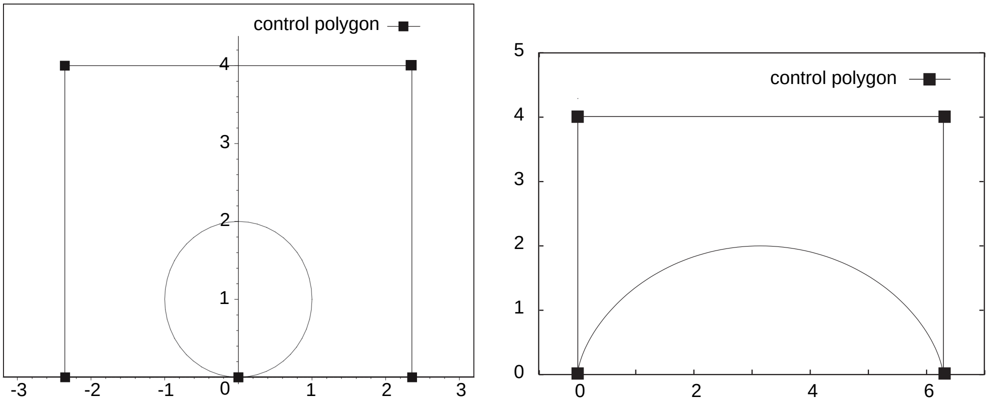


Figure 2: Control polygon of a circle (left) and a complete cycloid arc (right) in $\bar{T}_{1/2}$

Figure 2 (right) shows the cycloid $(t-\sin t, 1-\cos t)$, $t \in [0, 2\pi]$, and its control polygon

$$\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0\\4\end{pmatrix}\begin{pmatrix}2\pi\\4\end{pmatrix}\begin{pmatrix}2\pi\\0\end{pmatrix}\begin{pmatrix}2\pi\\0\end{pmatrix}.$$

Quadratic curves can also be represented in this space. In fact, the parabola $y = x^2$, $x \in [-1, 1]$, can be represented in this space by the parametric curve $(\cos(t/2), (1+\cos t)/2)$, $t \in [0, 2\pi]$, which, in view of Table 1, has the following control polygon

$$\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix}.$$

Other remarkable parametric curves which can be represented in this space are

$$\begin{aligned}&\text{Cardioid:} && (a(2\cos(t/2)+1+\cos t), a(2\sin(t/2)+\sin t)),\\ &\text{Deltoid:} && (a(2\cos(t/2)+\cos t), a(2\sin(t/2)-\sin t)),\\ &\text{Trifolium:} && (a\cos(3t/4)\cos(t/4), a\cos(3t/4)\sin(t/4)).\end{aligned}$$

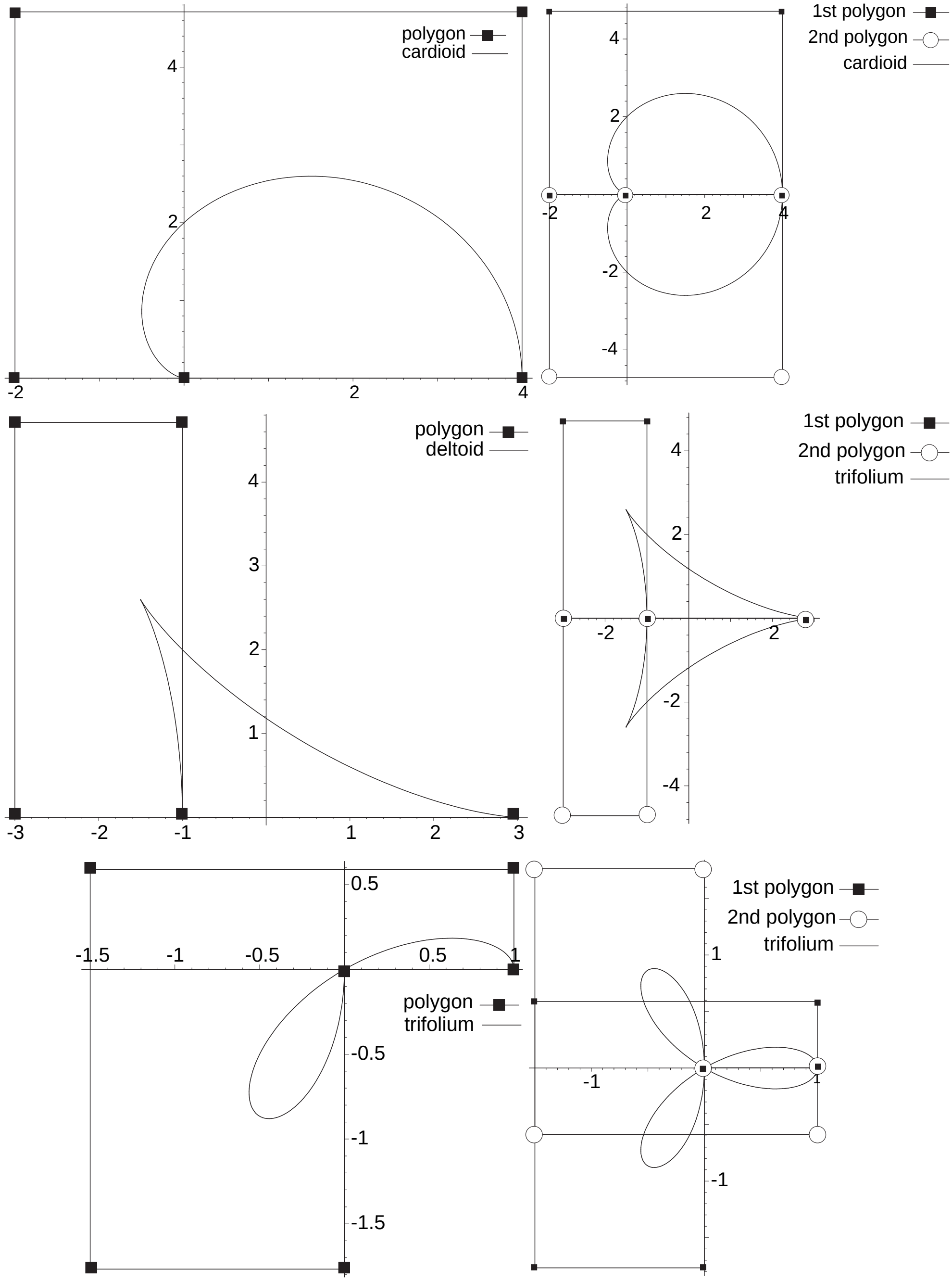


Figure 3: Cardioid, Deltoid and Trifolium

Let us observe that these curves are defined on the parameter interval $[0, 4\pi]$. The representation of $(x(t), y(t))$, $t \in [0, 4\pi]$, $x, y \in \bar{T}_{1/2}$, requires two control polygons. The first one is needed to represent the curve $(x(t), y(t))$, $t \in [0, 2\pi]$, and the second one to represent $(x(t + 2\pi), y(t + 2\pi))$, $t \in [0, 2\pi]$. Since the space $\bar{T}_{1/2}$ is invariant under translations, the functions $x(t + 2\pi), y(t + 2\pi)$, $t \in [0, 2\pi]$, also belong to $\bar{T}_{1/2}$. In view of the symmetry of these curves, $x(4\pi - t) = x(t)$, $y(4\pi - t) = -y(t)$, we have that the control polygon $\tilde{P}_0 \cdots \tilde{P}_5$, $\tilde{P}_i = (\tilde{x}_i, \tilde{y}_i)$, $i = 0, \ldots, 5$, of the curve $x(t + 2\pi), y(t + 2\pi)$, $t \in [0, 2\pi]$, can be expressed in terms of the control polygon $P_0 \cdots P_5$, $P_i = (x_i, y_i)$, $i = 0, \ldots, 5$, of $x(t), y(t)$, $t \in [0, 2\pi]$, by

$$\tilde{x}_i = x_{5-i}, \quad \tilde{y}_i = -y_{5-i}, \quad i = 0, \ldots, 5.$$

Figure 3 shows the cardiod (top), deltoid (middle) and Descartes' trifolium (bottom), corresponding to the parameter value $a = 1$ and their corresponding couple of control polygons with respect to the normalized B-basis of $\bar{T}_{1/2}$.

Acknowledgements

Partially supported by the Spanish Research Grant MTM2006-03388 and by Gobierno de Aragón

References

[1] Carnicer, J. M., Mainar, E., and Peña, J. M. Shape preserving representations and optimality of the Bernstein basis. *Adv. Comput. Math. 1* (1993), 173–196.

[2] Carnicer, J. M., Mainar, E., and Peña, J. M. Totally positive bases for shape preserving curve design and optimality of B-splines. *Comput. Aided Geom. Design 11* (1994), 635–656.

[3] Carnicer, J. M., Mainar, E., and Peña, J. M. Critical length for design purposes and extended chebyshev spaces. *Constr. Approx. 20* (2004), 55–71.

[4] Carnicer, J. M., Mainar, E., and Peña, J. M. Shape preservation regions for six-dimensional spaces. *Adv. Comput. Math. 26* (2007), 121–136.

[5] Goodman, T. N. T. Shape preserving representations. In *Mathematical Methods in CAGD* (Boston, 1989), Academic Press, pp. 333–351.

[6] Peña, J. M. Shape preserving representations for trigonometric polynomial curves. *Comput. Aided Geom. Design 14* (1997), 5–11.

J. M. Carnicer and J. M. Peña
Departamento de Matemática Aplicada/IUMA
Universidad de Zaragoza
50009 Zaragoza, Spain
carnicer@unizar.es, jmpena@unizar.es

E. Mainar
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria
39005 Santander, Spain
mainare@unican.es

Monografías Matemáticas *García de Galdeano* **35**, 99–106 (2010)

Some methods based on cubic splines to solve a reaction-diffusion problem: uniform convergence for global solution and normalized flux

Carmelo Clavero

Abstract. In this paper we combine the classical cubic spline with two different finite difference schemes to find an approximation to the global solution and the global normalized flux of a singularly perturbed boundary value problem of reaction-diffusion type. We prove that if the schemes are constructed on a slight modification of a piecewise uniform Shishkin mesh, then the numerical solutions are uniformly convergent for both the global solution and the global normalized flux. We give theoretical error bounds showing the order of uniform convergence of the methods and we display some numerical examples corroborating in practice these orders of convergence.

Keywords: Reaction-diffusion problems, modified Shishkin mesh, cubic spline, global solution, global normalized flux.

AMS classification: 65L10, 65L12, 65L20.

§1. Introduction

We consider the singularly perturbed reaction–diffusion two-point boundary-value problem

$$\begin{aligned} &Lu(x) \equiv -\varepsilon u''(x) + b(x)u(x) = f(x), \quad x \in D = (0,1),\\ &u(0) = A, \quad u(1) = B, \end{aligned} \tag{1}$$

where $\varepsilon > 0$ is a small parameter and b, f are sufficiently smooth functions such that $b(x) \geq \beta > 0$ on $\overline{D} = [0,1]$. Under these assumptions it is well known (see [4]) that (1) has an unique solution satisfying

$$|u^{(k)}(x)| \leq C\left(1 + \varepsilon^{-k/2} e(x,x,\beta,\varepsilon)\right), \quad 0 \leq k \leq j+1. \tag{2}$$

where $e(\xi_1,\xi_2,\beta,\varepsilon) = \exp(-\sqrt{\beta}\xi_1/\sqrt{\varepsilon}) + \exp(-\sqrt{\beta}(1-\xi_2)/\sqrt{\varepsilon})$, and the value of j depends on the smoothness of data b and f. Bounds (2) give the asymptotic behavior of the exact solution of (1) with respect to the diffusion parameter ε, showing the presence of boundary layers at both end points on $\overline{D}$.

To approximate the solution of (1) it is essential to devise efficient methods, giving good approximations for any value of the diffusion parameter ε, i.e., uniformly convergent methods. Many numerical methods having this property are developed in last years (see for instance [3, 5, 7, 8]), showing in some cases uniform convergence only at the nodal points and in other cases also uniform convergence for the global solution on $\overline{D}$. In this paper we extend the results of [6] by modifying the original piecewise uniform Shishkin mesh. We construct some methods giving good approximations for the global solution and the global normalized flux, by using a classical cubic spline based on the numerical solutions at mesh points.

The paper is organized as follows: in Section 2, we present the numerical methods used to solve (1) and we define the numerical cubic spline associated to the numerical solutions at mesh points. In Section 3 we prove the uniform convergence of the cubic spline in the approximation of both the global solution and the global normalized flux. Finally, in Section 4 we show some results obtained by the numerical methods in a particular example, corroborating in practice the theoretical results. Henceforth, C denotes any positive constant independent of the diffusion parameter ε and the discretization parameter N. C can take different values at different places.

§2. The finite difference schemes

The first step to define the finite difference scheme is to construct the mesh. Then, following [4], the domain $\overline{D}$ is divided into three subintervals as $\overline{D} = [0,\sigma) \cup [\sigma, 1-\sigma] \cup (1-\sigma, 1]$, where σ is the transition parameter given by

$$\sigma = \min\left\{1/4, \sigma_0 \sqrt{\varepsilon} \ln N\right\}, \tag{3}$$

and σ_0 is a positive constant. On the subintervals $[0,\sigma]$ and $[1-\sigma, 1]$ an uniform mesh with $N/4$ mesh intervals are placed, while $[\sigma, 1-\sigma]$ has an uniform mesh with $N/2$ mesh intervals. Obviously the mesh is uniform when $\sigma = 1/4$. The mesh size in $[\sigma, 1-\sigma]$ is $H = 2(1-2\sigma)/N$, and in $[0,\sigma] \bigcup [1-\sigma, 1]$ it is $h = 4\sigma/N$. Let $\overline{D}^N \equiv \{x_i : 0 = x_0 < \cdots < x_N = 1\}$ be the mesh and we denote by $h_{i+1} = x_{i+1} - x_i,\ i = 0, 1, \ldots, N-1$.

For the exact values $u(x_i)$ $i = 0, \ldots, N$, of the function u at the nodal points, it s well known that there exists an interpolating cubic spline $s(x)$ given by

$$\begin{aligned} s(x) = {} & \frac{(x_{i+1}-x)^3}{6h_{i+1}} M_i + \frac{(x-x_i)^3}{6h_{i+1}} M_{i+1} + \left(u_i - \frac{h_{i+1}^2}{6} M_i\right)\left(\frac{x_{i+1}-x}{h_{i+1}}\right) + \\ & + \left(u_{i+1} - \frac{h_{i+1}^2}{6} M_{i+1}\right)\left(\frac{x-x_i}{h_{i+1}}\right),\ x_i \le x \le x_{i+1},\ i = 0, \ldots, N-1, \end{aligned} \tag{4}$$

where $u_i = u(x_i)$, $M_i = u''(x_i)$, $i = 0, \ldots, N$. From this cubic spline the approximation to the global normalized flux is obtained by $\sqrt{\varepsilon} s'(x)$.

To calculate a numerical cubic spline, we can use the discrete solution $U_i,\ i = 0, \ldots N$, given by a finite difference scheme at mesh points, and then, defining $\overline{M}_i = (b_i U_i - f_i)/\varepsilon$, $i = 0, \ldots N$, the numerical cubic spline is defined as

$$\begin{aligned} S(x) = {} & \frac{(x_{i+1}-x)^3}{6h_{i+1}} \overline{M}_i + \frac{(x-x_i)^3}{6h_{i+1}} \overline{M}_{i+1} + \left(U_i - \frac{h_{i+1}^2}{6} \overline{M}_i\right)\left(\frac{x_{i+1}-x}{h_{i+1}}\right) + \\ & + \left(U_{i+1} - \frac{h_{i+1}^2}{6} \overline{M}_{i+1}\right)\left(\frac{x-x_i}{h_{i+1}}\right),\ x_i \le x \le x_{i+1},\ i = 0, \ldots, N-1. \end{aligned} \tag{5}$$

This spline gives an approximation to the exact solution of the boundary value problem (1) at the whole domain $\overline{D}$ and also an approximation to the normalized flux by using $\sqrt{\varepsilon}S'(x)$.

To obtain the uniform convergence for the global solution and the global normalized flux, it will be necessary to use a slight modification of the original Shsihkin mesh. Following the original idea of Surla (see [8]), we define a new parameter $\overline{H} = \sqrt{\varepsilon/\beta}N\ln N$ and we construct a modified Shishkin mesh as follows. If $H/2 \leq \overline{H}$ the mesh is the same that the original Shishkin mesh; on the other hand, when $H/2 > \overline{H}$, we introduce two new points, $\overline{x}_{N/4} = x_{N/4} + \overline{H}$ and $\overline{x}_{3N/4} = x_{3N/4} - \overline{H}$. So, in this case the number of mesh points is $N_1 = N + 2$, and they are given by

$$x_i = \begin{cases} ih, \quad i = 0, 1, \dots, N/4, \\ \sigma + \overline{H}, \quad i = N/4 + 1, \\ \sigma + (i - 1 - N/4)H, \quad i = N/4 + 2, \dots, 3N/4, \\ 1 - \sigma - \overline{H}, \quad i = 3N/4 + 1, \\ 1 - \sigma, \quad i = 3N/4 + 2, \\ 1 - \sigma + (i - 3N/4 + 2)h, \quad i = 3N/4 + 3, \dots, N_1 \end{cases} \tag{6}$$

On this modified Shishkin mesh we consider two different finite difference schemes. The first one, constructed in [6], is a hybrid scheme defined as

$$\begin{aligned} &L^N U_i^N \equiv r_i^- U_{i-1}^N + r_i^c U_i^N + r_i^+ U_{i+1}^N = q_i^- f_{i-1} + q_i^c f_i + q_i^+ f_{i+1}, 1 \leq i \leq N-1, \\ &U_0^N = A, \quad U_N^N = B, \end{aligned} \tag{7}$$

where for indices $i = 1, \dots, N/4 - 1$ and also $3N/4 + 1, \dots, N - 1$, the coefficients of the scheme are given by

$$\begin{aligned} r_i^- &= \frac{-3\varepsilon}{h_i(h_i + h_{i+1})} + \frac{h_i}{2(h_i + h_{i+1})} b_{i-1}, \quad r_i^c = \frac{3\varepsilon}{h_i h_{i+1}} + b_i, \\ r_i^+ &= \frac{-3\varepsilon}{h_{i+1}(h_i + h_{i+1})} + \frac{h_{i+1}}{2(h_i + h_{i+1})} b_{i+1}, \\ q_i^- &= \frac{h_i}{2(h_i + h_{i+1})}, \quad q_i^c = 1, \quad q_i^+ = \frac{h_{i+1}}{2(h_i + h_{i+1})}, \end{aligned} \tag{8}$$

and for indices $i = N/4, \dots, 3N/4$, the coefficients are now given by

$$\begin{aligned} &r_i^- = \frac{-2\varepsilon}{h_i(h_i + h_{i+1})}, \quad r_i^c = \frac{2\varepsilon}{h_i h_{i+1}} + b_i, \quad r_i^+ = \frac{-2\varepsilon}{h_{i+1}(h_i + h_{i+1})}, \\ &q_i^- = 0, \quad q_i^c = 1, \quad q_i^+ = 0. \end{aligned} \tag{9}$$

The second method is the HOC (High Order Compact) scheme constructed in [3], which is defined as

$$\begin{gathered} L_{\varepsilon,N}[U_i] \equiv r_i^- U_{i-1} + r_i^c U_i + r_i^+ U_{i+1} = Q_N(f_i), \ 1 \leq i \leq N-1, \\ U_0 = A, \quad U_N = B, \end{gathered} \tag{10}$$

where the coefficients are given by

$$\begin{aligned}
r_i^- &= \frac{-2\varepsilon}{(h_i+h_{i+1})h_i} - \delta_{i,N/4}\frac{(h_{i+1}-h_i)b_i}{3h_i} - \frac{(h_{i+1}^3+h_i^3)b_i'}{6(h_i+h_{i+1})h_i}\,\mathrm{sgn}\, b_i',\\
r_i^+ &= \frac{-2\varepsilon}{(h_i+h_{i+1})h_{i+1}} + \delta_{i,3N/4}\frac{(h_{i+1}-h_i)b_i}{3h_{i+1}} + \frac{(h_{i+1}^3+h_i^3)b_i'}{6(h_i+h_{i+1})h_i}(1-\ \mathrm{sgn}\, b_i'),\\
r_i^c &= -r_i^- - r_i^+ + Q_N^2(b_i),
\end{aligned} \tag{11}$$

whit sgn $z_i = 1$, if $z_i \geq 0$ and sgn $z_i = 0$, if $z_i < 0$, $\delta_{il} = 1$ if $i = l$, $\delta_{il} = 0$ if $i \neq l$ and

$$\begin{aligned}
Q_N(z_i) \equiv z_i &+ \frac{h_{i+1}-h_i}{3}\left(z_i' + \frac{b_i z_i}{2\varepsilon}\left(\delta_{i,N/4}h_i - \delta_{i,3N/4}h_{i+1}\right)\right) +\\
&+ \frac{h_{i+1}^3+h_i^3}{12(h_i+h_{i+1})}\left(z_i'' + \frac{b_i z_i}{\varepsilon} + \frac{b_i' z_i}{\varepsilon}(h_i \mathrm{sgn}\, b_i' - (1-\mathrm{sgn}\, b_i')h_{i+1})\right).
\end{aligned} \tag{12}$$

§3. Uniform convergence for the global solution and the normalized flux

In this section we give the main results showing the uniform convergence for the global solution and for the global normalized flux, using the cubic spline together with the two finite difference schemes previously defined.

Theorem 1. *Let $u(x)$ be the solution of* (1) *and $S(x)$ be the numerical spline given in* (5)*, based on the solution of the finite difference scheme* (7)–(9) *constructed on the modified Shishkin mesh* (6)*. Then, the error satisfies*

$$|S(x) - u(x)| \leq \left(N^{-2}\ln^2 N + N^{3-\sqrt{\beta}\sigma_0}\ln^3 N\right), \quad \forall x \in \overline{D}. \tag{13}$$

Proof. We only give the main ideas of the proof; for full details see [2]. Let $x_i \leq x \leq x_{i+1}$, $i = 0, \ldots, N-1$, be; then, using (4), (5), Taylor expansions and the bounds (2) for the derivatives of the exact solution u, we can prove that

$$\begin{aligned}
&|s(x) - u(x)| \leq C h_{i+1}^3\left(1 + \varepsilon^{-3/2}e(x_i, x_i, \beta, \varepsilon)\right), \quad \text{if } x \leq 1/2,\\
&|s(x) - u(x)| \leq C h_{i+1}^3\left(1 + \varepsilon^{-3/2}e(x_{i+1}, x_{i+1}, \beta, \varepsilon)\right), \quad \text{if } x \geq 1/2,\\
&|s(x) - S(x)| \leq C\left(1 + b^* h_{i+1}^2/\varepsilon\right)\max\{|u_i - U_i|, |u_{i+1} - U_{i+1}|\},
\end{aligned} \tag{14}$$

where $b^* = \max\limits_{x \in D} b(x)$. Then, if $\sigma = 1/4$ and $\varepsilon^{-1/2} \leq C \ln N$, it is straightforward to obtain that

$$|S(x) - u(x)| \leq C\left(N^{-3}\ln^3 N + N^{-\sqrt{\beta}\sigma_0}\right). \tag{15}$$

On the other hand, when $1/4 > \sigma_0\sqrt{\varepsilon}\ln N$, we distinguish several cases depending on the location of the mesh point x_i, concretely when x_i is inside the boundary layer, outside the layer or x_i is one of the transition points σ or $1-\sigma$. From the uniform stability of the hybrid scheme, easily we have $|s(x) - S(x)| \leq C\left(1 + b^* h_{i+1}^2/\varepsilon\right)|\tau_i|$, where the local error at x_i satisfies (see [6])

$\tau_i = (\varepsilon/H^2)\,(R_3(x_i, x_{i+1}, u) + R_3(x_i, x_{i-1}, u))$, where $R_n(a, p, g) = (1/n!)\displaystyle\int_a^p (p-\xi)g^{(n+1)}(\xi)d\xi$ denotes the remainder of the Taylor expansion. Using the integral form for the remainder, integrating by parts and taking into account that $e(x_j, x_j, \beta, \varepsilon) \le N^{-\sqrt{\beta}\sigma_0}$, $j = i-1, i, i+1$, it is possible to prove the required result. □

Theorem 2. *Let* $\sqrt{\varepsilon}u'(x)$ *be the normalized flux of* (1) *and* $\sqrt{\varepsilon}S'(x)$ *be the normalized flux obtained from the cubic spline based on the numerical solution of the finite difference scheme* (7)–(9) *constructed on the modified Shishkin mesh* (6)*. Then, for any* $x \in \overline{D}$*, it holds*

$$\sqrt{\varepsilon}\left|S'(x) - u'(x)\right| \le \begin{cases} C(N^{-2}\ln^3 N + N^{3-\sqrt{\beta}\sigma_0}\ln^3 N), & \text{if } N^{-1} > \sqrt{\varepsilon}, \\ C(N^{-1}\sqrt{\varepsilon}\ln^2 N + N^{1-\sqrt{\beta}\sigma_0}), & \text{if } N^{-1} \le \sqrt{\varepsilon}. \end{cases} \tag{16}$$

Proof. The proof follows similar ideas to these ones of Theorem 1. Again we take $x_i \le x \le x_{i+1}$, $i = 0, \ldots, N-1$. Now it is possible to obtain that

$$\begin{aligned} &\sqrt{\varepsilon}\left|s'(x) - u'(x)\right| \le C\sqrt{\varepsilon}h_{i+1}^3\left(1 + \varepsilon^{-2}e(x_i, x_i, \beta, \varepsilon)\right), \quad \text{if } x \le 1/2, \\ &\sqrt{\varepsilon}\left|s'(x) - u'(x)\right| \le C\sqrt{\varepsilon}h_{i+1}^3\left(1 + \varepsilon^{-2}e(x_{i+1}, x_{i+1}, \beta, \varepsilon)\right), \quad \text{if } x \ge 1/2, \\ &\sqrt{\varepsilon}\left|s'(x) - S'(x)\right| \le C\left(\sqrt{\varepsilon}/h_{i+1} + b^* h_{i+1}/\sqrt{\varepsilon}\right)\max\{|u_i - U_i|, |u_{i+1} - U_{i+1}|\}. \end{aligned} \tag{17}$$

Then, using Taylor expansions with the remainder in integral form and distinguishing the cases when the mesh is or non uniform and depending on the location of the mesh point in the domain (inside the layer, outside the layer or the transition points), it is not difficult to prove the required result. □

Remark 1. From Theorems 1 and 2 we see that if $\sqrt{\beta}\sigma_0 \ge 5$, then the global solution has order of uniform convergence $O(N^{-2}\ln^3 N)$ and the global normalized flux has almost second order of uniform convergence except for $N^{-1} \le \sqrt{\varepsilon}$, which is less interesting in practice. Our computational results in the next section show that even when $N^{-1} \le \sqrt{\varepsilon}$ the results show the same orders of convergence than for $N^{-1} > \sqrt{\varepsilon}$.

The two following theorems prove the uniform convergence for the global solution and for the global normalized flux, when the HOC scheme is used. Their proof is similar to this one of the two previous theorems for the hybrid scheme; full details can be found in [1].

Theorem 3. *Let* u *the solution of* (1) *and* S *the numerical spline given in* (5)*, based on the solution of the finite difference scheme* (10)–(12) *constructed on the modified Shishkin mesh* (6)*. Then, the error satisfies*

$$|S(x) - u(x)| \le \left(N^{-4}\ln^4 N + N^{4-\sqrt{\beta}\sigma_0}\ln^4 N\right), \quad \forall x \in \overline{D}. \tag{18}$$

Theorem 4. *Let* $\sqrt{\varepsilon}u'$ *be the normalized flux of* (1) *and* $\sqrt{\varepsilon}S'$ *be the normalized flux obtained from the cubic spline approximations, based on the numerical solution of the finite difference scheme* (10)–(12) *constructed on the modified Shishkin mesh* (6)*. Then, for any* $x \in \overline{D}$*, it holds*

$$\sqrt{\varepsilon}\left|S'(x) - u'(x))\right| \le \begin{cases} C(N^{-4}\ln^4 N + N^{4-\sqrt{\beta}\sigma_0}), & \text{if } x = (x_i + x_{i+1})/2,\ N^{-1} > \sqrt{\varepsilon}, \\ (N^{-3}\sqrt{\varepsilon}\ln^4 N + N^{1-\sqrt{\beta}\sigma_0}), & \text{if } x = (x_i + x_{i+1})/2,\ N^{-1} \le \sqrt{\varepsilon}, \\ C(N^{-3}\sqrt{\varepsilon}\ln^4 N + N^{3-\sqrt{\beta}\sigma_0}\ln^3 N), & \text{in other case}. \end{cases} \tag{19}$$

Method	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$	$N = 2048$
hybrid	$7.2314E{-1}$	$2.0974E{-1}$	$5.1976E{-2}$	$1.4784E{-2}$	$4.5414E{-3}$	$1.4013E{-3}$	$4.2698E{-4}$	$1.2885E{-4}$
scheme	1.7857	2.0127	1.8138	1.7028	1.6964	1.7145	1.7285	
HOC	$3.0110E{+0}$	$8.8361E{-1}$	$1.9046E{-1}$	$3.1506E{-2}$	$4.2628E{-3}$	$4.9632E{-4}$	$5.1766E{-5}$	$4.9917E{-6}$
scheme	1.7687	2.2139	2.5958	2.8858	3.1025	3.2612	3.3744	

Table 1: Uniform errors and uniform orders for the global solution

Remark 2. From Theorems 3 and 4 it follows that if $\sqrt{\beta}\sigma_0 \geq 8$, then the approximation to the global solution has order of uniform convergence $O(N^{-4}\ln^4 N)$, and the approximation to the global normalized flux has almost fourth order of uniform convergence at midpoints and almost third order in the rest, for all cases except for $N^{-1} \leq \sqrt{\varepsilon}$. Again, the computational results show that if $N^{-1} \leq \sqrt{\varepsilon}$, the same orders of convergence than in the case $N^{-1} > \sqrt{\varepsilon}$ are obtained.

§4. Numerical Experiments

To illustrate the efficacy of our numerical methods, we solve the problem

$$-\varepsilon u''(x) + (1 + x^2 + \cos(\pi x))u(x) = 1 + x^{4.5} + \sin(\pi x),\ x \in (0,1),\ u(0) = 0, \quad u(1) = 0,$$

for which the exact solution is unknown. We are only interested in the errors outside the mesh points; then, in the tables we will show the errors at midpoints, $x = (x_i + x_{i+1})/2$, of the corresponding modified Shishkin mesh. To approximate the maximum errors at midpoints we use a variant of the double mesh principle (see [4]). The idea is to calculate the numerical solution U^N on the modified Shishkin mesh $\overline{D}^N$ and also the numerical solution $\widetilde{U}^N$ on a new mesh $\widetilde{D}^N$, for which the transition parameter is now given by $\widetilde{\sigma} = \min\left\{1/4, \sigma_0\sqrt{\varepsilon}\ln(N/2)\right\}$.

This slightly altered value of σ will ensure that the positions of transition points remain the same in meshes $\overline{D}^N$ and $\widetilde{D}^{2N}$ and the midpoints $x = (x_i + x_{i+1})/2$ of the mesh $\overline{D}^N$ are also mesh points of the mesh $\widetilde{D}^{2N}$. Then the errors at midpoints are obtained by $E_\varepsilon^N = \max_x |S^N(x) - \widetilde{S}^{2N}(x)|$, $E^N = \max_\varepsilon E_\varepsilon^N$, where S_N and $\widetilde{S}_{2N}$ are the splines defined by (5) on the meshes $\overline{D}^N$ and $\widetilde{D}^{2N}$ respectively. From these errors, the numerical orders of convergence and the uniform orders of convergence are given by $p_\varepsilon^N = \log_2\left(E_\varepsilon^N/E_\varepsilon^{2N}\right)$, $p^N = \log_2\left(E^N/E^{2N}\right)$. We show the results on the range of values $\varepsilon = 2^0, 2^{-2}, 2^{-4}, \ldots, 2^{-48}$.

Table 1 displays the results for the hybrid and the HOC scheme. For each scheme, the first row gives the uniform maximum errors E^N and the second one the uniform orders of convergence p^N. From this table we deduce the almost second order of uniform convergence for the hybrid scheme and the fourth order of uniform convergence, except by the logarithmic factor, for the HOC scheme, in agreements with Theorems 1 and 3 respectively.

To compare the efficacy of the methods, we show the results obtained by using the scheme developed in [8], based on a spline collocation method. Table 2 displays the results obtained in this case; from it we see that if the diffusion parameter ε is not very small then the results are good confirming the almost second order of uniform convergence. Nevertheless, for ε sufficiently small the maximum errors do not stabilize for any value of the discretization parameter N, and therefore the method does not show the uniform convergence for the

Method	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$	$N = 2048$
$\varepsilon = 2^{-8}$	$4.4875E{-2}$	$1.1582E{-2}$	$2.9268E{-3}$	$7.3339E{-4}$	$1.8337E{-4}$	$4.5846E{-5}$	$1.1461E{-5}$	$2.8654E{-6}$
	1.9540	1.9846	1.9966	1.9998	1.9999	2.0000	2.0000	
$\varepsilon = 2^{-16}$	$4.9999E{-1}$	$1.9545E{-1}$	$6.2321E{-2}$	$2.6428E{-2}$	$8.6093E{-3}$	$2.7160E{-3}$	$8.4123E{-4}$	$2.5446E{-4}$
	1.3551	1.6490	1.2377	1.6181	1.6644	1.6909	1.7250	
$\varepsilon = 2^{-24}$	$5.0940E{-1}$	$1.9767E{-1}$	$6.2740E{-2}$	$2.6489E{-2}$	$8.6265E{-3}$	$2.7211E{-3}$	$8.4278E{-4}$	$2.5493E{-4}$
	1.3657	1.6556	1.2440	1.6185	1.6646	1.6909	1.7251	
$\varepsilon = 2^{-32}$	$1.2115E{+0}$	$3.9149E{-1}$	$6.2766E{-2}$	$2.6493E{-2}$	$8.6276E{-3}$	$2.7214E{-3}$	$8.4288E{-4}$	$2.5496E{-4}$
	1.6297	2.6409	1.2444	1.6186	1.6646	1.6909	1.7251	
$\varepsilon = 2^{-40}$	$1.5892E{+1}$	$5.3076E{+0}$	$4.4122E{-1}$	$4.7986E{-2}$	$8.6277E{-3}$	$2.7214E{-3}$	$8.4288E{-4}$	$2.5496E{-4}$
	1.5822	3.5885	3.2008	2.4756	1.6646	1.6909	1.7251	
$\varepsilon = 2^{-44}$	$6.2879E{+1}$	$2.1047E{+1}$	$1.7229E{+0}$	$1.8829E{-1}$	$2.1247E{-2}$	$2.7214E{-3}$	$8.4288E{-4}$	$2.5496E{-4}$
	1.5789	3.6108	3.1938	3.1476	2.9649	1.6909	1.7251	
$\varepsilon = 2^{-48}$	$2.5082E{+2}$	$8.4008E{+1}$	$6.8499E{+0}$	$7.4981E{-1}$	$8.9050E{-2}$	$9.8585E{-3}$	$8.4288E{-4}$	$2.5675E{-4}$
	1.5781	3.6164	3.1915	3.0738	3.1752	3.5480	1.7150	

Table 2: Uniform errors and uniform orders for the global solution using the Surla scheme

Method	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$	$N = 2048$
hybrid	$4.8352E{-2}$	$4.5769E{-2}$	$3.3284E{-2}$	$1.8024E{-2}$	$7.8996E{-3}$	$2.9894E{-3}$	$1.0135E{-3}$	$3.2552E{-4}$
scheme	0.0792	0.4595	0.8849	1.1900	1.4019	1.5605	1.6386	
HOC	$2.7893E{-1}$	$1.1928E{-1}$	$4.2171E{-2}$	$1.2092E{-2}$	$2.7963E{-3}$	$5.5628E{-4}$	$1.0096E{-4}$	$1.7301E{-5}$
scheme	1.2256	1.5000	1.8022	2.1125	2.3296	2.4620	2.5449	

Table 3: Uniform errors and uniform orders for the global normalized flux

global solution. So, we can conclude that this method is considerably worse than these ones developed in this paper.

To approximate the errors associated to the normalized flux, again only at midpoints $x = (x_i + x_{i+1})/2$ of the modified Shishkin mesh, the first idea is to use the derivatives of the numerical splines S_N and $\widetilde{S}_{2N}$ defined on the meshes $\overline{D}^N$ and $\widetilde{D}^{2N}$ respectively; then, we calculate $F_\varepsilon^N = \max_x \sqrt{\varepsilon}\left|S'_N(x) - \widetilde{S}'_{2N}(x)\right|$, $F^N = \max_\varepsilon F_\varepsilon^N$. From these values, the order of convergence and the ε-uniform order of convergence for the flux are calculated by $q_\varepsilon^N = \log_2\left(F_\varepsilon^N/F_\varepsilon^{2N}\right)$, $q^N = \log_2\left(F^N/F^{2N}\right)$.

From Table 3 we cannot observe the predicted almost fourth order of convergence for the normalized flux at midpoints. The reason is related with the use of the double mesh principle, because the midpoint of one mesh is becoming the nodal point in doubling the mesh. Then, to find the errors for the normalized flux we use a second numerical idea. We consider a new mesh $\overline{\overline{D}}^N$ where the mesh points are $\overline{\overline{x}}_{3i} = x_i$, $\overline{\overline{x}}_{3i+1} = x_i + h_{i+1}/3$, $\overline{\overline{x}}_{3i+2} = x_i + 2h_{i+1}/3$, $i = 0, 1, \ldots, N-1$, and $\overline{\overline{x}}_{3N} = x_N$. We denote by $\overline{\overline{U}}^{3N}$ the numerical solution on this mesh and $\overline{\overline{S}}_{3N}$ the corresponding cubic spline. Then, the error associated to the normalized flux at any point x which is not a mesh point, is calculated by $\sqrt{\varepsilon}\left|S'_N(x) - \overline{\overline{S}}'_{3N}(x)\right|$. Table 4 displays the results obtained by using this idea; from it we deduce the almost fourth order of uniform convergence according with Theorem 4.

Method	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$	$N = 2048$
HOC	$2.2766E{-}1$	$7.7506E{-}2$	$2.0308E{-}2$	$4.2751E{-}3$	$7.6856E{-}4$	$1.0860E{-}4$	$1.2598E{-}5$	$1.2857E{-}6$
scheme	1.5545	1.9323	2.2480	2.4757	2.8231	3.1078	3.2925	

Table 4: Uniform errors and uniform orders for the global normalized flux

Acknowledgements

This research was partially supported by the project MEC/FEDER MTM2007-63204 and the Diputación General de Aragón.

References

[1] Bawa, R. K., and Clavero, C. Higher order global solution and normalized flux for singularly perturbed reaction-diffusion problems. *Appl. Math Comp. 216*, 7 (2010), 2058–2068.

[2] Clavero, C., Bawa, R. K., and Natesan, S. A robust second-order numerical method for global solution and global normalized flux of singularly perturbed self-adjoint boundary-value problems. *Int. J. Comput. Math. 86*, 10–11 (2009), 1731—–1745.

[3] Gracia, J., Lisbona, F., and Clavero, C. High order ε-uniform methods for singularly perturbed reaction-diffusion problems. *Lecture Notes in Computer Science 1988* (2001), 350–358.

[4] Miller, J. J. H., O'Riordan, E., and Shishkin, G. I. *Fitted numerical methods for singular perturbation problems*. World Scientific, Singapore, 1996.

[5] Natesan, S., and Bawa, R. K. Second-order numerical schemes for singularly perturbed reaction–diffusion robin problems. *Journal Numerical Analysis, Industrial and Applied Mathematics 2* (2007), 177–192.

[6] Natesan, S., Bawa, R. K., and Clavero, C. A uniformly convergent method for global solution and normalized flux of singular perturbation problems of reaction-diffusion type. *International Journal of Information and Systems Sciences 3* (2007), 207–221.

[7] Stajanovic, M. Global convergence method for singularly perturbed boundary value problems. *International J. Comp. Appl. Math. 181* (2005), 326–335.

[8] Surla, K., and Uzelac, Z. A uniformly accurate spline collocation method for a normalized flux. *International J. Comp. Appl. Math. 166* (2004), 291–305.

Carmelo Clavero
Department of Applied Mathematics. CPS
C/ María de Luna, 3
50018 Zaragoza
clavero@unizar.es

Monografías Matemáticas *García de Galdeano* **35**, 107–116 (2010)

SYMMETRIC AND ROW SCALES PARTIAL PIVOTING STRATEGIES

V. Cortés and J. M. Peña

Abstract. Row and symmetric scaled partial pivoting strategies present nice stability properties for some classes of matrices. In this paper both kinds of strategies are compared. Following [17], the average normalized growth factor for random matrices associated to Gauss elimination with scaled partial pivoting strategies for several norms is approximated by power functions. For nonsingular M-matrices, an economic implementation of the symmetric scaled partial pivoting for the 1-norm is presented.

Keywords: Gauss elimination, scaled pivoting, growth factor, conditioning, M-matrices.

AMS classification: 65F05, 65F35, 65G50.

§1. Introduction

Several pivoting strategies for Gauss elimination, such as partial and complete pivoting, have been deeply studied. Their growth factor has been analyzed from several points of view. The nice behaviour of a pivoting strategy introduced recently, and called rook pivoting, has been analyzed in several papers (see, for instance, [4] and [13]–[15]). This paper considers scaled partial pivoting strategies, which present very nice properties when dealing with some important classes of matrices, as we shall recall and show in this paper. These pivoting strategies have been frequently used in the literature and even in basic books such as [3]. In [14], it is established that row scaled partial pivoting is generally successful when the larger elements of the coefficient matrix of a linear system $Ax = b$ are uniformly distributed across its rows and columns. One of the attractive features of scaled partial pivoting (SPP) is that the accuracy of the computed solution of a linear system by SPP is essentially independent of row scaling of the coefficient matrix. Thus, if the matrix is ill-conditioned due to bad row scaling, then a highly accurate solution can usually be obtained with SPP. A nice explanation of the advantages of SPP comes from the underlying hyperplane geometry of Gauss elimination (see [7] and [14]), as recalled in Section 2.

There are two types of SPP strategies: row SPP and symmetric SPP strategies (see definitions in Section 2). In this paper we compare the properties of these two types of strategies. Besides, there is scarce literature about their stability properties when applied to general or random matrices. This is another topic considered in this paper. Rice (see [16, p. 44]) and Poole and Neal [13] noted that if the elements of the coefficient matrix of a linear system are of uniform size, the computations are more robust. In [17] and [5], the average normalized growth factor for random matrices has been analyzed for several pivoting strategies different from SPP. We analyze the average normalized growth factor for SPP strategies.

In general, a drawback of SPP is its high computational cost. It requires $O(n^3)$ elementary operations in addition to the computational cost of the Gauss elimination of an $n \times n$ matrix.

However, its implementation for special classes of matrices can have lower computational cost.

Let us now mention two classes of matrices playing an important role in many applications where SPP strategies present very nice properties. In the first case (with sign regular matrices), the implementation of the pivoting strategy was performed in [10] and the good properties correspond to row SPP. The second case (with M-matrices) is a novelty of this paper and now the good properties correspond to symmetric SPP. The two classes of matrices are:

- Nonsingular sign regular matrices. An $n \times n$ matrix A is sign regular if, for each k with $1 \leq k \leq n$, all minors of order k have the same sign. Due to their variation diminishing properties, these matrices present important applications in many fields, such as Approximation Theory, Statistics or Computer Aided Geometric Design (see references in [10]). In [10] it was proved that row SPP for any strictly monotone vector norm can be implemented increasing the computational cost of Gauss elimination with $O(n)$ elementary operations, a cost considerably lower than that of partial pivoting. In addition the growth factor is optimal (see Corollary 2.4).
- Nonsingular M-matrices. A nonsingular matrix A is an M-matrix if it has positive diagonal entries, nonpositive off-diagonal entries and A^{-1} is nonnegative. Nonsingular M-matrices present many applications to Numerical Analysis, Dynamic Systems, Economics and Linear Programming, among other fields. In Section 3, we show how to implement with low computational cost (increasing the computational cost of Gauss elimination with $O(n^2)$ elementary operations) the symmetric SPP for $\|\cdot\|_1$ in the class of nonsingular M-matrices. We also show that the growth factor is optimal.

We now introduce some basic notations. Given $k, l \in \{1, 2, \ldots, n\}$, let α (resp., β) be any increasing sequence of k (resp., l) positive integers less than or equal to n. Let A be a real square matrix of order n. Then we denote by $A[\alpha|\beta]$ the $k \times l$ submatrix of A containing rows numbered by α and columns numbered by β. Besides let $A[\alpha] := A[\alpha|\alpha]$. Gauss elimination transforms a linear system $Ax = b$ into an equivalent upper triangular linear system $Ux = c$. Gauss elimination with a given pivoting strategy, for nonsingular matrices A, consists of a succession of at most $n - 1$ major steps resulting in a sequence of matrices as follows:

$$A = A^{(1)} \longrightarrow \tilde{A}^{(1)} \longrightarrow A^{(2)} \longrightarrow \tilde{A}^{(2)} \longrightarrow \cdots \longrightarrow A^{(n)} = \tilde{A}^{(n)} = U,$$

where $A^{(t)} = (a_{ij}^{(t)})_{1 \leq i,j \leq n}$ has zeros below its main diagonal in the first $t-1$ columns. The matrix $\tilde{A}^{(t)} = (\tilde{a}_{ij}^{(t)})_{1 \leq i,j \leq n}$ is obtained from the matrix $A^{(t)}$ by reordering the rows and/or columns $t, t+1, \ldots, n$ of $A^{(t)}$ according to the given pivoting strategy and satisfying $\tilde{a}_{tt}^{(t)} \neq 0$. To obtain $A^{(t+1)}$ from $\tilde{A}^{(t)}$ we produce zeros in column t below the *pivot element* $\tilde{a}_{tt}^{(t)}$ by subtracting multiples of row t from the rows beneath it. Rows $1, 2, \ldots, t$ are not altered. If P is the permutation matrix associated to the pivoting strategy and $B := PA$, then the Gauss elimination of B can be performed without row exchanges and we say that we have performed a *row pivoting strategy*. Finally, if $B = P^T AP$ we say that we have performed a *symmetric pivoting strategy*. In Section 4.2.9 of [8], symmetric pivoting strategies are applied to symmetric matrices. However, in this paper we can apply them to unsymmetric matrices.

§2. Row SPP strategies versus symmetric SPP strategies

A *row* (resp., *symmetric*) *scaled partial pivoting* strategy for a norm $\|\cdot\|$ consists of an implicit scaling by the norm $\|\cdot\|$ followed by partial (resp., symmetric and partial) pivoting. Let $r_i^{(t)}$ denote the ith row ($t \le i \le n$) of the submatrix $A^{(t)}[t, t+1, \dots, n]$. For each t ($1 \le t \le n-1$), these strategies look for the first integer i_t ($t \le i_t \le n$) satisfying

$$\frac{|a_{i_t t}^{(t)}|}{\|r_{i_t}^{(t)}\|} = \max_{t \le i \le n} \frac{|a_{it}^{(t)}|}{\|r_i^{(t)}\|}$$

(resp.,

$$\frac{|a_{i_t i_t}^{(t)}|}{\|r_{i_t}^{(t)}\|} = \max_{t \le i \le n} \frac{|a_{ii}^{(t)}|}{\|r_i^{(t)}\|}).$$

We shall deal with monotone vector norms. As examples of monotone vector norms, we can consider the vector norms $\|\cdot\|_2, \|\cdot\|_1, \|\cdot\|_\infty$. In the particular case of $\|\cdot\|_2$, the associated SPP strategy for Gauss elimination is called Euclidean scaled partial pivoting (ESPP) and has a nice geometric interpretation remarked in [13]. This strategy leads to a triangular system where the hyperplane of $\mathbf{R^n}$ associated to its ith equation ($i = 1, 2, \dots, n$) is well oriented with respect to the x_i-axis. We mean that, in step i, we select as the ith hyperplane the one which is the most orthogonal to the x_i-axis (observe that the strategy is based on direction cosines).

Let us compare row SPP and symmetric SPP strategies with respect to theoretical bounds for the growth factor. Given a matrix M, $|M|$ will denote the matrix whose entries are given by the absolute values of the entries of M. The growth factor is an indicator of the stability of Gauss elimination. Given an $n \times n$ nonsingular matrix A, let us consider the growth factor given by

$$\rho_n(A) := \frac{\| |L|\,|U| \|_\infty}{\|A\|_\infty}, \tag{1}$$

where LU is the triangular factorization of the matrix $B = PAQ$ and P, Q are the permutation matrices associated to the pivoting strategy. Amodio and Mazzia (see [2] p. 398) introduced the number

$$\rho_n^N(A) := \frac{\max_t \|A^{(t)}\|_\infty}{\|A\|_\infty} \tag{2}$$

and have shown its nice behavior for the error analysis of Gauss elimination.

In Corollary 4.2 of [12] it was found an upper bound for the growth factor associated to row SPP strategy for $\|\cdot\|_1$: it satisfies $\rho_n^N(A) \le 2^{n-1}$, analogously to the theoretical bound satisfied by partial pivoting (see [2]). The following example shows that, in contrast, the growth factor of symmetric SPP strategies can be arbitrarily large even for 2×2 matrices.

Example 1. Let us consider $\varepsilon > 0$, the matrix A and the upper triangular matrix U obtaining after applying Gauss elimination with any symmetric SPP strategy (which does not produce row and column exchanges):

$$A = \begin{pmatrix} \varepsilon & 1 \\ 1 & \varepsilon \end{pmatrix}, \quad U = \begin{pmatrix} \varepsilon & 1 \\ 0 & \varepsilon - 1/\varepsilon \end{pmatrix}.$$

Observe that $\rho_2^N(A) = \frac{(1/\varepsilon)-\varepsilon}{1+\varepsilon} = \frac{1-\varepsilon}{\varepsilon}$ is arbitrarily large.

Let us mention that we shall see at the end of this section that the growth factor of symmetric SPP strategies for random matrices is not as catastrophic as in the previous example.

In Theorem 2.2 of [9] it was proved that, given a nonsingular matrix A, if there exists a permutation matrix P such that the LU-factorization of the matrix $B = PA$ satisfies $|LU| = |L|\,|U|$, then P is associated with the row scaled partial pivoting for any strictly monotone vector norm. This can be used to derive nice backward error bounds (see [9]). This happens, for instance, with the class of sign-regular matrices, as shown in [10]. The following result shows that the growth factors defined in (1) and (2) are optimal under the previous hypothesis.

Proposition 1. *Let A be an $n \times n$ nonsingular matrix. If there exists a permutation matrix P such that the LU-factorization of the matrix PA satisfies $|LU| = |L|\,|U|$, then P is associated to the row SPP for a strictly monotone vector norm $\|\cdot\|$ and this strategy satisfies*

$$\rho_n(A) = 1, \qquad \rho_n^N(A) = 1.$$

Proof. The first part of the proposition is consequence of Theorem 2.2 of [9]. The result $\rho_n(A) = 1$ follows from the hypothesis $|LU| = |L|\,|U|$.

Since P is the permutation matrix associated to the row SPP strategy, the Gauss elimination of $B := PA$ can be performed without row exchanges and so, if $B = LU$ with L a lower triangular matrix with unit diagonal and U a nonsingular upper triangular matrix, then

$$B^{(t)}[t,\dots,n] = L[t,\dots,n]U[t,\dots,n]$$

and

$$B^{(t)}[1,\dots,t-1|1,\dots,n] = U[1,\dots,t-1|1,\dots,n].$$

From the previous formulas, we can conclude that $\|A^{(t)}\|_\infty = \|B^{(t)}\|_\infty \le \||L|\,|U|\|_\infty = \|A\|_\infty$ for all $t = 1,\dots,n-1$. Thus, $\rho_n^N(A) = 1$. □

An analogous result to Proposition 1 does not hold for symmetric SPP strategies.

Example 2. The following nonsingular matrix A has associated a permutation matrix P such that the LU-factorization of the matrix PAP^T satisfies $|PAP^T| = |L|\,|U|$:

$$A = \begin{pmatrix} 1 & 1 & 4 \\ 1/2 & 2 & 3 \\ 4 & 3 & 20 \end{pmatrix},\ P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ L = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 3/2 & 13/3 & 1 \end{pmatrix},\ U = \begin{pmatrix} 2 & 1/2 & 3 \\ 0 & 3/4 & 5/2 \\ 0 & 0 & 14/3 \end{pmatrix}.$$

However, P is not associated to the symmetric SPP strategy for $\|\cdot\|_1$. This strategy is associated to the permutation matrix Q and it can also be checked that $|QAQ^T| \neq |\tilde{L}||\tilde{U}|$, where $\tilde{L}\tilde{U}$ is the LU-factorization of the matrix QAQ^T:

$$Q = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ \tilde{L} = \begin{pmatrix} 1 & 0 & 0 \\ 3/20 & 1 & 0 \\ 1/5 & 8/31 & 1 \end{pmatrix},\ \tilde{U} = \begin{pmatrix} 20 & 3 & 4 \\ 0 & 31/20 & -1/10 \\ 0 & 0 & 7/31 \end{pmatrix}.$$

As an application of Proposition 1, let us see that the pivoting strategy introduced in [10] for nonsingular sign regular matrices and called first-last pivoting presents optimal growth factors (1) and (2). Let us also recall that this pivoting strategy increases the computational cost of Gauss elimination in only $O(n)$ elementary operations.

Corollary 2. *Let A be an $n \times n$ nonsingular sign regular matrix. The growth factors (1) and (2) of the first-last pivoting satisfy*

$$\rho_n(A) = 1, \qquad \rho_n^N(A) = 1. \tag{3}$$

Proof. By Corollary 3.5 of [10], the permutation matrix P associated with the first-last pivoting strategy satisfies that PA admits an LU-decomposition $PA = LU$ with $|PA| = |L|\,|U|$. By Corollary 3.6 of [9], this matrix P coincides with the permutation matrix associated with any scaled partial pivoting strategy for a strictly monotone vector norm. Then, by Proposition 2.2, the growth factors (1) and (2) of the first-last pivoting satisfy (3). □

Another good property for backward stability is diagonal dominance. In fact, nice stability properties satisfied when the resultant matrix U is diagonally dominant by rows are described in [11]. In Section 3, we shall see that this happens when we apply symmetric SPP to a nonsingular M-matrix.

As commented in the introduction, SPP strategies present very good stability properties for some special classes of matrices capable of good properties in this sense. Here we analyze the behavior for random matrices. The behavior is worse than with partial pivoting but better than with other strategies considered in [17].

Stability of Gauss elimination with partial pivoting *on average* was analyzed through numerical experiments in [17]. In [5], the stability on average was studied for some pivoting strategies intermediate between partial pivoting and rook pivoting (see [15], [4]). Here we consider the stability on average of row SPP strategies. Following [17], we have considered matrices whose elements are independent samples of the standard normal distribution of mean 0 and variance 1. In the numerical experiments these $n \times n$ matrices are selected at random, with the sample size N diminishing with n to keep the computing time within reasonable bounds. A typical set of dimensions and sample sizes are listed below, although for some of our experiments the samples were larger:

dimension n	2	4	8	16	32	64	128	256	512	1024
sample size N	4096	2048	1024	512	256	128	64	32	20	10

We also modify the classical definition of growth factor due to Wilkinson dividing by the standard deviation σ_A of the initial element distribution:

$$\hat{\rho} := \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\sigma_A},$$

which will be called the *average normalized growth factor*.

In [17] it was shown that the average normalized growth factor of the partial pivoting and complete pivoting for random $n \times n$ matrices was very close to $n^{2/3}$ and $n^{1/2}$, respectively. Now, let us show in Figure 1 and Table 1 the average normalized growth factor of some row scaled partial pivoting strategies: $\hat{\rho}_2$ (corresponding to row SPP for $\|\cdot\|_2$), $\hat{\rho}_1$ (corresponding to row SPP for $\|\cdot\|_1$) and $\hat{\rho}_\infty$ (corresponding to row SPP for $\|\cdot\|_\infty$). The calculations were performed with MATLAB.

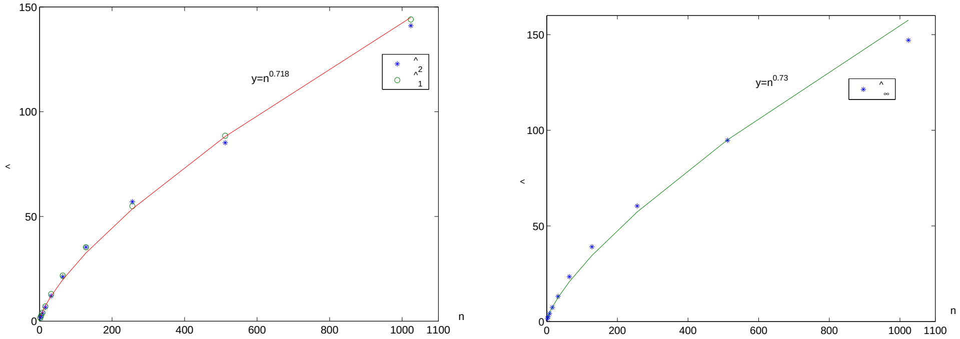


Figure 1: Approximations for $\hat{\rho}_2$ and $\hat{\rho}_1$ (left) and for $\hat{\rho}_\infty$ (right)

n	$\hat{\rho}_2$	$\hat{\rho}_1$	$n^{0.718}$
2	1.5695	1.5763	1.6449
4	2.4725	2.4966	2.7057
8	3.8001	4.0399	4.4506
16	6.6124	7.0371	7.3208
32	12.0579	12.9469	12.0420
64	21.1924	21.7841	19.8078
128	35.3150	35.2997	32.5819
256	57.0067	54.9150	53.5940
512	85.2017	88.5132	88.1568
1024	141.0891	144.0571	145.0091

n	$\hat{\rho}_\infty$	$n^{0.73}$
2	1.5534	1.6586
4	2.5188	2.7511
8	4.1648	4.5631
16	7.3439	7.5685
32	13.1250	12.5533
64	23.4446	20.8215
128	39.0697	34.5353
256	60.4346	57.2816
512	94.7647	95.0095
1024	147.0777	157.5865

Table 1: Approximations for $\hat{\rho}_2$ and $\hat{\rho}_1$ (left) and for $\hat{\rho}_\infty$ (right)

n	$\tilde{\rho}_1$	$n^{0.73}$	$\tilde{\rho}_2$	$n^{0.728}$	$\tilde{\rho}_\infty$	$n^{0.734}$
2	2.9984	1.6586	2.9867	1.6563	2.9397	1.6632
4	5.0854	2.7511	5.0893	2.7435	4.8774	2.7664
8	8.0621	4.5631	7.6822	4.5441	8.0890	4.6012
16	12.4662	7.5685	12.5019	7.5266	12.4775	7.6529
32	19.9420	12.5533	19.2773	12.4666	19.6611	12.7286
64	29.4259	20.8215	29.7842	20.6490	28.8675	21.1707
128	47.7391	34.5353	45.2706	34.2018	48.2033	35.2121
256	75.1349	57.2816	73.1623	56.6498	72.1990	58.5663
512	107.4418	95.0095	104.2726	93.8315	101.2777	97.4101
1024	139.9168	157.5865	139.6121	155.4169	149.4781	162.0168

Table 2: Approximation for $\tilde{\rho}_1, \tilde{\rho}_2, \tilde{\rho}_\infty$

In Figure 1, we observe a very slightly better behavior in the cases of norms $\|\cdot\|_1$ and $\|\cdot\|_2$ than with $\|\cdot\|_\infty$, and slightly worse bounds than with partial pivoting. The average normalized growth factor can be approximated by $n^{0.718}$ for the two first norms and by $n^{0.73}$ for $\|\cdot\|_\infty$.

Now, let us calculate the average normalized growth factor of some symmetric scaled partial pivoting strategies: $\tilde{\rho}_1$ (corresponding to symmetric SPP for $\|\cdot\|_1$), $\tilde{\rho}_2$ (corresponding to symmetric SPP for $\|\cdot\|_2$) and $\tilde{\rho}_\infty$ (corresponding to symmetric SPP for $\|\cdot\|_\infty$). The results are given in Table 2. We note that, in the numerical experiments with symmetric scaled partial pivoting, we have refused the test matrices that have some submatrix $A^{(k)}[k,\dots,n]$ of their elimination process with null diagonal.

In Tables 1 and 2, we also observe that the behavior of the average normalized growth factor for SPP strategies is nice as we previously expected because of the introduction comments for matrices with uniform elements (see Rice (see [16] p. 44) and Poole and Neal [13]). However, if we analyze the obtained approximations of the average normalized growth factor, then we can say that row SPP strategies work better than symmetric SPP strategies for random matrices.

§3. An economic implementation of symmetric SPP for nonsingular M-matrices

In general, a disadvantage of SPP versus PP is the computational cost because SPP pivoting strategies require $O(n^3)$ (instead of $O(n^2)$) elementary operations in addition to the cost of Gauss elimination. However, for special classes of matrices SPP strategies can require lower computational cost. This already happened with the class of sign regular matrices, for which an implementation of row SPP for $\|\cdot\|_1$ with less computational cost than PP was presented in [10]. This section is devoted to the important class of M-matrices.

A nonsingular $n\times n$ matrix A is an *M-matrix* if it has positive diagonal entries, nonpositive off-diagonal entries and A^{-1} is nonnegative. M-matrices have very important applications, for instance, in iterative methods in numerical analysis, in the analysis of dynamical systems, in economics and in mathematical programming. Let us see that we can implement for a nonsingular $n \times n$ M-matrix the symmetric SPP for $\|\cdot\|_1$ increasing the computational cost of Gauss elimination in only $O(n^2)$ elementary operations, and that the corresponding growth factor (2) is optimal. In Proposition 4.7 of [11], a similar computational cost was obtained but with a pivoting strategy which was not a SPP strategy.

Theorem 3. *Let A be a nonsingular $n\times n$ M-matrix and let us consider the linear system $Ax = b$. Then the symmetric scaled partial pivoting for the norm $\|\cdot\|_1$ leads to an upper triangular matrix diagonally dominant by rows and can be implemented with a computational cost which adds $\frac{3}{2}(n^2-n)$ sums, $\frac{1}{2}(n^2-n)$ multiplications, $\frac{1}{2}(n^2-n)$ divisions and $\frac{1}{2}(n^2+n)$ comparisons to the computational cost of Gauss elimination without row or column exchanges. Moreover, the growth factor (2) of this strategy satisfies that*

$$\rho_n^N(A) = 1. \tag{4}$$

Proof. By Proposition 4.3 (i) of [11] and Proposition 4.5 of [11], symmetric SPP for $\|\cdot\|_1$ applied to a nonsingular M-matrix leads to an upper triangular matrix U diagonally dominant by rows. Then, by Proposition 3.1 of [11], $\|A^{(t)}\|_\infty \le \|A\|_\infty$ for all $t \in \{1,\dots,n\}$. So, (4) holds.

For each t $(1 \le t \le n-1)$, the symmetric scaled partial pivoting for $\|\cdot\|_1$ chooses as pivot of the tth step the first integer i_t $(t \le i_t \le n)$ such that

$$|a^{(t)}_{i_ti_t}|/(\textstyle\sum_{j\ge t}|a^{(t)}_{i_tj}|) = \max\limits_{t\le i\le n}(|a^{(t)}_{ii}|/(\textstyle\sum_{j\ge t}|a^{(t)}_{it}|)).$$

Let us recall that if we perform a row permutation and the same column permutation in a nonsingular M-matrix we again obtain a nonsingular M-matrix and that the Schur complements of nonsingular M-matrices are also M-matrices (cf. [6]). So, when applying a symmetric pivoting strategy to a nonsingular M-matrix A the resulting matrices $A^{(t)}[t,\dots,n]$ are also nonsingular M-matrices and have positive diagonal entries. Clearly, i_t is also the first integer between t and n such that

$$\frac{\sum_{j\ge t}|a^{(t)}_{i_tj}|}{a^{(t)}_{i_ti_t}} = \min_{t\le i\le n}\frac{\sum_{j\ge t}|a^{(t)}_{ij}|}{a^{(t)}_{ii}} = \min_{t\le i\le n}\left(1+\frac{\sum_{j\ge t}^{j\ne i}|a^{(t)}_{ij}|}{a^{(t)}_{ii}}\right),$$

which in turn coincides with the first integer between t and n such that

$$1-\frac{\sum_{j\ge t}^{j\ne i_t}|a^{(t)}_{ij}|}{a^{(t)}_{i_ti_t}} = \max_{t\le i\le n}\left(1-\frac{\sum_{j\ge t}^{j\ne i}|a^{(t)}_{ij}|}{a^{(t)}_{ii}}\right) = \max_{t\le i\le n}\frac{a^{(t)}_{ii}-\sum_{j\ge t}^{j\ne i}|a^{(t)}_{ij}|}{a^{(t)}_{ii}}. \tag{5}$$

Taking into account that each matrix $A^{(t)}[t,\dots,n]$ is a nonsingular M-matrix, we know that it has positive diagonal elements and nonpositive off-diagonal entries and so we conclude from (5) that i_t is also the first integer between t and n such that

$$\frac{\sum_{j\ge t}a^{(t)}_{i_tj}}{a^{(t)}_{i_ti_t}} = \max_{t\le i\le n}\frac{\sum_{j\ge t}a^{(t)}_{ij}}{a^{(t)}_{ii}}. \tag{6}$$

It remains to see that we can calculate the indices i_t $(1 \leq t \leq n-1)$ with the number of additional elementary operations mentioned above. Let $e := (1, \ldots, 1)^T$ and $z := Ae$. As usual, we also denote $A^{(1)} := A$, $z^{(1)} := z$. By (6) for $t = 1$, the first index i_1 such that

$$\frac{z_{i_1}}{a_{i_1 i_1}} = \max_{t \leq i \leq n} \frac{z_i}{a_{ii}}$$

determines the pivot row i_1 and the permutation matrix P_1 such that $\tilde{A}^{(1)} = P_1^T A P_1$. The solution of the augmented matrix $(\tilde{A}^{(1)}; P_1^T b, P_1^T z)$ is $(P_1^T x, e)$. If we perform one step of Gauss elimination we arrive at the augmented matrix $(A^{(2)}; b^{(2)}, z^{(2)})$ and we have that $A^{(2)} e = z^{(2)}$. Then, by (6) for $t = 2$, the first index $i_2 \in \{2, \ldots, n\}$ such that

$$\frac{z_{i_2}^{(2)}}{a_{i_2 i_2}^{(2)}} = \max_{2 \leq i \leq n} \frac{z_i^{(2)}}{a_{ii}^{(2)}}$$

determines the pivot row i_2. Iterating this procedure, we conclude that the computational cost of the pivoting strategy corresponds to the extra calculations for obtaining the right side z (given by the row sums of A), for transforming it by Gauss elimination into

$$z^{(2)}[2, \ldots, n], \ldots, z^{(n-1)}[n-1, n],$$

for calculating the quotients $z_i^{(k)}/a_{ii}^{(k)}$ $(k = 1, \ldots, n-1)$ and for choosing the largest component

$$\frac{z_{i_k}^{(k)}}{a_{i_k i_k}^{(k)}} = \max_{k \leq i \leq n} \frac{z_i^{(k)}}{a_{ii}^{(k)}}$$

in each step k. □

Let us observe that, in many applications (as shown in [1]), the row sums (that is, the vector z of the proof of the previous theorem) are natural parameters. In this case, we even can reduce the computational cost of the pivoting strategy in $n^2 - n$ sums.

Acknowledgements

Research partially supported by the Spanish Research Grant MTM2006-03388 and by Gobierno de Aragón and Fondo Social Europeo.

References

[1] Alfa, A. S.; Xue, J., and Ye, Q. Entrywise perturbation theory for diagonally dominant m-matrices with applications. *Numer. Math. 90* (2002), 401–414.

[2] Amodio, P., and Mazzia, F. A new approach to backward error analysis of lu factorization. *BIT 39* (1999), 385–402.

[3] Burden, R. L., and Faires, J. D. *Numerical Analysis*, vol. 40 of *Sixth Edition*. International Thomson Publishing, 1996.

[4] Chang, X.-W. Some features of gaussian elimination with rook pivoting. *BIT 42* (2002), 66–83.

[5] Cortés, V., and Peña, J. M. Growth factor and expected growth factor of some pivoting strategies. *J. Comput. Appl. Math. 202* (2007), 292–303.

[6] Fan, K. Note on *m*-matrices. *Quart. J. Math. Oxford Ser. 11* (1961), 43–49.

[7] Gasca, M., and Peña, J. M. Scaled pivoting for gaussian and neville elimination for totally positive systems. *Appl. Numer. Math. 13* (1993), 345–355.

[8] Golub, G. H., and Van Loan, C. F. *Matrix Computations*. 3rd ed. The Johns Hopkins University Press, London, 1996.

[9] Peña, J. M. Pivoting strategies leading to small bounds of the errors for certain linear systems. *IMA J. Numer. Anal. 16* (1996), 141–153.

[10] Peña, J. M. Backward stability of a pivoting strategy for sign-regular linear systems. *BIT 37* (1997), 910–924.

[11] Peña, J. M. Pivoting strategies leading to diagonal dominance by rows. *Numer. Math. 81* (1998), 293–304.

[12] Peña, J. M. Scaled pivots and scaled partial pivoting strategies. *SIAM J. Numer. Anal. 41* (2003), 1022–1031.

[13] Poole, G., and Neal, L. A geometric analysis of gaussian elimination i. *Linear Algebra Appl. 149* (1991), 249–272.

[14] Poole, G., and Neal, L. Gaussian elimination: When is scaling beneficial? *Linear Algebra Appl. 162–164* (1992), 309–324.

[15] Poole, G., and Neal, L. The rook's pivoting strategy. *J. Comput. Appl. Math. 123* (2000), 353–369.

[16] Rice, J. R. *Matrix Computations and Mathematical Software*. McGraw-Hill, New York, 1981.

[17] Trefethen, L. N., and Schreiber, R. S. Average case stability of gaussian elimination. *Matrix Anal. Appl. 11* (1990), 335–360.

Vanesa Cortés
Departamento de Matem·tica Aplicada
Universidad de Zaragoza
50009 Zaragoza
vcortes@unizar.es

Juan Manuel Peña
Departamento de Matem·tica Aplicada
Universidad de Zaragoza
50009 Zaragoza
jmpena@unizar.es

Monografías Matemáticas *García de Galdeano* **35**, 117–124 (2010)

On the hydrostatic Stokes approximation with non homogeneous Dirichlet conditions

Fabien Dahoumane

Abstract. We deal with the hydrostatic Stokes approximation with non homogeneous Dirichlet boundary conditions. While investigated the homogeneous case, we build a shifting operator of boundary values related to the divergence operator, and solve the non homogeneous problem in a domain with sidewalls.

Keywords: Hydrostatic approximation, De Rham's lemma, shifting operator, primitive equations, non homogeneous Dirichlet conditions.

AMS classification: 35Q30, 35B40, 76D05, 34C35.

§1. Introduction

Let us consider $\Omega \subset \mathbb{R}^3$ a bounded domain defined by

$$\Omega = \{ x = (x', x_3) \in \mathbb{R}^3 \mid x' \in \omega \text{ and } -h(x') < x_3 < 0 \}, \tag{1}$$

where $\omega \subset \mathbb{R}^2$ is a bounded Lipschitz-continuous domain and h, defined in ω, is a mapping satisfying the following assumption.

Assumption 1. *The mapping h is positive and Lipschitz-continuous on ω. Besides, there is a constant $\alpha > 0$ such that*

$$\inf_{x' \in \omega} h(x') \geqslant \alpha. \tag{2}$$

Therefore, Ω has a Lipschitz-continuous boundary Γ splitted into three parts, each one with a positive measure: the surface Γ_S, the bottom Γ_B , and sidewalls Γ_L, defined by:

$$\begin{aligned} \Gamma_S &= \omega \times \{0\}, \quad \Gamma_B = \{(x', -h(x')) \mid x' \in \omega\}, \\ \Gamma_L &= \{ x \in \mathbb{R}^3 \mid x' \in \partial\omega \text{ and } -h(x') < x_3 < 0 \}. \end{aligned}$$

Finally, we denote by $\boldsymbol{n}$ the unit external vector normal to Γ. Below, the drawing of the domain Ω.

Let $\boldsymbol{f}' = (f_1, f_2) : \Omega \to \mathbb{R}^2$, $\Phi : \Omega \to \mathbb{R}$, and $\boldsymbol{g} = (\boldsymbol{g}', g_3) : \Gamma \to \mathbb{R}^3$ be given functions, Φ and $\boldsymbol{g}$ satisfying adequate compatibility conditions (see (7)). In this paper, we study the hydrostatic Stokes approximation consisting in seeking $\boldsymbol{u} : \Omega \to \mathbb{R}^3$ and $p : \omega \to \mathbb{R}$

$$(\mathcal{SH}) \begin{cases} -\Delta \boldsymbol{u}' + \nabla' p = \boldsymbol{f}', \quad \partial_3 p = 0, \quad \nabla \cdot \boldsymbol{u} = \Phi \quad \text{in } \Omega, \\ \qquad\qquad\qquad\qquad \boldsymbol{u}' = \boldsymbol{g}', \quad u_3 n_3 = g_3 \quad \text{on } \Gamma. \end{cases}$$

Here $\nabla' = (\partial_{x_1}, \partial_{x_2})$ denotes the gradient operator with respect to the variables x_1 and x_2.

When Φ and g_3 are identically equal to 0, some authors have considered $(\mathcal{SH})$ as a reduced Stokes-type system. Indeed, let us consider the case of homogeneous conditions. The simplifications of $(\mathcal{SH})$ come from the hydrostatic pressure hypothesis:

$$\frac{\partial p}{\partial x_3} = 0 \text{ in } \Omega, \tag{3}$$

ensuring that p_S, the pressure at $x_3 = 0$, is in fact the real unknown. Moreover, by integrating with respect to x_3 the incompressibility equation:

$$\nabla \cdot \boldsymbol{u} = 0 \text{ in } \Omega, \tag{4}$$

and taking into account the boundary conditions over u_3, it appears that the vertical velocity u_3 is given by the horizontal velocity $\boldsymbol{u}'$. In this case, the equations of $(\mathcal{SH})$ can be reduced to the following system:

$$\begin{cases} -\Delta \boldsymbol{u}' + \nabla' p_S = \boldsymbol{f}' & \text{in } \Omega, \\ \nabla' \cdot \displaystyle\int_{-h(x')}^{0} \boldsymbol{u}'(x', x_3)\, dx_3 = 0 & \text{in } \omega, \\ \boldsymbol{u}' = 0 & \text{on } \Gamma. \end{cases} \tag{5}$$

Then, we get back to u_3 and the global pressure p by setting

$$x \in \Omega, \quad u_3(x) = \int_{x_3}^{0} \nabla' \cdot \boldsymbol{u}'(x', \xi)\, d\xi, \quad p(x) = p_S(x'). \tag{6}$$

However, studying (5) yields real difficulties when the mapping h vanishes on $\partial\omega$. Previous works dealing with (5) use assumption (2). Weak solutions to (5) was investigated in [5, 4]. Results of [5, 4] are then reviewed in [3], where the author deals with some models close to (5).

The purpose of the paper is to present a proof of the following thoerem, in a simplified case. The complete proof is given in [1]. Before, we introduce the space

$$\boldsymbol{X} = H^1(\Omega)^2 \times H(\partial_{x_3}, \Omega),$$

and its hilbertian norm $\|\boldsymbol{u}\|_X = \left(\|\boldsymbol{u}'\|^2_{H^1(\Omega)^2} + \|u_3\|^2_{H(\partial_{x_3},\Omega)}\right)^{1/2}$, where $H(\partial_{x_3}, \Omega)$ is defined in Subsection 2.2.

Theorem 2. *Assume assumption* (2). *Let* $\boldsymbol{f}' \in H^{-1}(\Omega)^2$, $\Phi \in L^2(\Omega)$, $\boldsymbol{g}' \in H^{1/2}(\Gamma)^2$ *and* $g_3 \in L^2(\Gamma)$ *such that* $g_3 = 0$ *on* Γ_L, *and satisfying the following compatibility condition:*

$$\int_\Gamma \boldsymbol{g}' \cdot \boldsymbol{n}'\, d\sigma + \int_\Gamma g_3\, d\sigma = \int_\Omega \Phi\, dx. \tag{7}$$

Then, there is a unique pair $(\boldsymbol{u}, p) \in \boldsymbol{X} \times (L^2(\Omega)/\mathbb{R})$ *solution to Problem* $(\mathcal{SH})$ *and satisfying the estimate,*

$$\|\boldsymbol{u}\|_X + \|p\|_{L^2(\Omega)/\mathbb{R}} \leqslant C\left\{\|\boldsymbol{f}'\|_{H^{-1}(\Omega)^2} + \|\Phi\|_{L^2(\Omega)} + \|\boldsymbol{g}'\|_{H^{1/2}(\Gamma)^2} + \|g_3\|_{L^2(\Gamma)}\right\}, \tag{8}$$

where $C > 0$ *is a constant depending only on* Ω.

The outline of the paper is as follows. In Section 2 we set the appropriate functional framework. In particular, we recall the definition and structure of the anisotropic space $H(\partial_{x_3}, \Omega)$, which is the adapted space for u_3. Moreover, we introduce the usual integration operators M and F (see (14) and (15)), useful in our study, to provide an adapted lemma of De Rham (see Lemma 7). Finally, we prove Theorem 2 in Section 3.

§2. Functional framework

We assume the reader to be familiar with the classical notations and properties of Lebesgue and Sobolev spaces on a regular open set.

2.1. Computations of surface integrals

For any function $\mu : \Gamma \to \mathbb{R}$, we define the functions μ_S or $(\mu)_S$ and μ_B or $(\mu)_B$ by setting

$$x' \in \omega, \quad \mu_S(x') = \mu(x', 0), \quad \mu_B(x') = \mu(x', -h(x')).$$

We start with an important tool which enables us to replace any integrals defined on Γ_S and Γ_B by one defined on ω.

Lemma 3. *The mapping* $\mu \mapsto (\mu_S, \mu_B)$ *is linear and continuous from* $L^2(\Gamma)$ *into* $L^2(\omega)^2$. *Moreover, one has by definition of the measure* $d\sigma$*:*

$$\int_{\Gamma_S} \mu \, d\sigma = \int_\omega \mu_S \, dx' \quad \textit{and} \quad \int_{\Gamma_B} \mu \, d\sigma = \int_\omega \mu_B \sqrt{1 + |\nabla h|^2} dx'. \tag{9}$$

Proof. This result follows from straightforward calculating. □

Remark 1. Notice that the integrals in (9) are well defined since ω is bounded. Next, the third component of the normal n_3 satisfies $n_3 = 1$ on Γ_S, $n_3 = 0$ on Γ_L and $(n_3)_B(1 + |\nabla h|^2)^{1/2} = -1$ on ω. Moreover, $(n_i)_B(1 + |\nabla h|^2)^{1/2} = -\partial_{x_i} h$ in ω. Therefore,

$$\forall \mu \in L^2(\Gamma), \quad \int_\Gamma \mu n_3 \, d\sigma = \int_\omega \mu_S \, dx' - \int_\omega \mu_B \, dx'. \tag{10}$$

$$\int_{\Gamma_B} \mu n_i \, d\sigma = - \int_\omega \mu \frac{\partial h}{\partial x_i} \, dx'. \tag{11}$$

2.2. The anisotropic space $H(\partial_{x_3}, \Omega)$

Let us recall here some useful results that can be found in [6]. Set

$$H(\partial_{x_3}, \Omega) = \Big\{ u \in L^2(\Omega) \, \Big| \, \frac{\partial u}{\partial x_3} \in L^2(\Omega) \Big\},$$

which is a Hilbert space endowed with norm $\|u\|_{H(\partial_{x_3}, \Omega)} = \left(\|u\|^2_{L^2(\Omega)} + \left\|\partial_{x_3} u\right\|^2_{L^2(\Omega)} \right)^{1/2}$. For any $u \in H(\partial_{x_3}, \Omega)$, we have $u n_3 \in H^{-1/2}(\Gamma)$. Then, setting

$$H_0(\partial_{x_3}, \Omega) = \Big\{ u \in L^2(\Omega) \, \Big| \, \frac{\partial u}{\partial x_3} \in L^2(\Omega) \text{ and } u n_3 = 0 \Big\},$$

the following Green's formula holds

$$\forall u \in H(\partial_{x_3}, \Omega),\ \forall v \in H_0(\partial_{x_3}, \Omega), \qquad \int_\Omega u \frac{\partial v}{\partial x_3}\, dx = -\int_\Omega v \frac{\partial u}{\partial x_3}\, dx, \tag{12}$$

as well as the Poincaré's Inequality

$$\forall u \in H_0(\partial_{x_3}, \Omega), \quad \|u\|_{L^2(\Omega)} \leqslant \|h\|_{L^\infty(\omega)} \left\| \frac{\partial u}{\partial x_3} \right\|_{L^2(\Omega)}. \tag{13}$$

2.3. Definition and properties of the operators M and F.

Let u be a function defined in Ω. We consider the following operators

$$x' \in \omega, \quad Mu(x') = \int_{-h(x')}^{0} u(x',\, x_3)\, dx_3, \tag{14}$$

$$x = (x',\, x_3) \in \Omega, \quad Fu(x) = \int_{x_3}^{0} u(x',\, \xi)\, d\xi, \quad Gu(x) = \int_{-h(x')}^{x_3} u(x',\, \xi)\, d\xi. \tag{15}$$

Proposition 4. *The operator M is linear and continuous from $L^2(\Omega)$ into $L^2(\omega)$, and from $H^1(\Omega)$ into $H^1(\omega)$. Then, one has for $i = 1, 2$:*

$$\forall u \in H^1(\Omega), \quad \frac{\partial}{\partial x_i}(Mu) = M\Big(\frac{\partial u}{\partial x_i}\Big) + \frac{\partial h}{\partial x_i} u_B \ in\ \omega; \tag{16}$$

$$\forall u \in H_0^1(\Omega), \quad \frac{\partial}{\partial x_i}(Mu) = M\Big(\frac{\partial u}{\partial x_i}\Big) \ in\ \omega. \tag{17}$$

Moreover, the following relation holds:

$$\forall u \in H_0(\partial_{x_3}, \Omega), \quad M\Big(\frac{\partial u}{\partial x_3}\Big) = 0 \ in\ \omega. \tag{18}$$

Proof. Let $u \in L^2(\Omega)$. By applying Fubini's Theorem, we deduce that $Mu \in L^2(\omega)$ and $\|Mu\|_{L^2(\omega)} \leqslant \|h\|_{L^\infty(\omega)} \|u\|_{L^2(\Omega)}$. Therefore, the mapping M is linear and continuous from $L^2(\Omega)$ into $L^2(\omega)$. Next, for u in $H^1(\Omega)$ and $i = 1, 2$, one has for any $\psi \in \mathcal{D}(\omega)$:

$$\int_\omega Mu \frac{\partial \psi}{\partial x_i}\, dx' = \int_\Omega u \frac{\partial \psi}{\partial x_i}\, dx = -\int_\Omega \frac{\partial u}{\partial x_i} \psi\, dx + \int_\Gamma u\psi\, n_i\, d\sigma.$$

Then, (11) gives

$$\int_{\Gamma_B} u\psi\, n_i\, d\sigma = -\int_\omega u_B \psi \frac{\partial h}{\partial x_i}\, dx', \tag{19}$$

since ψ does not depend on x_3 and since $\psi = 0$ on Γ_L. Thus

$$\int_\omega Mu \frac{\partial \psi}{\partial x_i}\, dx' = -\int_\omega \left[M\Big(\frac{\partial u}{\partial x_i}\Big) + u_B \frac{\partial h}{\partial x_i} \right] \psi\, dx'.$$

Thus (16) holds in $\mathcal{D}'(\omega)$. From Proposition 3 and the fact that h is Lipschitz-continuous, (16) holds in $L^2(\omega)$. The same arguments prove that M is a linear mapping from $H^1(\Omega)$ in $H^1(\omega)$. When u belongs to $H_0^1(\Omega)$, the function u_B vanishes on ω. Therefore, we get (17). Finally, (18) follows from a computation using relation (12). □

Proposition 5. *The operator F is linear and continuous from $L^2(\Omega)$ into $L^2(\Omega)$ and G is the adjoint operator to F. Next, the operator F is continuous from $L^2(\Omega)$ into $H(\partial_{x_3}, \Omega)$, and*

$$\forall u \in L^2(\Omega), \quad \frac{\partial}{\partial x_3}(Fu) = -u \textit{ in } \Omega. \tag{20}$$

Moreover, the following relation holds:

$$\forall u \in H_0(\partial_{x_3}, \Omega), \quad F\Big(\frac{\partial u}{\partial x_3}\Big) = -u \textit{ in } \Omega. \tag{21}$$

Proof. Let $u \in L^2(\Omega)$. Thanks to Fubini's Theorem, we deduce that $Fu \in L^2(\Omega)$ and from Poincaré's Inequality we have $\|Fu\|_{L^2(\Omega)} \leqslant \|h\|_\infty \|u\|_{L^2(\Omega)}$ by . Hence F is linear and continuous from $L^2(\Omega)$ into $L^2(\Omega)$. Again Fubini's Theorem ensures that

$$\forall u,\, v \in L^2(\Omega), \quad \int_\Omega v\, Fu\, dx = \int_\Omega u\, Gv\, dx. \tag{22}$$

Next, (22) gives that for any $\varphi \in \mathcal{D}(\Omega)$,

$$\int_\Omega \frac{\partial \varphi}{\partial x_3}\, Fu\, dx = \int_\Omega u\, G\Big(\frac{\partial \varphi}{\partial x_3}\Big)\, dx = \int_\Omega u\varphi\, dx.$$

Hence (20) holds in $\mathcal{D}'(\Omega)$ and $\partial_{x_3}(Fu) \in L^2(\Omega)$. Moreover, we deduce from above that the operator F is continuous from $L^2(\Omega)$ into $H(\partial_{x_3}, \Omega)$. Finally, we use the same arguments as above and relation (12) to prove (21). □

Remark 2. Let $u \in H^1(\Omega)$ and $\varphi \in \mathcal{D}(\Omega)$. Thanks to Proposition 5 and (10), one gets:

$$\begin{aligned} \int_\Omega G(\frac{\partial u}{\partial x_3})\varphi\, dx &= \int_\Omega u\varphi\, dx + \int_{\Gamma_S \cup \Gamma_B} u n_3\, F\varphi\, d\sigma \\ &= \int_\Omega u\varphi\, dx + \int_\omega u_S\, (F\varphi)_S\, dx' - \int_\omega u_B (F\varphi)_B\, dx'. \end{aligned}$$

By observing that $(F\varphi)_S = 0$ and $(F\varphi)_B = M\varphi$ in ω, one has

$$\int_\Omega G\Big(\frac{\partial u}{\partial x_3}\Big)\varphi\, dx = \int_\Omega u\varphi\, dx - \int_\Omega u_B\varphi\, dx,$$

which provides that,

$$\forall u \in H^1(\Omega), \quad G\Big(\frac{\partial u}{\partial x_3}\Big) = u - \widetilde{u_B} \text{ in } \Omega. \tag{23}$$

We conclude this subsection by giving additional properties on M and F. Precisely, we prove the following relation between the operators M and F.

Proposition 6. *Let $u \in L^2(\Omega)$. Then, the following assertions are equivalent:*

(i) $Mu = 0$ *in* $L^2(\omega)$.

(ii) $(Fu)n_3 = 0$ *in* $H^{-1/2}(\Gamma)$.

Proof. Given $u \in L^2(\Omega)$, Proposition 5 ensure that $(Fu)n_3$ is in $H^{-1/2}(\Gamma)$. Next, (23) gives for any $v \in H^1(\Omega)$:

$$\langle (Fu)\, n_3, v\rangle_{H^{-1/2}(\Gamma),H^{1/2}(\Gamma)} = \int_\Omega \frac{\partial v}{\partial x_3}\, Fu\, dx - \int_\Omega uv\, dx \;= \int_\Omega u\, G\Big(\frac{\partial v}{\partial x_3}\Big)\, dx - \int_\Omega uv\, dx$$
$$= \int_\Omega u\,(v - \widetilde{v_B})\, dx - \int_\Omega uv\, dx.$$

Therefore, one obtains a relation between F and M:

$$\forall (u, v) \in L^2(\Omega) \times H^1(\Omega), \quad \langle (Fu)\, n_3, v\rangle = -\int_\omega v_B\, Mu\, dx', \tag{24}$$

which proves that (i) implies (ii). Conversely, for any ψ in $\mathcal{D}(\omega)$ and applying (24) with $v = \psi$, we get

$$\int_\omega \psi\, Mu\, dx' = \int_\omega v_B\, Mu\, dx' = -\langle (Fu)\, n_3, v\rangle = 0.$$

Then (ii) implies (i): this completes the proof of Proposition 6. □

2.4. Some properties related to the mean divergence operator

For any vector field $\boldsymbol{v} = (v_1,\, v_2,\, v_3)$, we define

$$\nabla' \cdot \boldsymbol{Mu'} = \sum_{i=1,2} \partial_{x_i}(Mu_i),$$

and the corresponding space $\boldsymbol{V}_M = \left\{ \boldsymbol{v}' \in H_0^1(\Omega)^2 \mid \nabla' \cdot M\boldsymbol{v}' = 0 \text{ in } \omega \right\}$.

Lemma 7. *If $\boldsymbol{f}' \in H^{-1}(\Omega)^2$ satisfies*

$$\forall \boldsymbol{v}' \in \boldsymbol{V}_M, \quad \langle \boldsymbol{f}',\, \boldsymbol{v}'\rangle_{H^{-1}(\Omega)^2, H_0^1(\Omega)^2} = 0,$$

then, there is $q \in L^2(\omega)/\mathbb{R}$ such that $\nabla' \widetilde{q} = \boldsymbol{f}'$ in Ω. Moreover, there is a constant $C > 0$ depending only on Ω such that

$$\|q\|_{L^2(\omega)/\mathbb{R}} \leqslant C\, \|\nabla \widetilde{q}\|_{H^{-1}(\Omega)}\,. \tag{25}$$

Proof. Let us set $\boldsymbol{f} = (\boldsymbol{f}',\, 0)$. Let $\boldsymbol{v} \in H_0^1(\Omega)^3$ such that $\nabla \cdot \boldsymbol{v} = 0$. Thanks to (17) and (18) one has $\boldsymbol{v}' \in \boldsymbol{V}_M$. Therefore, using results from [2] from pages 22-25, there is a unique function p in $L^2(\Omega)/\mathbb{R}$ such that $\nabla p = \boldsymbol{f}$. Then, since $\partial_{x_3} p = 0$ in Ω, there is $q \in L^2(\omega)/\mathbb{R}$, such that $p = \widetilde{q}$ in Ω. Thus q satisfies $\nabla' \widetilde{q} = \boldsymbol{f}'$ in Ω. □

§3. Resolution of Problem $(\mathcal{SH})$ with homogeneous Dirichlet conditions

Proposition 8. *Let $\boldsymbol{f}' \in L^2(\Omega)^2$ and assume that Φ and $\boldsymbol{g}$ are identically equal to 0. Then, Problem $(\mathcal{SH})$ has a at least solution $(\boldsymbol{u},\, p)$ in the space $\boldsymbol{X} \times (L^2(\Omega)/\mathbb{R})$.*

Proof. Let us consider the solution $(\boldsymbol{u},\, p)$ related to the data $\boldsymbol{f}' = 0$. We multiply the first equation of $(\mathcal{SH})$ by $\boldsymbol{u}'$. Then, using (12) and since $\nabla \cdot \boldsymbol{u} = 0$ and $\partial_{x_3} p = 0$ in Ω, one has

$$\int_\Omega \nabla \boldsymbol{u}' : \nabla \boldsymbol{u}'\, dx = \int_\Omega p\, \nabla' \cdot \boldsymbol{u}'\, dx = -\int_\Omega p\, \frac{\partial u_3}{\partial x_3}\, dx = \int_\Omega u_3\, \frac{\partial p}{\partial x_3}\, dx = 0.$$

Therefore $\nabla \boldsymbol{u}' = 0$ in Ω and, since Ω is connected, $\boldsymbol{u}' = 0$ in Ω. As $\nabla \cdot \boldsymbol{u} = 0$ in Ω, we deduce that $\partial_{x_3} u_3 = 0$ in Ω, and from the inequality (13) we get $u_3 = 0$ in Ω. Next, since $\nabla' p = \Delta \boldsymbol{u}' = 0$ in Ω, one obtains that $\nabla p = 0$ in Ω, hence $p = 0$ in Ω. Finally, the solution related to the data $\boldsymbol{f}' = 0$ is $\boldsymbol{u} = 0$ and $p = 0$, which proves that Problem $(\mathcal{SH})$ has at least one solution in $\boldsymbol{X} \times (L^2(\Omega)/\mathbb{R})$. □

Theorem 9. *Let $\boldsymbol{f}'$ in $H^{-1}(\Omega)^2$ and assume that Φ and $\boldsymbol{g}$ are identically equal to 0. Then, Problem $(\mathcal{SH})$ has a unique solution $(\boldsymbol{u},\, p)$ in the space $\boldsymbol{X} \times (L^2(\Omega)/\mathbb{R})$. Moreover, there is a constant $C > 0$ such that*

$$\|\boldsymbol{u}'\|_{H^1(\Omega)^2} + \|u_3\|_{H(\partial_{x_3},\Omega)} + \|p\|_{L^2(\Omega)/\mathbb{R}} \leqslant C\|\boldsymbol{f}'\|_{H^{-1}(\Omega)^2}. \tag{26}$$

To prove Theorem 9, we need Lemma 7 and the proposition stated below.

Lemma 10. *Let $\boldsymbol{u} = (\boldsymbol{u}',\, u_3)$ with $\boldsymbol{u}'$ in $H^1_0(\Omega)^2$ and u_3 in $H(\partial_{x_3}, \Omega)$. Then the following assertions are equivalent*

(i) $\nabla \cdot \boldsymbol{u} = 0$ *in* Ω, $\quad u_3 n_3 = 0$ *in* $H^{-1/2}(\Gamma)$.

(ii) $\nabla' \cdot (M\boldsymbol{u}') = 0$ *in* ω, $\quad u_3 = F(\nabla' \cdot \boldsymbol{u}')$ *in* Ω.

Proof. Assume that (i) holds. Then, (18) and (21) yield

$$M(\nabla' \cdot \boldsymbol{u}') = 0 \quad \text{and} \quad u_3 = F(\nabla' \cdot \boldsymbol{u}').$$

Moreover, thanks to (17) one has $M(\nabla' \cdot \boldsymbol{u}') = \nabla' \cdot M\boldsymbol{u}'$, from which follows (ii). Conversely, one has by (20), $\nabla \cdot \boldsymbol{u} = 0$. Since $M(\nabla' \cdot \boldsymbol{u}') = 0$, Proposition 6 ensures that $n_3 F(\nabla' \cdot \boldsymbol{u}') = 0$ in $H^{-1/2}(\Gamma)$. Hence $u_3 n_3 = 0$ in $H^{-1/2}(\Gamma)$. □

From Lemma 10 and the fact that p does not depend on x_3, solving Problem $(\mathcal{SH})$ reduces to solve the following problem:

$$\text{Find } (\boldsymbol{u}',\, p_S) \in H^1_0(\Omega)^2 \times (L^2(\omega)/\mathbb{R}) \text{ such that:}$$
$$\begin{cases} -\Delta \boldsymbol{u}' + \nabla' p_S = \boldsymbol{f}' & \text{in } \Omega, \\ \nabla' \cdot M\boldsymbol{u}' = 0 & \text{in } \omega, \\ \boldsymbol{u}' = 0 & \text{on } \Gamma. \end{cases} \tag{27}$$

We get back to p and u_3 thanks to (6). The existence and uniqueness of the solution to (27) is given by the following proposition.

Proposition 11. *Let $\boldsymbol{f}'$ in $H^{-1}(\Omega)^2$. There is a unique solution $(\boldsymbol{u}', p_S)$ in the space $H_0^1(\Omega)^2 \times (L^2(\omega)/\mathbb{R})$ to Problem* (27). *Moreover, there is a constant $C > 0$ such that*

$$\|\boldsymbol{u}'\|_{H^1(\Omega)^2} + \|p_S\|_{L^2(\omega)/\mathbb{R}} \leqslant C\,\|\boldsymbol{f}'\|_{H^{-1}(\Omega)^2}. \tag{28}$$

Proof. Any solution $(\boldsymbol{u}', p_S)$ in the space $H_0^1(\Omega)^2 \times (L^2(\omega)/\mathbb{R})$ satisfies the following variational formulation:

$$\forall \boldsymbol{v}' \in \boldsymbol{V}_M, \int_\Omega \nabla \boldsymbol{u}' : \nabla \boldsymbol{v}'\, dx = \langle \boldsymbol{f}', \boldsymbol{v}' \rangle_{H^{-1}(\Omega)^2, H_0^1(\Omega)^2} \,. \tag{29}$$

Conversely, any solution $\boldsymbol{u}' \in \boldsymbol{V}_M$ to (29) is such that

$$\forall \boldsymbol{v}' \in \boldsymbol{V}_M, \quad \langle -\Delta \boldsymbol{u}' - \boldsymbol{f}', \boldsymbol{v}' \rangle_{H^{-1}(\Omega)^2, H_0^1(\Omega)^2} = 0.$$

Therefore, Lemma 7 provides a unique p_S in $(L^2(\omega)/\mathbb{R})$ such that $(\boldsymbol{u}', p_S)$ is a solution to (27). Then, by Lax-Milgram's lemma, there is a unique $\boldsymbol{u}'$ in $\boldsymbol{V}_M$ satisfying (29) and $\|\nabla \boldsymbol{u}'\|_{L^2(\Omega)} \leqslant C\|\boldsymbol{f}'\|_{H^{-1}(\Omega)^2}$, hence $\|\boldsymbol{u}'\|_{H^1(\Omega)^2} \leqslant C\,\|\boldsymbol{f}'\|_{H^{-1}(\Omega)^2}$ by Poincaré's Inequality, where $C > 0$ denotes is a constant depending only on Ω. To finish, we deduce (28) from (25) since

$$\|p_S\|_{L^2(\omega)/\mathbb{R}} \leqslant C\left\|\nabla \widetilde{p_S}\right\|_{L^2(\Omega)} \leqslant C\|\boldsymbol{f}'\|_{H^{-1}(\Omega)^2}. \qquad \square$$

Thanks to Proposition 11 and Lemma 10, $(\mathcal{SH})$ admits a unique solution $(\boldsymbol{u}, p) \in \boldsymbol{X} \times (L^2(\Omega)/\mathbb{R})$. Combining results from Proposition 11 and Proposition 5, we get (8). This complete the proof of Theorem 8.

References

[1] Dahoumane, F., Amrouche, C., and Vallet, G. On the hydrostatic Stokes approximation with non homogeneous boundary conditions. *DEA* (2009, to appear).

[2] Girault, V., and Raviart, P. A. *Finite element methods for the Navier-Stokes equations*, vol. 5 of *Springer Series in Computational Mathematics*. Springer Verlag, 1986.

[3] Lewandowski, R. *Analyse mathématique et océanographie*. Masson, 1997.

[4] Lions, J.-L., Temam, R., and Wang, S. New formulation of the primitive equations of the atmosphere and applications. *Nonlinearity 5* (1992), 1007–1053.

[5] Lions, J.-L., Temam, R., and Wang, S. On the equations of the large scale ocean. *Nonlinearity 5* (1992), 237–288.

[6] Temam, R. Sur la stabilité et la convergence de la méthode des pas fractionnaires. *Ann. Math. Pura ed Applicata LXXIX* (1968), 191–379.

Dahoumane Fabien
Laboratoire de Mathématiques Appliquées
Université de pau et des pays de l'Adour
I.P.R.A, B.P. 1155
64130 Pau Cedex, France
fabien.dahoumane@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **35**, 125–135 (2010)

SIMULATION OF RAINFALL EVENTS AND OVERLAND FLOW

Olivier Delestre and François James

Abstract. We are interested in simulating overland flow on agricultural fields during rainfall events. The model considered is the shallow water system (or Saint-Venant equations) without infiltration, complemented with a friction term. In this context, we definitely have to cope with dry/wet interfaces and water inflow on dry soil. We present a simplified one-dimensional model, discretized with a well-balanced finite volume method, and we describe the specific additional features needed to deal with dry/wet transitions and steady-state solutions due to topography and friction. The method as well as the choice of the friction term are tested and discussed both on analytical solutions and experimental results.

Keywords: Shallow water equations, finite volume schemes, well-balanced schemes, hydrostatic reconstruction, friction laws, rainfall hydrograph, analytical solution, dry/wet transitions.

AMS classification: 76M12, 74G05, 74G75, 35L65, 20C20.

Introduction

Rain on agricultural fields can yield to overland flow. This flow may have some undesirable effects. At the field scale, we can have soil erosion and pollutant transport. Downstream the fields, roads and houses may be damaged. To prevent these effects, control measures can be taken, such as grass strips. But one must know how the water is flowing in order to place efficiently these developments. In the spirit of [6, 7], we try to model these phenomenon by using the shallow water (or Saint-Venant) equations. Efficient numerical simulations are of great help in this context, because field measurements, such as velocities or water heights, are very difficult to obtain, especially during the rain event, which is quite unpredictible.

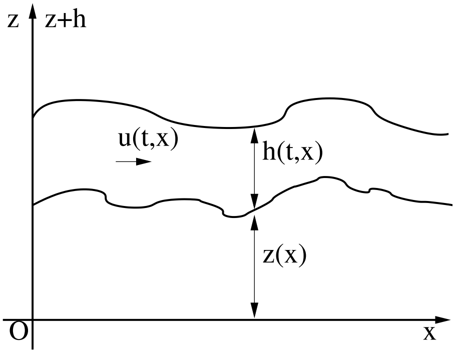
The aim of this paper is not to give a complete account on the problem, which has to be thaught of as a multi-scale problem: one has to deal with roughness induced at the decimeter scale (e.g. by furrows on agricultural surfaces), flows at the scale of ten square meters, which is the scale of the numerical topography data, and also the agricultural field itself, whose surface is of the order of the hectare. We give here a short review of the shallow water equations, with emphasis on some specific aspects in this context. Namely, since the rain is an intermittent phenomenon, we definitely have to cope with dry/wet transitions, a problem analogous to the vacuum apparition in gas dynamics. More classically in shallow water problems, we have to take into account carefully the interactions between the soil topography and the friction of water on the soil, which eventually lead to steady-state solutions that have to be computed accurately.

For this introduction to the topic, we deliberately use a simplified model, firstly by considering one-dimensional flows. This is enough to understand the ideas of the numerical methods, which can be developed in two space dimensions on a rectangular mesh. Next, from a more practical viewpoint, we neglect importants phenomena, which deserve a complete modelling: infiltration and soil erosion. Infiltration appears as a supplementary source term in the shallow water equations, and can be treated quite easily, when a relevant model is chosen. Erosion is a much more complex problem.

We begin by a short review of the shallow water system, recalling a few basic properties. Next, we describe numerical methods adapted to the situation, in particular we discuss briefly the discretization of the friction terms. Finally, we give several illustrations of the results. First we justify the choice of the method by comparison with analytical solutions. Next, we show an attempt of recovering experimental results, with a clear evidence that the choice of the friction laws is not obvious. The last section is devoted to an unstability phenomenon wich occurs when perturbating steady-state solutions (with rain for instance): the so-called roll-waves.

§1. Model

The model we consider here are the so-called shallow-water equations, which are convenient for small heights of water, according to the following scheme



The unknowns are here the velocity of the water $u(t,x)$, and its height $h(t,x)$. The shape of the bottom is also called the topography, it is a given function z. For our specific application, the model has to be complemented by taking int account friction on the soil and rain. Therefore the equations are

$$\partial_t h + \partial_x(hu) = R(t), \qquad \partial_t(hu) + \partial_x\Big(hu^2 + \frac{gh^2}{2}\Big) = -gh\Big(\partial_x z + S_f\Big), \tag{1}$$

where g is the gravity constant, $R(t)$ the rain intensity, assumed constant in space, and $S_f(h,u)$ the friction term. Notice that infiltration in the soil can be accounted by a source term in the first equation like $R(t) - I(t,x)$, where I is a given function. We shall denote by $q = hu$ the water flow, or discharge. The typical practical configuration we consider is a channel with finite length L, so that the system must be set on the interval $]0, L[$, and complemented with boundary conditions at inflow and outflow we do not detail here, see an example in Section 3.

Concerning the friction term, it is a given function of h and u, two examples widely used in hydrology (see for instance [6, 7, 8, 9]) are the Manning and the Darcy-Weisbach friction laws, which are given respectively by

$$S_f = -\frac{k^2 u|u|}{h^{4/3}} = -\frac{k^2 q|q|}{h^{10/3}}, \qquad S_f = -\frac{ku|u|}{8gh} = -\frac{kq|q|}{8gh^3}, \tag{2}$$

where $k > 0$ stands for the roughness coefficient. Both laws are derived from empirical considerations, in particular in the context of pipelines. The problem of their relevance in the present context of overland flow is difficult.

The system can be rewritten in a more compact form by setting

$$U = \begin{pmatrix} h \\ q \end{pmatrix}, \quad F(U) = \begin{pmatrix} q \\ q^2/h + gh^2/2 \end{pmatrix}, \quad B = \begin{pmatrix} R \\ -gh\left(\partial_x z + S_f\right) \end{pmatrix}.$$

We obtain therefore

$$\partial_t U + \partial_x F(U) = \partial_t U + F'(U)\partial_x U = B.$$

The system is by definition hyperbolic if the matrix $F'(U)$ admits a basis of eigenvectors with real eigenvalues, strictly hyperbolic if the eigenvalues are distinct. An easy computation shows that the shallow water system is strictly hyperbolic provided $h > 0$, with eigenvalues $\lambda_-(U) = u - \sqrt{gh}, \quad \lambda_+(U) = u + \sqrt{gh}$. When $h = 0$, the system is no longer hyperbolic, actually it is rather meaningless, since $h = 0$ means that there is no water, so that the velocity u cannot be defined. This is exactly the problem of the vacuum in the Euler equations of fluid mechanics, and leads to severe numerical problems, which cannot be avoided in our context since we consider rain on dry soils.

At this point, we introduce an important quantity, the so-called Froude number

$$Fr = \frac{u}{\sqrt{gh}}. \tag{3}$$

This dimensionless number plays the same role as the Mach number in fluid mechanics, and allows to classify the flows:

- $Fr < 1$ subcritical flow, as in a river (corresponding to subsonic flow in fluid mechanics);
- $Fr > 1$ supercritical flow, as in a torrent (subsonic flow);
- $Fr = 1$ critical flow (transonic flow).

The differences between these flows can be easily experimented by observing the surface waves obtained by throwing a stone in a river.

§2. Numerical method

The shallow water system is discretized by a finite volume method on a fixed time-space grid. A time step $\Delta t > 0$ and a space step $\Delta x > 0$ are fixed, we set $x_i = i\Delta x$, and the interval $]x_i - \Delta x/2, x_i + \Delta x/2[$ will be referred to as the cell i. The finite volume scheme can be written in a compact form as

$$\frac{d}{dt}U_i + \frac{1}{\Delta t}(\mathcal{F}_{i+1/2} - \mathcal{F}_{i-1/2}) = S_i, \tag{4}$$

where the vector U_i is an approximation of the conservative variables in the cell i, $\mathcal{F}_{i+1/2}$ is the numerical flux at the interface between cells i and $i+1$, and S_i a discretization of the source term. Boundary conditions are treated by the method of characteristics (see [4]). The scheme is completely determined once the numerical flux and the source term discretization have been fixed. These choices are not independant one from the other.

Indeed it is well-known that source terms in hyperbolic systems of conservation laws give rise to serious problems. The main difficulty is to find schemes that preserve equilibria (steady-states solutions). In system (1), the main problems are due to

– topography: pools, lakes;

– friction terms: balance between kinematics and friction.

The rain source term can be treated by a second-order accurate Strang type splitting.

Schemes that preserve equilibria are known as well-balanced schemes. The strategy to obtain such schemes consists in choosing first a consistent numerical flux for the system without source terms. Next, a correction is given to take into account equilibria. The reader can find all the details and a large bibliography in the book [3]. We merely give a sketch of the method here, with emphasis on the problem of friction. The numerical flux is the so-called HLL flux, and the order 2 is obtained in space by a MUSCL type reconstruction, in time by Runge Kutta (Heun) (see [3] for details). Notice that dry/wet transitions imply a specific reconstruction for the water height, not only for the velocity as usual (see [1]).

First we consider the equilibria for topography. They are given by

$$hu = Cst, \qquad u^2/2 + g(h+z) = Cst.$$

However a complete resolution of these equations would lead to a far too time consuming scheme. Thus, following [3, 1, 2], we limit ourselves to the equibria at rest:

$$u = 0, \qquad g(h+z) = Cst.$$

This procedure is known as the (second order) hydrostatic reconstruction, and it turns out to give good results at an acceptable numerical cost. We refer the reader interested into details to the preceding references.

Now we turn to friction terms, which can be treated by two different means. The first one aims at building a well-balanced scheme for friction as well as topography, is the apparent topography method, introduced by [3]. It consists in building an modified topography z_{app} which takes into account the friction, as follows:

$$z_{app} = z - b, \qquad \text{with } \partial_x b = S_f.$$

We proceed then exactly as before, with this new topography (detailed computations for the friction laws (2) can be found in [5]). This gives rise to a scheme which computes neatly equilibrium states, but is not completely satisfactory on transition solutions, as we shall see in the next section.

Therefore we turned to a splitting method, and we chose the semi-implicit treatment proposed in [4], not only because it preserves steady states at rest, but also for its stability. For the Darcy-Weisbach friction law (2)-right, it writes

$$q_i^{n+1} + \frac{f|q_i^n|q_i^{n+1}}{8h_i^n h_i^{n+1}}\Delta t = q_i^n + \frac{\Delta t}{\Delta x_i}(\mathcal{F}_{i+1/2G} - \mathcal{F}_{i-1/2D}),$$

where the right-hand side is nothing more than the discharge obtained at each step of the second order in time Runge-Kutta reconstruction. Notice also the simplicity of the method, which gives an explicit value for q_i^{n+1}. Now we illustrate these ideas on a set of analytical solutions.

§3. Analytical solutions

Here we present briefly an adaptation to the $1-d$ case and our friction laws of an idea presented in [8, 9] for pseudo two dimensional cases. At steady states, we have $\partial_t h = \partial_t u = \partial_t q = 0$, thus the mass-conservation equation gives $q = cst$ and we get the equation

$$\partial_x z = \left(\frac{q^2}{gh^3} - 1\right)\partial_x h + S_f(q,h) \tag{5}$$

where $S_f(q,h)$ depends on the friction law chosen, for instance (2). For any given value of the constants k and q, once we are given an explicit expression for $h(x)$, then formula (5) allows us to compute the topography corresponding to this steady state and this water height. Other friction laws can of course be chosen.

As an example, we consider a channel of length 1000 m, with a specified water height $h(x)$ given by

$$h(x) = \left(\frac{4}{g}\right)^{1/3}\left(1 + \frac{1}{2}\exp\left(-16\left(\frac{x}{1000} - \frac{1}{2}\right)^2\right)\right).$$

The friction model is the Manning law, with roughness coefficient $k = 0.033$. The topography is calculated iteratively thanks to (5). To make use of the shallow water system, we have now to impose boundary conditions. Since the flow is subcritical both at inflow $x = 0$ and outflow $x = 1000$, we have to impose the value of one quantity at inflow and one at outflow. We choose to put a discharge of $q = 2$ m^2/s at inflow and a water height corresponding to the value of $h(1000)$ downstream.

We first compare the results obtained by the apparent topography and the semi-implicit scheme in preserving the equilibrium state. It turns out that both methods preserve correctly the steady state along time, as is evidenced by fig. 1.

Since for our application we are particularly interested in non-stationary solutions, we have considered an initially dry soil and the upstream discharge $q = 2$ m^2/s, and computed the unsteady solution up to equilibrium. Both methods (apparent topography and semi-implicit treatment) converge towards the steady state, with slightly better results with the apparent topography method. However, before the steady state is reached, we have a wet/dry transition (fig. 2). We note that the apparent topography method is not adapted to this transition: we have a peak in the velocity profile (fig. 2-left), which appears also in the water height profile. With the semi-implicit treatment, the water height profile is very clean (fig. 2-right).

In figure 3, two more examples of computation of steady states are displayed, both with sub- and supercritical inflows and outflows, and using the semi-implicit method. The numerical scheme deals in particular with transition from one regime to the other, including hydraulic jumps (fig. 3-right).

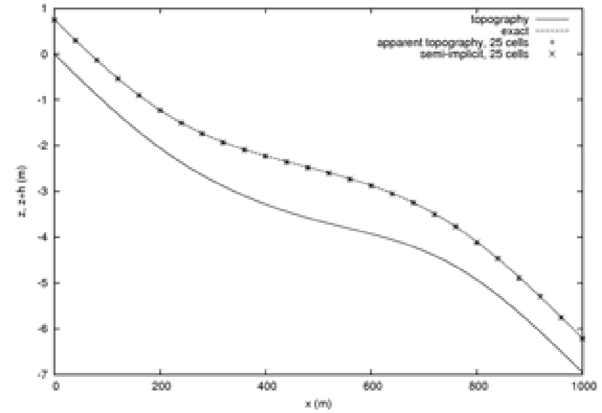


Figure 1: Steady state solution, subcritical inflow and outflow: apparent topography +, semi-implicit ×, analytical −.

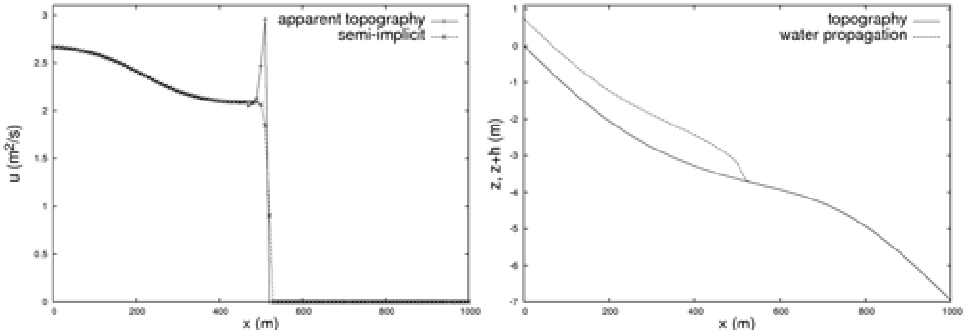


Figure 2: Left: water front velocities at $t = 200$ s: apparent topography (+), semi-implicit treatment (×). Right: water front height at $t = 200s.$, semi-implicit

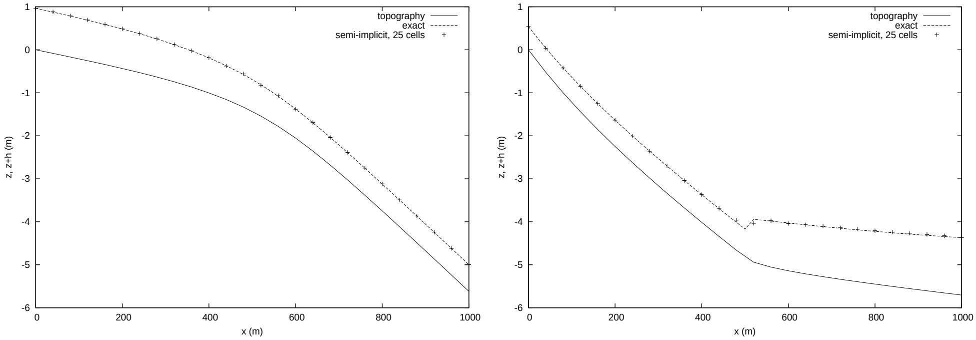


Figure 3: Steady state solution, numerical (symbols) vs analytical (lines). Left: subcritical inflow and supercritical outflow, right: supercritical inflow and subcritical outflow.

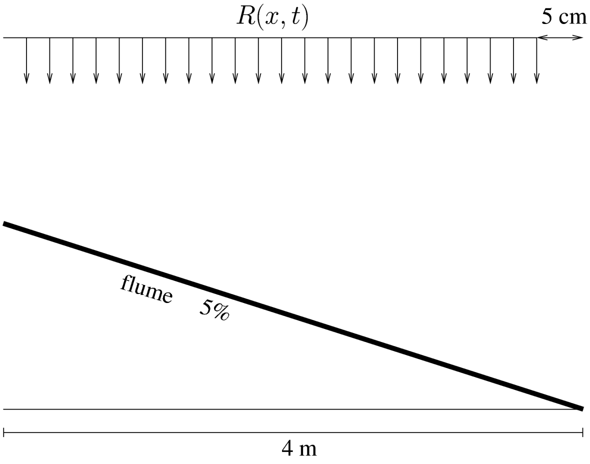


Figure 4: Experimental configuration.

§4. Rainfall hydrograph test

In this section we present another test case, based on experimental measurements realized thanks to the ANR project METHODE in a flume at the rain simulation facility at INRA-Orléans. The flume is 4 m long with a slope of 5% (fig. 4). The simulation duration is 250 s. The rainfall intensity $R(x,t)$ is described by

$$R(x,t) = \begin{cases} 50 \text{ mm/h} & \text{if } (x,t) \in [0, 3.95 \text{ m}] \times [5, 125 \text{ s}], \\ 0 & \text{otherwise.} \end{cases}$$

For this test, dry/wet transitions are involved, since on the one hand there is no rain on the last 5 cm of the flume, on the other hand rain falls on a dry soil. The measured output is an hydrograph, that is a plot of the discharge versus time (see fig. 5).

The mathematical model for this ideal overland flow is the following. We consider a uniform plane catchment whose overall length in the direction of flow is L. The surface roughness and slope are assumed to be constant in space and time. The friction law is the Darcy-Weisbach one. We consider a constant rainfall excess such that

$$R(x,t) = \begin{cases} I & \text{for } 0 \leq t \leq t_d, 0 \leq x \leq L, \\ 0 & \text{otherwise,} \end{cases}$$

where I is the rainfall intensity and t_d is the duration of the rainfall excess. First we compute some explicit "naive" analytical solution to the problem. We notice that three phases can clearly be identified on the hydrograph: a first non-steady step at the beginning of the rainfall event, then a steady-state and lastly another non-steady step when rain stops. The first and the second step solutions can be computed explicitly, and the "naive" solution is obtained by assuming a simple concatenation of the two parts (we refer to [5] for the detailed computations).

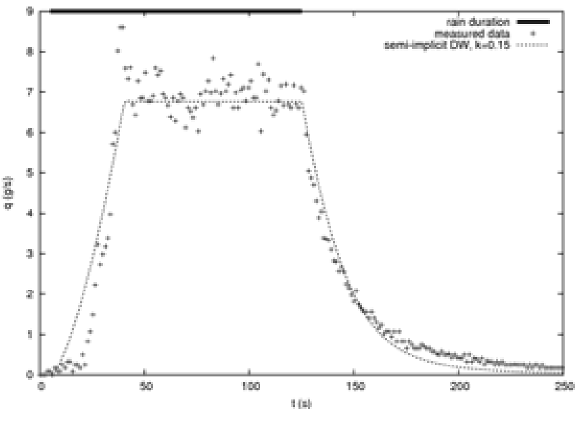


Figure 5: Comparison between experimental measures (+) and numerical results (−).

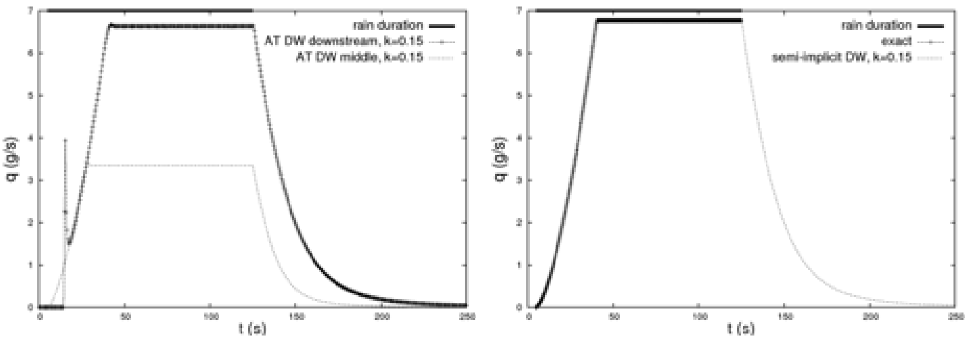


Figure 6: Computed rainfall hydrographs for Darcy-Weisbach's law (DW). Left: apparent topography method (AT). Right: semi-implicit scheme.

At first we compare numerical results with the analytical "naive" solution. Once again, with (fig. 6-a) we show that with the apparent topography method, we get a peak on the discharge downstream that we do not get far from this transition. With the semi-implicit method, we do not have this peak (fig. 6-b). This treatment gives good results close to the "naive" exact solution. The hydrograph is well calculated (fig. 6-b), notice here the computed hydrograph at the middle of the flume, a quantity hardly accessible by experiment.

Next, we propose a comparison between experimental measurements and numerical simulation (fig. 5), obtained with the Darcy-Weisbach friction law. We obtain a reasonable agreement, but it turns out that it is impossible to fit correctly the shape of both the increasing and decreasing parts of the hydrograph. This indicates clearly that the model has to be modified, for instance by choosing alternative friction laws, but this is beyond the scope of this paper.

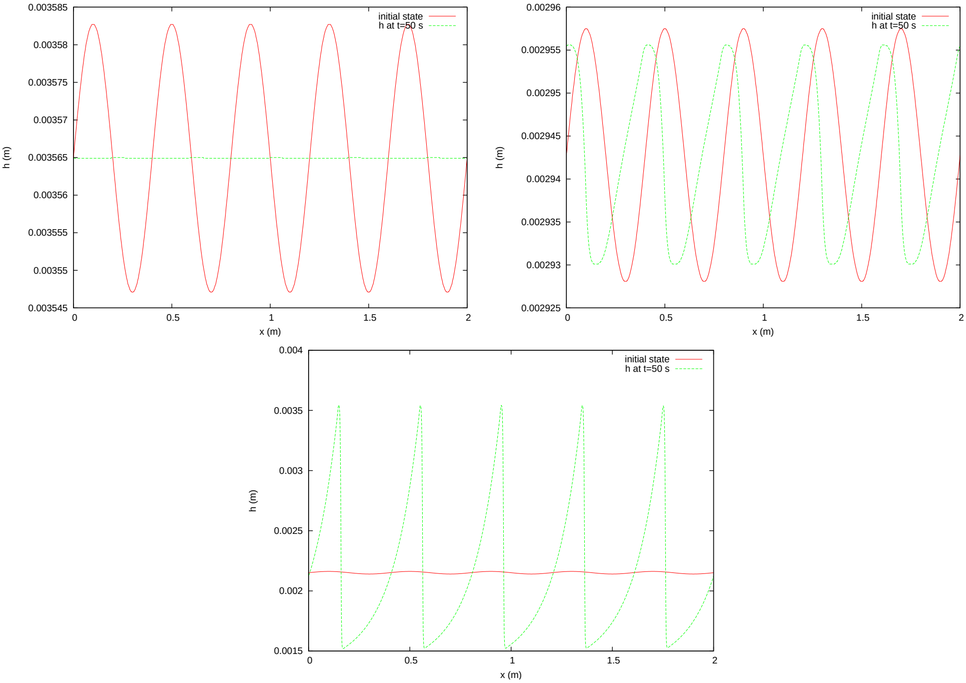


Figure 7: Perturbed initial and final ($t = 200\,\mathrm{s}$) states for Fr=1.5 (top left), Fr=2 (top right), Fr=3.7 (bottom).

§5. Roll waves

This section is devoted to some examples of the so-called “roll-waves”, a phenomenon which results from the competition between topography and friction. Several steady regimes turn out to be unstable, a slight perturbation generating a periodic travelling wave with shocks (hydraulic jumps). In ref [10], Que and Xu gather a set of explicit computations in the simple case of a constant steady states in inclined open channels with constant slope. They provide a precise analysis for the linear stability, proving in particular the following criterion: the initial constant state is linearly stable if and only if the Froude number (3) is smaller than 2.

We recover here these results, using the semi-implicit scheme described above, together with hydrostatic reconstruction. The initial height of water is different for each case, but the amplitude of the perturbation is the same. The “final states” showed here are computed at time $t = 200\,\mathrm{s}$, since it turns out that the solution is stabilized at this time. All cases are perfectly computed, the convergence rates for different values of the Froude number are given in figure 8.

Comparisons between the initial perturbation and the final state are displayed in fig. 7. For $Fr = 2$, the initial state is supposed to be exactly stable, the smaller amplitude of the final result is due to the numerical diffusion. Notice the nonlinear effects (fig. 7, top right). For $Fr < 2$ (top left), the initial perturbation completely disappears, for $Fr > 2$ (bottom), a

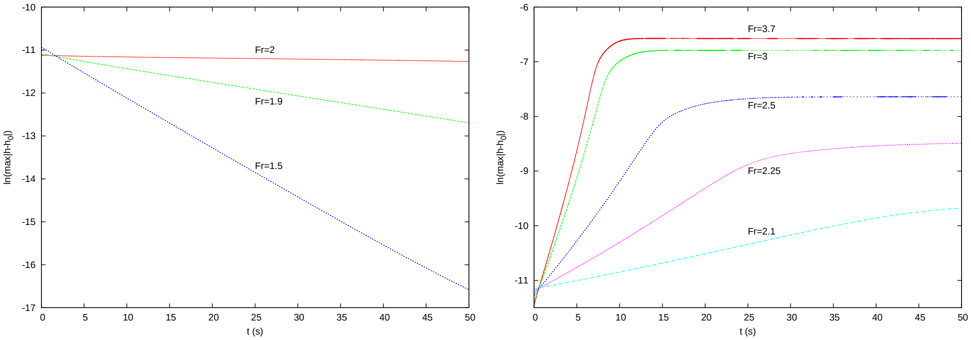


Figure 8: Convergence rate to the final state, Fr≤ 2 (left), Fr>2 (right).

roll-wave appears, whose amplitude depends on the initial state (see fig. 8).

Conclusion

This preliminary study of overland flow due to rainfall events clearly enlights several specific difficulties. First, from the numerical point of view, it seems that the apparent topography method, which was designed in order to catch steady states, is not adapted for wet/dry transitions. The semi-implicit treatment seems to be better in the problems we consider and gives good results compared to experimental data. Next, the model itself has to be improved, in particular regarding the empirical friction laws we used, which were not developed in this hydrological context. Finally, more realistic situations require infiltration and two-dimensional simulations, which are in progress and already validated on analytical solutions. This will be again compared with experimental data, as for the flume test.

Acknowledgements

This work was partially supported by the ANR grant "METHODE" #ANR-07-BLAN-0232.

References

[1] Audusse, E. *Modélisation hyperbolique et analyse numérique pour les écoulements en eaux peu profondes*. PhD thesis, Université Paris VI – Pierre et Marie Curie, 14Sept. 2004. 196 pages.

[2] Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., and Perthame, B. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput. 25*, 6 (2004), 2050–2065. doi:10.1137/S1064827503431090.

[3] Bouchut, F. *Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources*, vol. 2/2004. Birkhäuser Basel, 2004. doi:10.1007/b95203.

[4] Bristeau, M.-O., and Coussin, B. Boundary conditions for the shallow water equations solved by kinetic schemes. Tech. Rep. 4282, INRIA, Oct. 2001.

[5] Delestre, O., Cordier, S., James, F., and Darboux, F. Simulation of Rain-Water Overland-Flow. MSC 76M12,74G05,74G75,35L65,20C20. Available from: `http://hal.archives-ouvertes.fr/hal-00343721/en/`.

[6] Esteves, M., Faucher, X., Galle, S., and Vauclin, M. Overland flow and infiltration modelling for small plots during unsteady rain : numerical results versus observed values. *Journal of Hydrology 228* (2000), 265–282.

[7] Fiedler, R. F., and Ramirez, J. A. A numerical method for simulating discontinuous shallow flow over an infiltrating surface. *International Journal for Numerical Methods in Fluids 32* (2000), 219–240.

[8] MacDonald, I., Baines, M. J., Nichols, N. K., and Samuels, P. G. Steady open channel test problems with analytic solutions. Tech. Rep. 3, Department of Mathematics–University of Reading, 1995.

[9] MacDonald, I., Baines, M. J., Nichols, N. K., and Samuels, P. G. Analytic benchmark solutions for open-channel flows. *Journal of Hydraulic Engineering 123*, 11 (November 1997), 1041–1045.

[10] Que, Y.-T., and Xu, K. The numerical study of roll-waves in inclined open channels and solitary wave run-up. *Internat. J. Numer. Methods Fluids 50*, 9 (2006), 1003–1027.

Olivier Delestre and François James
Mathématiques, Applications et Physique Mathématique d'Orléans - UMR CNRS 6628 – Fédération
Denis Poisson - FR CNRS 2964 Université d'Orléans – BP 6759
45067 Orléans cedex 2 – FRANCE
`olivier.delestre@etu.univ-orleans.fr` and `francois.james@univ-orleans.fr`

Monografías Matemáticas *García de Galdeano* **35**, 137–144 (2010)

Newton-like methods for operators with bounded second Fréchet derivative

J. A. Ezquerro, M. A. Hernández and N. Romero

Abstract. We present a new semilocal convergence result for a family of Newton-like methods that generalizes the usual iterations of third order of convergence in Banach spaces. To do this, fewer convergence conditions are required than all the existing ones until now. We apply this analysis to a nonlinear integral equation of mixed Hammerstein type.

Keywords: Nonlinear equations in Banach spaces, iterative methods, semilocal convergence theorem, nonlinear integral equation.

AMS classification: AMS 45G10, 47H99, 65J15.

§1. Introduction

In order to solve nonlinear equations in Banach spaces, $F(x) = 0$, where $F : \Omega \subseteq X \to Y$, X, Y Banach spaces and Ω a nonempty open convex subset, we consider in this paper the family of Newton-like iterations of third order of convergence, which is based on the following algorithm, [4]:

$$\begin{cases} x_{n+1} = x_n - H(L_F(x_n))[F'(x_n)]^{-1}F(x_n), & n \geq 0, \\ H(L_F(x)) = I + \dfrac{1}{2}L_F(x) + \displaystyle\sum_{k\geq 2} A_k L_F(x)^k, & A_k \in \mathbb{R}^+, \quad k \geq 2, \end{cases} \tag{1}$$

where $x_0 \in \Omega$, I is the identity operator on X, $L_F(x)$ is the degree of logarithmic convexity defined by $L_F(x) = F'(x)^{-1}F''(x)\left[F'(x)^{-1}F(x)\right] \in \mathcal{L}(X)$ [5], where $\mathcal{L}(X)$ is the set of bounded linear operators from X into X, provided that $[F'(x)]^{-1}$ exists at each step x_n, and

$$\sum_{k\geq 0} A_k t^k < +\infty, \qquad \text{for} \quad |t| < r. \tag{2}$$

The operators $F'(x)$ and $F''(x)$ denote respectively the first and the second Fréchet-derivatives of the operator F.

The convergence of third-order methods has been examined extensively by several authors, and most of them can be written from (1), see [4]. Basic results concerning the convergence of these methods have been published under assumptions of Newton-Kantorovich type. For example, Safiev presents in [6] a convergence theorem for the Halley method under the following conditions:

(C1) There exists a point $x_0 \in \Omega$ where the operator $\Gamma_0 = [F'(x_0)]^{-1} \in \mathcal{L}(Y, X)$ is defined and $\|\Gamma_0\| \leq \beta$,

(C2) $\|\Gamma_0 F(x_0)\| \leq \eta$,

(C3) $\|F''(x)\| \leq M, \quad \forall x \in \Omega$,

(C4) $\|F'''(x)\| \leq N, \quad \forall x \in \Omega$.

The convergence conditions given by Safiev can be modified by replacing the strongest one $\|F'''(x)\| \leq N$, $x \in \Omega$, by

$$\left\|F''(x) - F''(y)\right\| \leq K \|x - y\|, \quad K \geq 0, \quad x, y \in \Omega, \tag{3}$$

or the milder one

$$\left\|F''(x) - F''(y)\right\| \leq L \|x - y\|^p, \quad L \geq 0, \quad p \in [0, 1], \quad x, y \in \Omega. \tag{4}$$

(See [1]). These two last conditions mean that F'' is Lipschitz continuous in Ω and F'' is (L, p)-Hölder continuous in Ω, respectively.

According to the above, the number of equations that can be solved is limited. For instance, we cannot analyze the convergence of iteration (1) to a solution of equations where sums of operators, which satisfy (3) or (4), are involved, as for example in the following nonlinear integral operator of mixed Hammerstein type [3]

$$F(x)(s) = x(s) + \sum_{i=1}^{m} \int_a^b k_i(s, t)\ell_i(x(t))\, dt - u(s), \quad s \in [a, b], \tag{5}$$

where $-\infty < a < b < \infty$, u, ℓ_i, k_i, for $i = 1, 2, \ldots, m$, are known functions, $\ell_i''(x(t))$ is (L_i, p_i)-Hölder continuous in Ω, for $i = 1, 2, \ldots, m$, and x is a solution to be determined. This types of operator appear, for instance, in dynamic models of chemical reactors, see [2].

In order to consider more general situation as the aforesaid we relax the previous convergence conditions. We then reconsider the convergence of (1) in Banach spaces by assuming only (C1), (C2) and (C3), so that fewer convergence conditions are required.

The paper follows this scheme. In Section 2, we present a technique based on recurrence relations to establish the semilocal convergence of iterative processes (1) to a solution of the equation $F(x) = 0$. Besides, we establish domains of existence and uniqueness of solution. In Section 3, we apply the previous analysis to the following nonlinear equation, where an operator of type (5) is involved:

$$F(x)(s) = x(s) - 1 - \int_0^1 G(s, t)\left(x(t)^{5/2} + x(t)^3/5\right) dt = 0, \quad s \in [0, 1], \tag{6}$$

where $x \in C[0, 1]$, $s, t \in [0, 1]$, and the kernel G is the Green function

$$G(s, t) = \begin{cases} (1 - s)t, & t \leq s, \\ s(1 - t), & s \leq t. \end{cases}$$

Throughout the paper we denote

$$\overline{B(x, r)} = \{y \in X; \|y - x\| \leq r\} \quad \text{and} \quad B(x, r) = \{y \in X; \|y - x\| < r\}.$$

§2. Analysis of the convergence and order of convergence

In this section, we present two analysis of the convergence of (1). In the first one, we guarantee that the order of convergence is locally at least three. In the second one, a more general analysis of the semilocal convergence is given, where conditions (C1)–(C3) are only required.

2.1. Local convergence

Next result proves that the order of convergence of iterations (1) is locally at least three.

Theorem 1. *Suppose that $F : \Omega \subseteq X \to Y$ is a sufficiently differentiable operator on a non-empty open convex domain Ω. If F has a simple root $x^* \in \Omega$, $[F'(x)]^{-1}$ exists in a neighborhood of x^* and x_0 is sufficiently close to x^*, then iterations (1) have order of convergence at least three.*

2.2. Semilocal convergence

We consider the problem of approximating a solution x^* of a nonlinear equation

$$F(x) = 0, \tag{7}$$

with $F : \Omega \subseteq X \to Y$, X, Y Banach spaces, Ω a nonempty open convex subset and we assume that F is a twice continuously Fréchet-differentiable operator satisfying conditions (C1), (C2) and (C3). Using a technique based on recurrence relations, we prove that under conditions (C1)–(C3), iterative processes (1) are convergent to a solution of (7). Also, we find the domains where the solution is located and unique.

2.2.1. A system of recurrence relations

Firstly, we construct a system of recurrence relations which allows to establish the convergence of iterations (1).

Notice that, from initial conditions (C1)–(C3), we have

$$\|L_F(x_0)\| \leq \|\Gamma_0\|\,\|F''(x_0)\|\,\|\Gamma_0 F(x_0)\| \leq M\beta\eta.$$

We then define the parameter $a_0 = M\beta\eta$. If $a_0 < r$, where r is the radius of convergence of series (2), then x_1 is well defined, since $H(L_F(x_0))$ exists and

$$\left\|H(L_F(x_0))\right\| \leq \sum_{k\geq 0} \left\|A_k L_F(x_0)^k\right\| \leq \sum_{k\geq 0} A_k a_0^k.$$

Moreover,

$$\|x_1 - x_0\| \leq \left\|H(L_F(x_0))\right\| \|\Gamma_0 F(x_0)\| \leq \left(\sum_{k\geq 0} A_k a_0^k\right) \|\Gamma_0 F(x_0)\|.$$

Throughout this section we consider the auxiliary real functions

$$h(t) = 1 + \frac{1}{2}t + \sum_{k\geq 2} A_k t^k, \quad f(t) = \frac{1}{1 - th(t)}, \quad g(t) = h(t)\Big(1 + \frac{t}{2}h(t)\Big) - 1, \tag{8}$$

and, from a_0, the following sequence is constructed

$$a_{n+1} = a_n f(a_n)^2 g(a_n). \tag{9}$$

Next, we establish the recurrence relations that allow to prove the convergence of iterations (1) under conditions (C1)–(C3). The proofs of the recurrence relations follow as in [4].

Lemma 2. *Let us suppose that x_0, $x_n \in \Omega$, for $n \in \mathbb{N}$. If $a_0 < r$, $a_0 h(a_0) < 1$, and $a_0 f(a_0)^2 g(a_0) < 1$, then the following relations are verified:*

(I) *There exists* $\Gamma_n = [F'(x_n)]^{-1}$ *and* $\|\Gamma_n\| \leq f(a_{n-1})\,\|\Gamma_{n-1}\|$,

(II) $\|\Gamma_n F(x_n)\| \leq f(a_{n-1}) g(a_{n-1}) \|\Gamma_{n-1} F(x_{n-1})\|$,

(III) $\|\Gamma_n\|\,\|F''(x_n)\|\,\|\Gamma_n F(x_n)\| \leq a_n$ *and there exists* $H(L_F(x_n))$,

(IV) $\|x_{n+1} - x_n\| \leq h(a_n)\,\|\Gamma_n F(x_n)\|$,

(V) $\|x_{n+1} - x_0\| \leq h(a_0)\left(\sum_{k=0}^{n} (f(a_0)g(a_0))^k\right)\|\Gamma_0 F(x_0)\|$.

2.2.2. Main result

Next, we give some results where some properties of real functions (8) and real sequence $\{a_n\}$ are shown, from which we establish the semilocal convergence of third-order iterations (1).

Lemma 3. *Let h, f and g be the real functions given in (8). If $a_0 h(a_0) < 1$ and $f(a_0)^2 g(a_0) < 1$, then $\{a_n\}$, given in (9), is a decreasing sequence. Moreover, $f(a_0)g(a_0) < 1$.*

Lemma 4. *Let h, f and g be the real functions given in (8) and $\gamma = a_1/a_0$. If $a_0 h(a_0) < 1$ and $f(a_0)^2 g(a_0) < 1$. Then, the following properties are satisfied:*

(i) *Sequence $\{a_n\}$, given in (9), is decreasing,*

(ii) $\gamma = f(a_0)^2 g(a_0) \in (0,1)$,

(iii) $f(\gamma t) < f(t)$, $g(\gamma t) < \gamma g(t)$, $\forall t > 0$ *and* $\forall \gamma \in (0,1)$.

(iv) $a_{n+1} < \gamma^{2^n} a_n < \gamma^{2^{n+1}-1} a_0$, $\forall n \geq 1$,

(v) $f(a_{n+1}) g(a_{n+1}) < \gamma^{2^{n+1}} \Delta$, *where* $\Delta = 1/f(a_0)$, $\forall n \geq 1$.

After that, we suppose now that $\Gamma_0 = F'(x_0)^{-1} \in \mathcal{L}(Y, X)$ exists at some $x_0 \in \Omega$, where $\mathcal{L}(Y, X)$ is the set of bounded linear operators from Y into X. The following semilocal convergence result is then obtained.

Theorem 5. *Let $F : \Omega \subseteq X \to Y$ be a twice continuously differentiable operator on a non-empty open convex domain Ω satisfying conditions (C1)–(C3). If conditions $a_0 < r$, $a_0 h(a_0) < 1$, $f(a_0)^2 g(a_0) < 1$ and $B(x_0, R) \subset \Omega$, where $R = \frac{h(a_0)\eta}{1 - f(a_0)g(a_0)}$, are satisfied, then iterations (1), starting from x_0, generate a sequence $\{x_n\}$ that converges to a solution $x^* \in \overline{B(x_0, R)}$ of (7). Moreover, x^* is unique in $\Omega_0 = B(x_0, \frac{2}{M\beta} - R) \cap \Omega$.*

Proof. By hypotheses, we have $x_0 \in \Omega$ and, taking into account (C1)–(C3), we obtain $x_1 \in \overline{B(x_0, R)} \subset \Omega$. Besides, from lemma 2 and by following an inductive procedure, we obtain $x_n \in \overline{B(x_0, R)} \subset \Omega$, for all $n \in \mathbb{N}$. Now, from recurrence relations given in Lemma 2, it follows

$$\|x_{k+1} - x_k\| \leq h(a_k)\,\|\Gamma_k F(x_k)\| < h\Big(\gamma^{2^k - 1} a_0\Big)\eta\, \gamma^{2^k - 1} \Delta^k,$$

where $\Delta = 1/f(a_0)$.

To establish the convergence of iterations (1) we prove that the sequence $\{x_n\}$ is a Cauchy one. To do this, we consider $n, m \in \mathbb{N}$, and

$$\|x_{n+m} - x_n\| \leq \sum_{k=n}^{n+m-1} \|x_{k+1} - x_k\| < h\Big(\gamma^{2^n-1} a_0\Big)\eta\, \gamma^{2^n-1}\Delta^n \left(\sum_{k=0}^{m-1} \gamma^{2^n[2^k-1]}\Delta^k\right).$$

Now, by applying Bernoulli's inequality, $(1 + x)^k > 1 + kx$, with $x = 1$, we have $2^k - 1 > k$ and therefore

$$\|x_{n+m} - x_n\| < h\Big(\gamma^{2^n-1} a_0\Big)\eta\, \gamma^{2^n-1}\Delta^n \frac{1 - \left(\gamma^{2^n}\Delta\right)^m}{1 - \gamma^{2^n}\Delta}, \tag{10}$$

for all $n, m \in \mathbb{N}$ and $\{x_n\}$ is then a Cauchy sequence.

Taking $n = 0$ in (10), it follows $x_m \in B(x_0, R)$, for all $m \in \mathbb{N}$, since

$$\|x_m - x_0\| < h(a_0)\frac{1 - \left(\gamma\Delta\right)^m}{1 - \gamma\Delta}\eta < R.$$

Besides, $\{x_n\}$ converges to the solution $x^* = \lim\limits_{n\to\infty} x_n$ of (7) and, by taking $n = 0$ and $m \to \infty$ in (10), it follows

$$\|x^* - x_0\| \leq h(a_0)\ \frac{\eta}{1 - f(a_0)g(a_0)} = R$$

and $x^* \in \overline{B(x_0, R)}$. On the other hand, we notice that $\{\|F'(x_n)\|\}$ is a bounded sequence, since

$$\left\|F'(x_n)\right\| \leq Mh(a_0)\left(\sum_{k=0}^{n-1}(f(a_0)g(a_0))^k\right)\eta + \|F'(x_0)\|.$$

Thus, from $\|F(x_n)\| \leq \|F'(x_n)\|\,\|\Gamma_n F(x_n)\|$ and $\lim_{n\to\infty}\|\Gamma_n F(x_n)\| = 0$, we have that $\lim_{n\to\infty}\|F(x_n)\| = 0$. By the continuity of F, we obtain that $F(x^*) = 0$ and x^* is therefore a solution of equation (7).

To prove the uniqueness of x^*, we follows the same procedure as in [4]. □

§3. Application

We now illustrate the previous study with an application where the following nonlinear integral operator of mixed Hammerstein type (6) is involved. Note that (6) is such that $F : C^+[0,1] \subseteq C[0,1] \to C[0,1]$, where $C^+[0,1] = \{x \in C[0,1] \mid x(t) \geq 0, t \in [0,1]\}$. Observe that the second Fréchet derivatives of operator (6) is

$$[F''(x)yz](s) = -\int_0^1 G(s,t)\left(\frac{15}{4}x(t)^{1/2} + \frac{6}{5}x(t)\right)z(t)y(t)\,dt. \tag{11}$$

Notice that F'' does nor satisfy (3) neither (4), but the conditions of Theorem 5 does, so that a solution of the corresponding equation $F(x) = 0$ can be approximated by any iteration (1).

	Chebyshev	Super-Halley	Halley
R	$0.449411\ldots$	$0.534408\ldots$	$0.482791\ldots$
$\frac{2}{M\beta}-R$	$1.53039\ldots$	$1.44539\ldots$	$1.49701\ldots$

Table 1: Ratios of existence and uniqueness of solution of the Hammerstein equation

We use the max-norm and take into account that a solution x^* of (6) in $C^+[0,1]$ must satisfy

$$\|x^*\| - \frac{\|x^*\|^{5/2}}{8} - \frac{\|x^*\|^3}{40} - 1 \leq 0,$$

i.e., $\|x^*\| \leq \rho_1 = 1.28982\ldots$ or $\|x^*\| \geq \rho_2 = 2.28537\ldots$, where ρ_1 and ρ_2 are the positive roots of the real equation $z - z^{5/2}/8 - z^3/40 - 1 = 0$. Taking now into account (6), it is needed that $x^*(s) \geq 0$, then if we look for a solution such that $\|x^*\| < \rho_1$, we can consider for example $\Omega = B(x_0(s), 1) \subseteq C^+[0,1]$ as a non-empty open convex domain.

If we choose $x_0(s) = 1$, then $\|I - F'(x_0)\| \leq 31/80 < 1$, Γ_0 is defined, $\|\Gamma_0\| \leq 80/49$ and $\|\Gamma_0 F(x_0)\| \leq 12/49 = \eta$. From (11), it follows

$$\left\|F''(x)\right\| \leq 99/160 = M, \quad \forall x \in \Omega, \quad \text{and} \quad a_0 = 0.247397\ldots.$$

Now, by applying Theorem 5, the convergence of iterations (1) to a solution x^* of equation (6) is guaranteed, starting from $x_0(s) = 1$. Notice that the domains of existence,

$$\{u \in C[0,1] \mid \|u-1\| \leq R\} \subseteq \Omega,$$

and uniqueness,

$$\left\{u \in C[0,1] \,\middle|\, \|u-1\| < \frac{2}{M\beta} - R\right\},$$

of solution depend on the corresponding iteration (1) applied to solve the equation. For example, if (1) is reduced to the Chebyshev ($A_k = 0$ for all $k \geq 2$), Super-Halley ($A_k = 1/2$ for all $k \geq 2$) and Halley ($A_k = 1/2^k$ for all $k \geq 2$) methods, the radii of these domains appear in Table 1. To obtain a numerical solution of $F(x) = 0$, we first discretize the problem and approach the integral by a Gauss-Legendre quadrature formula with 8 knots. If we denote $x_i = x(t_i)$, $i = 1, 2, \ldots, 8$, equation $F(x) = 0$ becomes the following nonlinear system:

$$x_i = 1 + \frac{1}{2}\sum_{j=1}^{8} \beta_j G(t_i, t_j)\left(x_j^{\frac{5}{2}} + \frac{x_j^3}{5}\right), \quad i = 1, 2, \ldots 8,$$

where the weights β_j and the knots t_j are known. Now, if we denote

$$a_{ij} = \begin{cases} \frac{1}{2}(1-t_i)t_j\beta_j, & j \leq i, \\ \frac{1}{2}t_i(1-t_j)\beta_j, & i \leq j, \end{cases}$$

then we can write the previous nonlinear system as

$$x_i = 1 + \sum_{j=1}^{8} a_{ij}\left(x_j^{\frac{5}{2}} + \frac{x_j^3}{5}\right), \quad i = 1, 2, \ldots, 8. \tag{12}$$

x_1^*	1.01779...	x_3^*	1.17636...	x_5^*	1.24150...	x_7^*	1.08589...
x_2^*	1.08589...	x_4^*	1.24150...	x_6^*	1.17636...	x_8^*	1.01779...

Table 2: Numerical solution $\overline{x}^*$ of equation (13)

n	Chebyshev-like	Super-Halley	Halley
0	$2.41500\ldots \times 10^{-1}$	$2.41500\ldots \times 10^{-1}$	$2.41500\ldots \times 10^{-1}$
1	$1.71318\ldots \times 10^{-3}$	$1.26433\ldots \times 10^{-3}$	$2.87962\ldots \times 10^{-3}$
2	$3.71075\ldots \times 10^{-10}$	$1.47479\ldots \times 10^{-10}$	$6.76345\ldots \times 10^{-9}$

Table 3: Errors for the Chebyshev-like, Super-Halley and Halley methods

Moreover, denoting $\overline{x} = (x_1, \ldots, x_8)^T$, $\overline{1} = (1, \ldots, 1)^T$ and $A = (a_{ij})$, we write nonlinear system (12) in the matrix form:

$$F(\overline{x}) = \overline{x} - \overline{1} - A\left(\overline{x}^{5/2} + \overline{x}^3/5\right) = \overline{0}. \tag{13}$$

Therefore,

$$F'(\overline{x})(\overline{y}) = \left[I - A\left(\frac{5}{2}D_{3/2}(\overline{x}) + \frac{3}{5}D_2(\overline{x})\right)\right]\overline{y}, \quad \forall \overline{y} \in \mathbb{R}^8,$$

where $D_k(\overline{x})$ denotes the diagonal matrix with the components of the vector $(x_1^k, x_2^k, \ldots x_n^k)$ in the diagonal, and F'' is the bilinear operator defined by

$$F''(\overline{x})(\overline{y}, \overline{z}) = -\frac{3}{20}A\left((25x_1^{1/2} + 8x_1)z_1y_1, \ldots, (25x_8^{1/2} + 8x_8)z_8y_8\right)^t, \quad \forall \overline{y}, \overline{z} \in \mathbb{R}^8.$$

To find an approximation of a solution of equation (13), we use for example the Chebyshev-like method given by (1), with $A_2 = 1/2$ and $A_k = 0$, for all $k > 2$.

On the other hand, we denote the n-th iteration by $\overline{x}_n = (x_1^{(n)}, x_2^{(n)}, \ldots, x_8^{(n)})^t$. If we choose $x_i^{(0)} = 1$, for $i = 1, 2, \ldots, 8$, after six iterations applying Chebyshev-like method and using stopping criteron $\left\|\overline{x}_n - \overline{x}^*\right\| < 10^{-15}$, we obtain the numerical solution $\overline{x}^* = (x_1^*, x_2^*, \ldots, x_8^*)$ of (13) given in Table 2, and considering the same stopping criterion we obtain the errors $\left\|\overline{x}_n - \overline{x}^*\right\|$, which appear in Table 3 when (1) is reduced to the Chebyshev-like, Super-Halley and Halley methods.

Acknowledgements

Preparation of this paper is partly supported by the Ministry of Education and Science (MTM 2008-01952) and the Government of the Riojan Autonomous Community (Colabora 2009/04).

References

[1] Argyros, I. K. The Halley method in Banach spaces and the Ptak error estimates. *Rev. Acad. Cienc. Zaragoza 2* (1997), 31–41.

[2] Bruns, D. D., and Bailey, J. E. Nonlinear feedback control for operating a nonisothermal CSTR near an unstable steady state. *Chem. Eng. Sci. 32* (1977), 257–264.

[3] Ganesh, M., and Joshi, M. C. Numerical solvability of Hammerstein integral equations of mixed type. *IMA J. Numer. Anal. 11* (1991), 21–31.

[4] Hernández, M. A., and Romero, N. On a chatacterization of some Newton-like methods of *R*-order at least three. *J. Comput. Appl. Math. 183* (2005), 53–66.

[5] Hernández, M. A., and Salanova, M. A. Indices of convexity and concavity: Application to Halley method. *Appl. Math. Comput. 103* (1999), 27–49.

[6] Safiev, R. A. On some iterative processes (in russian). *Zh. Vychisl. Mat. i Mat. Fiz., (Translated into English by L.B. Rall as MRC Technical Summary Report, No. 649, Univ. Wisconsin-Madison, 1966) 4* (1964), 139–143.

José Antonio Ezquerro, Miguel Ángel Hernández and Natalia Romero
University of La Rioja.
Departamento de Matemáticas y Computación.
C/ Luis de Ulloa s/n.
26004 Logroño. Spain.
jezquer@unirioja.es, mahernan@unirioja.es and natalia.romero@unirioja.es

Monografías Matemáticas *García de Galdeano* **35**, 145–153 (2010)

Development of efficient geometric multigrid algorithms by LFA for systems of partial differential equations on triangular grids

F. J. Gaspar, J. L. Gracia, F. J. Lisbona and C. Rodrigo

Abstract. In this paper Local Fourier Analysis (LFA) for multigrid methods on triangular grids is extended to the case of systems of PDEs. In particular, it is performed for the problem of planar elasticity, although its application to other systems is straightforward. Analogously to the scalar case, this analysis is based on an expression of the Fourier transform in new coordinate systems, both in space and in frequency variables, associated with reciprocal bases. LFA is particularly valuable for systems of PDEs, since it is often much more difficult to identify the correct multigrid components than for a scalar problem. For the discrete elasticity operator obtained with linear finite elements, different collective smoothers like three-color smoother and some zebra-type smoothers are analyzed. LFA results for these smoothers are presented.

Keywords: Geometric multigrid, Fourier analysis, three-color smoother, triangular grids, elasticity.

AMS classification: 65N55, 65F10, 65N30.

§1. Introduction

Planar elasticity models the displacements of an elastic body $\Omega \subset \mathbb{R}^2$, subject to a force density $\mathbf{f}$, with respect to its original configuration. These displacements are described by means of a vector function $\mathbf{u} = (u, v)$, which is the solution of the following system of equations

$$\mathbf{L}\,\mathbf{u} = -\mu\boldsymbol{\Delta}\mathbf{u} - (\lambda + \mu)\mathrm{grad}(\mathrm{div}\,\mathbf{u}) = \mathbf{f}, \quad \text{in } \Omega,$$

where $\boldsymbol{\Delta}$ is the vector Laplace operator, λ and μ are the so-called Lamè's coefficients, and $\mathbf{f} = (f_1, f_2) \in (L^2(\Omega))^2$. Here, a discretization by linear finite elements of this elasticity operator is considered,

$$\mathbf{L}_h = \begin{pmatrix} L_h^{u,u} & L_h^{u,v} \\ L_h^{v,u} & L_h^{v,v} \end{pmatrix} = \begin{pmatrix} -(\lambda + 2\mu)(\partial_{xx})_h - \mu(\partial_{yy})_h & -(\lambda+\mu)(\partial_{xy})_h \\ -(\lambda+\mu)(\partial_{xy})_h & -\mu(\partial_{xx})_h - (\lambda + 2\mu)(\partial_{yy})_h \end{pmatrix}.$$

The algebraic linear equation system arising from this discretization will be solved by means of a geometric multigrid algorithm, due to the fact that these methods are among the most efficient numerical algorithms for solving this kind of systems. In geometric multigrid, a hierarchy of grids must be proposed. For an irregular domain, it is very common to apply regular refinement to an unstructured input grid; in this way, a hierarchy of globally unstructured grids is generated that is suitable for use with geometric multigrid. So, we are interested in the framework of hierarchical hybrid grids (HHG) which was presented in [1]. The coarsest mesh is assumed rough enough in order to fit the geometry of the domain. Once this coarse triangulation is given, each triangle is divided into four congruent triangles connecting the midpoints of their edges, and so forth until the mesh has the desired fine scale to approximate the solution of the problem. In this way, a nested hierarchy of grids is obtained.

As it is well-known, the construction of an efficient multigrid method is strongly dependent on the choice of its components, which have to be selected so that they efficiently interplay with each other. Especially, the choice of a suitable smoother is an important feature for the design of an efficient multigrid method. In this paper, linear interpolation has been chosen, the restriction operator has been taken as its adjoint and the discrete operator corresponding to each mesh results from the direct discretization of the problem. Moreover, collective three-color smoother and some collective line-wise smoothers of zebra-type are proposed as relaxing methods.

In order to choose suitable components for a multigrid method, LFA is used, due to its being a powerful tool for the design of efficient multigrid methods. This analysis is mainly based on the Fourier transform and was introduced by Brandt [2]. A good introduction can be found in the books by Trottenberg et al. [4] and Wienands and Joppich [5]. This technique has been widely used in the framework of discretizations on rectangular grids, and recently a generalization to triangular grids has been proposed in [3]. The key fact for carrying out this generalization is to write the Fourier transform using coordinates in non-orthogonal bases fitting the new structure of the grid. In order to extend LFA to the case of the planar elasticity system, a new expression of the Fourier transform for vector functions is considered. To study multigrid methods in the framework of HHG, the LFA proposed here is applied to each input triangle in such a way that the global behavior of the method will depend on the quality of the chosen local components.

The organization of the paper is as follows. In Section 2, the way in which LFA can be performed on non-orthogonal grids for systems of PDEs is explained. In section 3, the relaxation methods considered, three-color and zebra-type smoothers are presented. Finally, in Section 4, some LFA results are shown in order to choose the relaxation method of the multigrid algorithm which is more suitable for different grid geometries.

§2. Fourier analysis on non-orthogonal grids

A non-orthogonal unitary basis of $\mathbb{R}^2$ is established: $\{\mathbf{e}'_1, \mathbf{e}'_2\}$ with $0 < \gamma < \pi$ being the angle between the vectors of the basis. It is also considered its reciprocal basis $\{\mathbf{e}''_1, \mathbf{e}''_2\}$, i.e., $(\mathbf{e}'_i, \mathbf{e}''_j) = \delta_{ij}$, $1 \le i, j \le 2$, where $(\cdot, \cdot)$ is the usual inner product in $\mathbb{R}^2$ and δ_{ij} is the Kronecker's delta, see Figure 1. The coordinates of a point in these bases, $\{\mathbf{e}'_1, \mathbf{e}'_2\}$ and $\{\mathbf{e}''_1, \mathbf{e}''_2\}$, are $\mathbf{y}' = (y'_1, y'_2)$ and $\mathbf{y}'' = (y''_1, y''_2)$, respectively, just like $\mathbf{y} = (y_1, y_2)$ in the canonical basis $\{\mathbf{e}_1, \mathbf{e}_2\}$.

By applying the changes of variables $\mathbf{x} = \mathbf{F}(\mathbf{x}')$ and $\boldsymbol{\theta} = \mathbf{G}(\boldsymbol{\theta}'')$ to the usual Fourier transform formula, the Fourier transform and its corresponding back transformation formula

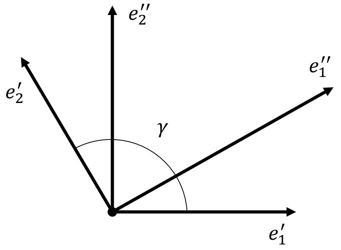


Figure 1: Reciprocal bases in $\mathbb{R}^2$.

with coordinates in a non-orthogonal basis, result in the following

$$\hat{\mathbf{u}}(\mathbf{G}(\boldsymbol{\theta}'')) = \frac{\sin\gamma}{2\pi}\int_{\mathbb{R}^2} e^{-i\mathbf{G}(\boldsymbol{\theta}'')\cdot\mathbf{F}(\mathbf{x}')}\mathbf{u}(\mathbf{F}(\mathbf{x}'))\,\mathrm{d}\mathbf{x}',$$
$$\mathbf{u}(\mathbf{F}(\mathbf{x}')) = \frac{1}{2\pi\sin\gamma}\int_{\mathbb{R}^2} e^{i\mathbf{G}(\boldsymbol{\theta}'')\cdot\mathbf{F}(\mathbf{x}')}\hat{\mathbf{u}}(\mathbf{G}(\boldsymbol{\theta}''))\,\mathrm{d}\boldsymbol{\theta}''.$$

Since the new bases are reciprocal bases, the inner product $\mathbf{G}(\boldsymbol{\theta}'')\cdot\mathbf{F}(\mathbf{x}')$ is given by $\theta_1''x_1' + \theta_2''x_2'$. Using previous expressions, a discrete Fourier transform for non-rectangular grids can be introduced. With this purpose, a uniform infinite grid is defined:

$$G_h = \{\mathbf{x}' = (x_1', x_2') \mid x_i' = k_i h_i,\ k_i \in \mathbb{Z},\ i = 1,2\},$$

where $\mathbf{h} = (h_1, h_2)$ is a grid spacing. Now, for a vector grid function $\mathbf{u}_h$, the discrete Fourier transform and its back Fourier transformation can be defined by

$$\begin{aligned}\hat{\mathbf{u}}_h(\boldsymbol{\theta}'') &= \frac{h_1h_2\sin\gamma}{2\pi}\sum_{\mathbf{x}'\in G_h} e^{-i(\theta_1''x_1'+\theta_2''x_2')}\mathbf{u}_h(\mathbf{x}'),\\ \mathbf{u}_h(\mathbf{x}') &= \frac{1}{2\pi\sin\gamma}\int_{\boldsymbol{\Theta}_h} e^{i(\theta_1''x_1'+\theta_2''x_2')}\hat{\mathbf{u}}_h(\boldsymbol{\theta}'')\mathrm{d}\boldsymbol{\theta}'',\end{aligned}\tag{1}$$

where $\boldsymbol{\theta}'' = (\theta_1'', \theta_2'') \in \boldsymbol{\Theta}_h = (-\pi/h_1, \pi/h_1] \times (-\pi/h_2, \pi/h_2]$ are the coordinates of the point $\theta_1''\mathbf{e}_1'' + \theta_2''\mathbf{e}_2''$ in the frequency space. Considering the scalar Fourier modes, $\varphi_h(\boldsymbol{\theta}'', \mathbf{x}') = \mathrm{e}^{i\theta_1''x_1'}\,\mathrm{e}^{i\theta_2''x_2'}$, their vector counterparts are $\boldsymbol{\varphi}_h(\boldsymbol{\theta}'', \mathbf{x}') := (\varphi_h(\boldsymbol{\theta}'', \mathbf{x}'), \varphi_h(\boldsymbol{\theta}'', \mathbf{x}'))^t$, with $\mathbf{x}' \in G_h$, and $\boldsymbol{\theta}'' \in \boldsymbol{\Theta}_h$. They give rise to the Fourier space, $\mathcal{F}(G_h) = \mathrm{span}\{\boldsymbol{\varphi}_h(\boldsymbol{\theta}'', \cdot) \mid \boldsymbol{\theta}'' \in \boldsymbol{\Theta}_h\}$. From (1), it follows that each discrete function $\mathbf{u}_h(\mathbf{x}') \in (l_h^2(G_h))^2$ can be written as a formal linear combination of the Fourier modes, which are linearly independent discrete functions.

Due to the fact that the grid and the frequency space are referred to as reciprocal bases, the Fourier modes have a formal expression, in terms of $\boldsymbol{\theta}''$ and $\mathbf{x}'$, similar to those in Cartesian coordinates. Therefore, the Local Fourier analysis on non-rectangular grids can be performed straightforwardly.

Let $\mathcal{T}_h$ be a regular triangular grid on a fixed coarse triangle $\mathcal{T}$; see left picture of Figure 2. $\mathcal{T}_h$ is extended to the infinite grid G_h given before, where $\mathbf{e}_1'$ and $\mathbf{e}_2'$ are unit vectors indicating

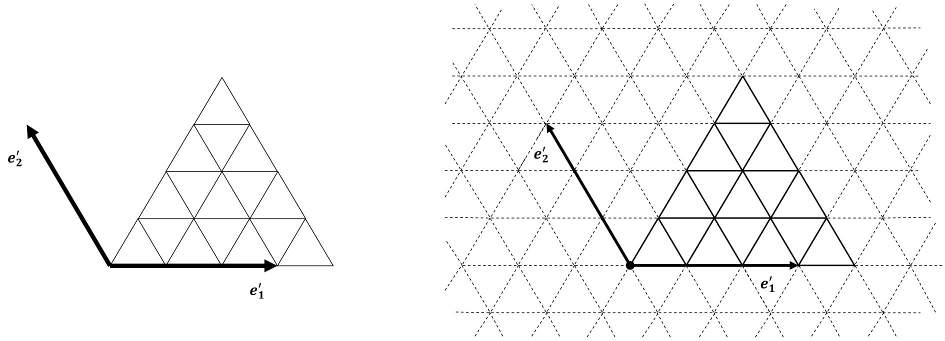


Figure 2: Regular triangular grid on a fixed coarse triangle $\mathcal{T}$ and its extension to an infinite grid.

the direction of two of the edges of $\mathcal{T}$, and such that $\mathcal{T}_h = G_h \bigcap \mathcal{T}$, see right picture of Figure 2. Neglecting boundary conditions and/or connections with other neighboring triangles of the coarsest grid, the discrete problem $\mathbf{L}_h\mathbf{u}_h = \mathbf{f}_h$ can be extended to the whole grid G_h. As it is well-known, vector Fourier modes $\boldsymbol{\varphi}_h(\boldsymbol{\theta}'', \mathbf{x}')$ are formal eigenfunctions of the discrete operator $\mathbf{L}_h$. More precisely, it is fulfilled

$$\mathbf{L}_h\boldsymbol{\varphi}_h(\boldsymbol{\theta}'', \mathbf{x}') = \widetilde{\mathbf{L}}_h(\boldsymbol{\theta}'')\boldsymbol{\varphi}_h(\boldsymbol{\theta}'', \mathbf{x}'), \quad \widetilde{\mathbf{L}}_h(\boldsymbol{\theta}'') = \begin{pmatrix} \widetilde{L}_h^{u,u}(\boldsymbol{\theta}'') & \widetilde{L}_h^{u,v}(\boldsymbol{\theta}'') \\ \widetilde{L}_h^{v,u}(\boldsymbol{\theta}'') & \widetilde{L}_h^{v,v}(\boldsymbol{\theta}'') \end{pmatrix},$$

where matrix $\widetilde{\mathbf{L}}_h(\boldsymbol{\theta}'')$ is the Fourier symbol of $\mathbf{L}_h$.

Using standard coarsening, high and low frequency components on G_h are distinguished, in the way that the subset of low frequencies is $\boldsymbol{\Theta}_{2h} = (-\pi/2h_1, \pi/2h_1] \times (-\pi/2h_2, \pi/2h_2]$, and the subset of high frequencies is $\boldsymbol{\Theta}_h \setminus \boldsymbol{\Theta}_{2h}$.

From these definitions LFA smoothing and two-grid analysis can be performed as in rectangular grids, and smoothing factors for the relaxing methods μ and two-grid convergence factors ρ, which give the asymptotic convergence behavior of the method, can be well defined.

§3. Relaxing methods

Collective three-color smoother and some collective line-wise smoothers are proposed as relaxing methods. These smoothers appear as a natural extension to triangular grids of some smoothers widely used on rectangular grids, as red-black Gauss-Seidel and line-wise relaxations of zebra type.

3.1. Three–color smoother

To apply three-color smoother, the grid associated with a fixed refinement level η of a triangle T of the coarsest triangulation,

$$G_{T,h} = \{\mathbf{x}' = (x_1', x_2') \mid x_j' = k_j h_j,\ k_j \in \mathbb{Z},\ j = 1, 2,\ k_1 = 0, \ldots, 2^\eta,\ k_2 = 0, \ldots, k_1\}, \quad (2)$$

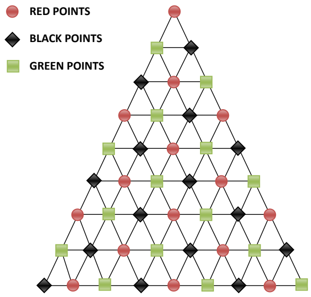


Figure 3: Three-color smoother.

is split into three disjoint subgrids,

$$G^i_{T,h} = \{\mathbf{x}' = (x'_1, x'_2) \in G_{T,h} \mid x'_j = k_j h_j,\ j = 1,2,\ k_1 + k_2 = i \pmod 3\}, \quad i = 0,1,2,$$

each of them associated with a different color, as shown in Figure 3, so that the unknowns of the same color have no direct connection with each other.

The complete three-color smoothing operator is given by the product of three partial operators, $\mathbf{S}_h(\omega) = \mathbf{S}^2_h(\omega)\mathbf{S}^1_h(\omega)\mathbf{S}^0_h(\omega)$. In each partial relaxation step, only the grid points of $G^i_{T,h}$ are processed, whereas the remaining points are not treated, i.e.

$$\mathbf{S}^i_h(\omega)\mathbf{v}_h(\mathbf{x}') = \begin{cases} [(\mathbf{I}_h - \omega\mathbf{D}_h^{-1}\mathbf{L}_h)\mathbf{v}_h](\mathbf{x}'), & \mathbf{x}' \in G^i_{T,h}, \\ \mathbf{v}_h(\mathbf{x}'), & \mathbf{x}' \in G_{T,h} \setminus G^i_{T,h}, \end{cases}$$

where $\mathbf{D}_h$ is the diagonal part of the discrete operator $\mathbf{L}_h$, $\mathbf{I}_h$ is the identity operator and ω is a relaxation parameter.

3.2. Zebra-type smoothers

For triangular grids, three different zebra smoothers can be defined on a triangle as shown in Figure 4. They consist of two half steps. In the first half-step, odd lines parallel to the edges of the triangle are processed, whereas even lines are relaxed in the second step, in which the updated approximations on the odd lines are used. They will be denoted as zebra-red, zebra-black and zebra-green smoothers, since they correspond to each of the vertices of the triangle.

In order to perform these smoothers, a splitting of the grid $G_{T,h}$ into two different subsets $G^{even}_{T,h}$ and $G^{odd}_{T,h}$ is necessary. For each of the zebra smoothers these subgrids are defined in a different way, and the corresponding distinction between them is specified in Table 1, where k_1 and k_2 are the indices of the grid points given in (2). Thus, these three smoothers $\mathbf{S}^{zR}_h$, $\mathbf{S}^{zB}_h$ and $\mathbf{S}^{zG}_h$ are defined by the product of two partial operators. For example, if zebra-red smoother is considered, $\mathbf{S}^{zR}_h = \mathbf{S}^{zR-even}_h \mathbf{S}^{zR-odd}_h$ where $\mathbf{S}^{zR-even}_h$ is in charge of relaxing the points in $G^{even}_{T,h}$ and $\mathbf{S}^{zR-odd}_h$ is responsible for the points in $G^{odd}_{T,h}$. These smoothers are preferred to the

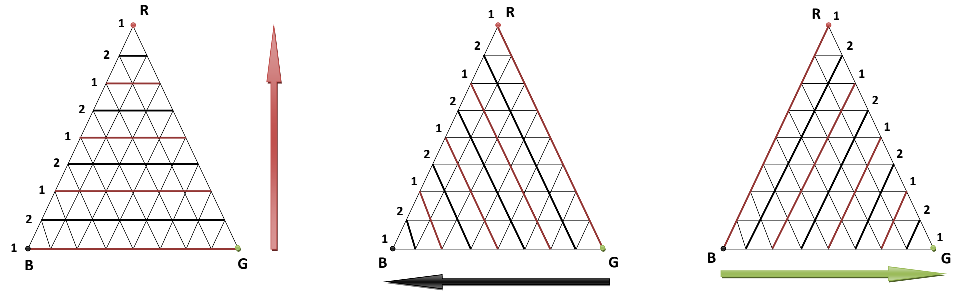


Figure 4: Zebra line smoothers: approximations at points marked by 1 are updated in the first half-step of the relaxation, and those marked by 2 in the second.

Relaxation	$G^{even}_{T,h}$	$G^{odd}_{T,h}$
Zebra-red	k_2 even	k_2 odd
Zebra-black	k_1 even	k_1 odd
Zebra-green	$k_1 + k_2$ even	$k_1 + k_2$ odd

Table 1: Characterization of subgrids $G^{even}_{T,h}$ and $G^{odd}_{T,h}$ for different zebra smoothers.

lexicographic line-wise Gauss-Seidel because in spite of having the same computational cost, smoothing factors corresponding to zebra smoothers are better than those of lexicographic line-wise relaxations as we will see further on.

§4. Fourier analysis results

It is easy to see that $\widetilde{\mathbf{L}}_{R,h} = R\, \widetilde{\mathbf{L}}_h R^t$, where $\widetilde{\mathbf{L}}_h$ and $\widetilde{\mathbf{L}}_{R,h}$ are the LFA symbols of the discrete operators associated with two grids, one obtained by rotating the other. Thus, it is fulfilled that these LFA symbols are similar and therefore LFA results obtained for these two grids are completely identical. Due to this property it is possible to restrict the analysis to triangles that sit on the x-axis of the Cartesian coordinate system.

This section focuses on analyzing different smoothers for the posed problem, while the components of the coarse–grid correction are taken as the standard ones as we have mentioned before.

One of the proposals here is the three–color smoother. In order to support the choice of this smoother as a good option for some geometries, some results obtained comparing it with the point-wise Gauss-Seidel are presented. For an equilateral triangle, these results appear in Table 2, where their two–grid convergence factors ρ and also the experimentally measured W–cycle convergence factors, denoted by ρ_h and obtained with a zero right-hand side and a random initial guess, are shown in order to observe that convergence factors are very well predicted by LFA. It can be observed that three-color smoother provides the best convergence factors between the two smoothers.

However, the three–color smoother is not robust over all angles, that is, the highly satis-

	Gauss–Seidel		Three–color smoother	
ν_1, ν_2	$\rho(\nu_1, \nu_2)$	$\rho_h(\nu_1, \nu_2)$	$\rho(\nu_1, \nu_2)$	$\rho_h(\nu_1, \nu_2)$
$1, 0$	0.516	0.506	0.422	0.422
$1, 1$	0.257	0.255	0.173	0.172
$2, 1$	0.172	0.172	0.097	0.095
$2, 2$	0.113	0.113	0.073	0.072

Table 2: Two–grid convergence factors ρ and measured W–cycle convergence rates ρ_h for equilateral triangles.

	Equilateral		Isosceles (75°)		Isosceles (85°)	
ν_1, ν_2	$\mu^{\nu_1+\nu_2}$	$\rho(\nu_1, \nu_2)$	$\mu^{\nu_1+\nu_2}$	$\rho(\nu_1, \nu_2)$	$\mu^{\nu_1+\nu_2}$	$\rho(\nu_1, \nu_2)$
$1, 0$	0.503	0.422	0.811	0.814	0.976	0.977
$1, 1$	0.253	0.173	0.657	0.661	0.954	0.955
$2, 1$	0.127	0.097	0.533	0.536	0.932	0.934
$2, 2$	0.064	0.073	0.432	0.435	0.911	0.913

Table 3: LFA smoothing and two–grid factors for different triangles with three-color smoother.

factory factors obtained for equilateral triangles worsen when one of the angles of the triangle is small. This behavior can be seen in Table 3, where smoothing and two-grid factors obtained with this smoother are shown for some representative triangles.

To overcome this difficulty, three zebra-type smoothers, associated with the three vertices of the triangle, are proposed. These zebra–type smoothers are preferred to the lexicographic block–line Gauss–Seidel smoothers because, despite having the same computational cost, they are more suitable for parallel implementation and their two-grid convergence factors are better, as we can see in Table 4 for an isosceles triangle with common angle 85^o. Each of these zebra–type smoothers is highly efficient when the angle corresponding to the vertex of its color is sufficiently small. This is shown in Table 5, where smoothing and two-grid factors for some representative triangles are shown.

As a final remark, it is observed that, depending on the geometry of the triangles, it is possible to improve the convergence factors of three-color and zebra-type smoothers by means of a relaxation parameter, whereas for the point-wise Gauss-Seidel and lexicographic line-wise smoothers there is no improvement. For instance, in the case of equilateral triangles the obtained convergence factor for three-color smoother is about 0.422 for $\nu_1 = 1$, $\nu_2 = 0$, and it can enhance to 0.303 taking a damping parameter $\omega = 1.1$.

§5. Conclusions

A Local Fourier Analysis for multigrid methods on triangular grids for the problem of planar elasticity has been presented. Analogously to the scalar case, the key point of this analysis is to introduce an expression of the Fourier transform in new coordinate systems, both in space and in frequency variables, associated with reciprocal bases. This analysis makes highly

	Lexicographic line-wise smoother		Zebra–type smoother	
ν_1, ν_2	$\rho(\nu_1, \nu_2)$	$\rho_h(\nu_1, \nu_2)$	$\rho(\nu_1, \nu_2)$	$\rho_h(\nu_1, \nu_2)$
$1,0$	0.333	0.331	0.143	0.142
$1,1$	0.151	0.145	0.071	0.069
$2,1$	0.094	0.094	0.047	0.046
$2,2$	0.063	0.062	0.036	0.034

Table 4: LFA two–grid convergence factors and measured W–cycle convergence rates ρ_h for isosceles triangles with common angle 85^o.

	Equilateral		Isosceles (75°)		Isosceles (85°)	
ν_1, ν_2	$\mu^{\nu_1+\nu_2}$	$\rho(\nu_1, \nu_2)$	$\mu^{\nu_1+\nu_2}$	$\rho(\nu_1, \nu_2)$	$\mu^{\nu_1+\nu_2}$	$\rho(\nu_1, \nu_2)$
$1,0$	0.535	0.404	0.387	0.165	0.265	0.143
$1,1$	0.226	0.164	0.096	0.072	0.053	0.071
$2,1$	0.104	0.088	0.034	0.047	0.034	0.047
$2,2$	0.049	0.067	0.025	0.035	0.025	0.036

Table 5: LFA smoothing and two–grid factors for equilateral and isosceles triangles with zebra–type smoother.

accurate predictions of the performance of a multigrid algorithm and as a consequence, also the choice of the adequate components of the method for a given problem. In this paper LFA has been applied to study the planar elasticity system, and with the help of this analysis a three-color smoother and some zebra-type smoothers are proposed to obtain an efficient multigrid algorithm to solve this problem.

Acknowledgements

This research has been partially supported by FEDER/MCYT Projects MTM2007-63204 and the DGA (Grupo consolidado PDIE).

References

[1] Bergen, B., Gradl, T., Hülsemann, F., and Rüde, U. A massively parallel multigrid method for finite elements. *Comput. Sci. Eng. 8* (2006), 56–62.

[2] Brandt, A. Multi-level adaptive solutions to boundary-value problems. *Comput. Sci. Eng. 31* (1977), 333–390.

[3] Gaspar, F. J., Gracia, J. L., and Lisbona, F. J. Fourier analysis for multigrid methods on triangular grids. *SIAM J. Sci. Comput. 31* (2009), 2081–2102.

[4] Trottenberg, U., Oosterlee, C. W., and Schüller, A. *Multigrid.* Academic Press, New York, 2001.

[5] Wienands, R., and Joppich, W. *Practical Fourier analysis for multigrid methods.* Chapman and Hall/CRC Press, 2005.

F. J. Gaspar, J. L. Gracia, F. J. Lisbona and C. Rodrigo
Department of Applied Mathematics
University of Zaragoza
fjgaspar@unizar.es, jlgracia@unizar.es, lisbona@unizar.es and carmenr@unizar.es

Monografías Matemáticas *García de Galdeano* **35**, 155–162 (2010)

α-THEORY FOR NEWTON-MOSER METHOD

José M. Gutiérrez, Miguel A. Hernández and Natalia Romero

Abstract. We study the semilocal convergence of Newton-Moser method to solve nonlinear equations $F(x) = 0$ defined in Banach spaces. The method defines a sequence $\{x_n\}$ that under appropriate conditions converges to a solution of the aforesaid equation. In fact, by following the known as α-theory, we give conditions on the starting point x_0 and on the derivatives of the operator F in order to establish such convergence. Finally, as an application, we apply this theory to the study of a kind of integral equations.

Keywords: Newton's method, Moser's method, semilocal convergence.

AMS classification: 45G10, 47H17, 65J15.

§1. Introduction

Newton-Moser method is a method to numerically solve nonlinear equations. In order to consider the more general case, let us consider a nonlinear equation

$$F(x) = 0, \tag{1}$$

where F is an operator defined between two Banach spaces X and Y. Let us assume that x^* is a simple root of (1).

Newton-Moser method is an iterative method defined by

$$\begin{cases} x_{n+1} = x_n - B_n F(x_n), \quad n \geq 0, \\ B_{n+1} = 2B_n - B_n F'(x_{n+1}) B_n, \quad n \geq 0, \end{cases} \tag{2}$$

where x_0 is a given point in X and B_0 is a given linear operator from Y to X.

The method exhibits several attractive features. First, it avoids the calculus of inverse operators that appears in Newton's method, $x_{n+1} = x_n - F'(x_n)^{-1}F(x_n)$, $n \geq 0$. So it is not necessary to solve a linear equation at each iteration. Second, it has quadratic convergence, the same as Newton's method. Third, in addition to solve the nonlinear equation (1), the method produces successive approximations $\{B_n\}$ to the value of $F'(x^*)^{-1}$, being x^* a solution of (1). This property is very helpful when one investigates the sensitivity of the solution to small perturbations.

We find the origin of the method in a Moser's work [6] for investigating the stability of the N-body problem in Celestial Mechanics. The main difficulty in this, and similar problems involving small divisors, is the solution of a system of nonlinear partial differential equations. In fact, Moser proposed the following method

$$\begin{cases} x_{n+1} = x_n - A_n F(x_n), \quad n \geq 0, \\ A_{n+1} = A_n - A_n(F'(x_n)A_n - I), \quad n \geq 0, \end{cases} \tag{3}$$

for a given $x_0 \in X$, a given $A_0 \in \mathcal{L}(Y, X)$, the set of linear operators from Y to X, and where I is the identity operator in X.

Notice that the first equation is similar to Newton's method, but replacing the operator $F'(x_n)^{-1}$ by a linear operator A_n. The second equation is Newton's method applied to equation $g_n(A) = 0$ where $g_n : \mathcal{L}(Y, X) \to \mathcal{L}(X, Y)$ is defined by $g_n(A) = A^{-1} - F'(x_n)$. So $\{A_n\}$ gives us an approximation of $F'(x_n)^{-1}$.

Method (3), firstly proposed by Moser, has a rate of convergence of $(1 + \sqrt{5})/2$ for simple roots. However, the variant (2) later introduced by Ulm [9] reaches quadratic convergence. Notice that in (2) $F'(x_{n+1})$ appears instead of $F'(x_n)$.

Since then, method (2) has been also considered by other authors. For instance, Hald [4] showed the quadratic convergence of the method. Later, Petzeltova [7] studied the convergence of the method under Kantorovich-type conditions.

Recently, in [2] a system of recurrence relations is given in order to analyze the convergence of Newton-Moser method (2) under estimations at one point. This theory, introduced by Smale [8], is an alternative to Kantorovich theory [5] to study the semilocal convergence of iterative processes to solve nonlinear equations. Roughly speaking, if x_0 is an initial value such that the sequence $\{x_n\}$ satisfies

$$\|x_n - x^*\| \le \left(\frac{1}{2}\right)^{2^n - 1} \|x_0 - x^*\|,$$

then x_0 is said to be an approximate zero of F. The following conditions were introduced by Smale [8] in order to prove that x_0 is an approximated zero

$$\left\|F'(x_0)^{-1}F(x_0)\right\| \le \beta, \tag{4a}$$

$$\sup_{k\ge 2}\left(\frac{1}{k!}\left\|F'(x_0)^{-1}F^{(k)}(x_0)\right\|\right)^{1/(k-1)} \le \gamma, \tag{4b}$$

$$\alpha = \beta\gamma \le 3 - 2\sqrt{2}. \tag{4c}$$

Wang and Zhao [10] pointed that condition (4) is too restrictive. Instead of (4) they assume

$$\left\|F'(x_0)^{-1}F(x_0)\right\| \le \beta, \tag{5a}$$

$$\frac{1}{k!}\left\|F'(x_0)^{-1}F^{(k)}(x_0)\right\| \le \gamma_k, \ k \ge 2, \tag{5b}$$

$$\left\{\begin{array}{l}\text{the equation } \phi(t) = 0 \text{ has at least a positive}\\ \text{solution, where } \phi(t) = \beta - t + \sum_{k\ge 2}\gamma_k t^k.\end{array}\right. \tag{5c}$$

In [2] the semilocal convergence of Newton-Moser method is established from a system of recurrence relations. However, a majorizing function, as the given in (5c), is not provided. In this paper we present a majorizing function for Newton-Moser method and we give an analysis of its convergence by following the patterns of the α-theory introduced by Smale. The semilocal convergence hypothesis and the main theorem are shown in section 2.

§2. Semilocal convergence results (α-theory)

In this section we study the semilocal convergence of Newton-Moser method (2) to solve the nonlinear equation (1). Let us assume that F is a nonlinear operator defined from an open subset Ω in a Banach space X to another Banach space Y. Let $x_0 \in \Omega$ be a given point and $B_0 \in \mathcal{L}(Y, X)$ a given linear operator defined from Y to X.

Instead the aforesaid conditions (4) or (5), we consider the following ones:

$$\|B_0 F(x_0)\| \leq \gamma_0, \tag{6a}$$

$$\|I - B_0 F'(x_0)\| \leq \beta < 1, \tag{6b}$$

$$\|B_0 F^{(j)}(x_0)\| \leq \gamma_j, \text{ for } j \geq 2, \tag{6c}$$

$$\begin{cases} \text{there exists } R > 0 \text{ such that the series} \\ \sum_{j\geq 2} \gamma_j t^j / j! \text{ is convergent for } t \in [0, R), \end{cases} \tag{6d}$$

$$f(\hat{t}) < 0, \tag{6e}$$

where $\hat{t}$ is the absolute minimum of the function

$$f(t) = \gamma_0 + (\beta - 1)t + \sum_{j\geq 2} \frac{1}{j!}\gamma_j t^j, \quad t \geq 0. \tag{7}$$

In addition, we consider the following scalar sequence

$$\begin{cases} t_0 = 0, \quad b_0 = -1, \\ t_{n+1} = t_n - b_n f(t_n), \\ b_{n+1} = 2b_n - b_n f'(t_{n+1}) b_n. \end{cases} \tag{8}$$

Condition (6e) allows us to say that function $f(t)$ defined in (7) has at least one positive root. Let us denote t^* the smallest positive solution of $f(t) = 0$. With the rest of conditions in (6), (7), (8), we can show that $\{t_n\}$ is an increasing monotone sequence to t^* and

$$\|x_{n+1} - x_n\| \leq t_{n+1} - t_n, \; n \geq 0. \tag{9}$$

Consequently, as $\{t_n\}$ is a convergent sequence and $\{x_n\}$ is a sequence defined in a Banach space, $\{x_n\}$ converges to a limit x^*, that can be shown it is a solution of the nonlinear equation (1).

In a more explicit way, the aforementioned comments are shown in the following results.

Theorem 1. *Let us consider the scalar sequences $\{t_n\}$ and $\{b_n\}$ defined in (8). Then the following relations hold:*

1. $b_n < 0$.
2. $b_n f'(t_n) < 1$.
3. $t_n < t_{n+1} < t^*$, *where t^* is the smallest positive root of (7).*

Proof. Firstly we notice that $f''(t) > 0$ for $t > 0$. Then, as $f'(0) = \beta - 1 < 0$ and $\lim_{t\to\infty} f(t) = \infty$, there exists a only value $\hat{t} \in (0, \infty)$ such that $f(\hat{t}) = 0$. Then, condition (6d) guarantees the existence of positive roots of function $f(t)$ defined in (7).

Now we prove the aforementioned are true for $n \geq 0$ by following an inductive reasoning. For $n = 0$ these relations are obviously true. If we suppose they are true for a given value of n, then $b_{n+1} = b_n(2 - b_n f'(t_{n+1})) < 0$, since $b_n f'(t_{n+1}) < b_n f'(t_n) < 1$.

In addition, as $(1 - b_n f'(t_{n+1}))^2 > 0$, then $b_{n+1} f'(t_{n+1}) = 2b_n f'(t_{n+1}) - b_n^2 f'(t_{n+1})^2 < 1$. Now we have $t_{n+2} - t_{n+1} = -b_{n+1} f(t_{n+1}) > 0$ and finally,

$$t^* - t_{n+2} = (1 - b_{n+1} f'(\eta_{n+1}))(t^* - t_{n+1}),$$

for $\eta_{n+1} \in (t_{n+1}, t^*)$. As $b_{n+1} f'(\eta_{n+1}) < b_{n+1} f'(t_{n+1}) < 1$, we conclude $t^* - t_{n+2} > 0$ and the induction is completed. □

Theorem 2. *Under conditions (6), the scalar sequence $\{t_n\}$ defined in (8) is a majorizing function for $\{x_n\}$ defined in (2), that is,*

$$\|x_{n+1} - x_n\| \leq t_{n+1} - t_n,\ n \geq 0. \tag{10}$$

Consequently, $\{x_n\}$ converges to a limit x^*.

Proof. Formula (10) can be proved by following an inductive reasoning. In fact, we can prove that the following inequalities hold for $n \geq 0$:

(I) $\|I - B_n F'(x_n)\| \leq 1 - b_n f'(t_n)$.

(II) $\|B_n F(x_n)\| \leq -b_n f(t_n)$.

(III) $\|B_n F^{(j)}(x_n)\| \leq -b_n f^{(j)}(t_n)$, $j \geq 2$.

Notice that (II) is equivalent to (10).

The aforesaid inequalities are clear for $n = 0$, just by taking into account (6). Now, if we assume they are true for $0, 1, \ldots, n$, then we can prove they are also true for $n + 1$.

Firstly, by (2), we have the following relationships:

$$\begin{aligned} I - B_{n+1} F'(x_{n+1}) &= (I - B_n F'(x_{n+1}))^2, \\ I - B_n F'(x_{n+1}) &= I - B_n F'(x_n) - \sum_{j\geq 1} \frac{1}{j!} B_n F^{(j+1)}(x_n)(x_{n+1} - x_n)^j, \\ \|I - B_n F'(x_{n+1})\| &\leq 1 - b_n f'(t_{n+1}), \\ \|I - B_{n+1} F'(x_{n+1})\| &\leq (1 - b_n f'(t_{n+1}))^2 = 1 - b_{n+1} f'(t_{n+1}). \end{aligned} \tag{11}$$

Then, (I) happens for $n + 1$.

Secondly,

$$B_n F(x_{n+1}) = (I - B_n F'(x_n)) B_n F(x_n) + \sum_{j\geq 2} \frac{1}{j!} B_n F^{(j)}(x_n)(x_{n+1} - x_n)^j.$$

Consequently,

$$\|B_nF(x_{n+1})\| \leq (1 - b_nf'(t_n))(-b_nf(t_n)) + \sum_{j\geq 2} \frac{1}{j!}(-b_nf^{(j)}(t_n))(t_{n+1} - t_n)^j$$
$$= -b_nf(t_n)) - b_nf'(t_n)(t_{n+1} - t_n) + \sum_{j\geq 2} \frac{1}{j!}(-b_nf^{(j)}(t_n))(t_{n+1} - t_n)^j = -b_nf(t_{n+1}).$$

Then, by taking norms in $B_{n+1}F(x_{n+1}) = (2I - B_nF'(x_{n+1}))B_nF(x_{n+1})$, we show that (II) also holds for $n + 1$. In fact,

$$\|B_{n+1}F(x_{n+1})\| \leq -(2 - b_nf'(t_{n+1})(b_nf(t_{n+1})) = -b_{n+1}f(t_{n+1}).$$

Finally,

$$\|B_{n+1}F^{(j)}(x_{n+1})\| \leq (2 - b_nf'(t_{n+1})) \sum_{k\geq 0} \frac{1}{k!}(-b_nf^{(k+j)}(t_n))(t_{n+1} - t_n)^k$$
$$= -(2 - b_nf'(t_{n+1}))(b_nf^{(j)}(t_{n+1})) = -b_{n+1}f^{(j)}(t_{n+1}).$$

Then (III) also holds and the induction is complete.

Now, as $\{t_n\}$ is a increasing sequence that converges to t^*, and the sequence $\{x_n\}$ is defined in a Banach space, $\{x_n\}$ converges to a limit x^*. □

Theorem 3. *Let x^* be the limit of the sequence $\{x_n\}$ defined in (2). Then, if $\|B_0\| \leq 1$, x^* is a solution of (1), that is $F(x^*) = 0$.*

Proof. Notice that $\|B_0\| \leq 1 = -b_0$. Then, taking into account (11) and the relationship $B_n = (I + (I - B_{n-1}F'(x_n))B_{n-1}$, we can show that $\|B_n\| \leq -b_n$ for $n \geq 0$.

In addition, as $B_{n+1} - B_n = ((I - B_nF'(x_{n+1}))B_n$, we have $\|B_{n+1} - B_n\| \leq b_{n+1} - b_n$ for $n \geq 0$ and then $\{B_n\}$ is a Cauchy sequence. Consequently, there exists a linear operator B^* such that $B^* = \lim_{n\to\infty} B_n$, $B^*F'(x^*) = I$. Then (see [5, Th. 2, p. 153]) there exists $F'(x^*)^{-1}$ and $\|F'(x^*)^{-1}\| \leq -1/f'(t^*)$. This fact, together with (II) in the proof of Theorem 2 guarantees that $F(x^*) = 0$. □

§3. Application to Fredholm integral equations

In this section we consider the following integral equation:

$$x(t) = z(t) + \lambda \int_a^b k(t, s)H(x(s))\,ds, \quad t \in [a, b],$$

where z is a given continuous function, H is an analytic function, k is a kernel continuous in its two variables and λ is a real parameter. This equation can be written as a equation $F(x) = 0$, where $F : X \to X$ is an operator defined on $X = C[a, b]$, the space of continuous functions in the interval $[a, b]$. The expression of such operator is the following:

$$F(x)(t) = x(t) - z(t) - \lambda \int_a^b k(t, s)H(\phi(s))\,ds, \quad t \in [a, b]. \tag{12}$$

In the space of continuous functions in $[a, b]$ we consider the max-norm:

$$\|g\| = \max_{t\in[a,b]} |g(t)|, \ g \in C[a, b].$$

For the kernel k we define

$$\|k\| = \max_{t\in[a,b]} \int_a^b |k(t, s)|\, ds.$$

In [3] Newton's method has been considered for studying the solution of (12). The two main problems of using Newton's method for solving a nonlinear equation is the choice of the initial approximation x_0 and the calculus of the inverses $F'(x_k)^{-1}$ (or the corresponding solution of a linear equation) at each step. In [3] the initial approximation is chosen as $x_0(t) = z(t)$ and then it is established a set of values for the parameter λ in order equation (12) has a solution. An estimate for the norm of $F'(x_0)^{-1}$ is also given.

Now, in this section we use Newton-Moser method (2) for studying the solution of (12). We consider the same choice for the initial approximation, that is $x_0(t) = z(t)$, but the calculus of $F'(x_0)^{-1}$ it is not required now.

To construct the majorizing function (7) we need to calculate the parameters γ_0, β and γ_j, $j \geq 2$, given in (6), by taking as starting point the function $x_0 = z$. The derivatives of order j of (12) are j-linear operators from the space X^j on X given by:

$$F'(x)[y_1](t) = y_1(t) - \lambda \int_a^b k(t, s)H'(x(s))y_1(s)\, ds,$$

$$F^{(j)}(x)[y_1, \dots, y_j](t) = -\lambda \int_a^b k(t, s)H^{(j)}(x(s))y_1(t)\cdots y_j(t)\, ds, \ j \geq 2.$$

Now we consider a particular integral equation of type (12). We take $x_0(t) = z(t)$ and $B_0 = I$, the identity operator, as starting values for Newton-Moser method (2) and we study the existence of solutions for the corresponding majorizing equation $f(t) = 0$, with f defined in (7). Notice that different convergence results could be obtained under different choices for $x_0(t)$ and B_0.

Let us consider the nonlinear integral equation

$$F(x)(t) = x(t) - 1 - \lambda \int_0^1 \cos(\pi st)x(s)^m\, ds. \tag{13}$$

We take $x_0(t) = 1$ for all $t \in [0, 1]$ and $B_0 = I$. Then, $\gamma_0 = |\lambda|, \beta = m|\lambda|$ and

$$\gamma_j = \begin{cases} |\lambda| m(m-1)\cdots(m-j+1), & \text{if } 2 \leq j \leq m, \\ 0 & \text{if } j > m. \end{cases}$$

Consequently the majorizing function (7) is given by

$$f(t) = |\lambda| + (m|\lambda| - 1)t + |\lambda| \sum_{j=2}^{m} \binom{m}{j} t^j = |\lambda|(1 + t)^m - t.$$

n	Newton-Moser method (2)	ρ
1	1.180118×10^{-1}	1.78711
2	2.14224×10^{-3}	1.84597
3	3.57016×10^{-5}	1.92453
4	1.30970×10^{-8}	1.96938
5	2.24399×10^{-15}	1.98755

Table 1: Error estimates (10) and the computational order of convergence (14)

If $m|\lambda| < 1$ this function has an absolute minimum $\hat{t} = -1 + (m\,|\lambda|)^{-1/(m-1)}$ and, in addition, $f(\hat{t}) < 0$.

Then, according with the results of the previous section, we have established a result on the existence of solution for equations (13). In fact, if $|\lambda| < 1/m$, the integral equation (13) has a solution. In addition, this solution can be approximated by using Newton-Moser method (2) starting with $x_0(t) = 1$ and $B_0 = I$.

For instance, if we consider $m = 5$ and $\lambda = \frac{1}{20}$ then, function

$$f(t) = \frac{1}{20}\left(1 - 15t + 10t^2 + 10t^3 + 5t^4 + t^5\right),$$

is the majorizing function of sequence $\{x_n\}$ and, $t^* = 0.0701898$ is the smallest positive root of f.

Using the majorizing sequence $\{t_n\}$, we show in Table 1 a priori error estimates (10) and the computational order of convergence [1]:

$$\rho \approx \ln \frac{\|t_{n+1} - t^*\|}{\|t_n - t^*\|} / \ln \frac{\|t_n - t^*\|}{\|t_{n-1} - t^*\|}, \qquad n \in \mathbb{N}, \tag{14}$$

when Newton-Moser method (2) is applied to solve equation (13).

Now, from Theorem 3 the integral equation (13) has a solution x^* in $B(1, 0.0701898)$ which is the limit of the iterations of Newton-Moser method (2) starting with $x_0(t) = 1$ and $B_0 = I$:

$$\begin{aligned}
x_1(t) &= 1 + 0.015915493\, t^{-1} \sin(3.14159\, t),\\
x_2(t) &= 1 + 0.017615759\, t^{-1} \sin(3.14159\, t),\\
x_3(t) &= 1 + 0.017633935\, t^{-1} \sin(3.14159\, t),\\
x_4(t) &= 1 + 0.017633938\, t^{-1} \sin(3.14159\, t).
\end{aligned}$$

Considering iteration $x_4(t)$ as a numerical solution x^* of integral equation (13) and the computational order of convergence:

$$\rho_n \approx \ln \frac{\|x_{n+1}(t) - x^*\|}{\|x_n(t) - x^*\|} \Bigg/ \ln \frac{\|x_n(t) - x^*\|}{\|x_{n-1}(t) - x^*\|}, \qquad n \in \mathbb{N}, \tag{15}$$

Newton-Moser method reach computationally the R-order of convergence at least two. In fact, $\rho_1 = 1.95368$ and $\rho_2 = 1.97401$.

Acknowledgements

This paper has been supported by the Spanish Ministry of Science and Technology (Ref. MTM2008-01952/MTM).

References

[1] Grau-Sánchez, M., and Noguera, M. A variant of Cauchy's method with accelerated fifth-order convergence. *Appl. Math. Lett. 17* (2004), 509–517.

[2] Gutiérrez, J. M., Hernández, M. A., and Romero, N. A note on a modification of Moser's method. *Journal of Complexity 24* (2008), 185–197.

[3] Gutiérrez, J. M., Hernández, M. A., and Salanova, M. A. α-theory for nonlinear Fredholm integral equations. *Grazer Mathematische Berichte 346* (2004), 187–196.

[4] Hald, O. H. On a Newton-Moser type method. *Numer. Math. 23* (1975), 411–425.

[5] Kantorovich, L. V. *Functional analysis*. Pergamon Press, Oxford, 1982.

[6] Moser, J. *Stable and random motions in dynamical systems with special emphasis on celestian mechanics*, vol. 77 of *Herman Weil Lectures, Annals of Mathematics Studies*. Princeton Univ. Press, Princeton, New Jersey, 1973.

[7] Petzeltova, H. Remark on Newton-Moser type method. *Commentationes Mathematicae Universitatis Carolinae 21* (1980), 719–725.

[8] Smale, S. *Newton's method estimates from data at one point*. The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics. Springer-Verlag, New York, 1986.

[9] Ulm, S. On iterative methods with successive aproximation of the inverse operator (Russian). *Izv. Akad Nauk Est. SSR 16* (1967), 403–411.

[10] Wang, D. R., and Zhao, F. G. The theory of Smale's point estimation and its applications. *J. Comput. Appl. Math. 60* (1995), 253–269.

José M. Gutiérrez, Miguel A. Hernández and Natalia Romero
Dep. Mathematics and Computation, University of La Rioja
C/ Luis de Ulloa s/n
26004 Logroño, Spain
`{jmguti, mahernan, natalia.romero}@unirioja.es`

Monografías Matemáticas *García de Galdeano* **35**, 163–170 (2010)

STABILITY OF EQUATORIAL AND HALO ORBITS AROUND A NON-SPHERICAL MAGNETIC PLANET

Manuel Iñarrea, Víctor Lanchares, Jesús Palacián,
Ana Isabel Pascual, José Pablo Salas and Patricia Yanguas

Abstract. The presence of micron size dust particles is frequent in the solar system and their dynamics have attracted the attention of researchers from the very beginning. Indeed, dusty rings of giant planets can be modeled by very simple models that take into account the movement of a single particle. One of these models is the so-called generalized Störmer problem, where a charged particle is supposed to orbit a spherical planet with magnetosphere. In this case, it is known the presence of equatorial and circular halo orbits as well as their stability. However, planets are not perfect spheres, and their oblateness must be taken into account. The aim of this paper is to show how the oblateness of the body affects the existence and stability of equatorial and halo orbits.

Keywords: Planetary magnetosphere, Störmer problem, equatorial and halo orbits, oblateness, equilibria, stability.

AMS classification: 70F05, 70F15, 70H08, 70H12, 70H14.

§1. Introduction

The understanding of the dynamics of planetary tiny dusty rings is usually studied by means of a single particle model named, in the literature, the generalized Störmer problem [1]. This model describes the dynamics of a dust particle orbiting a rotating magnetic planet and it takes into account the gravitational and magnetic effects. In this work we will consider that the magnetic field is a perfect magnetic dipole aligned along the north–south poles of the planet and the planet's magnetosphere is a rigid conducting plasma which rotates with the same angular velocity as the planet, which entails that the charge is subject to a corotational electric field. Finally we will suppose that the planet is not spherical. This last assumption introduces an additional perturbation to the previous works of Howard et al. [4, 3], Dullin et al. [1], Grotta–Ragazzo et al. [2], where the spherical case is considered. The aim of this paper is to study the influence of the oblateness coefficient in the existence and stability of circular orbits paralell to the equator or lying in it, that is to say, halo orbits and equatorial orbits respectively.

The paper is structured in three sections. The first one includes the Hamiltonian formulation of the problem. Second section is devoted to analyze the existence of equatorial and halo orbits. Some results about the stability of circular orbits appear in the third section.

Now, we start with the formulation of the problem. After using dimensionless cylindrical coordinates and momenta $(\rho, z, \phi, P_\rho, P_z, P_\phi)$ and adding the influence of the oblateness to the generalized Störmer problem, the system can be modeled by the following two degrees of freedom Hamiltonian function (see [1, 5, 6] for details)

$$\mathcal{H} = \frac{1}{2}\left(P_\rho^2 + P_z^2 + \frac{P_\phi^2}{\rho^2}\right) - \frac{1}{r} - \delta\frac{P_\phi}{r^3} + \frac{\delta^2}{2}\frac{\rho^2}{r^6} + \delta\beta\frac{\rho^2}{r^3} + 3J_2\frac{z^2}{2r^5} - \frac{J_2}{2r^3}, \tag{1}$$

where r is the distance of the particle to the center of the mass of the planet and lengths and time are expressed, respectively, in units of the planetary radius R and the Keplerian frequency $w_K = \sqrt{M/R^2}$. The problem depends on five parameters. Three external parameters: δ, the ratio between magnetic and Keplerian interactions (charge–mass ratio); β, the ratio between electrostatic and Keplerian interactions; and J_2, the oblateness coefficient of the planet. The sign of J_2 indicates if the planet is oblate ($J_2 > 0$) or prolate ($J_2 < 0$). The other two parameters are internal ones: P_ϕ, the angular momentum and $\mathcal{H} = \mathcal{E}$, the energy of the system.

Circular periodic trajectories appear as equilibria of the Hamiltonian system

$$\dot\rho = \frac{\partial\mathcal{H}}{\partial P_\rho}, \qquad \dot z = \frac{\partial\mathcal{H}}{\partial P_z}, \qquad \dot P_\rho = -\frac{\partial\mathcal{H}}{\partial\rho}, \qquad \dot P_z = -\frac{\partial\mathcal{H}}{\partial z},$$

or equivalently, as critical points of the generalized potential energy function (called effective potential) that can be written as

$$U_{eff} = \frac{P_\phi^2}{2\rho^2} - \frac{1}{r} - \delta\,\frac{P_\phi}{r^3} + \frac{\delta^2}{2}\frac{\rho^2}{r^6} + \delta\beta\frac{\rho^2}{r^3} + 3J_2\frac{z^2}{2r^5} - \frac{J_2}{2r^3}.$$

As it is usual in the literature, we introduce the particle angular velocity,

$$\omega = \dot\phi = \frac{\partial\mathcal{H}}{\partial P_\phi} = \frac{P_\phi}{\rho^2} - \frac{\delta}{r^3},$$

to eliminate P_ϕ, because ω is a more interesting parameter from the point of view of applications. To simplify the calculations, we also move to spherical variables (r, θ, ϕ) given by

$$\rho = r\sin\theta, \quad z = r\cos\theta, \quad \theta \in [0, \pi/2].$$

With these changes, circular orbits are obtained as the solutions of the nonlinear system of equations

$$\begin{cases} -6J_2 + 2r^2 + (9J_2 - 2\delta(\beta-\omega)r^2 - 2r^5\omega^2)\sin^2\theta = 0, \\ (-3J_2 + 2\delta(\beta-\omega)r^2 - r^5\omega^2)\sin 2\theta = 0. \end{cases} \tag{2}$$

Two types of equilibria, or circular orbits, appear depending on whether $\sin 2\theta$ is equal to zero or not. The first one occurs for $\sin 2\theta = 0$ and then $\theta = 0$ or $\theta = \pi/2$. If $\theta = 0$ then $\rho = 0$ and it constitutes a degenerate case, only meaningful for $J_2 > 1/3$. If $\theta = \pi/2$ we find the equatorial orbits. The second case takes place for $\sin 2\theta \neq 0$, and it gives rise to circular orbits parallel to the equator, also called halo orbits.

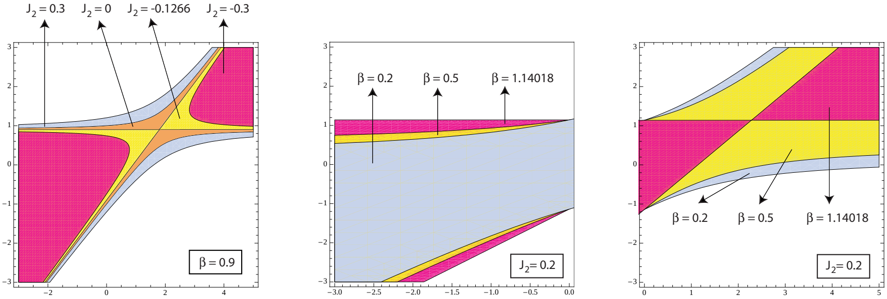


Figure 1: Regions of existence of equatorial orbits fixed β (left figure) and fixed J_2 (central and right figure) in the δ–ω plane.

§2. Circular orbits

An important question is to establish the conditions under which each type of circular orbit exists. We start discussing the case of equatorial orbits.

2.1. Equatorial orbits

As $\theta = \pi/2$, the second equation of the nonlinear system (2) is always verified and, for the first equation to be satisfied, r must be a positive real root of the following polynomial equation in the variable r

$$3\,J_2 + 2(1 - \beta\delta + \delta\omega)\,r^2 - 2\,\omega^2\,r^5 = 0. \tag{3}$$

Each positive real root of (3) corresponds to an equatorial orbit. Langbort [7] and Dullin et al. [1] have studied, respectively, some particular cases, when the particle is not charged ($\delta = 0$) and when the planet is spherical ($J_2 = 0$). Assuming $\delta \neq 0$ and $J_2 \neq 0$, some different results about the existence of equatorial orbits are obtained. They can be summarized in the following propositions (for details the reader is referred to [6]).

Proposition 1. *The region of existence of equatorial orbits enlarges for increasing values of J_2, fixed β. If J_2 is fixed, the region of existence enlarges or diminishes with β depending on the sign of the charge of the particle.*

Proof. The proof of this proposition, as well as the subsequent ones, is based on the analysis of the discriminant of the polynomial in equation (3), and on the fact that the radius of the orbit must be greater than one to be meaningful. Therefore, two curves appear delimiting the region of existence of equatorial orbits:

$$3125 J_2^3\omega^4 + 32(1 - \beta\delta + \delta\omega)^5 = 0, \tag{4}$$

$$2 - 2\beta\delta + 3J_2 + 2\delta\omega - 2\omega^2 = 0. \tag{5}$$

A detailed discussion of (4) and (5), in terms of the parameters β and J_2, yields the desired result. An illustration is given in Figure 1. □

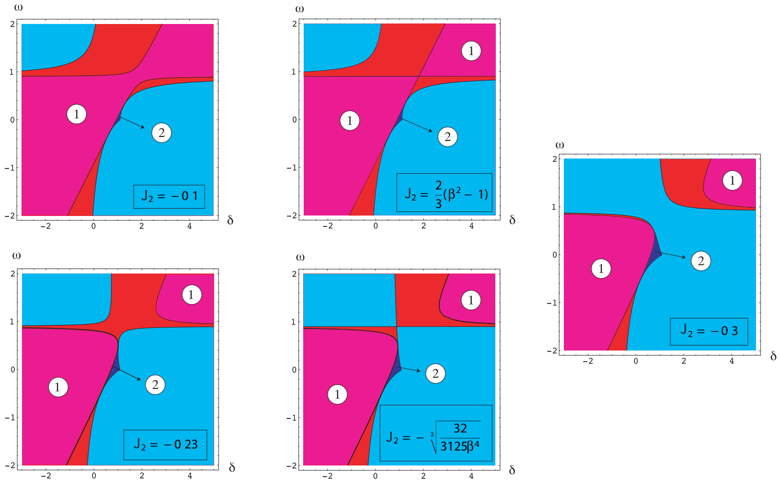


Figure 2: Number of equatorial orbits for $\beta = 0.9$ and different values of $J_2 < 0$.

In the proof of Proposition 1 it can be seen that for a prolate planet there may exist two positive real roots of the equation (3). In this case, it is interesting to know when these two roots give rise to two meaningful equatorial orbits, with radius greater than one. In this sense we obtain the following result.

Proposition 2. *If $J_2 < 0$, there is a region where two equatorial orbits with $r > 1$ exist at the same time.*

Proof. The region is defined by the contact points of the limiting curves (4) and (5). As the contact points are function of β and J_2, this region varies as the parameters change, as it is showed in Figure 2. □

It is worth noting that the region with two equatorial orbits is, in general, small in comparison with the region with only one equatorial orbit.

2.2. Halo orbits

The discussion about the existence of halo orbits is more difficult. Now, as $\theta \neq \pi/2$, none of the equations (2) vanishes identically and we are left to the equivalent system:

$$-3J_2 + 2\delta(\beta - \omega)r^2 - r^5\omega^2 = 0, \tag{6}$$

$$\sin^2\theta = \frac{6J_2 - 2r^2}{9J_2 - 2\delta(\beta - \omega)r^2 - 2r^5\omega^2)}. \tag{7}$$

The influence of J_2 in the region of existence of halo orbits can be summarized in the following two propositions.

Proposition 3. *The region of existence of halo orbits diminishes as J_2 increases for fixed β. Besides, if $J_2 > 0$, there is a range of charge-mass ratio not allowed for a particle to be in*

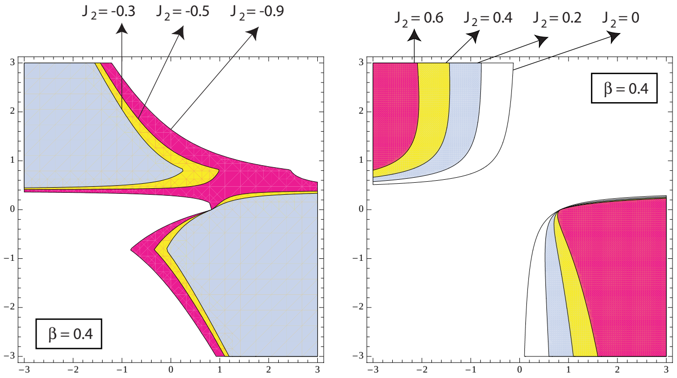


Figure 3: Region of existence of halo orbits for fixed β.

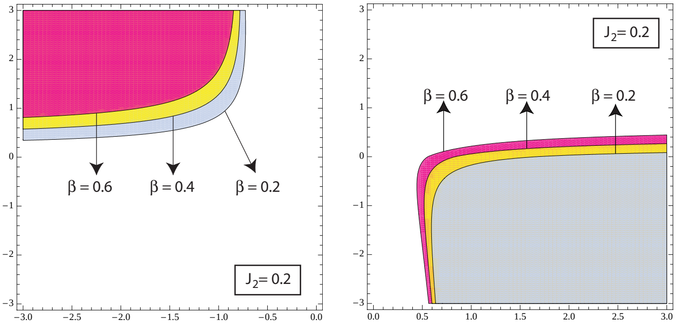


Figure 4: Region of existence of halo orbits for fixed J_2.

halo orbit. If J_2 is fixed, the region enlarges or diminishes with β depending on the sign of the charge.

Proof. The result follows from analysis of polynomial equation (6) and equation (7), taking into account that $0 \leq \sin^2\theta \leq 1$ and $r > 1$. As in the equatorial case, we find several limiting curves which depend on the parameters β and J_2. We arrive to the desired conclusion by the discussion of the limiting curves. Figures 3 and 4 illustrate the results, where Figure 3 shows the difference between oblate and prolate cases. Note how the gap of charge-mass ratios increases with the oblateness. □

Proposition 3 considers halo orbits with $r > 1$, which is a strong constrain for non-spherical bodies. Besides, it focuses on the existence of at least one halo orbit, but not in the number of them. Next proposition solves these aspects, which are illustrated in Figure 5.

Proposition 4. *There is a region where two halo orbits exist at the same time if $J_2 > 1/3$. This limit reduces to $J_2 > 1/8$ if the body is a homogeneous ellipsoid of revolution.*

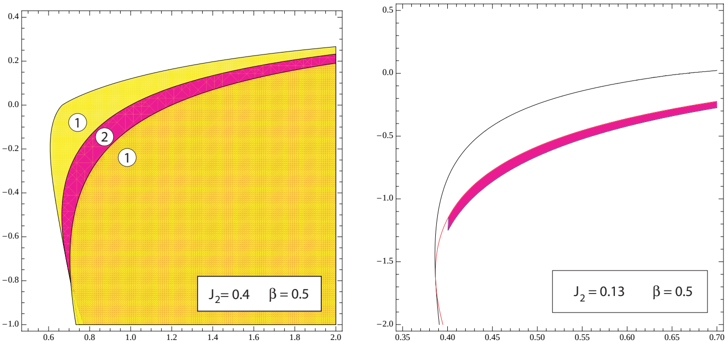


Figure 5: Region of existence of halo orbits for an oblate planet.

§3. Stability

Beyond the existence of circular orbits, their stability is an important question, as it determines the persistence of them with time. In this way, the stability follows from their character as critical points of the effective potential. Thus, if the Hessian matrix has two positive eigenvalues at the corresponding equilibrium, it is stable. The entries of the Hessian matrix are given by the second order partial derivatives of the effective potential

$$\begin{aligned}
\frac{\partial^2 U_{\mathit{eff}}}{\partial r^2} &= \frac{(\delta^2 + 2\beta\delta r^3 - 6\delta\omega r^3 + 3\omega^2 r^6)\sin^2\theta + 18J_2 r\cos^2\theta - 6J_2 r - 2r^3}{r^6},\\
\frac{\partial^2 U_{\mathit{eff}}}{\partial \theta^2} &= \frac{2(\delta + \omega r^3)^2 + [2\delta^2 - 3J_2 r + \omega^2 r^6 + 2\delta(\beta + \omega)r^3]\cos 2\theta}{r^4},\\
\frac{\partial^2 U_{\mathit{eff}}}{\partial r\partial \theta} &= \frac{-2\delta^2 + 9J_2 r - 2\delta(\beta - 2\omega)r^3 + 2\omega^2 r^6}{r^5}\sin\theta\cos\theta.
\end{aligned}$$

3.1. Stability of equatorial orbits

Here we will only discuss the stability of equatorial orbits. In this case the crossed derivative vanishes and the stability decouples in the radial direction (along the equator) and the vertical direction (away the equator), given by the eigenvalues

$$\begin{aligned}
\lambda_r &= \delta^2 - 6J_2 r - 2r^3 + 2\beta\delta r^3 - 6\delta r^3\omega + 3r^6\omega^2,\\
\lambda_\theta &= 3J_2 - 2\beta\delta r^2 + 2\delta r^2\omega + r^5\omega^2.
\end{aligned}$$

Exploiting the idea that if $\lambda_r = 0$ or $\lambda_\theta = 0$, a change in the stability occurs, we arrive to the following results which are illustrated in Figures 6, 7 and 8.

Proposition 5. *The area of radial stability enlarges when J_2 decreases, fixed β. The contrary for the stability away the equator.*

Proposition 6. *For fixed J_2, the region of stability, both radial and away the equator, enlarges if β increases and the charge is negative. For positive charged particles the region of stability away the equator diminishes for increasing β.*

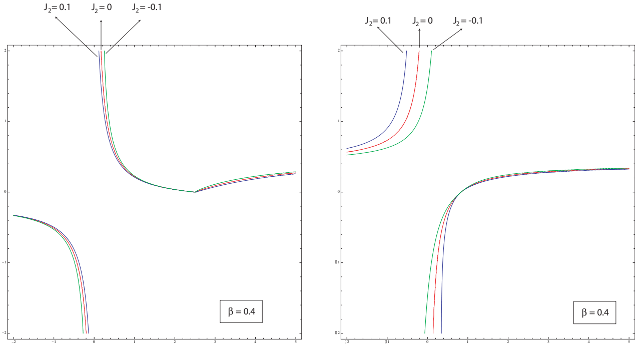


Figure 6: Regions of radial and vertical stability (respectively, left and right graphs) for equatorial orbits fixed β.

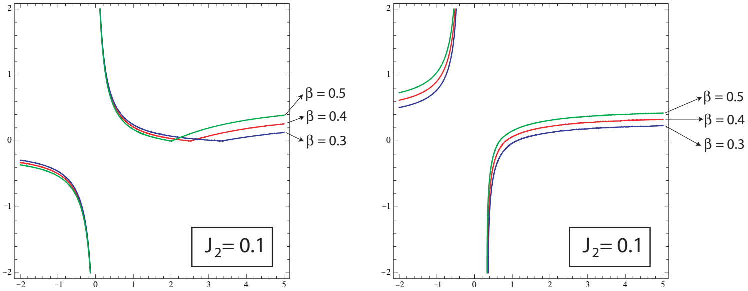


Figure 7: Regions of radial and vertical stability (respectively, left and right graphs) for equatorial orbits fixed J_2.

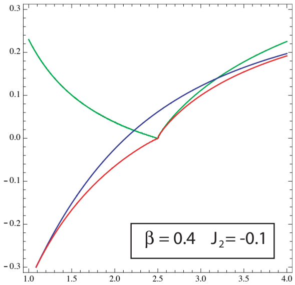


Figure 8: Changes of stability in the region with two equatorial orbits.

Proposition 7. *In the region with two equatorial orbits, the larger one suffers two changes of stability in the radial direction for positive charged particles.*

Proposition 8. *If $\lambda_r = 0$, a saddle-center bifurcation takes place, whereas, if $\lambda_\theta = 0$, there is a pitchfork bifurcation.*

Acknowledgements

We thank financial support from Ministerio de Ciencia e Innovación (Project #MTM2008–03818/MTM) from Spain.

References

[1] Dullin, H. R., Horányi, M., and Howard, J. E. Generalizations of the Störmer problem for dust grain orbits. *Physica D 171* (2002), 178–195. doi:10.1016/S0167-2789(02)00550-X.

[2] Grotta-Ragazzo, C., Kulesza, M., and Salomão, P. A. S. Equatorial dynamics of charged particles in planetary magnetospheres. *Physica D 225* (2007), 169–183. doi:10.1016/j.physd.2006.10.009.

[3] Howard, J. E., Dullin, H. R., and Horányi, M. Stability of halo orbits. *Phys. Rev. Lett. 84* (2000), 3244–3247. doi:10.1103/PhysRevLett.84.3244.

[4] Howard, J. E., Horányi, M., and Stewart, G. R. Global dynamics of charged dust particles in planetary magnetospheres. *Phys. Rev. Lett. 83* (1999), 3993–3996. doi:10.1103/PhysRevLett.83.3993.

[5] Iñarrea, M., Lanchares, V., Palacián, J., Pascual, A. I., Salas, J. P., and Yanguas, P. The Keplerian regime of charged particles in planetary magnetospheres. *Physica D 197* (2004), 242–268. doi:10.1016/j.physd.2004.07.009.

[6] Iñarrea, M., Lanchares, V., Palacián, J., Pascual, A. I., Salas, J. P., and Yanguas, P. The effect of J_2 on equatorial and halo orbits around a magnetic planet. *Chaos, Solitons and Fractals* (2008). doi:10.1016/j.chaos.2008.11.016.

[7] Langbort, C. Bifurcation of relative equilibria in the main problem of artificial satellite theory for a prolate body. *Celest. Mech. Dyn. Astr. 84* (2002), 369–385. doi:10.1023/A:1021185011071.

Manuel Iñarrea, Víctor Lanchares, Ana Isabel Pascual and José Pablo Salas.
Grupo de Dinámica No Lineal.
Universidad de La Rioja.
manuel.inarrea@unirioja.es, vlancha@unirioja.es, aipasc@unirioja.es,
josepablo.salas@unirioja.es

Jesús Palacián and Patricia Yanguas.
Departamento de Ingeniería Matemática e Informática.
Universidad Pública de Navarra.
palacian@unavarra.es, yanguas@unavarra.es

Monografías Matemáticas *García de Galdeano* **35**, 171–179 (2010)

ARBITRARY HIGH ORDER SCHEMES FOR THE SOLUTION OF THE LINEAR ADVECTION EQUATION

B. Latorre, P. García-Navarro, J. Murillo and J. Burguete

Abstract. In this work an arbitrary high order formulation for solving the linear advection equation with constant coefficients is presented. The conservative formulation is explicit, one step, and provides information at the sub-mesh scale. High order accuracy in space and time is achieved by means of polynomial representation of the states in each cell and conservative functional approximation of the exact solution of the advection equation. Altough high order methods are widely used when high precision of the numerical results is required, in this work we study the use of high order methods to compute faster than first order methods when low or middle precision of the numerical results is required. Numerical results for one-dimensional problems using schemes up to order of accuracy 5 are presented.

Keywords: Linear advection, high-order schemes, CFA, computational efficiency, Legendre polynomials.

§1. Introduction

The study of mixing in fluid flows is a complex problem involving different phenomena that can be faced under different degrees of approximation. For many hydraulic and environmental applications it is widely assumed that the fate of a tracer concentration can be modeled by means of the differential tracer mass conservation equation. This contains information on the mechanism of advection by the average flow velocity as well as molecular diffusion and turbulence mixing. The latter are often formulated as general diffusion or dispersion terms in the equation [6]. Using the mass conservation equation for the fluid flow it can be simplified to get the most widely used non-conservative form known as the convection-dispersion equation.

It is possible to solve numerically the advection-diffusion equation by discretizing the complete equation or by solving separately the advection and the diffusion. In this work, letting aside the technique that could be applied to the diffusive part, we are concerned with an efficient and accurate treatment of the convective part.

Finite volume methods rely on an integral formulation of the conservation law. The evolution of the cell averaged value of the conserved variable is evaluated through an estimation of the fluxes at the cell edges. Godunov's method, for instance, evaluates the flux using the exact solution of the Riemann problem at the edge [3]. Several approximate Riemann solvers have been proposed in order to generate efficient schemes [5], [4]. High order finite volume methods use also high order spatial reconstrucions at the grid cells in order to improve the evaluation of the numerical fluxes and Taylor time series or Runge-Kutta time integrations.

The limiting procedure to avoid oscillations follows the same criteria than in finite differences. A well known second order finite volume method is the MUSCL-Hancock scheme [10]. Ben-Artzi [1] and Toro [8] proposed methods based on the solution of the generalized Riemann plobem (piecewise polynomial), up to second order [1] and arbitrary high order (ADER) [8] [7] [9] in space and time. The ADER formulation offers the main advantage of requiring a single time step, however, it is associated to the necessity of following the steps of high order reconstruction, high order limitation, flux estimation and finally the storage of a single cell averaged value. This is a disadvantage common to all finite volume methods.

Dumbser [2] introduced the ADER methodolgy in the discontinuos Galerkin finite element framework. The variable is represented at every cell as a linear combination of basis functions leading to high order discretization avoiding the reconstruction step. Time accuracy is achieved in a single step thanks to the ADER formulation that transforms time derivatives into space derivatives. The finite element formulation projects the conservation law on every basis function in the time-space domain and evaluates the resulting integrals by means of quadrature formulae, allowing the resolution in both structured and unstructured grids.

This work presents a high order discretization technique analogous to the finite element methodology, based on Legendre polynomials. High order in space is achieved thanks to the polynomial representation within each cell avoiding the reconstruction step. High order in time is achieved in a single step by a conservative functional approximation (CFA technique) of the exact solution of the conservation law and a direct evaluation of the resulting integrals that does not require quadrature formulae. The resulting scheme is efficient in the sense that it requires a minimum number of mathematical operations. The scheme of order N to solve the linear advection equation in D dimensions requires $2DN^{2D}$ multiplications and additions per cell to calculate the evolution of the variable in a time step.

§2. Spatial discretization and sub-grid information

The numerical schemes considered in this work start from the basis that the information stored in a cell is a functional approximation, of a certain order of accuracy, of the spatial distribution of the variable within that cell. The numerical scheme is built to calculate the system evolution by providing an approximation of the new spatial distribution of the variable, at every cell, in the next time step.

The spatial representation of the variables at every grid cell is based on the mathematical concept of Hilbert space, an infinite-dimensional function space, defined over a spatial domain. The particular basis functions used in the present work are Legendre polynomials $P_n(x)$, defined in the spatial domain $x \in [-1, 1]$. Their orthogonality is given by the property:

$$\langle P_m, P_n \rangle = \int_{-1}^{1} P_m(x) P_n(x)\, dx = \frac{2}{2n-1}\delta_{m,n} \tag{1}$$

with the norm

$$||P_n|| = \sqrt{\langle P_n, P_n \rangle} = \sqrt{\frac{2}{2n-1}}. \tag{2}$$

One useful property of the Legendre polynomials in the context of the present work is that, for $n > 1$ they do not have net area:

$$\int_{-1}^{1} P_n(x)\,dx = \int_{-1}^{1} P_n(x)P_1(x)\,dx = \langle P_n, P_1\rangle = 2\,\delta_{n,1}. \tag{3}$$

An approximation of order of accuracy N of a square-integrable function $g(x)$ in the interval $x \in [-1, 1]$ can be obtained as linear combination of the N first Legendre polynomials:

$$\bar{g}(x) = \sum_{n=1}^{N} g_n P_n(x) \approx g(x), \tag{4}$$

with the coefficients

$$g_n = \frac{2n-1}{2}\langle g, P_n\rangle. \tag{5}$$

It is important to note that the mentioned approximation is conservative. To prove this property, fundamental to the conservative character of the numerical scheme, equation (4) is integrated using (5) and (3):

$$\int_{-1}^{1} \bar{g}(x)\,dx = \sum_{n=1}^{N} g_n \int_{-1}^{1} P_n(x)\,dx = \sum_{n=1}^{N} g_n\langle P_n, P_1\rangle = 2g_1 = \int_{-1}^{1} g(x)\,dx. \tag{6}$$

Due to the restrictions on the domain where the Hilbert space is defined, it is necessary, to move from a global to a local coordinate system within each cell, that is adapted to the domain of orthogonality of the basis functions. For that purpose x' will be used to denote the global coordinate and x to denote the local, cell adapted, coordinate.

To summarize, for the scheme of order of accuracy N, the state of the system at time t will be represented by means of the storage of N numbers at every grid cell ${}^{n}q_i^{t}$ representing the spatial distribution of the variable as linear combination of the Legendre polynomials

$$q_i(x,t) = \sum_{n=1}^{N} {}^{n}q_i^{t} P_n(x). \tag{7}$$

2.1. Time evolution

The form to build a numerical scheme able to solve the time evolution of a given spatial distribution based on the Legendre polynomial representation is next presented. The linear advection equation in one dimension for a function $q(x', t)$ is

$$\partial_t q(x', t) + \lambda\partial_x q(x', t) = 0, \tag{8}$$

with λ constant and positive. Instead of seeking an approximation of the individual terms in the equation, the existing exact solution is used as the basis of the advective method, that can be expressed in local coordinate

$$q(x, t + \Delta t) = q(x - 2c, t), \tag{9}$$

where c is the CFL number

$$c = \frac{|\lambda|\Delta t}{\Delta x'}. \tag{10}$$

Hence, starting from the known initial conditions, already expressed in local coordinates (7), we are interested in an expression for the solution at time $t = t + \Delta t$.

To achieve that, the pure transport formulated by (9) of the initial condition (7) is performed leading to the exact solution in cell i:

$$\widetilde{q}_i(x, t + \Delta t) = \begin{cases} q_{i-1}(x + 2 - 2c, t) & \text{if } -1 < x < 2c - 1, \\ q_i(x - 2c, t) & \text{if } 2c - 1 < x < 1. \end{cases} \tag{11}$$

Finally a conservative functional approximation of (11) is performed in order to reach the updated set of coefficients in the grid cell:

$$q_i(x, t + \Delta t) = \sum_{n=1}^{N} {}^{n}q_i^{t+\Delta t} P_n(x), \tag{12}$$

with the coefficients

$${}^{n}q_i^{t+\Delta t} = \frac{2n-1}{2} \int_{-1}^{1} \widetilde{q}_i(x, t + \Delta t) P_n(x)\, dx. \tag{13}$$

2.2. Updating scheme

The exact calculation of the integrals present in the definition of the coefficients (13) leads to the following updating numerical scheme of order accuracy N:

$${}^{n}q_i^{t+\Delta t} = \sum_{j=1}^{N} {}^{j}q_{i-1}^{t} L_{n,j}(c) + \sum_{j=1}^{N} {}^{j}q_i^{t} R_{n,j}(c), \tag{14}$$

where c is the CFL number (10) used to evaluate the left matrix L:

$$L(c) = T(c, 1 - c) \tag{15}$$

and the right matrix R:

$$R(c) = T(1 - c, -c). \tag{16}$$

Both L and R matrix are written in terms of a translation matrix T:

$$T_{i,j}(a, b) = a\,\frac{2j-1}{2} \sum_{n=1}^{min(i,j)} \frac{2}{2n-1} A_{i,n}(a, b) A_{j,n}(a, -b), \tag{17}$$

where A is a matrix with the property

$$P_i(ax + b) = \sum_{n=1}^{N} A_{i,n}(a, b) P_n(x). \tag{18}$$

To illustrate the procedure, the expression of the auxiliary matrix A up to third order of accuracy is next provided:

$$A(a,b) = \begin{pmatrix} 1 & 0 & 0 \\ b & a & 0 \\ d & 3ab & a^2 \end{pmatrix}, \quad d = \frac{a^2 + 3b^2 - 1}{2}, \tag{19}$$

and the corresponding T matrix

$$T(a,b) = a \begin{pmatrix} 1 & -3b & 5d \\ b & a^2 - 3b^2 & 5\left(bd - a^2 b\right) \\ d & 3\left(a^2 b - bd\right) & a^4 - 15a^2b^2 + 5d^2 \end{pmatrix}, \quad d = \frac{a^2 + 3b^2 - 1}{2}. \tag{20}$$

The numerical scheme (14) is conditionally stable provided that $c \leq 1$ for all orders of accuracy. This is dictated by the advection rule assumed (11) that only holds for $c \leq 1$. The scheme is exact when $c = 1$ in the sense that it transports exactly the polynomial distribution within every grid cell.

It is worth noting that in a simulation with fixed Δt and Δx, matrices L and R are constant. The scheme of order of accuracy N requires in this case $2N^2$ multiplications and additions per cell and time step. Taking in to account that the maximum time step allowed is independent of the order of accuracy, the total computational cost grows with the order of accuracy at a rate N^2.

§3. Numerical tests

3.1. Test 1

A Gaussian initial distribution is used as first test case to quantify the actual order of accuracy of the different approximations:

$$q(x', 0) = e^{-(x'-1)^2/0.05}. \tag{21}$$

A domain $x' \in [0, 2]$ is considered and periodical boundary conditions assumed. The advection velocity is chosen $\lambda = 1$ and a simulation time of $t = 20$ is performed so that the Gaussian distribution crosses 10 times the computational domain. The exact solution is the initial distribution at the same location. Using $c = 0.95$, simulations with different cell size Δx have been performed and their error has been evaluated using the L_1 norm as follows:

$$L_1 = \frac{1}{2M} \sum_{i=1}^{M} \int_{-1}^{1} |q_i(x) - q_{Exact}(x)| \, dx, \tag{22}$$

where M represent the cell number and the integrals in (22) are numerically computed.

The L_1 error norm corresponding to numerical schemes of orders of accuracy 2, 3, 4 and 5 versus the grid spacing is represented in logarithmic scale in Figure 1. In all the cases, the results have been fit to a straight line and the slope found is indicated in the figure.

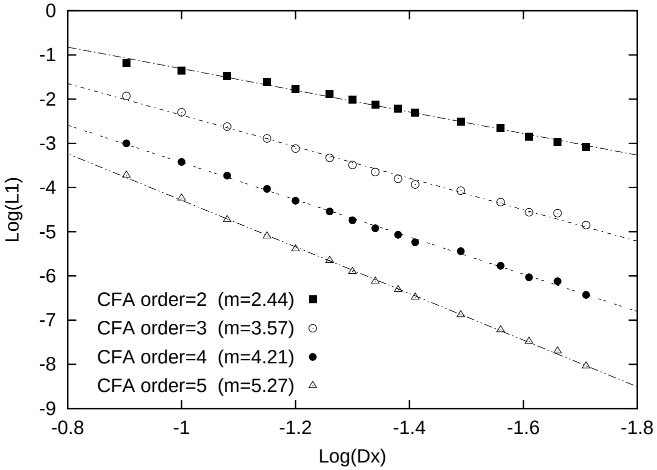


Figure 1: Error convergence of the CFA methods of order of accuracy 2, 3, 4 and 5 in the numerical test 1.

3.2. Test 2

The same initial distribution (21) and the same domain $x' \in [0, 2]$ are considered with the focus on the computational efficiency of the different schemes. We are interested in comparing the computational time required by the different schemes to achieve a target computational accuracy. The error of the schemes is quantified using the L_1 norm (22). Figure 2 displays the shape of the numerical solutions corresponding to values of $\log(L_1) = -1$ and $\log(L_1) = -2$ compared to the exact solution in order to state the acceptable size of this error in practical applications.

Two sets of computations, corresponding to a simulation time of $t = 20$ and $t = 200$, have been made and are plotted in Figures 3 a) and 3 b) respectively. These figures are a representation of the L_1 errors produced by the CFA schemes of orders of accuracy 1, 2, 3, 4 and 5 as a function of the computational time used, in logarithmic scale. In general, given a desired maximum error there is a scheme of a certain order of accuracy able to provide the solution at the lowest cost. Also, the figures show that if, when starting by a first order approach the numerical error is excessive, refining the grid is never the best option to reduce the error. This tendency continues and indicates that, for longer simulations (either in larger domains) higher order approximations gain in relative efficiency.

§4. Conclusions

A new approximation well suited for high order advection simulation has been presented. The scheme is explicit and based on a single updating step. Piecewise polynomial spatial discretization using Legendre polynomials provides the required spatial accuracy and the subgrid information. The updating scheme is built from the functional approximation of the

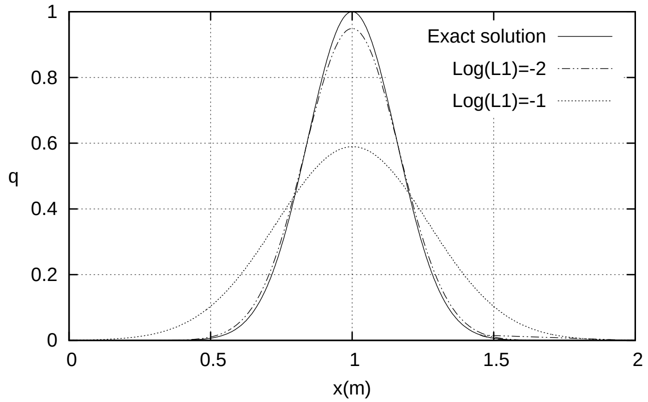


Figure 2: Shape of the numerical solutions corresponding to values of $\log(L_1) = -1$ and $\log(L_1) = -2$ compared to the exact solution in order to visualize the size of the L_1 error in Test case 2.

exact solution of the advection equation and a direct evaluation of the resulting integrals.

The numerical details have been provided and the schemes have been validated using a set of numerical experiments. Some of the test cases have been oriented to the convergence analysis of the schemes of different order of accuracy with an special interest in the computational efficiency of the different options. The results from the schemes from 1^{st} to 5^{th} order of accuracy have been presented showing that, in general, given a desired maximum error, there is a scheme of a certain order of accuracy able to provide the solution at the lowest cost. It is also worth remarking that if when starting by a first order approach the numerical error is excessive, refining the grid is never the best option to reduce the error. It is also true that for longer simulations (either in larger domains) higher order approximations gain in relative efficiency.

Acknowledgements

This research work has been partially funded by the Spanish Ministry of Science and Education under research project CGL2005-07059-C02-02.

References

[1] Ben-Artzi, M., and Falcovitz, J. A second-order Godunov-type scheme forcompressible fluid dynamics. *J. Comput. Phys. 55* (1984), 1–32.

[2] Dumbser, M., Balsara, D. S., Toro, E. F., and Munz, C.-D. A unified framework for the construction of one-step finite volume and discontinuous galerkin schemes on unstructured meshes. *J. Comput. Phys. 227*, 18 (2008), 8209–8253. doi:http://dx.doi.org/10.1016/j.jcp.2008.05.025.

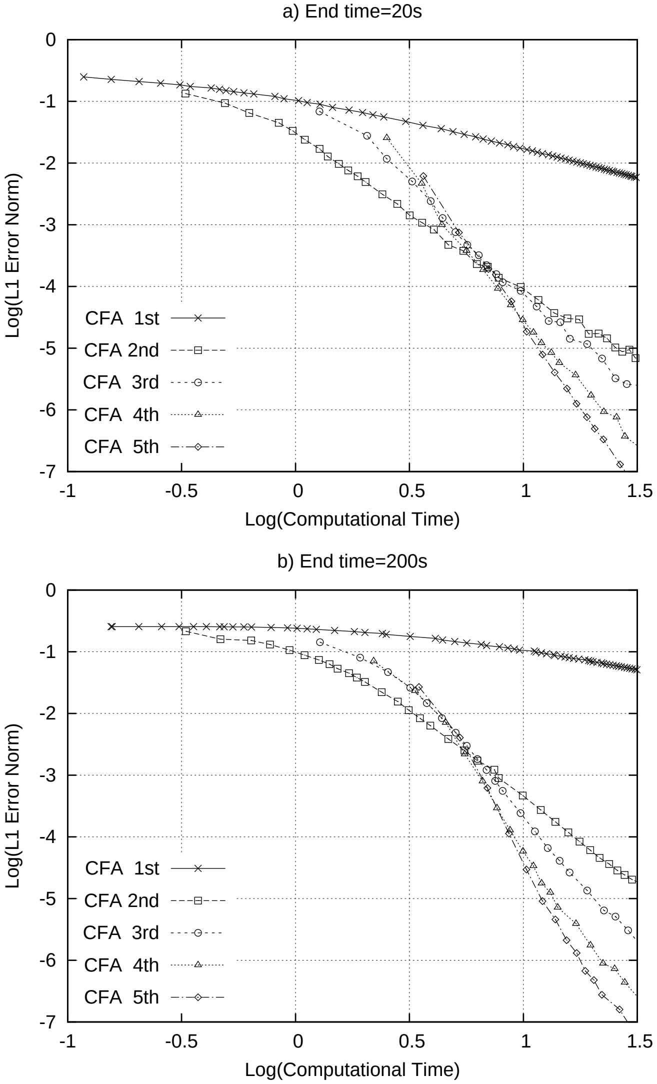


Figure 3: L_1 error norm produced by the CFA schemes of orders of accuracy 1, 2, 3, 4 and 5 as a function of the computational time used in numerical Test 2.

[3] Godunov, S. K. A finite difference method for the computation of discontinious solutions of the equation of fluid dynamics. *Mat. Sb. 47* (1959), 357–393.

[4] Harten, A., Lax, P. D., and van Leer, B. On upstreaming differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev. 25* (1983), 35–61.

[5] Roe, P. L. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys. 43* (1981), 357–372.

[6] Rutherford, J. C. *River Mixing*. John Wiley Sons Ltd, 1994.

[7] Titarev, V. A., and Toro, E. F. ADER: Arbitrary high order godunov approach. *J. Comput. Phys. 17* (2002), 609–618.

[8] Toro, E. F., Millington, R. C., and Nejad, L. A. M. *Towards very high order Godunov schemes*. Kluwer/Plenum Academic Publishers, 2001.

[9] Toro, E. F., and Titarev, V. A. ADER schemes for scalar non-linear hyperbolic conservation laws with source terms in three-space dimensions. *J. Comput. Phys. 202*, 1 (2005), 196–215. doi:http://dx.doi.org/10.1016/j.jcp.2004.06.014.

[10] van Leer, B. Towards the ultimate conservative difference scheme V: A second order sequel to Godunov's method. *J. Comput. Phys. 32* (1979), 101–136.

Borja Latorre
Mecánica de Fluidos
Centro Politécnico Superior
María de Luna 3. 50015 Zaragoza. Spain
borja.latorre@unizar.es

Monografías Matemáticas *García de Galdeano* **35**, 181–188 (2010)

SOLVING ONE-DIMENSIONAL LINEAR BOUNDARY VALUE PROBLEMS BY MULTI-POINT TAYLOR POLYNOMIALS. APPLICATIONS TO SPECIAL FUNCTIONS

José Luis López, Ester Pérez Sinusía and Nico M. Temme

Abstract. We consider second order linear differential equations of the form $\varphi(x)y'' + f(x)y' + g(x)y = h(x)$ in a real finite interval I with mixed Dirichlet and Neumann boundary data and a representation of its solution $y(x)$ by a multi-point Taylor expansion. The number and location of the base points of that expansion are conveniently chosen to guarantee that the expansion is uniformly convergent $\forall x \in I$. We propose several algorithms to approximate the multi-point Taylor polynomials of the solution based on the power series method for initial value problems. We show that multi-point Taylor polynomials are adequate to approximate the solution when the singularities of the coefficient functions of the differential equation are close to the interval I. We apply this technique to the approximation of several special functions.

Keywords: Second order linear differential equations, boundary value problem, multi-point Taylor expansions, special functions.

AMS classification: 34A25, 34B05, 41A58.

§1. Introduction

Let us consider boundary value problems of the form

$$\begin{cases} \varphi(x)y'' + f(x)y' + g(x)y = h(x) & \text{in } (-1,1), \\ MY = N, \end{cases} \tag{1}$$

where

$$M = \begin{pmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{33} \end{pmatrix}, \quad N = \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}, \quad Y^T = \left(y(-1), y'(-1), y(1), y'(1)\right),$$

M_{ij} and N_i are real numbers and $\operatorname{rank}(M) = 2$. We assume that (1) has a unique solution.

Different methods for approximating the solution of this kind of problems have been developed in the literature. Among these methods, the Taylor polynomial method is one of the most used tools. In the last few years, several authors have revisited this method and proposed new algorithms ([1, 6]). In the case in which it is possible to find a disk of convergence where the coefficient functions φ, f, g and h are analytic, the interval $[-1, 1]$ is contained inside that disk and $\varphi(x)$ does not vanish in that disk, the basic idea of the method proposed by Sezer and Kesan ([1, 6]) is the following. We consider the finite part of the Taylor expansion of the solution y at $x = c$:

$$y(x) \simeq y_n(x) := \sum_{k=0}^{n} a_k (x-c)^k$$

and equate to zero the Taylor coefficients at $x = c$ of $R(x) := \varphi(x)y_n''(x) + f(x)y_n'(x) + g(x)y_n(x) - h(x)$ up to the order $n-2$. Thus, we obtain a system of $n-1$ linear equations for the $n+1$ unknowns $a_0, a_1, a_2, \ldots, a_n$. The system is complemented with the two linear equations $MY = N$. We obtain then a linear system of $n+1$ equations and $n+1$ unknowns $a_0, a_1, a_2, \ldots, a_n$, whose solution gives an approximation to the Taylor polynomial $y_n(x)$, and then an approximation of the solution $y(x)$ of (1) ([1, 6]).

When $[-1, 1]$ is not included in the disk $D_r(c)$, we can take several points c_k (typically along the interval $[-1, 1]$) in such a way that $[-1, 1] \subset \cup_k D_{r_k}(c_k)$. Then, we use a Taylor expansion of the solution at every such point $x = c_k$ and match these expansions at intersecting disks $D_{r_k}(c_k)$ [5, Sec. 7]. In this way, we obtain an approximation of the solution of (1) in the form of a piecewise polynomial in several subintervals of $[-1, 1]$. Although this method gives an analytic approximation to the solution, this approximation is not uniform in the whole interval $[-1, 1]$ because it has a different polynomial representation over different subintervals $[-1, 1] \cap D_{r_k}(c_k)$. Besides this, the coefficients of the Taylor polynomial in every subinterval are determined by the coefficients of the Taylor polynomial in the adjacent subintervals, and this matching of expansions translates into numerical errors.

Thus, our purpose in this work is to show that multi-point Taylor polynomials [3, 4] combined with the method proposed in [1, 6] are adequate to approximate the solution of these equations in the case in which it is not possible to find a disk of convergence containing the interval of integration. Besides this, we show that this approximation provides a convergent expansion of the solution uniformly valid in the whole interval.

The paper is organized as follows. Section 2 presents a new method by considering two-point Taylor expansions instead of the classical Taylor expansion. We illustrate the technique with different examples. As a straightforward generalization of the two-point Taylor approximation, Section 3 includes an approximation by an $n-$point Taylor expansion.

§2. A Taylor expansion of the solution at the two extreme points

Let us consider a two-point Taylor expansion of the solution of (1) at the base points $x = \pm 1$ ([3]):

$$y(x) = \sum_{k=0}^{\infty} [a_k + b_k x](x^2 - 1)^k, \tag{2}$$

where the (unique) two-point Taylor coefficients a_k and b_k are related to the derivatives of y at $x = \pm 1$ ([3]).

We denote the Cassini oval in the complex plane with foci at $x = \pm 1$ and Cassini radius r by $O_r = \left\{ z \in \mathbb{C} \mid |z^2 - 1| = r \right\}$ and the Cassini disk by $\mathcal{D}_r = \left\{ z \in \mathbb{C} \mid |z^2 - 1| < r \right\}$. When $r > 1$, O_r is a single oval, when $r = 1$ it is a lemniscate, and when $r < 1$ it consists of two

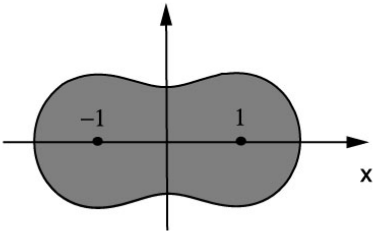


Figure 1: Graph of the Cassini disk when $r > 1$.

small ovals around the points ±1. When we assume $r > 1$, the interval $[-1, 1]$ is lying inside $\mathcal{D}_r$ (see Figure 1).

Suppose that the functions φ, f, g and h are analytic in the Cassini disk $\mathcal{D}_r$, $r > 1$, and $\varphi \neq 0$ in $\mathcal{D}_r$. We propose the following algorithm to approximate the unique solution y of (1).

Algorithm 1. The method of Frobenius assures that the unique solution y of (1) is analytic in the Cassini disk $\mathcal{D}_r$. Then, it is shown in [3] and [4] that y admits a two-point Taylor expansion of the form (2). From (2) we have

$$y'(x) = \sum_{k=0}^{\infty} \Big\{ \big[(2k+1)b_k + 2(k+1)b_{k+1}\big] + 2(k+1)a_{k+1}x \Big\} (x^2-1)^k,$$

$$\begin{aligned} y'(x) &= \sum_{k=0}^{\infty} \Big\{ \big[(2k+1)b_k + 2(k+1)b_{k+1}\big] + 2(k+1)a_{k+1}x \Big\} (x^2-1)^k, \\ y''(x) &= \sum_{k=0}^{\infty} 2(k+1) \Big\{ \big[(2k+1)a_{k+1} + 2(k+2)a_{k+2}\big] + \big[(2k+3)b_{k+1} + 2(k+2)b_{k+2}\big]x \Big\} \\ &\quad \times (x^2-1)^k. \end{aligned} \tag{3}$$

Using the above two-point Taylor expansions of y, y' and y'', we equate to zero the two-point Taylor coefficients of $R(x) := \varphi(x)y'' + f(x)y' + g(x)y - h(x)$ at $x = \pm 1$. We obtain in this way a_k and b_k, $k = 2, 3, 4, \ldots$, from a system of two recursions of the form:

$$\begin{cases} a_k = \displaystyle\sum_{j=0}^{k-1} [\alpha_{k,j} a_j + \beta_{k,j} b_j] + \gamma_k, \\ b_k = \displaystyle\sum_{j=0}^{k-1} [\alpha'_{k,j} a_j + \beta'_{k,j} b_j] + \gamma'_k, \end{cases} \qquad k = 2, 3, 4, \ldots, \tag{4}$$

where the coefficients $\alpha_{k,j}$, $\beta_{k,j}$, γ_k, $\alpha'_{k,j}$, $\beta'_{k,j}$, γ'_k depend on the two-point Taylor coefficients of φ, f, g and h at $x = \pm 1$. The computation of the coefficients a_k, b_k, $k = 2, 3, 4, \ldots$, requires the initial seed a_0, b_0, a_1 and b_1. From these recurrence relations we obtain the two-point Taylor coefficients a_k and b_k, $k = 2, 3, 4, \ldots$, of y at $x = \pm 1$ as an affine combination of the four first coefficients a_0, b_0, a_1 and b_1. We have

$$\begin{cases} a_k = A_k a_0 + B_k b_0 + C_k a_1 + D_k b_1 + E_k, \\ b_k = F_k a_0 + G_k b_0 + H_k a_1 + I_k b_1 + J_k, \end{cases} \qquad k = 2, 3, 4, \ldots, \tag{5}$$

where the coefficients A_k, $B_k, \ldots, J_k$ are functions of $\alpha_{k,j}$, $\beta_{k,j}$, γ_k, $\alpha'_{k,j}$, $\beta'_{k,j}$, γ'_k. The parameters a_0, b_0, a_1 and b_1 are linked by the equations $MY = N$, with

$$Y^T = (a_0 - b_0, b_0 + 2b_1 - 2a_1, a_0 + b_0, b_0 + 2b_1 + 2a_1). \tag{6}$$

This means that only two of these four parameters are free; suppose, for example, that a_1 and b_1 are free (if we choose another pair of parameters as free parameters we can proceed in a similar manner). Then, every two-point Taylor coefficient a_k and b_k, $k = 2, 3, 4, \ldots$, is an affine combination of only a_1 and b_1.

Every pair (a_1, b_1) gives rise to a different function y given by (2)-(5). Formally, all of these functions y are solutions of (1). But this problem has a unique solution, and then it must happen that the series (2) is convergent only for one pair (a_1, b_1), the one that gives rise to the unique solution of (1). The series (2) must be divergent for any other pair (a_1, b_1).

The correct values (a_1, b_1) may be then obtained by imposing the convergence of (2). In practice, we obtain an approximation $(\tilde{a}_1, \tilde{b}_1)$ of (a_1, b_1) by solving the two linear equations $a_{n+1} = b_{n+1} = 0$ (a_{n+1} and b_{n+1} are affine combinations of a_1 and b_1). Doing this we are imposing implicitly that (2) is convergent when we approximate this infinite series by

$$y_n(x) := \sum_{k=0}^{n} [a_k + b_k x](x^2 - 1)^k. \tag{7}$$

Once we have obtained the approximation $(\tilde{a}_1, \tilde{b}_1)$, we obtain from $MY = N$ an approximation $(\tilde{a}_0, \tilde{b}_0)$ of (a_0, b_0) and then, from (5), we obtain the approximations $\tilde{a}_k$ and $\tilde{b}_k$, $k = 2, 3, 4, \ldots$ of a_k and b_k as affine combinations of $\tilde{a}_1$ and $\tilde{b}_1$ and hence, the approximate two-point Taylor polynomial

$$\tilde{y}_n(x) := \sum_{k=0}^{n} [\tilde{a}_k + \tilde{b}_k x](x^2 - 1)^k. \tag{8}$$

Algorithm 1 can be reformulated in a more appropriate computational form. For further information, we refer to [2].

Example 1. Consider the boundary value problem

$$\begin{cases} \left(x^2 + \dfrac{1}{4}\right) y''(x) + i\left[c - (a + b + 1)\left(\dfrac{1}{2} + ix\right)\right] y'(x) + ab\, y(x) = 0, & x \in (-1, 1), \\ y(-1) = {}_2F_1(a, b; c; 1/2 - i), \quad y(1) = {}_2F_1(a, b; c; 1/2 + i). \end{cases}$$

We have $M_{11} = M_{23} = 1$ and the remaining $M_{ij} = 0$; $N_1 = {}_2F_1(a, b; c; 1/2 - i)$, $N_2 = {}_2F_1(a, b; c; 1/2 + i)$, $\varphi(x) = x^2 + 1/4$, $f(x) = i(c - (a + b + 1)(1/2 + ix))$, $g(x) = ab$ and $h(x) = 0$. The unique solution of this problem is the hypergeometric function: $y(x) = {}_2F_1(a, b; c; 1/2 + ix)$.

The coefficient functions are entire functions, but the function $\varphi(x) = (x^2 + \frac{1}{4})$ vanishes at $x = \pm 1/2i$. Thus, this function is nonvanishing in the Cassini disk $\mathcal{D}_r$ with foci at $x = \pm 1$ for any $1 < r < \sqrt{5}/2$.

We have

$$\begin{cases} y(-1) = a_0 - b_0 = {}_2F_1(a, b; c; 1/2 - i), \\ y(1) = a_0 + b_0 = {}_2F_1(a, b; c; 1/2 + i), \end{cases}$$

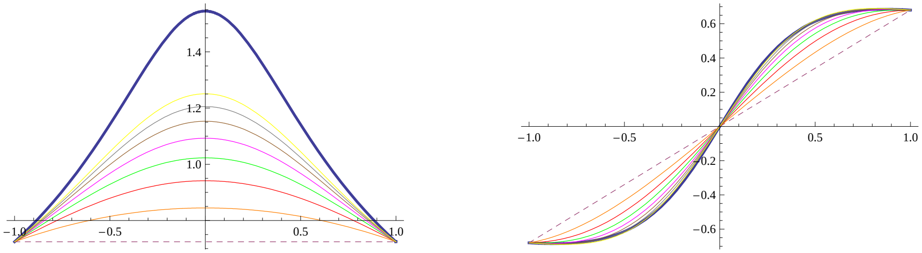


Figure 2: Graph of the real part and the imaginary part of the exact solution ${}_2F_1(a,b;c;1/2+ix)$ (blue) and the approximations $\widetilde{y}_n$, $n=0,1,\dots,7$ for $a=1$, $b=2$ and $c=3$.

thus,

$$\begin{cases} a_0 = \dfrac{{}_2F_1(a,b;c;1/2+i) + {}_2F_1(a,b;c;1/2-i)}{2}, \\ b_0 = \dfrac{{}_2F_1(a,b;c;1/2+i) - {}_2F_1(a,b;c;1/2-i)}{2}. \end{cases}$$

The two-point Taylor expansions of the coefficient functions are finite in this example:

$$\varphi(x) = \left[\frac{5}{4} + 0\cdot x\right] + [1 + 0\cdot x]\,(x^2-1),$$

$$f(x) = \left[i\left(c - \frac{a+b+1}{2}\right) + (a+b+1)x\right], \qquad g(x) = [ab + 0\cdot x]\,,$$

and then, the recursions are, for $k=0,1,2,\dots$,

$$\begin{aligned} &5(k+1)(k+2)a_{k+2} + \frac{1}{2}(k+1)(4a+4b+9+18k)a_{k+1} - i(k+1)(a+b+1-2c)b_{k+1} \\ &\quad + (a+2k)(b+2k)a_k - \frac{i}{2}(2k+1)(a+b+1-2c)b_k = 0, \\ &5(k+1)(k+2)b_{k+2} + \frac{1}{2}(k+1)(4a+4b+19+18k)b_{k+1} - i(k+1)(a+b+1-2c)a_{k+1} \\ &\quad + (1+a+2k)(1+b+2k)b_k = 0, \end{aligned}$$

with a_0 and b_0 given above and a_1 and b_1 free.

For several values of $n \in \mathbb{N}$, we solve the equations $a_{n+1} = b_{n+1} = 0$ for a_1 and b_1 and obtain the approximate values $\widetilde{a}_1$ and $\widetilde{b}_1$. From the above recursions and using the exact values of a_0 and b_0 and the approximate $\widetilde{a}_1$ and $\widetilde{b}_1$ we obtain the approximate Taylor polynomial. Figure 2 shows the approximation $\widetilde{y}_n(x)$ of $y(x)$ for some values of n and a, b and c.

Example 2. As an example of an oscillatory function we consider the boundary value problem

$$\begin{cases} [a+b+(b-a)x]^2\,y'' + (b-a)\,[a+b+(b-a)x]\,y' \\ \qquad + (b-a)^2\left[\left(\dfrac{a+b+(b-a)x}{2}\right)^2 - \alpha^2\right]y = 0, \quad x\in(-1,1), \\ y(-1) = J_\alpha(a), \qquad y(1) = J_\alpha(b), \end{cases} \tag{9}$$

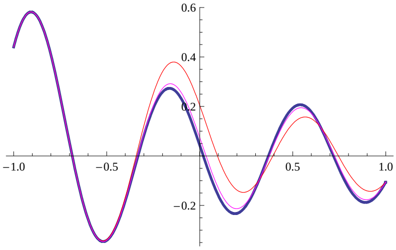


Figure 3: Plot of the exact solution $y(x) = J_\alpha(\frac{a+b+(b-a)x}{2})$ (thick blue) of (9) and the approximations $\tilde{y}_{10}(x)$ (red) and $\tilde{y}_{11}(x)$ (magenta) with $a = 1$, $b = 19$ and $\alpha = 1$.

with $0 < a < b$. We have $M_{11} = M_{23} = 1$ and the remaining $M_{ij} = 0$; $N_1 = J_\alpha(a)$, $N_2 = J_\alpha(b)$, $\varphi(x) = (a+b+(b-a)x)^2$, $f(x) = (b-a)[a+b+(b-a)x]$, $g(x) = (b-a)^2[(\frac{a+b+(b-a)x}{2})^2 - \alpha^2]$ and $h(x) = 0$. We consider the base points ±1. The unique solution of this problem is the Bessel function: $y(x) = J_\alpha(\frac{a+b+(b-a)x}{2})$.

For several $n \in \mathbb{N}$, we seek for an approximation $\tilde{y}_n(x)$ of the two-point Taylor polynomial $y_n(x)$ of $y(x)$ using Algorithm 1. Figure 3 illustrates the approximation $y(x) \simeq \tilde{y}_n(x)$ for some values of n, a, b and α.

Other examples in which two-point Taylor polynomials may be applied can be found in [2].

§3. A Taylor expansion of the solution at n points

When the Cassini disk $\mathcal{D}_r$ of analyticity of the coefficient functions of (1) with foci $x = \pm 1$ does not contain the interval $[-1, 1]$, we may consider an $n-$point Taylor expansion with $n > 2$ (see [4]). When those base points are conveniently chosen, we facilitate the inclusion of the interval $[-1, 1]$ in the generalized Cassini disk of convergence of the $n-$point Taylor expansion.

In general, if the coefficient functions have more singular points P_1, P_2, $P_3, \ldots$, close to the interval $[-1, 1]$, then we should consider a multi-point Taylor expansion with more base points such that the region of convergence avoids those singular points and contains the interval $[-1, 1]$. When we take more base points for the multi-point Taylor expansion, we squeeze the convergence region of the expansion avoiding the singular points P_k and including the interval $[-1, 1]$ in this region [4] (see Figure 4). The generalization of Algorithm 1 from two-point Taylor expansions to the $n-$point Taylor expansion case is straightforward. For further information, we refer to [2].

We illustrate the idea for $n = 3$ with the following example.

Example 3. Consider the boundary value problem

$$\begin{cases} \left[1-(x-ia)^2\right]y'' - 2(x-ia)y' + \left[\nu(\nu+1) - \dfrac{\mu^2}{1-(x-ia)^2}\right]y(x) = 0, \quad x \in (-1,1), \\ y(-1) = P_\nu^\mu(-1-ia), \qquad y(1) = P_\nu^\mu(1-ia), \end{cases} \tag{10}$$

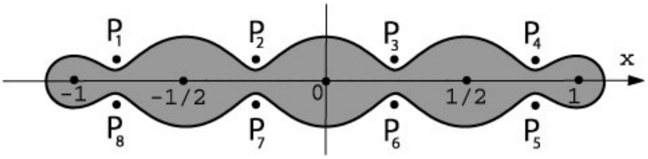


Figure 4: Typical portrait of the convergence region of a five-point Taylor expansion at the five base points $x = \pm 1$, $x = \pm 1/2$ and $x = 0$.

with $0 < a$ and $P_\nu^\mu(z)$ the Legendre function of the second kind with $\nu, \mu \in \mathbb{C}$. The coefficient functions are entire functions, but the function $\varphi(x) = [1 - (x - ia)^2]$ vanishes at $x = \pm 1 - ia$. If $a < (\sqrt{5} - 2)^{1/2}$, we cannot find a Cassini oval with foci at $x = \pm 1$ that contains the interval $[-1, 1]$ and that does not contain the points $x = \pm 1 - ia$. Hence, we cannot apply the method of Section 2. We consider then a three-point Taylor approximation for y with base points $x = \pm 1$ and $x = 0$ (see [4]) in the form

$$y(x) = \sum_{k=0}^{\infty} [a_k + b_k x + c_k x^2] x^k (x^2 - 1)^k.$$

This expansion is convergent in the region (see [4]) $\mathcal{E}_r = \left\{ z \in \mathbb{C} \,|\, |z(z^2 - 1)| < r \right\}$ with $r \leq a\sqrt{a^4 + 5a^2 + 4}$, that does not contain the points $x = \pm 1 - ia$. Moreover, this region contains the interval $[-1, 1]$ when $r > 2/(3\sqrt{3})$ (see Figure 5).

Figure 6 shows the approximation for some values of ν, μ and a.

Acknowledgements

J. L. López and E. Pérez Sinusía acknowledge financial support from *Dirección General de Ciencia y Tecnología*, project MTM2007-63772 and *Instituto de Biocomputación y Física de Sistemas Complejos de la Universidad de Zaragoza*. N.M.Temme acknowledges financial support from *Ministerio de Educación y Ciencia*, project MTM2006–09050.

References

[1] Kesan, C. Taylor polynomial solutions of linear differential equations. *Appl. Math. Comp. 142*, 1 (2003), 155–165.

[2] López, J. L., Pérez Sinusía, E., and Temme, N. M. Multi-point Taylor approximations in one-dimensional linear boundary value problems. *Appl. Math. Comp. 207* (2009), 519–527.

[3] López, J. L., and Temme, N. M. Two-point Taylor expansions of analytic functions. *Stud. Appl. Math. 109*, 4 (2002), 297–311.

[4] López, J. L., and Temme, N. M. Multi-point Taylor expansions of analytic functions. *Trans. Amer. Math. Soc. 356*, 11 (2004), 4323–4342.

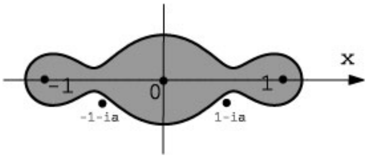


Figure 5: The region $\mathcal{E}_r$ of convergence of Example 3 contains the real interval $[-1, 1]$ and it does not contain the zeros $\pm 1 - ia$ of the function φ if $2/(3\sqrt{3}) < r \le a\sqrt{a^4 + 5a^2 + 4}$.

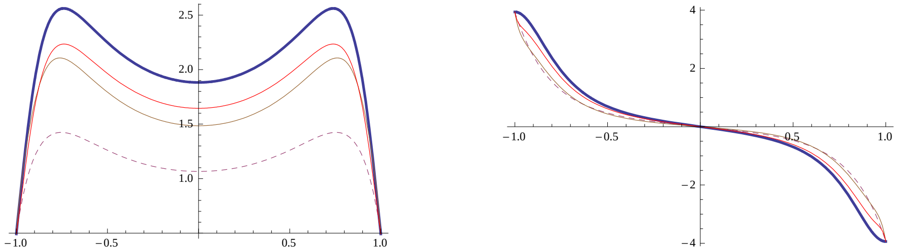


Figure 6: Plot of the exact solution $y(x) = P_\nu^\mu(x - ia)$ (thick blue) of (10) and the approximations $\tilde{y}_3(x)$ (dashed), $\tilde{y}_{12}(x)$ (brown) and $\tilde{y}_{20}(x)$ (red) for $a = 1/4$, $\nu = 1$ and $\mu = 2$.

[5] Olde Daalhuis, A. B., and Olver, F. W. J. On the asymptotic and numerical solution of linear ordinary differential equations. *SIAM Rev. 40*, 3 (1998), 463–495.

[6] Sezer, M. A method for the approximate solution of the second-order linear differential equations in terms of Taylor polynomials. *Int. J. Math. Edu. Sci. Technol. 27* (1996), 821–834.

José Luis López
Dpto. de Ingeniería Matemática e Informática
Universidad Pública de Navarra
Campus de Arrosadía
31006 Pamplona, Spain
jl.lopez@unavarra.es

Ester Pérez Sinusía
Departamento de Matemática Aplicada
Universidad de Zaragoza
C/ María de Luna
50018 Zaragoza, Spain
ester.perez@unizar.es

Nico M. Temme
Centrum Wiskunde & Informatica (CWI)
Science Park 123
1098 XG Amsterdam, The Netherlands
Nico.Temme@cwi.nl

Monografías Matemáticas *García de Galdeano* **35**, 189–196 (2010)

MATHEMATICAL ANALYSIS OF A STRATIGRAPHIC MODEL FOR THE LARGE SCALE DEPOSITIONAL TRANSPORT PROCESSES

Mohamed Salem Louly

Abstract. This work deals with the study of a stratigraphic model for the formation of geological basins under a maximal erosion rate constraint. It leads to an original class of conservation laws in order to simulate the large scale depositional transport processes of sediments.The main feature of the mathematical framework is characterized by a global constraint on the time-derivative of the unknown (the theoretical topography) and a non-linear transport term.Various theoretical results and research procedures are presented for solving the monolithologic column case: existence and uniqueness of an approximating sequence via an implicit time-discretization scheme, lack of compactness results and open problems for passing to the limits in the discretized processes (double weak convergence in the diffusive term).The locally hyperbolic behavior of the model is illustrated by constructing realistic traveling-waves solutions and by pointing out some surprising properties (barrier effects and dead-zones).

Keywords: Stratigraphic modelling, limited weathering, inverse problem.

AMS classification: 35K20, 35K85, 35Q72.

§1. Introduction

Let us consider a sedimentary basin, with base Ω, bounded domain in $\mathbb{R}^d$, $(d = 1, 2)$, supposed horizontal at the zero topographic level. One denotes by $S := S(t, x)$, $(t, x) \in Q$, with $Q := [0, T] \times \Omega$, $T > 0$, the sediments height, and, by $\vec{V} := \vec{V}(x, S)$, a transport term, and finally, one denotes by $\vec{q}$ the flow of matter.

According to the Darcy law, we propose the following equality for sediment flux:

$$\vec{q} \quad = -\lambda[\nabla S + \vec{V}(x, S)],$$

where λ is a proportionality coefficient; it is playing the role of a flux limiter, with the relevant law of state in the non-standard form, according to [1, 4]:

$$\frac{\partial S}{\partial t} + E \geq 0, \quad 0 \leq \lambda \leq 1 \quad \text{and} \quad (1-\lambda)\left(\frac{\partial S}{\partial t} + E\right) = 0 \quad \text{a.e. in } Q, \tag{1}$$

where E is a positive constant. Then, to express such a constraint, the limiter λ will be in $\mathcal{H}(\frac{\partial S}{\partial t} + E)$, where $\mathcal{H}$ denotes the maximal graph of Heaviside.

On the other hand, the mass conservation law gives us

$$\frac{\partial S}{\partial t} + \operatorname{div}\vec{q} = 0 \quad \text{a.e. in } Q.$$

Therefore,

$$\frac{\partial S}{\partial t} - \operatorname{div}\left\{\lambda\left(\frac{\partial S}{\partial t} + E\right)\left[\nabla S + \vec{V}(x,S)\right]\right\} = 0, \quad \text{a.e. in } Q.$$

It is a degenerate non-linear hyperbolic equation. Thus, we formulate the model in the following way (homogeneous Dirichlet conditions on the boundary):

$$(\mathcal{P})\begin{cases}\text{Find } (S,\lambda) \in L^\infty(0,T;H_0^1(\Omega)) \cap H^1(Q) \times L^\infty(Q) \cap \mathcal{H}\left(\frac{\partial S}{\partial t} + E\right) \\ \text{such that, for almost any } t \in]0,T[\text{ and for all } v \in H_0^1(\Omega), \\ \displaystyle\int_\Omega \frac{\partial S}{\partial t} v\,dx + \int_\Omega \lambda[\nabla S + \vec{V}(x,S)]\cdot\nabla v\,dx = 0, \\ \displaystyle\frac{\partial S}{\partial t} + E \geq 0 \text{ a.e. in } Q \text{ if } d = 2, \quad S(0,\cdot) = S_0 \text{ in } H_0^1(\Omega),\end{cases}$$

where $\vec{V} \in W^{1,\infty}(\Omega\times\mathbb{R},\mathbb{R}^2)$. Note that if $d = 1$, thanks to the Saks lemma,

$$\left(\frac{\partial S}{\partial t} + E\right)^- \leq -\frac{\partial}{\partial x}\left[\lambda\left(\frac{\partial S}{\partial x} + V(x,S)\right)\right]1_{\{\frac{\partial S}{\partial t}+E<0\}} = 0.$$

If $d = 2$, the argument fails for the divergence operator.

Firstly, we consider the lipschitz approximation of the Heaviside function:

$$\lambda_\varepsilon : r \in \mathbb{R} \longmapsto \lambda_\varepsilon(r) = \min\left(\frac{r^+}{\varepsilon}, 1\right), \quad \varepsilon > 0.$$

Then, we introduce the approximating formulation, solving in the following sense:

$$\begin{cases}\text{For } S_0 \in H_0^1(\Omega), \text{ find } S_\varepsilon \in L^2(0,T;H_0^1(\Omega)) \text{ such that } \frac{\partial S_\varepsilon}{\partial t} \in L^2(0,T;L^2(\Omega)), \\ \text{verifying, for all } v \in H_0^1(\Omega) \text{ and for almost any } t \in]0,T[\,, \\ \displaystyle\int_\Omega \frac{\partial S_\varepsilon}{\partial t} v\,dx + \int_\Omega \lambda_\varepsilon\left(\frac{\partial S_\varepsilon}{\partial t} + E\right)\left[\nabla S_\varepsilon + \vec{V}(x,S_\varepsilon)\right]\cdot\nabla v\,dx = 0, \\ \displaystyle\frac{\partial S_\varepsilon}{\partial t} + E \geq 0 \text{ a.e. in } Q(\text{if } d = 2), \quad S_\varepsilon(0,\cdot) = S_0 \text{ a.e. in } \Omega.\end{cases}$$

§2. An implicit time discretization method

Let $S_0 \in H_0^1(\Omega)$ and a real $h > 0$. Then we consider the variational problem

$$(\mathcal{P}_\varepsilon)\begin{cases}\text{Find } S_\varepsilon \text{ in } H_0^1(\Omega) \text{ such that, for any } v \in H_0^1(\Omega), \\ \displaystyle\int_\Omega \frac{S_\varepsilon - S_0}{h} v\,dx + \int_\Omega \lambda_\varepsilon\left(\frac{S_\varepsilon - S_0}{h} + E\right)\left[\nabla S_\varepsilon + \vec{V}(x,S_\varepsilon)\right]\cdot\nabla v\,dx = 0.\end{cases}$$

It is a degenerate problem. Then we use an artificial viscosity method: let $\delta \in \,]0,1[$ and consider the nondegenerate stationary problem

$$\left(\mathcal{P}_\varepsilon^\delta\right)\begin{cases}\text{Find } S_\varepsilon^\delta \text{ in } H_0^1(\Omega) \text{ such that, for all } v \in H_0^1(\Omega),\\ \displaystyle\int_\Omega \frac{S_\varepsilon^\delta - S_0}{h} v\,dx + \int_\Omega \Big[\lambda_\varepsilon(\frac{S_\varepsilon^\delta - S_0}{h} + E) + \delta\Big]\Big[\nabla S_\varepsilon^\delta + \vec{V}(x, S_\varepsilon^\delta)\Big]\cdot\nabla v\,dx = 0.\end{cases}$$

To resolve this problem, we use Schauder-Tikhonov fixed point theorem in separable Hilbert spaces. Exactly, we make the following step: for any $g \in H_0^1(\Omega)$, let us introduce the linear problem

$$\left(\mathcal{P}_\varepsilon^{\delta,g}\right)\begin{cases}\text{Find } S_\varepsilon^{\delta,g} \text{ in } H_0^1(\Omega) \text{ such that, for any } v \in H_0^1(\Omega),\\ \displaystyle\int_\Omega \frac{S_\varepsilon^{\delta,g} - S_0}{h} v\,dx + \int_\Omega \Big[\lambda_\varepsilon(\frac{g - S_0}{h} + E) + \delta\Big]\Big[\nabla S_\varepsilon^{\delta,g} + \vec{V}(x, g)\Big]\cdot\nabla v\,dx = 0,\end{cases}$$

which possesses a unique solution according to the theorem of Lax-Milgram. Then, we can define the mapping

$$\psi : g \in H_0^1(\Omega) \longrightarrow S_\varepsilon^{\delta,g} \in H_0^1(\Omega)$$

and we show that by Poincaré's inequality

$$\left\|S_\varepsilon^{\delta,g}\right\|_{H_0^1(\Omega} \le \frac{C(\Omega)}{\delta h}\,\|S_0\|_{H_0^1(\Omega)} + \frac{2}{\delta}\left\|\vec{V}\right\|_\infty \sqrt{\text{meas}(\Omega)} = r_\delta,$$

that is

$$\psi(H_0^1(\Omega)) \subset \overline{B}_{H_0^1(\Omega)}(0, r_\delta).$$

Then, we show that ψ is sequentially weakly continuous in $\overline{B}$, what implies, according to Schauder, that ψ admits a fixed point. As a consequence, the problem $\left(\mathcal{P}_\varepsilon^\delta\right)$ has one solution. Next, let us have a look for $\delta \to 0^+$. We show, by taking the test function $v = \int_0^{S_\varepsilon^\delta - S_0} \frac{du}{\lambda_\varepsilon(u/h+E)+\delta}$ in the formulation $\left(\mathcal{P}_\varepsilon^\delta\right)$, that

$$\left\|S_\varepsilon^\delta\right\|_{H_0^1(\Omega)} \le 2\|S_0\|_{H_0^1(\Omega)} + \|\vec{V}\|_\infty \sqrt{\text{meas}(\Omega)}.$$

That is, the sequence $(S_\varepsilon^\delta)_{\delta>0}$ is bounded in $H_0^1(\Omega)$ and, up to a subsequence, we obtain the necessary convergences for the passage to the limit when δ aims towards 0 in $\left(\mathcal{P}_\varepsilon^\delta\right)$. Thus, a solution of $(\mathcal{P}_\varepsilon)$ exists. We have then the following proposition.

Proposition 1. *Let $S^0 \in H_0^1(\Omega)$, $N > 0$ and $h = T/N$. Then, there exists a sequence $(S_\varepsilon^i)_{1\le i\le N}$ such that, for all $1 \le i \le N$ and for all $v \in H_0^1(\Omega)$,*

$$\int_\Omega \frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} v\,dx + \int_\Omega \lambda_\varepsilon(\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} + E)\Big[\nabla S_\varepsilon^i + \vec{V}(x, S_\varepsilon^i)\Big]\cdot\nabla v\,dx = 0. \tag{2}$$

Moreover,

$$\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} + E \ge 0 \quad \text{a.e. in } \Omega, \text{ where } S_\varepsilon^0 = S^0. \tag{3}$$

Note that the inequality (3) is obvious by taking $v = \left(\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} + E\right)^-$. Moreover, we also assert the uniqueness of the solution, in every iteration:

Proposition 2. *The solution given by the Proposition 1 is unique in every iteration.*

Proof. It is enough to show this property of uniqueness for the first iteration. For it, we put

$$w_\varepsilon = \frac{S_\varepsilon^1 - S_0}{h} + E \qquad \text{and} \qquad A_\varepsilon(r) = \int_0^r \lambda_\varepsilon(u)\,du,\ r \geq 0.$$

Then, for all $v \in H_0^1(\Omega)$, we have, according to (2), that

$$0 = \int_\Omega (w_\varepsilon - E)v\,dx + \int_\Omega \Big\{h\nabla A_\varepsilon(w_\varepsilon) + \lambda_\varepsilon(w_\varepsilon)\Big[\nabla(S_0 - hE) + \vec{V}(x, h(w_\varepsilon - E) + S_0)\Big]\Big\} \cdot \nabla v\,dx.$$

We consider possibly two solutions S_ε^1 and $\widehat{S}_\varepsilon^1$ in the statement of the problem, and w_ε and $\widehat{w}_\varepsilon$, the associated expressions. It thus occurs that

$$\begin{aligned} 0 = & \int_\Omega (w_\varepsilon - \widehat{w}_\varepsilon)v\,dx + h\int_\Omega \nabla\Big[A_\varepsilon(w_\varepsilon) - A_\varepsilon(\widehat{w}_\varepsilon)\Big] \cdot \nabla v\,dx \\ & + \int_\Omega (\lambda_\varepsilon(w_\varepsilon) - \lambda_\varepsilon(\widehat{w}_\varepsilon))\,\nabla(S_0 - hE) \cdot \nabla v\,dx \\ & + \int_\Omega (\lambda_\varepsilon(w_\varepsilon) - \lambda_\varepsilon(\widehat{w}_\varepsilon))\,\vec{V}(x, h(w_\varepsilon - E) + S_0) \cdot \nabla v\,dx \\ & + \int_\Omega \lambda_\varepsilon(\widehat{w}_\varepsilon)\Big[\vec{V}(x, h(w_\varepsilon - E) + S_0) - \vec{V}(x, h(\widehat{w}_\varepsilon - E) + S_0)\Big] \cdot \nabla v\,dx. \end{aligned}$$

By taking $v = p_\eta\,(A_\varepsilon(w_\varepsilon) - A_\varepsilon(\widehat{w}_\varepsilon))$ with

$$p_\eta(r) = 1_{]\eta,+\infty[}(r) + \ln\frac{er}{\eta}1_{[\frac{\eta}{e},\eta]}(r), \quad \eta > 0,$$

and noticing that the functions A_ε^{-1} and $\lambda_\varepsilon \circ A_\varepsilon^{-1}$, defined on $\mathbb{R}_+$, are Hölder-continuous of order $\frac{1}{2}$, we obtain that

$$\begin{aligned} \left|\int_\Omega (w_\varepsilon - \widehat{w}_\varepsilon)p_\eta\,(A_\varepsilon(w_\varepsilon) - A_\varepsilon(\widehat{w}_\varepsilon))\;dx\right| \leq & \frac{3C_1}{4h}\int_{\Omega\cap\{\frac{\eta}{e} \leq A_\varepsilon(w_\varepsilon) - A_\varepsilon(\widehat{w}_\varepsilon) \leq \eta\}} |\nabla(S_0 - hE)|^2\;dx \\ & + \frac{3}{4}\Big(\frac{C_1\|\vec{V}\|_\infty^2}{h} + C_2\Big)\int_{\Omega\cap\{\frac{\eta}{e} \leq A_\varepsilon(w_\varepsilon) - A_\varepsilon(\widehat{w}_\varepsilon) \leq \eta\}} dx. \end{aligned}$$

By the Lebesgue dominated convergence, we obtain finally, when $\eta \to 0^+$, that

$$\int_\Omega (w_\varepsilon - \widehat{w}_\varepsilon)^+\,dx = 0,$$

that is, $w_\varepsilon = \widehat{w}_\varepsilon$ and thus $S_\varepsilon^1 = \widehat{S}_\varepsilon^1$ a.e. in Ω. □

To take into account the constraints (1), one needs to know about ε going to 0^+, because $\lambda_\varepsilon \notin \mathcal{H}$. That is the object of the following section.

§3. The discrete differential inclusion

Proposition 3. *For $S^0 \in H_0^1(\Omega)$, $N > 0$ and $h = T/N$, there exists a sequence $(S^i)_{1\le i\le N} \subset H_0^1(\Omega)$ and there exists a sequence $(\lambda^i)_i$, with $\lambda^i \in \mathcal{H}\left(\frac{S^i-S^{i-1}}{h} + E\right)$, such that, for all $1 \le i \le N$ and for all $v \in H_0^1(\Omega)$,*

$$\int_\Omega \frac{S^i - S^{i-1}}{h} v\,dx + \int_\Omega \lambda^i \Big[\nabla S^i + \vec{V}(x, S^i)\Big] \cdot \nabla v\,dx = 0,$$
$$\frac{S^i - S^{i-1}}{h} + E \ge 0 \qquad \textit{a.e. in } \Omega.$$

Proof. Taking the test function $v = \int_0^{\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h}} \frac{du}{\lambda_\varepsilon(u+E)+\eta}$, with $\eta > 0$, in the formulation $(\mathcal{P}_\varepsilon)$, one gets that

$$\|S_\varepsilon^i\|_{H_0^1(\Omega)} \le 2\|S_\varepsilon^{i-1}\|_{H_0^1(\Omega)} + \|\vec{V}\|_\infty \sqrt{\text{meas}(\Omega)}.$$

And, by usual arguments (the Marcus and Mizel's lemma, the Lebesgue dominated convergence and the Green's formula), we obtain

$$\left\|\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h}\right\|_{L^2(\Omega)}^2 + \frac{1}{h}\,\|S_\varepsilon^i\|_{H_0^1(\Omega)}^2 - \frac{1}{h}\,\left\|S_\varepsilon^{i-1}\right\|_{H_0^1(\Omega)}^2 + \frac{1}{h}\,\left\|S_\varepsilon^i - S_\varepsilon^{i-1}\right\|_{H_0^1(\Omega)}^2 \le C_1 + C_2\,\|S_\varepsilon^i\|_{H_0^1(\Omega)}^2.$$

By addition, it follows

$$\sum_{i=1}^n h\left\|\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h}\right\|_{L^2(\Omega)}^2 + \|S_\varepsilon^n\|_{H_0^1(\Omega)}^2 + \sum_{i=1}^n h\,\|S_\varepsilon^i - S_\varepsilon^{i-1}\|_{H_0^1(\Omega)}^2$$
$$\le \|S_0\|_{H_0^1(\Omega)}^2 + C_1 T + C_2 \sum_{i=1}^n h\,\|S_\varepsilon^i\|_{H_0^1(\Omega)}^2.$$

So, thanks to Gronwall's lemma, one has

$$\|S_\varepsilon^i\|_{H_0^1(\Omega)}^2 \le \left(2T(C_1 + 2C_2\|\vec{V}\|_\infty^2\,\text{meas}(\Omega)) + \|S^0\|_{H_0^1(\Omega)}^2\right)\exp(1 + 8TC_2).$$

Therefore, by compactness, when $\varepsilon \to 0^+$, (up to a subsequence)

$$\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} + E \to \frac{S^i - S^{i-1}}{h} + E \qquad \text{a.e. in } \Omega,$$
$$\lambda_\varepsilon\Big(\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} + E\Big) \rightharpoonup^* \lambda^i \qquad \text{weakly* in } L^\infty(\Omega).$$

Set

$$A^i = \left\{x \in \Omega \,\middle|\, \Big(\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} + E\Big)(x) \to \Big(\frac{S^i - S^{i-1}}{h} + E\Big)(x) > 0\right\}.$$

Then $\lambda_\varepsilon\left(\frac{S_\varepsilon^i - S_\varepsilon^{i-1}}{h} + E\right) \to 1$ in A^i when $\varepsilon \to 0^+$, and $\lambda^i = 1$ in A^i. Thus, $\lambda^i \in \mathcal{H}\left(\frac{S^i - S^{i-1}}{h} + E\right)$, and one concludes. □

§4. The continuous formulation ($h \to 0^+$)

Usually, one takes the function defined on $[0,T]$ by

$$\widetilde{S}^h(t) = \sum_{i=1}^{N}\left[\frac{S^i - S^{i-1}}{h}(t-(i-1)h) + S^{i-1}\right] 1_{](i-1)h,ih]}(t), \quad \widetilde{S}^h(0) = S_0.$$

Then, the previous discretization is equivalent to the approximating equation

$$\int_Q \frac{\partial \widetilde{S}^h}{\partial t} v\, dx\, dt + \int_Q \lambda^h\Big[\nabla\widetilde{S}^h + \vec{V}(x,\widetilde{S}^h)\Big]\cdot\nabla v\, dx\, dt = o(h),$$

where $\lambda^h = \sum_{i=1}^N \lambda^i 1_{](i-1)h,ih]} \in \mathcal{H}\Big(\frac{\partial\widetilde{S}^h}{\partial t} + E\Big)$, $v \in L^2(0,T;H_0^1(\Omega))$.

Here there is an open problem, that is that we cannot pass to the limit into the term $\lambda^h\nabla\widetilde{S}^h$ when $h \longrightarrow 0^+$, because there is a double weak convergence conjecture.

§5. The 1-D case

In the 1-D case, it is proved in [5] that $\{\lambda^h\}$ is bounded in $L^\infty(Q) \cap \bar{B}V(Q)$. That implies that λ^h converges *a.e.* in Q [3], and then, thanks to Lebesgue's Theorem, the problem of passing to the limit is resolved. It leads to an original Bernoulli problem [2].

5.1. The sedimentation phenomena in marine transport

5.1.1. An hyperbolic behaviour

If

$$a\frac{\partial^2 S}{\partial t^2} + 2b\frac{\partial^2 S}{\partial t\partial x} + c\frac{\partial^2 S}{\partial x^2} + d\frac{\partial S}{\partial t} + e\frac{\partial S}{\partial x} = g,\ (t,x) \in\ ]0,T[\times\Omega,$$

then, this equation is said degenerate hyperbolic if $b^2 - ac \geq 0$. Here, we have actually

$$b^2 - ac = \frac{1}{4}\left(\lambda'(\frac{\partial S}{\partial t}(t,x))\Big[\frac{\partial S}{\partial x}(t,x) + V(x,S(t,x))\Big]\right)^2 \geq 0.$$

5.1.2. The sedimentation in marine transport

Let us consider $V = 0$ for $S \geq 0$, $E = 0$, so therefore $\frac{\partial S}{\partial t} \geq 0$ a.e. in Q.

Proposition 4. *Any solution S of the problem $(\mathcal{P})$ is such that*

(i) $\lambda(\frac{\partial S}{\partial t})\nabla S^+ = 0$ *a.e. in* Q,

(ii) for all $t \geq 0$, $S^+(t,\cdot) = S_0^+$ *a.e. in* Ω;

and, as a consequence (hyperbolic behavior),

(iii) for all $t \geq 0$, $S(t,\cdot) = S_0$ *a.e. in* $\{x \in \Omega \mid S_0(x) \geq 0\}$, *and for all* $t \geq 0$, $S(t,\cdot) \geq S_0$ *a.e. in* $\{x \in \Omega \mid S_0(x) \leq 0\}$.

5.1.3. Dead-zones and isolation effect

Let us consider $V = 0$ for $S \geq 0$, $E = 0$.

Corollary 5. *Assume that there exists a compact set K and an open set ω such that*

$$K \subset \omega \subset \Omega \text{ and } \omega \setminus K \subset \{S_0 \geq 0\}.$$

Then, for any $t \geq 0$, $S(t, \cdot) = S_0$ a.e. in ω.

Proof. Consider v in $H_0^1(\omega)$ such that $1_K \leq v \leq 1$ (lemma of Urysohn), one gets $\int_K \frac{\partial S}{\partial t}\, dx \leq 0$. Therefore, $\frac{\partial S}{\partial t} = 0$ a.e. in Ω. □

5.2. Traveling waves with mobile obstacle (finite speed)

Following [1, 4], we are looking for solutions of the form

$$S(t,x) = h(\xi), \quad \lambda(x,t) = \lambda(\xi)$$

where $\xi = x + \mu t$, $\mu > 0$, $x \in]0, L[$, $L > 0$, $t \in]0, T[$, $T > 0$, and $Q =]0, T[\times]0, L[$, subjected to a mobile obstacle of the form

$$E(t,x) = E^* 1_{Q\cap\{0\leq x+\mu t\leq \xi_0\}}(t,x) + E^{**} 1_{Q\cap\{\xi_0 < x+\mu t\leq \xi_1\}}(t,x),$$

with $E^{**} > E^* > 0$ and $0 < \xi_0 < \xi_1$.

Then, supposing $V = V(S)$, the free boundary problem

$$\begin{cases} \text{Search } (S,\lambda), \text{ with } \lambda \in \mathcal{H}(\dfrac{\partial S}{\partial t} + E) \text{ such that} \\ \dfrac{\partial S}{\partial t} - \dfrac{\partial}{\partial x}\left[\lambda\left(\dfrac{\partial S}{\partial x} + V(S)\right)\right] = 0 \quad \text{in } Q, \\ S(0,\cdot) = S_0 \quad \text{in }]0, L[\,, \end{cases}$$

may be rewritten by the non-ordinary differential system

$$\begin{cases} \text{Search } \xi \mapsto h(\xi) \text{ and } \xi \mapsto \lambda(\xi) \text{ such that} \\ \mu h'(\xi) - \{\lambda(\xi)\,[h'(\xi) + V(h(\xi))]\}' = 0, \quad \xi > 0,\ h(0) = S_0(0), \\ 0 \leq \lambda(\xi) \leq 1, \quad \mu h'(\xi) + E(\xi) \geq 0,\ (1-\lambda(\xi))(\mu h'(\xi) + E(\xi)) = 0. \end{cases}$$

Then, we deduce that

$$\exists C_0 \in \mathbb{R} \mid \forall \xi > 0,\ \mu h(\xi) - \lambda(\xi)\left[h'(\xi) + V(h(\xi))\right] = C_0.$$

- *First phase*: Inactive constraint in $Q_0 = Q \cap \{0 \leq x + \mu t \leq \xi_0\}$.
  We assume that

  $$V(S) = aS,\ a \in \mathbb{R},\ |a| < \mu,\ \lambda(\xi) = 1,\ \mu h'(\xi) + E^* \geq 0,$$

  and we obtain explicitly

  $$h(\xi) = S_0(0) + \frac{E^*}{\mu(\mu - a)}\, e^{-(\mu-a)\xi_0}\left[1 - e^{(\mu-a)\xi}\right],\ \lambda(\xi) = 1,$$
  $$\mu h'(\xi) + E^* > 0 \text{ for } \xi \in [0, \xi_0[,\ \mu h'(\xi_0) + E^* = 0.$$

- *The second phase*: Activation of the constraint

$$\mu h'(\xi) + E^{**} = 0,\ 0 \leq \lambda(\xi) \leq 1 \quad \text{if } \xi \in [\xi_0, \xi_1],$$

and we obtain after calculations the explicit formulae

$$h(\xi) = -\frac{E^{**}}{\mu}(\xi - \xi_0) + S_0(0) - \frac{E^*}{\mu(\mu - a)}\left[1 - e^{-(\mu-a)\xi_0}\right],$$
$$\lambda(\xi) = \frac{-E^{**}(\xi - \xi_0) - \frac{E^*}{\mu} + ah(\xi_0)}{-\frac{aE^{**}}{\mu}(\xi - \xi_0) + ah(\xi_0) - \frac{E^{**}}{\mu}},$$

for $\xi \in [\xi_0, \xi_1]$.

§6. Conclusion and open problems

The model states a mathematical description for the coupling of a diffusion- transport phenomenon and the weather limited erosion through a dynamic-slope approach. The emergence of hyperbolic zones and free boundaries leads to an original conservation law whose study remains still open, in particular the way to formulate the well-posedness of the problem in an appropriate framework.

References

[1] Antontsev, S. N., Gagneux, G., Luce, R., and Vallet, G. New unilateral problems in stratigraphy. *M2AN Math. Model. Numer. Anal. 40*, 4 (2006), 765–784. Available from: http://dx.doi.org/10.1051/m2an:2006029, doi:10.1051/m2an:2006029.

[2] Beurling, A. On free-boundary problems for the Laplace equation. In *Seminars on Analytic Functions* (Princenton, 1958), Institute for Advanced Study, pp. 248–263.

[3] Evans, L. C., and Gariepy, R. F. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[4] Gagneux, G., and Vallet, G. Sur des problèmes d'asservissements stratigraphiques. *ESAIM Control Optim. Calc. Var. 8* (2002), 715–739 (electronic). A tribute to J. L. Lions. Available from: http://dx.doi.org/10.1051/cocv:2002055, doi:10.1051/cocv:2002055.

[5] Vallet, G. Sur une loi de conservation issue de la géologie. *C. R. Math. Acad. Sci. Paris 337*, 8 (2003), 559–564. Available from: http://dx.doi.org/10.1016/j.crma.2003.08.012, doi:10.1016/j.crma.2003.08.012.

Mohamed-Salem Louly
Laboratoire de Mathématiques Appliquées - Université de Pau et des Pays de l'Adour
UMR-CNRS 5142, I.P.R.A, B.P. 1155, 64013 Pau Cedex, France
mohamedsalemlouly@yahoo.fr

Monografías Matemáticas *García de Galdeano* **35**, 197–204 (2010)

Analytical and numerical methods in stratigraphy

Amar Mokrani, Abdelaziz Taakili and Guy Vallet

Abstract. In this paper, we are interested in a mathematical problem arising from the modelling of maximal erosion rates in geological stratigraphy. The problem is nonlinear with a diffusion coefficient that is a nonlinear function of u and $\partial_t u$. Moreover, the problem degenerates in order to take implicitly into account a constraint on $\partial_t u$. Our aim in this paper is to present a survey of the results exposed in the oral communication and written in the PhD theses [11] and [13].

Keywords: Stratigraphy, discontinuous Galerkin method, constraint, pseudoparabolic.

AMS classification: 35K70, 65N30, 65N12..

§1. Introduction and mathematical model

This work deals with the study of a mathematical model arising from the modelling of geological basin formation. It takes into account sedimentation, transport and accumulation, erosion phenomena, and others. The original mathematical aspect of this model is the imposition of a constraint on the time-derivative of the unknown u.

Let us consider a sedimentary basin and denote by Ω its basis. It is assumed to be a fixed bounded domain of $\mathbb{R}^N$ $(N = 1, 2)$ with a Lipschitz boundary Γ. As usual, for any positive T, we set $Q =]0, T[\times \Omega$. In the model, u, the sediments thickness, naturally satisfies the mass balance equation

$$\partial_t u + \operatorname{div}(\vec{q}) = f \text{ in } Q, \tag{1}$$

where f denotes a source term (modelling, for example, of suspension matter in the sea that sediments in the domain). According to Darcy-Barenblatt's law (see [8] for example), the flux $\vec{q}$ is given by the relation

$$\vec{q} = -\lambda K(u)\nabla(u + \tau\partial_t u), \tag{2}$$

where λ is a parameter to be defined later and τ is a positive time-scaled parameter.

In a sedimentary basin formation process, sediments must first be produced *in situ* by weathering effects prior to be transported by surfacing erosion. Thus, a constraint on a maximum erosion rate $-\partial_t u \leq E$ in Q has to be introduced (see [9]), where E is non-negative. It takes into account the composition, the structure and the age of the sediments. In their paper, the authors consider a flux limiter λ, $0 \leq \lambda \leq 1$ that satisfies

$$\partial_t u + E \geq 0, \ (1 - \lambda)(\partial_t u + E) = 0, \ \text{a.e in } Q. \tag{3}$$

Following [2], one remarks that, as soon as $f + E \geq 0$, for all $u \in L^2(0, T; H_0^1(\Omega))$ with $\partial_t u \in L^2(0, T; H_0^1(\Omega))$, (1)–(3) is equivalent to the following formulation:

$$\partial_t u - \operatorname{div}[\lambda K(u)\nabla(u + \tau\partial_t u)] = f \text{ in } Q, \ \lambda \in H(\partial_t u + E) \cap L^\infty(Q). \tag{4}$$

Here homogeneous Dirichlet boundary conditions on u and $\partial_t u$ are considered, $u(0,.) = u_0 \in H^1_0(\Omega)$, $E \in L^\infty(0,T;H^1(\Omega))$, $f \in L^\infty(0,T;L^2(\Omega))$ and H denotes the maximal monotone graph of the Heaviside function.

Indeed, if $f + E \geq 0$ is assumed, using the admissible test function $(\partial_t u + E)^-$ (where $x^- = -\min(0,x)$ for $x \in \mathbb{R}$) in (4), we get that $\partial_t u + E \geq 0$ a.e in Q since $\lambda \mathbf{1}_{\{\partial_t u + E < 0\}} = 0$ a.e., and therefore (3) and $\lambda \in H(\partial_t u + E)$ are equivalent assertions. Moreover, using that $\partial_t u + E \geq 0$, one has that

$$\lambda \nabla(u + \tau \partial_t u) = \lambda \nabla[u - \tau E + \tau(\partial_t u + E)] = \lambda \nabla(u - \tau E) + \tau \nabla(\partial_t u + E).$$

Thus, the problem (4) is equivalent to the following one:

$$\partial_t u - \operatorname{div}\left\{\lambda\, K(u)[\nabla u - \tau \nabla E]\right\} - \tau \operatorname{div}\left\{K(u)[\nabla \partial_t u + \nabla E]\right\} = f,\ \lambda \in H(\partial_t u + E) \cap L^\infty(Q). \quad (5)$$

Results of existence and uniqueness of a solution to such a problem is still an open question. Thus, a modified model where H is replaced by a continuous function a, for example the Yosida approximation of H, will be proposed. Such a problem has been analysed by S. N. Antontsev *et al.* [2] with $K \equiv 1$, $f \equiv 0$ and a constant E. Then, a result of existence and uniqueness of a solution is given in [11, 3] with a null source term and a time dependant function E. While a result of existence of a solution and a numerical scheme based on the discontinuous Galerkin methods (DgFem) are considered in [13, 4] with a source term, a space-time function E and $K = 1$.

Note that the above remark concerning the equivalence between (4) and (5) doesn't hold anymore if one replace λ by $a(\partial_t u + E)$ *i.e.* equivalence between (6) and (8). The two problems have not got the same nature. Indeed, thanks to the localisation methods proposed in the book of S.N. Antontsev *et al.* [1], it has been proved in [11], that under some hypothesis on a, any weak solution u to the 1-D problem:

$$\partial_t u - \partial_x\left\{K(u) a(t, \partial_t u)[\partial_x u + \tau \partial_{xt} u]\right\} = f \quad \text{in } Q =]0,T[\times\Omega \quad \text{with } \Omega =]0,1[, \quad (6)$$

with the boundary and initial conditions:

$$\partial_t u_{|_{t=0}} = 0, \quad u(0,x) = u_0(x),\ x \in \Omega, \quad \text{where} \quad u_0(x) = 0,\ x \in]0,\rho_0[,\ \ 0 < \rho_0 < 1, \quad (7)$$

there exist a positive $\delta > 0$ and $\rho(t) \in (0,\rho_0)$, such that, if $f(t,x) = 0$ in $]0,\delta[\times]0,\rho_0[$, then u satisfies the finite speed of propagation property: $u(t,x) = 0$ in $x \in]0,\rho(t)[$, $0 \leq t \leq \delta$.

But this locally hyperbolic behaviour is unknown concerning the pseudo-parabolic problem:

$$\partial_t u - \operatorname{div}\left\{a(\partial_t u + E)\, K(u)[\nabla u - \tau \nabla E]\right\} - \tau \operatorname{div}\left\{K(u)[\nabla \partial_t u + \nabla E]\right\} = f \text{ in } Q. \quad (8)$$

On the one hand, both (6) and (8) are approximations of the same problem when a converges toward the graph of Heaviside. On the other hand, they reveal distinct natures.

From now on, one would consider the pseudo-parabolic problem (8) with an initial height given by: $u(0,\cdot) = u_0$ in Ω, where $u_0 \in H^1_0(\Omega)$, and homogeneous Dirichlet condition for u and $\partial_t u$.

Contrarily to the perturbation (6), the main interest of (8) is that it will be possible to use the theorems of N.G. Meyers and J. Nečas, in order to obtain a more regular solution (i.e $u \in W^{1,\infty}(0,T;W_0^{1,p}(\Omega))$, with $p > 2$ as soon as $u_0 \in W_0^{1,p}$) and thus a uniqueness result.

Let us set the assumptions made on the data and the definition of a solution

$$\textbf{(H)}: \begin{cases} \tau > 0;\ E \in L^\infty(0,T;H^1(\Omega)), E \geq 0;\ f \in L^\infty(0,T;L^2(\Omega)),\ f + E \geq 0 \text{ in } Q; \\ a \in C^{0,\theta}(\mathbb{R}), \text{ with } \theta \geq 1/2,\ 0 \leq a \leq 1,\ \forall x \in \left]-\infty, 0\right],\ a(x) = 0; \\ K : \mathbb{R} \to \mathbb{R} \text{ is a Lipschitz function, with } 0 < K_{\min} \leq K \leq K_{\max}. \end{cases}$$

Definition 1. Under assumption **(H)**, a solution to problem (8) is any u in $W^{1,2}(0,T;H_0^1(\Omega))$ such that for any v in $H_0^1(\Omega)$ and for a.e. t in $]0,T[$,

$$\int_\Omega \partial_t u v\, dx + \int_\Omega K(u) a(\partial_t u + E)(\nabla u - \tau \nabla E)\cdot \nabla v\, dx + \tau \int_\Omega K(u)(\nabla \partial_t u + \nabla E)\cdot \nabla v\, dx = \int_\Omega f v\, dx$$

with the initial condition $u(t = 0) = u_0$ a.e. in Ω.

Since a vanishes on $\mathbb{R}^-$, as previously shown with λ, the constraint $\partial_t u + E \geq 0$ in Q is implicitly satisfied.

The sequel of this paper is organised as follows: in Section 2, we present the result of existence and uniqueness of the solution to the problem (8). Then, in Section 3, the DgFem for the model is introduced. It is construct in order to satisfy implicitly the constraint (3) in the lowest-order case. A last section is concerned by numerical results.

§2. Existence and uniqueness result

In this section, we present the result of existence and uniqueness of a solution to problem (8).

Theorem 1. *Assume* (**H**). *For any* $u_0 \in H_0^1(\Omega)$, *a solution u to the problem* (8) *exists in the sense of the definition* 1 *and is given in the space* $W^{1,\infty}(0,T;H_0^1(\Omega))$. *Moreover, the constraint* $\partial_t u + E \geq 0$ *a.e. in Q is implicitly satisfied.*

Following [4, 11, 13], the result is based on the implicit time discretization:

$$\int_\Omega \frac{u^k - u^{k-1}}{h} v + K(u^k)\Big[a\Big(\frac{u^k - u^{k-1}}{h} + E^k\Big)(\nabla u^k - \tau\nabla E^k) + \tau\Big(\nabla\frac{u^k - u^{k-1}}{h} + \nabla E^k\Big)\Big]\cdot\nabla v\, dx = \int_\Omega f^k v\, dx.$$

The existence of u^k comes from Schauder's fixed point theorem. Then, one notes that it is the unique solution in $H_0^1(\Omega)$ to the elliptic problem: $-\operatorname{div}[\alpha\nabla u^k] = f_0 - \operatorname{div}\vec{f}$, with non constant and symmetrical coefficients α and suitable f_0 and $\vec{f}$ (see [3] for more details).

By applying the theorem of Meyers [12], we get, following [11], that

Lemma 2. *Assuming that* $u^{k-1} \in W_0^{1,p_0}(\Omega)$ *and that* $f^k \in L^{p_0}(\Omega)$ *with* $p_0 > 2$, *there exists a real* $\overline{p}(p_0) > 2$, *depending on* p_0 *and* $\frac{K_{\max}(a_{\max}+\tau/h)}{(\tau/h)K_0}$, *and a positive constant* $C(\overline{p}(p_0))$ *such that*

$$u^k \in W_0^{1,\overline{p}(p_0)}(\Omega) \quad \text{and} \quad \|\nabla u^k\|_{L^{\overline{p}(p_0)}(\Omega)^N} \leq C(\overline{p}(p_0))\Big(\|u_0\|_{W_0^{1,p_0}(\Omega)}, \|f\|_{L^{p_0}(\Omega)}\Big).$$

Indeed, $K_0\tau \leq \alpha \leq K_{\max}[a_{\max}h + \tau]$ with $h << 1$. Since the dimension N is 1 or 2, estimation

$$\|\nabla \frac{u^k - u^{k-1}}{h}\|_{L^2(\Omega)^N} \leq C\Big(K_0, K_{\max}, a_{\max}, \tau, T, \|\nabla u_0\|_{L^2(\Omega)^N}, \|f\|_{L^2(\Omega)}\Big)$$

yields that $f_0 \in L^{p_0}(\Omega)$. Moreover, $\vec{f_1} \in L^{p_0}(\Omega)^N$, and Meyers' theorem [5] leads to the assertion.

Still following [11], by using the theorem of Nečas, the regularity $W_0^{1,p_0}(\Omega)$ can be obtained.

Lemma 3. *If $u_0 \in W_0^{1,p_0}(\Omega)$ with $p_0 > 2$, then, $u^k \in W_0^{1,p_0}(\Omega)$, for any $k = 1, \ldots, N$, and there exists a positive constant $C(p_0)$ such that*

$$\|\nabla u^k\|_{L^{p_0}(\Omega)^N} \leq C(p_0)\Big(\|u_0\|_{W_0^{1,p_0}(\Omega)}, \|f\|_{L^{p_0}(\Omega)}\Big).$$

Indeed, since $N \leq 2$, the Sobolev embedding ensures that α is a continuous function on $\overline{\Omega}$ and the expected regularity comes from Nečas theorem [7, 12]. Thus, the assertion yields by using the discrete Gronwall's lemma.

Thanks to those lemmata, following [13, 11], suitable *a priori* estimates hold to get that

Theorem 4. *If $u_0 \in W_0^{1,p_0}(\Omega)$ and $f \in L^2(0, T; L^{p_0}(\Omega))$, for a given $p_0 > 2$, then, any limit point u in $W^{1,\infty}(0, T; H_0^1(\Omega))$ of the implicit time discretized scheme satisfies moreover that $u \in W^{1,\infty}(0, T; W_0^{1,p_0}(\Omega))$.*

Then, one is able to adapt the demonstrations involving tri-linear terms and based on Hölder type inequalities to prove that

Theorem 5. *There exists a real $\tau^* \geq 0$, such that for any $\tau > \tau^*$, the problem (8) has a unique solution in the space $W^{1,\infty}(0, T; W_0^{1,p_0}(\Omega))$ with $p_0 > 2$. Moreover, the application: $u_0 \mapsto \partial_t u$ is a locally Lipschitz continuous function in the space $H_0^1(\Omega)$ to the space $L^\infty(0, T; H_0^1(\Omega))$.*

§3. Space DgFem discretization

In this section, we consider a numerical scheme for the computation of the semi-discretized problem. Our approach is based on the discontinuous Galerkin finite element method (DgFem). For convenience, we assume in the sequel that E is a non-negative constant, and that the function a is Lipschitz-continuous. Before discretizing the problem, some notations are collected.

We suppose that $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain and that h is a regular triangular mesh in a family of shape-uniform meshes [6].

We denote by $\mathcal{K}_h$ the set of triangles and by $\mathcal{S}_h$ the set of edges, divided into interior edges $\mathcal{S}_h^{int}$ and boundary edges $\mathcal{S}_h^{\partial}$. An interiori edge $S \in \mathcal{S}_h^{int}$ is shared by two triangles, we arbitrarily chose a normal n_S pointing from K^+ to K^-. For $p \in \mathbb{N}$, we define the discontinuous finite element space:

$$V_h^p = \Big\{v_h \in L^2(\Omega) \,:\, v_h|_K \in P^p \text{ for all } K \in \mathcal{K}_h\Big\},$$

where P^p denotes the space of polynomials functions of maximal degree p. Due to the discontinuity of the approximation space, the weak formulation reveals terms of jumps through the cell interfaces. We make use of the standard notations concerning the jumps and averages for $v_h \in V_h^p$, $S \in \mathcal{S}_h^{int}$, and $x \in S$:

$$v_h^{\pm}(x) = \lim_{\varepsilon \to 0^+} v_h(x \pm \varepsilon n_S) \quad \text{and} \quad [v_h]_S = v_h^+ - v_h^-.$$

For a boundary edge we set $[v_h]_S := v_h^-$.

In addition, for any bounded positive piece-wise continuous function κ with respect to h, we define the weighted average of $v_h \in V_h$ on an interior edge S as

$$\left\{\frac{\partial v_h}{\partial n}\right\}_{S,\kappa} = \frac{\kappa^- \kappa^+}{\kappa^+ + \kappa^-}\left(\frac{\partial v_h^+}{\partial n_S}\bigg|_S + \frac{\partial v_h^-}{\partial n_S}\bigg|_S\right).$$

Let us now consider the following time semi-discrete problem which reads: Find $u^{k+1} \in H_0^1(\Omega)$ such that for all v in $H_0^1(\Omega)$

$$\left\{\begin{aligned} &\frac{1}{\Delta t}\int_\Omega u^{k+1} v\,dx + \int_\Omega D(u^{k+1})\nabla u^{k+1}\cdot\nabla v\,dx + \frac{\tau}{\Delta t}\int_\Omega K(u^{k+1})\nabla u^{k+1}\cdot\nabla v\,dx \\ &\qquad = \int_\Omega f^{k+1} v\,dx + \frac{1}{\Delta t}\int_\Omega u^k v\,dx + \frac{\tau}{\Delta t}\int_\Omega K(u^{k+1})\nabla u^k\cdot\nabla v\,dx, \\ &\qquad\qquad u^0 = u_0 \ \text{ in } \Omega, \end{aligned}\right. \tag{9}$$

where $D(w) := a(\frac{w-u^k}{\Delta t} + E)K(w)$.

The discrete DgFem formulation of problem (9) reads: Find $u_h^{k+1} \in V_h^p$ such that for all $v_h \in V_h^p$

$$A(u_h^{k+1})(u_h^{k+1}, v_h) = L^k(u_h^{k+1})(v_h), \tag{10}$$

where the bilinear form A and the linear form L are given for $\rho_h \in V_h^p$ by

$$\begin{aligned} A(\rho_h)(u_h, v_h) = {}& \frac{1}{\Delta t}\int_\Omega u_h v_h\,dx + \sum_{K\in\mathcal{K}_h}\int_K\Big(D(\rho_h) + \frac{\tau}{\Delta t}K(\rho_h)\Big)\nabla u_h\cdot\nabla v_h\,dx \\ &+ \sum_{S\in\mathcal{S}_h}\int_S\left(\frac{\gamma}{h_S}\gamma_S[u_h][v_h] - \left\{\frac{\partial u_h}{\partial n_S}\right\}_{S,D}[v_h]_S - [u_h]_S\left\{\frac{\partial v_h}{\partial n_S}\right\}_{S,D}\right)ds, \\ &- \frac{\tau}{\Delta t}\sum_{S\in\mathcal{S}_h}\int_S\left(\left\{\frac{\partial u_h}{\partial n_S}\right\}_{S,K}[v_h]_S + [u_h]_S\left\{\frac{\partial v_h}{\partial n_S}\right\}_{S,K}\right)ds, \end{aligned}$$

and

$$\begin{aligned} L^k(\rho_h)(v_h) = {}& \int_\Omega f^{k+1} v_h\,dx + \frac{1}{\Delta t}\int_\Omega u_h^k v_h\,dx + \frac{\tau}{\Delta t}\sum_{K\in\mathcal{K}_h}\int_K K(\rho_h)\nabla u_h^k\cdot\nabla v_h\,dx \\ &+ \sum_{S\in\mathcal{S}_h}\int_S\left(\frac{\gamma}{h_S}\beta_S[u_h^k][v_h] - \frac{\tau}{\Delta t}\left\{\frac{\partial u_h^k}{\partial n_S}\right\}_{S,K}[v_h]_S - \frac{\tau}{\Delta t}[u_h^k]_S\left\{\frac{\partial v_h}{\partial n_S}\right\}_{S,K}\right)ds, \end{aligned}$$

where $\beta_S = \frac{2K^+K^-}{K^-+K^+}$, $\gamma_S = \beta_S + \frac{2D^+D^-}{D^++D^-}$ and $\gamma > 0$ has to be chosen large enough.

In the lowest-order case ($p = 0$), we assume that the triangulation h satisfies the classical angle condition given in [10]. In this case, the bilinear form A reduces to

$$A(\rho_h)(u_h, v_h) = \frac{1}{\Delta t}\int_\Omega u_h v_h \, dx + \sum_{S\in\mathcal{S}_h} \frac{1}{h_S}\int_S \gamma_S [u_h][v_h]\, ds.$$

As well as for the continuous formulation, we are able to say

Proposition 6. *Assume* (**H**)*, the problem* (10) *has at least one solution. In addition, if τ sufficiently large, the solution is unique and if $p = 0$, then, we have for $k = 0, \dots, N-1$,*

$$\frac{u_h^{k+1} - u_h^k}{\Delta t} + E^{k+1} \geq 0 \ a.e \ in \ \Omega.$$

Proof. The proof of this result is given with more detail in [4] in the case $K \equiv 1$, using hypothesis on K this result can be generalised to the case $K := K(u)$. □

§4. Numerical results

In the numerical example, we consider Problem (8) in the domain $\Omega =]-1, 1[^2$, for $0 \leq t \leq T$, with homogenous Dirichlet condition, a null source term and we consider the initial height $u_0(x, y) = -\sin(\pi x)\sin(\pi y)$ and $a(z) = a_\varepsilon(z) = \inf(1, [\frac{3z^2}{\varepsilon^2}(1 - \frac{2z}{3\varepsilon})]^+)$, $\varepsilon > 0$.

Assume that $K \equiv 1$, the meshes are obtained by uniform refined from a coarse mesh h_0, verifying the angle condition required for $p = 0$. In the practice, we use the algorithm of Newton with line search to solve the nonlinear system of equations. Numerical experiments show that, the convergence of Newton algorithm is very slow if the parameter τ is very small.

4.1. Mesh stability study

In this section, we study the stability behaviour of the mesh in the L^2 norm. Since there is no exact known solution for this problem, nor benchmark, the "error" would be understood by the L^2-difference between a calculate solution u_h and a reference solution, denote by u_h^*, obtained by solving the problem (8) using quadratic DgFem scheme in a fine mesh with 57344 elements.

For numerical runs, we choose $T = 1$, $\varepsilon = 10^{-1}$, $\Delta t = 10^{-1}$ and $\tau = 10^{-1}$. We represent in Table 1 the L^2-norm of this so called error as a function of h, at time $t = T$. This table confirms the $p + 1$-order behaviour of the scheme, excepted when $p = 2$ which may depends of the regularity of the solution.

4.2. Numerical simulations

In this section, we present the numerical solution obtained by using DgFem(0) scheme. Our attention is to test numerically the discrete version of the constraint. For numerical runs, we choose $\varepsilon = E = 10^{-1}$, $\Delta t = \tau = 10^{-1}$. Figure 1 represents the numerical solution at different time t and the corresponding discrete constraint

This numerical simulations show that, the constraint is satisfied in all the domain, which confirms our theoretical result.

Ne	$\|\|u_h^* - u_h\|\|_{L^2(\Omega)}$			convergence rate		
	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
896	$1.51e - 1$	$3.57e - 2$	$1.48e - 2$	–	–	–
3584	$6.95e - 2$	$9.47e - 3$	$1.48e - 3$	1.11	1.91	2.90
14336	$3.32e - 2$	$2.42e - 3$	$2.25e - 4$	1.06	1.96	2.70

Table 1: L^2 norm of the “error” with respect to h and p.

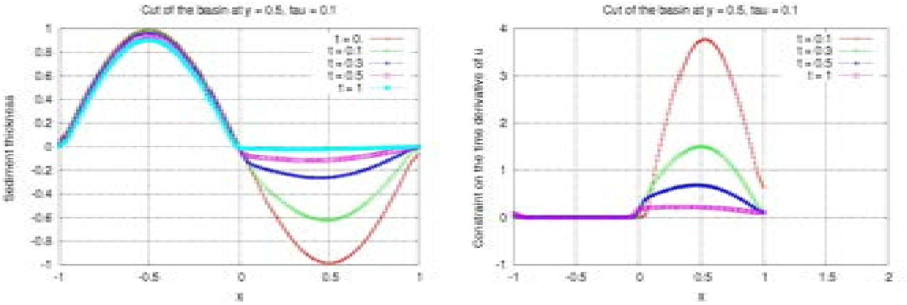


Figure 1: Vertical 1D cut at $y = 0.5$ of the numerical solution (left) and the constraint (right) with $\tau = 0.1$ and $p = 0$.

§5. Conclusion

In this survey, we have presented a result of existence and uniqueness of the solution to a realistic problem where the diffusion coefficient depends of the unknown u. However, many open questions still have to be treated, as to deal with the physics boundary conditions, *i.e.* nonhomogenuous Neumann boundary conditions on the inward part and unilateral boundary conditions on the outward one. Concerning the numerical aspect, we have presented a numerical scheme that implicitly takes into account the constraint on the time-derivative of the unknown. It's well known that a higher order scheme doesn't keep this important property, an adaptive algorithm that combines h-refinement with $p = 0$ and p-refinement has to be proposed in order to get a more accuracy, while still verifying the the monotonocity.

References

[1] Antontsev, S., Díaz, J., and Shmarev, S. *Energy Methods for Free Boundary Problems “Applications to Nonlinear PDEs and Fluid Mechanics”*. Progress in Nonlinear Diff. Equ. and Appl **48**. Basel, Birkhäuser, 2002.

[2] Antontsev, S. N., Gagneux, G., Luce, R., and Vallet, G. On a pseudoparabolic problem with constraint. *Differential & Integral Equations 19*, 12 (2006), 1391–1412.

[3] Antontsev, S. N., Gagneux, G., Mokrani, A., and Vallet, G. Stratigraphic modelling by the way of a pseudoparabolic problem with constraint. *Advances in Mathematical Sciences and Applications (Japan) 19* (2009).

[4] Becker, R., Taakili, A., and Vallet, G. A discontinuous galerkin method for a model from stratigraphy (sumitted).

[5] Bensoussan, A., Lions, J., and Papanicolaou, G. Asymptotic analysis for periodic structures. *North-holland, Amsterdam* (1978).

[6] Ciarlet, P. *The finite element method for elliptic problems.* Studies in Mathematics and its Applications. Vol. 4. Amsterdam - New York - Oxford: North-Holland Publishing Company., 1978.

[7] Clain, S. Elliptic operators of divergence type with hölder coefficients in fractional sobolev spaces. *Rendiconti di Matematica 17*, 7 (1997), 207–236.

[8] Cuesta, C., Duijn, C. J. V., , and Hulshof, J. Infiltration in porous media with dynamic capillary pressure: Travelling waves. *Eur. J. Appl. Math. 11*, 4 (2000), 381–397.

[9] Eymard, R., Gallouët, T., Granjeon, D., Masson, R., and Tran, Q. Multilithology model under maximum erosion rate constraint. *Internat. J. Numer. Methods Engrg. 60*, 2 (2004), 527–548.

[10] Eymard, R., Gallouet, T., and Herbin, R. *Finite volume methods.* 2000.

[11] Mokrani, A. *Problèmes pseudo-paraboliques à vitesse asservie. Applications en prospection pétrolière*. PhD thesis, Université de Pau, 2008.

[12] Nečas, J. Sur une loi de conservation issue de la géologie. *Collection Recherche et Mathématiques Appliquées*, 10 (1989).

[13] Taakili, A. *Méthode de Galerkin discontinue pour un modèle stratigraphique*. PhD thesis, Université de Pau, 2008.

Amar Mokrani
IECN, Université Henri Poincaré
BP 239-54506 Vandoeuvre-lès-Nancy
`Amar.Mokrani@iecn.u-nancy.fr`

Abdelaziz Taakili
INRIA, Paris-Rocquencourt
Bâtiment 13, 78150 Rocquencourt
`abdelaziz.taakili@inria.fr`

Guy Vallet
LMA de Pau UMR 5142
IPRA, BP 1155 64013 Pau Cedex
`guy.vallet@univ-pau.fr`

Monografías Matemáticas *García de Galdeano* **35**, 205–211 (2010)

A FRACTAL PROCEDURE FOR THE COMPUTATION OF MIXED INTERPOLANTS

María Antonia Navascués and María Victoria Sebastián

Abstract. We develop a procedure from the fractal methodology for the computation of an interpolant born from the cooperation of two functions of different nature. In particular, we define an Iterated Function System whose attractor is a fractal interpolant to a set of data with mixing properties. If the maps of the System are chosen in a suitable way, the approximant constructed is differentiable.
Since the degree of smoothness can be a priori fixed, the methodology described may be used in order to reduce the regularity of the classical interpolants as polynomial, splines, etc.

Keywords: Fractal interpolation functions, iterated function systems.

AMS classification: 28A80, 58C05, 65D05.

§1. Introduction

In this paper we propose a procedure for the definition of smooth fractal functions of interpolation, whose degree of regularity can be fixed a priori. The function is defined as the perturbation of a classical mapping with a criterion of proximity to another. In this way, the approximant constructed comes from the cooperation of two classical elements and the methodology of iterated funtion systems. After the construction, we give an upper bound of the uniform error committed on a compact interval.

In a second part we present an application of the procedure to the field of the numerical integration. In particular we propose a correction to the polynomial quadrature formulae for partitions with small number of points.

§2. Fractal Functions

In former papers ([4], [5]), we have studied an Iterated Function System $\{w_n(t,x)\}_{n=1}^N$ defined on the set $C = I \times \mathbb{R}$, where I is a compact interval, $I = [a,b] \subset \mathbb{R}$. The maps $w_n(t,x)$ are defined by

$$w_n(t,x) = (L_n(t), F_n(t,x)) \quad \forall\, n = 1,2,...,N,$$

where

$$\begin{cases} L_n(t) = a_n t + b_n, \\ F_n(t,x) = \alpha_n x + q_n(t). \end{cases} \tag{1}$$

The system is associated with a partition of the interval I

$$\Delta : a = t_0 < t_1 < \cdots < t_N = b.$$

The coefficients a_n and b_n are defined in terms of the nodes of the partition as

$$a_n = \frac{t_n - t_{n-1}}{t_N - t_0}, \qquad b_n = \frac{t_N t_{n-1} - t_0 t_n}{t_N - t_0}, \tag{2}$$

and $F_n(t,x)$ satisfies some Lipschitz conditions ([1]). The multiplier α_n is a vertical scale factor of the transformation, such that $-1 < \alpha_n < 1$. $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ is the scale vector.

Theorem 1. *[1, 2]: The iterated function system (IFS) defined above admits a unique attractor G. G is the graph of a continuous function $h : I \to \mathbb{R}$ interpolating the data ($h(t_n) = x_n$, for all $n = 0, 1, \ldots, N$).*

The previous function is called a fractal interpolation function (FIF) corresponding to $\{(L_n(t), F_n(t,x))\}_{n=1}^N$. It satisfies the functional equation:

$$h(t) = F_n(L_n^{-1}(t), h \circ L_n^{-1}(t)). \tag{3}$$

In this paper we study a particular case of a Fractal Interpolation Function (FIF). The maps q_n are defined as

$$q_n(t) = g \circ L_n(t) - \alpha_n b(t), \tag{4}$$

where g and b are continuous functions, $g, b : I \to \mathbb{R}$, such that $b(t_0) = g(t_0)$, $b(t_N) = g(t_N)$.

The attractor of the system is the graph of a continuous function $g^\alpha : I \to \mathbb{R}$ which interpolates to g at the nodes of the partition,

$$g^\alpha(t_n) = g(t_n) \qquad \forall\, n = 0, 1, \ldots, N. \tag{5}$$

The mapping g^α satisfies the functional equation (3)

$$g^\alpha(t) = g(t) + \alpha_n\, (g^\alpha - b) \circ L_n^{-1}(t) \qquad \forall\, t \in I_n. \tag{6}$$

Let $\mathcal{G}$ be the set of continuous functions

$$\mathcal{G} = \{f \in C[a,b] : f(t_0) = g(t_0), f(t_N) = g(t_N)\}.$$

$\mathcal{G}$ is a complete metric space with respect to the uniform norm. Define a mapping $T^\alpha : \mathcal{G} \to \mathcal{G}$ by

$$(T^\alpha f)(t) = F_n(L_n^{-1}(t), f \circ L_n^{-1}(t)). \tag{7}$$

for all $t \in [t_{n-1}, t_n]$, $n = 1, 2, \ldots, N$.

T^α is a contraction mapping on the metric space $(\mathcal{G}, \|\cdot\|_\infty)$ and possesses a unique fixed point on $\mathcal{G}$, that is the FIF g^α.

The uniform distance between g^α and g is bounded in terms of the scale vector ([6]) and the map b,

$$\|g^\alpha - g\|_\infty \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|g - b\|_\infty \tag{8}$$

where $\|\cdot\|_\infty$ is the uniform norm defined as

$$\|f\|_\infty = \max\{|f(t)| : t \in I\} \tag{9}$$

and

$$|\alpha|_\infty = \max\{|\alpha_n| : \ n = 1, 2, ..., N\} \tag{10}$$

is the contractivity factor of the transformation T^α.

Sufficient conditions for the smoothness of order p of g^α are (see the reference [7]):

$$g, b \in C^p(I) \quad \text{and} \quad \begin{cases} g^{(r)}(t_0) = b^{(r)}(t_0), \\ g^{(r)}(t_N) = b^{(r)}(t_N), \end{cases} \quad r = 0, 1, \ldots, p, \tag{11}$$

$$|\alpha|_\infty < \frac{1}{N^p},$$

$$\alpha_n = \text{cte} \qquad \forall\, n = 1, 2, \ldots, N.$$

In order to satisfy the condition (11), we can choose as b a Hermite polynomial osculating g at the extremes of the interval I.

§3. Correction of a classical interpolant with fractal methodology

In this section we present an interpolant born from the cooperation of two approximants of different nature, first developed in previous works [8, 3]. The fractal function is defined first as perturbation of one classical. The additional condition of proximity to another interpolant provides a problem of convex optimization whose solution is a fractal element with mixing properties.

Theorem 2 (Collage Theorem [2]). *Let (X, d) be a complete metric space and let T be a contraction map on X with contractivity factor $c \in [0, 1)$. Then, for any $f \in X$*

$$d(f, \tilde{f}) \leq \frac{1}{1-c}\, d(f, Tf),$$

where $\tilde{f}$ is the fixed point of T.

We consider two classical interpolants (S and P) of a set of data. We construct the fractal function P^α associated to P, defined in the previous section ($g = P$). Now we apply the collage theorem for $X = \mathcal{G}$, $f = S$, $\tilde{f} = P^\alpha$ and $T = T^\alpha$.

The distance here is the uniform metric and $T = T^\alpha$ is the contraction (7), so that $||T^\alpha S - S||_\infty < \varepsilon$ implies $||S - P^\alpha||_\infty < \frac{\varepsilon}{1-|\alpha|_\infty}$ and P^α will be a fractal interpolant close to S.

We look for a smooth function, for instance $P^\alpha \in C^1(I)$, and then we may set the problem of finding α^* solving the optimization

$$\min_\alpha ||T^\alpha S - S||_\infty = \min_\alpha c(\alpha)$$

where $|\alpha|_\infty \leq \delta < 1/N$, according to the condition 2 for the smoothness of P^α. The map b must have a contact of first order with P at the extremes of the interval.

The classical interpolants S (polynomial, spline) are piecewise smooth and consequently by the definition of T^α, $T^\alpha S - S$ also is. $c(\alpha)$ is non-differentiable in general, but its convexity can be proved and thus, the problem

$$(CP) \begin{cases} \min\limits_\alpha c(\alpha), \\ |\alpha|_\infty \leq \delta < 1/N, \end{cases}$$

is a constrained convex optimization problem. The existence of solution is clear if c is a continuous function as $\mathcal{B}_\delta = \{\alpha \in \mathbb{R}^N : |\alpha|_\infty \leq \delta < 1/N\}$ is a compact set of $\mathbb{R}^N$. In a previous paper [8] we proved that c is continuous, and (CP) convex, so that (CP) is a problem of constrained convex optimization with some solution.

If α^* is the optimum scale ($\alpha^* = \alpha_n$, $\forall n = 1, 2, \ldots, N$), the expression $c(\alpha^*)/(1 - |\alpha^*|_\infty)$ provides an upper bound of the uniform distance $\|P^{\alpha^*} - S\|_\infty$ according to the Collage Theorem.

Figures 1 and 2 display a polynomial interpolant P and a cubic spline S (respectively) to the set of data $D = \{(0, 1), (1/4, 5), (1/2, 2), (3/4, 4), (1, 3)\}$. Figure 3 shows the corresponding fractal P^{α^*} defined by the method described. The order of regularity is $p = 1$. The loss of smoothness can be observed.

The following result provides an upper bound of the uniform error of the fractal interpolant P^{α^*} with respect to the original function X.

Theorem 3. *If $X(t)$ is the original continuous function providing the interpolation data and α^* is the optimum scale, the following error estimate is obtained:*

$$\|X - P^{\alpha^*}\|_\infty \leq E_P + \frac{l^4}{(N-1)4!2^4}\|P^{(4)}\|_\infty, \tag{12}$$

where E_P is an upper bound of the interpolation error corresponding to P, l is the length of the interval I, $N+1$ is the number of points of the partition and b (4) is a Hermite polynomial with a contact of first order with P at the extremes of the interval.

Proof. It is clear that

$$\|X - P^{\alpha^*}\|_\infty \leq \|X - P\|_\infty + \|P - P^{\alpha^*}\|_\infty.$$

In the reference [7] (expression (2.53) for $k = 0$ and $p = 1$) it is proved that

$$\|P - P^{\alpha^*}\|_\infty \leq \frac{|\alpha^*|}{1 - |\alpha^*|}\,\frac{l^4}{4!\,2^4}\,\|P^{(4)}\|_\infty.$$

The inequality $|\alpha^*| < 1/N$ provides the bound proposed. □

§4. Fractal quadrature

The procedure described is applied now for the computation of a numerical integration. Let us denote M_0 the integral of the interpolant P^{α^*} on the interval I.

M_0 can be computed using the fixed point equation (6)

$$M_0 = \int_I P^{\alpha^*}(t)\,dt = \sum_{n=1}^{N}\int_{I_n}(\alpha_n P^{\alpha^*} \circ L_n^{-1}(t) + q_n \circ L_n^{-1}(t))\,dt,$$

that is to say

$$M_0 = \left(\sum_{n=1}^{N}\alpha_n \int_{I_n} P^{\alpha^*} \circ L_n^{-1}(t)\,dt\right) + Q_0,$$

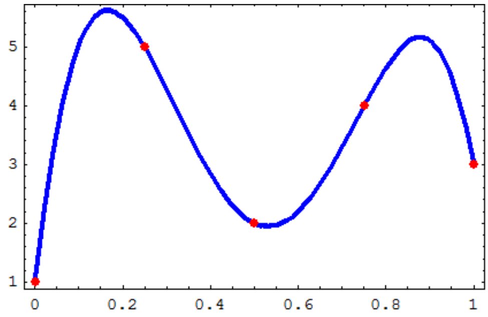


Figure 1: A polynomial interpolant P

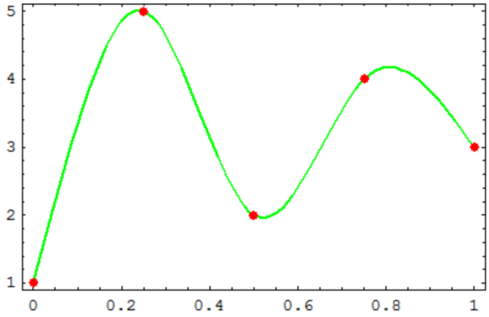


Figure 2: A cubic spline S

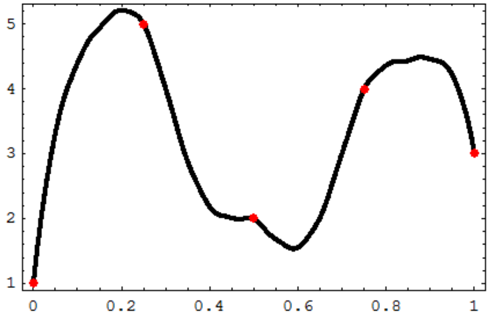


Figure 3: Fractal interpolant P^{α^*} computed from the maps of Figures 1 and 2

where

$$Q_0 = \int_I Q(t)\, dt \tag{13}$$

and

$$Q(t) = q_n \circ L_n^{-1}(t) \quad \text{if} \quad t \in I_n. \tag{14}$$

With the change $L_n^{-1}(t) = \widetilde{t}$, bearing in mind (1),

$$M_0 = \sum_{n=1}^{N} \alpha_n a_n M_0 + Q_0$$

and

$$M_0 = \frac{Q_0}{1 - \sum_{n=1}^{N} \alpha_n a_n}.$$

In this case,

$$q_n(t) = P \circ L_n(t) - \alpha_n b(t)$$

and thus

$$Q_0 = \int_I P(t)\, dt - \sum_{n=1}^{N} \alpha_n \int_{I_n} b \circ L_n^{-1}(t)\, dt.$$

With the same change $L_n^{-1}(t) = \widetilde{t}$,

$$Q_0 = C_0 - B_0\, (\sum_{n=1}^{N} \alpha_n a_n),$$

where C_0 is the polynomial quadrature

$$C_0 = \int_I P(t)\, dt$$

and

$$B_0 = \int_I b(t)\, dt.$$

Since $a_n = 1/N$ and $\alpha_n = \alpha^*$,

$$Q_0 = (C_0 - \alpha^* B_0)$$

and

$$M_0 = \frac{(C_0 - \alpha^* B_0)}{(1 - \alpha^*)}.$$

This formula introduces a slight modification to the quadrature C_0 corresponding to P.

Example 1. Let us consider the original function $X(t) = \frac{1}{1+25t^2}$ in the interval $[-1, 1]$ with the partition $\Delta : -1 < \frac{-2}{3} < \frac{-1}{3} < 0 < \frac{1}{3} < \frac{2}{3} < 1$. The value obtained for α^* in the optimization method described in the previous section is $\alpha^* = 0.04$. The polynomial quadrature gives $C_0 = 0.77407$, with an error of -0.224729. The correction M_0 provided by our procedure is $M_0 = 0.45459$, obtaining an error of 0.0947702, what improves the sought scalar.

Exact Value	C_0	Error C_0	α^*	M_0	Error M_0
0.54936	0.77409	-0.224729	0.04	0.45459	0.0947702

References

[1] Barnsley, M. F. Fractal functions and interpolation. *Constr. Approx. 2*, 4 (1986), 303–329.

[2] Barnsley, M. F. *Fractals Everywhere*. Academic Press, Inc., San Diego, 1988.

[3] La Torre, D., and Rocca, R. Approximating continuous functions by iterated functions systems and optimization. *Int. Math. J. 2*, 8 (2002), 801–811.

[4] Navascués, M. A. Fractal polynomial interpolation. *Z. Anal. Anwendungen 24*, 2 (2005), 401–418.

[5] Navascués, M. A. Fractal trigonometric approximation. *Electron. Trans. Numer. Anal. 20* (2005), 64–74.

[6] Navascués, M. A., and Sebastián, M. V. Fitting curves by fractal interpolation: an application to the quantification of cognitive brain processes. *In: Thinking in Patterns: Fractals and Related Phenomena in Nature, Novak, M.M.(ed.), World Scientific* (2004), 143–154.

[7] Navascués, M. A., and Sebastián, M. V. Smooth fractal interpolation. *Journal of Inequalities and Applications 78734* (2006), 1–20.

[8] Navascués, M. A., and Sebastián, M. V. Fractal-classic interpolants. *In: Convex and Fractal Geometry, Banach Center Publications 84* (2009), 173–180.

María Antonia Navascués and María Victoria Sebastián
Departamento de Matemática Aplicada
Universidad de Zaragoza
Campus Río Ebro
50018 Zaragoza, Spain
manavas@unizar.es and msebasti@unizar.es

Monografías Matemáticas *García de Galdeano* **35**, 213–219 (2010)

Axial Couette flow of second grade fluid due to a longitudinal time dependent shear stress

M. Nazar, M. Athar and W. Akhtar

Abstract. The axial flow of a second grade fluid through an infinite straight circular cylinder is considered. The flow of the fluid is due to the longitudinal shear stress that is prescribed on the boundary of the cylinder. The velocity field and the resulting shear stress are determined by means of the finite Hankel and Laplace transforms. The corresponding solutions for Newtonian fluids, performing the same motion, are obtained as limiting case from our general solutions. Graphical illustrations are presented for the velocity field and the shear stress for both the second grade and Newtonian fluids.

Keywords: Second grade fluids, velocity field, longitudinal shear stress.
AMS classification: 53B25, 53C15.

§1. Introduction

In many engineering fields, such as oil exploitation, polymer chemical industry, bio-engineering is necessary to study the non-Newtonian fluid flows. The second grade fluid is the common viscoelastic fluid in industrial fields, such as polymer solutions. The most exact solutions in this field correspond to the case when the velocity is given by the boundary. The first exact solutions for second grade fluids, in which a constant shear stress is given on the boundary, seem to be those of Bandelli and Rajagopal [2].

The aim of this paper is to study the flow of a second grade fluid in a circular infinite cylinder due to a longitudinal time dependent shear stress. We establish both the velocity field and the resulting shear stress corresponding to the motion of the fluid. These solutions can be easily specialized to give the solutions to the Newtonian fluids performing the same motion. Finally, for comparison, the profiles of the velocity $v(r,t)$ and the shear stress $\tau(r,t)$, for the Newtonian and second grade fluids are plotted as functions of r for different values of the time t.

§2. Governing equations

The constitutive equation of the second grade fluids is given by [4, 5, 6, 10]

$$\mathbf{T} = -p\mathbf{I} + \mu\mathbf{A}_1 + \alpha_1\mathbf{A}_2 + \alpha_2\mathbf{A}_1^2, \tag{1}$$

where $\mathbf{T}$ is the Cauchy stress tensor, p is the pressure, $\mathbf{I}$ is the unit tensor, μ is the dynamic viscosity, α_1 and α_2 are the normal stress moduli and $\mathbf{A_1}$, $\mathbf{A_2}$ are the kinematic tensors. $\mathbf{A_2}$ is defined by

$$\mathbf{A_2} = \dot{\mathbf{A}}_\mathbf{1} + \mathbf{A_1}(\operatorname{grad}\mathbf{v}) + (\operatorname{grad}\mathbf{v})^T\mathbf{A_1}, \tag{2}$$

where $\mathbf{v}$ is the velocity field, $\mathbf{A_1} = \operatorname{grad}\mathbf{v} + (\operatorname{grad}\mathbf{v})^T$ and the superscript T denotes the transpose operator. In cylindrical coordinates (r, θ, z), the velocity of the axial flow is given by [4]

$$\mathbf{v} = \mathbf{v}(r, t) = v(r, t)\mathbf{e_z}, \tag{3}$$

where $\mathbf{e_z}$ is the unit vector in the z-direction. For such flows the constraint of incompressibility is automatically satisfied. Since the velocity field is independent of θ and z, we also assume that the extra-stress tensor $\mathbf{S}$ is independent of these variables. Furthermore, if the fluid is assumed to be at rest at the moment $t = 0$, then

$$\mathbf{S}(r, 0) = \mathbf{0}. \tag{4}$$

Equalities (1), (2) and (3) lead to the constitutive relationship [2]

$$\tau(r, t) = (\mu + \alpha_1 \partial_t)\frac{\partial v(r, t)}{\partial r}, \tag{5}$$

where $\tau(r, t) = S_{rz}(r, t)$ is the shear stress which is different of zero.

In the absence of body forces and a pressure gradient in the $z-$direction, the balance of the linear momentum leads to the relevant equation

$$\rho\frac{\partial v(r, t)}{\partial t} = \left(\frac{\partial}{\partial r} + \frac{1}{r}\right)\tau(r, t), \tag{6}$$

where ρ is the constant density of the fluid.

Eliminating $\tau(r, t)$ among Eqs. (5) and (6), we attain to the governing equation

$$\frac{\partial v(r, t)}{\partial t} = (\nu + \alpha\partial_t)\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right)v(r, t), \tag{7}$$

where $\nu = \mu/\rho$ is the kinematic viscosity of the fluid and $\alpha = \alpha_1/\rho$.

§3. Axial flow through an infinite circular cylinder

Let us consider an incompressible second grade fluid at rest in an infinite circular cylinder of radius R. At time $t = 0^+$, the cylinder is suddenly pulled with a time dependent shear stress. Due to the shear, the fluid is gradually moved. It's velocity being of the form (3) and imposed initial and boundary conditions are

$$v(r, 0) = 0\ ; \quad r \in [0, R)\ , \tag{8}$$

$$\tau(R, t) = (\mu + \alpha_1\partial_t)\frac{\partial v(R, t)}{\partial r} = ft, \quad t > 0. \tag{9}$$

Applying Laplace transform to Eqs. (7), (9) and using (8), we obtain

$$q\bar{v}(r, q) = (\nu + \alpha q)\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right)\bar{v}(r, q), \tag{10}$$

$$\left.\frac{\partial\bar{v}(r, q)}{\partial r}\right|_{r=R} = \frac{f}{q^2(\mu + \alpha_1 q)}. \tag{11}$$

In order to obtain an analytical solution of the problem (10) and (11), the finite Hankel transform method is used. We define the Hankel transform of the function $\bar{v}(r,q)$ by [3]

$$\bar{v}_{H}(r_n,q) = \int_0^R r\bar{v}(r,q)J_0(rr_n)\,dr, \tag{12}$$

where r_n, $n = 1, 2, 3, ldots$, are the positive roots of the equation

$$J_1(Rr) = 0. \tag{13}$$

In the above relation, $J_\nu(\cdot)$ is the Bessel function of the first kind of order ν [7]. Multiplying now both sides of Eq. (10) by $rJ_0(rr_n)$, integrating then with respect to r from 0 to R and taking into account the condition (11) and the equality

$$\int_0^R r\Big(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\Big)\bar{v}(r,q)J_0(rr_n)dr = -r_n^2\bar{v}_{H}(r_n,q) + RJ_0(Rr_n)\frac{\partial\bar{v}(R,q)}{\partial r}\ ,$$

we find that

$$\bar{v}_{H}(r_n,q) = \frac{Rf}{\rho}J_0(Rr_n)\,\frac{1}{q^2(q+\alpha r_n^2 q + \nu r_n^2)}. \tag{14}$$

We rewrite Eq. (14) as

$$\bar{v}_{H}(r_n,q) = \bar{v}_{1H}(r_n,q) + \bar{v}_{2H}(r_n,q), \tag{15}$$

where

$$\bar{v}_{1H}(r_n,q) = \frac{RfJ_0(Rr_n)}{r_n^2}\,\frac{1}{q^2(\mu+\alpha_1 q)} \tag{16}$$

and

$$\bar{v}_{2H}(r_n,q) = -\frac{RfJ_0(Rr_n)}{q}\,\frac{1}{r_n^2(\mu+\alpha_1 q)(q+\alpha r_n^2 q+\nu r_n^2)}. \tag{17}$$

The inverse Hankel transform of the function $\bar{v}_{1H}(r_n,q)$ and $\bar{v}_{2H}(r_n,q)$ are

$$\bar{v}_1(r,q) = \frac{r^2 f}{2R}\,\frac{1}{q^2(\mu+\alpha_1 q)}, \quad \bar{v}_2(r,q) = \frac{2}{R^2}\sum_{n=1}^{\infty}\frac{J_0(rr_n)}{J_0^2(Rr_n)}\,\bar{v}_{2H}(r_n,q). \tag{18}$$

From (15)-(18) we find that the Laplace transform of the velocity $v(r,t)$, has the form

$$\bar{v}(r,q) = \frac{r^2 f}{2R}\,\frac{1}{q^2(\mu+\alpha_1 q)} - \frac{2f}{R}\sum_{n=1}^{\infty}\frac{J_0(rr_n)}{r_n^2 J_0(Rr_n)}\,\frac{1}{q(\mu+\alpha_1 q)(q+\alpha r_n^2 q+\nu r_n^2)}. \tag{19}$$

Applying the discrete inverse Laplace transform to Eq. (19), using the expansion

$$\frac{1}{q(q+\alpha r_n^2 q+\nu r_n^2)} = \frac{q^{-2}}{(1+\alpha r_n^2)+\nu r_n^2 q^{-1}} = \sum_{k=0}^{\infty}(-\nu r_n^2)^k\frac{q^{-k-2}}{(1+\alpha r_n^2)^{k+1}}, \tag{20}$$

the convolution theorem and the formulae

$$L^{-1}\Big\{\frac{1}{q^a}\Big\} = \frac{t^{a-1}}{\Gamma(a)},\ a>0, \quad L^{-1}\Big\{\frac{q^b}{(q^a-d)^c}\Big\} = G_{a,b,c}(d,t),\ \mathrm{Re}(ac-b)>0, \tag{21}$$

where $G_{a,b,c}(d,t)$ are the generalized G-functions defined as [8]

$$G_{a,b,c}(d,t) = \sum_{j=0}^{\infty} \frac{d^j\,\Gamma(c+j)}{\Gamma(c)\Gamma(j+1)} \frac{t^{(c+j)a-b-1}}{\Gamma[(c+j)a-b]}, \tag{22}$$

we find that

$$\begin{aligned} v(r,t) = \frac{fr^2}{2R}\left\{\frac{\alpha_1}{\mu^2}[\exp(-\frac{\mu t}{\alpha_1})-1]+\frac{t}{\mu}\right\} - \frac{2f}{\alpha_1 R}\sum_{n=1}^{\infty}\frac{J_0(rr_n)}{r_n^2 J_0(Rr_n)}\sum_{k=0}^{\infty}(-\nu r_n^2)^k \\ \times\int_0^t G_{1,0,1}(-\mu/\alpha_1,s)G_{0,-k-2,k+1}(-\alpha r_n^2,t-s)\,ds, \end{aligned} \tag{23}$$

which can be simplified by using the following relations

$$G_{0,-k-2,k+1}(-\alpha r_n^2,t) = \sum_{j=0}^{\infty}(-\alpha r_n^2)^j \frac{\Gamma(k+j+1)}{\Gamma(k+1)\Gamma(j+1)}\frac{t^{k+1}}{\Gamma(k+2)} = \frac{t^{k+1}}{(k+1)!}\frac{1}{(1+\alpha r_n^2)^{k+1}}, \tag{24}$$

$$\begin{aligned} \sum_{k=0}^{\infty}(-\nu r_n^2)^k G_{0,-k-2,k+1}(-\alpha r_n^2,t) &= \frac{-1}{\nu r_n^2}\sum_{k=0}^{\infty}\left(-\frac{\nu r_n^2 t}{1+\alpha r_n^2}\right)^{k+1}\frac{1}{(k+1)!} \\ &= \frac{1}{\nu r_n^2}\left[1-\exp\left(-\frac{\nu r_n^2 t}{1+\alpha r_n^2}\right)\right], \end{aligned} \tag{25}$$

and

$$G_{1,0,1}(-\mu/\alpha_1,t) = \exp(-\frac{\mu t}{\alpha_1}). \tag{26}$$

Now the velocity field $v(r,t)$ has form

$$v(r,t) = \frac{fr^2}{2\mu R}(t-\frac{\alpha_1}{\mu}) - \frac{2f}{\mu\nu R}\sum_{n=1}^{\infty}\left[1-(1+\alpha r_n^2)\exp\left(-\frac{\nu r_n^2 t}{1+\alpha r_n^2}\right)\right]\frac{J_0(rr_n)}{r_n^4 J_0(Rr_n)}. \tag{27}$$

§4. Calculation of the shear stress

Applying the Laplace transform to Eq. (5), we find that

$$\bar{\tau}(r,q) = (\mu+\alpha_1 q)\frac{\partial\bar{v}(r,t)}{\partial r}. \tag{28}$$

Differentiating Eq. (19) with respect to r and using the identity

$$\frac{d}{dr}J_0(rr_n) = -r_n J_1(rr_n),$$

we find $\bar{\tau}(r,q)$, after using Eq. (28)

$$\bar{\tau}(r,q) = \frac{fr}{Rq^2} + \frac{2f}{R}\sum_{n=1}^{\infty}\frac{J_1(rr_n)}{r_n J_0(Rr_n)}\frac{1}{q(q+\alpha r_n^2 q+\nu r_n^2)}. \tag{29}$$

Applying inverse Laplace transform to Eq. (29) by using (21) and (25), we get the shear stress $\tau(r,t)$

$$\tau(r,t) = \frac{frt}{R} + \frac{2f}{\nu R}\sum_{n=1}^{\infty}\frac{J_1(rr_n)}{r_n^3 J_0(Rr_n)}\Big[1 - \exp\Big(-\frac{\nu r_n^2 t}{1+\alpha r_n^2}\Big)\Big]. \tag{30}$$

§5. Limiting case $\alpha_1 \to 0$

Making $\alpha_1 \to 0$ into Eqs. (27) and (30), we obtain the velocity field

$$v(r,t) = \frac{fr^2 t}{2R\mu} - \frac{2f}{R\nu\mu}\sum_{n=1}^{\infty}\frac{J_0(rr_n)}{r_n^4 J_0(Rr_n)}\Big(1 - e^{-\nu r_n^2 t}\Big), \tag{31}$$

and the associated shear stress

$$\tau(r,t) = \frac{frt}{R} + \frac{2f}{\nu R}\sum_{n=1}^{\infty}\frac{J_1(rr_n)}{r_n^3 J_0(Rr_n)}\Big(1 - e^{-\nu r_n^2 t}\Big), \tag{32}$$

corresponding to a Newtonian fluid, performing the same motion.

Eqs. (31) and (32) are identical with those found by W. Akhtar *et al.* [1].

§6. Conclusions

In this paper, the velocity field and the associated shear stress corresponding to the axial flow of second grade fluids through a circular cylinder are determined. The motion is due to a longitudinal shear stress which is prescribed on the boundary of the cylinder. More exactly, at the moment $t = 0^+$ the cylinder is pulled with a time dependent shear stress along its axis.

The solutions determined by means of the Laplace and finite Hankel transforms satisfy all imposed initial and boundary conditions. The corresponding solutions for Newtonian fluids, performing the same motion, are obtained as limiting case from our solutions. Finally, in Figs. 1 and 2, the profiles of the velocity and shear stress of the second grade fluid (curves $v(r)$ and $\tau(r)$) and Newtonian fluid (curves $vN(r)$ and $\tau N(r)$) are plotted as function of r for different values of the time t. From these figures we have that for low values of the time t the second grade fluid flows slower than the Newtonian fluid and this difference disappear when the values of the time increase.

In all figures we consider $R = 0.1$, $f = 2$, $\rho = 1260$, $\mu = 1.48$, $\alpha = 80$. The units of parameters in Figs. 1 and 2 are from SI units and the roots r_n have been approximated with [9] $r_n = (4n-1)\pi/(4R)$.

Acknowledgements

The authors would like to thank Prof. Constantin Fetecau and Prof. Dumitru Vieru for their suggestions and constructive comments.

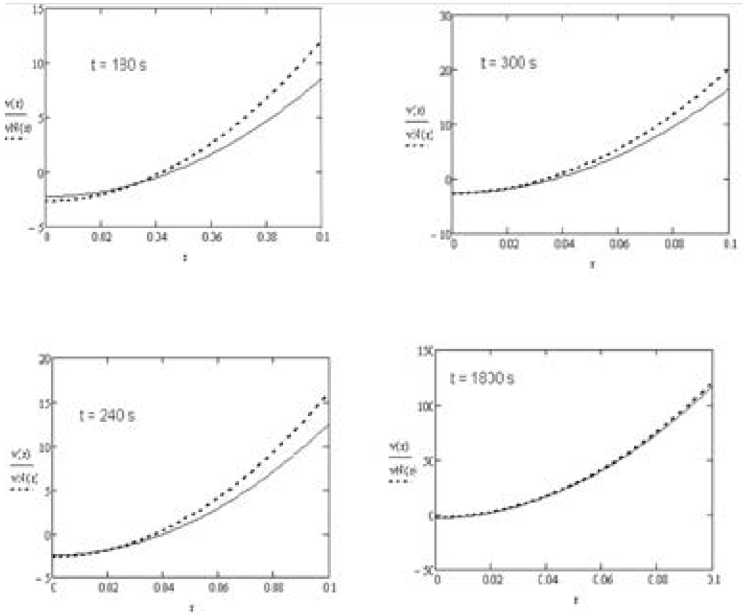


Figure 1: Velocity profiles $v(r)$ for different values of the time t: $v(r)$– the second grade fluid, $vN(r)$–the Newtonian fluid.

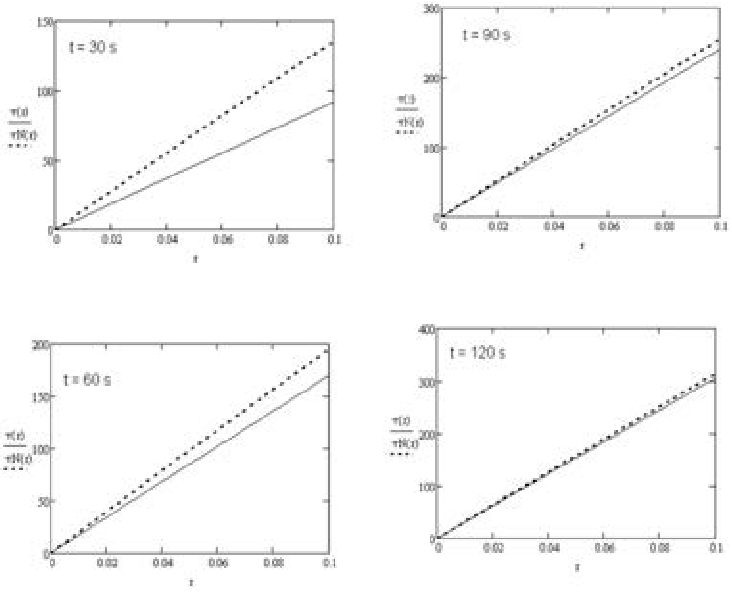


Figure 2: The profiles of the shear stress $\tau(r)$ for different values of the time t: $\tau(r)$– the second grade fluid, $\tau N(r)$– the Newtonian fluid.

References

[1] Akhtar, W., and Nazar, M. On the helical flow of newtonian fluids induced by time dependent shear. Published in this volume.

[2] Bandelli, R., and Rajagopal, K. R. Start-up flows of second grade fluids in domains with one finite dimension. *Internat. J. Non-Linear Mech. 30*, 6 (1995), 817–839.

[3] Debnath, L., and Bhatta, D. *Integral transforms and their applications*, second ed. Chapman & Hall/CRC, Boca Raton, FL, 2007.

[4] Fetecau, C., and Fetecau, C. On some axial Couette flows of non-Newtonian fluids. *Z. Angew. Math. Phys. 56*, 6 (2005), 1098–1106.

[5] Fetecau, C., and Fetecau, C. Starting solutions for the motion of a second grade fluid due to longitudinal and torsional oscillations of a circular cylinder. *Internat. J. Engrg. Sci. 44*, 11-12 (2006), 788–796.

[6] Hayat, T., Asghar, S., and Siddiqui, A. Some unsteady unidirectional flows of a non-Newtonian fluid. *Internat. J. Engrg. Sci. 38*, 3 (2000), 337–345.

[7] Kreyszig, E. *Advanced engineering mathematics*. John Wiley & Sons Inc., New York, 1999.

[8] Lorenzo, C. F., and Hartley, T. T. Generalized functions for the fractional calculus. Tech. Rep. NASA TP-1999-209424/REV1, National Aeronautics and Space Administration, October 1999.

[9] McLachlan, N. *Bessel Functions for Engineers*. Oxford University Press, London, 1955.

[10] Wang, S., and Xu, M. Axial Couette flow of two kinds of fractional viscoelastic fluids in an annulus. *Nonlinear Anal. Real World Appl. 10*, 2 (2009), 1087–1096.

Mudassar Nazar, Muhammad Athar and Waseem Akhtar
Abdus Salam School of Mathematical Sciences GC University
68-B New Muslim Town Lahore
PAKISTAN
mudassar_666@yahoo.com, athar_sms@yahoo.com and wasakh75@yahoo.com

Monografías Matemáticas *García de Galdeano* **35**, 221–227 (2010)

TESTING NUMERICAL METHODS FOR SOLVING INTEGRAL EQUATIONS

Miguel Pasadas and Miguel L. Rodríguez

Abstract. Many modeling problems in physics and in a variety of engineering fields lead to integral equations. We briefly describe the main classical techniques to obtain approximated solutions of them: Nyström methods and projection methods. Moreover, we introduce a new method to approximate the solution of integral equations based in a variational scheme. We test these techniques with numerical examples and we show several tables in order to measure the error obtained by the presented methods.

Keywords: Integral equations.

AMS classification: 65R20.

§1. Introduction

Integral equations are equations involving an unknown function which appears under an integral sign. The theory of integral equations has close contacts with many different areas of mathematics.

We consider the Fredholm integral equation of the second kind

$$f(t) = x(t) - \int_0^1 k(t,s)x(s)\,ds,\ 0 \le t \le 1. \tag{1}$$

It is known that the expression (1) in operator form can be written

$$f = x - \mathcal{K}x = (I - \mathcal{K})X.$$

Such equations occur widely in diverse areas of applied mathematics and physics, such as potential theory and radiation heat transfer but also some other equations reducible to it, and, in particular, the Lippman-Schwinger equation in potential scattering. In addition, many problems in the fields of differential equations can be recast as integral equations.

It is usually to impose to the operator $I - \mathcal{K}$ certain assumptions in order to establish the existence and uniqueness of solution of (1).

§2. Solving Fredholm integral equations of the second kind

The main numerical methods for solve these type of integral equations are Nyström methods or quadrature methods and the projection methods, based in approximate the numerical integral.

The projection methods with the collocation and Galerkin methods as special case, are a general tool that can also solve equations of the first kind. They are known as spectral methods and pseudospectral methods respectively.

The collocation method seeks an approximate solution from a finite dimensional space by requiring that equation (1) to satisfy only at a finite number of points, called collocation points. In collocation methods one can use e. g. interpolation functions in polynomial or spline spaces. Because of the better convergence properties of splines, spline collocation is superior to polynomial collocation.

The Galerkin method and Petrov–Galerkin method, with many variants [3], consists in finding a best approximation to the exact solution of (1) in a finite dimensional space by the minimizing of the so called energy functional. One of the advantages of the Petrov-Galerkin method is that it allows to achieve the same order of convergence as the Galerkin method with much less computational cost by choosing the test spaces to be spaces of piecewise polynomials of lower degree.

Another method which is employed for the solution of integral equations on smooth closed curves is *the qualocation method*. Qualocation method is a Petrov–Galerkin method in which the outer integrals are performed numerically by special quadrature rules.

We have been developed another method in order to approximate the solution of (1). The method is based in the minimization of a functional that involves (1) and that it is similar to the Petrov–Galerkin method.

Our aim in this work is to detail the computational part of distinct solving methods and to show different examples in order to compare them. Throughout this overview, we will assume that the integral equations have a unique solution to be determined.

§3. Preliminaries and notations

The Euclidean norm and inner product in $\mathbb{R}^n$ will be denoted respectively by $\langle\cdot\rangle$ and $\langle\cdot,\cdot\rangle$. Moreover, we designate by $H^k\,(0,1)$ the Sobolev space of order k, which is equipped with the inner product and norm

$$((u,v))_k = \sum_{i=0}^{k} \int_0^1 u^{(i)}(t)v^{(i)}(t)\,dt, \qquad \|u\|_k = (u,u)_k^{1/2},$$

the semi–inner products and semi-norms

$$(u,v)_j = \int_0^1 u^{(j)}(x)v^{(j)}(x)\,dx,\ |u|_j = (u,u)_j^{1/2},\ \forall j = 0,\dots,k.$$

Let $k(t,s) \in H^3(0,1) \times L^2(0,1)$ be a given function, and we designate by $\mathcal{K}$ the integral operator associated with $k(t,s)$,

$$\mathcal{K}u(t) = \int_0^1 \int_0^1 k(t,s)u(s)\,ds\,dt, \qquad \forall u \in H^3(0,1).$$

Finally, we assume that $f \in C(0,1)$ and $\mathcal{K}$ is a compact operator on $C(0,1)$, and that 1 is not an eigenvalue of $\mathcal{K}$.

§4. Mixed Variational Method

This method can be briefly described as follows. For each $h \in \mathbb{R}_+$ and $N \in \mathbb{N}$ fixed, with $h = 1/N$, let

$$\Delta_h = \{0 = t_0 < \cdots < t_N = 1\},\ t_i = i\,h,$$

be a subset of distinct points of $[0,1]$

We denote by $S(3,2;\Delta_h)$ the space of the splines of degree 3 and class 2 associated with Δ_h, i.e.

$$S(3,2;\Delta_h) = \{s \in C^2(0,1) : s|_{[t_{i-1},t_i]} \in \mathbb{P}_3[t_{i-1},t_i],\ i = 1,\ldots,N\}.$$

A basis of this finite dimensional space is given by B-splines functions.

Given $a_1 = 0$, $a_2 = p$, $a_3 = 1$, being p a knot of Δ_h, we define the operator $\rho : H^3(0,1) \to \mathbb{R}^3$ by

$$\rho v = ((I - \mathcal{K})v\,(a_i))_{i=1,2,3}$$

and let $\beta = (f(a_i))_{i=1,2,3}$.

For each $h \in \mathbb{R}^+$, we define

$$G_h = \{u \in X_h : \rho u = \beta\}$$

and the vectorial space

$$G_h^0 = \{u \in X_h : \rho u = \mathbf{0}\}.$$

It is said that x_h is an approximated solution of (1) if x_h is a solution of the problem

$$\begin{cases} x_h \in G_h, \\ \forall v \in G_h, \quad J(x_h) \leq J(v), \end{cases} \tag{2}$$

where J is the functional defined on $H^3(0,1)$ by

$$J(v) = |(I - \mathcal{K})v - f|_3^2\,.$$

The next result guarantees the existence and the uniqueness of the solution of Problem (2).

Theorem 1. *Problem (2) has a unique solution characterized as the unique solution of the following variational problem: find x_h such that*

$$\begin{cases} x_h \in G_h, \\ \forall v \in G_h^0, ((I - \mathcal{K})\sigma_N, (I - \mathcal{K})v)_3 = ((I - \mathcal{K})v, f)_3. \end{cases}$$

Proof. It is clear that G_N is a nonempty closed convex subset of $S(3,2;\Delta_N)$. Now, we consider the form $a : S(3,2;\Delta_N) \times S(3,2;\Delta_N) \to \mathbb{R}$ given by $a(u,v) = 2\,(((u,v)))$.
Note that the application a is bilinear, symmetric, continuous and coercive since $S(3,2;\Delta_N)$ is a finite dimensional space. Let $\varphi(v) = ((I - \mathcal{K})v, f)_3$ be a linear form, which is clearly continuous. Now, Stampacchia's Theorem (see [2]) can be applied and we conclude the proof. □

As a consequence of this result we can obtain that there exists a unique $(x_h, \tau) \in X_h \times \mathbb{R}^3$ such that for all $v \in X_h$

$$((I - \mathcal{K})x_h, (I - \mathcal{K})v)_3 + \langle \tau, \rho v \rangle = (f, (I - \mathcal{K})v)_3, \tag{3}$$

where x_h is the unique solution of Problem (2).

§5. Computation of the methods

We detail the computation of the methods described above. In all cases, we seek an approximate function $x_h(t) \in X_h$ where X_h is a finite dimensional subspace.

The approximated solution of the integral equation (1) can be written as

$$x_h(t) = \sum_{i=1}^{n} \alpha_i \phi_i(t), \quad t \in D,$$

where $\{\phi_1, \dots, \phi_h\}$ is a basis of X_h. The question is how we can determinate the unknown coefficients α_i, $i = 1, \dots, n$, by using the above methods.

Collocation method. We choose distinct node points $t_1, \dots, t_n \in [0, 1]$ and we impose that

$$f(t_i) - \lambda x_h(t_i) - \int_0^1 k(t_i, s) x_h(s)\, ds = 0, \quad i = 1, \dots, n.$$

Then, we have now a linear system of n equations with unknown α_i.

Galerkin method. We impose that

$$\langle (\lambda I - \mathcal{K}) x_h, \phi_j \rangle = \langle f, \phi_j \rangle, \quad \forall j = 1, \dots, n.$$

The coefficients α_i are determined by solving

$$\sum_{i=1}^{n} \alpha_i \left(\langle \lambda\, \phi_i, \phi_j \rangle - \langle \mathcal{K} \phi_i, \phi_j \rangle \right) = \langle f, \phi_j \rangle, \quad \forall j = 1, \dots, n.$$

Nyström method. It requires the choice of some approximate quadrature rule. By using a numerical scheme, from

$$\int_0^1 k(t, s) x_h(s)\, ds \approx \sum_{j=1}^{n} \omega_j k(t, t_j) x_h(t_j), \quad 0 \le j \le n,$$

we obtain the linear system

$$\lambda x_h(t_i) + k(t_i, t_j)\, x_h(t_j) = f(t_i), \quad i = 1, \dots, n.$$

Here the set ω_j are the weights of the quadrature rule, while the n points t_j are the abscissas. An interesting approach of this method can be found in [1].

Mixed Variational method. By replacing in (3), we have for all $v \in X_h$

$$\sum_{i=1}^{n} \alpha_i ((I - \mathcal{K})\phi_i, (I - \mathcal{K})v)_3 + \langle \tau, \rho v \rangle = (f, (I - \mathcal{K})v)_3,$$

subject to the restrictions

$$x_h(a_j) - \int_0^1 k(a_j, s)\, x_h(s)\, ds = f(a_j), \;\; j = 1, 2, 3.$$

Taking $v = \phi_i$, for $i = 1, \ldots, n$, we obtain a linear system of order $n + 3$ with the unknown $\alpha_1, \ldots, \alpha_n, \tau_1, \tau_2, \tau_3$. The matrix form of such system is

$$\begin{pmatrix} A & D \\ D^t & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \tau \end{pmatrix} = \begin{pmatrix} \widehat{f_1} \\ \widehat{f_2} \end{pmatrix},$$

where

$$A = \Big(((I - \mathcal{K})\phi_i, (I - \mathcal{K})\phi_j)_3 \Big)_{1 \le i,j \le n},$$

$$D = \left(\phi_i(a_j) - \int_0^1 k(a_j, s)\, \phi_i(s)\, ds \right)_{\substack{1 \le i \le n \\ 1 \le j \le 3}},$$

$$\widehat{f_1} = ((f, (I - \mathcal{K})\phi_i)_3)^t_{1 \le i \le n}, \qquad \widehat{f_2} = (f(a_i))^t_{1 \le i \le 3}.$$

§6. Numerical examples

We present several numerical experiments and we compare the results with the exact solution of the integral equation.

Now, we have chosen the space of cubic spline functions for the numerical experiments. For each $N \in \mathbb{N}$ let $h = 1/N$ and let

$$\Delta_N = \{0 = t_0 < \ldots < t_N = 1\}, \; t_i = i\,h, \; i = 0, \ldots, N,$$

be a subset of distinct points of $[0, 1]$.

A basis of this $S(3, 2; \Delta_N)$ is given by B-splines functions. We denote this basis as $\mathcal{B} = \{B_i : 1 \le i \le N + 3\}$. The general expression of $B_i(t)$ is

$$B_i(t) = \frac{1}{6h^3} \begin{cases} (t - t_{i-2})^3, & \text{if } t \in [t_{i-2}, t_{i-1}], \\ h^3 + 3h^2(t - t_{i-1}) + 3h(t - t_{i-1})^2 - 3(t - t_{i-1})^3, & \text{if } t \in [t_{i-1}, t_i], \\ h^3 + 3h^2(t_{i+1} - t) + 3h(t_{i+1} - t)^2 - 3(t_{i+1} - t)^3, & \text{if } t \in [t_i, t_{i+1}], \\ (t_{i+2} - t)^3, & \text{if } t \in [t_{i+1}, t_{i+2}], \\ 0, & \text{otherwise.} \end{cases}$$

We take $\phi_i(t) = B_i(t)$. In order to show the efficiency, we have computed an estimation in several spaces of the error $\|x_h(t) - x(t)\|$, where x is the exact solution of the integral equation (1) and x_h is the approximated solution.

Example 1. We consider the simple test equation

$$x(t) - \int_0^1 x(s)(t^2 - t - s^2 + s)\, ds = -2t^3 + 3t^2 - t, \quad t \in [0, 1].$$

with exact solution $x(t) = -2t^3 + 3t^2 - t$. See Table 1.

	Method		
Computed error in	Collocation	Galerkin	Variational
$H^0(0,1)$	0	0	0
$H^1(0,1)$	3.17(−14)	7.65(−10)	0
$H^2(0,1)$	8.16(−13)	2.47(−8)	2.45(−15)
$H^3(0,1)$	1.27(−11)	4.56(−7)	6.32(−15)
$C^0(0,1)$	6.35(−15)	1.08(−10)	0
$C^1(0,1)$	1.73(−13)	4.69(−9)	2.33(−15)
$C^2(0,1)$	3.71(−12)	1.20(−7)	5.46(−15)

Table 1: Table of the computed relative error for Example 1 for $N = 4$ equidistant knots.

	Method		
Computed error in	Collocation	Galerkin	Variational
$H^0(0,1)$	4.34(−8)	1.05(−8)	3.71(−6)
$H^1(0,1)$	3.06(−6)	3.02(−6)	2.35(−5)
$H^2(0,1)$	2.97(−4)	3.07(−4)	3.17(−4)
$H^3(0,1)$	3.44(−2)	3.45(−2)	3.21(−2)
$C^0(0,1)$	1.25(−7)	1.61(−7)	5.22(−6)
$C^1(0,1)$	6.85(−6)	9.44(−6)	6.01(−5)
$C^2(0,1)$	1.02(−3)	1.54(−3)	7.31(−4)

Table 2: Table of the computed relative error for Example 2 for $N = 4$ equidistant knots.

Example 2. The following integral equation

$$x(t) - \int_0^1 x(s)e^{-t-s}\,ds = e^t - e^{-t}, \quad t \in [0,1].$$

has the exact solution $x(t) = e^t$. See Table 2.

6.1. Future work

We plan to do research in the following items:

1. The study of the mixed variational method for integral equations with Cauchy kernels.
2. Numerical experiments in distinct finite dimensional spaces as wavelets spaces.
3. The extension of the mixed variational method to the two dimensional case.

Acknowledgements

This work has been supported by the Junta de Andalucía (Research group FQM/191) and by Spain's Ministerio de Educación y Ciencia (Research project MTM2008-00671).

References

[1] Ezzirani, A., and Guessab, A. A fast algorithm for Gaussian type quadrature formulae with mixed boundary conditions and some lumped mass spectral approximations. *Math. Comp. 68*, 225 (1999), 217–248. Available from: http://dx.doi.org/10.1090/S0025-5718-99-01001-7, doi:10.1090/S0025-5718-99-01001-7.

[2] Kinderlehrer, D., and Stampacchia, G. *An introduction to variational inequalities and their applications*, vol. 88 of *Pure and Applied Mathematics*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980.

[3] Sloan, I. H. Four variants of the Galerkin method for integral equations of the second kind. *IMA J. Numer. Anal. 4*, 1 (1984), 9–17. Available from: http://dx.doi.org/10.1093/imanum/4.1.9, doi:10.1093/imanum/4.1.9.

Miguel Pasadas and Miguel L. Rodriguez
Dept. Matemática Aplicada
E.T.S. Ingenieros de Caminos, Canales y Puertos
Campus de Fuentenueva, s/n
18071 Granada
mpasadas@ugr.es and miguelrg@ugr.es

Monografías Matemáticas *García de Galdeano* **35**, 229–236 (2010)

EFFICIENT SECOND-ORDER DISCRETIZATIONS FOR SEMILINEAR PARABOLIC PROBLEMS

Laura Portero, Andrés Arrarás and Juan Carlos Jorge

Abstract. This work is devoted to the efficient numerical solution of semilinear parabolic problems posed on two-dimensional domains. To this end, we first carry out a spatial semidiscretization that uses a mimetic finite difference scheme based on the support-operator method. The connection between mimetic finite difference techniques and mixed finite element methods is the key to proving second-order convergence for such a scheme. Next, we consider a splitting of the semidiscrete elliptic operator subordinate to a decomposition of the spatial domain into a set of overlapping subdomains. Within this framework, we apply a second-order linearly implicit fractional step Runge-Kutta method as the time integrator. Thus, the original problem is reduced to the solution of a set of linear systems per time step. Furthermore, such linear systems can be decomposed into a set of smaller subsystems that may be solved in parallel without iterative processing.

Keywords: Domain decomposition, linearly implicit fractional step method, mimetic finite difference method, mixed finite element method, semilinear parabolic problem.

AMS classification: 35K99, 65M12, 65Y05.

§1. Introduction

Let us consider the following semilinear parabolic initial-boundary value problem: find $\psi : \Omega \times [0, T] \to \mathbb{R}$ such that

$$\psi_t(\underline{x}, t) - \operatorname{div}(K(\underline{x}) \text{ grad } \psi) = g(t, \psi) + f(\underline{x}, t), \qquad (\underline{x}, t) \in \Omega \times (0, T], \tag{1a}$$

$$\psi(\underline{x}, 0) = \psi_0(\underline{x}), \qquad \underline{x} \in \Omega, \tag{1b}$$

$$(-K(\underline{x}) \text{ grad } \psi) \cdot \underline{n} = 0, \qquad (\underline{x}, t) \in \partial\Omega \times (0, T]. \tag{1c}$$

The spatial domain $\Omega \subseteq \mathbb{R}^2$ is assumed to be a bounded open set with boundary $\partial\Omega$ and $K \equiv K(\underline{x}) = \{k_{ij}(\underline{x})\}_{2\times 2}$ is a symmetric positive definite tensor. On the other hand, $g(t, \cdot)$ denotes a nonlinear function assumed to be Lipschitz in the second variable, $f \equiv f(\underline{x}, t)$ is a sufficiently smooth source/sink term and $\underline{n}$ is the outward unit vector normal to $\partial\Omega$. If we replace (1a) by an equivalent system of first-order equations, we obtain:

$$\psi_t + \operatorname{div} \underline{u} = g(\psi) + f, \qquad (\underline{x}, t) \in \Omega \times (0, T], \tag{2a}$$

$$\underline{u} = -K \text{ grad } \psi, \qquad (\underline{x}, t) \in \Omega \times (0, T], \tag{2b}$$

where $\underline{u} \equiv \underline{u}(\underline{x}, t)$ is a vector-valued function that we refer to as the flux.

This paper proposes a numerical approach for solving (1) which is based on the method of lines, thus combining a spatial semidiscretization with a time integration. For the first stage, the spatial domain Ω is discretized with a logically rectangular grid composed of quadrilateral elements and, then, a mimetic finite difference (MFD) method is used to approximate problem (2), (1b) and (1c). In section 2, we briefly describe the mimetic technique in the context of semilinear parabolic problems, extending the ideas proposed in [3] for the elliptic case. Following [2], a second-order convergence result in the approximation of ψ is obtained by establishing a suitable connection between MFD methods and mixed finite element (MFE) methods in Raviart-Thomas spaces.

Next, we carry out the time integration by means of a linearly implicit fractional step Runge-Kutta (FSRK) method. For that purpose, we assume Ω to be a rectangle and suppose that tensor K is diagonal and positive definite. In this setting, we construct a sufficiently smooth partition of unity subordinate to a suitable decomposition of the spatial domain and use it to define certain splittings for both the semidiscrete operator and the source/sink term (cf. [4]). The combination of such splittings with a linearly implicit FSRK method reduces the original problem to the solution of several linear systems per internal stage that can be easily parallelized. In section 3, we introduce an FSRK time integrator proposed in [5] in order to define the totally discrete scheme and show its second-order unconditional convergence. Finally, section 4 contains a numerical test that illustrates the theoretical results surveyed in the paper.

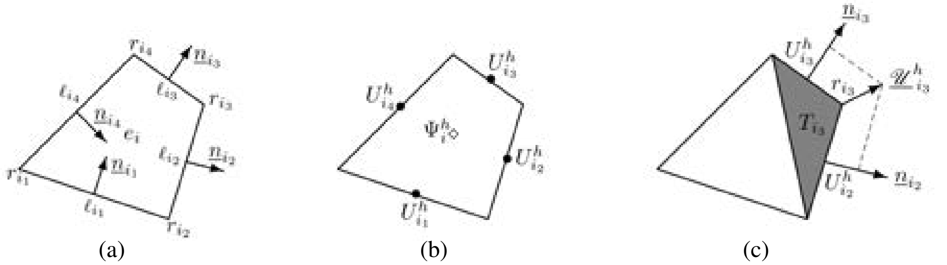
§2. Spatial semidiscretization

2.1. The mimetic finite difference method

Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into convex quadrilateral elements e, where $h = \max_{e \in \mathcal{T}_h} \operatorname{diam}(e)$ is the mesh size. In this work, $\mathcal{T}_h$ is assumed to be an h^2-uniform partition, i.e., each element is an h^2-parallelogram and any two adjacent elements represent an h^2-parallelogram (see [2]).

The MFD discretization may be outlined in four stages. The first one introduces the vector spaces of semidiscrete functions for both scalar and vector unknowns. On one hand, let $\mathcal{W}^h$ be the vector space of cell-centered semidiscrete scalar functions $\Psi^h = (\Psi_1^h, \Psi_2^h, \dots, \Psi_{N_e}^h)^T$, where N_e denotes the number of mesh elements. Here, $\Psi_i^h \equiv \Psi_i^h(t)$ is associated to the center of the i-th element e_i and provides an approximation to $\psi(\underline{x}, t)|_{e_i}$. On the other hand, we denote by $\mathcal{V}^h$ the vector space of edge-based semidiscrete vector functions $\underline{U}^h = (U_1^h, U_2^h, \dots, U_{N_\ell}^h)^T$, where N_ℓ is the number of mesh edges. In this case, $U_i^h \equiv U_i^h(t)$ is associated to the midpoint of the i-th mesh edge ℓ_i and provides an approximation to the normal component of vector $\underline{u}(\underline{x}, t)$ at ℓ_i (i.e., $\underline{u} \cdot \underline{n}_i$, where $\underline{n}_i$ is the unit vector normal to ℓ_i). Fig. 1(a) shows the local indexing of mesh vertices r_{i_j}, mesh edges ℓ_{i_j} and corresponding normal vectors $\underline{n}_{i_j}$, whereas Fig. 1(b) represents the discrete degrees of freedom for both scalar and vector functions at element e_i.

The second stage in the MFD method is to equip the previous vector spaces with appropriate inner products. The inner product on $\mathcal{W}^h$ is given by the expression $[\Psi^h, \Phi^h]_{\mathcal{W}^h} = \sum_{i=1}^{N_e} |e_i| \, \Psi_i^h \, \Phi_i^h$, where $\Psi^h, \Phi^h \in \mathcal{W}^h$ and $|e_i|$ denotes the area of the i-th element. For $\mathcal{V}^h$, we define the inner product to be $[\underline{U}^h, \underline{V}^h]_{\mathcal{V}^h} = \frac{1}{2} \sum_{i=1}^{N_e} \sum_{j=1}^{4} |T_{i_j}| \, K_i^{-1} \underline{\mathcal{U}}_{i_j}^h \cdot \underline{\mathcal{V}}_{i_j}^h$, where $\underline{U}^h, \underline{V}^h \in \mathcal{V}^h$, $|T_{i_j}|$ is the area of the triangle with vertices $r_{i_{j-1}}$, r_{i_j} and $r_{i_{j+1}}$ (with $r_0 = r_4$ and $r_5 = r_1$) and K_i is obtained from the evaluation of K at the center of the i-th element. The corner vec-



(a) (b) (c)

Figure 1: (a) Local indexing of vertices, edges and normal vectors at element e_i. (b) Discrete degrees of freedom for scalar and vector functions, Ψ^h and $\underline{U}^h$, at element e_i. (c) Construction of vector $\underline{\mathcal{U}}^h_{i_3}$ at vertex r_{i_3} of element e_i.

tors $\underline{\mathcal{U}}^h_{i_j}$ and $\underline{\mathcal{V}}^h_{i_j}$ are uniquely determined at the j-th vertex of e_i by using the corresponding components of $\underline{U}^h$ and $\underline{V}^h$, respectively. For instance, as displayed in Fig. 1(c), vector $\underline{\mathcal{U}}^h_{i_3}$ is obtained at node r_{i_3} as $\underline{\mathcal{U}}^h_{i_3} = U^h_{i_2}\,\underline{n}_{i_2} + U^h_{i_3}\,\underline{n}_{i_3}$, where $U^h_{i_2}$ and $U^h_{i_3}$ are those components of $\underline{U}^h$ associated with the edges ℓ_{i_2} and ℓ_{i_3}, respectively.

Once we have introduced the vector spaces of discrete functions and their corresponding inner products, the third stage in the MFD method consists of defining the discrete divergence operator, $\mathcal{D} : \mathcal{V}^h \to \mathcal{W}^h$, at the center of the i-th element, as $(\mathcal{D}\underline{U}^h)_i = |e_i|^{-1}(U^h_{i_2}\,|\ell_{i_2}| - U^h_{i_4}\,|\ell_{i_4}| + U^h_{i_3}\,|\ell_{i_3}| - U^h_{i_1}\,|\ell_{i_1}|)$, where $|\ell_{i_j}|$ denotes the length of the i-th edge, for $i = 1, 2, \ldots, N_e$ and $j = 1, 2, 3, 4$. Finally, in the fourth stage, we obtain the discrete flux operator, $\mathcal{G} : \mathcal{W}^h \to \mathcal{V}^h$, as the adjoint of $\mathcal{D}$ with respect to the inner products defined in the second stage, i.e., $\mathcal{G} = \mathcal{D}^*$ such that $[\mathcal{D}\underline{U}^h, \Psi^h]_{\mathcal{W}^h} \equiv [\underline{U}^h, \mathcal{G}\Psi^h]_{\mathcal{V}^h}$, for $\underline{U}^h \in \mathcal{V}^h$ and $\Psi^h \in \mathcal{W}^h$. This formula is a discrete version of Green's first identity. The relation of the previous inner products with the standard dot product permits us to obtain $\mathcal{G} = S^{-1}\mathcal{D}^{\dagger}M$, where M is a diagonal matrix given by $M = \operatorname{diag}\{|e_1|, |e_2|, \ldots, |e_{N_e}|\}$ and S is a symmetric positive-definite matrix with a 5-point stencil (cf. [3]).

The MFD method that approximates system (2), with initial and boundary data (1b) and (1c), can be written as follows: find $(\underline{U}^h, \Psi^h) : [0, T] \to \mathcal{V}^h \times \mathcal{W}^h$ such that

$$\Psi^h_t(t) + \mathcal{D}\underline{U}^h(t) = G^h(t, \Psi^h) + F^h(t), \qquad t \in (0, T], \tag{3a}$$

$$\underline{U}^h(t) = \mathcal{G}\Psi^h(t), \qquad t \in (0, T], \tag{3b}$$

$$\Psi^h(0) = \Psi^h_0, \tag{3c}$$

where vectors $G^h(t, \Psi^h)$ and $F^h(t)$ belong to $\mathcal{W}^h$ and their components are $G^h_i(t, \Psi^h) = g(t, \Psi^h_i)$ and $F^h_i(t) = |e_i|^{-1} \int_{e_i} f(\underline{x}, t)\, d\underline{x}$, respectively, for $i = 1, 2, \ldots, N_e$. Furthermore, Ψ^h_0 represents an adequate approximation to $\psi_0(\underline{x})$ to be specified later. Multiplying (3a), (3c) by $M\Phi^h$ and (3b) by $S\underline{V}^h$, we get, by omitting the time dependencies:

$$[\Psi^h_t, \Phi^h]_{\mathcal{W}^h} + [\mathcal{D}\underline{U}^h, \Phi^h]_{\mathcal{W}^h} = [G^h(\Psi^h), \Phi^h]_{\mathcal{W}^h} + [F^h, \Phi^h]_{\mathcal{W}^h}, \qquad \forall\, \Phi^h \in \mathcal{W}^h, \tag{4a}$$

$$[\underline{U}^h, \underline{V}^h]_{\mathcal{V}^h} = [\Psi^h, \mathcal{D}\underline{V}^h]_{\mathcal{W}^h}, \qquad \forall\, \underline{V}^h \in \mathcal{V}^h, \tag{4b}$$

$$[\Psi^h(0), \Phi^h]_{\mathcal{W}^h} = [\Psi^h_0, \Phi^h]_{\mathcal{W}^h}, \qquad \forall\, \Phi^h \in \mathcal{W}^h. \tag{4c}$$

This formulation will permit us to connect the MFD method described in the present subsection with a MFE method to be introduced next.

2.2. The mixed finite element method

Let us define $V = \left\{\underline{v} \in H(\mathrm{div};\Omega) : \underline{v} \cdot \underline{n} = 0 \text{ on } \partial\Omega\right\}$ and $W = L^2(\Omega)$, where $H(\mathrm{div};\Omega) = \{\underline{v} \in [L^2(\Omega)]^2 : \mathrm{div}\,\underline{v} \in L^2(\Omega)\}$. The variational formulation of system (2), (1b) and (1c) is: find $(\underline{u}, \psi) : [0,T] \to V \times W$ such that

$$(\psi_t, \phi) + (\mathrm{div}\,\underline{u}, \phi) = (g(\psi), \phi) + (f, \phi), \qquad \forall\, \phi \in W, \tag{5a}$$

$$a(\underline{u}, \underline{v}) = (\psi, \mathrm{div}\,\underline{v}), \qquad \forall\, \underline{v} \in V, \tag{5b}$$

$$(\psi(0), \phi) = (\psi_0, \phi), \qquad \forall\, \phi \in W, \tag{5c}$$

where $a(\cdot,\cdot)$ is a bilinear form given by $a(\underline{u}, \underline{v}) = \int_\Omega K^{-1}\, \underline{u} \cdot \underline{v}\, d\underline{x}$.

For the discretization of (5), recall that we consider an h^2-uniform partition $\mathcal{T}_h$ of $\overline{\Omega}$ consisting of convex quadrilaterals. Let $\tilde{e}$ be the reference unit square with vertices $(0,0)^T$, $(1,0)^T$, $(1,1)^T$ and $(0,1)^T$ and define a bilinear mapping $\mathcal{F}_e : \tilde{e} \to e$ which transforms the vertices of $\tilde{e}$ into the vertices of $e = \mathcal{F}_e(\tilde{e})$. If we denote by $r_i = (x_i, y_i)^T$, for $i = 1,2,3,4$, the corresponding vertices of e (counted counter-clockwise) and define $\underline{x} \equiv (x, y)$ and $\underline{\tilde{x}} \equiv (\tilde{x}, \tilde{y})$, we have: $\underline{x} = \mathcal{F}_e(\underline{\tilde{x}}) = r_1(1-\tilde{x})(1-\tilde{y}) + r_2\tilde{x}(1-\tilde{y}) + r_3\tilde{x}\tilde{y} + r_4(1-\tilde{x})\tilde{y}$. We shall denote by $J_e \equiv J_e(\underline{\tilde{x}})$ and $d_e \equiv d_e(\underline{\tilde{x}})$ the Jacobian matrix of $\mathcal{F}_e$ and its determinant, respectively.

Let us now consider the lowest-order Raviart-Thomas finite element spaces on the reference element $\tilde{e}$, given by $\tilde{V}_{\tilde{e}} = Q_{1,0}(\tilde{e}) \times Q_{0,1}(\tilde{e})$ and $\tilde{W}_{\tilde{e}} = Q_{0,0}(\tilde{e})$. Here, $\tilde{V}_{\tilde{e}} \times \tilde{W}_{\tilde{e}} \subset H(\mathrm{div};\tilde{e}) \times L^2(\tilde{e})$ and $Q_{m,n}(\tilde{e})$ refers to the space of polynomial fuctions on $\tilde{e}$ of degree at most m in $\tilde{x}$ and at most n in $\tilde{y}$. The corresponding spaces $V^h \times W^h \subset V \times W$ on $\mathcal{T}_h$ are given by $V^h = \{\underline{v} \in V : \underline{v}|_e = (d_e^{-1} J_e\, \underline{\tilde{v}}) \circ \mathcal{F}_e^{-1},\ \underline{\tilde{v}} \in \tilde{V}_{\tilde{e}}\ \forall\, e \in \mathcal{T}_h\}$ and $W^h = \{\phi \in W : \phi|_e = \tilde{\phi} \circ \mathcal{F}_e^{-1},\ \tilde{\phi} \in \tilde{W}_{\tilde{e}}\ \forall\, e \in \mathcal{T}_h\}$. The so-called velocity space V^h is a finite element subspace of $H(\mathrm{div};\Omega)$ which is defined on any convex quadrilateral e via the Piola transform.

Given the finite element spaces V^h and W^h, the MFE approximation to (5) reads: find $(\underline{u}^h, \psi^h) : [0,T] \to V^h \times W^h$ such that

$$(\psi_t^h, \phi^h) + (\mathrm{div}\,\underline{u}^h, \phi^h) = (g(\psi^h), \phi^h) + (f, \phi^h), \qquad \forall\, \phi^h \in W^h, \tag{6a}$$

$$a_h(\underline{u}^h, \underline{v}^h) = (\psi^h, \mathrm{div}\,\underline{v}^h), \qquad \forall\, \underline{v}^h \in V^h, \tag{6b}$$

$$(\psi^h(0), \phi^h) = (\psi_0^h, \phi^h), \qquad \forall\, \phi^h \in W^h, \tag{6c}$$

where $g(\psi^h)$ is a piecewise constant function such that $g(\psi^h)|_{e_i} = g(\psi^h(c_i))$, being c_i the center of element e_i, and ψ_0^h denotes the elliptic mixed finite element projection of ψ_0. Finally, $a_h(\cdot,\cdot)$ is a discrete bilinear form corresponding to the application of a numerical quadrature rule for computing $a(\cdot,\cdot)$ to be defined below.

Now, we are in condition to introduce the basic tool for the error analysis of the mimetic method described in the previous subsection. Recalling the definition of the MFD vector spaces, it is possible to establish an isometry between $\mathcal{W}^h$ and W^h, given by $\mathcal{I}_{\mathcal{W}^h} : \mathcal{W}^h \to W^h$, as well as an isomorphism between $\mathcal{V}^h$ and V^h, given by $\mathcal{I}_{\mathcal{V}^h} : \mathcal{V}^h \to V^h$ (cf. [2]). Taking into account these relationships, if we compare the MFD equations (4) with the MFE formulation (6), it is not difficult to prove that $(\operatorname{div} \underline{u}^h, \phi^h) = [\mathcal{D}\underline{U}^h, \Phi^h]_{\mathcal{W}^h}$ and $(\psi^h, \operatorname{div} \underline{v}^h) = [\Psi^h, \mathcal{D}\underline{V}^h]_{\mathcal{W}^h}$. Furthermore, the definition of Ψ^h, F^h, $G^h(\Psi^h)$ and Ψ^h_0 leads to the following equalities: $(\psi^h_t, \phi^h) = [\Psi^h_t, \Phi^h]_{\mathcal{W}^h}$, $(f, \phi^h) = [F^h, \Phi^h]_{\mathcal{W}^h}$, $(g^h(\psi^h), \phi^h) = [G^h(\Psi^h), \Phi^h]_{\mathcal{W}^h}$, $(\psi^h_0, \phi^h) = [\Psi^h_0, \Phi^h]_{\mathcal{W}^h}$. Finally, the equivalence between both formulations follows from the identity $a_h(\underline{u}^h, \underline{v}^h) \equiv [\underline{U}^h, \underline{V}^h]_{\mathcal{V}^h}$. Note that the quadrature rule $a_h(\cdot,\cdot)$ provides a coercive bilinear form, thus making problem (6) be well-posed.

We refer to [1] for a detailed description of the convergence analysis for the semidiscrete scheme. The main result from that work involves the classical L^2-projection operator $\mathcal{P}_h : W \to W^h$ and may be stated as follows.

Theorem 1. *Let $\mathcal{T}_h$ be an h^2-uniform quadrilateral partition of $\overline{\Omega}$ and let $\Psi^h(t)$ denote the MFD approximation to $\psi(\underline{x}, t)$. Under sufficient smoothness and compatibility conditions on data, if we set $\psi^h(t) = \mathcal{I}_{\mathcal{W}^h}(\Psi^h(t))$ and assume that $K \in (W^{1,\infty}(e))^{2\times 2}$ and $K^{-1} \in (W^{2,\infty}(e))^{2\times 2}$ for all $e \in \mathcal{T}_h$, then there exists a constant $C > 0$, independent of h, such that $\|\mathcal{P}_h\psi(\underline{x}, t) - \psi^h(t)\| \leqslant Ch^2$ for all $t \in [0, T]$.*

2.3. Mimetic finite differences on rectangular grids

Here and henceforth, we assume that the spatial domain Ω is a rectangle $(a, b) \times (c, d)$ and $\mathcal{T}_h$ is a rectangular mesh with $N_x \times N_y$ cells whose dimensions are $h_x = (b - a)/N_x$ and $h_y = (d - c)/N_y$. Moreover, $c_{i,j} = ((i - 1/2)h_x, (j - 1/2)h_y)$ denotes the center of the (i, j)-cell $e_{i,j}$. Finally, we assume that $K(\underline{x})$ is a diagonal 2×2 tensor, whose components satisfy $k_{11}(\underline{x})$, $k_{22}(\underline{x}) > 0$ for all $\underline{x} \in \Omega$.

If we define the restriction operator to the cell centers as $r_h : L^2(\Omega) \to \mathcal{W}^h$, then it holds that $r_h|_{W^h} \equiv \mathcal{I}^{-1}_{\mathcal{W}^h}$ and $\|r_h u^h\|_h = \|u^h\|$ for all $u^h \in W^h$, where $\|U^h\|_h \equiv [U^h, U^h]_{\mathcal{W}^h}$ denotes the discrete L^2-norm associated to $\mathcal{W}^h$. Inserting (3b) into (3a), the following differential system is obtained: find $\Psi^h : [0, T] \to \mathcal{W}^h$ such that

$$\Psi^h_t(t) - \mathcal{A}\Psi^h(t) = G^h(t, \Psi^h) + F^h(t), \qquad t \in (0, T], \tag{7a}$$

$$\Psi^h(0) = \Psi^h_0, \tag{7b}$$

where $\Psi^h(t) \equiv r_h\,(\psi^h(t))$ and $-\mathcal{A}\Psi^h(t) \equiv \mathcal{D}\,(\mathcal{G}\,\Psi^h(t))$ is the mimetic finite difference approximation to $-\operatorname{div}(K \operatorname{grad} \psi)$. Such an approximation uses the well-known harmonic average for the elements of K in the x- and y-direction, thus leading to a standard five-cell discretization on a rectangular grid (see [3] for details).

§3. Time integration

Let us decompose Ω into the union of m overlapping subdomains $\{\Omega_\ell\}_{\ell=1}^m$, each of which consists of a certain number of disjoint connected components, i.e., $\Omega = \bigcup_{\ell=1}^m \Omega_\ell$, where $\Omega_\ell = \bigcup_{j=1}^{m_\ell} \Omega_{\ell j}$ such that $\Omega_{\ell j} \cap \Omega_{\ell k} = \emptyset$ if $j \neq k$. By considering such a decomposition, we can define a partition of unity consisting of m smooth functions $\{\rho_\ell(\underline{x})\}_{\ell=1}^m$, with $\rho_\ell : \Omega \to [0, 1]$, such that $\sum_{\ell=1}^m \rho_\ell(\underline{x}) = 1$ for all $\underline{x} \in \Omega$ and $\operatorname{supp}(\rho_\ell) \equiv \Omega_\ell$, for $\ell = 1, 2, \ldots, m$.

Next, we introduce the splittings $\mathcal{A} = \sum_{\ell=1}^{m} \mathcal{A}_\ell$ and $F^h(t) = \sum_{\ell=1}^{m} F^h_\ell(t)$, such that:

$$
\begin{aligned}
(\mathcal{A}_\ell \Psi^h)_{(i,j)} = & \frac{\rho_\ell(c_{i+1,j})\,\tilde{k}_{11}(i+1,j)\,\frac{\Psi^h_{i+1,j}-\Psi^h_{i,j}}{h_x} - \rho_\ell(c_{i,j})\,\tilde{k}_{11}(i,j)\,\frac{\Psi^h_{i,j}-\Psi^h_{i-1,j}}{h_x}}{h_x} \\
& + \frac{\rho_\ell(c_{i,j+1})\,\tilde{k}_{22}(i,j+1)\,\frac{\Psi^h_{i,j+1}-\Psi^h_{i,j}}{h_y} - \rho_\ell(c_{i,j})\,\tilde{k}_{22}(i,j)\,\frac{\Psi^h_{i,j}-\Psi^h_{i,j-1}}{h_y}}{h_y},
\end{aligned}
\tag{8}
$$

where $\tilde{k}_{11}$ and $\tilde{k}_{22}$ denote the harmonic averages of k_{11} and k_{22} in the x- and y-direction, respectively, and $(F^h_\ell(t))_{(i,j)} = \rho_\ell(c_{i,j})|e_{i,j}|^{-1} \int_{e_{i,j}} f(\underline{x}, t)\, d\underline{x}$, being $|e_{i,j}|$ the area of cell $e_{i,j}$. Similar domain decomposition operator splittings have been previously used in [4]. Matrices $\{\mathcal{A}_\ell\}_{\ell=1}^{m}$ defined in (8) are block-tridiagonal, symmetric and non-positive definite, but they do not commute. From a theoretical point of view, this lack of commutativity requires the use of time integrators which are proven to be stable for non-commuting operators.

Following [5], let us now introduce a second-order fractional step method in order to reduce the semilinear stiff problem (7) to the solution of the following set of linear systems:

$$
\begin{cases}
\text{For } n = 0, 1, \ldots, N_T : \\
\quad \Psi^h_{n,1} = \Psi^h_n, \\
\quad \text{For } \ell = 2, 3, \ldots, 2m-1 : \\
\qquad \Psi^h_{n,\ell} = \Psi^h_{n,\ell-1} + \tau \displaystyle\sum_{k=\ell-1}^{\ell} d_k (\mathcal{A}_{i_k} \Psi^h_{n,k} + F^h_{i_k}(t_{n,k})) + \frac{\tau}{2}\, \Phi^h_{n,\ell}, \\
\quad \Psi^h_{n+1} = \Psi^h_{n,2m-1},
\end{cases}
\tag{9}
$$

where $\Phi^h_{n,2} = G^h(t_{n,1}, \Psi^h_{n,1})$, $\Phi^h_{n,2m-1} = 2\,G^h(t_{n,m}, \Psi^h_{n,m}) - G^h(t_{n,1}, \Psi^h_{n,1})$ and $\Phi^h_{n,\ell} \equiv 0$, for $\ell = 3, 4, \ldots, 2m-2$. The time step is denoted by τ and $N_T \equiv [T/\tau] - 1$. Subindex i_k is such that $i_k = k$, for $k = 1, 2, \ldots, m$, and $i_k = 2m-k$, for $k = m+1, m+2, \ldots, 2m-1$. On the other hand, the intermediate times are given by $t_{n,1} = t_n = n\tau$, $t_{n,k} = t_n + \frac{\tau}{2}$, for $k = 2, 3, \ldots, 2m-2$, and $t_{n,2m-1} = t_n + \tau$, while the coefficients of the internal stages are $d_1 = d_m = d_{2m-1} = \frac{1}{2}$ and $d_j = \frac{1}{4}$, for $j \in \{2, 3, \ldots, m-1\} \cup \{m+1, m+2, \ldots, 2m-2\}$. Finally, the totally discrete solution Ψ^h_{n+1} approximates $\Psi^h(t_{n+1})$. This method, which can be seen as a linearly implicit generalization of Peaceman-Rachford fractional step method, has been proven to be stable even for non-commuting matrices $\{\mathcal{A}_\ell\}_{\ell=1}^{m}$ (cf. [5]).

Note that the choice of a linearly implicit scheme like (9) entails an explicit treatment of the nonlinear semidiscrete function $G^h(t, \Psi^h)$. As a consequence, at each internal stage, we have to solve a linear system with associated matrix $(\mathcal{I} - \tau\, d\, \mathcal{A}_\ell)$, where $\mathcal{I}$ is the identity matrix of order $N_x N_y$, $d > 0$ and $\ell \in \{1, 2, \ldots, m\}$. Recalling that ρ_ℓ is considered in the definition of $\mathcal{A}_\ell$ and $\operatorname{supp}(\rho_\ell) \equiv \Omega_\ell$, such a linear system involves as many unknowns as the number of cell centers lying inside Ω_ℓ. Finally, since Ω_ℓ consists of the union of several disjoint components, the linear system to solve is, in fact, a collection of several uncoupled subsystems that can be solved in parallel. It is interesting to point out that, as a difference to classical domain decomposition techniques, our proposal does not require any Schwarz iteration procedures.

To finish the section, let us define the global error of the totally discrete scheme (9) at t_{n+1} as $\|E^h_{n+1}\|_h = \|r_h\,(\mathcal{P}_h\psi(\underline{x}, t_{n+1})) - \Psi^h_{n+1}\|_h$. A classical combination of suitable consistency and stability properties for the time integrator, together with the bound given in Theorem 1 for the spatial semidiscretization scheme, permits us to prove the following convergence result (cf. [1]).

Theorem 2. *Let* Ψ^h_{n+1} *be the solution of the totally discrete scheme* (9)*. Then, there exists a constant* $C > 0$*, independent of* h *and* τ*, such that* $\|E^h_{n+1}\|_h \leqslant C\,(h^2 + \tau^2)$ *for all* $n = 0, 1, \ldots, N_T$*.*

§4. A numerical example

Let us consider the semilinear parabolic problem given by (1), where $\Omega = (0, 1) \times (0, 1)$, $K(\underline{x}) = (1 + x^2 + y^2)\,\mathcal{I}$ and $g(t, \psi) = -\frac{1}{1+\psi^2}$. Data functions f and ψ_0 are defined in such a way that $\psi(\underline{x}, t) = e^{-t}(x^2(1 - x)^2 + y^2(1 - y)^2)$ is the exact solution of the problem.

We consider a rectangular mesh $\mathcal{T}_h$ that covers Ω with $N_x \times N_y$ cells and we carry out the spatial semidiscretization process described in section 2. In this case, $h = \max\{\frac{1}{N_x}, \frac{1}{N_y}\}$. Then, we define $I_1 \equiv (0, \frac{5}{16}] \cup [\frac{7}{16}, \frac{13}{16}]$ and $I_2 \equiv [\frac{3}{16}, \frac{9}{16}] \cup [\frac{11}{16}, 1)$ and we set out $\Omega_1 \equiv I_1 \times I_1$, $\Omega_2 \equiv I_2 \times I_1$, $\Omega_3 \equiv I_1 \times I_2$ and $\Omega_4 \equiv I_2 \times I_2$. Thus, the spatial domain $\Omega = \bigcup_{i=1}^4 \Omega_i$ is decomposed into $m = 4$ overlapping subdomains, each of which consists of 4 disjoint connected components. In order to define a smooth partition of unity $\{\rho_\ell\}_{\ell=1}^4$, we use suitable products of dilations and translations of a C^∞ function (cf. [1]). Finally, the integral averages of f over the mesh cells are computed by using the two-dimensional Simpson's rule.

In order to test the second-order convergence of the spatial semidiscretization scheme, we consider a small fixed time step $\tau = 10^{-5}$ and, starting from a mesh with 3×4 cells, we compute the global errors resulting from doubling the number of cells in each direction. Such errors, denoted by $E_{N_x,N_y,\tau}$, are measured in the maximum norm in time and the discrete L^2-norm in space and are displayed in the upper row of Table 1. From these results, we obtain the usual estimates for the order of convergence in space as $p_{N_x,N_y,\tau} = \log_2(E_{N_x,N_y,\tau}/E_{2N_x,2N_y,\tau})$. As shown in the lower row of Table 1, $p_{N_x,N_y,\tau}$ approaches 2 when h tends to 0, as predicted in Theorem 2.

Next, we compare the numerical solution obtained for a mesh size h and a time step τ with that obtained for the same mesh size and a time step $\tau/2$ (by using again the maximum norm in time and the discrete L^2-norm in space). When considering a small enough fixed h, this quantity estimates the global error in time and can be used to check the second-order convergence of the time integrator. In this case, we consider a fine spatial grid with 96×128 cells and, starting from a time step $\tau = 10^{-1}$, we compute the global errors resulting from halving the time step. From such errors, denoted by $\bar{E}_{N_x,N_y,\tau}$, we can estimate the orders of convergence in time as $\bar{p}_{N_x,N_y,\tau} = \log_2(\bar{E}_{N_x,N_y,\tau}/\bar{E}_{N_x,N_y,\tau/2})$. Table 2 shows the values of $\bar{E}_{N_x,N_y,\tau}$ (upper row) and $\bar{p}_{N_x,N_y,\tau}$ (lower row). Observe that the numerical orders of convergence also approach 2 when τ tends to 0, as stated in Theorem 2. Finally, it is important to note that, despite the fact that the nonlinear term is treated explicitly, the method shows an unconditionally stable behaviour.

(N_x, N_y)	(3,4)	(6,8)	(12,16)	(24,32)	(48,64)	(96,128)
$E_{N_x,N_y,\tau}$	6.526E-3	2.093E-3	5.511E-4	1.395E-4	3.499E-5	8.766E-6
$p_{N_x,N_y,\tau}$	1.641	1.925	1.982	1.995	1.997	–

Table 1: Global errors and numerical orders of convergence in space ($\tau = 10^{-5}$).

τ	0.1	$0.1 \cdot 2^{-1}$	$0.1 \cdot 2^{-2}$	$0.1 \cdot 2^{-3}$	$0.1 \cdot 2^{-4}$	$0.1 \cdot 2^{-5}$
$\bar{E}_{N_x,N_y,\tau}$	8.763E-2	4.523E-2	2.224E-2	1.020E-2	4.164E-3	1.559E-3
$\bar{p}_{N_x,N_y,\tau}$	0.954	1.013	1.135	1.292	1.417	1.672
τ	$0.1 \cdot 2^{-6}$	$0.1 \cdot 2^{-7}$	$0.1 \cdot 2^{-8}$	$0.1 \cdot 2^{-9}$	$0.1 \cdot 2^{-10}$	$0.1 \cdot 2^{-11}$
$\bar{E}_{N_x,N_y,\tau}$	4.893E-4	1.424E-4	4.036E-5	1.078E-5	2.775E-6	7.012E-7
$\bar{p}_{N_x,N_y,\tau}$	1.781	1.819	1.904	1.958	1.985	–

Table 2: Global errors and numerical orders of convergence in time ($N_x = 96$, $N_y = 128$).

Acknowledgements

This work is partially supported by the Spanish Ministry of Innovation and Science under Research Project MTM2007-63772 and by Government of Navarre under Research Projects 228/2008-2301 and CTP-07/R-10.

References

[1] Arrarás, A., Portero, L., and Jorge, J. C. Convergence of fractional step mimetic finite difference discretizations for semilinear parabolic problems. *Appl. Numer. Math. 60* (2010), 473–485.

[2] Berndt, M., Lipnikov, K., Shashkov, M., Wheeler, M. F., and Yotov, I. Superconvergence of the velocity in mimetic finite difference methods on quadrilaterals. *SIAM J. Numer. Anal. 43* (2005), 1728–1749.

[3] Hyman, J., Shashkov, M., and Steinberg, S. The numerical solution of diffusion problems in strongly heterogeneous non-isotropic materials. *J. Comput. Phys. 132* (1997), 130–148.

[4] Mathew, T. P., Polyakov, P. L., Russo, G., and Wang, J. Domain decomposition operator splittings for the solution of parabolic equations. *SIAM J. Sci. Comput. 19* (1998), 912–932.

[5] Portero, L., and Jorge, J. C. A new class of second order linearly implicit fractional step methods. *J. Comput. Appl. Math. 218* (2008), 603–615.

Laura Portero, Andrés Arrarás and Juan Carlos Jorge
Departamento de Ingeniería Matemática e Informática
Universidad Pública de Navarra
Campus de Arrosadía, 31006 - Pamplona, Spain
`{laura.portero,andres.arraras,jcjorge}@unavarra.es`

Monografías Matemáticas *García de Galdeano* **35**, 237–244 (2010)

EXISTENCE OF A SOLUTION TO A CLASS OF PSEUDOPARABOLIC PROBLEMS

Ngonn Seam and Guy Vallet

Abstract. In this paper we are interested, on the one hand, in problems involving a nonlinearity of form $f(\partial_t u)$; on the other hand, we are interested in Barenblatt's type equations [5] too.

By the way of an implicit time-discretization, we would prove the existence of a solution to the following problem: $f(\partial u_t) - \Delta\phi(u) - \epsilon\Delta\partial u_t = g$ with a Lipschitz-continuous function ϕ.

Keywords: Pseudoparabolic problems, existence results, time-discretization.

AMS classification: 35K65, 35K70.

§1. Introduction

In this paper, we are interested in the mathematical analysis of the pseudoparabolic Cauchy problem:

$$f(\partial_t u) - \Delta\phi(u) - \epsilon\Delta\partial_t u = g, \quad u(0, .) = u_0, \tag{1}$$

where f and ϕ are Lipschitz-continuous functions with f non-decreasing.

This study has its roots in the analysis of problems with a nonlinearity of form $f(\partial_t u)$. Such a term has been previously introduced by G. I. Barenblatt in [5] for elasto-plastic porous media. It has been revisited by S. N. Antontsev *et al.* [1, 2, 3, 4] or G. Vallet [8] concerning a constrained stratigraphic models in geology.

An implicit time-discretization scheme is used to prove the existence of a solution in a suitable functional space. As an application, by passing to the limits with respect to ϵ, one proves the existence of a solution to the Barenblatt's equation.

Let us consider in the sequel a bounded domain $\Omega \subset \mathbb{R}^d$ with a Lipschitz-boundary Γ. For any $T > 0$, let us denote Q a cylinder defined by $Q :=]0, T[\times \Omega$.
Moreover, one assumes that:

$$f \text{ is a non-decreasing Lipschitz-continuous function,} \tag{H_1}$$

$$\phi \text{ is a } C^1(\mathbb{R})\text{-Lipschitz-continuous function such that } \phi(0) = 0, \tag{H_2}$$

$$\epsilon > 0 \text{ and } u_0 \in H_0^1(\Omega), \tag{H_3}$$

$$g \in L^2(Q). \tag{H_4}$$

We shall write $M = \|\phi'\|_\infty$.

Let us define now what is a solution to our pseudoparabolic problem.

Definition 1. A solution to (1) is any $u \in H^1(0, T, H_0^1(\Omega))$ such that $u(0, \cdot) = u_0$ and, for all v in $H_0^1(\Omega)$,

$$\int_\Omega \left\{ f\left(\partial_t u\right) v + \phi'\left(u\right) \nabla u \nabla v + \epsilon \nabla \partial_t u \nabla v \right\} dx = \int_\Omega g v \, dx. \tag{2}$$

The main result of this paper is that

Theorem 1. *There exists a solution to Problem* (1).

§2. Existence of a solution

2.1. Semi-discretized processes

Consider a positive integer N and denote by $h = T/N$. In this section, we are interested in proving the existence of the sequence of approximation by the way of an implicit semi-discretization scheme.

Each step of the scheme consist in solving a nonlinear elliptic problem. In a first par, the case of a bounded f would be consider. Then, thanks to some truncation arguments, the general case would be obtained.

Proposition 2. *Under the hypothesis* (H_1) *to* (H_3) *and by assuming moreover that* f *is a bounded function, if* h *is small enough* ($h < \epsilon/(M + 1)$)*, for any* $g \in L^2(\Omega)$*, there exists an element* u *in* $H_0^1(\Omega)$ *such that, for all* v *in* $H_0^1(\Omega)$*,*

$$\int_\Omega f\left(\frac{u - u_0}{h}\right) v \, dx + \int_\Omega \phi'\left(u\right) \nabla u \nabla v, dx + \epsilon \int_\Omega \nabla \frac{u - u_0}{h} \nabla v \, dx = \int_\Omega g v \, dx. \tag{3}$$

This element is unique as soon as ϕ' *is a Lipschitz-continuous function.*

Proof. The existence of a solution of (2) is classically obtained by using the Schauder-Tikhonov fixed point theorem in the framework of separable reflexive B-spaces. In order to do it, let us denoted Ψ the mapping defined by $\Psi : H_0^1(\Omega) \to H_0^1(\Omega)$, $S \mapsto u_S$, where u_S is the unique solution of the following linear problem: find $u_S \in H_0^1(\Omega)$ such that, for all $v \in H_0^1(\Omega)$,

$$\int_\Omega \left(\phi'\left(S\right) + \frac{\epsilon}{h}\right) \nabla u_S \nabla v \, dx = \int_\Omega g v \, dx - \int_\Omega f\left(\frac{S - u_0}{h}\right) v \, dx + \frac{\epsilon}{h} \int_\Omega \nabla u_0 \nabla v \, dx. \tag{4}$$

As soon as $h < \epsilon/(M + 1)$, this linear problem is coercive in $H_0^1(\Omega)$. It is well-posed and Ψ exists. Choosing $v = u_S$ a test function, one gets that

$$\left\| u_{S_n} \right\|_{H_0^1(\Omega)} \leq C_1 = C\left(\Omega, \|f\|_\infty, g, \epsilon, u_0, h\right), \tag{5}$$

and Ψ conserve the closed ball $\bar{B}_{H_0^1(\Omega)}(0, C_1)$.

Let (S_n) be a sequence that converges weakly in $H_0^1(\Omega)$ towards S. Up to a subsequence still denoted in the same way, it can be assumed that S_n converges strongly in $L^2(\Omega)$ and *a.e.*

in Ω. Furthermore, the functions ϕ' and f are continuous and bounded, then owing to the theorem of Lebesgue, we can prove that, for all v in $H_0^1(\Omega)$,

$$\int_\Omega f\left(\frac{S_n - u_0}{h}\right) v\,dx \to \int_\Omega f\left(\frac{S - u_0}{h}\right) v\,dx \quad \text{and} \quad \phi'\left(S_n\right)\nabla v \to \phi'\left(S\right)\nabla v \quad \left(L^2(\Omega)\right)^d, \tag{6}$$

Moreover, according to (5), the sequence (u_{S_n}) is bounded in $H_0^1(\Omega)$. Thus, χ in $H_0^1(\Omega)$ exists, as well as a subsequence, still indexed by n, extracted from (u_{S_n}), such that, u_{S_n} converges weakly in $H_0^1(\Omega)$ toward χ. Then, we have that

$$\nabla u_{S_n} \rightharpoonup \nabla\chi \quad \text{in } \left(L^2(\Omega)\right)^d \quad \text{and} \quad \nabla\frac{u_{S_n} - u_0}{h} \rightharpoonup \nabla\frac{\chi - u_0}{h} \quad \text{in } \left(L^2(\Omega)\right)^d. \tag{7}$$

Passing to the limits in (4) with S_n by using (6) and (7), we obtain that χ is a solution to problem (4) with S. By uniqueness of such a solution, one gets that $\chi = u_S$.

Thus by a compactness argument, all the sequences converge weakly in $H_0^1(\Omega)$ toward u_S, *i.e.* $u_{S_n} \rightharpoonup u_S$ weakly in $H_0^1(\Omega)$. Then the mapping Ψ is sequentially weakly weakly continuous in $H_0^1(\Omega)$. Thus the fixed point theorem of Schauder-Tikhonov proves that Ψ has at most a fixed point; *i.e.* there exists S in $H_0^1(\Omega)$ such that $u_S = S$ and a solution to (3) exists.

Let us prove now that this solution is unique. Let us consider $\widehat{u}$ another solution of (3). Thus we obtain by subtraction, for all v in $H_0^1(\Omega)$,

$$\begin{aligned}0 = &\int_\Omega \left[f\left(\frac{u - u_0}{h}\right) - f\left(\frac{\widehat{u} - u_0}{h}\right)\right] v\,dx + \int_\Omega \left(\phi'\left(u\right) + \frac{\epsilon}{h}\right)\nabla\left(u - \widehat{u}\right)\nabla v\,dx\\ &+ \int_\Omega \left(\phi'\left(u\right) - \phi'\left(\widehat{u}\right)\right)\nabla\widehat{u}\nabla v\,dx.\end{aligned} \tag{8}$$

For a giving $\mu > 0$, let us denote by $p_\mu(r) = (r - \mu)^+/r$; p_μ is non-decreasing Lipschitz function with $p'_\mu(r) = \frac{\mu}{r^2}\mathbf{1}_{\{r>\mu\}}$.

Therefore, as $v = p_\mu\left(u - \widehat{u}\right)$ is a suitable test function, its comes that

$$\begin{aligned}0 = &\int_\Omega \left[f\left(\frac{u - u_0}{h}\right) - f\left(\frac{\widehat{u} - u_0}{h}\right)\right] p_\mu\left(u - \widehat{u}\right)dx + \mu\int_{\{u-\widehat{u}>\mu\}} \left(\phi'\left(u\right) + \frac{\epsilon}{h}\right)\frac{\left|\nabla\left(u - \widehat{u}\right)\right|^2}{|u - \widehat{u}|^2}dx\\ &+ \mu\int_{\{u-\widehat{u}>\mu\}} \frac{\phi'\left(u\right) - \phi'\left(\widehat{u}\right)}{|u - \widehat{u}|^2}\nabla\widehat{u}.\nabla\left(u - \widehat{u}\right)dx.\end{aligned}$$

Since f is a non-decreasing function and as $h \leq \epsilon/(M + 1)$, it comes that

$$\begin{aligned}\int_{\{u-\widehat{u}>\mu\}} \frac{\left|\nabla\left(u - \widehat{u}\right)\right|^2}{|u - \widehat{u}|^2}dx &\leq \int_{\{u-\widehat{u}>\mu\}} \frac{\left|\phi'\left(u\right) - \phi'\left(\widehat{u}\right)\right|^2}{2|u - \widehat{u}|^2}|\nabla\widehat{u}|^2 dx + \int_{\{u-\widehat{u}>\mu\}} \frac{|\nabla(u - \widehat{u}|^2)}{2|u - \widehat{u}|^2}dx\\ &\leq \int_{\{u-\widehat{u}>\mu\}} \frac{\left|\phi'\left(u\right) - \phi'\left(\widehat{u}\right)\right|^2}{|u - \widehat{u}|^2}|\nabla\widehat{u}|^2 dx \leq \|\phi''\|_\infty \int_\Omega |\nabla\widehat{u}|^2 dx.\end{aligned}$$

Let us denote by $F_\mu(r) = \ln\left(1 + (r - \mu)^+/\mu\right)$. F_μ is a Lipchitz-continuous function, $F_\mu\left(u - \widehat{u}\right) \in H_0^1(\Omega)$ and one gets that

$$\int_\Omega \left|\nabla F_\mu\left(u - \widehat{u}\right)\right|^2 dx \leq \left\|\phi''\right\|_\infty \int_\Omega |\nabla\widehat{u}|^2\,dx.$$

Thanks to Poincaré inequality, the sequence $\left(F_\mu\left(u-\widehat{u}\right)\right)_\mu$ is bounded in $L^2(\Omega)$ independently of μ. Note that the sequence $\left(F_{1/n}\left(u-\widehat{u}\right)\right)_n$ is non-decreasing, and converges almost everywhere in $\mathbb{R}\cup\{+\infty\}$ to $+\infty\,\mathbf{1}_{\{u-\widehat{u}>0\}}$. Hence, the theorem of Beppo Levi leads to $\operatorname{meas}\left(\{u>\widehat{u}\}\right)=0$. Then $(u-\widehat{u})^+=0$, *i.e* $u\le\widehat{u}$.

Permutating u and $\widehat{u}$ thereinbefore gives $\widehat{u}\le u$ as well and the solution is unique. □

Proposition 3. *Under the hypothesis* (H_1) *to* (H_3)*, if* h *is small enough* ($h<\epsilon/(M+1)$)*, for any* $g\in L^2(\Omega)$*, there exists an element* u *in* $H^1_0(\Omega)$ *such that, for all* v *in* $H^1_0(\Omega)$*,*

$$\int_\Omega f\left(\frac{u-u_0}{h}\right)v\,dx+\int_\Omega\nabla\phi\left(u\right)\nabla v\,dx+\epsilon\int_\Omega\nabla\frac{u-u_0}{h}\nabla v\,dx=\int_\Omega gv\,dx. \tag{9}$$

This element is unique as soon as ϕ' *is a Lipschitz-continuous function.*

Proof. The proof of the uniqueness result of the solution is identical to the one proposed previously.

Concerning the result of existence, consider for any positive n, $f_n=\max\left(-n,\min\left(n,f\right)\right)$. The corresponding solutions, given by the above proposition, are denoted by u_n. Applying the test function $v=(u_n-u_0)/h$ to (3), one gets that

$$\begin{aligned}\int_\Omega\left[f_n\left(\frac{u_n-u_0}{h}\right)-f_n(0)\right]\frac{u_n-u_0}{h}\,dx+\int_\Omega\left[h\phi'\left(u_n\right)+\epsilon\right]\left|\nabla\frac{u_n-u_0}{h}\right|^2\,dx\\ \le\int_\Omega\left[g-f_n(0)\right]\frac{u_n-u_0}{h}\,dx-\int_\Omega\phi'\left(u_n\right)\nabla u_0\nabla\frac{u_n-u_0}{h}\,dx\\ \le\left[\|g-f_n(0)\|_{L^2(\Omega)}+M\|u_0\|_{H^1_0(\Omega)}\right].\left\|\frac{u_n-u_0}{h}\right\|_{H^1_0(\Omega)}.\end{aligned}$$

Since f is non-decreasing, f_n too, $h<\epsilon/(M+1)$ and thanks to Poincaré's inequality, one gets that

$$\left\|\frac{u_n-u_0}{h}\right\|_{H^1_0(\Omega)}\le\|g\|_{L^2(\Omega)}+|f(0)|\sqrt{\operatorname{meas}(\Omega)}+M\|u_0\|_{H^1_0(\Omega)}. \tag{10}$$

Therefore, a sub-sequence still indexed by n can be extracted, such that u_n converges in $H^1_0(\Omega)$ weakly to u, strongly in $L^2(\Omega)$ and *a.e.* in Ω. Moreover, one has that

$$\left\|f_n(\frac{u_n-u_0}{h})\right\|_{H^1_0(\Omega)}\le\left\|f'\right\|_\infty\left[\|g\|_{L^2(\Omega)}+|f(0)|\sqrt{\operatorname{meas}(\Omega)}+M\|u_0\|_{H^1_0(\Omega)}\right]. \tag{11}$$

Since $f_n(\frac{u_n-u_0}{h})$ converges a.e. to $f(\frac{u-u_0}{h})$, it ensures that $f\left(u_n\right)$ converges in $L^2(\Omega)$ toward $f(u)$ (and weakly in $H^1(\Omega)$). Furthermore, since ϕ is a Lipschitz-continous function, $\phi(u_n)$ converges weakly to $\phi(u)$ in $L^2(\Omega)$, and, passing to the limits in the variational formulation stating u_n, one gets (9). □

Inductively, the following result can be proved:

Theorem 4. *Let us consider $N \in \mathbb{N}^*$ with $N > T(M+1)/\epsilon$, $h = T/N$ and $(g^k) \subset L^2(\Omega)$. Then, under the hypothesis (H_1)–(H_3), there exists a sequence $(u^k)_k$ in $H^1_0(\Omega)$ with $u^0 = u_0$ and such that, for all $v \in H^1_0(\Omega)$,*

$$\int_\Omega f\left(\frac{u^{k+1}-u^k}{h}\right) v\,dx + \int_\Omega \nabla\phi\left(u^{k+1}\right)\nabla v\,dx + \epsilon\int_\Omega \nabla\frac{u^{k+1}-u^k}{h}\nabla v\,dx = \int_\Omega g^{k+1} v\,dx. \tag{12}$$

This sequence is unique as soon as ϕ' is a Lipschitz-continuous function.

2.2. Existence of a solution

In order to prove the existence of a solution, let us introduce some notations. For any sequence v^k, let us denote in the sequel

$$v^h = \sum_{k=0}^{N-1} v^{k+1}\mathbf{1}_{[t_k,t_{k+1}[} \quad \text{and} \quad \widetilde{v}^h = \sum_{k=0}^{N-1}\left[\frac{v^{k+1}-v^k}{h}(t-t_k)+v^k\right]\mathbf{1}_{[t_k,t_{k+1}[},$$

where $t_k = kh$ and

$$g^h = \sum_{k=0}^{N-1}\frac{1}{h}\int_{kh}^{(k+1)h} g(t,\cdot)dt\,\mathbf{1}_{[t_k,t_{k+1}[}.$$

Lemma 5. *Assume that $h < \epsilon/(M+1)$. Then,*

(i) The sequence (u^h) is bounded in $L^\infty(0,T;H^1_0(\Omega))$ and $(\widetilde{u}^h)$ is bounded in $H^1(0,T;H^1_0(\Omega)) \cap L^\infty(0,T;H^1_0(\Omega))$.

(ii) There exists $C > 0$ such that for all t in $[0,T[$, $\left\|\widetilde{u}^h(t)-u^h(t)\right\|_{H^1_0(\Omega)} \le C\sqrt{h}$.

(iii) There exists a set Z of full measure in $]0,T[$ such that, for any t in Z, $\partial_t\widetilde{u}^h(t)$ is bounded in $H^1_0(\Omega)$.

Proof. Thanks to (10), one has that

$$\left\|\frac{u^{k+1}-u^k}{h}\right\|_{H^1_0(\Omega)} \le \left\|g^{k+1}\right\|_{L^2(\Omega)} + |f(0)|\sqrt{\text{meas}(\Omega)} + M\left\|u^k\right\|_{H^1_0(\Omega)}, \tag{13}$$

and, if $k > 0$,

$$\left\|\frac{u^{k+1}-u^k}{h}\right\|_{H^1_0(\Omega)} \le \left\|g^{k+1}\right\|_{L^2(\Omega)} + C + M\left\|u_0\right\|_{H^1_0(\Omega)} + Mh\sum_{i=0}^{k-1}\left\|\frac{u^{i+1}-u^i}{h}\right\|_{H^1_0(\Omega)}. \tag{14}$$

Then, one gets that

$$\begin{aligned}\sum_{k=0}^{n} h\left\|\frac{u^{k+1}-u^k}{h}\right\|^2_{H^1_0(\Omega)} &\le 4\sum_{k=0}^{n} h\left\|g^{k+1}\right\|^2_{L^2(\Omega)} + C(u_0)T + 4M^2h^2\sum_{k=1}^{n} h\left[\sum_{i=0}^{k-1}\left\|\frac{u^{i+1}-u^i}{h}\right\|_{H^1_0(\Omega)}\right]^2 \\ &\le C(g,u_0) + 4M^2Th\sum_{k=1}^{n}\sum_{i=0}^{k-1} h\left\|\frac{u^{i+1}-u^i}{h}\right\|^2_{H^1_0(\Omega)} \le C(g,u_0)e^{4M^2T},\end{aligned}$$

thanks to the discrete Gronwall lemma. This yields

$$\sum_{k=0}^{N-1} \left\|u^{k+1} - u^k\right\|^2_{H^1_0(\Omega)} \leq hC(g, u_0)e^{4M^2T}, \tag{15}$$

and (i)–(ii) hold.

Moreover, (14) yields, for any $t \in]t_k, t_{k+1}[$, to

$$\left\|\partial_t \widetilde{u}^h(t)\right\|^2_{H^1_0(\Omega)} \leq 4 \left\|g^h(t)\right\|^2_{L^2(\Omega)} + C(u_0) + 4M^2C(g, u_0)e^{4M^2T}. \tag{16}$$

If moreover t belongs to the set of Lebesgue of g in $L^2(0, T; L^2(\Omega))$, $\partial_t \widetilde{u}^h(t)$ is bounded in $H^1_0(\Omega)$ and (iii) holds. □

Theorem 6. *Under the hypotheses* (H_1)*–*(H_4)*, there exists* u *in* $H^1\left(0, T; H^1_0(\Omega)\right)$ *such that, for all* v *in* $H^1_0(\Omega)$*,*

$$\int_\Omega f(\partial_t u)\, v\, dx + \int_\Omega \nabla\phi(u) \nabla v\, dx\epsilon + \int_\Omega \nabla\partial_t u \nabla v dx = \int_\Omega gv\, dx, \tag{17}$$

with $u(0, \cdot) = u_0$*.*

Proof. Leading from Lemma 5-(i), there exists u in $H^1(0, T; H^1_0(\Omega))$, such that, up to a sub-sequences still denoted in the same way, one may assume that $\widetilde{u}^h$ converges to u weakly in $H^1(0, T; H^1_0(\Omega))$. Then, for any t in $[0, T]$, $\widetilde{u}^h(t)$ converges weakly in $H^1_0(\Omega)$ toward $u(t)$. Then, Lemma 5-(ii) ensures that $u^h(t)$ converges weakly to $u(t)$ in $H^1_0(\Omega)$. Moreover, since ϕ is a Lipschitz-countinuous function, $\phi(u^h(t))$ converges weakly to $\phi(u(t))$ in $H^1_0(\Omega)$ too.

Thanks to Lemma 5-(iii), for any t in Z, up to a sub-sequence indexed by h_t, $\partial_t \widetilde{u}^{h_t}(t)$ converges weakly in $H^1_0(\Omega)$ towards a given $\xi(t)$ and strongly in $L^2(\Omega)$.

Then, there exists k such that (12) leads, for any $v \in H^1_0(\Omega)$, to

$$\int_\Omega f\left(\partial_t \widetilde{u}^{h_t}(t)\right) v\, dx + \int_\Omega \nabla\phi\left(u^{h_t}(t)\right) \nabla v\, dx + \epsilon \int_\Omega \nabla\partial_t \widetilde{u}^{h_t}(t) \nabla v\, dx = \int_\Omega g^{h_t}(t) v\, dx. \tag{18}$$

By passing to the limits in the above equation, on gets that $\xi(t)$ is a solution in in $H^1_0(\Omega)$ to the variational problem:

$$\forall v \in H^1_0(\Omega), \int_\Omega f(\xi(t))\, v\, dx + \epsilon \int_\Omega \nabla\xi(t) \nabla v\, dx = \int_\Omega gv dx - \int_\Omega \phi'(u(t)) \nabla u(t) \nabla v\, dx. \tag{19}$$

Then, since f is non-decreasing, this implies that such a solution is unique. As $\partial_t \widetilde{u}^h(t)$ is a bounded sequence in $H^1_0(\Omega)$, one concludes that $\partial_t \widetilde{u}^h(t)$ converges toward $\xi(t)$ weakly in $H^1_0(\Omega)$.

Therefore, $\xi :]0, T[\rightarrow H^1_0(\Omega)$ is a weakly measurable function. Then, thanks to the theorem of Pettis ([9, p. 131]), it is a measurable function.

For any v in $H_0^1(\Omega)$, $\int_\Omega \nabla\partial_t u^h(t)\nabla v\,dx$ converges *a.e.* in $]0,T[$ toward $\int_\Omega \nabla\xi(t)\nabla v\,dx$. Since $\left|\int_\Omega \nabla\partial_t\widetilde{u}^h(t)\nabla v\,dx\right| \leq \left\|\partial_t\widetilde{u}^h(t)\right\|_{H_0^1(\Omega)}\|v\|_{H_0^1(\Omega)}$, it is bounded in $L^2(0,T)$ and [7, Lemma 1.3, p.12] ensures that

$$\forall\alpha\in L^2(0,T),\ \int_0^T\int_\Omega \alpha(t)\nabla\partial_t\widetilde{u}^h(t).\nabla v\,dx\,dt \to \int_0^T\int_\Omega \alpha(t)\nabla\xi(t).\nabla v\,dx\,dt.$$

Since $(\partial_t\widetilde{u}^h)$ is bounded in $L^2(0,T;H_0^1(\Omega))$, an argument of density leads to the weak convergence in $L^2(0,T;H_0^1(\Omega))$ of $\partial_t\widetilde{u}^h$ toward ξ. Thus by uniqueness of the weak limit, one obtains that $\partial_t u = \xi$ and that there exists a solution. □

§3. Application to Barenblatt's equation

As an application, let us return to the existence of a solution to Barenblatt's equation:

$$f(\partial_t u) - \Delta u = g,$$

where $f(r) = r$ if $r > 0$ and $f(r) = \alpha r$ $(\alpha > 0)$ if $r \leq 0$, with $\alpha \neq 1$ *a priori*.

Our method consists in passing to the limits in the pseudoparabolic problem (2) with respect to ϵ toward 0, when $\phi = Id$, g in $L^2(Q)$ and u_0 in $H_0^1(\Omega)$.

By using the test function $v = \partial_t u_\epsilon$ in (2), we obtain, for any t, the following estimate:

$$\int_{\Omega\times]0,t[} f(\partial_t u_\epsilon)\partial_t u_\epsilon + \epsilon|\nabla\partial_t u_\epsilon|^2\,dx + \frac{1}{2}\int_\Omega |\nabla u_\epsilon(t)|^2\,dx = \int_{\Omega\times]0,t[} g\partial_t u_\epsilon\,dx + \frac{1}{2}\int_\Omega |\nabla u_0|^2\,dx. \quad (20)$$

Thus, the sequence (u_ϵ) is bounded in $H^1(Q)\cap L^\infty(0,T;H_0^1(\Omega))$ as well as $(f(\partial_t u_\epsilon))$ in $L^2(Q)$. Indeed, for all t,

$$\min(1,\alpha)\int_{]0,t[\times\Omega} |\partial_t u_\epsilon|^2\,dx\,dt + \frac{1}{2}\int_\Omega |\nabla u_\epsilon(t)|^2\,dx \leq \frac{1}{2}\int_\Omega |\nabla u_0|^2\,dx + \int_{]0,t[\times\Omega} g\partial_t u_\epsilon\,dx\,dt.$$

Up to a sub-sequence still indexed by ϵ, one assumes that there exists u in $H^1(Q)\cap L^\infty(0,T;H_0^1(\Omega))$, weak limit in $H^1(Q)$ and weak-* limit in $L^\infty(0,T;H_0^1(\Omega))$ of (u_ϵ); as well as χ, weak limit in $L^2(Q)$ of $f(\partial_t u_\epsilon)$.

On the one hand, one has $\chi - \Delta u = g$, *i.e.* $\partial_t u - \Delta u = g + \partial_t u - \chi := h$. Since $h \in L^2(Q)$ with the initial condition in $H_0^1(\Omega)$, one gets

$$\int_Q |\partial_t u|^2\,dx\,dt + \frac{1}{2}\int_\Omega |\nabla u(T)|^2\,dx = \frac{1}{2}\int_\Omega |\nabla u_0|^2\,dx + \int_Q [g + \partial_t u - \chi]\partial_t u\,dx\,dt. \quad (21)$$

On the other hand, since $(u_\epsilon(T))$ bounded in $H_0^1(\Omega)$ and as $u_\epsilon(T)$ converges toward $u(T)$ in $L^2(\Omega)$, it converges weakly in $H_0^1(\Omega)$ and passing to the limits in (20) yields

$$\limsup_{\epsilon\to 0}\int_Q f(\partial_t u_\epsilon)\partial_t u_\epsilon\,dx\,dt + \frac{1}{2}\int_\Omega |\nabla u(T)|^2\,dx \leq \frac{1}{2}\int_\Omega |\nabla u_0|^2\,dx + \int_Q g\partial_t u\,dx\,dt.$$

Thus, $\limsup\epsilon\to 0\int_Q f(\partial_t u_\epsilon)\partial_t u_\epsilon\,dx\,dt \leq \int_Q \chi\partial_t u\,dx\,dt$. Then, according to H. Brézis [6, Prop. 2.5, p. 27], $\chi = f(\partial_t u)$ and u is a solution to the problem.

References

[1] Antontsev, S. N., Gagneux, G., Luce, R., and Vallet, G. New unilateral problems in stratigraphy. *M2AN Math. Model. Numer. Anal. 40*, 4 (2006), 765–784.

[2] Antontsev, S. N., Gagneux, G., Luce, R., and Vallet, G. A non-standard free boundary problem arising from stratigraphy. *Anal. Appl. (Singap.) 4*, 3 (2006), 209–236.

[3] Antontsev, S. N., Gagneux, G., Luce, R., and Vallet, G. On a pseudoparabolic problem with constraint. *Differential Integral Equations 19*, 12 (2006), 1391–1412.

[4] Antontsev, S. N., Gagneux, G., Mokrani, A., and Vallet, G. Stratigraphic modelling by the way of a pseudoparabolic problem with constraint. *Advances in Mathematical Science and Applications* (To appear).

[5] Barenblatt, G. I. Similarity, self-similarity, and intermediate asymptotics. *New York, London: Consultants Bureau. XVII* (1982).

[6] Brézis, H. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert.* North-Holland Mathematics Studies. 5. Notas de matematica (50). Amsterdam-London: North-Holland Publishing Comp.; New York: American Elsevier Publishing Comp., Inc. 183 p., 1973.

[7] Lions, J.-L. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod, 1969.

[8] Vallet, G. Sur une loi de conservation issue de la géologie. *C. R. Math. Acad. Sci. Paris 337*, 8 (2003), 559–564.

[9] Yosida, K. *Functional analysis*, fourth ed. Springer-Verlag, New York, 1974. Die Grundlehren der mathematischen Wissenschaften, Band 123.

Ngonn Seam and Guy Vallet
LMA, University of Pau
IPRA BP 1155 Pau Cedex (France)
seamngonn@yahoo.fr and guy.vallet@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **35**, 245–251 (2010)

On a stochastic nonlinear conservation law

Guy Vallet

Abstract. In this paper, we are interested in the stochastic viscous Buckley-Leverett equation with a Hölder continuous nonlinear function.

Keywords: Viscous Buckley-Leverett, stochastic perturbation, well-posedness.

AMS classification: 35K60, 35K65, 60H15.

§1. Introduction

In our presentation "On stochastic nonlinear conservation laws", at the Ninth International Conference Zaragoza-Pau on Applied Mathematics and Statistics, we have presented results of existence and uniqueness for the solution to parabolic and hyperbolic problems. These results were extracted from the publications G. Vallet [10] and G. Vallet and P. Wittbold [11]. In this paper, we would like to revisit the example of the formal stochastic viscous Buckley-Leverett equation

$$du - \epsilon\Delta u\,dt - \operatorname{div}(f(u)\vec{B})\,dt = h dw \quad \text{in } D \times]0,T[\times\Omega,$$

where f is assumed to be a Hölder continuous function.

In the sequel, one assumes that D is a bounded Lipschitz domain of $\mathbb{R}^d$, that T is a positive number and one denotes by $Q =]0,T[\times D$. Then, homogeneous Dirichlet would be considered.

Thereafter, $W = \{w_t, \mathcal{F}_t\,;\, 0 \le t \le T\}$ denotes a standard adapted one-dimensional continuous Brownian motion, defined on some probability space $(\Omega, \mathcal{F}, P)$, with the property that $w_0 = 0$. This assumption on W is made for convenience. Our aim is to adapt known methods for nonlinear PDE to noise perturbed ones.

Usually, the Buckley-Leverett equation is a transport equation used to model two-phase flow in porous media ($\epsilon = 0$). It can be obtained as the limit, when ϵ goes to 0, of the above viscous equation. Such a result can be found in G. Vallet and P. Wittbold [11] for a regular function f, but one needs the notion of entropy solution. Note that the model corresponds to a generalization to $d > 1$ of the Burger's equation too: *i.e.* $d = 1$ and $f(x) = x^2$.

The Burger's equation has been intensively studied in the literature with many extensions. Usually, the stochastic convolution is used. Let us mention, without exhaustiveness, G. Da Prato *et al.* [2, 3], W. Grecksch *et al.* [4] or I. Gyongy *et al.* [5] and M. Röckner *et al.* in [9] for a generalization of the classical.

Usually, Lipschitz or local-Lipschitz conditions are assumed on the function f. In this application we consider that f is a $1/2$-Hölder-continuous function (with $f(0) = 0$ since $\operatorname{div}\vec{B} = 0$). The method consists in using a Lipschitz-approximation of f and passing to the limits with respect to this approximation.

§2. Assumptions, definition of a solution and the main result

Let us assume in the sequel that

- $\vec{B} \in (L^\infty(D))^d$ with div $\vec{B} = 0$ a.e. in D,
- $f : \mathbb{R} \to \mathbb{R}$ is a 1/2 Hölder-continuous function with $f(0) = 0$,
- $h \in L^2(Q)$ and $u_0 \in L^2(D)$.

Denote by $V = H_0^1(D)$, endowed with $\|u\|_V = \left(\int_D |\nabla u|^2\, dx\right)^{1/2}$ the norm of Poincaré (*cf.* R. Adams [1, Th. 6.28, p.159]), by C_p the Poincaré's constant, *i.e.*, for all $v \in V$, $\|v\|_{L^2(D)} \leq C_p\|v\|_V$.

Our aim is then to give a result of existence and uniqueness of the variational solution to the above-mentioned problem. Let us fix in what sense such a solution is understood.

Definition 1. Any function u of $L^2(\Omega \times]0,T[\,; V)$ such that $\frac{\partial}{\partial t}\left[u - \int_0^t h(s,.)\, dw(s)\right]$, taken in the sense of the vectorial V'-valued distributions, belongs to $L^2(\Omega \times]0,T[\,; V')$ is a solution to our stochastic problem if u is $L^2(D)$-valued progressively measurable and if for t a.e. in $]0,T[$ and any test-function v of V, the variational formulation holds

$$0 = \left\langle \frac{\partial}{\partial t}\left[u - \int_0^t h(s,.,)\, dw(s)\right], v\right\rangle_{V',V} + \int_D \{\epsilon\nabla u.\nabla v + f(u)\vec{B}.\nabla v\}\, dx,$$

with the initial condition $u(0,.) = u_0$.

The results we want to prove is:

Theorem 1. *A unique solution in the sense of the above definition exists to the above stochastic Buckley-Leverett equation.*

§3. Proof of the result

For any positive M, consider $f_M = (f * \rho_M) \circ T_M$ where ρ_M denotes the usual mollifier sequence of support $1/M$ and $T_M(x) = \max[\min(x, M), -M]$. Then, f_M is a bounded, Lipschitz-continuous function and classical results yield the existence and uniqueness of the solution, denote by u_M, to the problem:

$$du_M - \epsilon\Delta u_M\, dt - \mathrm{div}(f_M(u_M)\vec{B})dt = hdw \quad \text{in } D \times]0,T[\, \times \Omega$$

for the same initial condition and the regularity required in the previous definition.

Such a result would be admitted; refer *e.g.* to G. Da Prato *et al.* [3], W. Grecksch *et al.* [4] or G. Vallet [10].

Thanks to the stochastic-energy equality, one has that a positive constant C exists such that, for any t,

$$E\int_D u_M^2(t)\, dx + 2E\int_0^t\int_D |\nabla u_M|^2\, dx\, ds + 2E\int_0^t\int_D f_M(u_M)\vec{B}.\nabla u_M\, dx\, ds = \int_0^t\int_D h^2\, dx\, ds.$$

Thus, one deduces that

$$E\int_D u_M^2(t)\, dx + 2E\int_Q |\nabla u_M|^2\, dx\, ds \leq C(h). \tag{1}$$

Moreover, for any v in $V \setminus \{0\}$,

$$\frac{\left|\left\langle \frac{\partial}{\partial t}\left[u_M - \int_0^t h\, dw(s)\right], v\right\rangle_{V',V}\right|}{\|v\|_V} \leq \|\nabla u_M\|_{L^2(D)} + \|\vec{B}\|_\infty c_P \|f_M(u_M)\|_{L^2(D)}.$$

Since

$$\begin{aligned}|f_M(u_M)|^2 &= \left|\int_{\mathbb{R}} f(T_M(u_M) - y)\rho_M(y)\, dy\right|^2 \leq \int_{\mathbb{R}} |f(T_M(u_M) - y)|^2 \rho_M(y)\, dy \\ &\leq c(f)\int_{\mathbb{R}} |T_M(u_M) - y|\rho_M(y)\, dy \leq c(f)(|T_M(u_M)| + 1) \leq c_1 u_M^2 + c_2,\end{aligned}$$

one deduces that

$$\frac{\left|\left\langle \frac{\partial}{\partial t}\left[u_M - \int_0^t h\, dw(s)\right], v\right\rangle_{V',V}\right|^2}{\|v\|_V^2} \leq 2\|\nabla u_M\|_{L^2(D)}^2 + 2\|\vec{B}\|_\infty^2 c_P^2 \left[c_1 \|u_M\|_{L^2(D)}^2 + c_2 \operatorname{meas}(D)\right]$$

and that

$$E\int_0^T \left\|\frac{\partial}{\partial t}\left[u_M - \int_0^t h\, dw(s)\right]\right\|_{V'}^2 dt \leq C(h). \tag{2}$$

Thus, one is able to assert the

Lemma 2. *Uniformly with respect to M and for any $t \in [0,T]$, the sequences $u_M(t)$, u_M and $\frac{\partial}{\partial t}\left[u_M - \int_0^t h\, dw(s)\right]$ are bounded respectively in $L^2(\Omega \times D)$, $L^2(\Omega \times]0,T[, V)$ and $L^2(\Omega \times]0,T[, V')$.*

Following J. U. Kim [6], denote, for any t, by

$$\Theta(u_M, t) = \sup_{s\in[0,t]} \|u_M(s)\|_{L^2(D)}^2 + \|u_M\|_{L^2(0,t;V)}^2 + \left\|\frac{\partial}{\partial t}\left[u_M - \int_0^{\cdot} h\, dw(s)\right]\right\|_{L^2(0,t,V')}^2,$$

$$\widetilde{\Omega}(t) = \bigcup_{L\geq 2}\bigcup_{M\geq 1}\bigcup_{k\geq m}\{\Theta(u_M,t) \leq L\} \qquad \text{and} \qquad \widetilde{\Omega} = \widetilde{\Omega}(T).$$

Thanks to the above lemma, one deduces that $P(\widetilde{\Omega}) = 1$. Then, for P-a.s. ω, a positive constant $L(\omega)$ and a sub-sequence denoted by u_{M_ω} exist such that $\{\Theta(u_{M_\omega}, T) \leq L(\omega)\}$. Therefore, there exist $u = u(\omega)$ in $L^2(0,T;V)$ with moreover $\frac{\partial}{\partial t}[u - \int_0^t h\, dw(s)]$ in $L^2(0,T,V')$ and a sub-sequence denoted by (u_k) such that u_k converges weakly to u in $L^2(0,T;V)$ and that $\frac{\partial}{\partial t}[u_k - \int_0^t h\, dw(s)]$ converges weakly to $\frac{\partial}{\partial t}[u - \int_0^t h\, dw(s)]$ in $L^2(0,T,V')$.

Moreover, thanks to Corollary 4, (u_k) converges in $L^2(0,T;L^2(D))$ and a.e. in Q since sub-sequences are considered, and in $C([0,T];H^{-1}(D))$. In particular, $u_0 = u_k(0)$ converges to $u(0)$ in V'.

Since $f_k^2(u_k) \leq c_1 u_k^2 + c_2$, it can be assumed, up to a sub-sequence denoted in the same way, that $f_k(u_k)$ converges weakly to some f_u in $L^2(Q)$. Note that, by construction, $f_k(u_k)$ converges a.e. in Q to $f(u)$. Then, it converges weakly to $f(u)$ in $L^2(Q)$ (Cf. J.-L. Lions [7, lemma 1.3, p.12]).

It follows that, for any v in V and t a.e. in $]0,T[$,

$$\left\langle \frac{\partial}{\partial t}[u - \int_0^t h\,dw], v \right\rangle_{V',V} + \int_D \nabla u.\nabla v + f(u)\vec{B}.\nabla v\,dx = 0.$$

If one denotes by $\hat{u}$ an other solution, for any v in V, one gets that

$$\left\langle \frac{\partial}{\partial t}[u - \hat{u}], v \right\rangle_{V',V} + \int_D \nabla[u-\hat{u}].\nabla v + [f(u) - f(\hat{u})]\vec{B}.\nabla v\,dx = 0.$$

For a given $\mu > 0$, set $v = p_\mu[u - \hat{u}]$ where $p_\mu(x) = 0$ if $x < \mu/e$, 1 if $x > \mu$ and $\ln(ex/\mu)$ else. Note that p_μ is a Lipschitz-continuous function and denote by $P_\mu = \int_0^x p_\mu(s)\,ds$. Then,

$$0 = \frac{d}{dt}\int_D P_\mu\,[u-\hat{u}]\,dx + \int_D p'_\mu\,[u-\hat{u}]\,|\nabla[u-\hat{u}]|^2 + [f(u)-f(\hat{u})]p'_\mu[u-\hat{u}]\vec{B}.\nabla[u-\hat{u}]\,dx.$$

And by construction,

$$\frac{d}{dt}\int_D P_\mu\,[u-\hat{u}]\,dx + \frac{1}{2}\int_D p'_\mu[u-\hat{u}]|\nabla[u-\hat{u}]|^2 \leq C\int_{\{\mu/e<u-\hat{u}<\mu\}} |u-\hat{u}|p'_\mu\,[u-\hat{u}]\,dx.$$

Thus,

$$\int_D P_\mu[u-\hat{u}]\,dx \leq C\,\mathrm{meas}\,(\{\mu/e < u - \hat{u} < \mu\}) + \int_D P_\mu[0]\,dx.$$

Passing to the limits leads to $u \leq \hat{u}$.

Since one is able to prove in the same way that $u \geq \hat{u}$, the solution to the above problem is unique and all the sequence (u_{M_ω}) converges.

Now, one needs to prove that u, generated by sub-sequences depending on ω, is adapted to the filtration and belongs to the stated spaces. In order to prove this, we propose to follow J. U. Kim's [6] arguments. Consider a closed ball B in $H^{-1}(D)$ and, for any positive integer n, $B_n = \bigcup_{v\in B}\bar{B}_{H^{-1}(D)}(v, 1/n)$. For a fixed t^*, note that

$$\widetilde{\Omega}\cap\{u(t^*)\in B\} = \widetilde{\Omega}\cap\Big[\bigcup_{L>0}\bigcap_{n>0}\bigcap_{k>0}\bigcup_{M\geq k}\{u_M(t^*)\in B_n\}\cap\{\Theta(u_M,t^*)\leq L\}\Big]. \tag{3}$$

Indeed, for any $\omega \in \widetilde{\Omega}\cap\{u(t^*)\in B\}$, (u_{M_ω}) satisfies $\Theta(u_{M_\omega}, t^*) \leq \Theta(u_{M_\omega}, T) \leq L(\omega)$. Moreover, since u_{M_ω} converges in $C([0,T], H^{-1}(D))$, ω belongs to the right hand side set.

Conversely, if ω belongs to the right hand side set, there exists $\bar{L}(\omega) > 0$ such that for any positive integer n, one is able to construct a sub-sequence $u_{\bar{M}_{\omega,n}}$ with $u_{\bar{M}_{\omega,n}}(t^*) \in B_n$ and $\Theta(u_{\bar{M}_{\omega,n}}, t^*) \leq \bar{L}(\omega)$.

Since, what has been done with u_{M_ω} in $]0,T[$ can be done again with $u_{\bar{M}_{\omega,n}}$ in $]0,t^*[$, the uniqueness result proved above yields the convergence of $u_{\bar{M}_{\omega,n}}$ to u. Therefore, $u(t^*) \in B_n$ for any n, and the result holds.

Thanks to the regularity of u_M, the left hand side of (3) is $\mathcal{F}_{t^*}$-measurable and $\{u(t^*)\in B\}$ is in $\mathcal{F}_{t^*}$. More generally, for any t in $[0,T]$, $\{u(t^*)\in F\} \in \mathcal{F}_t$ for any Borel subset F of $H^{-1}(D)$. Since u belongs to $C([0,T], H^{-1}(D))$, $\{(t,\omega),\ 0\leq t\leq t^*,\ u(t,\omega)\in F\} \in ([0,T])\times\mathcal{F}_{t^*}$ for each $F \in \mathcal{B}(H^{-1}(D))$ and any $t \in\,]0,T]$.

Since u belongs to $C_s([0,T], L^2(D))$, $u(t) \in L^2(D)$ for any t and thanks to lemmata 5 and 7 in the annexes, the same result of measurability holds for any $F \in \mathcal{B}(L^2(D))$; and u is progressively measurable as a $L^2(D)$ valued process.

Note that a similar argument could be used in $L^2(D)$ with the weak topology since u belongs to $C_s([0,T], L^2(D))$ with values in a fixed bounded subset of $L^2(D)$ and thanks to lemma 6 in the annexes.

Then, thanks to (1), (2) and the lemma of Fatou, on gets that

$$E\int_D u^2(t)\,dx + 2E\int_Q |\nabla u|^2\,dxds + E\int_0^T \left\|\frac{\partial}{\partial t}\left[u - \int_0^t h\,dw(s)\right]\right\|_{V'}^2 dt \leq C(h),$$

and a solution exists in the sense of the definition 1.

For the uniqueness of the solution, one has just to use the same method than the one given above, based on the approximation of the sgn^+ function by p_μ.

§4. Annexes

In this section we propose some classical tools used in this paper.

First, let us remind the theorem on Aubin-Simon:

Theorem 3 ([7, Th. 5.1, Th. 12.1 and (12.10)]). *Let us consider $1 < p \leq +\infty$, $1 \leq q \leq +\infty$, B_0, B_1 and B_2 three B-spaces such that the embedding of B_0 in B_1 is compact and the embedding of B_1 in B_2 is continuous. If (u_n) is a bounded sequence in $L^q(0,T;B_0)$ such that (du_n/dt) (the derivation is taken in the sense of vectorial distributions) is a bounded sequence in $L^p(0,T;B_2)$, then there exists a subsequence that converges in $L^q(0,T;B_1)$ and in $C([0,T];B_2)$.*

The following corollary is the main tool of compactness used in the paper:

Corollary 4. *Let (u_n) be a bounded sequence in $L^2(0,T;H^1_0(D)) \cap L^\infty(0;T;L^2(D))$ and $H \in C([0,T];L^2(D))$. If $(d(u_n - H)/dt)$ (the derivation is taken in the sense of vectorial distributions) is a bounded sequence in $L^2(0,T;H^{-1}(D))$ then there exists a subsequence (u_{n_k}) that converges in $L^2(0,T;L^2(D))$ and in $C([0,T];H^{-1}(D))$.*

Moreover, the limit is $C_s([0,T];L^2(D))$[1].

Proof. Since the embedding of $L^2(D)$ in $H^{-1}(D)$ is compact and since $(u_n - H)$ is bounded in $L^\infty(0;T;L^2(D))$, thanks to Aubin-Simon's theorem, there exists a subsequence $(u_{n_k} - H)$ that converges in $C([0,T];H^{-1}(D))$. In particular, (u_{n_k}) converges in $C([0,T];H^{-1}(D))$ too.

Thanks to the lemma of Lions ([7, Lemma 5.1]), for any positive ϵ, there exists a positive d_ϵ such that, for any n, p,

$$||u_{n+p} - u_n||_{L^2(0,T;L^2(D))} \leq \epsilon||u_{n+p} - u_n||_{L^2(0,T;H^1_0(D))} + d_\epsilon||u_{n+p} - u_n||_{L^2(0,T;H^{-1}(D))}.$$

Thus, since $||u_n||_{L^2(0,T;H^1_0(D))}$ is bounded, one gets that for any positive ϵ, there exists a positive d_ϵ such that, for any n, p,

$$||u_{n+p} - u_n||_{L^2(0,T;L^2(D))} \leq \frac{\epsilon}{2} + d_\epsilon||u_{n+p} - u_n||_{L^2(0,T;H^{-1}(D))}.$$

[1] $u \in C_s([0,T];X)$ if, for any $x^* \in X'$, $t \mapsto \langle x^*, u(t)\rangle$ is continuous.

Since (u_{n_k}) is a Cauchy sequence in $L^2(0,T;H^{-1}(D))$, a positive integer N exists such that $d_\epsilon\|u_{n_{k+p}} - u_{n_k}\|_{L^2(0,T;H^{-1}(D))} \leq \epsilon/2$ as soon as $n_k \geq N$. Then, (u_{n_k}) is a Cauchy sequence in $L^2(0,T;L^2(D))$ and it converges.

Obviously, the limit belongs to $L^\infty(0;T;L^2(D)) \cap C([0,T];H^{-1}(D))$, thus it belongs to $C_s([0,T];L^2(D))$ thanks to [8, Lemma 8.1, p.297]. □

Let us give now some lemmata concerning the measurability of vector-valued functions.

Lemma 5. *Assume that u is a function with values in $L^2(D)$ and $H^{-1}(D)$-measurable, then it is $L^2(D)$-measurable.*

Proof. Our argument is based on the theorem of Pettis in separable B-spaces [12].

If u is $H^{-1}(D)$-measurable, then it is weakly measurable. Thus, for any v in $H_0^1(D)$, $\langle u,v\rangle_{H^{-1},H_0^1}$ is a scalar measurable function. As u is a function with values in $L^2(D)$, $\langle u,v\rangle_{H^{-1},H_0^1} = \int_D uv\,dx$ and it is a scalar measurable function. Note that for any $v \in L^2(D)$, there exists $(v_n) \subset H_0^1(D)$ that converges toward v in $L^2(D)$. Thus, $\int_D uv_n\,dx$ converges a.e. toward $\int_D uv\,dx$ and it is a scalar measurable function. Therefore, u is weakly $L^2(D)$-measurable, thus $L^2(D)$-measurable. □

I would like to present an generalisation proposed by L. Thibault (personal communication) and based on the two following lemmata:

Lemma 6. *Let Y be a separable B-space. Then, the Borel sigma-algebra $\mathcal{B}(Y)$ when Y is endowed with the strong topology is the same than the Borel sigma-algebra $\mathcal{B}_w(Y)$ when Y is endowed with the weak topology. Moreover, $\mathcal{B}(Y)$ is the sigma-algebra generated by the closed balls of Y.*

Proof. First $\mathcal{B}_w(Y) \subset \mathcal{B}(Y)$ is obvious since the same inclusion holds for the topologies.

On the other hand, any closed ball $\bar{B}(a,r)$ in Y is $\sigma(Y,Y^*)$-closed since it is convex. In particular, $\bar{B}(a,r) \in \mathcal{B}_w(Y)$.

As any open ball is a countable reunion of closed ones, any open ball belongs to $\mathcal{B}_w(Y)$. Now, thanks to the separability of Y, any open subset of Y is a countable reunion of open balls. Then, any open subset of Y is an element of $\mathcal{B}_w(Y)$ and $\mathcal{B}(Y) \subset \mathcal{B}_w(Y)$. Note that this prove that $\mathcal{B}(Y)$ is generated by the closed balls too. □

Lemma 7. *Assume that $X \subset Y$ are separable B-spaces with X reflexive. If the embedding i of X in Y is continuous, then $\mathcal{B}(X) \subset \mathcal{B}(Y)$, where $\mathcal{B}(X)$ (resp. Y) denotes the Borel σ-algebra of X (resp. Y).*

Proof. Consider A a closed ball in X. Since X is assumed to be reflexive, A is $\sigma(X,X^*)$ compact. Moreover, the application i is $\sigma(X,X^*)$-$\sigma(Y,Y^*)$ continuous and then A is a compact set of Y for the topology $\sigma(Y,Y^*)$. Therefore, A is weakly closed in Y, it is closed and it belongs to $B(Y)$. The conclusion comes from the remark that $B(X)$ is generated by the closed balls of X. □

References

[1] Adams, R. A. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[2] Da Prato, G., Debussche, A., and Temam, R. Stochastic Burgers' equation. *NoDEA, Nonlinear Differ. Equ. Appl. 1*, 4 (1994), 389–402.

[3] Da Prato, G., and Zabczyk, J. *Stochastic equations in infinite dimensions*, vol. 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

[4] Grecksch, W., and Tudor, C. *Stochastic evolution equations*, vol. 85 of *Mathematical Research*. Akademie-Verlag, Berlin, 1995. A Hilbert space approach.

[5] Gyöngy, I., and Nualart, D. On the stochastic Burgers' equation in the real line. *Ann. Probab. 27*, 2 (1999), 782–802.

[6] Kim, J. U. On the stochastic porous medium equation. *J. Differential Equations 220*, 1 (2006), 163–194.

[7] Lions, J.-L. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod, 1969.

[8] Lions, J.-L., and Magenes, E. *Problèmes aux limites non homogènes et applications. Vol. 1, 2.* Paris: Dunod 1: XIX, 372 p.; 2: XV, 251 p., 1968.

[9] Röckner, M., and Sobol, Z. l^1-theory for the Kolmogorov operators of stochastic generalized Burgers equations. In *Quantum Information and Complexity* (2004), K. S. T. Hida and S. Si, Eds., vol. Proceedings of the 2003 Meijo Winter School and Conference, World Scientific, pp. 87–105.

[10] Vallet, G. Stochastic perturbation of nonlinear degenerete parabolic problems. *Differential and Integral Equations* (To appear).

[11] Vallet, G., and Wittbold, P. On a stochastic first order hyperbolic equation in a bounded domain. *Infinite Dimensional Analysis, Quantum Probability and Related Topics (IDAQP)* (To appear).

[12] Yosida, K. *Functional analysis*, fourth ed. Springer-Verlag, New York, 1974. Die Grundlehren der mathematischen Wissenschaften, Band 123.

Guy Vallet
LMA UMR-CNRS 5142
IPRA BP 1155
64013 Pau Cedex (FRANCE)
guy.vallet@univ-pau.fr

Statistics

Monografías Matemáticas *García de Galdeano* **35**, 255–261 (2010)

ON STOCHASTIC ORDERS AND AGING PROPERTIES IN GENERALIZED MIXTURES

F. G. Badía and M. D. Berrade

Abstract. In this work we deal with generalized discrete mixtures where the weights corresponding to some of the distributions in the mixture can be negative. We study closure conditions for some aging properties as well as stochastic ordering in mixtures of this type.

Keywords: Aging properties, mixtures, stochastic orders.

AMS classification: 60K10, 60E15, 62N05.

§1. Introduction and background

The study of both aging properties and stochastic orders constitutes a traditional research area in reliability, as it provides the suitable way to assess how systems wear-out thus resulting in increasingly frequent malfunctions and decreased reliability.

In addition many systems are likely to work under conditions that do not remain constant over time but are prone to experience random changes. If so, changing conditions are responsible for the uncertainty in the parameters of the distributions representing the lifetimes of systems that turn out to be better modeled by a mixture of distributions.

In this article we consider generalized mixtures whose reliability function, $\overline{F}^\star(x)$ is given as follows

$$\overline{F}^\star(x) = \sum_{i=1}^{N} p_i \overline{F}_i(x),$$

where $\sum_{i=1}^{N} p_i = 1$. $\overline{F}_i$ represents, for each $i = 1, 2, \ldots, N$, the reliability function corresponding to the distribution, X_i, in the mixture. $X^\star$ denotes the random variable representing the mixture. The term generalized mixtures refers to the fact that some of the weights, p_i, can be negative. [2] deals with negative weights in mixtures involving Weibull and inverse Weibull distributions.

Generalized mixtures can be useful when approximating the reliability behaviour of consecutive k-out-of-n systems, that is, the type of configuration where the system works only if at least k of the n components work.

Consider the time to failure of a system, X with $f(x)$, $F(x)$, and $\bar{F}(x)$ being, respectively, the corresponding density, distribution and reliability functions. When assessing its stochastic aging, the hazard rate, $r(x)$, and the mean residual life $m(x)$, defined below, constitute the traditional reliability measures.

$$
\begin{aligned}
r(x) &= \frac{f(x)}{\overline{F}(x)},\\
\mu(x) &= E[X - x \mid X > x] = \frac{\int_x^\infty \overline{F}(u)\,du}{\overline{F}(x)}.
\end{aligned}
$$

The failure rate is an indicator of the proneness to failure whereas the mean residual life represents the expected remaining life of a system that has survived up to x.

Lately new measures have emerged, namely, the reversed hazard rate $q(x)$ and the mean inactivity time $m(x)$ given as follows

$$
\begin{aligned}
q(x) &= \frac{f(x)}{F(x)},\\
m(x) &= E[x - X \mid X \leq x] = \frac{\int_0^x F(u)\,du}{F(x)}.
\end{aligned}
$$

Both are useful, for instance, in medical studies when it's known for example that someone is infected with a virus, say, VIH, and physicians' interest focuses on determining the infection time or the time elapsed since then.

The monotonic behavior of the foregoing functions constitutes a crucial issue in maintenance scheduling. Thus, the study of aging characteristics in mixtures and the conditions under which the monotonicity of the distributions in the mixture is preserved is a key research area. The preservation under mixtures of decreasing failure rate (DFR) distributions [6] turns out to be the seminal result. It's well known that a similar property doesn't hold in general for distributions with increasing failure rate (IFR), decreasing mean residual life (DMRL), decreasing reversed hazard rate (DRHR) and increasing mean inactivity time (IMIT). Therefore research focuses on conditions under which closure results remain valid for the foregoing nonparametric classes. Following this line [3], [5], and [4] provide remarkable results

The comparison of distribution and thus the stochastic order approach, emerges to determine which of several distributions appears to be preferable to the rest. Hence the stochastic ordering of mixtures is an interesting area to deal with. The works due to [7], [3], [5], and [4] also deal with this issue.

Section 2 and Section 3 contain, respectively, the results concerning the aging properties and the stochastic ordering in mixtures. Section 4 contains some concluding remarks.

§2. Preservation of aging properties under generalized mixtures

First we present some known results concerning the preservation of aging properties that motivate the study carried out in this article.

Proposition 1 (Cf. [4])**.** *Consider that $p_1 > 0$ and X_1 belongs to the increasing hazard rate class (IFR), whereas $p_1 \leq 0$ and X_i is decreasing hazard rate (DFR) for $i = 2, 3, \ldots, N$. Then, the mixture $X^\star$ is IFR.*

Proposition 2 (Cf. [5])**.** *Consider that $p_1 > 0$ and X_1 is decreasing mean residual life (DMRL), whereas $p_1 \leq 0$ and X_i is increasing mean residual life (IMRL) for $i = 2, 3, \ldots, N$. Then, the mixture $X^\star$ belongs to the DMRL class.*

Proposition 3 (Cf. [3])**.** *Consider that $p_1 > 0$ and X_1 has the increasing likelihood ratio property (ILR or logconcave), whereas $p_1 \leq 0$ and X_i has the decreasing likelihood ratio (DRL or logconvex) for $i = 2, 3, \ldots, N$. Then, the mixture, $X^\star$ is also IRL.*

Next, we introduce new results related to the reversed hazard rate and the mean inactivity time.

Proposition 4. *Consider that $p_1 > 0$ and X_1 belongs to the decreasing reversed hazard rate class (DRHR), whereas $p_1 \leq 0$ and X_i is increasing reversed hazard rate (IRHR) for $i = 2, 3, \ldots, N$. Then, the mixture, $X^\star$ is DRHR.*

Proof. The reversed failure rate of the mixture denoted by $q^\star$ is written as follows

$$q^\star(x) = \sum_{i=1}^{N} \overline{w}_i(x) q_i(x),$$

where

$$\overline{w}_i(x) = \frac{p_i F_i(x)}{\sum_{i=1}^{N} p_i F_i(x)} = \frac{p_i F_i(x)}{F^\star(x)}, \quad i = 1, 2, \ldots, N,$$

alternatively

$$q^\star(x) = \overline{w}_1(x) q_1(x) + (1 - \overline{w}_1(x)) q_0^\star(x),$$

with

$$q_0^\star(x) = \frac{\sum_{i=2}^{N} \frac{-p_i}{1-p_1} i f_i(x)}{\sum_{i=2}^{N} \frac{-p_i}{1-p_1} F_i(x)}.$$

In addition $q_0^\star$ is the reversed failure rate of a mixture of distributions, therefore it is an increasing function [1]. By taking derivatives we get

$$\frac{\overline{w}_1(x)}{dx} = \overline{w}_1(x)(1 - \overline{w}_1(x))(q_1(x) - q_0^\star(x)).$$

Hence

$$\frac{dq^\star(x)}{dx} = \overline{w}_1(x)(1 - \overline{w}_1(x))(q_1(x) - q_0^\star(x))^2 + \overline{w}_1(x)\frac{dq_1(x)}{dx} + (1 - \overline{w}_1(x))\frac{dq_0^\star(x)}{dx}.$$

It follows from the assumptions that $\overline{w_1}(x) \geq 0$, $1 - \overline{w_1}(x) \leq 0$, $\frac{dq_1(x)}{dx} \leq 0$ and $\frac{dq_0^\star(x)}{dx} \geq 0$. Thus, $\frac{dq^\star(x)}{dx} \leq 0$ and $X^\star$ is DRHR. □

Proposition 5. *Consider that $p_1 > 0$ and X_1 belongs to the increasing mean inactivity time (IMIT), whereas $p_1 \leq 0$ and X_i is decreasing mean inactivity time (DIMIT) for $i = 2, 3, \ldots, N$. Then, the mixture $X^\star$ is IMIT.*

Proof. The expression below denoted by $v^\star(x)$ corresponds to the mean inactivity time of the mixture

$$v^\star(x) = \frac{\sum_{i=1}^N p_i L_i(x) v_i(x)}{\sum_{i=1}^N p_i F_i(x)},$$

where

$$L_i(x) = \int_{-\infty}^x F_i(u)du.$$

The inverse value of the mean inactivity time of the mixture is given by

$$\frac{1}{v^\star(x)} = \alpha(x)\frac{1}{v_1(x)} + (1-\alpha(x))\frac{1}{v_0^\star(x)},$$

with

$$\alpha(x) = \frac{p_1 L_1(x)}{\sum_{i=1}^N p_i L_i(x)}$$

and

$$v_0^\star(x) = \frac{\sum_{i=2}^N \frac{-p_i}{1-p_1} i L_i(x)}{\sum_{i=2}^N \frac{-p_i}{1-p_1} F_i(x)}.$$

It can be stated that $v_0^\star(x)$ is a decreasing function because it represents the mean inactivity time of a mixture of distributions [1]. Moreover,

$$\frac{d\alpha(x)}{dx} = \alpha(x)(1-\alpha(x))\left(\frac{1}{v_1(x)} - \frac{1}{v_0^\star(x)}\right).$$

Therefore

$$\frac{d\frac{1}{v^\star(x)}}{dx} = \alpha(x)(1-\alpha(x))\left(\frac{1}{v_1(x)} - \frac{1}{v_0^\star(x)}\right)^2 + \alpha(x)\frac{d\frac{1}{v_1(x)}}{dx} + (1-\alpha(x))\frac{d\frac{1}{v_0^\star(x)}}{dx}.$$

From assumptions in Proposition 5 we conclude that $\alpha(x) \geq 0$, $1-\alpha(x) \leq 0$, $\frac{d(1/v_1(x))}{dx} \geq 0$ and $\frac{d(1/v_0^\star(x))}{dx} \geq 0$. Hence, $\frac{d(1/v^\star(x))}{dx} \leq 0$ and $v^\star(x)$ is an increasing function. Then, the result holds. □

§3. Stochastic order in generalized mixtures of two random variables

When comparing two distributions we can consider different stochastic orders depending of our knowledge of the underlying distributions. In what follows we present the basic definitions concerning several stochastic orders.

- Usual stochastic order: $X \leq_{st} Y$

$$\overline{F}_X(t) \leq \overline{F}_Y(t) \text{ for all } t \in (-\infty, \infty).$$

- Reversed hazard rate order: $X \leq_{RHR} Y$

$$q_X(t) \leq q_Y(t) \text{ for all } t \in (-\infty, \infty).$$

- Mean inactivity time order: $X \leq_{MIT} Y$

$$m_X(t) \geq m_Y(t) \text{ for all } t \in (-\infty, \infty).$$

- Convex order: $X \leq_{convex} Y$

$$E[\phi(X)] \leq E[\phi(Y)] \text{ for all convex functions } \phi : \mathbb{R} \to \mathbb{R}.$$

- Laplace transform order: $X \leq_{\mathrm{L}_t} Y$

$$E[e^{-tX}] \geq E[e^{-tY}] \text{ for all } t > 0.$$

The results of this section are based on that from [7]. We deal with mixtures of two distributions as follows

$$\overline{F}^{\star}_p(x) = p\overline{F}_1(x) + (1-p)\overline{F}_0(x),$$

analyzing the monotonic behavior of the mixture $X^{\star}_p$ under several stochastic orders provided that the corresponding variables X_0 and X_1 follow the same stochastic order.

We present two former results concerning the mean residual life and the likelihood ratio orders respectively.

Proposition 6 (Cf. [5])**.** *If $X_0 \leq_{MRL} X_1$ and $p \leq p'$, then $X^{\star}_p \leq_{MRL} X^{\star}_{p'}$.*

Proposition 7 (Cf. [3])**.** *If $X_0 \leq_{LR} X_1$ and $p \leq p'$, then $X^{\star}_p \leq_{LR} X^{\star}_{p'}$.*

Next we get additional results related to other stochastic orders.

Proposition 8. *If $X_0 \leq_{st} X_1$ and $p \leq p'$, then $X^{\star}_p \leq_{st} X^{\star}_{p'}$.*

Proof.

$$\frac{d\overline{F}^{\star}_p(x)}{dx} = \overline{F}_1(x) - \overline{F}_0(x) \geq 0.$$

Then, $X^{\star}_p \leq_{\mathrm{st}} X^{\star}_{p'}$ provided that $p \leq p'$. □

Proposition 9. *If $X_0 \leq_{RHR} X_1$ and $p \leq p'$, then $X^{\star}_p \leq_{RHR} X^{\star}_{p'}$*

Proof. The reversed hazard rate of the mixture is given next

$$q^{\star}_p(x) = \overline{w}_1(p,x)q_1(x)q_1(x) + (1 - \overline{w}_p(p,x))q_1(x),$$

where

$$\overline{w}_1(p,x)q_1(x) = \frac{pF_1(x)}{pF_1(x) + (1-p)F_0(x)} = \frac{pF_1(x)}{F^{\star}_p(x)}.$$

In addition

$$\frac{d\overline{w}_1(p,x)}{dp} = \frac{F_1(x)F_0(x)}{F^{\star}_p(x)^2}. \tag{1}$$

Hence

$$\frac{dq^{\star}_p(x)}{dp} = \frac{F_1(x)F_0(x)}{F^{\star}_p(x)^2}(q_1(x) - q_0(x)) \geq 0,$$

and the result holds. □

Proposition 10. *If* $X_0 \le_{MIT} X_1$ *and* $p \le p'$*, then* $X_p^\star \le_{MIT} X_{p'}^\star$.

Proof. The following expression represents the mean inactivity time of the mixture:

$$v_p^\star(x) = \overline{w}_1(p,x)v_1(x) + (1 - \overline{w}_p(p,x))v_0(x),$$

where

$$\overline{w}_1(p,x)q_1(x) = \frac{pF_1(x)}{pF_1(x) + (1-p)F_0(x)} = \frac{pF_1(x)}{F_p^\star(x)}$$

and $\overline{w}_1(p,x)$ as in (1). After taking derivatives we obtain

$$\frac{dv_p^\star(x)}{dp} = \frac{F_1(x)F_0(x)}{F_p^\star(x)^2}(v_1(x) - v_0(x)) \le 0.$$

Therefore, if $p \le p'$, then $X_p^\star \le_{MIT} X_{p'}^\star$. □

Proposition 11. *If* $X_0 \le_{convex} X_1$*, and* $p \le p'$*, then* $X_p^\star \le_{convex} X_{p'}^\star$.

Proof. Let f be a convex function, then

$$E(f(X_p^\star)) = pE(f(X_1)) + (1-p)E(f(X_0)).$$

Hence

$$\frac{dE(f(X_p^\star)}{dp} = E(f(X_1)) - E(f(X_0)) \ge 0.$$

Thus, if $p \le p'$, then $X_p^\star$ is less than $X_{p'}^\star$ under the convex stochastic order. □

Proposition 12. *If* $X_0 \le_{L_t} X_1$ *and* $p \le p'$*, then* $X_p^\star \le_{L_t} X_{p'}^\star$.

Proof. For $t \ge 0$ it is seen that

$$E(e^{-tX_p^\star}) = pE(e^{-tX_1}) + (1-p)E(e^{-tX_0}).$$

Therefore

$$\frac{dE(e^{-tX_p^\star})}{dp} = E(e^{-tX_1}) - E(e^{-tX_0}) \le 0$$

and the result holds. □

§4. Concluding remarks

The foregoing results aim at providing additional insight on aging phenomena in the context of reliability. According to this objective it can be noticed that the following inequality holds for all $p \in (0,1)$:

$$\begin{array}{ccccc} X_0 & \le_{st} & X_p^\star & \le_{st} & X_1, \\ X_0 & \le_{RHR} & X_p^\star & \le_{RHR} & X_1, \\ X_0 & \le_{MIT} & X_p^\star & \le_{MIT} & X_1, \\ X_0 & \le_{convex} & X_p^\star & \le_{convex} & X_1. \end{array}$$

That is, the mixture presents a “better aging” than the worst of the two components providing an spurious image of improvement. The results complete previous works due to [7], [3], [5], and [4] related to other aging classes and stochastic orders.

Acknowledgements

This work has been supported by research project MTM2007-63683 from the formerly Spanish Ministry of Education and Science.

References

[1] Badía, F., and Berrade, M. On the reversed hazard rate and mean inactivity time of mixtures. *Advances in Mathematical Modeling for Reliability, Tim Bedford et al. (Editors)* (2008), 103–110.

[2] Jiang, R., Zuo, M., and Li, H. Weibull and inverse weibull mixture models allowing negative weights. *Reliability Engineering and System Safety 66* (2008), 227–234.

[3] Navarro, J. Likelihood ratio ordering of order statistics, mixtures and systems. *Journal of Statistical Planning and Inference 138* (2008), 1242–1257.

[4] Navarro, J., Gullamón, A., and Ruiz, M. Bathtub-shaped hazard rate models obtained from generalized mixtures. *Applied Stochastic models in Business and Industry* (forthcoming).

[5] Navarro, J., and Hernández, P. Mean residual life functions of finite mixtures, order statistics and coherent systems. *Metrika 67* (2008), 277–298.

[6] Proschan, F. Theorical explanation of observed decreasing failure rate. *Technometrics 5* (1963), 375–383.

[7] Shaked, M., and Shanthikumar, J. *Stochastic Orders*. Springer Series in Statistics. Springer, 2007.

F.G. Badía and M.D. Berrade
Departamento de Métodos Estadísticos. Centro Politécnico Superior
María de Luna 3, 50018 Zaragoza, Spain
gbadia@unizar.es and berrade@unizar.es

Monografías Matemáticas *García de Galdeano* **35**, 263–270 (2010)

STATISTICAL INFERENCE IN THE STOCHASTIC GAMMA DIFFUSION PROCESS WITH EXTERNAL INFORMATION

R. Gutiérrez, A. Nafidi and R. Gutiérrez Sánchez

Abstract. In this work, we consider a new extension of the one-dimensional stochastic gamma diffusion process (cf. [11]) by introducing external time functions as exogenous factors, in the same way as exogenous factors have been introduced into lognormal process (cf. [14]), the Gompertz process (cf. [10]) and the Vasicek process (cf. [12]), among others. Firstly, we determine the probabilistic characteristics of the process as its analytical expression, the transition probability density function and the trend functions. Secondly, we study the statistical inference in this process: the parameters present in the model are studied by using the maximum likelihood estimation method on the basis of the discrete sampling, thus obtaining the expression of the likelihood estimators and their properties (statistical distribution, sufficiency and completeness), together with the confidence intervals of the parameters.

Keywords: Stochastic gamma diffusion process, exogenous factors, statistical inference in diffusion process.

AMS classification: 60J60, 62M05.

§1. Introduction

Stochastic processes are used in fields as diverse as physics, biology, economics and finance to model and analyze dynamic systems. One particular class of stochastic processes which has attracted considerable attention is that of diffusion processes. And one of the questions that has provoked greatest theoretical and practical interest concerning diffusions, and which has been the object of many studies in recent years, is the problem of establishing the corresponding statistical inference, a question that may be approached by the use of either continuous or discrete sampling. This inference, and in particular the estimation of parameters, has been studied in the general case by various authors, such as Bibby and Sorensen [3], Prakasa Rao [17], Ait-Sahalia [1] and Egorov et al. [4], among many others. And in the case of particular diffusions, it has been considered, for example, by Giovanis et al. [7] in the logistic case, Gutiérrez et al. [8] in the Gompertz case, Gutiérrez et al. [9] in the Rayleigh case and Forman et al. [6] in the case of Pearson diffusions, among other important diffusions.

Due to the need to use stochastic diffusions to accurately model real phenomena that are becoming more and more complex, various extensions of these processes have been considered, such as non-homogeneous extensions and, in particular non-homogeneous extensions with exogenous factors, which have been defined, studied and applied, for example, in the case of the lognormal process by Gutiérrez et al. [14], in the case of the Gompertz process by Gutiérrez et al. [10] and Ferrante [5], in the case of the Vasicek process by Gutiérrezet al. [12], and by Picchini et al. [16] in the case of the Brennan-Schwartz diffusion process, among others.

In the present study, based on the methodology established for the consideration of exogenous factors affecting drift , described in [14], [10] and [12], we define a new diffusion process with external information, modelled by time deterministic functions (exogenous variables) that affect the drift of the Gamma diffusion process, as studied in [13] and [11]. We go on to examine a new Gamma diffusion process with exogenous factors, investigating its main probabilistic characteristics and the corresponding statistical inference.

The remainder of the paper is organised as follows. In the next section, we first define the model and consider its probabilistic characterisations, such as the explicit expression, the probability transition density function (ptdf) and the moments (in particular the trend functions). In the third section, we study the statistical inference in the proposed process using discrete sampling, obtaining the likelihood estimators, their statistical properties and the confidence parameter intervals.

§2. The model and its basic probabilistic characteristics

2.1. The proposed model and their analytical expression

The model considered is the one dimensional process $\{x(t),\ t \in [t_1, T],\ t_1 > 0\}$ with values in $(0, \infty)$ and governed by the following Ito's stochastic differential equation (SDE)

$$dx(t) = a(t, x(t))dt + b^{1/2}(t, x(t))dw(t), \quad \mathrm{P}[x(t_1) = x_{t_1}] = 1, \tag{1}$$

where $a(t, x)$ and $b(t, x)$ are given by

$$a(t, x) = \left(\frac{\alpha}{t} - h(t)\right)x \quad \text{and} \quad b(t, x) = \sigma^2 x^2.$$

In the first coefficient $a(t, x)$, the function h is considered as a linear combination of the exogenous factors, and is given by

$$h(t) = \beta_0 + \sum_{i=1}^{q} \beta_i g_i(t)$$

where g_i (for $i = 1, \ldots, q$) are called exogenous factors (external information) and are a time-continuous function in $[t_1, T]$, α, β_i (for $i = 0, \ldots, q$) and $\sigma > 0$ are time-independent real parameters (to be estimated).

It can be proved that the functionals $a(t, x)$ and $b(t, x)$ are non-anticipative and satisfy the Lipschitz and the growth conditions, and consequently that there exists a unique, strong solution to Eq.(1) [see, for example, Liptser and Shiryayev [15], Theorem 4.6].

Furthermore, it is straightforward to show that these functionals are Borel measurable and satisfy the uniform Lipschitz condition and the c-Holder, in particular order 1 Holder, conditions (see, for example, Wong and Hajek [19], Propositions 4.1 and 7.1]. Consequently, there exists a separable, measurable and almost surely (a.s.) sample continuous diffusion process $\{x(t)\,;t\in[t_1,T]\}$ which is the unique (a.s.) solution to Ito's SDE Eq.(1) with infinitesimal moments (drift and diffusion coefficients) given, respectively, by $a(t,x)$ and $b(t,x)$.

2.2. The ptdf and moments of the model

The strong solution to Eq.(1) can be obtained by Ito's formula, transforming the latter using the function $y(t) = \log(x(t))$ to the following SDE

$$dy(t) = \left(\frac{\alpha}{t} - h(t) - \frac{\sigma^2}{2}\right)dt + \sigma dw(t); \quad y(t_1) = \log(x_{t_1}).$$

By integrating and substituting, we deduce that the analytical expression of the solution to the SDE Eq.(1) is

$$x(t) = x_{t_1}\left(\frac{t}{t_1}\right)^{\alpha}\exp\biggl(-\int_{t_1}^{t}\left(h(\tau) - \frac{\sigma^2}{2}\right)d\tau + \sigma(w(t) - w(t_1))\biggr),$$

then, $x(t)$ has a one-dimensional lognormal distribution $\Lambda_1[\mu(t_1,t,x_{t_1}),\sigma^2(t-t_1)]$, where $\mu(s,t,x)$ is given by

$$\mu(s,t,x) = \log(x) + \alpha\log(t/s) - (\beta_0 + \sigma^2/2)(t-s) - \sum_{i=1}^{q}\beta_i\int_s^t g_i(\tau)d\tau,$$

and therefore, the tpdf of the process has the following form

$$f(y,t \mid x,s) = \left[2\pi\sigma^2(t-s)\right]^{-1/2} y^{-1}\exp\left(-\frac{[\log(y) - \mu(s,t,x)]^2}{2\sigma^2(t-s)}\right). \tag{2}$$

Taking into account that $x(t) \mid x(s) = x_s$ is distributed as $\Lambda_1\left[\mu(s,t,x_s),\sigma^2(t-s)\right]$ and bearing in mind the properties of this distribution, the r-th conditional moment of the process is expressed by

$$\mathbb{E}\left[x^r(t)|x(s) = x_s\right] = \exp\biggl(r\mu(s,t,x_s) + \frac{r^2\sigma^2}{2}(t-s)\biggr).$$

Then, the conditional trend function ($r = 1$) of the process is

$$\mathbb{E}\left[x(t) \mid x(s) = x_s\right] = x_s\left(\frac{t}{s}\right)^{\alpha} e^{-\beta_0(t-s)-\sum_{i=1}^{q}\beta_i\int_s^t g_i(\tau)\,d\tau}.$$

Assuming the initial condition $\mathrm{P}(x(t_1) = x_1) = 1$, we obtain the trend function of the process

$$\mathbb{E}\left[x(t)\right] = x_{t_1}\left(\frac{t}{t_1}\right)^{\alpha} e^{-\beta_0(t-t_1)-\sum_{i=1}^{q}\beta_i\int_{t_1}^t g_i(\tau)\,d\tau}.$$

And the variance of the process is given by

$$Var\left[x(t)\right] = x_{t_1}^2\left(\frac{t}{1}\right)^{2\alpha} e^{-2\beta_0(t-t_1)-2\sum_{i=1}^{q}\beta_i\int_{t_1}^t g_i(\tau)d\tau}\left(e^{\sigma^2(t-t_1)} - 1\right).$$

§3. Statistical inference on the model

3.1. Likelihood parameter estimation

In the present study, with discrete sampling, we estimate the parameters α, σ^2 and β_i (for $i = 1, \ldots, q$) of the model by applying maximum likelihood estimation (MLE) methodology. Let us consider a discrete sampling of the process $x_1, \ldots, x_n$ for times $t_1, t_2, \ldots, t_n$ and assume an initial distribution $\mathrm{P}\left[x(t_1) = x_1\right] = 1$. Then the associated likelihood function can be obtained from Eq.(2) by the following expression

$$\mathbb{L}(x_1, \ldots, x_n, \alpha, \beta, \sigma^2) = \prod_{i=2}^{n} f\left(x_i, t_i \mid x_{i-1}, t_{i-1}\right).$$

An implementation based on the change of variable can be used in order to work with a known likelihood function and to calculate the maximum likelihood estimators in a simpler way. Consider the following transform: $v_i = (t_i - t_{i-1})^{-1/2}\left(\log(x_i) - \log(x_{i-1})\right)$, $i = 2, \ldots, n$, then, with the following reparametrization $\Gamma = \left(\alpha, -(\beta_0 + \sigma^2/2), -\beta_1, \ldots, -\beta_q\right)'$ and

$$u_i = (t_i - t_{i-1})^{-1/2}\left(\log(t_i/t_{t-1}), t_i - t_{i-1}, \int_{t_i}^{t_{i-1}} g_1(\tau)\, d\tau, \ldots, \int_{t_i}^{t_{i-1}} g_q(\tau)\, d\tau\right)'.$$

Then, the likelihood function for the transformed sample is

$$\mathbb{L}_{v_2,\ldots,v_n}(\Gamma, \sigma^2) = \left[2\pi\sigma^2\right]^{-(n-1)/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=2}^{n}(v_i - u_i'\Gamma)^2\right).$$

Let $\mathbf{V} = (v_2, \ldots, v_n)'$ and $\mathbf{U}$ be the $(q+2) \times (n-1)$ matrix, whose rank is $q+2$, and given by $\mathbf{U} = (\mathbf{u}_2, \ldots, \mathbf{u}_n)$. Then, the likelihood function can be rewritten in the following form:

$$\mathbb{L}_{\mathbf{V}}(\Gamma, \sigma^2) = \left[2\pi\sigma^2\right]^{-(n-1)/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{V} - \mathbf{U}'\Gamma)'(\mathbf{V} - \mathbf{U}'\Gamma)\right). \tag{3}$$

After calculating the derivatives of the log-likelihood function with respect to the parameter matrix Γ and the coefficient σ^2, the likelihood equations are

$$\mathbf{U}(\mathbf{V} - \mathbf{U}'\widehat{\Gamma}) = 0,$$
$$(n-1)\hat{\sigma}^2 = (\mathbf{V} - \mathbf{U}'\widehat{\Gamma})'(\mathbf{V} - \mathbf{U}'\widehat{\Gamma}).$$

From which, the likelihood estimators of the parameters are

$$\widehat{\Gamma} = (\mathbf{U}\mathbf{U}')^{-1}\mathbf{U}\mathbf{V},$$
$$(n-1)\hat{\sigma}^2 = \mathbf{V}'\mathrm{H}_{\mathbf{U}}\mathbf{V},$$

where $\mathrm{H}_{\mathbf{U}} = \mathrm{I}_{n-1} - \mathbf{U}'(\mathbf{U}\mathbf{U}')^{-1}\mathbf{U}$ is an idempotent symmetric matrix.

Remark 1. In the absence of exogenous factors (i.e: $g_i = 0$, for $i = 1, \ldots, q$), we obtain the stochastic Gamma diffusion process studied in Gutiérrezet al. [11, 13], and it can be shown that all the results established in the present study generalize those obtained in the two papers cited.

3.2. Properties of likelihood estimators

3.2.1. Distribution and independence of MLEs

Using Eq.(3), it can be deduced that $\mathbf{V} \sim \mathcal{N}_{n-1}\left[\mathbf{U}'\Gamma, \sigma^2\mathbf{I}_{n-1}\right]$.

The rank of $\mathbf{U}$ is $q+2$, Then, $(\mathbf{UU}')^{-1}\mathbf{U}$ has the same rank, and therefore, we have

$$\widehat{\Gamma} \sim \mathcal{N}_{q+2}\left[\Gamma, \sigma^2(\mathbf{UU}')^{-1}\right].$$

On the one hand, we have $\sigma^{-1}(\mathbf{V} - \mathbf{U}'\Gamma) \sim \mathcal{N}_{n-1}(0, \mathbf{I}_{n-1})$, as $\mathrm{H}_{\mathbf{U}}$ is idempotent, then by a known multivariate analysis result (see for example, [18, Theorem 2, p. 57]), we have

$$\sigma^{-2}(\mathbf{V} - \mathbf{U}'\Gamma)'\mathrm{H}_{\mathbf{U}}(\mathbf{V} - \mathbf{U}'\Gamma) \sim \chi^2_{\mathrm{rank}(\mathrm{H}_{\mathbf{U}})}.$$

And by taking into account that $\mathrm{H}_{\mathbf{U}}$ is symmetric and idempotent, we have $\mathrm{rank}(\mathrm{H}_{\mathbf{U}}) = \mathrm{tr}(\mathrm{H}_{\mathbf{U}}) = n - q - 3$, and therefore

$$\sigma^{-2}(\mathbf{V} - \mathbf{U}'\Gamma)'\mathrm{H}_{\mathbf{U}}\sigma^{-1}(\mathbf{V} - \mathbf{U}'\Gamma) = \sigma^{-2}\mathbf{V}'\mathrm{H}_{\mathbf{U}}\mathbf{V} \sim \chi^2_{n-q-3}.$$

From which, we deduce that

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-q-3)}.$$

On the other hand, as $(\mathbf{UU}')^{-1}\mathbf{U}\mathrm{H}_{\mathbf{U}} = 0$, then by Theorem 3 in [18, p. 59], we have $(\mathbf{UU}')^{-1}\mathbf{UV}$ and $\mathbf{V}'\mathrm{H}_{\mathbf{U}}\mathbf{V}$ are independently distributed, which means that $\hat{\Gamma}$ and $\hat{\sigma}^2$ are independently distributed.

3.2.2. Sufficiency and Completeness of MLEs

By substracting and adding $\mathbf{U}'\widehat{\Gamma}$ to $\mathbf{V} - \mathbf{U}'\Gamma$, expression Eq.(3) becomes

$$\mathbb{L}_{\mathbf{V}}(\Gamma, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n-1}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)\hat{\sigma}^2 + (\widehat{\Gamma} - \Gamma)'\mathbf{UU}'(\widehat{\Gamma} - \Gamma))\right]\right).$$

This shows that $\left(\widehat{\Gamma}, \hat{\sigma}^2\right)$ is conjointly sufficient for $\left(\Gamma, \sigma^2\right)$.

The completeness follows by means of similar reasoning to that established for the maximum likelihood estimators of the parameters of the multivariate normal distribution (see, for example, Anderson [2]).

Finally it can be deduced that the estimators $\widehat{\Gamma}$ and $\frac{(n-1)\hat{\sigma}^2}{(n-q-3)\sigma^2}$ are the UMVUE for the parameters Γ and σ^2 respectively.

3.3. Parameter confidence intervals

On the basis of the above results, it can be deduced that the $(1-\gamma)\%$ confidence interval for the parameter σ^2 is given, by

$$\left(\frac{(n-1)\hat{\sigma}^2}{\chi^2{}_{n-q-3,\frac{\gamma}{2}}}, \frac{(n-1)\hat{\sigma}^2}{\chi^2{}_{n-q-3,1-\frac{\gamma}{2}}}\right).$$

And the $(1-\gamma)\%$ concentration ellipsoid for $\Gamma_* = \left(-\beta_1, \ldots, -\beta_q\right)'$ is given by

$$\left(\Gamma_* - \hat{\Gamma}_*\right)\left[A_{(22)}\right]^{-1}\left(\Gamma_* - \hat{\Gamma}_*\right)' \le \frac{(n-1)q}{n-q-3}\hat{\sigma}^2 F_{q,n-q-3,\gamma},$$

where $\chi^2_{n,\gamma}$ and $F_{m,n,\gamma}$ are the upper 100γ per cent points of the χ^2 with n degrees of freedom and the F- distribution with m and n degrees of freedom, respectively, $A_{(22)}$ is $q \times q$-matrix and given in

$$(\mathbf{UU}')^{-1} = \begin{pmatrix} A_{(11)} & A_{(12)} \\ A_{(21)} & A_{(22)} \end{pmatrix}.$$

§4. Conclusions

The Gamma process, from the outset, is a non-homogenenous diffusion process, as its drift depends explicitly on the time t. In the present paper, we have introduced a new type of Gamma diffusion, including a second source of non-homogeneneity, which is derived from making the function $h(t)$, which forms part of the drift of the initial diffusion, depend on q exogenous factors, $g_1(t)$, $i = 1, \ldots, q$. These factors are external (or exogenous) to the process $x(t)$ itself, and act as "regressors" and thus the drift of the diffusion varies, as do its trend functions. In consequence, through an appropriate choice of such exogenous factors, it is possible to fit the Gamma diffusion introduced, and in particular its trend functions, to a real phenomenon, in a way that is more suitable in statistical terms than if this were done with the initial Gamma diffusion (without exogenous factors). This is so because, thanks to these factors, we can model the influence of certain exogenous factors on the dynamic behaviour of the endogenous variable $x(t)$.

This fit can be applied, in practice, to the Gamma diffusion examined in the present study because it has been possible to develop the basic results of statistical inference (the estimation and testing of hypotheses) for the model defined in Eq(1).Thus, we have a method for adjusting, and for analyzing the goodness of fit, that is suitable for practical implementation.

Acknowledgements

This work was supported partially by research project MTM 2008-05785, Ministerio de Educación y Ciencia, and project P06-FQM-02271, Junta de Andalucía, Spain.

References

[1] Ait-Sahalia, Y. Maximum likelihood estimation of discritely sampled diffusions: a closed- froma approximation approach. *Econometrica* (2002), 223–262.

[2] Anderson, T. *An introduction to multivariate statistical analysis, Second Edition.* New York, 1984.

[3] Bibby, B., and Sorensen, M. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli 1 (1/2)* (1995), 17–39.

[4] Egorov, A., Li, H., and Xu, Y. Maximum likelihood estimation of time-homogeneous diffusions. *Journal of Econometrics 114* (2003), 107–139.

[5] Ferrante, L., Bompade, S., Possati, L., Leone, L., and Montanari, M. A stochastic formulation of the Gompertzian growth model for in vitro bactericidad kinetics: Parameter estimation and extinction probability. *Biometrical Journal 47(3)* (2005), 306–318.

[6] Forman, J., and Sorensen, M. The pearson diffusions: A class of statistically tractable diffusion processes. *Scandinnavian Journal of Statistics 35* (2005), 438–465.

[7] Giovanis, A., and Skiadas, C. A stochastic logistic innovation diffusion-model studying the electricity consumption in greece and the united states. *Technological Forecasting and Social Change 61* (1999), 253–264.

[8] Gutiérrez, R., Gutiérrez-Sánchez, R., and Nafidi, A. Electricity consumption in morocco: Stochastic gompertz diffusion analysis with exogenous factors. *Applied Energy 83* (2006), 1139–1151.

[9] Gutiérrez, R., Gutiérrez-Sánchez, R., and Nafidi, A. Trend analysis and computational statistical estimation in a stochastic rayleigh model: Simulation and application. *Mathematics and Computers in Simulation 77* (2008), 209–217.

[10] Gutiérrez, R., Gutiérrez-Sánchez, R., and Nafidi, A. Trend analysis using nonhomogeneous stochastic diffusion processes. emission of CO_2; kyoto protocol in spain. *Stochastic Environmental Research and Risk Assessment 22* (2008), 57–66.

[11] Gutiérrez, R., Gutiérrez-Sánchez, R., and Nafidi, A. The trend of the total stock of the private car-petrol in spain: Stochastic modelling using a new gamma diffusion process. *Applied Energy 86* (2009), 18–24.

[12] Gutiérrez, R., Gutiérrez-Sánchez, R., Nafidi, A., and El Melliani, S. A non homogenous stochastic vasicek diffusion model: Application to global CO_2 emission in morocco. In *MAMERN-07* (2007), B. Amaziane, D. Barrera, and D. Sbibih, Eds., pp. 323–329.

[13] Gutiérrez, R., Gutiérrez-Sánchez, R., Nafidi, A., and Merbouha, A. A stochastic diffusion model based on the gamma density: Statistical inference. *Monografías del Seminario Matemático García de Galdeano 34* (2008), 117–125.

[14] Gutiérrez, R., Román, P., and Torres, F. Inference and first-passage time for the lognormal diffusion process with exogenous factors: application to modelling in economics. *Applied Stochastic Models in Business and Industry 15* (1999), 325–332.

[15] Lipter, R., and Shiryayev, A. *Statistics of Random Processes II. Applications*. Springer-Verlag, New York, 1978.

[16] Picchini, U., Ditlevsen, S., and De Gaetano, A. Maximum likelihood estimation of a time-inhomogeneous stochastic differential model of glucose dynamics. *Mathematical Medicine and Biology 25* (2008), 141–155.

[17] Prakasa-Rao, B. *Statistical inference for diffusion type processes*. Arnold, London and Oxford University Press, New York, 1999.

[18] Searle, S. *Linear Models*. John Wiley & Sons, 1971.

[19] Wong, E., and Hajek, B. *Stochastic processes in engineering systems.* Springer-Verlag, New York, 1985.

A. Nafidi
Université Hassan 1er
Ecole Supérieure de Technologie- Berrechid
B.P. 218, Berrechid, Maroc
nafidiah@ugr.es

R. Gutiérrez R. and R. Gutiérrez-Sánchez
Department of Statistics
and Operations Research
University of Granada, Facultad de Ciencias
Campus de Fuentenueva s/n,
18071 Granada, Spain
rgjaimez@ugr.es and ramongs@ugr.es

MONOGRAFÍAS DEL SEMINARIO MATEMÁTICO GARCÍA DE GALDEANO

Desde 2001, el Seminario ha retomado la publicación de la serie *Monografías* en un formato nuevo y con un espíritu más ambicioso. El propósito es que en ella se publiquen tesis doctorales dirigidas o elaboradas por miembros del Seminario, actas de congresos en cuya organización participe o colabore el Seminario, y monografías en general. En todos los casos, se someten al sistema habitual de arbitraje anónimo.

Los manuscritos o propuestas de publicaciones en esta serie deben remitirse a alguno de los miembros del Comité editorial. Los trabajos pueden estar redactados en español, francés o inglés.

Las monografías son recensionadas en *Mathematical Reviews* y en *Zentralblatt MATH.*

Últimos volúmenes de la serie:

21. A. Elipe y L. Floría (eds.): *III Jornadas de Mecánica Celeste*, 2001, ii + 202 pp., ISBN: 84-95480-21-2.

22. S. Serrano Pastor: *Modelos analíticos para órbitas de satélites artificiales de tipo quasi-spot*, 2001, vi + 76 pp., ISBN: 84-95480-35-2.

23. M. V. Sebastián Guerrero: *Dinámica no lineal de registros electrofisiológicos*, 2001, viii + 251 pp., ISBN: 84-95480-43-3.

24. Pedro J. Miana: *Cálculo funcional fraccionario asociado al problema de Cauchy*, 2002, 171 pp., ISBN: 84-95480-57-3.

25. Miguel Romance del Río: *Problemas sobre Análisis geométrico convexo*, 2002, xvii + 214 pp., ISBN: 84-95480-76-X.

26. Renato Álvarez-Nodarse: *Polinomios hipergeométricos y q-polinomios*, 2003, vi + 341 pp., ISBN: 84-7733-637-7.

27. M. Madaune-Tort, D. Trujillo, M. C. López de Silanes, M. Palacios y G. Sanz (eds.): *VII Jornadas Zaragoza-Pau de Matemática Aplicada y Estadística*, 2003, xxvi + 523 pp., ISBN: 84-96214-04-4.

28. Sergio Serrano Pastor: *Teorías analíticas del movimiento de un satélite artificial alrededor de un planeta. Ordenación asintótica del potencial en el espacio fásico*, 2003, 164 pp., ISBN: 84-7733-667-9.

29. Pilar Bolea Catalán: *El proceso de algebrización de organizaciones matemáticas escolares*, 2003, 260 pp., ISBN: 84-7733-674-1.

30. Natalia Boal Sánchez: *Algoritmos de reducción de potencial para el modelo posinomial de programación geométrica*, 2003, 232 pp., ISBN: 84-7733-667-9.

31. M. C. López de Silanes, M. Palacios, G. Sanz, J. J. Torrens, M. Madaune-Tort y D. Trujillo (eds.): *VIII Journées Zaragoza-Pau de Mathématiques Appliquées et de Statistiques*, 2004, xxvi + 578 pp., ISBN: 84-7733-720-9.

32. Carmen Godés Blanco: *Configuraciones de nodos en interpolación polinómica bivariada*, 2006, xii + 163 pp., ISBN: 84-7733-841-9.

33. M. Madaune-Tort, D. Trujillo, M. C. López de Silanes, M. Palacios, G. Sanz y J. J. Torrens (eds.): *Ninth International Conference Zaragoza-Pau on Applied Mathematics and Statistics*, 2006, xxxii + 440 pp., ISBN: 84-7733-871-X.

34. B. Lacruz, F. J. López, P. Mateo, C. Paroissin, A. Pérez-Palomares y G. Sanz (eds.): *Pyrenees International Workshop on Statistics, Probability and Operations Research, SPO 2007*, 2008, 205 pp., ISBN: 978-84-92521-18-0.