MONOGRAFÍAS MATEMÁTICAS
"GARCÍA DE GALDEANO"

2012

N.º 37

Eleventh International Conference Zaragoza-Pau on Applied Mathematics and Statistics

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

MONOGRAFÍAS MATEMÁTICAS
GARCÍA DE GALDEANO

Número **37**, 2012

Eleventh International Conference Zaragoza-Pau on Applied Mathematics and Statistics

Jaca (Spain), September 15–17, 2010

Editors

J. Giacomoni
M. Madaune-Tort
C. Paroissin
G. Vallet
Université de Pau et des Pays de l'Adour, France

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

J. J. Torrens
Universidad Pública de Navarra, Spain

JORNADAS ZARAGOZA–PAU DE MATEMÁTICA APLICADA Y ESTADÍSTICA (11ª. Jaca 2010. Jaca)

Eleventh International Conference Zaragoza–Pau on Applied Mathematics and Statistics : Jaca, Spain, September 15 – 17, 2010 / editors J. Giacomoni… [et al.]. — Zaragoza : Prensas Universitarias de Zaragoza : Instituto Universitario de Matemáticas y Aplicaciones, Universidad de Zaragoza, 2012

XXVI, 208 p. ; 24 cm. — (Monografías Matemáticas García de Galdeano ; 37)

ISBN 978-84-15538-15-8

Matemáticas-Congresos y asambleas

GIACOMONI, J.

51(063)

Monografías Matemáticas García de Galdeano nº 37
Marzo 2012
Universidad de Zaragoza

Imprime: Servicio de Publicaciones. Universidad de Zaragoza

D.L.: Z-790-2012
ISBN: 978-84-15538-15-8

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

XI Jornadas ZARAGOZA - PAU XI Journées

Eleventh International Conference Zaragoza-Pau on Applied Mathematics and Statistics

Residencia Universitaria de Jaca, September 15 – 17, 2010

http://pcmap.unizar.es/~jaca2010

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 69
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:
J. Giacomoni
Tel.: 559 40 75 53
jacques.giacomoni@univ-pau.fr

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

G. Vallet
Tel.: 559 40 75 57
guy.vallet@univ-pau.fr
Laboratoire de Mathématiques et leurs Applications de Pau

Contents

Published articles

Preface

The *International Conference Zaragoza-Pau on Applied Mathematics and Statistics* is organized normally every two years since 1989 by the *Departamento de Matemática Aplicada*, the *Departamento de Métodos Estadísticos*, both from the *Universidad de Zaragoza* (Spain), and the *Laboratoire de Mathématiques Appliquées et leurs Applications*, 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 Eleventh Conference took place in Jaca (Spain) from 15th to 17st September 2010. The official opening ceremony was graced by the presence of the Chancellor of the University of Zaragoza, 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, 87 mathematicians, coming from different universities, research institutes or the industrial sector, attended 13 plenary lectures, 39 contributed talks and a poster session with a total of 10 posters.

This edition had the pleasure of a special event. A mini-symposium in honour of Monique Madaune-Tort, Professor of the Université de Pau et des Pays de l'Adour. Monique is one of the pioneers of this conference and several other French-Spanish events in Mathematics. She belongs to the group of French and Spanish researchers who are deeply involved in the academic and scientific cooperation between the Université de Pau et des Pays de l'Adour and the Universidad de Zaragoza. In this mini-symposium, 11 invited conferences were held.

The principal talks were about theoretical and numerical analysis of deterministic models described by differential equations, statistics and stochastics processes, surface approximation and image analysis. At the same time, there was also a session devoted to Algebra and Geometry. These proceedings contain 1 paper based on the corresponding invited lectures along with 19 full length refereed research papers. In a special volume, 9 papers based on invited lectures given in the mini-symposium, as a special tribute to Monique Madaune-Tort, are published as a Monografía de la Real Academia de Ciencias de Zaragoza.

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 Eleventh Conference.

We wish to express our gratitude to Mohamed Amara (U. Pau), Enrique Artal (U. Zaragoza), Mehdi Badra (U. Pau), Roland Becker (U. Pau), Laurent Bordes (U. Pau), Mira Bozzini (U. Milano-Bicocca), Bénédicte Chassat-Alziary (U. Toulouse I), Marc Dambrine (U. Pau), Alberto Elduque (U. Zaragoza), Raúl Gouet (U. Chile), Laurent Lévi (U. Pau), Francisco Lisbona (U. Zaragoza), Miguel Pasadas (U. Granada), Juan Manuel Peña (U. Zaragoza), Tomas Sauer (U. Justus-Leibig-Geissen) and Jean Vallès (U. Pau), who, together with us, formed the Scientific Committee, and Mehdi Badra, Jacky Cresson, Marc Dambrine, Daniele Faenzi, Vicent Florens, Laurent Lévi and Marie-Laure Rius, from the Université de Pau et des Pays de l'Adour, 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 the others who helped in the organization of the Conference, in particular, Carmen Paniagua and José Manuel Palacios.

We finally acknowledge the assistance provided for the realization of the proceedings by the Instituto Universitario de Matemáticas y Aplicaciones, contained in 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.

The next edition of the Conference Zaragoza-Pau will be held in Jaca from 17th to 19th September 2012. All of you are cordially invited to participate in this event.

Pau and Zaragoza, Mars 2012
The Editors

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

Gerardo Sanz
Departamento de Métodos Estadísticos
Universidad de Zaragoza

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

Jacques Giacomoni
Monique Madaune-Tort
Christian Paroissin
Guy Vallet
Laboratoire de Mathématiques Appliquées et leurs Applications
Université de Pau et des Pays de l'Adour

Contributors

List of participants

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

ANDREIANOV, Boris
Laboratoire de Mathématiques,
Université de Franche-Compté,
16 route de Gray,
25030 Besançon Cedex, France.
boris.andreianov@univ-fcompte.fr

BAL, Kaushik
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.
kausbal@gmail.com

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

BARRAU, Nelly
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.
nelly.barrau@etud.univ-pau.fr

BARRIO, Roberto
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
rbarrio@unizar.es

BAUZET, Caroline
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.
carobauzet@hotmail.fr

BECKER, Roland
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.
roland.becker@univ-pau.fr

BLANCHET, Christian
U.F.R. Mathématiques, Université Paris Diderot,
175 rue de Chevaleret,
75013 Paris Cedex, France.
blanchet@math.jussieu.fr

BLESA, Fernando
Grupo de Mecánica Espacial,
Departamento de Física Aplicada, Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
fblesa@unizar.es

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

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

BOURDIN, Loïc
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.
bourdin.l@etud.univ-pau.fr

BRAACK, Malte
Mathematisches Seminar,
University of Kiel,
CAU Kiel,
Ludewig-Meyn-Str, 4,
D-24098 Kiel, Germany.
braack@math.uni-kiel.de

CAPATINA, Daniela
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.
daniela.capatina@univ-pau.fr

CARDOULIS, Laure
CEREMATH - Université Toulouse I,
Pl. du Doyen G. Marty,
31042 Toulouse Cedex, France.
laure.cardoulis@univ-tlse1.fr

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

CAUBET, 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.
fabien.caubet@etud.univ-pau.fr

CAVIEDES, Daniel
Departamento de Mecánica de Fluidos,
EINA, Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
daniel.caviedes@unizar.es

CHEDOM FOTSO, Donatien
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.
donatien.chedomfotso@univ-pau.fr

CLAVERO, Carmelo
Departamento de Matemática Aplicada,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
clavero@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

CUESTA, Elvira
Departamento de Matemática Aplicada,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
ecuesta@unizar.es

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

DENA, Ángeles
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
adena@unizar.es

DÍAZ, Jesús Ildefonso
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad Complutense de Madrid,
Plaza de las Ciencias 3,
28040 Madrid, Spain.
ildefonso.diaz@mat.ucm.es

DOMELEVO, Komla
Institut de Mathématiques de Toulouse,
Université Paul Sabatier,
118 route de Narbonne,
31062 Toulouse Cedex, France.
`komla.domelevo@math.univ-toulouse.fr`

EFENDIYEV, Messoud
Helmholtz Zentrum München and TU Manchen,
GmbH, Ingolstädter Landstrße 1,
D-85764 Neuherberg, Germany.
`messoud.efendiyev@helmholtz-muenchen.de`

FERREIRA, Consuelo
Departamento de Matemática Aplicada,
Facultad de Veterinaria,
Universidad de Zaragoza,
c/ Miguel Servet 117,
50013 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`

FORTES, Miguel Ángel
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Granada,
Campus de Fuentenueva s/n,
18071 Granada. Spain.
`mafortes@ugr.es`

GASPAR, Francisco
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
`fjgaspar@unizar.es`

GENOUD, François
Mathematical Institute,
University of Oxford,
24-29 St. Giles,
Oxford OX1 3LB, England.
`frgenoud@gmail.com`

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`

GONZÁLEZ, Daniel
Departamento de Matemáticas y Computación,
Universidad de la Rioja,
Edificio Vives, c/ Luis de Ulloa s/n,
26004 Logroño, Spain.
`daniel.gonzalez@unirioja.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`

HANEN, Hanna
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`

HERNÁNDEZ, Jesús
Departamento de Matemática Aplicada,
Universidad Autónoma de Madrid,
c/ Francisco Tomás y Valiente 7,
28049 Madrid, Spain.
`jesus.hernandez@uam.es`

IBÁÑEZ, María 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,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
dizquier@unizar.es

JASOR, Marie-Josée
Laboratoire de Mathématiques,
Université de Clermont-Ferrand,
24 Avenue des Landais,
BP 26, 63177 Aubière Cedex, France.
Marie-Josee.Jasor@
math.univ-bpclermont.fr

JODRÁ, Pedro
Departamento de Métodos Estadísticos,
EINA, Universidad de Zaragoza,
c/ María de Luna 3,
50018 Zaragoza, Spain.
pjodra@unizar.es

KOLB, Sébastien
CReA (Centre de Recherche de l'Armée de l'Air),
BA 701, 13661, Salon Air, France.
sebastien.kolb@inet.air.defense.gouv.fr

KOUIBIA, Abdelouahed
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Granada,
Campus de Fuentenueva s/n,
18071 Granada, Spain.
kouibia@ugr.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

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

LÉVI, 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.levi@univ-pau.fr

LISBONA, Francisco Javier
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
lisbona@unizar.es

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

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

MARCO-BUZUNÁRIZ, Miguel Ángel
Centro Universitario de la Defensa,
Academia General Militar,
Ctra. de Huesca s/n, 50090 Zaragoza, Spain.
mmarco@unizar.es

MARTÍN-MORALES, Jorge
Centro Universitario de la Defensa,
Academia General Militar,
Ctra. de Huesca s/n, 50090 Zaragoza, Spain.
jorge@unizar.es

MERCIER, Sophie
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.
sophie.mercier@univ-pau.fr

MERKER, Jochen
Institut für Mathematik,
Universität Rostock,
Raum 438 Ulmenstraße 69 (Haus 3),
18057 Rostock, Germany.
jochen.merker@uni-rostock.de

MESLAMENI, 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.
medmeslameni@univ-pau.fr

NAVASCUÉS, María Antonia
Departamento de Matemática Aplicada,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
manavas@unizar.es

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

PALACIOS, José Manuel
Departamento de Matemática Aplicada,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
josefoll@msm.com

PALACIOS, Manuel
Grupo de Mecánica Espacial,
Departamento de Matemática Aplicada,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
mpala@unizar.es

PANIAGUA, Carmen
Departamento de Matemática Aplicada,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
josefoll@msm.com

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 Matemática Aplicada,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
ester.perez@unizar.es

PIASECKI, Slawomir Stanislaw
Grupo de Mecánica Espacial,
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
piasek@unizar.es

POUTOUS, Cécile
CReA (Centre de Recherche de l'Armée de l'Air),
BA 701, 13661, Salon Air, France.
cecile.poutous@univ-pau.fr

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,
EINA, Universidad de Zaragoza,
c/ María de Luna 3, 50018 Zaragoza, Spain.
carmenr@unizar.es

RODRÍGUEZ, Marcos
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
marcos@unizar.es

SAUVY, Paul
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.
paul.sauvy@etud.univ-pau.fr

SCHINDLER, Ian
CEREMATH,
Université Toulouse I,
Manufacture des Tabacs,
21 Allées de Brienne,
31000 Toulouse. France.
ian.schindler@gmail.com

SEAM, Ngonn
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.
seamngonn@yahoo.fr

SEBASTIÁN, María Victoria
Centro Universitario de la Defensa,
Academia General Militar,
Ctra. de Huesca s/n, 50090 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

SHMAREV, Sergey
Departamento de Matemáticas,
Universidad de Oviedo,
c/ Calvo Sotelo s/n,
33007 Oviedo, Spain.
sergey.shmarev@gmail.com

STUART, Charles
Section de Mathématiques,
Faculté des Sciences de base,
EPFL, Chemin du Vieux Réservoir,
2 CH-1116 Cottens, Switzerland.
charles.stuart@epfl.ch

TAKÁČ, Peter
Institut für Mathematik,
Universität Rostock,
Universitätsplatz 1,
D-18055 Rostock, Germany.
peter.takac@uni-rostock.ge

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

TORT, Jacques
Institut de Mathématiques de Toulouse,
Université Paul Sabatier,
118 route de Narbonne,
31062 Toulouse Cedex, France.
jacques.tort@math.univ-toulouse.fr

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

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`

WEI, Na
CEREMATH - Université Toulouse I,
Pl. du Doyen G. Marty,
31042 Toulouse Cedex, France.
`nawei8328@gmail.com`

WITTBOLD, Petra
Facultät für Mathematik,
Universität Duisburg-Essen,
Universitätsstraße 2,
45117 Essen, Germany.
`petra.wittbold@uni-due.de`

ZIMMERMANN, Alexandra
TU Berlin,
MA 6-4 Straße des 17, Juni 135,
10623 Berlin, Germany.
`zmorzyns@tu-berlin.de`

Other communications

The following contributions have been published in volume no. 38 of the *Monografías de la Real Academia de Ciencias de Zaragoza*, entitled *A special tribute to Professor Monique Madaune-Tort*.

On the regularity for the Laplace equation and the Stokes system
C. Amrouche and M. A. Rodríguez Bellido

On the well-posedness for the coupling of multidimensional quasilinear diffusion-transport equations
G. Aguilar and L. Lévi

Option pricing for stocks with dividends: An analytic approach by PDEs
B. Alziary and P. Takáč

On intrinsic formulation and well-posedness of a singular limit of two-phase flow equations in porous media
B. Andreianov, R. Eymard, M. Ghilani and N. Marhraoui

Energy solutions of evolution equations with nonstandard growth conditions
S. Antontsev and S. Shmarev

On a class of quasilinear Barenblatt equations
C. Bauzet, J. Giacomoni and G. Vallet

Estimates of the solutions of some asymmetric problems defined on $\mathbb{R}^N$, $N \geq 3$
J. Fleckinger

Localized sequences of approximate critical points
C. A. Stuart

An inverse diffusion problem in a degenerate parabolic equation
J. Tort

⋆ ⋆ ⋆ ⋆ ⋆

The following contributions were also presented in the Conference Zaragoza-Pau, but they are not included in this book. Some will appear in other publications.

Integration of ODEs and variational equations: TIDES
A. Abad, R. Barrio, F. Blesa and M. Rodríguez

L^p theory for vector potentials and Stokes equations with non standard boundary conditions
C. Amrouche and N. E. H. Seloula

When are QBD models solvable by RG factorization and truncation
F. Avram and D. F. Chedom

Topologies, quasi-uniformities and quasi-pseudo-metrics
L. Barbet

Enhancement of the polynomial reproduction of multivariate quasi-interpolants
D. Barrera, A. Guessab, M. J. Ibáñez and O. Nouisser

Forman discrete Morse theory and gradient vector field
C. Blanchet

Staggered grid discretizations for the double porosity model
N. Boal, F. J. Gaspar, F. J. Lisbona and P. N. Vabishchevich

Stabilized finite element for Darcy flow and application to hydrothermal flow
M. Braack, J. Carpio Huertas and F. Schieweck

Remarks on the progressive iteration approximation property
J. M. Carnicer, J. Delgado and J. M. Peña

Numerical simulation of 1D transient flow in variably saturated soils
D. Caviedes-Voullième, P. García-Navarro and J. Murillo

Uniformly convergent methods for singularly perturbed problems of convection-diffusion and reaction-diffusion type
C. Clavero, J. L. Gracia and F. J. Lisbona

Necessary and sufficient positivity criterion for system of stochastic PDEs
J. Cresson

On generalized Ventcel's type boundary conditions for Laplace operator in a bounded domain
M. Dambrine

A quasilinear parabolic model for population evolution
A. Derlet, J. P. Gossez and P. Takáč

On very weak solutions of higher order equations
J. I. Diaz

Positive solutions for the p-Laplacian with Robin boundary conditions on irregular domains
P. Drabek and I. Schindler

Mathematical modelling of biofilm
M. Efendiyev

On a elliptic-parabolic integro-differential problem in L^1
P. Wittbold

Renormalized solutions for a nonlinear parabolic equation with variable exponents and L^1-data
A. Zimmermann

Monografías Matemáticas *García de Galdeano* **37**, 1–11 (2012)

VERY WEAK SOLUTIONS OF STOKES PROBLEM IN EXTERIOR DOMAIN

Chérif Amrouche and Mohamed Meslameni

Abstract. The existence and the uniqueness of very weak solutions of Stokes system are well known in the classical Sobolev spaces $\boldsymbol{W}^{m,p}(\Omega)$ when Ω is bounded (see [3]). When Ω is an exterior domain, a similar approach would fail (in particular because Poincare's inequalities do not hold in such domains). Therefore, a specific functional framework based on density arguments is necessary to do this work.

Keywords: Stokes equations, very weak solutions, weighted Sobolev spaces, exterior domain.

AMS classification: 35Q30, 76D03, 76D05, 76D07.

§1. Introduction

Let Ω' be a bounded connected open domain in $\mathbb{R}^3$ with boundary $\partial\Omega' = \Gamma$ of class $C^{1,1}$ representing an obstacle and let Ω its complement, i.e. $\Omega = \mathbb{R}^3 \setminus \overline{\Omega'}$. In this work, we are interested in the existence and the uniqueness of very weak solution concerning the Stokes problem in exterior domain:

$$-\Delta \boldsymbol{u} + \nabla q = \boldsymbol{f} \quad \text{and} \quad \nabla \cdot \boldsymbol{u} = h \ \text{ in } \ \Omega, \quad \boldsymbol{u} = \boldsymbol{g} \ \text{ on } \ \Gamma, \tag{$\mathcal{S}$}$$

where $\boldsymbol{u}$ denote the velocity and q the pressure and both are unknown, $\boldsymbol{f}$ the external forces, h the compressibility condition and $\boldsymbol{g}$ the boundary condition for the velocity, the three functions being known. This problem is well done in 2005 by R. Farwig [4], with data $\boldsymbol{f} = \operatorname{div} \mathbb{F}_0$, h and $\boldsymbol{g}$ satisfying

$$\mathbb{F}_0 \in \boldsymbol{L}^r(\Omega),\ h \in L^r(\Omega),\ \boldsymbol{g} \in \boldsymbol{W}^{-1/p,p}(\Gamma),\ 3 < p < \infty,\ \frac{1}{3} + \frac{1}{p} = \frac{1}{r}$$

yielding $\frac{3}{2} < r < 3$.

In this paper, we are interested in the following data:

$$\boldsymbol{f} = \operatorname{div} \mathbb{F}_0 + \nabla f_1, \quad h \in L^r(\Omega) \quad \text{and} \quad \boldsymbol{g} \in \boldsymbol{W}^{-1/p,p}(\Gamma),$$

with

$$\mathbb{F}_0 \in \boldsymbol{L}^r(\Omega),\ f_1 \in W_0^{-1,p}(\Omega),\ \frac{3}{2} < p < \infty, \quad \text{and} \quad \frac{1}{3} + \frac{1}{p} = \frac{1}{r},$$

or

$$\mathbb{F}_0 \in \boldsymbol{W}^{0,r}_{-1}(\Omega),\ f_1 \in W^{-1,p}_{-1}(\Omega) \quad \text{and} \quad h \in W^{0,r}_{-1}(\Omega),$$

with

$$\frac{3}{2} < p < \infty,\ p \neq 3 \quad \text{and} \quad \frac{1}{3} + \frac{1}{p} = \frac{1}{r}.$$

§2. Basic concepts on Sobolev spaces

Let $x = (x_1, x_2, x_3)$ be a typical point in $\mathbb{R}^3$ and let $r = |x| = (x_1^2+x_2^2+x_3^2)^{1/2}$ denote its distance to the origin. We define the weight function $\rho(x) = 1 + r$. For each $p \in \mathbb{R}$ and $1 < p < \infty$, the conjugate exponent p' is given by the relation $1/p + 1/p' = 1$. Then, for any nonnegative integers m and real numbers $p > 1$ and α, setting

$$k = k(m, p, \alpha) = \begin{cases} -1, & \text{if } \frac{3}{p} + \alpha \notin \{1, \dots, m\}, \\ m - \frac{3}{p} - \alpha, & \text{if } \frac{3}{p} + \alpha \in \{1, \dots, m\}, \end{cases}$$

we define the following space:

$$\begin{aligned} W^{m,p}_{\alpha}(\Omega) = \{\, & u \in \mathcal{D}'(\Omega); \\ & \forall\, \lambda \in \mathbb{N}^3 : \ 0 \le |\lambda| \le k,\ \rho^{\alpha-m+|\lambda|}(\ln(1+\rho))^{-1} D^{\lambda} u \in L^p(\Omega); \\ & \forall\, \lambda \in \mathbb{N}^3 : \ k+1 \le |\lambda| \le m, \rho^{\alpha-m+|\lambda|} D^{\lambda} u \in L^p(\Omega)\,\}. \end{aligned}$$

It is a reflexive Banach space equipped with its natural norm:

$$\begin{aligned} \|u\|_{W^{m,p}_{\alpha}(\Omega)} = \Big(& \sum_{0 \le |\lambda| \le k} \|\rho^{\alpha-m+|\lambda|}(\ln(1+\rho))^{-1} D^{\lambda} u\|^p_{L^p(\Omega)} \\ & + \sum_{k+1 \le |\lambda| \le m} \|\rho^{\alpha-m+|\lambda|} D^{\lambda} u\|^p_{L^p(\Omega)} \Big)^{1/p}. \end{aligned}$$

We note that the logarithmic weight only appears if $p = 3$ or $p = 3/2$ and all the local properties of $W^{1,p}_0(\Omega)$ (respectively, $W^{2,p}_0(\Omega)$) coincide with those of the corresponding classical Sobolev space $W^{1,p}(\Omega)$ (respectively, $W^{2,p}(\Omega)$). For $m = 1$ or $m = 2$ we set $\mathring{W}^{m,p}_{\alpha}(\Omega)$ as the adherence of $\mathcal{D}(\Omega)$ for the norm $\|\cdot\|_{W^{m,p}_{\alpha}(\Omega)}$. Then, the dual space of $\mathring{W}^{m,p}_{\alpha}(\Omega)$, denoting by $W^{-m,p'}_{-\alpha}(\Omega)$, is a space of distributions. When $\Omega = \mathbb{R}^3$, we have $\boldsymbol{W}^{1,p}_{\alpha}(\mathbb{R}^3) = \mathring{\boldsymbol{W}}^{1,p}_{\alpha}(\mathbb{R}^3)$.

If Ω is a Lipschitz exterior domain, then for $\alpha = 0$ we have

$$\mathring{\boldsymbol{W}}^{1,p}_0(\Omega) = \Big\{ \boldsymbol{v} \in \boldsymbol{W}^{1,p}_0(\Omega),\ \boldsymbol{v} = \boldsymbol{0} \ \text{ on } \ \partial\Omega \Big\},$$

and

$$\mathring{\boldsymbol{W}}^{2,p}_0(\Omega) = \Big\{ \boldsymbol{v} \in \boldsymbol{W}^{2,p}_0(\Omega),\ \boldsymbol{v} = \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{n}} = \boldsymbol{0} \ \text{ on } \ \partial\Omega \Big\},$$

where $\partial\boldsymbol{v}/\partial\boldsymbol{n}$ is the normal derivate of $\boldsymbol{v}$.

The spaces $\boldsymbol{W}^{1,p}_{\alpha}(\Omega)$ or $\boldsymbol{W}^{2,p}_{\alpha}(\Omega)$ sometimes contain some polynomial functions. We have for $m = 1$ or $m = 2$:

$$\mathcal{P}_j \subset \boldsymbol{W}^{m,p}_{\alpha}(\Omega) \quad \text{with} \quad \begin{cases} j = [m - (3/p + \alpha)], & \text{if } 3/p + \alpha \notin \mathbb{Z}^-, \\ j = -(3/p + \alpha), & \text{otherwise}, \end{cases}$$

where $[s]$ denotes the integer part of the real number s and $\mathcal{P}_j$ is the space of polynomials of degree less then j.

We recall the following Sobolev embeddings for $\alpha = 0$ or $\alpha = 1$

$$W^{1,p}_{\alpha}(\Omega) \hookrightarrow W^{0,p*}_{\alpha}(\Omega) \quad \text{where} \quad p* = \frac{3p}{3-p} \quad \text{and} \quad 1 < p < 3.$$

Consequently, by duality, we have

$$W^{0,q}_{-\alpha}(\Omega) \hookrightarrow W^{-1,p'}_{-\alpha}(\Omega) \quad \text{where} \quad q = \frac{3p'}{3+p'} \quad \text{and} \quad p' > 3/2.$$

On the other hand, if $3/p + \alpha \notin \{1, 2\}$, we have the following continuous embedding:

$$\boldsymbol{W}^{2,p}_{\alpha}(\Omega) \hookrightarrow \boldsymbol{W}^{1,p}_{\alpha-1}(\Omega) \hookrightarrow \boldsymbol{W}^{0,p}_{\alpha-2}(\Omega).$$

§3. Preliminary results

In the sequel, we need to introduce the following spaces:

$$\mathcal{D}_{\sigma}(\Omega) = \{\boldsymbol{\varphi} \in \mathcal{D}(\Omega);\ \nabla \cdot \boldsymbol{\varphi} = 0\} \quad \text{and} \quad \mathcal{D}_{\sigma}(\overline{\Omega}) = \left\{\boldsymbol{\varphi} \in \mathcal{D}(\overline{\Omega});\ \nabla \cdot \boldsymbol{\varphi} = 0\right\}.$$

Then, we show some density results that are essential for the proofs below. We begin by introducing the space

$$\boldsymbol{X}^{\ell}_{r,p}(\Omega) = \left\{\boldsymbol{\varphi} \in \mathring{\boldsymbol{W}}^{1,r}_{\ell}(\Omega);\ \nabla \cdot \boldsymbol{\varphi} \in \mathring{\boldsymbol{W}}^{1,p}_{\ell}(\Omega)\right\}.$$

Thank's to Poincaré-type inequality (see [2]), we can equipped this space with the following norm:

$$\|\boldsymbol{v}\|_{\boldsymbol{X}^{\ell}_{r,p}(\Omega)} = \sum_{1\leq i,j\leq 3} \left\| \frac{\partial \boldsymbol{v}_i}{\partial x_j} \right\|_{\boldsymbol{W}^{0,r}_{\ell}(\Omega)} + \|\nabla \cdot \boldsymbol{v}\|_{\boldsymbol{W}^{1,p}_{\ell}(\Omega)}.$$

Lemma 1. *Let Ω be a Lipschitz open set in $\mathbb{R}^3$ and suppose that $0 \leqslant 1/p - 1/r \leqslant 1/3$. We have the following properties:*

i) The space $\mathcal{D}(\Omega)$ is dense in $\boldsymbol{X}^1_{r,p}(\Omega)$ and, for all $\boldsymbol{q} \in \boldsymbol{W}^{-1,p}_{-1}(\Omega)$ and $\boldsymbol{\varphi} \in \boldsymbol{X}^1_{r',p'}(\Omega)$, we have

$$\langle \nabla \boldsymbol{q}, \boldsymbol{\varphi} \rangle_{[\boldsymbol{X}^1_{r',p'}(\Omega)]' \times \boldsymbol{X}^1_{r',p'}(\Omega)} = -\langle \boldsymbol{q}, \nabla \cdot \boldsymbol{\varphi} \rangle_{\boldsymbol{W}^{-1,p}_{-1}(\Omega) \times \mathring{\boldsymbol{W}}^{1,p'}_{1}(\Omega)}.$$

ii) If in addition $p \neq 3$ and $r \neq 3$, then the space $\mathcal{D}(\Omega)$ is dense in $\boldsymbol{X}^0_{r,p}(\Omega)$ and, for all $\boldsymbol{q} \in \boldsymbol{W}^{-1,p}_{0}(\Omega)$ and $\boldsymbol{\varphi} \in \boldsymbol{X}^0_{r',p'}(\Omega)$, we have

$$\langle \nabla \boldsymbol{q}, \boldsymbol{\varphi} \rangle_{[\boldsymbol{X}^0_{r',p'}(\Omega)]' \times \boldsymbol{X}^0_{r',p'}(\Omega)} = -\langle \boldsymbol{q}, \nabla \cdot \boldsymbol{\varphi} \rangle_{\boldsymbol{W}^{-1,p}_{0}(\Omega) \times \mathring{\boldsymbol{W}}^{1,p'}_{0}(\Omega)}.$$

Proof. The density of $\mathcal{D}(\Omega)$ in $\boldsymbol{X}^{\ell}_{r,p}(\Omega)$ relies on an adequate truncation procedure and regularization. The truncation function that we shall use has been defined by: $\varphi \in \mathcal{D}(\mathbb{R}^3)$ such that $0 \leqslant \varphi(t) \leqslant 1$ for any $t \in \mathbb{R}^3$, and

$$\varphi(t) = \begin{cases} 1, & \text{if } 0 \leqslant |t| \leqslant 1, \\ 0, & \text{if } |t| \geqslant 2. \end{cases}$$

Now, let $\boldsymbol{v} \in \boldsymbol{X}^{\ell}_{r,p}(\Omega)$ and $\widetilde{\boldsymbol{v}}$ be the extension by $\mathbf{0}$ of $\boldsymbol{v}$ to $\mathbb{R}^3$, then we have $\widetilde{\boldsymbol{v}} \in \boldsymbol{X}^{\ell}_{r,p}(\mathbb{R}^3)$. We begin to apply the cut off functions φ_k, defined on $\mathbb{R}^3$ for any $k \in \mathbb{N}^*$, by $\varphi_k(x) = \varphi(x/k)$. Set $\boldsymbol{v}_k = \varphi_k \widetilde{\boldsymbol{v}}$. It is easy to prove that $\boldsymbol{v}_k \to \widetilde{\boldsymbol{v}}$ in $\boldsymbol{X}^{\ell}_{r,p}(\mathbb{R}^3)$ when $k \to \infty$. Now, we start the regularization of our sequence $\boldsymbol{v}_k$. In a first step we consider that Ω' is strictly star-shaped with respect to one of its points which is taken to the origin. Under this assumption, we set $\boldsymbol{v}_{k,\theta}(x) = \boldsymbol{v}_k(\theta x)$ for any real number $\theta > 1$ and $x \in \mathbb{R}^3$. Then $\boldsymbol{v}_{k,\theta} \in \boldsymbol{X}^{\ell}_{r,p}(\mathbb{R}^3)$ and supp $\boldsymbol{v}_{k,\theta}$ is compact in Ω. Moreover

$$\lim_{\theta \to 1} \boldsymbol{v}_{k,\theta} = \boldsymbol{v}_k \ \text{ in } \ \boldsymbol{X}^{\ell}_{r,p}(\mathbb{R}^3).$$

Consequently, for any real number $\epsilon > 0$ small enough, the restriction of $\rho_\epsilon * \boldsymbol{v}_{k,\theta}$ to Ω belongs to $\mathcal{D}(\Omega)$ and

$$\lim_{\epsilon \to 0} \lim_{\theta \to 1} \lim_{k \to \infty} \rho_\epsilon * \boldsymbol{v}_{k,\theta} = \widetilde{\boldsymbol{v}} \ \text{ in } \ \boldsymbol{X}^{1}_{r,p}(\mathbb{R}^3),$$

where ρ_ϵ is a mollifier. Consequently, $\mathcal{D}(\Omega)$ is dense in $\boldsymbol{X}^{\ell}_{r,p}(\Omega)$. In the case where Ω' is only a Lipschitz open set in $\mathbb{R}^3$, we have to recover Ω' by a finite number of star open sets and partitions of unity. Clearly, it suffices to apply the above argument to each of these sets to derive the desired result on the entire domain. □

Remark 1. Observe that for $\boldsymbol{f} \in (\boldsymbol{X}^{\ell}_{r,p}(\Omega))'$ with $\ell = 1$ or $\ell = 0$, there exist $\mathbb{F}_0 = (f_{ij})_{1 \leq i,j \leq 3} \in \boldsymbol{W}^{0,r'}_{-\ell}(\Omega)$ and $f_1 \in W^{-1,p'}_{-\ell}(\Omega)$ such that:

$$\boldsymbol{f} = \nabla \cdot \mathbb{F}_0 + \nabla f_1. \tag{1}$$

Moreover,

$$\|\boldsymbol{f}\|_{[\boldsymbol{X}^{\ell}_{r,p}(\Omega)]'} = \max\Big\{ \big\| f_{ij} \big\|_{\boldsymbol{W}^{0,r'}_{-\ell}(\Omega)}, 1 \leq i,j \leq 3, \|f_1\|_{W^{-1,p'}_{-\ell}(\Omega)} \Big\}.$$

Conversely, if $\boldsymbol{f}$ satisfies (1), then $\boldsymbol{f} \in (\boldsymbol{X}^{\ell}_{r,p}(\Omega))'$.

Giving a meaning to the trace of a very weak solution of the Stokes problem is not trivial: remember that we are not in the classical variational framework. In this way, we need to introduce some spaces. First, we consider the space:

$$\boldsymbol{Y}_{p',\ell}(\Omega) = \Big\{ \boldsymbol{\psi} \in \boldsymbol{W}^{2,p'}_{\ell}(\Omega),\ \boldsymbol{\psi}|_\Gamma = 0, (\nabla \cdot \boldsymbol{\psi})|_\Gamma = 0 \Big\}.$$

The following lemma gives another characterization to the space $\boldsymbol{Y}_{p',\ell}(\Omega)$ very useful in the Green's formula defined in Corolllary 4.

Lemma 2. *We have the identity*

$$\boldsymbol{Y}_{p',\ell}(\Omega) = \left\{ \boldsymbol{\psi} \in \boldsymbol{W}^{2,p'}_{\ell}(\Omega),\ \boldsymbol{\psi}|_\Gamma = 0, \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{n}} \cdot \boldsymbol{n}|_\Gamma = 0 \right\} \tag{2}$$

and the range space of the normal derivative $\gamma_1 : \boldsymbol{Y}_{p',\ell}(\Omega) \longrightarrow \boldsymbol{W}^{1/p,p'}(\Gamma)$ *is*

$$\boldsymbol{Z}_{p'}(\Gamma) = \Big\{ \boldsymbol{z} \in \boldsymbol{W}^{1/p,p'}(\Gamma);\ \ \boldsymbol{z} \cdot \boldsymbol{n} = 0 \Big\}.$$

Proof. Let $\boldsymbol{u} \in \boldsymbol{W}_{\ell}^{2,p'}(\Omega)$ such that $\boldsymbol{u} = \boldsymbol{0}$ on Γ. Then $\operatorname{div} \boldsymbol{u} = (\partial \boldsymbol{u}/\partial \boldsymbol{n}) \cdot \boldsymbol{n}$ on Γ and the identity (2) holds. Moreover, it is clear that $\operatorname{Im}(\gamma_1) \subset \boldsymbol{Z}_{p'}(\Gamma)$. Conversely, let $\boldsymbol{\mu} \in \boldsymbol{Z}_{p'}(\Gamma)$. As Ω' is bounded, we can fix once for all a ball B_{R_o}, centered at the origin and with radius R_0, such that $\overline{\Omega'} \subset B_{R_o}$. Thus we have the existance of $\boldsymbol{u} \in \boldsymbol{W}^{2,p'}(\Omega_{R_0})$ such that $\boldsymbol{u} = \boldsymbol{0}$, $\partial \boldsymbol{u}/\partial \boldsymbol{n} = \boldsymbol{\mu}$ on $\Gamma \cup \partial B_{R_o}$ ($\Omega_{R_0} = \Omega \cap B_{R_0}$). The function $\boldsymbol{u}$ can be extended by zero outside B_{R_o} and owing to its boundary conditions on ∂B_{R_o}, the extended function, still denoted by $\boldsymbol{u}$, belongs to $\boldsymbol{W}_{\ell}^{2,p'}(\Omega)$, for any ℓ since its support is bounded. Since $\boldsymbol{\mu} \cdot \boldsymbol{n} = 0$ on Γ, we have $\boldsymbol{u} \in \boldsymbol{Y}_{p',\ell}(\Omega)$ and $\boldsymbol{\mu} \in \operatorname{Im}(\gamma_1)$. □

Secondly, we shall use the space:

$$\boldsymbol{T}_{r,p}^{\ell}(\Omega) = \left\{ \boldsymbol{v} \in \boldsymbol{W}_{-\ell}^{0,p}(\Omega);\ \Delta \boldsymbol{v} \in [\boldsymbol{X}_{r',p'}^{\ell}(\Omega)]' \right\},$$

equipped with the norm:

$$\|\boldsymbol{v}\|_{\boldsymbol{T}_{r,p}^{\ell}(\Omega)} = \|\boldsymbol{v}\|_{\boldsymbol{W}_{-\ell}^{0,p}(\Omega)} + \|\Delta \boldsymbol{v}\|_{[\boldsymbol{X}_{r',p'}^{\ell}(\Omega)]'}\,.$$

We also introduce the following space:

$$\boldsymbol{H}_{p,\ell}^{r}(\operatorname{div},\Omega) = \left\{ \boldsymbol{v} \in \boldsymbol{W}_{\ell-1}^{0,p}(\Omega);\ \nabla \cdot \boldsymbol{v} \in \boldsymbol{W}_{\ell-1}^{0,r}(\Omega) \right\}.$$

This space is equipped with the graph norm. The following lemma proves that the tangential trace of functions $\boldsymbol{v} \in \boldsymbol{T}_{r,p}^{\ell}(\Omega)$ belong to the dual space of $\boldsymbol{Z}_{p'}(\Gamma)$, wich is

$$(\boldsymbol{Z}_{p'}(\Gamma))' = \left\{ \mu \in \boldsymbol{W}^{-1/p,p}(\Gamma);\ \mu \cdot \boldsymbol{n} = 0 \right\}.$$

Observe that we can decompose $\boldsymbol{v}$ into its tangential and normal parts, that is: $\boldsymbol{v} = \boldsymbol{v}_\tau + (\boldsymbol{v} \cdot \boldsymbol{n})\boldsymbol{n}$. The proof of the following lemma is similar to the case of bounded domain (see [3]).

Lemma 3. *Suppose that* $3/2 < p < \infty$ *and* $1/p + 1/3 = 1/r$. *Then the space* $\mathcal{D}(\overline{\Omega})$ *is dense in* $\boldsymbol{T}_{r,p}^{0}(\Omega)$. *If in addition* $p \neq 3$, *we have* $\mathcal{D}(\overline{\Omega})$ *is dense in* $\boldsymbol{T}_{r,p}^{1}(\Omega)$.

Corollary 4. *Let* $3/2 < p < \infty$ *and* $1/p + 1/3 = 1/r$. *Then the mapping* $\gamma_\tau : \boldsymbol{v} \longrightarrow \boldsymbol{v}_\tau|_\Gamma$ *on the space* $\mathcal{D}(\overline{\Omega})$ *can be extended by continuity to a linear and continuous mapping, still denoted by* γ_τ, *from* $\boldsymbol{T}_{r,p}^{0}(\Omega)$ *into* $\boldsymbol{W}^{-1/p,p}(\Gamma)$, *and we have the Green formula: for any* $\boldsymbol{v} \in \boldsymbol{T}_{r,p}^{0}(\Omega)$ *and* $\boldsymbol{\psi} \in \boldsymbol{Y}_{p',0}(\Omega)$,

$$\langle \Delta \boldsymbol{v}, \boldsymbol{\psi} \rangle_{[\boldsymbol{X}_{r',p'}^{0}(\Omega)]' \times \boldsymbol{X}_{r',p'}^{0}(\Omega)} = \int_\Omega \boldsymbol{v} \cdot \Delta \boldsymbol{\psi}\, dx - \left\langle \boldsymbol{v}_\tau, \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{n}} \right\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma) \times \boldsymbol{W}^{1/p,p'}(\Gamma)}.$$

If in addition $p \neq 3$, *we have for any* $\boldsymbol{v} \in \boldsymbol{T}_{r,p}^{1}(\Omega)$ *and* $\boldsymbol{\psi} \in \boldsymbol{Y}_{p',1}(\Omega)$,

$$\langle \Delta \boldsymbol{v}, \boldsymbol{\psi} \rangle_{[\boldsymbol{X}_{r',p'}^{1}(\Omega)]' \times \boldsymbol{X}_{r',p'}^{1}(\Omega)} = \int_\Omega \boldsymbol{v} \cdot \Delta \boldsymbol{\psi}\, dx - \left\langle \boldsymbol{v}_\tau, \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{n}} \right\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma) \times \boldsymbol{W}^{1/p,p'}(\Gamma)}.$$

The following lemma gives a precise sense to the normal trace of functions $\boldsymbol{v} \in \boldsymbol{H}_{p,\ell}^{r}(\operatorname{div},\Omega)$ and the proof is very classical.

Lemma 5. *Let* Ω *be a Lipschitz open set in* $\mathbb{R}^3$. *Suppose that* $0 \leqslant 1/r - 1/p \leqslant 1/3$ *and* $\ell = 0$ *or* $\ell = 1$. *Then* *i) The space* $\mathcal{D}(\overline{\Omega})$ *is dense in* $\boldsymbol{H}^r_{p,\ell}(\mathrm{div},\Omega)$.

ii) The mapping $\gamma_n : \boldsymbol{v} \longrightarrow \boldsymbol{v}\cdot\boldsymbol{n}|_\Gamma$ *on the space* $\mathcal{D}(\overline{\Omega})$ *can be extended by continuity to a linear and continuous mapping, still denoted by* γ_n*, from* $\boldsymbol{H}^r_{p,\ell}(\mathrm{div},\Omega)$ *into* $\boldsymbol{W}^{-1/p,p}(\Gamma)$. *If in addition* $1/r - 1/p = 1/3$ *and* $3/2 < p < \infty$*, we have the following Green formula: for any* $\boldsymbol{v} \in \boldsymbol{H}^r_{p,\ell}(\mathrm{div},\Omega)$ *and* $\varphi \in W^{1,p'}_{1-\ell}(\Omega)$,

$$\int_\Omega \boldsymbol{v}\cdot\nabla\varphi\,dx + \int_\Omega \varphi\,\nabla\cdot\boldsymbol{v}\,dx = \langle \boldsymbol{v}\cdot\boldsymbol{n}, \varphi\rangle_{W^{-1/p,p}(\Gamma)\times W^{1/p,p'}(\Gamma)}\,.$$

§4. Very weak solutions in $\boldsymbol{L}^p(\Omega) \times W_0^{-1,p}(\Omega)$

In this section, we prove the existence and the uniqueness of very weak solutions to the Stokes problem via an argument of duality. We begin by specifying the meaning of very weak variational formulation.

Let

$$\boldsymbol{f} \in [\boldsymbol{X}^0_{r',p'}(\Omega)]',\ h \in L^r(\Omega) \quad \text{and} \quad \boldsymbol{g} \in \boldsymbol{W}^{-1/p,p}(\Gamma), \tag{3}$$

with

$$\frac{3}{2} < p < \infty \quad \text{and} \quad \frac{1}{p} + \frac{1}{3} = \frac{1}{r} \tag{A$_1$}$$

yielding $1 < r < 3$.

Definition 1 (Very weak solution for the Stokes problem)**.** Suppose that (A$_1$) is satisfied and let $\boldsymbol{f}$, h and $\boldsymbol{g}$ verifying (3). We say that $(\boldsymbol{u}, q) \in \boldsymbol{L}^p(\Omega) \times W_0^{-1,p}(\Omega)$ is a very weak solution of $(\mathcal{S})$ if the following equalities hold: For any $\boldsymbol{\varphi} \in \boldsymbol{Y}_{p',0}(\Omega)$ and $\pi \in W_0^{1,p'}(\Omega)$,

$$-\int_\Omega \boldsymbol{u}\cdot\Delta\boldsymbol{\varphi}\,dx - \langle q, \nabla\cdot\boldsymbol{\varphi}\rangle_{W_0^{-1,p}(\Omega)\times\mathring{W}_0^{1,p'}(\Omega)} = \langle \boldsymbol{f}, \boldsymbol{\varphi}\rangle_\Omega - \left\langle \boldsymbol{g}_\tau, \frac{\partial\boldsymbol{\varphi}}{\partial\boldsymbol{n}}\right\rangle_\Gamma, \tag{4}$$

$$\int_\Omega \boldsymbol{u}\cdot\nabla\pi\,dx = -\int_\Omega h\pi\,dx + \langle \boldsymbol{g}\cdot\boldsymbol{n}, \pi\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma)\times\boldsymbol{W}^{1/p,p'}(\Gamma)}\,, \tag{5}$$

where the dualities on Ω and Γ are defined by

$$\langle\cdot\,,\,\cdot\rangle_\Omega = \langle\cdot\,,\,\cdot\rangle_{[\boldsymbol{X}^0_{r',p'}(\Omega)]'\times\boldsymbol{X}^0_{r',p'}(\Omega)}\,, \quad \langle\cdot\,,\,\cdot\rangle_\Gamma = \langle\cdot\,,\,\cdot\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma)\times\boldsymbol{W}^{1/p,p'}(\Gamma)}\,.$$

Note that if $3/2 < p < \infty$ and $1/p + 1/3 = 1/r$, we have:

$$W_0^{1,p'}(\Omega) \hookrightarrow L^{r'}(\Omega) \quad \text{and} \quad \boldsymbol{Y}_{p',0}(\Omega) \hookrightarrow \boldsymbol{X}^0_{r',p'}(\Omega),$$

which means that all the brackets and integrals have a sense.

Proposition 6. *Under the assumptions of Definition 1, the following two statements are equivalent:*

i) $(\boldsymbol{u}, q) \in \boldsymbol{L}^p(\Omega) \times W_0^{-1,p}(\Omega)$ *is a very weak solution of* $(\mathcal{S})$,

ii) $(\boldsymbol{u}, q)$ *satisfies the system* $(\mathcal{S})$ *in the sense of distributions.*

Proof. **i) ⇒ ii)** Let $(\boldsymbol{u}, q) \in \boldsymbol{L}^p(\Omega) \times W_0^{-1,p}(\Omega)$ a very weak solution of $(\mathcal{S})$, then if we take $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{D}}(\Omega)$ and $\pi \in \mathcal{D}(\Omega)$ we can deduce by (4) and (5) that

$$-\Delta \boldsymbol{u} + \nabla \boldsymbol{q} = \boldsymbol{f} \ \text{ in } \ \Omega \quad \text{ and } \quad \nabla \cdot \boldsymbol{u} = h \ \text{ in } \ \Omega,$$

and that $\boldsymbol{u} \in \boldsymbol{T}^0_{r,p}(\Omega)$. Now let $\boldsymbol{\varphi} \in \boldsymbol{Y}_{p',0}(\Omega) \subset \boldsymbol{X}^0_{r',p'}(\Omega)$, then we have

$$\langle -\Delta \boldsymbol{u}, \boldsymbol{\varphi} \rangle_\Omega = \langle -\nabla \boldsymbol{q} + \boldsymbol{f}, \boldsymbol{\varphi} \rangle_\Omega \, .$$

As (A$_1$) is satisfied, it follows from Corollary 4 that

$$\langle -\Delta \boldsymbol{u}, \boldsymbol{\varphi} \rangle_\Omega = -\int_\Omega \boldsymbol{u} \cdot \Delta \boldsymbol{\varphi} \, dx + \left\langle \boldsymbol{u}_\tau, \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{n}} \right\rangle_\Gamma$$

and since $1/r - 1/p = 1/3$, it follows from Lemma 1 ii) that

$$\langle \nabla \boldsymbol{q}, \boldsymbol{\varphi} \rangle_\Omega = -\langle \boldsymbol{q}, \nabla \cdot \boldsymbol{\varphi} \rangle_{W_0^{-1,p}(\Omega) \times \mathring{W}_0^{1,p'}(\Omega)} \, .$$

Thus we have

$$-\int_\Omega \boldsymbol{u} \Delta \boldsymbol{\varphi} dx + \left\langle \boldsymbol{u}_\tau, \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{n}} \right\rangle_\Gamma = \langle \boldsymbol{q}, \nabla \cdot \boldsymbol{\varphi} \rangle_{W_0^{-1,p}(\Omega) \times \mathring{W}_0^{1,p'}(\Omega)} + \langle \boldsymbol{f}, \boldsymbol{\varphi} \rangle_\Omega \, ,$$

and we can deduce that for any $\boldsymbol{\varphi} \in \boldsymbol{Y}_{p',0}(\Omega)$

$$\left\langle \boldsymbol{u}_\tau, \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{n}} \right\rangle_\Gamma = \left\langle \boldsymbol{g}_\tau, \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{n}} \right\rangle_\Gamma .$$

Now let $\boldsymbol{\mu} \in \boldsymbol{W}^{1/p,p'}(\Gamma)$, then we have $\langle \boldsymbol{u}_\tau - \boldsymbol{g}_\tau, \boldsymbol{\mu} \rangle_\Gamma = \langle \boldsymbol{u}_\tau - \boldsymbol{g}_\tau, \boldsymbol{\mu}_\tau \rangle_\Gamma$. It is clear that $\boldsymbol{\mu}_\tau \in \boldsymbol{Z}_{p'}(\Gamma)$, thus it follows from Lemma 2 that there exists $\boldsymbol{\varphi} \in \boldsymbol{Y}_{p',0}(\Omega)$ such that $\partial \boldsymbol{\varphi} / \partial \boldsymbol{n} = \boldsymbol{\mu}_\tau$ on Γ. Then from this we can deduce that $\boldsymbol{u}_\tau = \boldsymbol{g}_\tau$ in $\boldsymbol{W}^{-1/p,p}(\Gamma)$. From the equation $\nabla \cdot \boldsymbol{u} = h$, we deduce that $\boldsymbol{u} \in \boldsymbol{H}^r_{p,1}(\text{div}, \Omega)$, then it follows from Lemma 5 ii), that for any $\boldsymbol{\pi} \in \boldsymbol{W}_0^{1,p'}(\Omega)$,

$$\langle \boldsymbol{u} \cdot \boldsymbol{n}, \boldsymbol{\pi} \rangle_\Gamma = \langle \boldsymbol{g} \cdot \boldsymbol{n}, \boldsymbol{\pi} \rangle_\Gamma \, .$$

Consequently $\boldsymbol{u} \cdot \boldsymbol{n} = \boldsymbol{g} \cdot \boldsymbol{n}$ in $\boldsymbol{W}^{-1/p,p}(\Gamma)$ and finally $\boldsymbol{u} = \boldsymbol{g}$ on Γ.

ii)⇒i) We suppose that $(\boldsymbol{u}, q)$ satisfies the system $(\mathcal{S})$ in the sense of distributions. Then for any $\boldsymbol{\varphi} \in \boldsymbol{Y}_{p',0}(\Omega) \hookrightarrow \boldsymbol{X}^0_{r',p'}(\Omega)$ we have

$$\langle -\Delta \boldsymbol{u}, \boldsymbol{\varphi} \rangle_\Omega = \langle \boldsymbol{f} - \nabla \, q, \boldsymbol{\varphi} \rangle_\Omega \, ,$$

Using Corollary 4 and Lemma 1 ii) we prove (4).

Now from the equation $\nabla \cdot \boldsymbol{u} = h$, we can deduce that for any $\pi \in W_0^{1,p'}(\Omega)$

$$\int_\Omega \pi \nabla \cdot \boldsymbol{u} \, dx = \int_\Omega h \pi \, dx,$$

this integral has a sense because we have $W_0^{1,p'}(\Omega) \hookrightarrow L^{r'}(\Omega)$. Using Lemma 5 ii) we deduce (5).

□

Before stating the theorem of the existense and the uniqueness of the very weak solution for Stokes problem, we need to introduce the following null spaces for $\alpha \in \{-1, 0, 1\}$ and $k \in \{0, 1, 2\}$:

$$\mathcal{N}_\alpha^{k,p}(\Omega) = \Big\{ (\boldsymbol{u}, \pi) \in \boldsymbol{W}_\alpha^{k,p}(\Omega) \times W_\alpha^{k-1,p}(\Omega);\ T(\boldsymbol{u}, \pi) = (\boldsymbol{0}, \boldsymbol{0}) \ \text{ in } \ \Omega \quad \text{and} \quad \boldsymbol{u}|_\Gamma = 0 \Big\},$$

with

$$T(\boldsymbol{u}, \pi) = (-\Delta\, \boldsymbol{u} + \nabla\, \pi, \operatorname{div} u).$$

If $p \notin \{3/2, 3\}$, we can prove that $\mathcal{N}_1^{2,p}(\Omega) = \mathcal{N}_0^{1,p}(\Omega) = \mathcal{N}_{-1}^{0,p}(\Omega)$. Note that if $\boldsymbol{u} \in \boldsymbol{W}_{-1}^{0,p}(\Omega)$ and $-\Delta\, \boldsymbol{u} + \nabla\, \pi = \boldsymbol{0}$ in Ω with $\pi \in W_{-1}^{-1,p}(\Omega)$, then the tangential component $\boldsymbol{u}_\tau$ of $\boldsymbol{u}$ belongs to $\boldsymbol{W}^{-1/p,p}(\Gamma)$ and if $\operatorname{div} \boldsymbol{u} = 0$ in Ω, then $\boldsymbol{u} \cdot \boldsymbol{n} \in \boldsymbol{W}^{-\frac{1}{p},p}(\Gamma)$. That means that $\boldsymbol{u} = \boldsymbol{0}$ on Γ makes sense.

Theorem 7. *Let Ω be an exterior domain with $C^{1,1}$ boundary and let p and r satisfy* (A_1) *and let $\boldsymbol{f}$, h and $\boldsymbol{g}$ satisfying* (3). *Then the Stokes problem* ($\mathcal{S}$) *has exactly one solution $\boldsymbol{u} \in \boldsymbol{L}^p(\Omega)$ and $\boldsymbol{q} \in \boldsymbol{W}_0^{-1,p}(\Omega)$ if and only if for any $(\boldsymbol{v}, \eta) \in \mathcal{N}_0^{2,p'}(\Omega)$:*

$$\langle \boldsymbol{f}, \boldsymbol{v} \rangle - \langle h, \eta \rangle + \langle \boldsymbol{g}, (\eta I - \nabla \boldsymbol{v}) \cdot \boldsymbol{n} \rangle_\Gamma = 0.$$

Moreover, there exists a constant $C > 0$ depending only on p and Ω such that:

$$\|\boldsymbol{u}\|_{\boldsymbol{L}^p(\Omega)} + \|q\|_{\boldsymbol{W}_0^{-1,p}(\Omega)} \leq C(\|\boldsymbol{f}\|_{[\boldsymbol{X}_{r',p'}^0(\Omega)]'} + \|h\|_{L^r(\Omega)} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)} \,.$$

Proof. It remains to consider the equivalent problem: Find $(\boldsymbol{u}, \boldsymbol{q}) \in \boldsymbol{L}^p(\Omega) \times \boldsymbol{W}_0^{-1,p}(\Omega)$ such that for any $\boldsymbol{w} \in \boldsymbol{Y}_{p',0}(\Omega)$ and $\pi \in W_0^{1,p'}(\Omega)$ it holds:

$$\int_\Omega \boldsymbol{u} \cdot (-\Delta \boldsymbol{w} + \nabla \pi)\, dx - \langle q, \nabla \cdot \boldsymbol{w} \rangle_{W_0^{-1,p}(\Omega) \times \mathring{W}_0^{1,p'}(\Omega)} = \langle \boldsymbol{f}, \boldsymbol{w} \rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{n}} \Big\rangle_\Gamma + \langle \boldsymbol{g} \cdot \boldsymbol{n}, \pi \rangle_\Gamma - \int_\Omega h\, \pi\, dx.$$

Let T be a linear form defined by:

$$\begin{aligned} T : \boldsymbol{L}^{p'}(\Omega) \times \mathring{W}_0^{1,p'}(\Omega) &\longrightarrow \mathbb{R} \\ (\boldsymbol{F}, \varphi) &\longmapsto \langle \boldsymbol{f}, \boldsymbol{w} \rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{n}} \Big\rangle_\Gamma + \langle \boldsymbol{g} \cdot \boldsymbol{n}, \pi \rangle_\Gamma - \int_\Omega h\, \pi\, dx, \end{aligned}$$

with $(\boldsymbol{w}, \pi) \in \boldsymbol{W}_0^{2,p'}(\Omega) \times \boldsymbol{W}_0^{1,p'}(\Omega)$ is a solution of the following Stokes problem:

$$-\Delta \boldsymbol{w} + \nabla\, \pi = \boldsymbol{F} \quad \text{and} \quad \nabla \cdot \boldsymbol{w} = \varphi \ \text{ in } \ \Omega, \quad \boldsymbol{w} = 0 \ \text{ on } \ \Gamma,$$

and satisfying the following estimate (see [1, Theorem 3.1]):

$$\inf_{(\boldsymbol{v},\eta) \in \mathcal{N}_0^{2,p'}(\Omega)} \Big(\|\boldsymbol{w} + \boldsymbol{v}\|_{\boldsymbol{W}_0^{2,p'}(\Omega)} + \|\pi + \eta\|_{W_0^{1,p'}(\Omega)} \Big) \leqslant C\Big(\|\boldsymbol{F}\|_{\boldsymbol{L}^{p'}(\Omega)} + \|\varphi\|_{W_0^{1,p'}(\Omega)} \Big). \tag{6}$$

Then we have for any pair $(\boldsymbol{F}, \varphi) \in \boldsymbol{L}^{p'}(\Omega) \times \mathring{W}_0^{1,p'}(\Omega)$ and for any $(\boldsymbol{v}, \eta) \in \mathcal{N}_0^{2,p'}(\Omega)$

$$\begin{aligned} &\Big| \langle \boldsymbol{f}, \boldsymbol{w} \rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{n}} \Big\rangle_\Gamma + \langle \boldsymbol{g} \cdot \boldsymbol{n}, \pi \rangle_\Gamma - \int_\Omega h\, \pi\, dx \Big| \\ &\quad = \Big| \langle \boldsymbol{f}, \boldsymbol{w} + \boldsymbol{v} \rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial (\boldsymbol{w} + \boldsymbol{v})}{\partial \boldsymbol{n}} \Big\rangle_\Gamma + \langle \boldsymbol{g} \cdot \boldsymbol{n}, \pi + \eta \rangle_\Gamma - \int_\Omega h\, (\pi + \eta)\, dx \Big| \\ &\quad \leq C\Big(\|\boldsymbol{f}\|_{[\boldsymbol{X}_{r',p'}^0(\Omega]'} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{-1/p,p}(\Omega)} + \|h\|_{L^r(\Omega)} \Big)\Big(\|\boldsymbol{w} + \boldsymbol{v}\|_{\boldsymbol{W}_0^{2,p'}(\Omega)} + \|\pi + \eta\|_{W_0^{1,p'}(\Omega)} \Big). \end{aligned}$$

Using (6), we prove that

$$\left|\langle \boldsymbol{f}, \boldsymbol{w}\rangle_\Omega - \left\langle \boldsymbol{g}_\tau, \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{n}}\right\rangle_\Gamma + \langle \boldsymbol{g}\cdot\boldsymbol{n}, \pi\rangle_\Gamma - \int_\Omega h\,\pi\,dx\right|$$
$$\leq C\Big(\|\boldsymbol{f}\|_{[\boldsymbol{X}^0_{r',p'}(\Omega]'} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{-1/p,p}(\Omega)} + \|h\|_{L^r(\Omega)}\Big)\Big(\|\boldsymbol{F}\|_{\boldsymbol{L}^{p'}(\Omega)} + \|\varphi\|_{W^{1,p'}_0(\Omega)}\Big),$$

from this we can deduce that the linear form T is continuous on $\boldsymbol{L}^{p'}(\Omega)\times W^{1,p'}_0(\Omega)$ and according to the Riesz' Theorem we deduce that there exists a unique $(\boldsymbol{u}, q) \in \boldsymbol{L}^p(\Omega)\times W^{-1,p}_0(\Omega)$ solution of $(\mathcal{S})$ satisfying the appropriate estimate. □

§5. Very weak solutions in $\boldsymbol{W}^{0,p}_{-1}(\Omega)\times W^{-1,p}_{-1}(\Omega)$

Here, we are interested in the case of the following assumptions:

$$\boldsymbol{f}\in[\boldsymbol{X}^1_{r',p'}(\Omega)]',\ h\in W^{0,r}_{-1}(\Omega)\quad\text{and}\quad \boldsymbol{g}\in\boldsymbol{W}^{-1/p,p}(\Gamma), \tag{7}$$

with

$$\frac{3}{2} < p < \infty,\quad p\neq 3\quad\text{and}\quad \frac{1}{p}+\frac{1}{3}=\frac{1}{r}, \tag{A$_2$}$$

yielding $1 < r < 3$.

Definition 2 (Very weak solution for the Stokes problem)**.** Suppose that (A$_2$) is satisfied and let $\boldsymbol{f}$, h and $\boldsymbol{g}$ satisfying (7). We say that $(\boldsymbol{u}, q)\in\boldsymbol{W}^{0,p}_{-1}(\Omega)\times W^{-1,p}_{-1}(\Omega)$ is a very weak solution of $(\mathcal{S})$ if the following equalities hold: For any $\boldsymbol{\varphi}\in\boldsymbol{Y}_{p',1}(\Omega)$ and $\pi\in W^{1,p'}_1(\Omega)$,

$$-\int_\Omega \boldsymbol{u}\cdot\Delta\boldsymbol{\varphi}\,dx - \langle q, \nabla\cdot\boldsymbol{\varphi}\rangle_{W^{-1,p}_{-1}(\Omega)\times\mathring{W}^{1,p'}_1(\Omega)} = \langle\boldsymbol{f},\boldsymbol{\varphi}\rangle_\Omega - \left\langle\boldsymbol{g}_\tau, \frac{\partial\boldsymbol{\varphi}}{\partial\boldsymbol{n}}\right\rangle_\Gamma,$$

$$\int_\Omega \boldsymbol{u}\cdot\nabla\pi\,dx = -\int_\Omega h\pi dx + \langle\boldsymbol{g}\cdot\boldsymbol{n},\pi\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma)\times\boldsymbol{W}^{1/p,p'}(\Gamma)},$$

where the dualities on Ω and Γ are defined by:

$$\langle\cdot,\cdot\rangle_\Omega = \langle\cdot,\cdot\rangle_{[\boldsymbol{X}^1_{r',p'}(\Omega)]'\times\boldsymbol{X}^1_{r',p'}(\Omega)},\quad \langle\cdot,\cdot\rangle_\Gamma = \langle\cdot,\cdot\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma)\times\boldsymbol{W}^{1/p,p'}(\Gamma)}.$$

Note that if $3/2 < p < \infty$ and $1/p + 1/3 = 1/r$, we have:

$$W^{1,p'}_1(\Omega)\hookrightarrow W^{0,r'}_1(\Omega),\quad\text{and}\quad \boldsymbol{Y}_{p',1}(\Omega)\hookrightarrow\boldsymbol{X}^1_{r',p'}(\Omega),$$

which means that all the brackets and integrals have a sense. As previously we prove under the assumption (A$_2$), that if $\boldsymbol{f}$, h and $\boldsymbol{g}$ satisfying (7), then $(\boldsymbol{u}, q)\in\boldsymbol{W}^{0,p}_{-1}(\Omega)\times W^{-1,p}_{-1}(\Omega)$ is a very weak solution of $(\mathcal{S})$ if and only if $(\boldsymbol{u}, q)$ satisfy the system $(\mathcal{S})$ in the sense of distributions.

Theorem 8. *Let Ω be an exterior domain with $C^{1,1}$ boundary and let p and r satisfy* (A$_2$) *and let $\boldsymbol{f}$, h and $\boldsymbol{g}$ satisfying* (7)*. Then the Stokes problem* $(\mathcal{S})$ *has a solution $\boldsymbol{u}\in\boldsymbol{W}^{0,p}_{-1}(\Omega)$ and $\boldsymbol{q}\in\boldsymbol{W}^{-1,p}_{-1}(\Omega)$ if and only if, for any $(\boldsymbol{v},\eta)\in\mathcal{N}^{2,p'}_1(\Omega)$:,*

$$\langle\boldsymbol{f},\boldsymbol{v}\rangle - \langle h,\eta\rangle + \langle\boldsymbol{g},(\eta\boldsymbol{I}-\nabla\boldsymbol{v})\cdot\boldsymbol{n}\rangle_\Gamma = 0.$$

In $\boldsymbol{W}^{0,p}_{-1}(\Omega) \times W^{-1,p}_{-1}(\Omega)$*, each solution is unique up to an element of* $\mathcal{N}^{0,p}_{-1}(\Omega)$ *and there exists a constant* $C > 0$ *depending only on* p *and* Ω *such that*

$$\inf_{(\boldsymbol{v},\eta)\in\mathcal{N}^{1,p}_{0}(\Omega)}\Big(\|\boldsymbol{u}+\boldsymbol{v}\|_{\boldsymbol{W}^{0,p}_{-1}(\Omega)} + \|q+\eta\|_{\boldsymbol{W}^{-1,p}_{-1}(\Omega)}\Big) \leq C\Big(\|\boldsymbol{f}\|_{[\boldsymbol{X}^{1}_{r',p'}(\Omega)]'} + \|h\|_{W^{0,r}_{-1}(\Omega)} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)}\Big).$$

Proof. It remains to consider the equivalent problem: Find $(\boldsymbol{u},\boldsymbol{q}) \in \boldsymbol{W}^{0,p}_{-1}(\Omega)\times\boldsymbol{W}^{-1,p}_{-1}(\Omega)$ such that for any $\boldsymbol{w} \in \boldsymbol{Y}_{p',0}(\Omega)$ and $\pi \in W^{1,p'}_{1}(\Omega)$ it holds:

$$\int_\Omega \boldsymbol{u}\cdot(-\Delta\boldsymbol{w} + \nabla\pi)\,dx - \langle q, \nabla\cdot\boldsymbol{w}\rangle_{W^{-1,p}_{-1}(\Omega)\times\mathring{W}^{1,p'}_{1}(\Omega)}$$
$$= \langle \boldsymbol{f}, \boldsymbol{w}\rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial\boldsymbol{w}}{\partial\boldsymbol{n}}\Big\rangle_\Gamma + \langle \boldsymbol{g}\cdot\boldsymbol{n}, \pi\rangle_\Gamma - \int_\Omega h\,\pi\,dx.$$

Let T be a linear form defined from $\Big(\boldsymbol{W}^{0,p'}_{1}(\Omega)\times\mathring{W}^{1,p'}_{1}(\Omega))\perp\mathcal{N}^{1,p}_{0}(\Omega)\Big)$ onto $\mathbb{R}$ by:

$$T(\boldsymbol{F},\varphi) = \langle \boldsymbol{f}, \boldsymbol{w}\rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial\boldsymbol{w}}{\partial\boldsymbol{n}}\Big\rangle_\Gamma + \langle \boldsymbol{g}\cdot\boldsymbol{n}, \pi\rangle_\Gamma - \int_\Omega h\,\pi\,dx,$$

with $(\boldsymbol{w},\pi) \in \boldsymbol{W}^{2,p'}_{1}(\Omega)\times\boldsymbol{W}^{1,p'}_{1}(\Omega)$ is a solution of the following Stokes problem:

$$-\Delta\boldsymbol{w} + \nabla\pi = \boldsymbol{F} \quad\text{and}\quad \nabla\cdot\boldsymbol{w} = \varphi \ \text{ in } \Omega, \quad \boldsymbol{w} = 0 \ \text{ on } \Gamma,$$

and satisfying the following estimate (see [1, Theorem 3.1]):

$$\inf_{(\boldsymbol{v},\eta)\in\mathcal{N}^{2,p'}_{1}(\Omega)}\Big(\|\boldsymbol{w}+\boldsymbol{v}\|_{\boldsymbol{W}^{2,p'}_{1}(\Omega)} + \|\pi+\eta\|_{W^{1,p'}_{1}(\Omega)}\Big) \leqslant C\Big(\|\boldsymbol{F}\|_{\boldsymbol{W}^{0,p'}_{1}(\Omega)} + \|\varphi\|_{W^{1,p'}_{1}(\Omega)}\Big). \tag{8}$$

Then for any pair $(\boldsymbol{F},\varphi) \in (\boldsymbol{W}^{0,p'}_{1}(\Omega)\times\mathring{W}^{1,p'}_{1}(\Omega))\perp\mathcal{N}^{1,p}_{0}(\Omega)$ and for any $(\boldsymbol{v},\eta)\in\mathcal{N}^{2,p'}_{1}(\Omega)$

$$\Big|\langle \boldsymbol{f}, \boldsymbol{w}\rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial\boldsymbol{w}}{\partial\boldsymbol{n}}\Big\rangle_\Gamma + \langle \boldsymbol{g}\cdot\boldsymbol{n}, \pi\rangle_\Gamma - \int_\Omega h\,\pi\,dx\Big|$$
$$= \Big|\langle \boldsymbol{f}, \boldsymbol{w}+\boldsymbol{v}\rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial(\boldsymbol{w}+\boldsymbol{v})}{\partial\boldsymbol{n}}\Big\rangle_\Gamma + \langle \boldsymbol{g}\cdot\boldsymbol{n}, \pi+\eta\rangle_\Gamma - \int_\Omega h\,(\pi+\eta)\,dx\Big|$$
$$\leq C\Big(\|\boldsymbol{f}\|_{[\boldsymbol{X}^{1}_{r',p'}(\Omega]'} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{-1/p,p}(\Omega)} + \|h\|_{W^{0,r}_{-1}(\Omega)}\Big)\Big(\|\boldsymbol{w}+\boldsymbol{v}\|_{\boldsymbol{W}^{2,p'}_{1}(\Omega)} + \|\pi+\eta\|_{W^{1,p'}_{1}(\Omega)}\Big).$$

Using (8), we prove that

$$\Big|\langle \boldsymbol{f}, \boldsymbol{w}\rangle_\Omega - \Big\langle \boldsymbol{g}_\tau, \frac{\partial\boldsymbol{w}}{\partial\boldsymbol{n}}\Big\rangle_\Gamma + \langle \boldsymbol{g}\cdot\boldsymbol{n}, \pi\rangle_\Gamma - \int_\Omega h\,\pi\,dx\Big|$$
$$\leq C\left(\|\boldsymbol{f}\|_{[\boldsymbol{X}_{p'}(\Omega]'} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{-1/p,p}(\Omega)} + \|h\|_{W^{0,r}_{-1}(\Omega)}\right)\left(\|\boldsymbol{F}\|_{\boldsymbol{W}^{0,p'}_{1}(\Omega)} + \|\varphi\|_{W^{1,p'}_{1}(\Omega)}\right),$$

From this we derive that the linear form T is continuous on $\Big(\boldsymbol{W}^{0,p'}_{1}(\Omega)\times\mathring{W}^{1,p'}_{1}(\Omega)\perp\mathcal{N}^{1,p}_{0}(\Omega)\Big)$ and according to the Riesz' Theorem, we deduce that there exists $(\boldsymbol{u},q) \in (\boldsymbol{W}^{0,p}_{-1}(\Omega)\times W^{-1,p}_{-1}(\Omega))$ solution of $(\mathcal{S})$ unique up to an element of $\mathcal{N}^{1,p}_{0}(\Omega)$ and satisfying the appropriate estimate. □

References

[1] Alliot, F., and Amrouche, F. Weak solutions for exterior Stokes problem in weighted Sobolev spaces. *Mathematical Methods in the Applied Sciences 23* (2000), 575–600.

[2] Amrouche, C., Girault, V., and Giroire, J. Dirichlet and Neumann exterior problems for the n-dimensional Laplace operator. An approach in weighted Sobolev spaces. *J. Math. Pures Appl. 76*, 1 (1997), 55–81.

[3] Amrouche, C., and Rodriguez-Bellido, M. A. Stokes, Ossen and Navier-Stokes equations with singular data. *Archive for Rational Mechanics and Analysis 199*, 2 (2011), 597–651.

[4] Farwig, R., Kozono, H., and Sohr, H. Very weak solutions of the Navier-Stokes equations in exterior domains with nonhomogeneous data. *J. Math. Soc. Japan 59*, 1 (2007), 127–150.

Chérif Amrouche and Mohamed Meslameni
Laboratoire de Mathématiques et de leurs Applications, CNRS UMR 5142
Université de Pau et des Pays de l'Adour
64013 Pau, FRANCE,
`cherif.amrouche@univ-pau.fr` and `mohamed.meslameni@univ-pau.fr`

Monografías Matemáticas *García de Galdeano* **37**, 13–23 (2012)

$\mathbf{Q}$-Resolutions and Intersection Numbers

Enrique Artal Bartolo, Jorge Martín-Morales,
Jorge Ortigas-Galindo

Abstract. In this paper we introduce the notion of embedded $\mathbf{Q}$-resolution, which is a special class of toric resolutions, and explain briefly how to compute it for plane curve singularities and obtain invariants from them. The main difference with standard resolutions is that we allow both the ambient space and the hypersurface to contain quotient singularities in some mild conditions. We develop an intersection theory on V-manifolds that allows us to calculate the intersection numbers of the exceptional divisors of the weighted blow-ups. An illustrative example is given at the end showing that the intersection matrix has the expected properties.

Keywords: Quotient singularity, intersection number, embedded $\mathbf{Q}$-resolution.

AMS classification: 32S25, 32S45.

Introduction

In Singularity Theory, resolution is the most important tool. In the embedded case, the starting point is a singular hypersurface; after a sequence of suitable blow-ups this hypersurface is replaced by a long list of smooth hypersurfaces (the strict transform and the exceptional divisors) which intersect in the simplest way (at any point we see coordinate hyperplanes for suitable local coordinates). This process can be very expensive from the computational point of view and, moreover, only a few amount of the obtained data is used for the understanding of the singularity. The experimental work shows that most of these data can be recovered if we allow some mild singularities to survive in the process (the quotient singularities). These *partial* resolutions, denoted as $\mathbf{Q}$-resolutions, can be obtained as a sequence of weighted blow-ups and their computational complexity is extremely lower compared with standard resolutions. Moreover, the process is optimal in the sense that we do not obtain useless data. To do this, we develop an intersection theory on varieties with quotient singularities and study weighted blow-ups at points. By using these tools we will be able to get a big amount of information about the singularity.

The paper is organized as follows. In §1 we give a general presentation of varieties with quotient singularities and list their basic properties; we introduce the main example, the weighted projective spaces. In §2 we describe the rational Weil and Cartier divisors on V-varieties and §3 introduces their intersection numbers. We discuss briefly in §4 the concepts of weighted blow-ups and embedded $\mathbf{Q}$-resolutions and their relationship with intersection theory. Finally an example on how to use $\mathbf{Q}$-resolutions to compute intersection numbers is given. Detailed proofs and further application will appear in a forthcoming work, see [1, 2, 6].

§1. V-manifolds and quotient singularities

Definition 1. A V-manifold of dimension n is a complex analytic space which admits an open covering $\{U_i\}$ such that U_i is analytically isomorphic to B_i/G_i where $B_i \subset \mathbb{C}^n$ is an open ball and G_i is a finite subgroup of $GL(n, \mathbb{C})$.

V-manifolds were introduced in [9] and have the same homological properties over $\mathbb{Q}$ as manifolds. For instance, they admit a Poincaré duality if they are compact and carry a pure Hodge structure if they are compact and Kähler, see [3]. They have been classified locally by Prill [8]. In this paper special attention is paid to V-manifolds where all groups G_i are abelian. In particular, the following notation is used.

Let $G := \mu_{d_1} \times \cdots \times \mu_{d_r}$ be an arbitrary finite abelian group written as a product of finite cyclic groups, that is, μ_{d_i} is the cyclic group of d_i-th roots of unity. Consider a matrix of weight vectors $A := (a_{ij})_{i,j} = [a_1 \,|\, \cdots \,|\, a_n] \in Mat(r \times n, \mathbb{Z})$ and the action

$$\begin{aligned}(\mu_{d_1} \times \cdots \times \mu_{d_r}) \times \mathbb{C}^n &\longrightarrow \mathbb{C}^n,\\ ((\xi_{d_1}, \ldots, \xi_{d_r}), (x_1, \ldots, x_n)) &\longmapsto (\xi_{d_1}^{a_{11}} \cdots \xi_{d_r}^{a_{r1}}\, x_1, \, \ldots \,, \xi_{d_1}^{a_{1n}} \cdots \xi_{d_r}^{a_{rn}}\, x_n).\end{aligned}$$

Note that the i-th row of the matrix A can be considered modulo d_i. The set of all orbits $\mathbb{C}^n/G$ is called *(cyclic) quotient space of type* $(d; A)$ and is denoted by

$$X(d; A) := X\left(\begin{array}{c|ccc} d_1 & a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ d_r & a_{r1} & \cdots & a_{rn} \end{array}\right).$$

The following result shows that the family of varieties which can locally be written like $X(d, A)$ is exactly the same as the family of V-manifolds with abelian quotient singularities.

Lemma 1. *Let G be a finite abelian subgroup of $GL(n, \mathbb{C})$. Then $\mathbb{C}^n/G$ is isomorphic to some quotient space $X(d; A)$.*

We finish this section with one of the classical examples of V-manifold, cf. [4], the *weighted projective spaces*.

Let $\omega = (q_0, \ldots, q_n)$ be a weight vector, that is, a finite set of positive integers. There is a natural action of the multiplicative group $\mathbb{C}^*$ on $\mathbb{C}^{n+1} \setminus \{0\}$ given by

$$(x_0, \ldots, x_n) \longmapsto (t^{q_0} x_0, \ldots, t^{q_n} x_n).$$

The set of orbits $(\mathbb{C}^{n+1} \setminus \{0\})/\mathbb{C}^*$ under this action is denoted by $\mathbb{P}^n_\omega$ and is called the *weighted projective space* of type ω. The class of a nonzero element $(x_0, \ldots, x_n) \in \mathbb{C}^{n+1}$ is denoted by $[x_0 : \ldots : x_n]_\omega$ and the weight vector is omitted if no ambiguity seems likely to arise. When $(q_0, \ldots, q_n) = (1, \ldots, 1)$ one obtains the usual projective space and the weight vector is always omitted.

As in the classical case, the weighted projective spaces can be endowed with an analytic structure. However, in general they contain cyclic quotient singularities. Consider the decomposition $\mathbb{P}^n_\omega = U_0 \cup \cdots \cup U_n$, where U_i is the open set consisting of all elements $[x_0 : \ldots : x_n]_\omega$ with $x_i \neq 0$. The map

$$\widetilde{\psi}_0 : \mathbb{C}^n \longrightarrow U_0, \quad \widetilde{\psi}_0(x_1, \ldots, x_n) := [1 : x_1 : \ldots : x_n]_\omega$$

is clearly a surjective analytic map but it is not a chart since injectivity fails. In fact, $[1 : x_1 : \ldots : x_n]_\omega = [1 : x'_1 : \ldots, x'_n]_\omega$ if and only if there exists $\xi \in \mu_{q_0}$ such that $x'_i = \xi^{q_i} x_i$ for all $i = 1, \ldots, n$. Hence the map above induces the isomorphism

$$\begin{aligned}\psi_0 : X(q_0; q_1, \ldots, q_n) &\longrightarrow U_0,\\ [(x_1, \ldots, x_n)] &\longmapsto [1 : x_1 : \ldots : x_n]_\omega.\end{aligned}$$

Analogously, $X(q_i; q_0, \ldots, \widehat{q_i}, \ldots, q_n) \cong U_i$ under the obvious analytic map. Therefore $\mathbb{P}^n_\omega$ is an analytic space with cyclic quotient singularities as claimed.

§2. Cartier and Weil divisors on V-manifolds: $\mathbb{Q}$-divisors

Given X a complex analytic surface, the intersection product $D \cdot E$ is well understood whenever D is a compact Weil divisor on X and E is a Cartier divisor on X. Over varieties with quotient singularities the notion of Cartier and Weil divisor coincide after tensoring with $\mathbb{Q}$, see Theorem 2 below. A rational intersection theory can be defined on this kind of varieties.

Let us start with X an irreducible complex analytic variety. As usual, consider $\mathcal{O}_X$ the structure sheaf of X and $\mathcal{K}_X$ the sheaf of total quotient rings of $\mathcal{O}_X$. Denote by $\mathcal{K}^*_X$ the (multiplicative) sheaf of invertible elements in $\mathcal{K}_X$. Similarly $\mathcal{O}^*_X$ is the sheaf of invertible elements in $\mathcal{O}_X$.

Definition 2. A *Cartier divisor* on X is a global section of the sheaf $\mathcal{K}^*_X/\mathcal{O}^*_X$, that is, an element in $\Gamma(X, \mathcal{K}^*_X/\mathcal{O}^*_X) = H^0(X, \mathcal{K}^*_X/\mathcal{O}^*_X)$. Any Cartier divisor can be represented by giving an open covering $\{U_i\}_{i\in I}$ of X and, for all $i \in I$, an element $f_i \in \Gamma(U_i, \mathcal{K}^*_X)$ such that

$$\frac{f_i}{f_j} \in \Gamma(U_i \cap U_j, \mathcal{O}^*_X), \quad \forall i, j \in I.$$

Two systems $\{(U_i, f_i)\}_{i\in I}$ and $\{(V_j, g_j)\}_{j\in J}$ represent the same Cartier divisor if and only if on $U_i \cap V_j$, f_i and g_j differ by a multiplicative factor in $\mathcal{O}_X(U_i \cap V_j)^*$. The abelian group of Cartier divisors on X is denoted by $\mathrm{CaDiv}(X)$. If $D := \{(U_i, f_i)\}_{i\in I}$ and $E := \{(V_j, g_j)\}_{j\in J}$ then $D + E = \{(U_i \cap V_j, f_i g_j)\}_{i\in I, j\in J}$.

Definition 3. A *Weil divisor* on X is a locally finite linear combination with integral coefficients of irreducible subvarieties of codimension one. The abelian group of Weil divisors on X is denoted by $\mathrm{WeDiv}(X)$.

Let $V \subset X$ be an irreducible subvariety of codimension one. It corresponds to a prime ideal in the ring of sections of any local complex model space meeting V. The *local ring of X along V*, denoted by $\mathcal{O}_{X,V}$, is the localization of such ring of sections at the corresponding prime ideal; it is a one-dimensional local domain. For a given $f \in \mathcal{O}_{X,V}$ define $\mathrm{ord}_V(f)$ to be $\mathrm{ord}_V(f) := \mathrm{length}_{\mathcal{O}_{X,V}} (\mathcal{O}_{X,V}/\langle f\rangle)$, where $\mathrm{length}_{\mathcal{O}_{X,V}}$ denotes the length as an $\mathcal{O}_{X,V}$-module.

Now if D is a Cartier divisor on X, one writes $\mathrm{ord}_V(D) = \mathrm{ord}_V(f_i)$ where f_i is a local equation of D on any open set U_i with $U_i \cap V \neq \emptyset$. This is well defined since f_i is uniquely determined up to multiplication by units and the order function is a homomorphism. Define the *associated Weil divisor* of a Cartier divisor D by setting

$$T_X : \mathrm{CaDiv}(X) \longrightarrow \mathrm{WeDiv}(X) : \quad D \longmapsto \sum_{V\subset X} \mathrm{ord}_V(D) \cdot [V],$$

where the sum is taken over all codimension one irreducible subvarieties V of X. By the additivity of the order function, the mapping T_X is a homomorphism of abelian groups.

Example 1. Let X be the surface in $\mathbb{C}^3$ defined by the equation $z^2 = xy$. The line $V = \{x = z = 0\}$ defines a Weil divisor which is not a Cartier divisor. The associated Weil divisor of $\{(X, x)\}$ is $T_X(\{(X, x)\}) = \sum_{Z\subset X} \operatorname{ord}_Z(x) \cdot [Z] = 2[V]$. Thus $[V]$ is principal as an element in $\operatorname{WeDiv}(X) \otimes_{\mathbb{Z}} \mathbb{Q}$ and corresponds to the $\mathbb{Q}$-Cartier divisor $\frac{1}{2}\{(X, x)\}$.

This fact can be interpreted as follows. First note that identifying our surface X with $X(2; 1, 1)$ under $[(x, y)] \mapsto (x^2, y^2, xy)$, the previous Weil divisor corresponds to $D = \{x = 0\}$. Although $f = x$ defines a zero set on $X(2; 1, 1)$, it does not induce a function on the quotient space. However, $x^2 : X(2; 1, 1) \to \mathbb{C}$ is a well-defined function and gives the same zero set as f. Hence as $\mathbb{Q}$-Cartier divisors one writes $D = \frac{1}{2}\{(X(2; 1, 1), x^2)\}$.

The preceding example illustrates the general behaviour of Cartier and Weil divisors on V-manifolds as the following result shows.

Theorem 2. *Let X be a V-manifold. Then the notion of Cartier and Weil divisor coincide over $\mathbb{Q}$. More precisely, the linear map*

$$T_X \otimes 1 : \operatorname{CaDiv}(X) \otimes_{\mathbb{Z}} \mathbb{Q} \longrightarrow \operatorname{WeDiv}(X) \otimes_{\mathbb{Z}} \mathbb{Q}$$

is an isomorphism of $\mathbb{Q}$-vector spaces. In particular, for a given Weil divisor D on X there exists $k \in \mathbb{Z}$ such that $kD \in$ CaDiv(X).

Definition 4. Let X be a V-manifold. The vector space of $\mathbb{Q}$-Cartier divisors is identified under T_X with the vector space of $\mathbb{Q}$-Weil divisors. A *$\mathbb{Q}$-divisor* on X is an element in $\operatorname{CaDiv}(X) \otimes_{\mathbb{Z}} \mathbb{Q} = \operatorname{WeDiv}(X) \otimes_{\mathbb{Z}} \mathbb{Q}$. The set of all $\mathbb{Q}$-divisors on X is denoted by $\mathbb{Q}$-Div(X).

The proof of the previous result is constructive. Let us summarize here how to write a Weil divisor as an element in $\operatorname{CaDiv}(X) \otimes_{\mathbb{Z}} \mathbb{Q}$ where X is an algebraic V-manifold.

1. Write $D = \sum_{i\in I} a_i[V_i] \in \operatorname{WeDiv}(X)$, where $a_i \in \mathbb{Z}$ and $V_i \subset X$ irreducible. Also choose $\{U_j\}_{j\in J}$ an open covering of X such that $U_j = B_j/G_j$ where $B_j \subset \mathbb{C}^n$ is an open ball and G_j is a **small**[1] finite subgroup of $GL(n, \mathbb{C})$.
2. For each $(i, j) \in I \times J$ choose a polynomial function $f_{i,j} : U_j \to \mathbb{C}$ satisfying the condition $[(f_{i,j})_x \in \mathcal{O}_{B_j,x}$ **reduced** $\forall x \in B_j]$ and such that $V_i \cap U_j = \{f_{i,j} = 0\}$. Then,
$$[V_i|_{U_j}] = \frac{1}{|G_j|}\left\{\left(U_j, f_{i,j}^{|G_j|}\right)\right\}.$$
3. Identifying $\{(U_j, f_{i,j}^{|G_j|})\}$ with its image under $\operatorname{CaDiv}(U_j) \hookrightarrow \operatorname{CaDiv}(X)$, one finally writes D as a sum of locally principal Cartier divisors over $\mathbb{Q}$,
$$D = \sum_{(i,j)\in I\times J} \frac{a_i}{|G_j|}\left\{\left(U_j, f_{i,j}^{|G_j|}\right)\right\}.$$

[1] A finite subgroup G of $GL(n, \mathbb{C})$ is called *small* if no element of G has 1 as an eigenvalue of multiplicity precisely $n - 1$, that is, G does not contain rotations around hyperplanes other than the identity.

§3. Rational intersection number on V-surfaces

Now we are going to develop an intersection theory on varieties with quotient singularities, without getting into technical details.

Definition 5. Let C be an irreducible analytic curve. Given a Weil divisor on C with finite support, $D = \sum_{i=1}^{r} n_i \cdot [P_i]$, its *degree* is defined to be $\deg(D) = \sum_{i=1}^{r} n_i \in \mathbb{Z}$. The *degree of a Cartier divisor* is the degree of its associated Weil divisor.

Definition 6. Let X be an analytic surface and consider $D_1 \in \operatorname{WeDiv}(X)$ and $D_2 \in \operatorname{CaDiv}(X)$. If D_1 is irreducible then the *intersection number* is defined as $D_1 \cdot D_2 := \deg\left(j_{D_1}^{*} D_2\right) \in \mathbb{Z}$, where $j_{D_1} : D_1 \hookrightarrow X$ denotes the inclusion and $j_{D_1}^{*}$ its pull-back functor. The expression above extends linearly if D_1 is a finite sum of irreducible divisors. This number is only well defined if either $D_1 \not\subseteq D_2$ and $D_1 \cap D_2$ is finite, or the divisor D_1 is compact, cf. [5, Ch. 2].

In the case $D_1 \not\subseteq D_2$, the number $(D_1 \cdot D_2)_P := \operatorname{ord}_P(j_{D_1}^{*} D_2)$ with $P \in D_1 \cap D_2$ is well defined too and it is called *local intersection number* at P.

Definition 7. Let X be a V-manifold of dimension 2 and consider $D_1, D_2 \in \mathbb{Q}\text{-}\operatorname{Div}(X)$. The *intersection number* is defined as $D_1 \cdot D_2 := (k_1 k_2)^{-1}\,(k_1 D_1 \cdot k_2 D_2) \in \mathbb{Q}$, where $k_1, k_2 \in \mathbb{Z}$ are chosen so that $k_1 D_1 \in \operatorname{WeDiv}(X)$ and $k_2 D_2 \in \operatorname{CaDiv}(X)$. Analogously, it is defined the *local intersection number* at $P \in D_1 \cap D_2$, if the condition $D_1 \not\subseteq D_2$ is satisfied.

In the following result the main usual properties of the intersection product are collected. Their proofs are straightforward since they are well known for the classical case (i.e. without tensoring with $\mathbb{Q}$), cf. [5], and our generalization is based on extending the classical definition to rational coefficients.

Proposition 3. *Let X be a V-manifold of dimension 2 and $D_1, D_2, D_3 \in \mathbb{Q} - \operatorname{Div}(X)$. Then the local and the global intersection numbers, provided the indicated operations make sense according to Definition 7, satisfy the following properties: ($\alpha \in \mathbb{Q}$, $P \in X$)*

1. *The intersection product is **bilinear** over $\mathbb{Q}$.*
2. ***Commutative:** If $D_1 \cdot D_2$ and $D_2 \cdot D_1$ are both defined, then $D_1 \cdot D_2 = D_2 \cdot D_1$. Analogously $(D_1 \cdot D_2)_P = (D_2 \cdot D_1)_P$ if both local numbers are defined.*
3. ***Non-negative:** Assume D_1 and D_2 are effective, irreducible and distinct. Then $D_1 \cdot D_2$ and $(D_1 \cdot D_2)_P$ are greater than or equal to zero. Moreover, $(D_1 \cdot D_2)_P = 0$ if and only if $P \notin |D_1| \cap |D_2|$, and hence $D_1 \cdot D_2 = 0$ if and only if $|D_1| \cap |D_2| = \emptyset$.*
4. ***Non-rational:** If $D_2 \in \operatorname{CaDiv}(X)$, $D_1 \in \operatorname{WeDiv}(X)$ then $D_1 \cdot D_2$ and $(D_1 \cdot D_2)_P$ are integral numbers. By the commutative property, the same holds if D_1 is a Cartier divisor and D_2 is a Weil divisor.*
5. ***$\mathbb{Q}$-Linear equivalence:** Assume D_1 has compact support. If D_2 and D_3 are $\mathbb{Q}$-linearly equivalent, i.e. $[D_2] = [D_3] \in \operatorname{Pic}(X) \otimes_{\mathbb{Z}} \mathbb{Q}$, then $D_1 \cdot D_2 = D_1 \cdot D_3$. Due to the commutativity, the roles of D_1 and D_2 can be exchanged.*
6. ***Normalization:** Let $\nu : \widetilde{|D_1|} \to |D_1|$ be the normalization of the support of D_1 and $j_{D_1} : |D_1| \hookrightarrow X$ the inclusion. Then $D_1 \cdot D_2 = \deg\left(j_{D_1} \circ \nu\right)^{*} D_2$. Observe that in this situation the normalization is a smooth complex analytic curve.*

7. ***Pull-back:*** *Let Y be another irreducible V-surface and let $F : Y \to X$ be a proper morphism. Given $D_1, D_2 \in \mathbb{Q} - \mathrm{Div}(X)$, if the intersection product $D_1 \cdot D_2$ is defined, then so is $F^*(D_1) \cdot F^*(D_2)$ and one has $F^*(D_1) \cdot F^*(D_2) = \deg(F)(D \cdot E)$.*

Remark 1. This rational intersection product was first introduced by Mumford for normal surfaces, see [7, pag. 17]. Our Definition 7 coincides with Mumford's because it has good behavior with respect to the pull-back. The main advantage is that ours does not involve a resolution of the ambient space and, for instance, this allowed us to easily find formulas for the self-intersection numbers of the exceptional divisors of weighted blow-ups, without computing any resolution, see Proposition 4 below.

The rest of this section is devoted to reviewing some classical results concerning the intersection multiplicity.

Classical blow-up at a smooth point. Let X be a smooth analytic surface. Let $\pi : \widehat{X} \to X$ be the classical blow-up at a (smooth) point P. Consider C and D two Cartier or Weil divisors on X with multiplicities m_C and m_D at P. Denote by E the exceptional divisor of π, and by $\widehat{C}$ (resp. $\widehat{D}$) the strict transform of C (resp. D). Then there are following equalities:

1. $E \cdot \pi^*(C) = 0, \quad \pi^*(C) = \widehat{C} + m_C E, \quad E \cdot \widehat{C} = m_C.$
2. $E^2 = -1, \quad \widehat{C} \cdot \widehat{D} = C \cdot D - m_C m_D, \quad \widehat{D}^2 = D^2 - m_D^2$ (when D is compact).

Note that the exceptional divisor has multiplicity 1 at every point. This is why one only has to subtract 1 for the self-intersection number of the exceptional divisors every time we blow up a point on them, when computing an embedded resolution on a plane curve.

Bézout's Theorem on $\mathbb{P}^2$. Every analytic Cartier or Weil divisor on $\mathbb{P}^2$ is algebraic and thus can be written as a difference of two effective divisors. On the other hand, every effective divisor is defined by a homogeneous polynomial. The *degree of an effective divisor on* $\mathbb{P}^2$ is the degree, $\deg(F)$, of the corresponding homogeneous polynomial. This degree map is extended linearly yielding a group homomorphism $\deg : \mathrm{Div}(\mathbb{P}^2) \to \mathbb{Z}$.

Let D_1, D_2 be two divisors on $\mathbb{P}^2$, then $D_1 \cdot D_2 = \deg(D_1)\deg(D_2)$. In particular the self-intersection number of a divisor D on $\mathbb{P}^2$ is $D^2 = \deg(D)^2$.

The rest of this paper is devoted to generalizing the classical results above to V-manifolds of dimension 2, weighted blow-ups, and quotient weighted projective planes, respectively.

§4. Weighted blow-ups and embedded Q-resolutions

Classically an embedded resolution of $\{f = 0\} \subset \mathbb{C}^n$ is a proper map $\pi : X \to (\mathbb{C}^n, 0)$ from a smooth variety X satisfying, among other conditions, that $\pi^{-1}(\{f = 0\})$ is a normal crossing divisor. To weaken the condition on the preimage of the singularity we allow the new ambient space X to contain abelian quotient singularities and the divisor $\pi^{-1}(\{f = 0\})$ to have "normal crossings" over this kind of varieties. This notion of normal crossing divisor on V-manifolds was first introduced by Steenbrink in [10].

Definition 8. A hypersurface D on a V-manifold X with abelian quotient singularities is said to be with $\mathbb{Q}$-*normal crossings* if it is locally isomorphic to the quotient of a normal crossing divisor under a group action of type $(d; A)$. That is, given $x \in X$, there is an isomorphism of germs $(X, x) \simeq (X(d; A), [\mathbf{0}])$ such that $(D, x) \subset (X, x)$ is identified under this morphism with a germ of the form $\left(\{[\mathbf{x}] \in X(d; A) \mid x_1^{m_1} \cdots x_k^{m_k} = 0\},\ [(0, \dots, 0)]\right)$.

Definition 9. Let $M = \mathbb{C}^{n+1}/G$ be an abelian quotient space. Consider $H \subset M$ an analytic subvariety of codimension one. An embedded **Q**-resolution of $(H, 0) \subset (M, 0)$ is a proper analytic map $\pi : X \to (M, 0)$ such that:

1. X is a V-manifold with abelian quotient singularities.
2. π is an isomorphism over $X \setminus \pi^{-1}(\mathrm{Sing}(H))$.
3. $\pi^{-1}(H)$ is a hypersurface with $\mathbb{Q}$-normal crossings on X.

Usually one uses weighted or toric blow-ups with smooth center as a tool for finding embedded **Q**-resolutions. Here we only discuss briefly the surface case. Let X be an analytic surface with abelian quotient singularities. Let us define the weighted blow-up $\pi : \widehat{X} \to X$ at a point $P \in X$ with respect to $\omega = (p, q)$. We distinguish two different situations.

(i) The point P is smooth. Assume $X = \mathbb{C}^2$ and $\pi = \pi_\omega : \widehat{\mathbb{C}}^2_\omega \to \mathbb{C}^2$ the weighted blow-up at the origin with respect to $\omega = (p, q)$,

$$\widehat{\mathbb{C}}^2_\omega := \left\{((x, y), [u : v]_\omega) \in \mathbb{C}^2 \times \mathbb{P}^1_\omega \mid (x, y) \in \overline{[u : v]_\omega}\right\}.$$

Here the condition about the closure means that $\exists t \in \mathbb{C}$, $x = t^p u$, $y = t^q v$. The new ambient space is covered as $\widehat{\mathbb{C}}^2_\omega = U_1 \cup U_2 = X(p; -1, q) \cup X(q;\ p, -1)$ and the charts are given by

$$\begin{array}{rl|rl} X(p; -1, q) \longrightarrow U_1, & & X(q;\ p, -1) \longrightarrow U_2, & \\ [(x, y)] \longmapsto ((x^p, x^q y), [1 : y]_\omega); & & [(x, y)] \longmapsto ((xy^p, y^q), [x : 1]_\omega). & \end{array}$$

The exceptional divisor $E = \pi_\omega^{-1}(0)$ is isomorphic to $\mathbb{P}^1_\omega$ which is in turn isomorphic to $\mathbb{P}^1$ under the map $[x : y]_\omega \mapsto [x^q : y^p]$. The singular points of $\widehat{\mathbb{C}}^2_\omega$ are cyclic quotient singularities located at the exceptional divisor. They actually coincide with the origins of the two charts.

(ii) The point P is of type $(d; a, b)$. Assume that $X = X(d; a, b)$. The group μ_d acts also on $\widehat{\mathbb{C}}^2_\omega$ and passes to the quotient yielding a map $\pi = \pi_{(d;a,b),\omega} : \widehat{X(d; a, b)}_\omega \to X(d; a, b)$, where by definition $\widehat{X(d; a, b)}_\omega := \widehat{\mathbb{C}}^2_\omega/\mu_d$. The new space is covered as

$$\widehat{X(d; a, b)}_\omega = \widehat{U}_1 \cup \widehat{U}_2 = X\left(\begin{array}{c|cc} p & -1 & q \\ pd & a & pb - qa \end{array}\right) \cup X\left(\begin{array}{c|cc} q & p & -1 \\ qd & qa - pb & b \end{array}\right)$$

and the charts are given by

$$\begin{array}{c|c} X\left(\begin{array}{c|cc} p & -1 & q \\ pd & a & pb - qa \end{array}\right) \longrightarrow \widehat{U}_1, & X\left(\begin{array}{c|cc} q & p & -1 \\ qd & qa - pb & b \end{array}\right) \longrightarrow \widehat{U}_2, \\ [(x, y)] \longmapsto [((x^p, x^q y), [1 : y]_\omega)]_{(d;a,b)}; & [(x, y)] \longmapsto [((xy^p, y^q), [x : 1]_\omega)]_{(d;a,b)}. \end{array}$$

The exceptional divisor $E = \pi^{-1}_{(d;a,b),\omega}(0)$ is identified with $\mathbb{P}^1_\omega(d; a, b) := \mathbb{P}^1_\omega/\mu_d$. Again the singular points are cyclic and correspond to the origins of the two charts.

Proposition 4. *Let X be a surface with abelian quotient singularities. Let $\pi : \widehat{X} \to X$ be the weighted blow-up at a point of type $(d; a, b)$ with respect to $\omega = (p, q)$. Assume $(d, a) = (d, b) = (p, q) = 1$ and write $e = \gcd(d, pb - qa)$.*

Consider two $\mathbb{Q}$-divisors C and D on X and, as usual, denote by E the exceptional divisor of π, and by $\widehat{C}$ (resp. $\widehat{D}$) the strict transform of C (resp. D). Let ν and μ the (p, q)-multiplicities of C and D at P, i.e. x (resp. y) has (p, q)-multiplicity p (resp. q). Then there are the following equalities:

1. $E \cdot \pi^*(C) = 0, \quad \pi^*(C) = \widehat{C} + \frac{\nu}{e} E, \quad E \cdot \widehat{C} = \frac{e\nu}{pqd}.$
2. $E^2 = -\frac{e^2}{pqd}, \quad \widehat{C} \cdot \widehat{D} = C \cdot D - \frac{\nu\mu}{pqd}, \quad \widehat{D}^2 = D^2 - \frac{\mu^2}{pqd}$ *(when D is compact).*

§5. Bézout's Theorem for Quotient Weighted Projective Planes

For a given weight vector $\omega = (p, q, r) \in \mathbb{N}^3$ and an action on $\mathbb{C}^3$ of type $(d; a, b, c)$, consider the quotient weighted projective plane $\mathbb{P}^2_\omega(d; a, b, c) := \mathbb{P}^2_\omega / \mu_d$ and the corresponding morphism $\tau_{(d;a,b,c),\omega} : \mathbb{P}^2 \to \mathbb{P}^2_\omega(d; a, b, c)$ defined by $\tau_{(d;a,b,c),\omega}([x : y : z]) = [x^p : y^q : z^r]_\omega$.

The space $\mathbb{P}^2_\omega(d; a, b, c)$ is a V-manifold with abelian quotient singularities; its charts are obtained as in Section 1. The *degree of a $\mathbb{Q}$-divisor on $\mathbb{P}^2_\omega(d; a, b, c)$* is the degree of its pull-back under the map $\tau_{(d;a,b,c),\omega}$, that is, by definition,

$$D \in \mathbb{Q}\text{-Div}\Big(\mathbb{P}^2_\omega(d; a, b, c)\Big), \quad \deg_\omega(D) := \deg\Big(\tau^*_{(d;a,b,c),\omega}(D)\Big).$$

Thus if $D = \{F = 0\}$ is a $\mathbb{Q}$-divisor on $\mathbb{P}^2_\omega(d; a, b, c)$ given by a ω-homogeneous polynomial that indeed defines a zero set on the quotient projective space, then $\deg_\omega(D)$ is the classical degree, denoted by $\deg_\omega(F)$, of a quasi-homogeneous polynomial.

Proposition 5. *Let us denote by m_1, m_2, m_3 the determinants of the three minors of order 2 of the matrix $\left(\begin{smallmatrix} p & q & r \\ a & b & c \end{smallmatrix}\right)$. Assume $\gcd(p, q, r) = 1$ and write $e = \gcd(d, m_1, m_2, m_3)$. Then the intersection number of two $\mathbb{Q}$-divisors, D_1 and D_2, on $\mathbb{P}^2_\omega(d; a, b, c)$ is*

$$D_1 \cdot D_2 = \frac{e}{dpqr} \deg_\omega(D_1) \deg_\omega(D_2) \in \mathbb{Q}.$$

Corollary 6. *Let X, Y, Z be the Weil divisors on $\mathbb{P}^2_\omega(d; a, b, c)$ given by $\{x = 0\}$, $\{y = 0\}$ and $\{z = 0\}$, respectively. Using the notation above one has:*

$$X^2 = \frac{ep}{dqr}, \quad Y^2 = \frac{eq}{dpr}, \quad Z = \frac{er}{dpq}, \quad X \cdot Y = \frac{e}{dr}, \quad X \cdot Z = \frac{e}{dq}, \quad Y \cdot Z = \frac{e}{dp}.$$

Remark 2. If $d = 1$, then e equals one too and the formulas become a bit simpler.

§6. Example of an Embedded Q-Resolution

Let us consider the following divisors on $\mathbb{C}^2$: $C_1 = \{((x^3 - y^2)^2 - x^4 y^3) = 0\}$, $C_2 = \{x^3 - y^2 = 0\}$, $C_3 = \{x^3 + y^2 = 0\}$, $C_4 = \{x = 0\}$ and $C_5 = \{y = 0\}$. We shall see that the local intersection

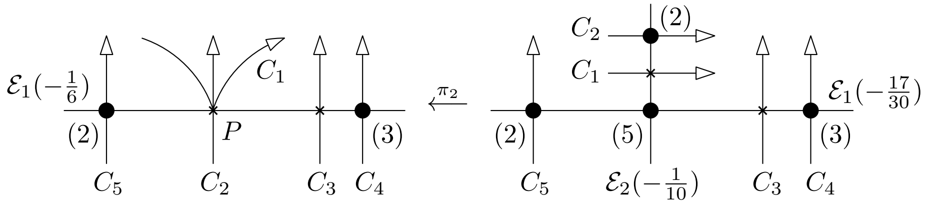

$\mathcal{E}_1(-\frac{1}{6})$
C_1
(2)
P
(3)
C_5
C_2
C_3
C_4
π_2
C_2
(2)
C_1
$\mathcal{E}_1(-\frac{17}{30})$
(2)
(5)
(3)
C_5
$\mathcal{E}_2(-\frac{1}{10})$
C_3
C_4

Figure 1: Embedded **Q**-resolution of $C = \bigcup_{i=1}^{5} C_i \subset \mathbb{C}^2$.

numbers $(C_i \cdot C_j)_0$, $i, j \in \{1, \dots, 5\}$, $i \neq j$, are encoded in the intersection matrix associated with any embedded **Q**-resolution of $C = \bigcup_{i=1}^{5} C_i$.

Let $\pi_1 : \mathbb{C}^2_{(2,3)} \to \mathbb{C}^2$ be the $(2,3)$-weighted blow-up at the origin. The new space has two cyclic quotient singular points of type $(2; 1, 1)$ and $(3; 1, 1)$ located at the exceptional divisor $\mathcal{E}_1$. The local equation of the total transform in the first chart is given by the function

$$x^{29}\,((1-y^2)^2 - x^5y^3)\,(1-y^2)\,(1+y^2)\,y : X(2;1,1) \longrightarrow \mathbb{C},$$

where $x = 0$ is the equation of the exceptional divisor and the other factors correspond in the same order to the strict transform of C_1, C_2, C_3, C_5 (denoted again by the same symbol). To study the strict transform of C_4 one needs the second chart, the details are left to the reader.

Hence $\mathcal{E}_1$ has multiplicity 29 and self-intersection number $-1/6$; it intersects transversally C_3, C_4 and C_5 at three different points, while it intersects C_1 and C_2 at the same smooth point P, different from the other three. The local equation of the divisor $\mathcal{E}_1 \cup C_2 \cup C_1$ at this point P is $x^{29}\,y\,(x^5 - y^2) = 0$, see Figure 1 below.

Let π_2 be the $(2,5)$-weighted blow-up at the point P above. The new ambient space has two singular points of type $(2; 1, 1)$ and $(5; 1, 2)$. The local equations of the total transform of $\mathcal{E}_1 \cup C_2 \cup C_1$ are given by the following two functions.

1st chart	2nd chart
$\underbrace{x^{73}}_{\mathcal{E}_2} \cdot \underbrace{y}_{C_2} \cdot \underbrace{(1-y^2)}_{C_1} \,:\, X(2;1,1) \longrightarrow \mathbb{C}$	$\underbrace{x^{29}}_{\mathcal{E}_1} \cdot \underbrace{y^{73}}_{\mathcal{E}_2} \cdot \underbrace{(x^5-1)}_{C_1} \,:\, X(2;1,1) \longrightarrow \mathbb{C}$

Thus the new exceptional divisor $\mathcal{E}_2$ has multiplicity 73 and intersects transversally the strict transform of C_1, C_2 and $\mathcal{E}_1$. Hence the composition $\pi_2 \circ \pi_1$ is an embedded **Q**-resolution of $C = \bigcup_{i=1}^{5} C_i \subset \mathbb{C}^2$. As for the self-intersection numbers, $\mathcal{E}_2^2 = -1/10$ and $\mathcal{E}_1^2 = -1/6 - 2^2/(1 \cdot 2 \cdot 5) = -17/30$. The following figure illustrates the whole process. The intersection matrix associated with the embedded **Q**-resolution obtained is $A = \begin{pmatrix} -17/30 & 1/5 \\ 1/5 & -1/10 \end{pmatrix}$ and $B = -A^{-1} = \begin{pmatrix} 6 & 12 \\ 12 & 34 \end{pmatrix}$.

Now one observes the intersection number is encoded in B as follows. For $i = 1, \dots, 5$, set $k_i \in \{1, \dots, 5\}$ such that $\emptyset \neq C_i \cap \mathcal{E}_{k_i} =: \{P_i\}$. Denote by $O(C_i)$ the order of the cyclic group acting on P_i. Then,

$$\big(C_i \cdot C_j\big)_0 = \frac{b_{k_i,k_j}}{O(C_i)\,O(C_j)}.$$

Looking at the figure one sees that $(k_1, \ldots, k_5) = (2, 2, 1, 1, 1)$ and $(O(C_1), \ldots, O(C_5)) = (1, 2, 1, 3, 2)$. Hence, for instance,

$$(C_1 \cdot C_2)_0 = \frac{b_{k_1,k_2}}{O(C_1)\, O(C_2)} = \frac{b_{22}}{1 \cdot 2} = \frac{34}{2} = 17,$$

which is indeed the intersection multiplicity at the origin of C_1 and C_2. Analogously for the other indices.

Remark 3. Consider the group action of type $(5; 2, 3)$ on $\mathbb{C}^2$. The previous plane curve C is invariant under this action and then it makes sense to compute an embedded **Q**-resolution of $\overline{C} := C/\mu_5 \subset X(5; 2, 3)$. Similar calculations as in the previous example, lead to a figure as the one obtained above with the following relevant differences:

- $\mathcal{E}_1 \cap \mathcal{E}_2$ is a smooth point.
- $\mathcal{E}_1$ (resp. $\mathcal{E}_2$) has self-intersection number $-17/6$ (resp. $-1/2$).
- The intersection matrix is $A' = \begin{pmatrix} -17/6 & 1 \\ 1 & -1/2 \end{pmatrix}$ and $B' = -(A')^{-1} = \begin{pmatrix} 6/5 & 12/5 \\ 12/5 & 34/5 \end{pmatrix}$.

Hence, for instance, $(\overline{C}_1 \cdot \overline{C}_2)_0 = b'_{22}/(1 \cdot 2) = (34/5)/2 = 17/5$, which is exactly the intersection number of the two curves, since that local number can also be computed as $(\overline{C}_1 \cdot \overline{C}_2)_0 = 5^{-1}(C_1 \cdot C_2)_0$.

Conclusion. The combinatorial and computational complexity of embedded **Q**-resolutions is much simpler than the one of the classical embedded resolutions, but they keep as much information as needed for the comprehension of the topology of the singularity. This will become clear in the second author's Ph.D. thesis. We will prove in a forthcoming paper another advantages of these embedded **Q**-resolutions, e.g. in the computation of abstract resolutions of surfaces via Jung method, see [1, 2, 6].

Acknowledgements

We wish to express our gratitude to J. I. Cogolludo for his constant support and motivation in our work over the years. All authors are partially supported by the Spanish projects MTM2010-2010-21740-C02-02 and "E15 Grupo Consolidado GeometrÌa" from the Government of Aragón. Also, the second author is partially supported by FQM-333.

References

[1] Artal Bartolo, E., MartÌn-Morales, J., and Ortigas-Galindo, J. Cartier and Weil divisors on varieties with quotient singularities. *ArXiv e-prints* (Apr. 2011).

[2] Artal Bartolo, E., MartÌn-Morales, J., and Ortigas-Galindo, J. Intersection theory on abelian-quotient V-surfaces and **Q**-resolutions. *ArXiv e-prints* (May 2011).

[3] Baily, Jr., W. L. The decomposition theorem for V-manifolds. *Amer. J. Math. 78* (1956), 862–888.

[4] Dolgachev, I. Weighted projective varieties. In *Group actions and vector fields (Vancouver, B.C., 1981)*, vol. 956 of *Lecture Notes in Math.* Springer, Berlin, 1982, pp. 34–71.

[5] Fulton, W. *Intersection theory*, second ed., vol. 2 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics].* Springer-Verlag, Berlin, 1998.

[6] MartÌn-Morales, J. Monodromy zeta function formula for embedded **Q**-resolutions. *In Preparation* (July 2011).

[7] Mumford, D. The topology of normal singularities of an algebraic surface and a criterion for simplicity. *Inst. Hautes Études Sci. Publ. Math.*, 9 (1961), 5–22.

[8] Prill, D. Local classification of quotients of complex manifolds by discontinuous groups. *Duke Math. J. 34* (1967), 375–386.

[9] Satake, I. On a generalization of the notion of manifold. *Proc. Nat. Acad. Sci. U.S.A. 42* (1956), 359–363.

[10] Steenbrink, J. H. M. Mixed Hodge structure on the vanishing cohomology. In *Real and complex singularities (Proc. Ninth Nordic Summer School/NAVF Sympos. Math., Oslo, 1976).* Sijthoff and Noordhoff, Alphen aan den Rijn, 1977, pp. 525–563.

Enrique Artal Bartolo and Jorge Ortigas-Galindo
Departamento de Matemáticas-IUMA
Universidad de Zaragoza
C/ Pedro Cerbuna 12, 50009, Zaragoza, Spain
`artal@unizar.es` and `jortigas@unizar.es`

Jorge Martín-Morales
Centro Universitario de la Defensa-IUMA
Academia General Militar
Ctra. de Huesca s/n. 50090, Zaragoza, Spain
`jorge@unizar.es`

Monografías Matemáticas *García de Galdeano* **37**, 25–35 (2012)

A REMARK ABOUT SYMMETRY OF SOLUTIONS TO SINGULAR EQUATIONS AND APPLICATIONS

Kaushik Bal and Jacques Giacomoni

Abstract. In this article we will use the moving plane method to discuss the symmetry of solution to an elliptic equation with singularity. Moreover by choosing a particular type of nonlinearity we will show some a priori estimates with the help of moving plane method.

Keywords: Symmetry, singularity, a priori estimate.

AMS classification: 35K55, 35J25, 35J65.

§1. Introduction

Suppose that Ω is a bounded domain in $\mathbb{R}^n$. Consider the equation

$$\begin{gathered} -\Delta u = \frac{1}{u^{\delta}} + f(u) \ \text{ in } \Omega, \\ u = 0 \ \text{ on } \partial\Omega, \ u > 0 \text{ in } \Omega, \end{gathered}$$

where $\delta > 0$ given and f is a locally lipchitz in $\mathbb{R}$. Extensive studies have been done on this equation in the past by many authors [1], [2], [5], [10], [12] and [13]. This kind of problem arises in the study of non-Newtonian fluids, boundary layer phenomena for viscous fluids as well as chemical heterogeneous chemical reactions.

In a famous paper [2] it was proved that equations of this kind admits a unique solution $u \in C^{2+\alpha}(\Omega) \cap C(\overline{\Omega})$. Moreover there exists positive constants R and Q s.t

$$Rp(d(x)) \leq u(x) \leq Qp(d(x))$$

near $\partial\Omega$, where $d(x) = \operatorname{dist}(x, \partial\Omega)$ and $p \in C([0, a]) \cap C^2((0, a])$ is the local solution of the problem

$$-p^{''} = g(p(s)), \quad p(s) > 0, \quad 0 < s < a, \quad p(0) = 0,$$

where $a > 0$ and g is a monotone decreasing continuous function.

In another famous paper [7] it was proved by the help of the moving plane method that if $u \in C(\overline{B}) \cap C^2(B)$ is a positive solution of

$$\begin{gathered} \Delta u + f(u) = 0 \ \text{ in } B \\ u = 0 \ \text{ on } \partial B \end{gathered}$$

where B is the unit ball and f is a locally lipchitz in $\mathbb{R}$. Then u is radially symmetric in B and $\frac{\partial u}{\partial r}(x) < 0$.

The original proof requires that solutions be C^2 up to the boundary. The main feature of our paper is to find the symmetry of the solution to the problem with singularity without any assumptions on the smoothness of the solutions up to the boundary. We also prove the existence of universal bounds for superlinear and singular problems following the idea of [9].

§2. Main results and preliminaries

Our main result is the following:

Theorem 1. *Suppose that Ω is a bounded domain which is convex in x_1 direction and symmetric with respect to the plane $x_1 = 0$. Suppose $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a positive solution of*

$$\begin{aligned} \Delta u + \frac{1}{u^\delta} + f(u) = 0 \;\; &\mathit{in}\, \Omega \\ u = 0 \;\; \mathit{on}\, \partial\Omega, \;\; u > 0 \; &\mathit{in}\, \Omega \end{aligned}$$

where $\delta > 0$ given and f is a locally lipchitz in $\mathbb{R}$. Then u is symmetric w.r.t x_1 and $D_{x_1}(x) < 0$ for any $x \in \Omega$ with $x_1 > 0$.

To proof the main theorem we need preliminary which we are going to state now. Let Ω be a bounded domain in $\mathbb{R}^n$. Consider the operator L in Ω

$$Lu = \sum_{i,j}^{n} a_{ij}(x) D_{ij}(x) u + \sum_{i}^{n} b_i(x) D_i u + c(x) u$$

for $u \in C^2(\Omega) \cap C(\overline{\Omega})$. We assume that a_{ij}, b_i and c are continuous in Ω. The coefficient matrix $A = (a_{ij})$ is positive definite everywhere in Ω. Likewise, we denote $D^* := (\det(A))^{1/n}$ as the geometric mean of the eigenvalues of A.

Definition 1. Define for every $u \in C^2(\Omega)$,

$$\Gamma^+(u) = \{y \in \Omega; u(x) \leq u(y) + Du(y).(x - y), x \in \Omega\}.$$

The set $\Gamma^+(u)$ is called the upper contact set of u and the Hessian matrix $(D^2 u)$ is nonpositive on $\Gamma^+(u)$.

Let us state a lemma from [11] (see Lemma 2.24) required to the proof of Alexandroff Maximum Principle.

Lemma 2. *Suppose $g \in L^1_{loc}(\Omega)$ is nonnegative. Then for any $u \in C^2(\Omega) \cap C(\overline{\Omega})$, there holds*

$$\int_{B_k(0)} g \leq \int_{\Gamma^+(u)} g(Du) \left|\det D^2 u\right|,$$

where $\Gamma^+(u)$ is the upper contact set of u, $B_k(0)$ is the ball with radius k and center 0 *and $k = (1/d)(\sup_\Omega u - \sup_{\partial\Omega} u^+)$, where d is the diameter of Ω.*

Now we give the Alexandroff Maximum Principle

Theorem 3. *Suppose $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfies $Lu \geq f$ in Ω with the following conditions*

$$\frac{|b|}{D^*}, \frac{f}{D^*} \in L^n(\Omega) \text{ and } c \leq 0 \text{ in } \Omega.$$

Then there holds

$$\sup_{\Omega} u \leq \sup_{\partial\Omega} u^+ + C \left\| \frac{f^-}{D^*} \right\|_{L^n(\Gamma^+(u))},$$

where C is a constant depend only on n, diam(Ω) and $\|f^-/D\|_{L^n(\Gamma^+(u))}$.

Note here that $c(x)$ is assumed to be only measurable and no assumption on the boundedness is required. We are providing the sketch of the proof for the convenience of the reader.

Proof. Without loss of generality we assume $u < 0$ on $\partial\Omega$. Set $\Omega^+ = \{u > 0\}$. Take $g(p) = (|p|^n + \mu^n)^{-1}$ and then let $\mu \to 0^+$.

Recall the area-formula for Du in $\Gamma^+ \cap \Omega^+ \subset \Omega$ gives

$$\int_{Du(\Gamma^+\cap\Omega^+)} \leq \int_{\Gamma^+\cap\Omega^+} g(Du) \left|\det(D^2(u))\right|,$$

where $D^2(u)$ is the Jacobian of the map $Du : \Omega \to \mathbb{R}^n$.

First we have,

$$\begin{gathered} -a_{ij}D_{ij}u \leq b_iD_iu + cu - f, \\ -a_{ij}D_{ij}u \leq b_iD_iu - f \text{ in } \Omega^+ = \{x; u(x) > 0\}, \\ -a_{ij}D_{ij}u \leq |b||Du| + f^-. \end{gathered}$$

Then by Cauchy inequality we have,

$$-a_{ij}D_{ij}u \leq 2\left(|b|^n + \frac{(f^-)^n}{\mu^n}\right)^{1/n} .(|Du|^n + \mu^n)^{1/n}.$$

So, by Lemma 2 and recalling that

$$\det(-D^2u) \leq \frac{1}{D}\left(\frac{-a_{ij}D_{ij}u}{n}\right)^n \text{ on } \Gamma^+,$$

where $D = \det(A)$, we have

$$\int_{B_k(0)} g \leq \frac{2^n}{n^n} \int_{\Gamma^+\cap\Omega^+} \frac{|b|^n + \mu^{-n}(f^-)^n}{D}.$$

Now evaluating the integral in the left-hand side we have,

$$\int_{B_k(0)} g = \frac{\omega_n}{n} \log\left(\frac{k^n}{\mu^n} + 1\right),$$

where ω_n is the volume of the unit ball in $\mathbb{R}^n$. Therefore we obtain

$$k^n \leq \mu^n \left\{\exp\left\{\frac{2^n}{\omega_n n^n}\left[\left\|\frac{b}{D^*}\right\|^n_{L^n(\Gamma^+\cap\Omega^+)} + \mu^{-n}\left\|\frac{f^-}{D^*}\right\|^n_{L^n(\Gamma^+\cap\Omega^+)}\right]\right\} - 1\right\}.$$

If $f \not\equiv 0$ then choose any $\mu > 0$ and then let $\mu \to 0$. This completes the proof. □

Next we give a statement of Hopf Maximum Principle and a Strong Maximum Principle adapted to our situation (see [11]). Let us assume the operator L as described above with the assumption that a_{ij}, b_i are continuous and hence bounded in $\overline{\Omega}$ and $c(x)$ is bounded below.

Then we have the following results:

Lemma 4 (Hopf Lemma)**.** *Let B an open ball in $\mathbb{R}^n$ with $x_0 \in \partial B$. Suppose $u \in C^2(B) \cap C(B \cup \{x_0\})$ satisfies $Lu \geq 0$ in B with $c(x) \leq 0$ and uniformly bounded in B. Assume in addition that*

$$u(x) < u(x_0) \quad \textit{for any } \ x \in B \textit{ and } u(x_0) \geq 0.$$

Then for each outward direction $\overline{\nu}$ and an outward normal direction $\overline{n}$ at x_0 with $\overline{\nu}.\overline{n} > 0$ there holds:

$$\liminf_{t \to 0^+} \frac{1}{t}[u(x_0) - u(x_0 - t\nu)] > 0.$$

Remark 1. If in addition $u \in C^2(\Omega) \cap C^1(\Omega \cup \{x_0\})$ then we have

$$\frac{\partial u}{\partial \nu}(x_0) > 0.$$

The proof of Lemma 4 can be found in [11]. From Lemma 4 we can prove the following strong maximum principle:

Theorem 5 (Strong Maximum Principle)**.** *Let Ω be a bounded and connected domain in $\mathbb{R}^n$. Suppose $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfies $Lu \geq 0$ in Ω with $c(x) \leq \underline{0}$. Then, the nonnegative maximum of u can be assumed only on $\partial\Omega$ unless u is constant in $\overline{\Omega}$.*

We adapt the proof given in [11].

Proof. Let M be the nonnegative maximum of u in $\overline{\Omega}$. Set $\Sigma := \{x \in \Omega;\ u(x) = M\}$. It is relatively closed in Ω. We want to show $\Sigma = \Omega$.

We prove by contradiction. If Σ is a proper set of Ω, then we may find an open ball $B \subset \Omega \setminus \Sigma$ with a point on its boundary belonging to Σ. (In fact, we may choose a point $p \in \Omega \setminus \Sigma$ such that $\mathrm{d}(p, \Sigma) < \mathrm{d}(p, \partial\Omega)$ first and then extend the ball. It hits Σ before hitting $\partial\Omega$). Suppose $x_0 \in \partial B \cap \Sigma$. Obviously we have $Lu \geq 0$ in B and

$$u(x) < u(x_0) \quad \text{for any } \ x \in B \text{ and } u(x_0) = M \geq 0.$$

Lemma 4 (note that c is bounded in B since by construction, $\overline{B} \subset \Omega$) implies $\frac{\partial u}{\partial \nu} > 0$ where ν is the outward normal direction at x_0 to the ball B. While x_0 is the interior maximal point of Ω, hence $Du(x_0) = 0$. This leads to a contradiction. □

A straightforward consequence of Theorem 5 is the following result:

Corollary 6 (Comparison Principle)**.** *Suppose $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfies $Lu \geq 0$ in Ω with $c(x) \leq 0$ in Ω. If $u \leq 0$ on $\partial\Omega$, then $u \leq 0$ in Ω. In fact, either $u < 0$ in Ω or $u \equiv 0$ in Ω.*

§3. Proof of the main result

Write $x = (x_1, y) \in \Omega$ for $y \in \mathbb{R}^{n-1}$. We will prove

$$u(x_1, y) < u(x_1^*, y) \text{ for any } x_1 > 0 \text{ and } x_1^* < x_1 \text{ with } x_1^* + x_1 > 0.$$

Then letting $x_1^* \to -x_1$, we get $u(x_1, y) \leq u(-x_1, y)$ for any x_1. Then by changing the direction $x_1 \to -x_1$,we get the symmetry.

We let a = sup x_1 for $(x_1, y) \in \Omega$ and for $0 < \lambda < a$, we define

$$\begin{aligned}
\Sigma_\lambda &= \{(x_1, \dots, x_n) \in \Omega \mid x_1 > \lambda\},\\
T_\lambda &= \{(x_1, \dots, x_n) \in \Omega \mid x_1 = \lambda\},\\
\Sigma'_\lambda &= \{(2\lambda - x_1, ..., x_n) \in \Omega \mid (x_1, \dots, x_n) \in \Sigma_\lambda\}.
\end{aligned}$$

Notice that Σ'_λ is the reflection of Σ_λ with respect to T_λ. In the following we denote by x_λ the image of x with respect to T_λ.

In Σ_λ, we define $w_\lambda(x) = u(x) - u(x_\lambda)$ for $x \in \Sigma_\lambda$. Then by Mean Value Theorem we have

$$\begin{aligned}
&\Delta w_\lambda + c(x, \lambda) w_\lambda - \frac{\delta w_\lambda}{u_\gamma^{\delta+1}} = 0 \text{ in } \Sigma_\lambda.\\
&w_\lambda \leq 0 \textit{ and } w_\lambda \neq 0 \text{ on } \partial\Sigma_\lambda.
\end{aligned} \tag{1}$$

where $u_\gamma(x) = u(x_\gamma)$ with x_γ is a suitable convex combination of x and x_λ and $c(x, \lambda)$ is a bounded function in Σ_λ .

We need to show $w_\lambda < 0$ in Σ_λ for any $\lambda \in (0, a)$. We divide the proof in three steps.

Step 1. For any λ close to a, we first show $w_\lambda \leq 0$, i.e we can actually start the moving plane. For λ close to a, we are rearranging (1) as:

$$\begin{aligned}
&\Delta w_\lambda - \left[c^-(x, \lambda) + \frac{\delta}{u_\gamma^{\delta+1}}\right] w_\lambda = -c^+(x, \lambda) w_\lambda \text{ in } \Sigma_\lambda,\\
&w_\lambda \leq 0 \text{ and } w_\lambda \neq 0 \text{ on } \partial\Sigma_\lambda.
\end{aligned}$$

Now, since $\sup_{\partial\Sigma_\lambda} w_\lambda = 0$, we have by Theorem 3 that for λ close to a,

$$\begin{aligned}
&\sup_{\Sigma_\lambda} w_\lambda \leq C(n, d)\, \|c^+ w_\lambda^+\|_{L^n(\Sigma_\lambda)},\\
&\sup_{\Sigma_\lambda} w_\lambda \leq C(n, d)\, \|c^+\|_{L^\infty(\Sigma_\lambda)} |\Sigma_\lambda|^{1/n} \sup_{\Sigma_\lambda} w_\lambda \leq \frac{1}{2} \sup_{\Sigma_\lambda} w_\lambda,
\end{aligned}$$

where d denotes the diameter of Ω. So we have $w_\lambda \leq 0$ for λ close to a.

Applying Corollary 6, we get $w_\lambda < 0$ in Σ_λ for λ close to a.

Step 2. Let (λ_0, a) be the largest interval of values of λ such that $w_\lambda < 0$ in Σ_λ. We want to show $\lambda_0 = 0$. If $\lambda_0 > 0$ by continuity $w_{\lambda_0} \leq 0$ in Σ_{λ_0} and $w_{\lambda_0} \neq 0$ on $\partial\Sigma_{\lambda_0}$. Now by Theorem 5 we have $w_\lambda < 0$ in Σ_{λ_0}. We will show that for a small $\epsilon > 0$ we have $w_{\lambda_0-\epsilon} < 0$ in $\Sigma_{\lambda_0-\epsilon}$, thus getting a contradiction that (λ_0, a) is the largest interval of values of λ such that $w_\lambda < 0$ in Σ_λ.

Fix $\theta > 0$ (to be determined). Let K be a closed subset in Σ_{λ_0} such that $|\Sigma_{\lambda_0-\epsilon} \setminus K| < \theta/2$. The fact $w_{\lambda_0} < 0$ in Σ_{λ_0} implies $w_{\lambda_0}(x) \leq -p < 0$ for any $x \in K$ and some $p > 0$. By continuity we have $w_{\lambda_0-\epsilon} < 0$ in K. For $\epsilon > 0$ small, $|\Sigma_{\lambda_0-\epsilon} \setminus K| < \delta$.

We choose δ in such a way that we may apply Theorem 3 to $w_{\lambda_0-\epsilon}$ in $\Sigma_{\lambda_0-\epsilon} \setminus K$. Hence we get $w_{\lambda_0-\epsilon} \leq 0$ in $\Sigma_{\lambda_0-\epsilon} \setminus K$.

Therefore we obtain that for any $\epsilon > 0$ small enough, we have $w_{\lambda_0-\epsilon}(x) \leq 0$ in $\Sigma_{\lambda_0-\epsilon}$. Again, using corollary 6, we get $w_{\lambda_0-\epsilon}(x) < 0$ in $\Sigma_{\lambda_0-\epsilon}$. Therefore, $\lambda_0 = 0$.

Step 3. We have $w_\lambda \leq 0$ for all $\lambda \in (0, a)$. Applying now Corollary 6 and Lemma 4 to the equation

$$\begin{aligned} &\Delta w_\lambda - \left[c^-(x, \lambda) + \frac{\delta}{u_\gamma^{\delta+1}}\right] w_\lambda = c^+(x, \lambda) w_\lambda \ \text{ in } \Sigma_\lambda, \\ &w_\lambda \leq 0 \ \text{ and } \ w_\lambda \neq 0 \ \text{ on } \ \partial\Sigma_\lambda, \end{aligned}$$

we have $w_\lambda < 0$ for $\lambda \in (0, a)$.
Note that w_λ admits its maximum along $\Sigma_\lambda \cap \Omega$. Again applying the next part of Lemma 4 we have

$$D_{x_1} w_\lambda|_{x_1=\lambda} = 2 D_{x_1} u_\lambda|_{x_1=\lambda} < 0.$$

The proof of Theorem 1 is now complete.

§4. Some a priori estimates

In this section we will produce some a priori results for (1) with the function f being replaced by a specific type of non-linearity. The equation is given by:

$$\begin{aligned} -\Delta u - \frac{1}{u^\delta} &= R(x) u^\alpha \ \text{ in } \Omega, \\ u = 0 \ \text{ on } \partial\Omega, &\ u > 0 \text{ in } \Omega, \end{aligned} \tag{2}$$

where R is continuous and strictly positive function in $\overline{\Omega}$ and $1 < \alpha < \frac{n+2}{n-2}$ with $\delta > 0$ is given.

We want to find some a priori estimates on the solutions of the above equation i.e., we show a uniform bound for the solutions and we achieve that goal with the help of a blow-up technique in a compact subset of Ω. For the rest of the domain, we apply Theorem 1 for deriving a uniform bound of solutions in a neighborhood of $\partial\Omega$.

We start by a lemma which is a global result of Liouville type (see [8]).

Lemma 7. *Let $u(x)$ be a non-negative C^2 solution of*

$$\Delta u + u^\alpha = 0 \ in \ \mathbb{R}^n \tag{3}$$

with $1 < \alpha < (n+2)/(n-2)$. *Then* $u(x) \equiv 0$.

Remark 2. Our main result is for f depending only on u but the same thing holds for $f(x, u)$ with f is a locally lipchitz w.r.t the second variable and continuous w.r.t the first variable.

To prove the result we need few lemmata. First we state here a result of [2].

Lemma 8. *Consider the equation given by*

$$-\Delta u = \frac{1}{u^\delta} \text{ in } \Omega$$
$$u = 0 \text{ on } \partial\Omega.$$

Then there exists unique solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$. Moreover we can find $0 < c_0 \le c_1$ such that

1. *For $0 < \delta < 1$, we have $c_0 d(x) \le u \le c_1 d(x)$.*
2. *For $\delta = 1$, we have $c_0 d(x) \ln (A/d(x))^{1/2} \le u \le c_1 d(x) \ln (A/d(x))^{1/2}$ where $A > 1$ is large enough.*
3. *For $\delta > 1$, we have $c_0\{d(x)\}^{2/(\delta+1)} \le u \le c_1\{d(x)\}^{2/(\delta+1)}$.*

The above result together with the comparison principle show that any non trivial solution u to (2) satisfies $u(x) \ge cd(x)$ with $c > 0$ independent of u. Next we state a strong comparison principle (see [6] for the extension in the case of quasilinear elliptic operators):

Lemma 9. *Let $u, v(\ge 0) \in C^2(\Omega) \cap C(\overline{\Omega})$ and satisfies*

$$-\Delta u - u^{-\delta} = f,$$
$$-\Delta v - v^{-\delta} = g,$$

with $u = v = 0$ on $\partial\Omega$, $0 < \beta < 1$ with $f, g \in C(\overline{\Omega})$ such that $0 \le f \le g$ pointwise everywhere in Ω and $f \not\equiv g$. Then $0 < u < v$ in Ω.

Now we are ready to proceed to the main result of this section:

Theorem 10. *Suppose that Ω is a bounded domain which is strictly convex. Suppose $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a positive solution of*

$$\begin{aligned} -\Delta u - \frac{1}{u^\delta} &= R(x)u^\alpha \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{4}$$

where $\delta > 0$, $1 < \alpha < (n + 2)/(n - 2)$ and R is continuous and strictly positive function in $\overline{\Omega}$. Then $u(x) < C$ for some uniform constant C where C only depends α and Ω .

Proof. We are going to divide the domain into two parts given by:

$$\Omega_\eta = \{x \in \Omega \mid \text{dist}(x, \partial\Omega) \ge \eta\},$$
$$\Omega \setminus \Omega_\eta = \{x \in \Omega \mid \text{dist}(x, \partial\Omega) < \eta\},$$

where $\eta > 0$ is small enough.

We proof the theorem by contradiction. Let on the contrary there exists a sequence of solutions $u^k(x)$ of (4) and a sequence of points $P_k \in \Omega_\eta$ such that $M_k = \sup_\Omega u^k(x) = u^k(P_k) \to +\infty$ as $k \to +\infty$.

We first prove that $P_k \to P \in \Omega_\eta$. For that, we apply the moving plane method as in the previous section . Applying the method used for the proof of Theorem 1 (see also [3]) and the convexity of Ω (precisely, we move the hyperplane in a direction close to the outward normal in a neighborhood of any point of the boundary), we have a $H > 0$ (depending on the domain and independent of k) and a $T > 0$ such that:

$u_k(x - t\gamma)$ is decreasing for $t \in [0, T]$ for $\gamma \in \mathbb{R}^n$ satisfying $|\gamma| = 1$ and $(\gamma.n(x)) \geq H$, $n(x)$ is the unit normal to $\partial\Omega$ at x and for $x \in \partial\Omega$.

The fact that $u_k(x - t\gamma)$ is non-decreasing in t for x, t and γ decribed above we have to positive numbers α_1 and α_2 both depending on Ω such that, for any x belonging to $\Omega \setminus \Omega_{\alpha_2} = \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) < \alpha_2\}$, we have a measurable set I_x with

- $|I_x| \geq \alpha_1$,
- $I_x \subset \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) \geq \alpha_2/2\}$,
- $u_k(\kappa) \geq u_k(x)$ for all κ in I_x.

Then, multiplying the equation satisfied by u_k by the L^1-normalised positive eigenfunction ϕ_1 associated to the first eigenvalue,

$$\lambda_1(\Omega) := \inf_{u \in H_0^1(\Omega),\, u \not\equiv 0} \frac{\int_\Omega |\nabla u|^2}{\int_\Omega u^2},$$

we get that

$$\lambda_1(\Omega) \int_\Omega u_k \phi_1 \mathrm{d}x = \int_\Omega \frac{\phi_1}{u_k^\delta} \,\mathrm{d}x + \int_\Omega R(x) u_k^\alpha \phi_1 \,\mathrm{d}x.$$

Observing that, for any $\ell > \lambda_1(\Omega)$, there exists $C > 0$ such that

$$\frac{1}{t^\delta} + R(x) t^\alpha \geq \ell t - C \quad \text{for any } \ t \in \mathbb{R}^+ \text{ and uniformly for } x \in \Omega.$$

Then, fixing $\ell > \lambda_1(\Omega)$, it follows that

$$(\ell - \lambda_1(\Omega)) \int_\Omega u_k \phi_1 \leq C.$$

Thus, from above, we get for $x \in \Omega \setminus \Omega_{\alpha_2}$

$$u_k(x) \int_{I_x} \phi_1 \mathrm{d}x \leq \int_{I_x} u_k \phi_1 \leq C.$$

Then, $u_k(x) \leq C/|I_x|^{1/2} \leq C/\alpha_1$ for $x \in \Omega \setminus \Omega_{\alpha_2}$. Therefore, $\operatorname{dist}(M_k, \partial\Omega) \geq \alpha_2$. We now apply the blow-up analysis of [9].

Let $B_R(a)$ denote a ball with radius R and centre $a \in \mathbb{R}^n$. Let λ_k be a sequence of positive numbers(to be defined later) and $y = (x - P_k)/\lambda_k$. Define the scaled function

$$v_k(y) = \lambda_k^{2/(\alpha-2)} u_k(x).$$

We choose λ_k so that $\lambda_k^{2/(\alpha-2)} M_k = 1$. Since $M_k \to +\infty$, we have $\lambda_k \to 0$ as $k \to +\infty$. For large k, $v_k(y)$ is well-defined in $B_{\eta/\lambda_k}(0)$, and

$$\sup_{y \in B_{\eta/\lambda_k}(0)} v_k(y) = v_k(0) = 1.$$

Moreover, $v_k(y)$ satisfies in $B_{\eta/\lambda_k}(0)$ the following equations:

$$\begin{aligned} -\lambda_k^{-2\alpha/(\alpha-1)} \Delta u_k - \lambda_k^{2\delta/(\alpha-1)} [v_k]^{-\delta} &= R(\lambda_k y + P_k) \lambda_k^{-2\alpha/(\alpha-1)} [v_k]^{\alpha}, \\ -\Delta v_k &= \lambda_k^{2(\alpha+\delta)/(\alpha-1)} [v_k]^{-\delta} + R(\lambda_k y + P_k)[v_k]^{\alpha}. \end{aligned}$$

From Lemma 9, we have $u_k \geq c_0 \{d(x)\}^{\alpha}$, where α depends on δ. Again by Lemma 10 we have $v_k \geq \lambda_k^{2/(\alpha-2)} u_k$. Combining these two results we have $v^k \geq p(>0)$ in Ω_η with p depending upon η and δ

Therefore given any radius R such that $B_R(0) \subset B_{\frac{\eta}{\lambda_k}}(0)$ we can, by elliptic L^p estimates, find uniform bounds for $\|v_k\|_{W^{2,p}(B_R(0))}$. Choosing p large we obtain by Morrey's embedding theorem that $\|v_k\|_{C^{1,\beta}(B_R(0))}$ for $0 < \beta < 1$ is also uniformly bounded. So for any sequence $k \to +\infty$, there exists a subsequence $k_j \to +\infty$ such that $v^{k_j} \to v$ in $W^{2,p} \cap C^{1,\beta}$, $p > n$ on $B_R(0)$. By Holder Continuity $v(0) = 1$ again since $R(\lambda_k y + P_k) \to R(P)$ as $k \to +\infty$, we have that

$$\begin{aligned} -\Delta v &= R(P) v^{\alpha}, \\ v(0) &= 1. \end{aligned}$$

We claim that v is well-defined in all of $\mathbb{R}^n$ and $v_{k_j} \to v$ in $W^{2,p} \cap C^{1,\beta}$, $p > n$ on compact subsets. To show this we consider $B_R(0) \subset B'_R(0)$. Repeating the above argument with $B'_R(0)$, the subsequence $v^{k'_j}$ has a convergent subsequence $v^{k'_j} \to v'$ on $B'_R(0)$, v' satisfies $\lambda_k^{2/(\alpha-2)} M^k = 1$ and if restricted to $B_R(0)$ gives v. By unique continuation, the entire original sequence converges and v is well defined. By Lemma 4, we have $v = 0$ in $\mathbb{R}^n$, a contradiction since $v(0) = 1$.

This completes the proof. □

The existence of a priori bounds together with the theory of global bifurcation in the context of singular problems (see [12] and the extension for more singular nonlinearities [4]) can be used to prove existence of multiple solutions. Precisely, let us consider the following problem where $\lambda \in \mathbb{R}^+$ is a parameter:

$$\begin{aligned} -\Delta u &= \lambda \left(\frac{1}{u^{\delta}} + R(x) u^{\alpha} \right) \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \ u > 0 \text{ in } \Omega. \end{aligned} \tag{5}$$

In particular, we can prove the following result:

Theorem 11. *Let $\delta \in (0,3)$ and $1 < \alpha < (n+2)/(n-2)$. Then, there exists an unbounded connected set $C \subset \mathbb{R}^+ \times \left(L^\infty(\Omega) \cap H_0^1(\Omega)\right)$ of solutions (λ, u) to (5) such that*

(i) there exists $\Lambda > 0$ such that $\Pi_{\mathbb{R}} C = [0, \Lambda]$;

(ii) for any $\lambda \in (0, \Lambda)$, there exists two solutions (λ, u_λ) and (λ, v_λ) belonging to C and such that $u_\lambda < v_\lambda$ in Ω.

The above theorem can be proved by showing that the conected component set of the minimal solutions curve admits a turning point at $\lambda = \Lambda$ and from the existence of universal bounds at $\lambda > 0$ bends back to $\lambda = 0$ where the branch admits an asymptotic bifurcation point.

Acknowledgements

We want to thank Prof. Jesús Hernández for a fruitful discussion of the topic.

References

[1] Coclite, M. M., and Palmieri, G. On a singular nonlinear Dirichlet problem. *Comm. Partial Differential Equations 14*, 10 (1989), 1315–1327.

[2] Crandall, M. G., Rabinowitz, P. H., and Tartar, L. On a Dirichlet problem with a singular nonlinearity. *Comm. Partial Differential Equations 2*, 2 (1977), 193–222.

[3] de Figueiredo, D. G., Lions, P.-L., and Nussbaum, R. D. A priori estimates and existence of positive solutions of semilinear elliptic equations. *J. Math. Pures Appl. (9) 61*, 1 (1982), 41–63.

[4] Dhanya, R., Giacomoni, J., Prashanth, S., and Saoudi, K. Global bifurcation and local multiplicity results for elliptic equations with singular nonlinearity of super exponential growth in two dimensions. To appear.

[5] Giacomoni, J., and Saoudi, K. Multiplicity of positive solutions for a singular and critical problem. *Nonlinear Anal. 71*, 9 (2009), 4060–4077.

[6] Giacomoni, J., Schindler, I., and Takáč, P. Sobolev versus Hölder local minimizers and existence of multiple solutions for a singular quasilinear equation. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) 6*, 1 (2007), 117–158.

[7] Gidas, B., Ni, W. M., and Nirenberg, L. Symmetry and related properties via the maximum principle. *Comm. Math. Phys. 68*, 3 (1979), 209–243.

[8] Gidas, B., and Spruck, J. Global and local behaviour of positive solutions of nonlinear elliptic equations. *Comm. Pure Appl. Math. 34*, 4 (1981), 525–598.

[9] Gidas, B., and Spruck, J. A priori bounds for positive solutions of nonlinear elliptic equations. *Comm. Partial Differential Equations 6*, 8 (1981), 883–901.

[10] Haitao, Y. Multiplicity and asymptotic behavior of positive solutions for a singular semilinear elliptic problem. *J. Differential Equations 189*, 2 (2003), 487–512.

[11] Han, Q., and Lin, F. *Elliptic partial differential equations*, vol. 1 of *Courant Lecture Notes in Mathematics*. New York University Courant Institute of Mathematical Sciences, New York, 1997.

[12] Hernández, J., Mancebo, F. J., and Vega, J. M. On the linearization of some singular, nonlinear elliptic problems and applications. *Ann. Inst. H. Poincaré Anal. Non Linéaire 19*, 6 (2002), 777–813.

[13] Sun, Y., Wu, S., and Long, Y. Combined effects of singular and superlinear nonlinearities in some singular boundary value problems. *J. Differential Equations 176*, 2 (2001), 511–531.

Kaushik Bal and Jacques Giacomoni
LMAP (UMR 5142), Bat. IPRA
Université de Pau et des Pays de l'Adour
Avenue de l'Université, 64013 cedex Pau, France
`kausbal@gmail.com` and `jacques.giacomoni@univ-pau.fr`

Monografías Matemáticas *García de Galdeano* **37**, 37–46 (2012)

Numerical simulation of anisothermal Newtonian flows

Nelly Barrau and Daniela Capatina

Abstract. We are interested in the finite element approximation of the Navier-Stokes equations with variable density and with heat transfer. We discuss the choice of compatible discretizations and we investigate the stability of the Jacobian matrix in a simplified framework. We propose to introduce the mass flux and to use Raviart-Thomas elements for its discretization, nonconforming elements for the velocity and a DG method for the temperature. Finally, some numerical tests are presented.

Keywords: Compressible Navier-Stokes equations, anisothermal flow, finite elements, stabilization.

AMS classification: 65M60, 76M10, 80A20.

§1. Introduction

We are interested here in the approximation of 2D anisothermal flows for Newtonian fluids. The governing equations are the momentum, mass and energy conservation laws together with the constitutive equation and a state equation, in a polygonal domain $\Omega \subset \mathbb{R}^2$:

$$\begin{cases} \rho\left(\dfrac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) - \operatorname{div}\underline{\tau} + \nabla p = \mathbf{f}, \\ \dfrac{\partial \rho}{\partial t} + \operatorname{div}(\rho\mathbf{v}) = 0, \\ \rho C_p\left(\dfrac{\partial T}{\partial t} + \mathbf{v}\cdot\nabla T\right) - k\triangle T = Q, \\ \rho = \rho(p,T), \\ \underline{\tau} = 2\eta\underline{D}(\mathbf{v}). \end{cases} \tag{1}$$

We close the system by imposing initial and boundary conditions. The unknowns are the velocity $\mathbf{v}$, the stress tensor $\underline{\tau}$, the pressure p, the temperature T and the density ρ. The viscosity η, the thermal conductivity k and the heat capacity C_p are given constants.

This preliminary study is devoted to the development of a stable finite element approximation of problem (1) and to its implementation in the C++ library Concha. The further goal is the extension to more complex anisothermal flows, for instance to compressible gases or to viscoelastic non-Newtonian fluids. Therefore, we propose to keep the density as an unknown of the problem in order to allow the treatment of different state equations, such as $p = \rho RT$ for a gas or $\rho = \rho_0(1 - \beta(T - T_0))$ for a polymeric liquid, with R the gas constant, β the dilatation coefficient and ρ_0 , T_0 some reference values.

As regards the constitutive law, it is obvious that in the Newtonian case the stress tensor can be eliminated from the equations, which is no longer possible when dealing with non-Newtonian fluids. For instance, for the polymeric liquids which we want to treat in the future the constitutive law can be usually written as follows:

$$\lambda\left(\frac{\partial}{\partial t}\underline{\tau} + \mathbf{v}\cdot\nabla\underline{\tau} - \underline{\tau}\nabla\mathbf{v}^T - \nabla\mathbf{v}\underline{\tau}\right) + \underline{\tau} + f(\underline{\tau}) = 2\eta\underline{D}(\mathbf{v})$$

and yields, at constant density and constant temperature, a three-fields formulation in $(\mathbf{v}, p, \underline{\tau})$. This aspect has been treated in the incompressible isothermal case in [3]. Here, we only focus on the velocity-pressure formulation for Newtonian fluids.

§2. Choice of compatible discretizations

We present in the sequel some numerical difficulties related to the approximation of (1), as well as our choice of discretization.

2.1. Incompressible Navier-Stokes equations

We begin by considering the stationary Stokes equations:

$$\begin{cases} -\eta\triangle\mathbf{v} + \nabla p = \mathbf{f} & \text{in } \Omega, \\ \operatorname{div}\mathbf{v} = 0 & \text{in } \Omega, \end{cases} \tag{2}$$

with homogeneous Dirichlet boundary conditions, for simplicity of presentation.

Its finite element discretization is very well studied in the literature and several methods exist, each one with its own advantages and disadvantages. Thus, one may employ finite element spaces for the velocity and the pressure which satisfy an inf-sup condition (see [2] for a review), or choose the two discrete spaces independently but then add a stabilization term in order to ensure the uniform coercivity of the matrix. Completely discontinuous discrete spaces can also be employed, leading to a discontinuous Galerkin (DG) method which is known to be flexible but quite expensive from a computational point of view.

Among the inf-sup stable spaces, there are the conforming and the nonconforming approximations. We have chosen to use here low-order nonconforming finite elements either on triangles or on quadrilaterals, due to their well-known stability and their reduced stencil. Note that in the triangular case, the mass matrix is diagonal and we recover a divergence free discrete velocity. These spaces also present certain advantages concerning the adaptivity. We are using Crouzeix-Raviart [1] elements on triangles, respectively Rannacher-Turek [6] elements on quadrilaterals, whose degrees of freedom are the mean values across the edges. The finite dimensional spaces for the velocity are defined as follows:

$$\mathbf{V}_h^{CR} = \left\{\mathbf{v} \in \mathbf{L}^2(\Omega)\,;\ \forall\, T \in \mathcal{T}_h,\ \mathbf{v}|_T \in \mathbf{P}_1,\ \forall\, e \in \mathcal{S}_h,\ \int_e [\mathbf{v}]\,\mathrm{ds} = 0\right\},$$

$$\mathbf{V}_h^{RT} = \left\{\mathbf{v} \in \mathbf{L}^2(\Omega)\,;\ \forall\, T \in \mathcal{T}_h,\ \mathbf{v}|_T \in \mathbf{Q}_T,\ \forall\, e \in \mathcal{S}_h,\ \int_e [\mathbf{v}]\,\mathrm{ds} = 0\right\},$$

where $\mathbf{Q}_T = (Q_T)^2$ with $Q_T = \{v\,;\, v \circ \Psi_T \in \hat{Q}^{\mathrm{rot}}\}$, $\hat{Q}^{\mathrm{rot}} = \mathrm{vect}\{1, \hat{x}, \hat{y}, \hat{x}^2 - \hat{y}^2\}$ and $\Psi_T : \hat{T} \to T$ the bilinear one-to-one transformation of the square $\hat{T} = [-1, 1]^2$. We employ the usual notation $[\cdot]$ for the jump across en edge $e \in \mathcal{S}_h$ of the triangulation; the jump is equal to the trace if $e \subset \partial\Omega$. The pressure is looked for in

$$M_h = \left\{p \in L^2_0(\Omega);\ \forall\, T \in \mathcal{T}_h,\ p_{|T} \in P_0\right\}.$$

As regards now the instationary Navier-Stokes equations, it is well-known that the discretization of the additionnal nonlinear term $\mathbf{v} \cdot \nabla\mathbf{v}$ is more delicate since it necessitates stabilization. Several schemes such as SUPG, LPS or edge stabilization were proposed in the literature and are implemented in the library Concha. The approximation of the time derivative $\partial\mathbf{v}/\partial t$ is more standard, and several schemes (implicit and explicit Euler, Crank-Nicolson, BDF) are available in Concha. These aspects will be detailed in the next section, since their treatment is specific to the change of variables that we propose in the compressible case.

2.2. Compressible Navier-Stokes equations

The density ρ is now an additionnal unknown, and we have to solve the following system :

$$\left\{\begin{aligned} \rho\left(\frac{\partial\mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) - \eta\triangle\mathbf{v} + \nabla p = \mathbf{f} \quad & \text{in } \Omega,\\ \frac{\partial\rho}{\partial t} + \mathrm{div}(\rho\mathbf{v}) = 0 \quad & \text{in } \Omega,\\ \rho = \rho(p) \quad & \text{in } \Omega. \end{aligned}\right. \tag{3}$$

The numerical treatment of the convective term $\rho\mathbf{v} \cdot \nabla\mathbf{v}$ necessitates the design of adequate stabilization techniques, and is still an active and open research topic. To tackle it, we have chosen here to introduce the mass flux as an auxiliary variable $\mathbf{G} = \rho\mathbf{v}$ belonging to $\mathbf{H}(\mathrm{div}, \Omega)$. For its discretization, we employ lowest-order Raviart-Thomas finite elements (see [7]), which are known to be $\mathbf{H}(\mathrm{div}, \Omega)$-conforming. More precisely, $\mathbf{G}_h$ is looked for in the space $\mathbf{W}_h = \{\mathbf{w} \in \mathbf{H}(\mathrm{div}, \Omega)\,;\, \forall\, T \in \mathcal{T}_h,\ \mathbf{w}|_T \in \mathbf{RT}_0\}$ where $\mathbf{RT}_0$ is defined as follows: $\mathbf{RT}_0 = \mathbf{P}_0 \oplus \mathbf{x}P_0$ on triangles, respectively $\mathbf{RT}_0 = P_1[x] \times P_1[y]$ on quadrilaterals. The degrees of freedom are the normal fluxes across the edges of the triangulation. It is useful to recall that the interpolation operator $\mathcal{E}_h$ of [7] satisfies, besides classical errors estimates, the following properties on every $T \in \mathcal{T}_h$ and $e \in \mathcal{S}_h$ respectively:

$$\mathrm{div}\,(\mathcal{E}_h\mathbf{w}) = \pi_0\,(\mathrm{div}\,\mathbf{w})\,, \qquad \mathcal{E}_h\mathbf{w}\cdot\mathbf{n} = \pi_0(\mathbf{w}\cdot\mathbf{n}),$$

where π_0 is the L^2-orthogonal projection on P_0.

We are next interested in the stability of the discrete steady problem. To highlight the structure of the corresponding operator, let us consider a simple state equation, let's say $\rho = C\,p$ with C constant:

$$\left\{\begin{aligned} -\eta\triangle\mathbf{v} + \nabla p = \mathbf{f} \quad & \text{in } \Omega,\\ -Cp\mathbf{v} + \mathbf{G} = \mathbf{0} \quad & \text{in } \Omega,\\ \mathrm{div}\,\mathbf{G} = 0 \quad & \text{in } \Omega. \end{aligned}\right.$$

We apply Newton's method and at each iterate, we obtain the variational formulation:

$$\begin{cases} (\mathbf{v}_h,\, \mathbf{G}_h,\, p_h) \in \mathbf{V}_h \times \mathbf{W}_h \times M_h, \\ \forall \mathbf{v}'_h \in \mathbf{V}_h,\ \eta \displaystyle\int_{\Omega_h} \nabla \mathbf{v}_h \cdot \nabla \mathbf{v}'_h \,\mathrm{dx} - \int_{\Omega} p_h \operatorname{div} \mathbf{v}'_h \,\mathrm{dx} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v}'_h \,\mathrm{dx}, \\ \forall \mathbf{G}'_h \in \mathbf{W}_h,\ -C \displaystyle\int_{\Omega} p_h^n \mathbf{v}_h \cdot \mathbf{G}'_h \,\mathrm{dx} - C \int_{\Omega} p_h \mathbf{v}_h^n \cdot \mathbf{G}'_h \,\mathrm{dx} + \int_{\Omega} \mathbf{G}_h \cdot \mathbf{G}'_h \,\mathrm{dx} = 0, \\ \forall p'_h \in M_h,\ \displaystyle\int_{\Omega} p'_h \operatorname{div} \mathbf{G}_h \,\mathrm{dx} = 0. \end{cases}$$

The corresponding Jacobian matrix can be written as follows:

$$\mathcal{J} = \begin{pmatrix} A & 0 & \vdots & B_1 \\ A_1 & I & \vdots & B_2 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & B_3 & \vdots & 0 \end{pmatrix} = \begin{pmatrix} \mathcal{A} & \mathcal{B}_1 \\ \mathcal{B}_2 & 0 \end{pmatrix},$$

with $\mathcal{B}_1 \neq \mathcal{B}_2^T$ and $\mathcal{A}$ non-symmetric. In order to show that $\mathcal{J}$ is invertible, we shall apply a generalization of the Babuska-Brezzi theorem which was given by Nicolaides in [5]. We have then to check three discrete inf-sup conditions on $\mathcal{B}_1$, $\mathcal{B}_2$ and $\mathcal{A}$ respectively, the latter one on $\operatorname{Ker} \mathcal{B}_2 \times \operatorname{Ker} \mathcal{B}_1$.

Proposition 1. *There exists* $\beta_1 > 0$ *independent of h such that*

$$\inf_{p \in M_h} \sup_{(\mathbf{v}, \mathbf{G}) \in \mathbf{V}_h \times \mathbf{W}_h} \frac{-\int_\Omega p \operatorname{div} \mathbf{v} \,\mathrm{dx} - C \int_\Omega p \mathbf{v}_h^n \cdot \mathbf{G} \,\mathrm{dx}}{\|p\|_{0,\Omega} \left(|\mathbf{v}|_{1,h} + \|\mathbf{G}\|_{\mathbf{H}(\mathrm{div},\Omega)} \right)} \geq \beta_1 .$$

Proof. The proof is identical to the one of the classical inf-sup condition for the two-fields formulation of the Stokes problem on $\mathbf{V}_h \times M_h$ (see for instance [2]), by taking $\mathbf{G} = \mathbf{0}$. □

Proposition 2. *There exists* $\beta_2 > 0$ *independent of h such that*

$$\inf_{p \in M_h} \sup_{\mathbf{G} \in \mathbf{W}_h} \frac{-\int_\Omega p \operatorname{div} \mathbf{G} \,\mathrm{dx}}{\|p\|_{0,\Omega} \|\mathbf{G}\|_{\mathbf{H}(\mathrm{div},\Omega)}} \geq \beta_2 .$$

Proof. The proof is well-known, see [7]. For $p \in M_h$, one considers the auxiliary problem:

$$\begin{cases} -\triangle z = p & \text{in } \Omega, \\ z = 0 & \text{on } \partial\Omega, \end{cases}$$

and takes $\mathbf{w} = \nabla z$ which belongs, thanks to the regularity of the Laplace operator, to $\mathbf{H}^a(\Omega)$ with $a > 1/2$. Let then the Raviart-Thomas interpolate $\mathbf{G} = \mathcal{E}_h \mathbf{w}$. According to the properties of $\mathcal{E}_h$, one has $\operatorname{div} \mathbf{G} = -p$ and $\|\mathbf{G}\|_{\mathbf{H}(\mathrm{div},\Omega)} \leq c \|p\|_{0,\Omega}$, which implies the uniform inf-sup condition on $\mathcal{B}_2$. □

Proposition 3. *There exists* $\alpha > 0$ *such that:*

$$\inf_{(\mathbf{v},\mathbf{G})\in \operatorname{Ker}\mathcal{B}_2}\ \sup_{(\mathbf{v}',\mathbf{G}')\in \operatorname{Ker}\mathcal{B}_1} \frac{\mathcal{A}\left((\mathbf{v},\mathbf{G}),(\mathbf{v}',\mathbf{G}')\right)}{\|(\mathbf{v},\mathbf{G})\|_{\mathbf{H}_0^1(\Omega)\times\mathbf{H}(\mathrm{div},\Omega)}\,\|(\mathbf{v}',\mathbf{G}')\|_{\mathbf{H}_0^1(\Omega)\times\mathbf{H}(\mathrm{div},\Omega)}} \geq \alpha,$$
$$\inf_{(\mathbf{v}',\mathbf{G}')\in \operatorname{Ker}\mathcal{B}_1}\ \sup_{(\mathbf{v},\mathbf{G})\in \operatorname{Ker}\mathcal{B}_2} \frac{\mathcal{A}\left((\mathbf{v},\mathbf{G}),(\mathbf{v}',\mathbf{G}')\right)}{\|(\mathbf{v},\mathbf{G})\|_{\mathbf{H}_0^1(\Omega)\times\mathbf{H}(\mathrm{div},\Omega)}\,\|(\mathbf{v}',\mathbf{G}')\|_{\mathbf{H}_0^1(\Omega)\times\mathbf{H}(\mathrm{div},\Omega)}} > 0.$$

Proof. These two inf-sup conditions translate the fact that the matrix $\mathcal{A}$ is invertible on $\operatorname{Ker}\mathcal{B}_2 \times \operatorname{Ker}\mathcal{B}_1$. Since $\mathcal{A} = \left(\begin{smallmatrix} A & 0 \\ A_1 & I \end{smallmatrix}\right)$ is block triangular, it is therefore sufficient to show the invertibility of A. Thanks to the discrete Poincaré inequality on the nonconforming spaces, A is uniformly invertible on the whole space $\mathbf{V}_h$. Thus, the statement is established. □

Let us now discuss the complete system (3). One may choose between two options: write the particular derivative of the first equation in conservative form $\partial\mathbf{G}/\partial t + \operatorname{div}(\mathbf{G}\otimes\mathbf{v})$, or keep $\rho\partial\mathbf{v}/\partial t + (\mathbf{G}\cdot\nabla)\mathbf{v}$. We have chosen here the latter variant. For the discretization of the convective term, we propose the stabilization:

$$\int_{\mathcal{T}_h} (\mathbf{G}_h\cdot\nabla)\mathbf{v}_h\cdot\mathbf{v}'_h\,\mathrm{dx} \approx -\int_{\mathcal{T}_h}\Big((\operatorname{div}\mathbf{G}_h)\mathbf{v}_h\cdot\mathbf{v}'_h + (\mathbf{G}_h\cdot\nabla)\mathbf{v}'_h\cdot\mathbf{v}_h\Big)\mathrm{dx} + \int_{\mathcal{S}_h}\mathbf{F}_e\left(\mathbf{G}_h,\mathbf{v}_h\right)\cdot\left[\mathbf{v}'_h\right]\,\mathrm{ds},$$

where $\mathbf{F}_e(\mathbf{G}_h,\mathbf{v}_h) = (\mathbf{G}_h\cdot\mathbf{n}_e)^+\,\mathbf{v}_h^{\mathrm{in}} + (\mathbf{G}_h\cdot\mathbf{n}_e)^-\,\mathbf{v}_h^{\mathrm{ex}}$ represents the numerical flux and $\mathbf{n}_e$ is a unit normal to the edge e . For a given piecewise continuous function φ, we have denoted $\varphi^{\mathrm{ex}}(\mathbf{x}) = \lim_{\varepsilon\to 0}\varphi(\mathbf{x}-\varepsilon\mathbf{n}_e)$, $\varphi^{\mathrm{in}}(\mathbf{x}) = \lim_{\varepsilon\to 0}\varphi(\mathbf{x}+\varepsilon\mathbf{n}_e)$ and $[\varphi] = \varphi^{\mathrm{ex}}-\varphi^{\mathrm{in}}$. Then we end up with another matrix $\mathcal{A}^* = \left(\begin{smallmatrix} A^* & A_2 \\ A_1 & I \end{smallmatrix}\right)$ instead of $\mathcal{A}$, for which the inf-sup conditions of Proposition 3 should be established. Note that for $\operatorname{div}\mathbf{G}_h = 0$, the diffusion-convection operator A^* is uniformly coercive on $\mathbf{V}_h$ since one can show that

$$A^*(\mathbf{v}_h,\mathbf{v}_h) = \eta\,|\mathbf{v}_h|_{1,h}^2 + \frac{1}{2}\int_{\mathcal{S}_h}|\mathbf{G}_h\cdot\mathbf{n}_e|\,[\mathbf{v}_h]\cdot[\mathbf{v}_h]\,\mathrm{ds}.$$

For the discretization of the time derivative $\rho\partial\mathbf{v}/\partial t$, we have employed the BDF (*Backward Differential Formula*) scheme of order 2, for its robustness and stability. The variable at t_{n+1} is expressed in terms of the solutions at the two previous time steps as follows:

$$\rho_{n+1}\frac{\partial\mathbf{v}_{n+1}}{\partial t} \approx \rho_{n+1}\left(\frac{1}{\triangle t}\left(\frac{3}{2}\mathbf{v}_{n+1} - 2\mathbf{v}_n + \frac{1}{2}\mathbf{v}_{n-1}\right) + O\left(\triangle t^2\right)\right).$$

The coercivity of the diagonal blocks corresponding to the velocity and the pressure is thus enhanced, but the block $\mathcal{B}_1$ is also modified and a new inf-sup condition should be satisfied.

2.3. Anisothermal flow

Taking into account the thermodynamics is essential in order to obtain realistic simulations.

The energy equation is convection-dominated due to the large value of the heat capacity coefficient C_p. We have chosen to employ a DG method for its discretization, which is known to be well-adapted to such problems (see for instance Lesaint and Raviart [4]). In order to reduce the computational cost and also because $k \ll 1$ while $\rho C_p \approx 10^6$, we use here piecewise constant elements for T. Thus, the discrete diffusion operator on T is reduced to the stabilization term on the edges while the convective term $\mathbf{G}\cdot\nabla T$ is approximated similarly to [4]. The density is approximated by the same finite elements as the temperature.

For the analysis of the corresponding discrete problem, one could apply twice the general results of [5]. To illustrate this, let us consider for the sake of simplicity the steady case and let us neglect the convection in the momentum law. Then the governing equations are:

$$\left\{\begin{aligned} -\eta\triangle\mathbf{v} + \nabla p &= \mathbf{f},\\ \mathbf{G} - \rho\mathbf{v} &= \mathbf{0},\\ -k\triangle T + C_p\mathbf{G}\cdot\nabla T &= 0,\\ \rho + \rho_0\beta T &= \rho_0\left(1 + \beta T_0\right),\\ \operatorname{div}\mathbf{G} &= 0, \end{aligned}\right.$$

and the Jacobian matrix of the discrete problem in the unknowns $(\mathbf{v}, \mathbf{G}, T, \rho, p)$ can be written as follows

$$\mathcal{J}' = \begin{pmatrix} \mathcal{A}' & \mathcal{B}_1' \\ & \\ \mathcal{B}_2' & 0 \end{pmatrix}, \text{ with: } \mathcal{A}' = \begin{pmatrix} A & 0 & 0 & 0 \\ A_1 & I & 0 & B_1 \\ 0 & B_2 & D & 0 \\ 0 & 0 & B_3 & I \end{pmatrix}, \quad \mathcal{B}_1' = \begin{pmatrix} B_4 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathcal{B}_2' = \begin{pmatrix} 0 \\ B_5 \\ 0 \\ 0 \end{pmatrix}^T.$$

The inf-sup conditions on $\mathcal{B}_1'$ and $\mathcal{B}_2'$ are the same as in the previous section, so one only has to check the inf-sup condition for $\mathcal{A}'$ on $\operatorname{Ker}\mathcal{B}_2' \times \operatorname{Ker}\mathcal{B}_1'$ in order to conclude that $\mathcal{J}'$ is invertible. For this purpose, one can decompose $\mathcal{A}'$ in mixed form as follows

$$\mathcal{A}' = \begin{pmatrix} \mathcal{A}'' & \mathcal{B}_1'' \\ & \\ \mathcal{B}_2'' & C \end{pmatrix},$$

where

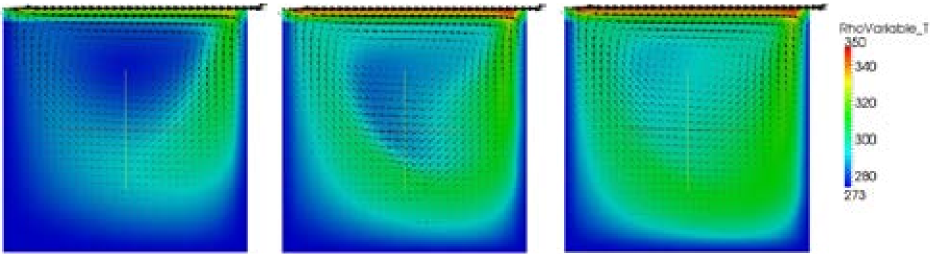
$$\mathcal{A}'' = \begin{pmatrix} A & 0 & 0 \\ A_1 & I & 0 \\ 0 & B_2 & D \end{pmatrix}, \quad \mathcal{B}_1'' = \begin{pmatrix} 0 \\ B_1 \\ 0 \end{pmatrix}, \quad \mathcal{B}_2'' = \begin{pmatrix} 0 \\ 0 \\ B_3 \end{pmatrix}^T, \quad C = I.$$

Since C is clearly positive definite, we can establish inf-sup conditions for $\mathcal{A}''$, $\mathcal{B}_1''$ and $\mathcal{B}_2''$. Note that the latter is obvious, since B_3 corresponds to $\int_\Omega (\rho_0\beta) T\rho \,\mathrm{dx}$. Moreover, A being the nonconforming diffusion operator on $\mathbf{v}$ and D the DG diffusion-convection operator on T, they are uniformly coercive, so that the block triangular matrix $\mathcal{A}''$ is clearly invertible.

In the unsteady case, the time-discretization enforces the coercivity of the diagonal blocks A and D, but also modifies the bilinear forms $\mathcal{B}_2'$ and $\mathcal{B}_1''$. Finally, when taking into account the convective term $\mathbf{G}\cdot\nabla\mathbf{v}$, the first line of $\mathcal{A}''$ is modified and the matrix is no longer block triangular; nevertheless, the new diagonal block A is still uniformly coercive.

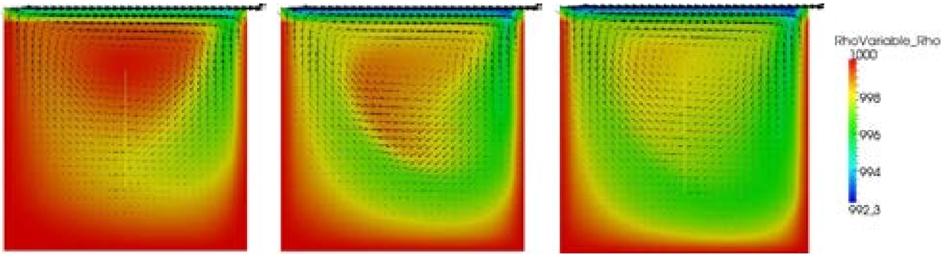
§3. Numerical experiments

We present some of our first numerical results, carried out on two academic tests. Two different fluids have been considered, a polymer with a high viscosity and a liquid with physical



(a) $t = 60\,\mathrm{s}$ (b) $t = 90\,\mathrm{s}$ (c) $t = 120\,\mathrm{s}$

Figure 1: Polymer flow: temperature at different time steps



(a) $t = 60\,\mathrm{s}$ (b) $t = 90\,\mathrm{s}$ (c) $t = 120\,\mathrm{s}$

Figure 2: Polymer flow: density at different time steps

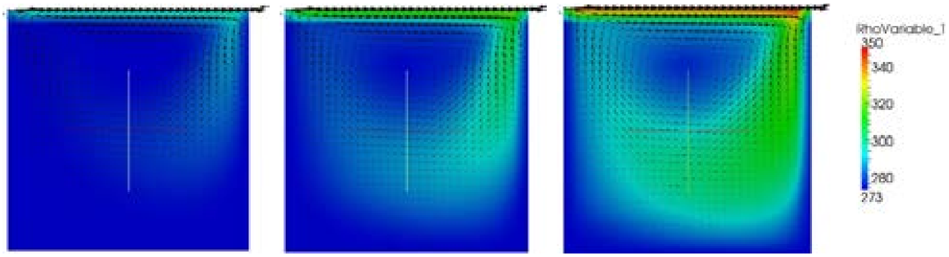
properties similar to those of water. We have taken into account the gravity force and we have considered an affine dependence of the density on the temperature, which corresponds to the case of polymers and which yields the state equation: $\rho = \rho_0\,(1 - \beta\,(T - T_0))$. The next tests are carried out on quadrilateral meshes. The parameters which are common to the numerical experiments are given in the table below:

Parameter	Value
ρ_0: initial density	$1000\,\mathrm{kg/m^3}$
T_0: initial temperature	$273\,\mathrm{K}$
β: dilatation coefficient	$10^{-4}\,\mathrm{m^3/kg\cdot K}$

3.1. Driven cavity: polymer flow

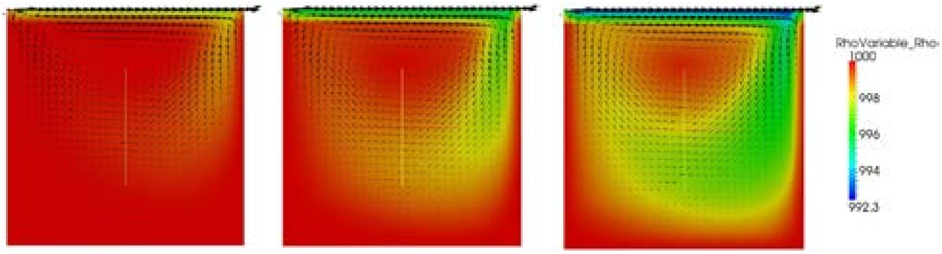
We consider first the driven cavity test in a square Ω. We impose a velocity $\mathbf{v} = (0.03, 0)\,\mathrm{m/s}$ on the top boundary and $\mathbf{0}$ elsewhere, while the temperature equals $350\,\mathrm{K}$ on the top and $273\,\mathrm{K}$ elsewhere. The parameters specific to a polymeric liquid are: the heat capacity $C_p = 2000\,\mathrm{J/kg\cdot K}$, the thermal conductivity $k = 0.05\,\mathrm{W/m\cdot K}$ and the viscosity $\eta = 1000\,\mathrm{Pa\cdot s}$.

One can see in Figures 1 and 2 the evolution of the temperature and of the density. The results of the simulation are physically acceptable. The vortex drags the warm fluid towards the bottom. Due to the gravity force, this one raises slowly to the top and thus it warms the fluid situated between the upper edge and the warm convected fluid.



(a) $t = 25\,\mathrm{s}$ (b) $t = 50\,\mathrm{s}$ (c) $t = 75\,\mathrm{s}$

Figure 3: Water flow: temperature at different time-steps



(a) $t = 25\,\mathrm{s}$ (b) $t = 50\,\mathrm{s}$ (c) $t = 75\,\mathrm{s}$

Figure 4: Water flow: density at different time steps

3.2. Driven cavity: water flow

The specific parameters are now: $C_p = 4186\,\mathrm{J/kg \cdot K}$, $k = 0.6\,\mathrm{W/m \cdot K}$ and $\eta = 0.001\,\mathrm{Pa \cdot s}$. We show in Figures 3 and 4 the temperature and the density (as well as the velocity field) at different time steps. Since the water has a turbulent flow, the stabilization employed in this case is not so efficient; we couldn't simulate a time interval as long as previously.

3.3. Confined flow

The domain is now a rectangle of sides 12 cm and 4 cm. We consider the polymeric liquid previously described and we impose $\mathbf{v} = \mathbf{0}$ on the whole boundary, a constant temperature 273 K on the top and a temperature depending on time and on the abscissa x on the bottom: $T(x,t) = 273 + 100t + 250x$ if $T < 350$ and $T(x,t) = 350$ otherwise. On the vertical boundaries, a homogeneous Neumann condition is set for the temperature.

We show in Figures 5 and 6 the first component of the velocity and the density, as well as the streamlines, at the end of the simulation. Due to the gravity force and to the non-symmetric boundary condition on the bottom, the heat goes up slowly and generates a velocity field.

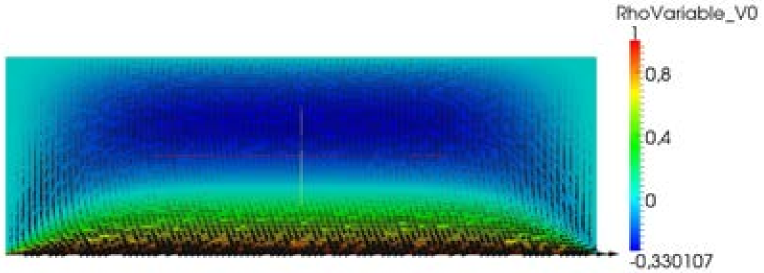


Figure 5: Confined flow: first component of the velocity at the end of simulation

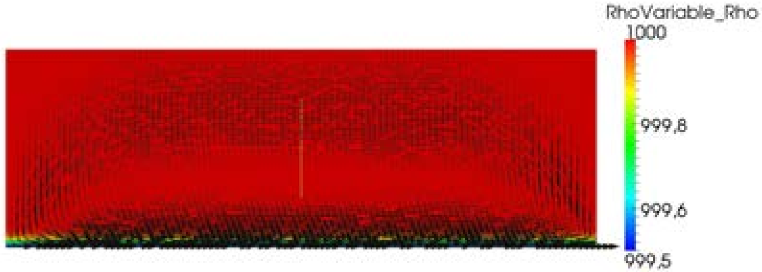


Figure 6: Confined flow: density at the end of simulation

§4. Conclusion and further developments

The anisothermal Navier-Stokes model set up in this work is a first step towards the numerical simulation of more complex flows with heat tranfer, by using the library Concha. We have proposed a finite element method based on the introduction of an additionnal unknown, the mass flux, and investigated the stability of the Jacobian matrix in a simplified framework. In perspective, this study should be extended to a more general case. It will also be interesting to compare this approach with the classical one, written only in the primitive variables.

Although the considered model presents some simplifications (simplified state equation, absence of viscous dissipation in the energy equation), it contains the main difficulties related to this type of problem: compressibility, turbulent flow, dominant convection, significant number of unknowns etc. From a numerical point of view, its treatment necessitated the enrichment of the library Concha in order to take into account a variable density, as well as the implementation of a specific stabilization for certain nonlinear convective terms.

The first numerical results are encouraging, and show that the code gives physically acceptable results. More numerical experiments and comparisons with other softwares such as PolyFlow® or OpenFoam should be carried out in order to further validate the code. As future improvements, we think of using adaptive time steps, iterative solvers and also a local elimination procedure for the mass flux, which amounts to a different stabilization of $\rho\mathbf{v} \cdot \nabla\mathbf{v}$.

Acknowledgements

The first author thanks INRIA for financial support during her training period.

References

[1] Crouzeix, M., and Raviart, P.-A. Conforming and non-conforming finite element methods for solving the stationary Stokes equations. *RAIRO Anal. Numer 7* (1973), 33–76.

[2] Girault, V., and Raviart, P.-A. *Finite Element Methods for Navier-Stokes Equations.* Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1986.

[3] Joie, J. *Simulation numérique des écoulements de liquides polymères.* PhD. Thesis. Université de Pau, 2010.

[4] Lesaint, P., and Raviart, P.-A. On a finite element method for solving the neutron transport. *Mathematical Aspects of Finite Elements in Partial Differential Equations* (1974), 89–123.

[5] Nicolaides, R. Existence, uniqueness and approximation for generalized saddle point problems. *SIAM J. Numer. Anal. 19* (1982), 349–357.

[6] Rannacher, R., and Turek, S. Simple nonconforming quadrilateral Stokes elements. *Numer. Meth. Partial Diff. Equations 8* (1992), 97–111.

[7] Roberts, J., and Thomas, J.-M. *Mixed and hybrid finite element methods*, vol. II of *Handbook of Numerical Analysis*. J.-L. Lions Ed., North-Holland, Amsterdam, 1989.

Nelly Barrau and Daniela Capatina
LMA & EPI Concha
Université de Pau & INRIA
IPRA BP 1155, Av. de l'Université
64013 PAU CEDEX, FRANCE
nelly.barrau@etud.univ-pau.fr and daniela.capatina@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **37**, 47–57 (2012)

ANALYSIS OF BIFURCATIONS APPEARING IN THE NONLINEAR HELICOPTER FLIGHT DYNAMICS

P-M. Basset, S. Kolb and C. Poutous

Abstract. The bifurcation theory is interested in the changes of the qualitative structure of dynamical system solutions when control parameters are varied. It is exploited here in order to analyse the highly nonlinear flight dynamics of a helicopter. This feature comes from the couplings between different constituting elements and physical variables and also from the overwhelming role of the main rotor whose dynamics is inherently quite complex. After describing the framework of the physical model i.e. the states, the control parameters and the dynamics, the appearing bifurcations are here analysed and characterised mathematically. Then the influence of the present nonlinearities over the global helicopter behaviour is assessed.

It is first shown that the formalism of a system of differential algebraic equations is here required so as to impose some algebraic constraints on some translational and rotational velocities, thus avoiding any inappropriate divergent movement. On the one hand, a bifurcation of equilibrium points associated to a real eigenvalue is linked to the vortex ring state phenomenon which occurs during steep descent flight. In this case, jumps and hysteresis reveal to be responsible for the dangerousness of such a situation. On the other hand, bifurcations of periodic orbits are observed and evaluated as triggering harmful pilot-aircraft couplings. Their type and characteristics are determined. To put in a nutshell, the nonlinear rotorcraft dynamics gives raise to interesting bifurcations whose description and characterisation need to be successfully performed in order to help avoiding dangerous configuration and recovering from these last ones.

Keywords: Bifurcation theory, dynamical systems, flight dynamics.

AMS classification: 34K18, 34K20.

Introduction

Helicopter flight dynamics is highly nonlinear because of its complex rotor dynamics and of the numerous physical couplings. In this paper, several types of bifurcations are studied and related to bifurcations of equilibrium points and of periodic orbits. The focus is stressed on the mathematical aspects of the analysis of a real rotorcraft behaviour.

First the mathematical framework must be defined and the type of mathematical equations involved must be made explicit. We can notice that the classical formulation of a system of ordinary differential equations does not fit well with this issue and that a system of differential algebraic equations must be employed. Secondly concrete bifurcations will be examined. On the one hand, a fold bifurcation of equilibrium points is diagnosed as underlying the vortex ring state phenomenon and gives raise to a hysteresis dynamics. On the other hand, bifurcations of periodic orbits trigger a jump in the oscillation amplitude and are responsible for harmful rotorcraft-pilot couplings. Finally it is shown how efficient the bifurcation theory can be for the analysis of nonlinear rotorcraft flight dynamics. This study presents the adaptations necessary in order to employ such a methodology in this particular case. It also gives concrete mathematical propositions and statements relative to helicopter flight dynamics.

§1. Mathematical modelling and numerical aspects for the helicopter flight dynamics analysis

Before being able to analyse such a dynamical system, it is first necessary to define the mathematical model i.e. to describe the dynamical system and the type of equations involved (and perhaps to complete also the modelling of the helicopter flight dynamics). Then numerical algorithms need to be developed and employed so as to calculate the characteristic loci of bifurcation theory.

1.1. Mathematical model

In order to define a dynamical system (1), we must specify the vector of state variables $X \in \mathbb{R}^n$, the vector of command variables (or control parameters) $U \in \mathbb{R}^k$ and the vector field corresponding to the state dynamics $F \in C^\infty\left(\mathbb{R}^n \times \mathbb{R}^k, \mathbb{R}^n\right)$ with $n, k \in \mathbb{N}^*$.

$$\dot{X} = F\,(X, U)\,. \tag{1}$$

First the state variables describing the rotorcraft flight dynamics are:

$$X = \left(U_{hel}, V_{hel}, W_{hel}, P_{hel}, Q_{hel}, R_{hel}, \phi, \theta, Vim_{MR}, Vim_{TR}\right). \tag{2}$$

They correspond on the one hand to the classical variables of flight dynamics, i.e. U_{hel}, V_{hel}, W_{hel}, P_{hel}, Q_{hel}, R_{hel}, ϕ and θ, which are the translational velocities, the rotational velocities and the Euler angles. On the other hand, (Vim_{MR}, Vim_{TR}) are the (mean) induced velocities of the main and tail rotors which are specific rotorcraft variables. Secondly the helicopter has four controls:

$$U = (DT0, DTC, DTS, DTA)\,. \tag{3}$$

The three first ones command the main rotor i.e. $DT0$ is the collective pitch, DTC the lateral cyclic pitch, DTS the longitudinal cyclic pitch, whereas the last one DTA is the collective pitch of the tail rotor. Thirdly the expression of dynamics function F results here from the fundamental principle of dynamics and from aerodynamics modelling works. It corresponds to

$$F\,(X, U) = \left(\dot{U}_{hel}, \dot{V}_{hel}, \dot{W}_{hel}, \dot{P}_{hel}, \dot{Q}_{hel}, \dot{R}_{hel}, \dot{\phi}, \dot{\theta}, \dot{V}im_{MR}, \dot{V}im_{TR}\right) \tag{4}$$

The physical model derives from Newton's laws of motion and is written in the body-fixed axes at the centre of gravity of the rotorcraft [8].

$$\begin{aligned}
\dot{U}_{hel} &= -\left(W_{hel} \cdot Q_{hel} - V_{hel} \cdot R_{hel}\right) + \frac{F_X}{M_{hel}} - g \sin\theta, \\
\dot{V}_{hel} &= -\left(U_{hel} \cdot R_{hel} - W_{hel} \cdot P_{hel}\right) + \frac{F_Y}{M_{hel}} + g \cos\theta \sin\phi, \\
\dot{W}_{hel} &= -\left(V_{hel} \cdot P_{hel} - U_{hel} \cdot Q_{hel}\right) + \frac{F_Z}{M_{hel}} + g \cos\theta \cos\phi, \\
I_{XX}\dot{P}_{hel} &= \left(I_{YY} - I_{ZZ}\right) Q_{hel} \cdot R_{hel} + I_{XZ}\left(\dot{R}_{hel} + P_{hel} \cdot Q_{hel}\right) + M_X, \\
I_{YY}\dot{Q}_{hel} &= \left(I_{ZZ} - I_{XX}\right) R_{hel} \cdot P_{hel} + I_{XZ}\left(R_{hel}^2 - P_{hel}^2\right) + M_Y, \\
I_{ZZ}\dot{R}_{hel} &= \left(I_{XX} - I_{YY}\right) P_{hel} \cdot Q_{hel} + I_{XZ}\left(\dot{P}_{hel} - Q_{hel} \cdot R_{hel}\right) + M_Z.
\end{aligned} \tag{5}$$

The forces (F_X, F_Y, F_Z) and the moments (M_X, M_Y, M_Z) contains the contributions of the main rotor, the tail rotor, the fuselage, the horizontal tailplane, the vertical fin. These external forces are taken into account in addition to the weight (mass M_{hel}). $I_{XX}, I_{YY}, I_{ZZ}, I_{XZ}$ are the fuselage moments of inertia along the body reference axes.

The corner stone of the computation procedures linked to dynamical system problems consists in a so-called continuation algorithm. Such a software was developed for example by P. Guicheteau at ONERA for his studies on nonlinear fixed-wing aircraft flight dynamics [6, 7]. The continuation algorithm consists basically in the repetition of four steps: seeking a point on the solution curve, evaluating the tangent direction (Jacobian matrix calculation), predicting a new point and correcting the predicted point such that the calculated point is effectively on the curve. The characteristic loci can and must always be expressed under the form of an implicit system of n equations and $(n + 1)$ variables (with $n \in \mathbb{N}^*$). As a consequence, there can only be one single control parameter.

In this study, the vortex ring state phenomenon will be examined. As a consequence, the focus is stressed on the dynamics along the vertical axis and the influence of a descent rate variation. The main rotor collective pitch $DT0$ which mainly governs V_Z and which determines the main rotor thrust is therefore selected as control parameter U. Unfortunately for a helicopter, all the physical variables are often coupled. When the collective pitch $DT0$ is reduced, the main rotor (torque) moment decreases also. But since the tail rotor still creates the same (anti-torque) moment as before, the helicopter begins to turn. To stabilise the yaw rate R_{hel} and to prevent the helicopter from turning, it is necessary to change the value of the tail rotor collective pitch DTA.

By imposing R_{hel} to zero by means of an additional algebraic constraint, the adapted trim value of the tail rotor collective pitch DTA is indirectly calculated. For equivalent reasons, the lateral velocity V_Y is forced to zero by determining the required lateral cyclic pitch angle DTC and the longitudinal cyclic pitch angle DTS is chosen such that the forward velocity V_X is equal to a fixed forward velocity V_{H0} (null here). Finally the movement can be imposed in a vertical plane by means of the following system of algebraic equations:

$$\begin{cases}
R_{hel}\left(X, DT0, DTC, DTS, \mathbf{DTA}\right) = 0, \\
V_X\left(X, DT0, DTC, \mathbf{DTS}, DTA\right) = 0, \\
V_Y\left(X, DT0, \mathbf{DTC}, DTS, DTA\right) = 0.
\end{cases} \tag{6}$$

Thus, as a partial conclusion of this modelling part, for a helicopter flight dynamics problem, it can be stated that the description must be made by means of a system of **differential algebraic equations (DAE)**. The classical formulation under the form of a system of (autonomous) differential equations is not convenient.

1.2. Local bifurcations

The current study deals with local bifurcations of vector fields whose analysis is accomplished by examining the vector field in the neighbourhood of the (degenerate) equilibrium points or periodic orbits. The associated theory makes the assumption that considering the linearised system or truncated Taylor series of the vector fields allows to draw directly a conclusion for the nonlinear problem and the global (asymptotic) behaviour of its solutions. The methodology relies partly on the theorem 1.

Theorem 1 (Hartman-Grobman). *If $D_X F\left(\bar{X}\right)$ has no zero or purely imaginary eigenvalues then there is a homeomorphism h defined on some neighbourhood $\mathcal{U}$ of $\bar{X} \in \mathbb{R}^n$ locally taking orbits of the nonlinear flow to those of the linear flow $e^{tD_X F(\bar{X})}$. The homeomorphism preserves the sense of orbits and can also be chosen to preserve parametrisation by time.*

Definition 1. If no eigenvalues of the Jacobian matrix $D_X F\left(\bar{X}\right)$ has a zero real part, then $\bar{X}$ is called a **hyperbolic** fixed point.

In such a situation, linearisation is sufficient to determine the asymptotic behaviour of solutions. On the contrary, when one of the eigenvalues has got a zero real part, the fixed point is said to be **nonhyperbolic**. Then it may be necessary to calculate higher order coefficients in the Taylor series and to evaluate the dynamics on the center manifold [5] so as to be able to conclude about the asymptotic behaviour of the overall system.

After describing the mathematical model and presenting elements concerning bifurcation theory, a concrete case is then examined with the help of this methodology.

§2. Fold bifurcation of equilibrium points, hysteresis (Vortex Ring State)

For a phenomenon such as the vortex ring state, the nonlinear behaviour comes from the nonlinear evolution of the induced velocity of the main rotor during descent flight. Indeed for a certain range of descent velocity (and forward speed), the rotor enters in its own wake and a doughnut-shaped ring appears around the rotor disk. The induced velocity of the main rotor increases strongly but its thrust is falling off and is not sufficient any more to stabilise the rotorcraft. The heave dynamics is affected, it corresponds approximatively to a change of sign of the derivative $Z_w = \partial \dot{W}_{hel} / \partial W_{hel}$ of the vertical dynamics and thus a behavioural change of the solution of the equation

$$\dot{W}_{hel} - Z_w W_{hel} = 0, \tag{7}$$

with the approximate analytic expression [8]

$$Z_W = -\frac{2 C_{Z_\alpha} \mathcal{A}_{blade}\, \rho \left(\Omega R\right) V_i / \left(\Omega R\right)}{\left(16 V_i / \left(\Omega R\right) + C_{Z_\alpha} N_{blade}\, c / \left(\pi R\right)\right) M_{hel}} \tag{8}$$

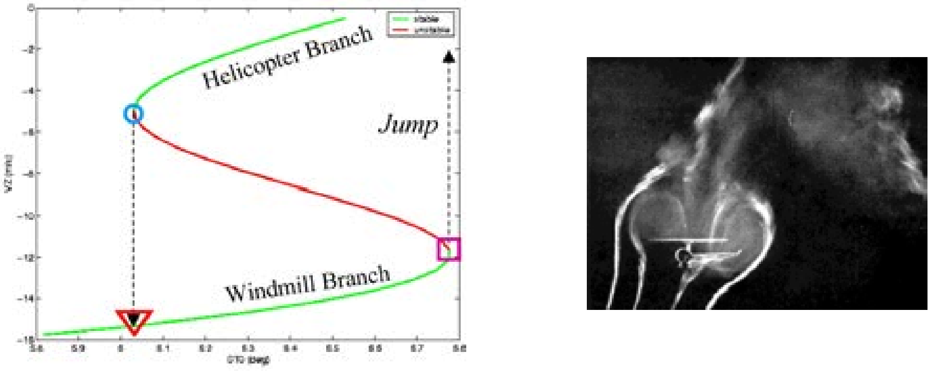


Figure 1: Bifurcation diagram associated to the vortex ring state [3] and presenting the descent rate V_Z in function of the collective pitch $DT0$

in hover or vertical flight and

$$Z_W = -\frac{C_{Z_\alpha} \mathcal{A}_{blade}\, \rho\, (\Omega R)}{2 M_{hel}} \, \frac{4}{(8 V_X / (\Omega R) + C_{Z_\alpha} N_{blade}\, c / (\pi R))} \tag{9}$$

in forward flight where C_{Z_α} is the blade lift curve slope, $\mathcal{A}_{blade}$ the blade surface, M_{hel} the total mass, N_{blade} the blade number, c the blade chord, Ω the nominal main rotor speed, R the main rotor radius and ρ the air density.

The bifurcation theory is interested in the determination of the bifurcation diagram (locus of equilibrium points), the locus of the bifurcation points and the equilibria surface. They are calculated in the following sections.

2.1. Locus of equilibrium points (bifurcation diagram)

The continuation algorithm allows to compute the bifurcation diagram of the system made of equations (1), (2), (3) and (6). Its result is shown in Figure 1.

According to [5], the “generic” saddle-node bifurcation looks qualitatively like the family of equations $\dot{x} = u - x^2$ in the zero eigenvector direction (and with hyperbolic behaviour in the complementary directions).

In Figure 1, the equilibrium curve contains two stable branches (green) and in the middle of them an unstable branch (red). The two bifurcations are linked to a zero real eigenvalue and are turning points [7]. For the range of control parameters $DT0$ between the two critical values, there are three equilibrium points whereas outside this region there is only one single equilibrium point. Such a bifurcation diagram is the typical one of a **hysteresis**. For flight dynamics engineers, the flight regimes at low descent rates is called “helicopter branch” and the one at high descent rates is named “windmill branch”.

Concretely when the system is in a steady configuration near the bifurcation point, a little variation of the control parameter induces a situation where the system does not succeed any more in stabilising itself. As a consequence, a jump occurs on the other branch of equilibria

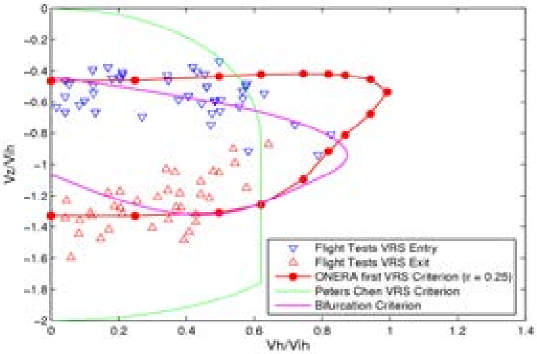


Figure 2: Boundaries of the vortex ring state region

(and isn't reversible with little opposite variations). From the viewpoint of flight dynamics, this jump from the helicopter branch to the windmill branch shows the loss of stability and the sudden increase in descent rate of the helicopter which appear when entering in vortex ring state.

After having determined the locus of equilibrium points and the type of dynamics associated to this bifurcation diagram, it is interesting to compute the locus of bifurcations points. It provides the analysis with new powerful information.

2.2. Locus of bifurcation points

The locus of bifurcation points is composed of the equilibria for which a behavioural change occurs (such as stability loss) that is to say here equilibria such that one real eigenvalue of the Jacobian matrix is equal to zero. The associated mathematical criterion is $\det\left(D_X\dot{X}\right) = 0$ which can also be written with the notation employed in the equation (1):

$$\det\left(D_X F\left(X, U\right)\right) = 0. \tag{10}$$

The continuation algorithm permits to compute the locus of bifurcation points by solving the following system of equations whose control parameters are V_{H0} and V_{Z0}:

$$\begin{cases} \dot{X} = F\left(X, U\right), \\ \det\left(D_X F\left(X, U\right)\right) = 0, \\ R_{hel}\left(X, DT0, DTC, DTS, \mathbf{DTA}\right) = 0, \\ V_X\left(X, DT0, DTC, \mathbf{DTS}, DTA\right) = V_{H0}, \\ V_Y\left(X, DT0, \mathbf{DTC}, DTS, DTA\right) = 0, \\ V_Z\left(X, \mathbf{DT0}, DTC, DTS, DTA\right) = V_{Z0}. \end{cases} \tag{11}$$

In Figure 2, the locus of the bifurcation points (labelled "Bifurcation Criterion" and purple-coloured) is compared with data resulting from flight tests organised by ONERA at

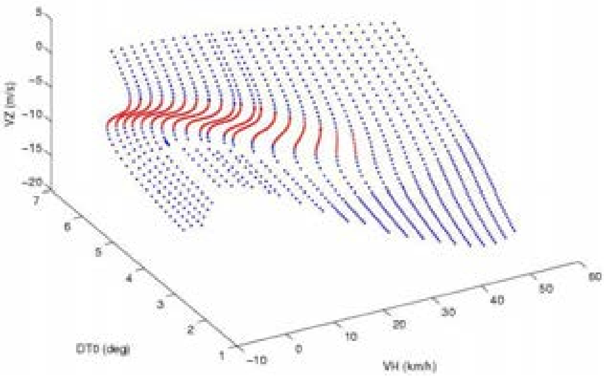


Figure 3: Surface of equilibria near the vortex ring state region

the French flight test centre of Istres and other criteria delimiting the VRS zone. The sudden drops are represented with blue triangles and the stabilisation points with red triangles.

As observed in the diagram presenting the forward velocity V_H and the descent rate V_Z normalised by the main rotor induced velocity in hover V_{ih} (cf. again Figure 2), the locus of bifurcation points fits well with the flight tests and predicts well the zone of instabilities.

Besides the gap for the lower frontier can be explained by the fact that the flight tests diagnose the conditions for which the aircraft stabilises after the drop whereas the bifurcation point represents the conditions for which the jump from the windmill branch to the helicopter branch occurs. The first point has got a bigger descent rate than the second one.

Moreover another relevant information can be obtained by scrutinising the surface of equilibrium points.

2.3. Surface of equilibria

Practically the surface of equilibria is actually determined by calculating the loci of equilibrium points for several longitudinal velocity V_H. The algebraic equations are (12) and the control parameters are $DT0$ and V_{H0}:

$$\begin{cases} R_{hel}\left(X, DT0, DTC, DTS, \mathbf{DTA}\right) = 0, \\ V_X\left(X, DT0, DTC, \mathbf{DTS}, DTA\right) = V_{H0}, \\ V_Y\left(X, DT0, \mathbf{DTC}, DTS, DTA\right) = 0, \end{cases} \tag{12}$$

The surface of equilibria in the neighbourhood of the vortex ring state is exposed in Figure 3. From the point of view of dynamical system theory, such a surface is called a **cusp** (cf. [5, page 355] or [9, pages 344-346]). There is a turning fold [7] i.e. a zone with three equilibrium points and another one with only one single equilibrium point. By considering the surface where the stable blue points are distinguished from the unstable red ones and by examining the configuration, an escape strategy can be deduced. When the aircraft jumps from the helicopter branch to the (windmill) branch with high descent rate, it is in a zone with three equilibriums. By increasing its forward velocity, the unstable zone reduces itself and disappears at the end. The helicopter is then in a zone with only one single stable equilibrium which means in a safe situation.

Conclusion about the analysis of a real bifurcation of equilibria in the case of the vortex ring state phenomenon

The mathematical formulation as a system of differential algebraic equations (DAE) seems to be necessary for the description and analysis of rotorcraft flight dynamics. Indeed many variables are coupled and some algebraic constraints must be added in order to avoid senseless configurations. As far as nonlinear analysis is concerned, on the one hand, the bifurcation theory reveals an underlying hysteresis phenomenon triggered by saddle-node bifurcations of equilibrium points. On the other hand, the locus of bifurcations points proves to be a relevant criterion in order to delimit the zone of instabilities linked to the vortex ring state [2].

This first part was devoted to the thorough analysis of a bifurcation of equilibria associated to a real eigenvalue and corresponding to the phenomenon of vortex ring state. After describing the mathematical model, the bifurcation diagram, the surface of equilibria and the locus of the bifurcation points were determined and interpreted from the both points of view of a mathematician and a flight dynamics engineer. In the next part, a bifurcation of periodic orbits will be studied. The mathematical model comes from the representation of a rotorcraft command channel with the use of the describing function theory.

§3. Bifurcation of limit cycles (Pilot-Induced Oscillations)

In order to perform the analysis of the rotorcraft command channel, the describing function method is employed and some elements about its mathematical justification is first introduced. Then the equations associated to the flight control system are made explicit. Finally the solution is computed thanks to the continuation algorithm and the results are interpreted with the bifurcation theory formalism.

3.1. Methodology

In order to exploit the describing function method [4], two conditions must hold. The first one states that there must be a clearly identifiable nonlinear element which can be isolated from the linear part whereas the second one stipulates that the linear part must behave like a low-pass filter. For the closed-loop system, the determination of the existence of possible periodic orbits and of their first-harmonic properties requires to solve the harmonic balance equation:

$$1 + L(j\omega) \cdot N(A, \omega) = 0, \tag{13}$$

where ω is the pulsation of the possible limit cycle, A the amplitude of its first harmonic, $N(A, \omega)$ the describing function of the nonlinear element (i.e. the rate-limited actuator) and $L(j\omega)$ the linear part including the bare airframe, the pilot and the linear actuators.

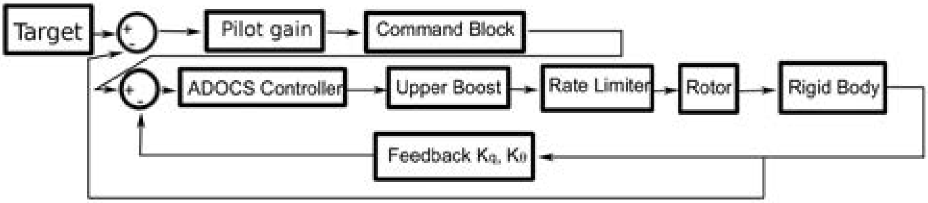


Figure 4: Closed-loop ADOCS command channel

Some elements relative to the mathematical foundation of the describing function method were explained in the previous section. The handled flight control system is exposed before beginning its concrete examination.

3.2. Command channel

The rotorcraft command channel presented in Figure 4 is the longitudinal one of the ADOCS helicopter as described by the NASA technical memorandum [10]. It contains the command block which filters the possible too aggressive pilot inputs, the blocks modelling the dynamics of the rotor and of the fuselage and the feedback loop block. The displacement velocity of the swashplates which command directly the motion of the main rotor blades is limited to 10 inches/s. This last one is here responsible for the observed nonlinear behaviour.

The longitudinal flight control system is analysed by means of the describing function method. According to (13), the equation (14) requires to be solved so as to diagnose the possible existence of a periodic orbit and to estimate the amplitude A and phase delay ϕ of its first harmonic for various values of input oscillation amplitude θ_c and for a pilot gain $K_p = 1$ (fixed nervousness here):

$$\begin{aligned}\Big(1 + \mathit{Rotor} \cdot \mathit{RigidBody} \cdot N(A,\omega) \cdot \mathit{Actuator} \cdot \big(K_p \cdot \mathit{CommandBlock} + \mathit{Feedback}\big)\Big)& \\ \times A \exp(j\phi) = \mathit{Actuator} \cdot \mathit{CommandBlock} \cdot \theta_c.& \quad (14)\end{aligned}$$

The characterisation of a saddle-node bifurcation of periodic orbits can be found in [9] and indeed it can be observed that Figure 5 is typical of a **saddle-node bifurcation of periodic orbits** [1, 9]. When the reference amplitude is increased from 0.33 rad to 0.34 rad, the amplitude of the entry state of the rate limiter jumps from 6 to 10.

Concretely the sudden increase may surprise, disturb greatly the pilot which does not succeed any more in controlling the aircraft, what leads to a risky situation.

Conclusion

During this study, two different phenomena coming from the field of rotorcraft flight dynamics were dealt with. Their underlying dynamics is governed by different types of bifurcations.

A fold bifurcation of equilibrium points of real eigenvalue proves to be responsible for a sudden jump of one branch of equilibria (helicopter branch) to another one (windmill branch)

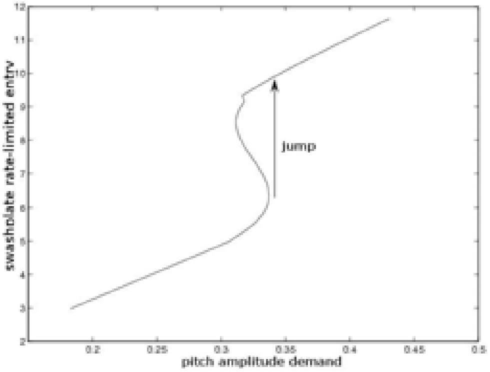


Figure 5: Jump of the oscillation amplitude

for a little variation of the (collective pitch) control. According to the bifurcation diagram, the underlying dynamics is a hysteresis. This last one explains mathematically the appearance of the vortex ring state phenomenon.

As far as pilot induced oscillations are concerned, a saddle-node bifurcation of periodic orbits is here observed. Amongst others, they imply jumps in amplitude of the periodic orbits and trigger some flying qualities cliffs.

Several important results were presented in this research paper. The detection of real bifurcations allows to delimit successfully the region of vortex ring state. The determination of the existence and of the properties of a bifurcation of limit cycles shows that the command channel of the ADOCS helicopter demonstrator which was adapted for this study is likely to have some flying qualities cliffs.

As a conclusion, the bifurcation theory reveals to be a useful tool for the nonlinear analysis of rotorcraft flight dynamics. It provides criteria helping delimiting dangerous regions of flight or detecting changes of flying qualities.

References

[1] Alcala, I., Gordillo, F., and Aracil, J. Phase compensation design for prevention of pio due to actuator rate saturation. In *American Control Conference* (Boston, Massachusetts, 2004).

[2] Basset, P.-M., Chen, C., Prasad, J. V. R., and Kolb, S. Prediction of vortex ring state boundary of a helicopter in descending flight by simulation. *Journal of the American Helicopter Society 23*, 2 (2008), 139–151.

[3] Drees, J. M., and Hendal, W. P. The field of flow through a helicopter rotor obtained from wind tunnel smoke tests. Tech. Rep. A.1205, National Luchtvaart Laboratorum, 1953.

[4] Gelb, A., and Vander Velde, W. E. *Multiple-Input Describing Functions and Nonlinear System Design*. McGraw-Hill, 1968.

[5] Guckenheimer, J., and Holmes, P. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, vol. 40 of *Classics in Applied Mathematics*. Springer-Verlag, Philadelphia, 2002. Firstly published by North-Holland, Amsterdam, 1978.

[6] Guicheteau, P. Bifurcation theory: A tool for nonlinear dynamics. *Philosophical Transactions: Mathematical and Engineering Sciences, Nonlinear Flight Dynamics of High Performance Aircraft 356* (1998), 2182–2202.

[7] Kubicek, M., and Marek, M. *Computational Methods in Bifurcation Theory and Dissipative Structures*. Springer-Verlag, 1983.

[8] Padfield, G. D. *Helicopter Flight Dynamics*. Blackwell Science Ltd, 1996.

[9] Perko, L. *Differential Equations and Dynamical Systems*, vol. 7 of *Texts in Applied Mathematics*. Springer-Verlag, Philadelphia, 2002.

[10] Tischler, M. B. Digital control of highly augmented combat rotorcraft. Technical Memorandum 88346, NASA, May 1987.

Pierre-Marie Basset
DCSD
ONERA
BA 701, F-13661 SALON AIR
pierre-marie.basset@onera.fr

Sébastien Kolb and Cécile Poutous
Flight dynamics team
French Air Force Research Centre
BA 701, F-13661 SALON AIR
sebastien.kolb@inet.air.defense.gouv.fr and cecile.poutous@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **37**, 59–67 (2012)

CONCEPTS OF THE FINITE ELEMENT LIBRARY *Concha*

Roland Becker and David Trujillo

Abstract. We describe the concepts underlying the finite element library *Concha*. The library intends to provide tools for implementing general finite element methods based on continuous, non-conforming, and discontinuous finite element spaces defined on different types of meshes in two and three space dimensions. It also provides different stabilization methods, a posteriori error estimators, and local mesh refinement. Concrete examples are shown for incompressible and compressible flows described by the Navier-Stokes and Euler equations, respectively.

Keywords: Finite element methods, adaptivity, C++.

AMS classification: 65N30, 65N50, 65N55, 65M30, 65M50, 65M55.

§1. Introduction

The finite element library *Concha* is developed by the team with the same name, Concha `https://sites.google.com/site/conchapau`, which is an “équipe” supported by the University of Pau and INRIA Bordeaux-Sud Ouest since 2008. The objective of this interdisciplinary team is the development and analysis of algorithms and efficient software for the simulation of complex flow problem. We are specially interested in modern discretization methods (adaptivity, high-order and stabilized methods) and in goal-oriented simulation tools (prediction of physical quantities, numerical sensitivities, and optimization). *Concha* is the common computing platform used to implement our algorithms and to perform numerical experiments. For the moment, it has been used to perform the following tasks:

- Numerical simulation of viscoelastic flows,
- Study of adaptive mesh-refinement algorithms,
- Numerical simulation of incompressible flows with heat transfer,
- Study of discontinuous finite element methods for the Euler equations,
- Numerical experiments of stabilized finite element methods for convection-dominated problems.

This article is organized as follows: in Section 2 we describe the general purposes of the library concerning physical models and finite element techniques. Section 3 is devoted to adaptive mesh refinement algorithms, which are then used in Section 4 in order to illustrate a typical work flow. The concepts of the C++-part of the library are described in Section 5. Sections 6 and 7 present typical examples for the Navier-Stokes and Euler equations. Finally, further development is outlined in Section 8.

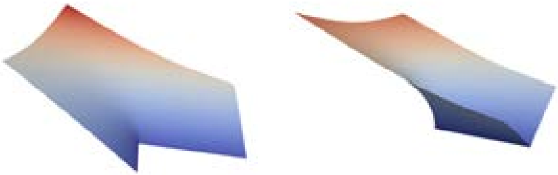


Figure 1: Geometrical singularities: solution of the Poisson equation for a domain with reentrant corner (left) and slit domain (right).

§2. Purpose of the library

The purpose of the library is to provide tools for the development of finite element methods in fluid mechanics. It is intended to become an academic tool able to tackle applications related to industrial problems from different areas. Therefore a high level of abstraction is required, at the same time with respect to physical models and numerical algorithms. Ideally, the addition of new equations or the change of constitutive laws should be simple. At the same time, switching to another discretization or solution method should be possible with a minimum amount of programming work. The benefits of object oriented programming with respect to these objectives are clear nowadays, and we have chosen to develop the core of the library in the C++-language. This part of the library is organized in different layers, varying in generality. In addition, a large effort was made to orthogonalize as much as possible the different computational tasks. We therefore use different executables, which communicate by files and are piloted by scripts written in the python-language. This also allows a simple use of external tools for mesh generation and refinement, as well as solution algorithms.

We consider unstructured meshes containing triangles, quadrilaterals, tetrahedra, and hexaedra, allowed to contain hanging nodes. The library contains tools for the construction of the following finite element spaces:

- Continuous finite element spaces P^k, Q^k,
- Non-conforming finite element spaces (Crouzeix-Raviart, Rannacher-Turek),
- Vector-spaces (Raviart-Thomas elements),
- Completely discontinuous finite elements.

In summary, the guidelines of our library are to a) reuse code as much as possible, b) orthogonalize different computational parts, and c) guarantee flexibility with respect to methods and models. The chosen technical tools for this are a) use of different executables, b) use of inheritance and polymorphism, c) project-oriented design.

§3. Adaptive finite element methods

Local mesh refinement has become an important tool in finite element simulations, since it allows to recover optimal convergence rates in many situations, where a loss of regularity

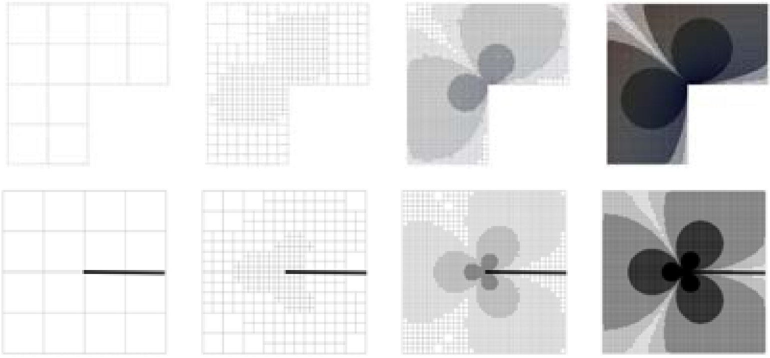


Figure 2: Geometrical singularities: sequences of meshes generated for a domain with reentrant corner (above) and slit domain (below).

leads to slow convergence under uniform mesh refinement. Sources for such a loss of regularity are corners in the geometry of the computational domain, see Figure 1, boundary and internal layers, see Figure 4.

The domain singularities lead to singular higher-order derivatives of the solutions, which do not allow for optimal order interpolation error estimates; but the situation is even worth: the error is transported by the differential operator and slows down convergence away from the corner, the well-known ’pollution effect’.

In order to recover the optimal order of convergence, the mesh has to be refined at the corner, and this local refinement has to be done with a certain concentration of mesh points depending on the desired accuracy.

Such sequences of meshes can be constructed in automatic way by an adaptive finite element method, see Figure 2.

An adaptive finite element method is based on a local refinement algorithm and an a posteriori error estimator. The algorithm is initialized by construction of the initial mesh and the choice of certain paramaters. It consists of an iterative loop, which performs, at each step, the following tasks:

$$\textbf{Solve} \longrightarrow \textbf{Estimate} \longrightarrow\textbf{Mark} \longrightarrow\textbf{Refine} \longrightarrow \cdots$$

The algorithm, which is completed by a stopping criterion, generates a sequence of meshes $(h_k)_{k\geq 1}$ and discrete solutions $(u_k)_{k\geq 1}$. Each mesh h_k is a member of the family of admissible meshes $\mathcal{H}$, defined through the initial mesh h_1 and the refinement algorithm; each solution u_k, generated by **Solve**, lies in the finite element space V_k depending on mesh h_k and the chosen finite element method. In addition, we also have a sequence of estimators $(\eta_k)_{k\geq 1}$, generated by **Estimate**, errors $(e_k)_{k\geq 1}$, and sets of marked cells $(\mathcal{M}_k)_{k\geq 1}$. The set of marked cells $\mathcal{M}_k$ is a subset of the cells of h_k, denoted by $\mathcal{K}_k$. It is generated by **Mark** using information from the estimator η_k and serves as an input to the local mesh refinement algorithm **Refine**. In any

	input	output
Initialization	domain geometry, external mesh	ConchaMesh
Solve	ConchaMesh	computed solution
Estimate	computed solution	cell-wise error indicators
Mark	cell-wise error indicators	set of marked cells
Refine	ConchaMesh and cell-wise error indicators	refined ConchaMesh

Table 1: Executables and their in- and output.

event, the local refinement algorithm is supposed to verify the following complexity estimate:

$$\#\mathcal{K}_n \leq \#\mathcal{K}_1 + C \sum_{k=1}^{n-1} \#\mathcal{M}_k, \tag{1}$$

where C is a mesh-independent constant. The estimate (1) is necessary for any complexity estimate.

For simple model problems, the number of cells generated by the adaptive algorithm can be related to the achieved accuracy. More precisely, let ε_k be the norm of the error at iteration k and $N_k = \#\mathcal{K}_k$ be the number of cells. Then recent results prove that $N_k \approx \varepsilon_k^{-1/s}$ where $s > 0$ is the speed of convergence. For this, a certain regularity on the continuous solution u of the problem has to be made. This assumption basically states that for given $\varepsilon > 0$ there exists a mesh $h \in \mathcal{H}$ such that $N \approx \varepsilon^{-1/s}$. For example, for two-dimensional elliptic problems and piece-wise linear approximation, the assumption holds with $s = 1/2$.

The optimality of an adaptive finite element method thus states that, if ever u can be approximated with the help of $\mathcal{H}$ at speed s, the algorithm automatically selects a sequence of meshes, that leads to convergence with speed s. Such results have first been obtained for continuous finite element approximations of the Poisson problem by [8, 16]. They have been generalized to mixed and non-conforming finite elements [5, 7] and to the Stokes equations [6].

§4. Work flow: pilotage of executables

As an example, we consider the adaptive algorithm of the preceding section. In our implementation, each task corresponds to an executable, which takes certain parameters and data as an input and produces other data. The whole loop is then written in python. This allows for a simple treatment of the parameters and eases modification of the scripts, avoiding compilation. We end up with the ingredients detailed in Table 1.

We remark that the Mark-executable is independent of the precise form of the estimator or problem at hand. A typical strategy for marking is the bulk criterion: Find $\mathcal{M}$ with minimal cardinality such that $\sum_{K\in\mathcal{M}} \eta_K^2 \geq \theta \sum_{K\in\mathcal{K}} \eta_K^2$ for a given parameter $0 < \theta < 1$. Other marking strategies are available, especially in the case that several estimators are computed, for example an additional data approximation term.

The Solve- and Estimate-executables must know about the finite element method and need to connect the solution data to these spaces. They are therefore based on common parts of the C++-library. These executables use parameter-files which are adapted appropriately by the python script.

The Refine-executable is written in the mesh part of our library. It uses a standard pointer-based data-structure to represent the mesh in tree form. There are also tools to convert external mesh-types into our data structures. This procedure allows us the use of external mesh tools without changing the solvers.

§5. Structure of the C++-code

In order to avoid code duplication and allow for abstraction, the C++-part of the library heavily depends on inheritance and polymorphism. The library basically presents classes which can be adapted by the developer. The following classes play a fundamental role in our design: *Loop*, *Solver*, *Model*, *Variable*, *Integrator*, and *Application*. The role of *Loop* is to define abstract algorithms such as iterative solution of a nonlinear system of equations, *StaticLoop*, time-stepping for dynamic problems, *DynamicLoop*, or computation of postprocessing, *PostProcessLoop*. The essential memory, that is vectors and matrices, are stored in the *Solver*. The last class also provides implementation of Newton-type algorithms and time discretization. The physical model and its finite element representation are described in the class *Model*. Its task is to define the set of variables representing the physical unknowns and the different terms of the equations in variational form. A *Variable* gives the name and size of a physical quantity together with its finite element space and some other useful information, as for example its output format. The core of the variational formulation is described by the *Integrators*. An *Integrator* defines a set of output and input variables and provides the implementation of the integrals used in the computation of residuals and Jacobians. All output variables that are not unknowns are considered as *PostprocessVariables*, which are not involved in the solution of the system of equations, but produce other values, such as error estimators and physical functionals. Finally, the class *Application* describes the variable part of the problem: boundary conditions, different forms of right-hand sides and possibly fixes some physical and numerical paramaters.

§6. Example: Incompressible viscous flows

Here we consider the stationary Navier-Stokes equations in a bounded two-dimensional domain Ω for the set of physical unknowns $u = (v, p)$ consisting of velocity vector v and pressure p:

$$\begin{cases} \rho v \cdot \nabla v - \mu \Delta v + \nabla p = 0 & \text{in } \Omega, \\ \operatorname{div} v = 0 & \text{in } \Omega, \\ v = v_D & \text{in } \Gamma_D, \\ \mu \partial_n v - pn = -p_D n & \text{in } \Gamma_N, \end{cases}$$

where ρ and μ are positive constants and the boundary $\partial\Omega$ is cut in a Dirichlet part Γ_D and a Neumann part Γ_N. In addition v_D and p_N are given data representing for example in and outflow data.

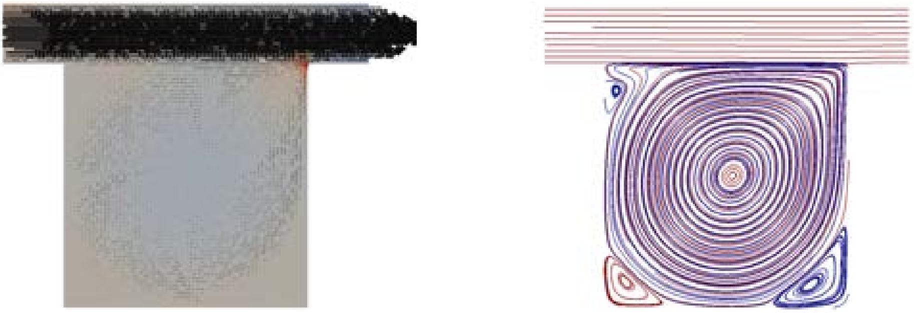


Figure 3: Driven cavity: domain and stream lines.

As an example we consider a driven cavity problem in the domain $\Omega =]-1,1[\times]-1,1[\,\cup\,]-1.5,1.5[\times]1,1.5[$. At the left inflow a parabolic inflow with maximal velocity 1.0 is given, the outflow is described by the Neumann-type condition. The viscosity is $\mu = 0.000025$. We use a stabilized Taylor-Hood scheme. The domain and streamlines of the velocity field are shown in Figure 3.

§7. Example: Compressible inviscid flows

We denote by u the vector of physical variables, i.e. $u = (\rho, \rho v, \rho E)$ where ρ, v, and E are the density, the velocity field, and the total energy. The pressure is related to ρ and E by the ideal gas law. We write the system of equations as

$$u_t + \operatorname{div} f(u) = 0,$$

completed by a set of appropriate initial and boundary conditions. The flux function f is given by $f(u) := (f_1(u), f_2(u))$, where

$$\begin{aligned} f_1(u) &= (u_1, u_1v_1 + p, u_1v_2, (u_3 + p)v_1),\\ f_2(u) &= (u_2, u_2v_1, u_2v_2 + p, (u_3 + p)v_2). \end{aligned}$$

The discontinuous finite element method is based on a piecewise polynomial approximation over a mesh h (either triangular or quadrilateral). The set of cells of h is denoted by $\mathcal{K}_h$ and the set of interior sides by $\mathcal{S}_h$; the set of boundary sides is denoted by $\mathcal{S}_h^{\partial}$. In addition, we denote by T_K the transformation of a reference cell to the physical cell K (it is linear in the case of triangles and bilinear in the case of quadrilaterals) and by R^k the set of polynomials of either total or maximal degree k (P^k for triangles and Q^k for quadrilaterals). The discontinuous finite element space is then defined as

$$V_h^k := \{v_h \in L^2(\Omega) : v_h|_K \circ T_K \in (R^k(K)^4 \quad \forall K \in \mathcal{K}_h\}.$$

The discrete variational formulation now reads: Find $u_h \in V_h$ such that for all $v_h \in V_h$:

$$a_h(u_h)(v_h) = l_h(v_h) \quad \forall v_h \in V_h,$$

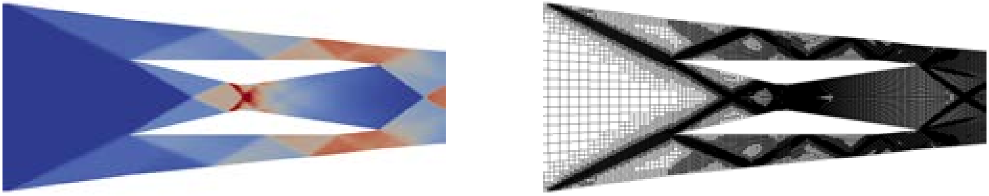


Figure 4: The scramjet test case: density and locally refined mesh.

where l_h is linear functional representing the inflow data and the form a_h is composed of three terms corresponding tho the mesh cells, interior sides, and boundary sides:

$$a_h(u_h)(v_h) = a_h^{\mathcal{K}_h}(u_h)(v_h) + a_h^{\mathcal{S}_h}(u_h)(v_h) + a_h^{\mathcal{S}_h^\partial}(u_h)(v_h).$$

The three terms are given by

$$\begin{aligned} a_h^{\mathcal{K}_h}(u_h)(v_h) &:= -\sum_{K\in\mathcal{K}_h}\int_K f(u_h) : \nabla v_h\, dx, \\ a_h^{\mathcal{S}_h}(u_h)(v_h) &:= \int_{\mathcal{S}_h} F(u_h, n_S)\cdot [v_h]\, ds, \\ a_h^{\mathcal{S}_h^\partial}(u_h)(v_h) &:= \int_{\mathcal{S}_h^\partial} \Phi(u, u_d, n_S)\cdot v\, ds. \end{aligned}$$

Here Φ and F are numerical fluxes representing the boundary conditions and the inter-element continuity. As numerical flux, we use here the Vijayasundaram flux. As a test case, we consider the scramjet configuration [11]. A steady supersonic flow enters the computational domain at Mach number 3 and hits two sharp-cornered internal obstacles. This configuration leads to multiple shock wave reflections. A typical solution and a locally refined mesh are shown in Figure 4.

§8. Further developement

The further development of the library is oriented towards the following topics:

Multigrid solvers

We develop multigrid solvers for the resolution of the linear systems arising in the Newton algorithm to solve the nonlinear problems. To this end, a hierarchy of meshes is created by maximal derefinement of the finest locally refined mesh, as described in [1].

Nitsche Extended Finite Element Method

NXFEM is a variational formulation of XFEM based on Nitsche's method [3, 13]. It allows for accurate discretization of interface problems on non-matching meshes, and can be used for fictitious domain approaches [3].

New stabilized FEM

As an alternative to SUPG, new stabilization techniques have been developed recently [2, 10, 4]. We are interested in a comparative study with respect to other methods such as the so-called discontinuous Galerkin methods.

Parallelization

The parallelization of the library is an ongoing projected, supported by INRIA in form of the ADT (action de développement technologique) Ampli.

Sharp error estimators

Based on the reconstruction of locally conservative fluxes, it is possible to derive sharp error estimators [17, 15, 12, 9].

Goal-oriented error estimation and sensitivity computations

Goal-oriented error estimation is an important tool for numerical simulation, since it allows to directly control the error in the computation of physical quantities. We are working on an automatization of the solution of the additional problems, which are required in this approach. Our techniques will also allow to compute sensitivities with respect to certain physical or modeling parameters.

Robust finite element discretizations for all-Mach-number flows

The efficient solution of flow at arbitrary Mach numbers remains a challenging problem. Although stable discretizations based on physical or entropy variables are known [14], some important questions such as the efficient solution are still unclear.

Robust finite element discretizations for high Reynolds numbers

High Reynolds number flows require in practice the use of some kind of turbulence models. We are interested in the development of variational multi-scale methods related to stabilization.

Robust finite element discretizations for high Weissenberg numbers

Another ongoing project is the development of robust methods for high Weissenberg numbers in viscoelastic flows.

Acknowledgements

We thank our colleagues Daniela Capatina, Didier Graebling, Robert Luce, and Eric Schall, who are all involved in the development of *Concha*.

References

[1] Becker, R., and Braack, M. Multigrid techniques for finite elements on locally refined meshes. *Numer. Linear Algebra Appl. 7*, 6 (2000), 363–379. Numerical linear algebra methods for computational fluid flow problems.

[2] Becker, R., and Braack, M. A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo 38*, 4 (2001), 173–199.

[3] BECKER, R., BURMAN, E., AND HANSBO, P. A Nitsche extended finite element method for incompressible elasticity with discontinuous modulus of elasticity. *Comput. Methods Appl. Mech. Engrg. 198*, 41-44 (2009), 3352–3360.

[4] BECKER, R., BURMAN, E., AND HANSBO, P. A finite element time relaxation method. *C. R. Acad. Sci. Paris* (2011).

[5] BECKER, R., AND MAO, S. An optimally convergent adaptive mixed finite element method. *Numer. Math. 111*, 1 (2008), 35–54.

[6] BECKER, R., AND MAO, S. Quasi-optimality of adaptive non-conforming finite element methods for the Stokes equations. *SIAM J Numer. Anal.* (2011).

[7] BECKER, R., MAO, S., AND SHI, Z.-C. A convergent nonconforming adaptive finite element method with optimal complexity. *SIAM J Numer. Anal. 47*, 6 (2010), 4639–4659.

[8] BINEV, P., DAHMEN, W., AND DEVORE, R. Adaptive finite element methods with convergence rates. *Numer. Math. 97*, 2 (2004), 219–268.

[9] BRAESS, D., AND SCHÖBERL, J. Equilibrated residual error estimator for edge elements. *Math. Comp. 77*, 262 (2008), 651–672.

[10] BURMAN, E., AND ERN, A. A continuous finite element method with face penalty to approximate Friedrichs' systems. *M2AN Math. Model. Numer. Anal. 41*, 1 (2007), 55–76.

[11] DIAZ, M., HECHT, F., AND MOHAMMADI, B. New progress in anisotropic grid adaptation for inviscid and viscous flows simulations. In *Proceedings of the 4th Annual International Meshing Roundtable* (1995), Sandia National Laboratories.

[12] ERN, A., NICAISE, S., AND VOHRALÍK, M. An accurate $H(\mathrm{div})$ flux reconstruction for discontinuous Galerkin approximations of elliptic problems. *C. R. Math. Acad. Sci. Paris 345*, 12 (2007), 709–712.

[13] HANSBO, A., AND HANSBO, P. An unfitted finite element method, based on nitsche's method, for elliptic interface problems. *Comput. Methods Appl. Mech. Eng. 191*, 47-48 (2002), 5537–5552.

[14] HAUKE, G., AND HUGHES, T. J. R. A comparative study of different sets of variables for solving compressible and incompressible flows. *Comput. Methods Appl. Mech. Engrg. 153*, 1-2 (1998), 1–44.

[15] LUCE, R., AND WOHLMUTH, B. A local a posteriori error estimator based on equilibrated fluxes. *SIAM J. Numer. Anal. 42*, 4 (2004), 1394–1414.

[16] STEVENSON, R. Optimality of a standard adaptive finite element method. *Found. Comput. Math. 7*, 2 (2007), 245–269.

[17] TRUJILLO, D. Mixed primal dual method for nuclear waste disposal far field simulation. *Computer Geosciences 8* (2004), 173–185.

Roland Becker and David Trujillo
Université de Pau et des Pays de l'Adour
IPRA-LMA BP 1155
F-64013 Pau Cedex
`roland.becker@univ-pau.fr` and `david.trujillo@univ-pau.fr`

Monografías Matemáticas *García de Galdeano* **37**, 69–78 (2012)

VARIATIONAL INTEGRATORS OF FRACTIONAL LAGRANGIAN SYSTEMS IN THE FRAMEWORK OF DISCRETE EMBEDDINGS

Loïc Bourdin

Abstract. This paper is a summary of the theory of discrete embeddings introduced in [5]. A discrete embedding is an algebraic procedure associating a numerical scheme to a given ordinary differential equation. Lagrangian systems possess a variational structure called Lagrangian structure. We are specially interested in the conservation at the discrete level of this Lagrangian structure by discrete embeddings. We then replace in this framework the variational integrators developed in [10, Chapter VI.6] and in [12]. Finally, we extend the notion of discrete embeddings and variational integrators to fractional Lagrangian systems.

Keywords: Lagrangian systems, variational integrator, fractional calculus.

AMS classification: 70H03, 37K05, 26A33.

Introduction

The theoretical framework of embeddings of dynamical systems is initiated by Cresson and Darses in [7]. A review of the subject is given in [6]. An embedding of an ordinary or partial differential equation is a way to give a sense to this equation over a larger set of solutions. As an example, the stochastic embedding developed in [7] allows to give a meaning of a differential equation over the set of stochastic processes.

We are specially interested in Lagrangian systems covering a large set of dynamical behaviors and widely used in classical mechanics, [2]. These systems possess a variational structure called Lagrangian structure, i.e. their solutions correspond to critical points of Lagrangian functionals, [2, p. 57]. The Lagrangian structure is intrinsic and induces strong constraints on the qualitative behavior of the solutions. The conservation of this structure by embedding seems then important. In [7], the authors construct stochastic embeddings which preserve the variational structure of Lagrangian systems, i.e. the generalized solutions are also characterized as critical points of generalized Lagrangian functionals.

This paper is a summary of the theory of discrete embeddings introduced in [5] where, as in [7], we are interested in the conservation of the Lagrangian structure of Lagrangian systems. We then refer to [5] for more details and for the proof of some results.

A discrete embedding is an algebraic procedure associating a numerical scheme to a given differential equation, in particular to a given Lagrangian system. On the other hand, defining a discrete embedding induces a discretization of the Lagrangian functional associated and we can develop a discrete calculus of variations on this one: this leads to a numerical scheme called variational integrator. The variational integrators, developed in [10, Chapter VI.6] and [12], are then numerical schemes for Lagrangian systems preserving their variational structures.

Thus, we propose the following definition: a discrete embedding is said to be coherent if the two discrete versions obtained (the direct one and the variational integrator) of a Lagrangian system coincide. Hence, a coherent discrete embedding conserves at the discrete level the Lagrangian structure of a Lagrangian system.

Recently, many studies have been devoted to fractional Lagrangian systems, [1, 7]. They arise for example in fractional optimal control theory, [9]. They are difficult to solve explicitly, it is then interesting to develop efficient numerical schemes to such systems.

Some preliminary results on fractional discrete operators and on the discretization of fractional Euler-Lagrange equations have been discussed by several authors, [3, 4, 8]. In this paper, we extend the discrete embedding point of view, the corresponding problem of coherence and the associated notion of variational integrator to the fractional case.

The paper is organized as follows. In Section 1, we define the notion of discrete embeddings of differential equations. Section 2 recalls definitions and results concerning Lagrangian systems and we apply the previous theory of discrete embeddings to Lagrangian systems. Then, we recall the strategy of variational integrators of Lagrangian systems in the framework of discrete embeddings and we finally present the problem of coherence of a discrete embedding. Section 3 is devoted to the extension of discrete embeddings to the fractional case.

§1. Notion of discrete embeddings

In this paper, we consider classical and fractional differential systems in $\mathbb{R}^d$ where $d \in \mathbb{N}^*$ is the dimension. The trajectories of these systems are curves q in $C^0([a,b],\mathbb{R}^d)$ where $a < b$ are two reals. For smooth enough functions q, we denote $\dot{q} = dq/dt$ and $\ddot{q} = d^2q/dt^2$.

1.1. Discrete embeddings

Definition 1. Defining a discrete embedding means giving a discrete version of the following elements: the curves $q \in C^0([a,b],\mathbb{R}^d)$, the derivative operator d/dt and the functionals $a : C^0([a,b],\mathbb{R}^d) \longrightarrow \mathbb{R}$. More precisely, it means giving:

- an application $q \longmapsto q^h$ where $q^h \in (\mathbb{R}^d)^{m_1}$,
- a discrete operator $\Delta : (\mathbb{R}^d)^{m_1} \longrightarrow (\mathbb{R}^d)^{m_2}$ discretizing the differential operator d/dt,
- an application $a \longmapsto a^h$ where $a^h : (\mathbb{R}^d)^{m_1} \longrightarrow \mathbb{R}$,

where $m_1, m_2 \in \mathbb{N}^*$.

In order to illustrate Definition 1, we define *backward and forward finite differences embeddings*. For all the rest of the paper, we fix $\sigma = \pm$ and $N \in \mathbb{N}^*$. We denote by $h = (b-a)/N$ the step size of the discretization and $\tau = (t_k)_{k=0,\dots,N}$ the following partition of $[a,b]$:

$$\forall k = 0,\dots,N, \quad t_k = a + kh.$$

Definition 2 (Case $\sigma = -$)**.** We call backward finite differences embedding denoted by $FDE-$ the definition of the following elements: the application

$$\begin{aligned} \mathrm{disc} : C^0([a,b],\mathbb{R}^d) &\longrightarrow (\mathbb{R}^d)^{N+1} \\ q &\longmapsto (q(t_k))_{k=0,\dots,N}, \end{aligned}$$

and the discrete operator

$$\begin{aligned} \Delta_- : \qquad (\mathbb{R}^d)^{N+1} &\longrightarrow (\mathbb{R}^d)^N \\ Q = (Q_k)_{k=0,\dots,N} &\longmapsto \left(\frac{Q_k - Q_{k-1}}{h}\right)_{k=1,..,N}. \end{aligned}$$

Definition 3 (Case $\sigma = +$)**.** We call forward finite differences embedding denoted by $FDE+$ the definition of the following elements: the application disc and the discrete operator

$$\begin{aligned} \Delta_+ : \qquad (\mathbb{R}^d)^{N+1} &\longrightarrow (\mathbb{R}^d)^N \\ Q = (Q_k)_{k=0,\dots,N} &\longmapsto \left(\frac{Q_k - Q_{k+1}}{h}\right)_{k=0,..,N-1}. \end{aligned}$$

Let us notice that the discrete analogous of d/dt in $FDE\sigma$ is then $-\sigma\Delta_\sigma$. We use these notations in order to be uniform with the fractional notations (see Section 3).

1.2. Direct discrete embeddings

Defining a discrete embedding allows us to define a direct discrete version of a given differential equation:

Definition 4. Let be fixed a discrete embedding as defined in Definition 1 and let (E) be an ordinary differential equation of unknown $q \in C^0([a,b],\mathbb{R}^d)$ given by:

$$O(q) = 0, \tag{E}$$

where O is a differential operator shaped as $O = \sum_i f_i(\cdot)(d/dt)^i \circ g_i(\cdot)$ where f_i, g_i are functions. Then, the direct discrete embedding of (E) is (E_h) the system of equations of unknown $q^h \in (\mathbb{R}^d)^{m_1}$ given by:

$$O^h(q^h) = 0, \tag{E_h}$$

where O^h is the discretized operator of O given by $O^h = \sum_i f_i(\cdot)\Delta^i \circ g_i(\cdot)$.

As an example, we consider the Newton's equation with friction of unknown $q \in C^0([a,b],\mathbb{R}^d)$ given by:

$$\forall t \in [a,b], \quad \ddot{q}(t) + \dot{q}(t) + q(t) = 0. \tag{NE}$$

Then, the direct discrete embedding of (NE) with respect to $FDE-$ is (NE_h) the system of equations of unknown $Q \in (\mathbb{R}^d)^{N+1}$ given by:

$$\forall k = 2,\dots,N, \quad \frac{Q_k - 2Q_{k-1} + Q_{k-2}}{h^2} + \frac{Q_k - Q_{k-1}}{h} + Q_k = 0. \tag{NE_h}$$

The direct discrete embedding of an ordinary differential equation is strongly dependent on the form of the differential operator O (and not on its equivalence class). The process $O \longrightarrow O^h$ is not an application. For example, the discretized operator O^h of $O = d/dt \circ \sin(\cdot) = d/dt(\cdot)\ \cos(\cdot)$ is different depending on the writing of O.

1.3. Direct discrete embeddings of Lagrangian systems

We recall now classical definitions and theorems concerning Lagrangian systems. We refer to [2] for a detailed study and for a detailed proof of Theorem 1.

Definition 5. A Lagrangian functional is an application defined by:

$$\begin{aligned} \mathcal{L} : C^2([a,b],\mathbb{R}^d) &\longrightarrow \mathbb{R} \\ q &\longmapsto \int_a^b L(q(t),\dot{q}(t),t)\,dt, \end{aligned}$$

where L is a Lagrangian i.e. a C^2 application defined by:

$$\begin{aligned} L : \mathbb{R}^d \times \mathbb{R}^d \times [a,b] &\longrightarrow \mathbb{R} \\ (x,v,t) &\longmapsto L(x,v,t). \end{aligned}$$

An *extremal* (or *critical point*) of a Lagrangian functional $\mathcal{L}$ is a trajectory q such that $D\mathcal{L}(q)(w) = 0$ for any *variations* w (i.e. $w \in C^2([a,b],\mathbb{R}^d)$, $w(a) = w(b) = 0$), where $D\mathcal{L}(q)(w)$ is the differential of $\mathcal{L}$ in q along the direction w. Extremals of a Lagrangian functional can be characterized as solution of a differential equation of order 2:

Theorem 1 (Variational principle)**.** *Let $\mathcal{L}$ be a Lagrangian functional associated to the Lagrangian L and let $q \in C^2([a,b],\mathbb{R}^d)$. Then, q is an extremal of $\mathcal{L}$ if and only if q is solution of the Euler-Lagrange equation given by:*

$$\forall t \in \,]a,b[\,,\quad \frac{\partial L}{\partial x}(q(t),\dot{q}(t),t) - \frac{d}{dt}\left(\frac{\partial L}{\partial v}(q(t),\dot{q}(t),t)\right) = 0. \tag{EL}$$

We now apply definitions of Section 1 on Lagrangian systems.

Proposition 2. *Let L be a Lagrangian and let* (EL) *be its associated Euler-Lagrange equation. The direct discrete embedding of* (EL) *with respect to FDEσ is given by:*

$$\frac{\partial L}{\partial x}(Q,-\sigma\Delta_\sigma Q,\tau) + \sigma\Delta_\sigma\left(\frac{\partial L}{\partial v}(Q,-\sigma\Delta_\sigma Q,\tau)\right) = 0, \quad Q \in (\mathbb{R}^d)^{N+1}. \tag{1}$$

We refer to [5] for a concrete example illustrating Theorem 1 and Proposition 2.

§2. Discrete embeddings and variational integrators of Lagrangian systems

A direct discrete embedding is only based on the form of the differential operator which is dependent of the coordinates system and consequently is not intrinsic. Then, a natural question arises: *what can be said about the conservation of intrinsic properties of a differential equation by a discrete embedding?* This paper is devoted to the conservation by discrete embeddings of the Lagrangian structure of Lagrangian systems. More precisely, Theorem 1 shows that (EL) possesses a variational structure: *the direct discrete embedding being a procedure mainly algebraic, does* (1) *possess a variational structure too?* It is *not always true*.

However, a variational integrator, developed in [10, Chapter VI.6] and in [12], is a discretization of a Lagrangian system preserving its variational structure. Indeed, it is based on the discrete analogous of the variational principle on a discrete version of the associated Lagrangian functional.

In our framework, the discretization of the Lagrangian functional is induced by giving a discrete embedding.

2.1. Discrete Lagrangian functionals and discrete calculus of variations

In this subsection, as an example, we are going to work exclusively in the framework of *FDEσ*. Giving *FDEσ* induces the discretization of a Lagrangian functional as long as a quadrature formula is fixed in order to approximate integrals. We choose the usual σ-quadrature formula of Gauss: for a continuous function f on $[a,b]$, we discretize $\int_a^b f(t)dt$ by $h\sum_{k\in I_\sigma} f(t_k)$ where $I_+ = \{0,\dots,N-1\}$ and $I_- = \{1,\dots,N\}$.

This process defines the *Gauss finite differences embedding* denoted by *Gauss-FDEσ*. Such a choice allows to keep at the discrete level the following fundamental result:

$$\int_a^b \dot{q}(t)\,dt = q(b) - q(a) \quad \xrightarrow{\textit{Gauss-FDE}\sigma} \quad h\sum_{k\in I_\sigma}(-\sigma\Delta_\sigma Q)_k = Q_N - Q_0.$$

Proposition 3. *Let $\mathcal{L}$ be a Lagrangian functional associated to a Lagrangian L. The discrete Lagrangian functional associated to $\mathcal{L}$ with respect to Gauss-FDEσ is given by:*

$$\begin{array}{rcl} \mathcal{L}_h^\sigma : & (\mathbb{R}^d)^{N+1} & \longrightarrow \mathbb{R} \\ & Q = (Q_k)_{k=0,\dots,N} & \longmapsto h\displaystyle\sum_{k\in I_\sigma} L(Q_k, (-\sigma\Delta_\sigma Q)_k, t_k). \end{array}$$

Once the discrete version of the Lagrangian functional is formulated, we can develop a discrete calculus of variations on it: this leads to a variational integrator. Let $\mathcal{L}$ be a Lagrangian functional and $\mathcal{L}_h^\sigma$ the discrete Lagrangian functional associated with respect to *Gauss-FDEσ*. A *discrete extremal* (or *discrete critical point*) of $\mathcal{L}_h^\sigma$ is an element Q in $(\mathbb{R}^d)^{N+1}$ such that $D\mathcal{L}_h^\sigma(Q)(W) = 0$ for any *discrete variations* W (i.e. $W \in (\mathbb{R}^d)^{N+1}$, $W_0 = W_N = 0$). Discrete extremals of $\mathcal{L}_h^\sigma$ can be characterized as solution of a system of equations:

Theorem 4 (Discrete variational principle)**.** *Let $\mathcal{L}_h^\sigma$ be the discrete Lagrangian functional associated to the Lagrangian L with respect to Gauss-FDEσ. Then, Q in $(\mathbb{R}^d)^{N+1}$ is a discrete extremal of $\mathcal{L}_h^\sigma$ if and only if Q is solution of the following system of equations (called discrete Euler-Lagrange equation) given by:*

$$\frac{\partial L}{\partial x}(Q, -\sigma\Delta_\sigma Q, \tau) - \sigma\Delta_{-\sigma}\left(\frac{\partial L}{\partial v}(Q, -\sigma\Delta_\sigma Q, \tau)\right) = 0, \quad Q \in (\mathbb{R}^d)^{N+1}. \qquad (EL_h^\sigma)$$

(EL_h^σ) is obtained from (EL) by variational integrator. Its variational origin allows us to say that it possesses a Lagrangian structure. Then, we have conservation at the discrete level of the Lagrangian structure by variational integrator.

Let us note that an asymmetry appears in (EL_h^σ): indeed, we have a composition between the two discrete operators Δ_+ and Δ_-. We notice that this asymmetry does not appear in the continuous space in (EL).

2.2. Problem of coherence of a discrete embedding

Hence, defining a discrete embedding leads to two discrete versions of an Euler-Lagrange equation: the first one (1) obtained by direct discrete embedding and the second one (EL_h^σ) corresponding to a variational integrator. The direct discrete embedding is an algebraic procedure (respecting for example the law of semi-group of the differential operator d/dt). On the contrary, a variational integrator is mainly based on a dynamical approach via the extremals of a functional.

However, we are interested in the conservation at the discrete level of the Lagrangian structure of Lagrangian systems. We then propose the following definition: a discrete embedding is said to be *coherent* if the two numerical schemes coincide. Precisely, a discrete embedding is coherent if it makes the following diagram commutative:

$$\begin{tikzcd}
\text{Lagrangian functional} \arrow[r, "\text{Functional discretization}"] \arrow[d, "\text{Variational principle}"'] & \text{Discrete Lagrangian functional} \arrow[d, "\text{Discr. var. principle}"] \\
\text{Euler-Lagrange equation} \arrow[r, "\text{Direct discrete embedding}"'] \arrow[r, bend left, "\text{Variational integrator}"] & \text{Numerical scheme}
\end{tikzcd}$$

Thus, a coherent discrete embedding provides a direct discrete version of a Lagrangian system preserving its Lagrangian structure.

The previous study leads to a default of coherence of *Gauss-FDEσ*. Indeed, algorithms obtained by direct discrete embedding (1) and obtained by discrete variational principle (EL_h^σ) do not coincide. The problem is to understand *why there is not asymmetry appearing in the direct discrete embedding?* It seems that we miss dynamical informations in the formulation of Lagrangian systems at the continuous level which are pointed up in the discrete space with the asymmetric discrete operators $(-\sigma\Delta_\sigma)_{\sigma=\pm}$. Nevertheless, this default of coherence can be corrected using a different writing of the initial Euler-Lagrange equation.

2.3. Rewriting of the Euler-Lagrange equation and discrete embeddings

The usual way to derive differential equations in Physics is to built a continuous model using discrete data. However, this process gives only an information in one direction of time. As a consequence, a discrete evaluation of the velocity corresponds in general at the continuous level to the evaluation of the right or left derivative. In general, we replace the right (or left) derivative by the classical derivative d/dt. However, this procedure assumes that the underlying solution is differentiable. This assumption is not only related to the regularity of the solutions but also to the reversibility of the systems (the right and left derivatives are equal). In this section, we introduce asymmetric Lagrangian systems which are obtained with functionals depending only on left or only on right derivatives. We prove in this case that *Gauss-FDEσ* is coherent.

Definition 6. For $f : [a,b] \longrightarrow \mathbb{R}^d$ smooth enough function, we denote:

$$\forall t \in]a,b], \quad d_- f(t) = \lim_{h \to 0^+} \frac{f(t) - f(t-h)}{h}$$

and

$$\forall t \in [a,b[, \quad d_+ f(t) = \lim_{h \to 0^+} \frac{f(t) - f(t+h)}{h}.$$

Although we have $d_- f = -d_+ f = \dot{f}$ for a differentiable function f, it is interesting to use these notations in order to keep dynamical informations.

Definition 7. An asymmetric Lagrangian functional is an application:

$$\begin{aligned} \mathcal{L}^\sigma : C^2([a,b],\mathbb{R}^d) &\longrightarrow \mathbb{R} \\ q &\longmapsto \int_a^b L(q(t), -\sigma d_\sigma q(t), t)\, dt, \end{aligned}$$

where L is a Lagrangian.

Then, by calculus of variations, we obtain the following characterization of the extremals of an asymmetric Lagrangian functional:

Theorem 5 (Variational principle)**.** *Let $\mathcal{L}^\sigma$ be an asymmetric Lagrangian functional associated to the Lagrangian L and let $q \in C^2([a,b],\mathbb{R}^d)$. Then, q is an extremal of $\mathcal{L}^\sigma$ if and only if q is solution of the asymmetric Euler-Lagrange equation:*

$$\forall t \in \,]a,b[\,, \quad \frac{\partial L}{\partial x}(q(t), -\sigma d_\sigma q(t), t) - \sigma d_{-\sigma}\left(\frac{\partial L}{\partial v}(q(t), -\sigma d_\sigma q(t), t)\right) = 0. \qquad (EL^\sigma)$$

Hence, (EL^σ) possesses a variational structure. *Is it conserved by discrete embeddings?* In order to embed (EL^σ), we have to discretize two differential operators at the same time. We then define the following asymmetric version of *Gauss-FDEσ*:

Definition 8. We call the asymmetric version of *Gauss-FDEσ* the definition of the following elements: the application disc, the σ-quadrature formula of Gauss and the discrete operators Δ_- and Δ_+ discretizing respectively the operators d_- and d_+.

Proposition 6. *The asymmetric version of Gauss-FDEσ is a coherent discrete embedding. Indeed, the direct discrete embedding and the variational integrator of (EL^σ) in the framework of the asymmetric Gauss-FDEσ lead to the same numerical scheme: (EL^σ_h).*

We notice that the rewriting (EL^σ) of (EL) provides additional dynamical informations which allows the asymmetric *Gauss-FDEσ* to unify the algebraic and the dynamical approaches in the discretization of a Lagrangian system. Moreover, this rewriting can be justified by the fractional calculus as we will see in Section 3.

§3. Discrete embeddings and variational integrators of fractional Lagrangian systems

3.1. Fractional derivatives and fractional Lagrangian systems

Fractional calculus is the generalization of the derivative notion to real orders. We refer to [11, 14] for many different ways generalizing this notion. For the whole paper, we fix $0 < \alpha < 1$ and for any $r \in \mathbb{N}^*$, we denote by $\alpha_r = (-\alpha)(1-\alpha)\cdots(r-1-\alpha)/r!$ and $\alpha_0 = 1$. We are going to use the classical notions of Grünwald-Letnikov. The following definition is extracted from [13].

Definition 9. Let f be an element of $C^1([a,b],\mathbb{R}^d)$. The Grünwald-Letnikov fractional left derivative of order α with inferior limit a of f is:

$$\forall t \in \,]a,b]\,,\quad D^\alpha_- f(t) = \lim_{\substack{h\to 0\\ nh=t-a}} \frac{1}{h^\alpha}\sum_{r=0}^{n} \alpha_r f(t-rh)$$

and the Grünwald-Letnikov fractional right derivative of order α with superior limit b of f is:

$$\forall t \in [a,b[\,,\quad D^\alpha_+ f(t) = \lim_{\substack{h\to 0\\ nh=b-t}} \frac{1}{h^\alpha}\sum_{r=0}^{n} \alpha_r f(t+rh).$$

Recently, an important activity has been devoted to fractional Lagrangian systems for the purpose of optimal control, mechanics, engineering and Physics, [1, 3, 9]. We recall definitions and results concerning these fractional systems, we refer to [1] for a detailed study and for a detailed proof of Theorem 7.

Definition 10. A fractional Lagrangian functional of order α is an application defined by:

$$\begin{aligned}\mathcal{L}^{\sigma,\alpha} : C^2([a,b],\mathbb{R}^d) &\longrightarrow \mathbb{R}\\ q &\longmapsto \int_a^b L(q(t), -\sigma D^\alpha_\sigma q(t), t)\,dt,\end{aligned}$$

where L is a Lagrangian.

We can give a characterization of extremals of a fractional Lagrangian functional as solutions of a fractional differential equation:

Theorem 7 (Variational principle)**.** *Let $\mathcal{L}^{\sigma,\alpha}$ be a fractional Lagrangian functional of order α associated to the Lagrangian L and let q be an element of $C^2([a,b],\mathbb{R}^d)$. Then, q is an extremal of $\mathcal{L}^{\sigma,\alpha}$ if and only if q is solution of the fractional Euler-Lagrange equation:*

$$\forall t \in]a,b[,\quad \frac{\partial L}{\partial x}(q(t), -\sigma D^\alpha_\sigma q(t), t) - \sigma D^\alpha_{-\sigma}\left(\frac{\partial L}{\partial v}(q(t), -\sigma D^\alpha_\sigma q(t), t)\right) = 0. \qquad (EL^{\sigma,\alpha})$$

We refer to [1] for a detailed proof. Hence, in the fractional case, we find an asymmetry again making a link with the asymmetric rewriting of (EL) into (EL^σ).

As in the classical case, we conclude that $(EL^{\sigma,\alpha})$ possesses a Lagrangian structure and we are iterested by its conservation at the discrete level by discrete embeddings.

3.2. Discrete embeddings of fractional Lagrangian systems

There exist many studies concerning the discretization of fractional differential equations but without the point of view of discrete embeddings. We refer to [3, 4]. By referring to the notion of Grünwald-Letnikov [8], we give the following definition:

Definition 11. The Gauss Grünwald-Letnikov embedding denoted by *Gauss-GLEσ* is the definition of the following elements: the application disc, the σ-quadrature formula of Gauss and the discrete operators

$$\begin{aligned} \Delta_-^\alpha : \qquad & (\mathbb{R}^d)^{N+1} \longrightarrow (\mathbb{R}^d)^N \\ Q = (Q_k)_{k=0,\dots,N} & \longmapsto \left(\frac{1}{h^\alpha} \sum_{r=0}^{k} \alpha_r Q_{k-r} \right)_{k=1,\dots,N} \end{aligned}$$

and

$$\begin{aligned} \Delta_+^\alpha : \qquad & (\mathbb{R}^d)^{N+1} \longrightarrow (\mathbb{R}^d)^N \\ Q = (Q_k)_{k=0,\dots,N} & \longmapsto \left(\frac{1}{h^\alpha} \sum_{r=0}^{N-k} \alpha_r Q_{k+r} \right)_{k=0,\dots,N-1} . \end{aligned}$$

These discrete operators are respectively the discrete versions of D_-^α and D_+^α.

We are first interested in the variational integrator of $(EL^{\sigma,\alpha})$ in the framework of *Gauss-GLEσ*. Giving *Gauss-GLEσ* allows us to formulate the discrete version of a fractional Lagrangian functional:

Proposition 8. *Let $\mathcal{L}^{\sigma,\alpha}$ be the fractional Lagrangian functional associated to the Lagrangian L. The discrete fractional Lagrangian functional associated to $\mathcal{L}^{\sigma,\alpha}$ with respect to Gauss-GLEσ is given by:*

$$\begin{aligned} \mathcal{L}_h^{\sigma,\alpha} : \qquad & (\mathbb{R}^d)^{N+1} \longrightarrow \mathbb{R} \\ Q = (Q_k)_{k=0,\dots,N} & \longmapsto h \sum_{k \in I_\sigma} L(Q_k, (-\sigma \Delta_\sigma^\alpha Q)_k, t_k). \end{aligned}$$

Then, discrete extremals of the discrete fractional Lagrangian functional can be characterized as solutions of a system of equations:

Theorem 9 (Discrete variational principle)**.** *Let $\mathcal{L}_h^{\sigma,\alpha}$ be a discrete fractional Lagrangian functional associated to the Lagrangian L with respect to Gauss-GLEσ. Then, Q in $(\mathbb{R}^d)^{N+1}$ is a discrete extremal of $\mathcal{L}_h^{\sigma,\alpha}$ if and only if Q is solution of the following system of equations, called the discrete fractional Euler-Lagrange equation:*

$$\frac{\partial L}{\partial x}(Q, -\sigma \Delta_\sigma^\alpha Q, \tau) - \sigma \Delta_{-\sigma}^\alpha \left(\frac{\partial L}{\partial v}(Q, -\sigma \Delta_\sigma^\alpha Q, \tau) \right) = 0, \quad Q \in (\mathbb{R}^d)^{N+1}. \qquad (EL_h^{\sigma,\alpha})$$

We conclude with the following proposition:

Proposition 10. *Gauss-GLEσ is a coherent discrete embedding. Indeed, the direct discrete embedding and the variational integrator of $(EL^{\sigma,\alpha})$ in the framework of Gauss-GLEσ lead to the same numerical scheme: $(EL_h^{\sigma,\alpha})$.*

References

[1] Agrawal, O. Formulation of Euler-Lagrange equations for fractional variational problems. *J. Math. Anal. Appl. 272*, 1 (2002), 368–379.

[2] Arnold, V. *Mathematical Methods of Classical Mechanics.* Graduate Texts in Mathematics. Springer-Verlag New York Inc, New York, USA, 1979.

[3] Baleanu, D., Defterli, O., and Agrawal, O. A central difference numerical scheme for fractional optimal control problems. *J. Vib. Control 15*, 4 (2009), 583–597.

[4] Bastos, N., Ferreira, R., and Torres, D. Discrete-time fractional variational problems. *Signal Processing 91*, 3 (2011), 513 – 524. Advances in Fractional Signals and Systems.

[5] Bourdin, L., Cresson, J., Greff, I., and Inizan, P. Variational integrators on fractional lagrangian systems in the framework of discrete embeddings. *arXiv:1103.0465v1 [math.DS]* (2011).

[6] Cresson, J. Introduction to embedding of lagrangian systems. *International journal for biomathematics and biostatistics 1*, 1 (2010), 23–31.

[7] Cresson, J., and Darses, S. Stochastic embedding of dynamical systems. *J. Math. Phys. 48*, 7 (2007), 072703, 54.

[8] Dubois, F., Galucio, A.-C., and Point, N. Introduction à la dérivation fractionnaire. théorie et applications. *Série des Techniques de l'ingénieur* (2009).

[9] Frederico, G., and Torres, D. Fractional conservation laws in optimal control theory. *Nonlinear Dynam. 53*, 3 (2008), 215–222.

[10] Hairer, E., Lubich, C., and Wanner, G. *Geometric numerical integration*, second ed., vol. 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2006. Structure-preserving algorithms for ordinary differential equations.

[11] Hilfer, R. Threefold introduction to fractional derivatives. *Anomalous Transport: Foundations and Applications* (2008).

[12] Marsden, J., and West, M. Discrete mechanics and variational integrators. *Acta Numer. 10* (2001), 357–514.

[13] Podlubny, I. *Fractional differential equations*, vol. 198 of *Mathematics in Science and Engineering*. Academic Press Inc., San Diego, CA, 1999. An introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications.

[14] Samko, S., Kilbas, A., and Marichev, O. *Fractional integrals and derivatives*. Gordon and Breach Science Publishers, Yverdon, 1993. Theory and applications, Translated from the 1987 Russian original.

Loïc Bourdin
LMA, University of Pau
Postal address IPRA BP 1155 Pau Cedex (France)
`bourdin.l@etud.univ-pau.fr`

Monografías Matemáticas *García de Galdeano* **37**, 79–90 (2012)

An application of Carleman inequalities for a curved quantum guide

Laure Cardoulis

Abstract. We consider in this paper the Schrödinger operator $-i\partial_t - \Delta$ on a curved quantum guide in $\mathbb{R}^2$ for which the reference curve is asymptotically straight. Using an adapted Carleman estimate, we establish a local estimation result for the curvature with a single observation.

Keywords: Schrödinger Operators, quantum guide, curvature, Carleman estimate, inverse problem.

AMS classification: 35J10.

§1. Introduction

Let $\Omega \subset \mathbb{R}^2$ be a curved quantum guide with a fixed width $d > 0$ and let $T > 0$. We consider the Schrödinger operator

$$H := -i\partial_t - \Delta \text{ in } \Omega \times (0, T).$$

We proceed as in [8] and [4]. We denote by $\Gamma = (\Gamma_1, \Gamma_2)$ the function which characterizes the reference curve and by $N = (N_1, N_2)$ the outgoing normal. We denote by

$$\Omega_1 := \mathbb{R} \times (d, 2d).$$

Each point (x, y) of Ω is described by the curvilinear coordinates (s, u) as follows:

$$f : \Omega_1 \longrightarrow \Omega \quad \text{with} \quad (x, y) = f(s, u) = \Gamma(s) + uN(s). \tag{1}$$

We assume $\Gamma_1'(s)^2 + \Gamma_2'(s)^2 = 1$ and we recall that the signed curvature γ of Γ is defined by $\gamma(s) = -\Gamma_1''(s)\Gamma_2'(s) + \Gamma_2''(s)\Gamma_1'(s)$, named so because $|\gamma(s)|$ represents the curvature of the reference curve at s. We assume throughout this paper that:

Assumption 1.

- $\gamma \in C^3(\mathbb{R})$, $\gamma^{(k)} \in L^\infty(\mathbb{R})$ *for each* $k = 0, 1, 2, 3$, *where* $\gamma^{(k)}$ *denotes the k-th derivatives of* γ.
- $\gamma(s) \to 0$ *as* $|s| \to \infty$ *and* $1 - 2d\|\gamma\|_\infty > 0$, *where* $\|\gamma\|_\infty := \sup_{s\in\mathbb{R}} |\gamma(s)| = \|\gamma\|_{L^\infty(\mathbb{R})}$.

Note that, by the inverse function theorem, the map f defined by (1) is a diffeomorphism provided $1 - u\gamma(s) \neq 0$, for all u, s, which is guaranteed by Assumption 1. The curvilinear coordinates (s, u) are locally orthogonal so the metric in Ω is expressed with respect to them through a diagonal metric tensor $\begin{pmatrix} (1-u\gamma(s))^2 & 0 \\ 0 & 1 \end{pmatrix}$. The transition to the curvilinear coordinates

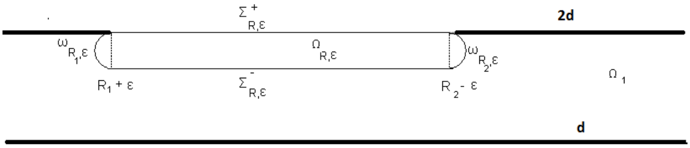


Figure 1: Geometry of the problem

represents an isometric map of $L^2(\Omega)$ to $L^2(\Omega_1, g^{1/2}\ dsdu)$, where $g(s,u) := (1 - u\gamma(s))^2$ is the Jacobian $\partial(x,y)/\partial(s,u)$. Therefore we can replace the operator H (with the metric $dx\, dy$ on Ω) by the operator H_g (with the metric $g^{1/2} ds\, du$ on Ω_1), where

$$H_g := -i\partial_t - g^{-1/2}\partial_s(g^{-1/2}\partial_s) - g^{-1/2}\partial_u(g^{1/2}\partial_u).$$

Then we can rewrite the operator H_g into a Schrödinger-type operator (with the metric $ds\, du$ on Ω_1). Indeed, using the unitary transformation $U_g(\psi) = g^{1/4}\psi$, setting $H_\gamma := U_g H_g U_g^{-1}$, we get

$$H_\gamma = -i\partial_t - \partial_s(c_\gamma(s,u)\partial_s) - \partial_u^2 + V_\gamma(s,u)$$

with

$$c_\gamma(s,u) = \frac{1}{(1 - u\gamma(s))^2} \tag{2}$$

and

$$V_\gamma(s,u) = -\frac{\gamma^2(s)}{4(1-u\gamma(s))^2} - \frac{u\gamma''(s)}{2(1-u\gamma(s))^3} - \frac{5u^2\gamma'^2(s)}{4(1-u\gamma(s))^4}.$$

Let $R := (R_1, R_2) \in \mathbb{R}^2$ and $\epsilon > 0$. We denote by

$$\Omega_{R,\epsilon} := \omega_{R,\epsilon} \cup (]R_1 + \epsilon, R_2 - \epsilon[\times]2d - 2\epsilon, 2d[)$$

a regular bounded domain in Ω_1, with

$$\begin{aligned}
&\omega_{R,\epsilon} := \omega_{R_1,\epsilon} \cup \omega_{R_2,\epsilon},\\
&\omega_{R_1,\epsilon} := \{(s,u) \in \mathbb{R}^2,\ R_1 < s < R_1 + \epsilon,\ 2d - 2\epsilon < u < 2d,\ (s - R_1 - \epsilon)^2 + (u - 2d + \epsilon)^2 < \epsilon\},\\
&\omega_{R_2,\epsilon} := \{(s,u) \in \mathbb{R}^2,\ R_2 - \epsilon < s < R_2,\ 2d - 2\epsilon < u < 2d,\ (s - R_2 + \epsilon)^2 + (u - 2d + \epsilon)^2 < \epsilon\}.
\end{aligned}$$

Note that $\omega_{R_1,\epsilon}$ and $\omega_{R_2,\epsilon}$ are half-balls and let (see Figure 1)

$$\begin{aligned}
\Sigma^+_{R,\epsilon} &:= [R_1 + \epsilon, R_2 - \epsilon] \times \{2d\}, & \Gamma_{R,\epsilon} &:= \partial\Omega_{R,\epsilon} - \Sigma^-_{R,\epsilon},\\
\Sigma^-_{R,\epsilon} &:= [R_1 + \epsilon, R_2 - \epsilon] \times \{2d - 2\epsilon\}, & \Gamma_\epsilon &:= (\partial\omega_{R_1,\epsilon} \cup \partial\omega_{R_2,\epsilon}) \cap \partial\Omega_{R,\epsilon}.
\end{aligned}$$

We now consider the following Schrödinger equation

$$\begin{cases}
H_\gamma z := -i\partial_t z(s,u,t) - \partial_s(c_\gamma(s,u)\partial_s z(s,u,t)) - \partial_u^2 z(s,u,t) + V_\gamma(s,u)z(s,u,t) = 0,\\
(s,u,t) \in \Omega_{R,\epsilon} \times (0,T),\\
z(s,u,t) = l(x,y,t),\ (s,u) \in \partial\Omega_{R,\epsilon},\ t \in (0,T),\\
z(s,u,0) = z_0(s,u),\ (s,u) \in \Omega_{R,\epsilon}.
\end{cases} \tag{3}$$

Our problem can be stated as follows: Is it possible to determine the curvature γ from the measurement of $\partial_\nu(\partial_t z)$ on $\Sigma^+_{R,\epsilon}$?

Let z, depending on ϵ (resp. $\widetilde{z}$, depending on ϵ too) be a solution of (3) associated with (γ, l, z_0) (resp. $(\widetilde{\gamma}, l, z_0)$). We assume that z_0 is a real-valued function and that $(\gamma - \widetilde{\gamma})(s) \neq 0$ and $(\gamma' - \widetilde{\gamma}')(s) \neq 0$ for all $s \in [R_1, R_2]$. Our main result is

$$\|\gamma - \widetilde{\gamma}\|^2_{L^2(\Omega_{R,\epsilon})} \leq C\|\partial_\nu(\partial_t z - \partial_t \widetilde{z})\|^2_{L^2(\Sigma^+_{R,\epsilon}\times(-T,T))} + C\epsilon,$$

where C is a positive constant which depends on d, T and where the above norms are weighted Sobolev norms.

This paper gives a quantum mechanics application of an inverse problem and we use for that the important tool of Carleman estimates. Indeed, the method of Carleman inequalities has been introduced in the field of inverse problems by Bukhgeim and Klibanov [2, 3, 11, 12, 13, 14] and constitutes a very efficient tool to derive observability estimates. Note also that even if the spectral properties of curved quantum guides have been intensively studied for several years (see [7, 8, 9] e.g.), up to our knowledge there are few results for inverse problems associated with curved quantum guide (see [4]). The main difficulty here is to recover the curvature γ via two coefficients c_γ and V_γ. Few results have already been obtained for the simultaneous identification of two coefficients with one observation and these two coefficients were not linked up (see [6]). This is not the case here where the coefficients c_γ and V_γ both depend on γ. Another difficulty when we work with Carleman estimates is the existence of the weight function $\widetilde{\beta}$ (see Assumption 2). And usually this imposes restrictive conditions for the diffusion coefficient i.e. in our case for c_γ and therefore for γ. This is why, due to these two difficulties which come from our model (a curved guide with an asymptotically straight curvature γ), we work in the subdomain $\Omega_{R,\epsilon}$ instead of the whole strip Ω_1 and we get an additional term $C\epsilon$ in the right hand side of our main result (which was not the case in [5,6]). This paper is organized as follows: Section 2 is devoted to the Carleman inequality adapted to our problem. In Section 3 we state and prove our main result.

§2. Carleman inequality

In this section we obtain a Carleman estimate for a function q equal to zero on $\partial\Omega_{R,\epsilon}\times(-T,T)$ and solution of the Schrödinger equation $H_\gamma q \in L^2(\Omega_{R,\epsilon} \times (-T,T))$. We prove a Carleman estimate for q with a single observation acting on $\Gamma_{R,\epsilon} \times (-T,T)$ in the right-hand side of the estimate. Note that this estimate is quite similar to the one obtained in [1] or [5] but the computations are different. Indeed the weight function $\widetilde{\beta}$ does not satisfy the same pseudo-convexity assumptions (see Assumption 2(iii)). This is the main difference compared to [5] and this is due to the particular form of the operator H_γ where the diffusion coefficient c_γ only appears in the derivatives respect to s.

We use the following notations

$$c := c_\gamma, \quad \nabla_c\beta := \begin{pmatrix} \sqrt{c}\partial_s\beta \\ \partial_u\beta \end{pmatrix} \quad \text{and} \quad \nu_c := \begin{pmatrix} \sqrt{c}\partial_s\nu \\ \partial_u\nu \end{pmatrix},$$

where ν denotes the unit outward normal to $\partial\Omega_{R,\epsilon}$ and we proceed as in [1] or [5]. Let $\widetilde{\beta} := \widetilde{\beta}(s,u)$ be a positive function such that there exists positive constants β_0 and C_{pc} which satisfy:

Assumption 2.

(i) $\widetilde{\beta} \in C^4(\overline{\Omega_{R,\epsilon}})$, *and* $\widetilde{\beta}(s,u) \geq 0$ *for all* $(s,u) \in \Omega_{R,\epsilon}$.

(ii) $|\nabla_c \widetilde{\beta}| \geq \beta_0 > 0$ *in* $\Omega_{R,\epsilon}$, *and* $\nabla_c \widetilde{\beta} \cdot \nu_c \leq 0$ *in* $\Sigma^-_{R,\epsilon}$.

(iii) $2\,\mathrm{Re}\, D^2_c \widetilde{\beta}(\xi, \overline{\xi}) - \frac{1}{c} \nabla_c c \cdot \nabla_c \widetilde{\beta} |\xi_1|^2 + 2|\nabla_c \widetilde{\beta} \cdot \xi|^2 \geq C_{pc} |\xi|^2$ *for all* $\xi = (\xi_1, \xi_2) \in \mathbb{C}$, *where*

$$D^2_c \widetilde{\beta} = \begin{pmatrix} \partial_s(c\partial_s \widetilde{\beta}) & \sqrt{c}\, \partial^2_{su} \widetilde{\beta} \\ \frac{1}{\sqrt{c}}\, \partial_u(c\partial_s \widetilde{\beta}) & \partial^2_u \widetilde{\beta} \end{pmatrix}. \tag{4}$$

This assumption imposes restrictive conditions for the choice of the coefficient $c := c_\gamma$ and thus for the curvature γ in connection with the function $\widetilde{\beta}$ as in [5, 6]. Note that there exists functions satisfying such conditions. Indeed if we assume that $\widetilde{\beta}(s,u) := \beta_1(s) + \beta_2(u)$, these conditions can be written in the following form:

$$A := 2\partial_s(c\partial_s\beta_1) - c\partial_s c\partial_s\beta_1 - \partial_u c\partial_u\beta_2 + 2c(\partial_s\beta_1)^2 \geq cst > 0 \text{ and } 2AC - B^2 \geq cst > 0,$$

with $B := (1/\sqrt{c})\partial_u c\partial_s\beta_1 + 2\sqrt{c}\partial_s\beta_1\partial_u\beta_2$ and $C := \partial^2_u\beta_2 + (\partial_u\beta_2)^2$. For example if $\widetilde{\beta}(s,u) = e^s + e^u$, these two last conditions become

$$A = (1 - u\gamma(s))^{-3}[(2 - c(s,u))2u\gamma'(s)e^s - 2\gamma(s)e^u] + 2c(s,u)(e^s + e^{2s})$$

and

$$\begin{aligned} 2AC - B^2 = 4c(s,u)\Big[&(2 - c(s,u))u\gamma'(s)(1 - u\gamma(s))^{-1}e^s(e^u + e^{2u}) \\ &- \gamma(s)(1 - u\gamma(s))^{-1}e^u(e^u + e^{2u}) + e^s e^u(1 + e^s + e^u) \\ &- \gamma(s)(1 - u\gamma(s))^{-1}e^{2s}(\gamma(s)(1 - u\gamma(s))^{-1} + 2e^u)\Big]. \end{aligned}$$

We have $A \geq cst > 0$ and $2AC - B^2 \geq cst > 0$ for any curvature γ in

$$\{\gamma \in C^1(\mathbb{R}),\ \gamma' \geq 0,\ \gamma \leq 0,\ (1 - 2d\|\gamma\|_\infty)^{-2} < 2,\ \gamma(s) > -2e^{2d-2\epsilon}(1 - 2d\|\gamma\|_\infty)\ \forall s \in [R_1, R_2]\}.$$

Similar restrictive conditions upon the function c in connection with the function $\widetilde{\beta}$ have also been highlighted for the hyperbolic case in [13, 14].

Then we define $\beta = \widetilde{\beta} + K$ with $K = m\|\widetilde{\beta}\|_{L^\infty(\Omega_{R,\epsilon})}$ and $m > 1$. For $\lambda > 0$ we define on $\Omega_{R,\epsilon} \times (-T, T)$ the functions ϕ and η by

$$\phi(s,u,t) = \frac{e^{\lambda\beta(s,u)}}{(T-t)(T+t)} \quad \text{and} \quad \eta(s,u,t) = \frac{e^{2\lambda K} - e^{\lambda\beta(s,u)}}{(T-t)(T+t)}. \tag{5}$$

For $S > 0$ we set $\psi = e^{-S\eta}q$ and $M\psi := e^{-S\eta}H_\gamma q$. Following [1], we write $M\psi - V_\gamma\psi = M_1\psi + M_2\psi$, with

$$M_1\psi := -i\partial_t\psi - \Delta_c\psi - S^2\lambda^2\phi^2\psi|\nabla_c\beta|^2, \tag{6}$$

$$M_2\psi := -iS\,\partial_t\eta\psi + 2S\lambda\phi\nabla_c\beta \cdot \nabla_c\overline{\psi} + S\lambda^2\phi\psi|\nabla_c\beta|^2 + S\lambda\phi\psi\Delta_c\beta, \tag{7}$$

where

$$\nabla_c\beta := \begin{pmatrix} \sqrt{c}\partial_s\beta \\ \partial_u\beta \end{pmatrix}, \quad \Delta_c\beta := \partial_s(c\partial_s\beta) + \partial^2_u\beta, \quad \nabla_c\beta \cdot \nabla_c\psi = c\partial_s\beta\partial_s\overline{\psi} + \partial_u\beta\partial_u\overline{\psi}.$$

Then the following result holds:

Theorem 3. *Let H_γ, M, M_1, M_2 be the operators defined as above. Assume that Assumptions 1 and 2 are satisfied. Then, there exist $\Lambda_0 > 0$, $S_0 > 0$ and a positive constant C depending on T such that, for any $\lambda > \Lambda_0$ and any $S > S_0$,*

$$S\lambda\int_{\Omega_{R,\epsilon}\times(-T,T)}\phi|\nabla q|^2e^{-2S\eta}+S^3\lambda^4\int_{\Omega_{R,\epsilon}\times(-T,T)}\phi^3|q|^2e^{-2S\eta}+\int_{\Omega_{R,\epsilon}\times(-T,T)}|M_1(e^{-S\eta}q)|^2$$
$$+\int_{\Omega_{R,\epsilon}\times(-T,T)}|M_2(e^{-S\eta}q)|^2\leq C\int_{\Omega_{R,\epsilon}\times(-T,T)}|H_\gamma q|^2e^{-2S\eta}+CS\lambda\int_{\Gamma_{R,\epsilon}\times(-T,T)}\phi|\partial_\nu q|^2e^{-2S\eta}$$

for all q satisfying $H_\gamma q\in L^2(\Omega_{R,\epsilon}\times(-T,T))$, $q\in L^2(-T,T;H^1_0(\Omega_{R,\epsilon}))$, $\partial_\nu q=\nabla q\cdot\nu$, and $\partial_\nu q\in L^2(-T,T;L^2(\Gamma_{R,\epsilon}))$.

Proof. We proceed as in [1], [5] or [6]. We have:

$$\int_{\Omega_{R,\epsilon}\times(-T,T)}|M\psi-V_\gamma\psi|^2=\int_{\Omega_{R,\epsilon}\times(-T,T)}(|M_1\psi|^2+|M_2\psi|^2)+2\operatorname{Re}\int_{\Omega_{R,\epsilon}\times(-T,T)}M_1\psi\overline{M_2\psi}. \tag{8}$$

Multiplying each term of $M_1\psi$ by each term of $\overline{M_2\psi}$ (see (6) and (7)), we will calculate under the following form:

$$\operatorname{Re}\int_{\Omega_{R,\epsilon}\times(-T,T)}M_1\psi\overline{M_2\psi}=I_{11}+I_{12}+I_{13}+I_{21}+I_{22}+I_{23}+I_{31}+I_{32}+I_{33}. \tag{9}$$

We denote by $Q:=\Omega_{R,\epsilon}\times(-T,T)$. We obtain by integrating by parts:

$$I_{11}=\operatorname{Re}\int_Q(-i\partial_t\psi)\overline{(-iS\,\partial_t\eta\psi)}=-\frac{S}{2}\int_Q\partial_t^2\eta\,|\psi|^2. \tag{10}$$

Since $I_{12}=\operatorname{Re}\int_Q(-i\partial_t\psi)2S\lambda\phi\nabla_c\beta\cdot\nabla_c\psi=S\lambda\operatorname{Im}\int_Q\phi\partial_t\psi\nabla_c\beta\cdot\nabla_c\psi-S\lambda\operatorname{Im}\int_Q\phi\partial_t\overline{\psi}\nabla_c\beta\cdot\nabla_c\overline{\psi}$, integrating by parts in time for the first term and in space for the second term, we get

$$I_{12}=S\lambda^2\operatorname{Im}\int_Q\phi\partial_t\overline{\psi}\psi|\nabla_c\beta|^2+S\lambda\operatorname{Im}\int_Q\phi\partial_t\overline{\psi}\psi\Delta_c\beta-S\lambda\operatorname{Im}\int_Q\partial_t\phi\psi\nabla_c\beta\cdot\nabla_c\psi. \tag{11}$$

Moreover, $I_{13}=\operatorname{Re}\int_Q(-i\partial_t\psi)[S\lambda^2\phi\overline{\psi}|\nabla_c\beta|^2+S\lambda\phi\overline{\psi}\Delta_c\beta]$ becomes

$$I_{13}=-S\lambda\operatorname{Im}\int_Q\phi\partial_t\overline{\psi}\psi\Delta_c\beta-S\lambda^2\operatorname{Im}\int_Q\phi\partial_t\overline{\psi}\psi|\nabla_c\beta|^2 \tag{12}$$

and integrating by parts in space we have

$$I_{21}=\operatorname{Re}\int_Q(-\Delta_c\psi)\overline{(-iS\,\partial_t\eta\psi)}=-S\lambda\operatorname{Im}\int_Q\partial_t\phi\psi\nabla_c\beta\cdot\nabla_c\psi. \tag{13}$$

So from (11)–(13) note that $I_{12}+I_{13}+I_{21}=-2S\lambda\operatorname{Im}\int_Q\partial_t\phi\psi\nabla_c\beta\cdot\nabla_c\psi$. Furthermore, $I_{22}=\operatorname{Re}\int_Q(-\Delta_c\psi)2S\lambda\phi\nabla_c\beta\cdot\nabla_c\psi$. By integrating by parts twice in space we obtain that

$$I_{22}=2S\lambda^2\int_Q\phi|\nabla_c\beta\cdot\nabla_c\overline{\psi}|^2+2S\lambda\operatorname{Re}\int_Q\phi D_c^2\beta(\nabla_c\psi,\nabla_c\overline{\psi})$$

$$\begin{aligned} & - S\lambda \int_Q \phi \frac{1}{c}\nabla_c c \cdot \nabla_c\beta |\sqrt{c}\partial_s\psi|^2 - S\lambda \int_{\partial\Omega_R\times(-T,T)} \phi\nabla_c\beta\cdot\nu_c|\nabla_c\psi|^2 \\ & - S\lambda^2 \int_Q \phi|\nabla_c\beta|^2|\nabla_c\psi|^2 - S\lambda\int_Q \phi\Delta_c\beta|\nabla_c\psi|^2, \end{aligned} \tag{14}$$

with $D_c^2\beta$ defined by (4). We have also $I_{23} = \mathrm{Re}\int_Q(-\Delta_c\psi)[S\lambda^2\phi\overline{\psi}|\nabla_c\beta|^2 + S\lambda\phi\overline{\psi}\Delta_c\beta]$ and, by integrations by parts twice in space, we obtain:

$$\begin{aligned} I_{23} = {} & S\lambda\int_Q \phi|\nabla_c\psi|^2\Delta_c\beta - S\lambda^3\int_Q|\psi|^2\phi|\nabla_c\beta|^2\Delta_c\beta - S\lambda^2\int_Q|\psi|^2\phi\nabla_c\beta\cdot\nabla_c(\Delta_c\beta) \\ & - \frac{S\lambda^2}{2}\int_Q|\psi|^2\phi\Delta_c(|\nabla_c\beta|^2) - \frac{S\lambda^2}{2}\int_Q|\psi|^2\phi(\Delta_c\beta)^2 - \frac{S\lambda}{2}\int_Q|\psi|^2\phi\Delta_c(\Delta_c\beta) \\ & - \frac{S\lambda^4}{2}\int_Q|\psi|^2\phi|\nabla_c\beta|^4 - S\lambda^3\int_Q|\psi|^2\phi\nabla_c\beta\cdot\nabla_c(|\nabla_c\beta|^2) + S\lambda^2\int_Q\phi|\nabla_c\beta|^2|\nabla_c\psi|^2. \end{aligned} \tag{15}$$

And we obviously have

$$I_{31} = \mathrm{Re}\int_Q(-S^2\lambda^2\phi^2\psi|\nabla_c\beta|^2)\overline{(-iS\partial_t\eta\psi)} = 0. \tag{16}$$

Moreover

$$\begin{aligned} I_{32} & = \mathrm{Re}\int_Q(-S^2\lambda^2\phi^2\psi|\nabla_c\beta|^2)2S\lambda\phi\nabla_c\beta\cdot\nabla_c\psi \\ & = S^3\lambda^3\int_Q\phi^3|\psi|^2(\nabla_c\beta\cdot\nabla_c(|\nabla_c\beta|^2) + |\nabla_c\beta|^2\Delta_c\beta) + 3S^3\lambda^4\int_Q\phi^3|\nabla_c\beta|^4|\psi|^2, \end{aligned} \tag{17}$$

$$\begin{aligned} I_{33} & = \mathrm{Re}\int_Q(-S^2\lambda^2\phi^2\psi|\nabla_c\beta|^2)[S\lambda^2\phi\overline{\psi}|\nabla_c\beta|^2 + S\lambda\phi\overline{\psi}\Delta_c\beta] \\ & = -S^3\lambda^3\int_Q\phi^3|\psi|^2|\nabla_c\beta|^2\Delta_c\beta - S^3\lambda^4\int_Q\phi^3|\nabla_c\beta|^4|\psi|^2. \end{aligned} \tag{18}$$

Therefore, from (10) to (18), (9) becomes

$$\begin{aligned} \mathrm{Re}\int_Q M_1\psi\overline{M_2\psi} = {} & -\frac{S}{2}\int_Q\partial_t^2\eta|\psi|^2 - S\lambda\int_Q\phi\frac{1}{c}\nabla_c c\cdot\nabla_c\beta|\sqrt{c}\partial_s\psi|^2 \\ & - 2S\lambda\,\mathrm{Im}\int_Q\partial_t\phi\psi\nabla_c\beta\cdot\nabla_c\psi + 2S\lambda^2\int_Q\phi|\nabla_c\beta\cdot\nabla_c\psi|^2 \\ & + 2S\lambda\,\mathrm{Re}\int_Q\phi D_c^2\beta(\nabla_c\overline{\psi},\nabla_c\psi) - S\lambda\int_{\partial\Omega_{R,\epsilon}\times(-T,T)}\phi|\nabla_c\psi|^2\nabla_c\beta\cdot\nu_c \\ & - S\lambda^3\int_Q\phi|\psi|^2|\nabla_c\beta|^2\Delta_c\beta - S\lambda^2\int_Q\phi|\psi|^2\nabla_c\beta\cdot\nabla_c(\Delta_c\beta) \\ & - \frac{S\lambda^2}{2}\int_Q\phi|\psi|^2(\Delta_c\beta)^2 - S\lambda^3\int_Q\phi|\psi|^2\nabla_c\beta\cdot\nabla_c(|\nabla_c\beta|^2) \\ & - \frac{S\lambda^2}{2}\int_Q\phi|\psi|^2\Delta_c(|\nabla_c\beta|^2) - \frac{S\lambda}{2}\int_Q\phi|\psi|^2\Delta_c(\Delta_c\beta) \end{aligned} \tag{19}$$

$$-\frac{S\lambda^4}{2}\int_Q \phi|\psi|^2|\nabla_c\beta|^4 + 2S^3\lambda^4\int_Q \phi^3|\psi|^2|\nabla_c\beta|^4 + S^3\lambda^3\int_Q \phi^3|\psi|^2\nabla_c\beta\cdot\nabla_c(|\nabla_c\beta|^2). \tag{20}$$

Now, if we call by X the terms in (19) which are neglectable with respect to the quatities $S\lambda^2\int_Q \phi|\nabla_c\beta\cdot\nabla_c\psi|^2$ or $S^3\lambda^4\int_Q \phi^3|\psi|^2|\nabla_c\beta|^4$, we get:

$$\begin{aligned} X = &-\frac{S}{2}\int_Q \partial_t^2\eta|\psi|^2 - 2S\lambda\,\mathrm{Im}\int_Q \partial_t\phi\psi\nabla_c\beta\cdot\nabla_c\psi - S\lambda^3\int_Q \phi|\psi|^2|\nabla_c\beta|^2\Delta_c\beta \\ &- S\lambda^2\int_Q \phi|\psi|^2\nabla_c\beta\cdot\nabla_c(\Delta_c\beta) - \frac{S\lambda^2}{2}\int_Q \phi|\psi|^2(\Delta_c\beta)^2 - \frac{S\lambda}{2}\int_Q \phi|\psi|^2\Delta_c(\Delta_c\beta) \\ &- \frac{S\lambda^4}{2}\int_Q \phi|\psi|^2|\nabla_c\beta|^4 - S\lambda^3\int_Q \phi|\psi|^2\nabla_c\beta\cdot\nabla_c(|\nabla_c\beta|^2) - \frac{S\lambda^2}{2}\int_Q \phi|\psi|^2\Delta_c(|\nabla_c\beta|^2) \\ &+ S^3\lambda^3\int_Q \phi^3|\psi|^2\nabla_c\beta\cdot\nabla_c(|\nabla_c\beta|^2). \end{aligned}$$

So (19) becomes

$$\begin{aligned} \mathrm{Re}\int_Q M_1\psi\overline{M_2\psi} = X &+ 2S\lambda^2\int_Q \phi|\nabla_c\beta\cdot\nabla_c\psi|^2 + 2S\lambda\,\mathrm{Re}\int_Q \phi D_c^2\beta(\nabla_c\overline{\psi},\nabla_c\psi) \\ &- S\lambda\int_Q \phi\frac{1}{c}\nabla_c c\cdot\nabla_c\beta|\sqrt{c}\partial_s\psi|^2 - S\lambda\int_{\partial\Omega_{R,\epsilon}\times(-T,T)} \phi|\nabla_c\psi|^2\nabla_c\beta\cdot\nu_c \\ &+ 2S^3\lambda^4\int_Q \phi^3|\psi|^2|\nabla_c\beta|^4 \end{aligned}$$

and there exists a positive constant k such that

$$|X| \le kS\lambda^4\int_Q \phi|\psi|^2 + kS^3\lambda^3\int_Q \phi^3|\psi|^2 + kS\lambda\int_Q \phi|\nabla_c\beta\cdot\nabla_c\psi|^2. \tag{21}$$

Moreover, from (8), (21) and Assumption 2, we get

$$\begin{aligned} &\int_Q [|M_1\psi|^2 + |M_2\psi|^2] + 4S\lambda^2\int_Q \phi|\nabla_c\beta\cdot\nabla_c\psi|^2 - 2S\lambda\int_Q \phi\frac{1}{c}\nabla_c c\cdot\nabla_c\beta|\sqrt{c}\partial_s\psi|^2 \\ &\quad + 4S\lambda\,\mathrm{Re}\int_Q \phi D_c^2\beta(\nabla_c\overline{\psi},\nabla_c\psi) + 4S^3\lambda^4\beta_0^4\int_Q \phi^3|\psi|^2 \\ &\le Cs\lambda\int_{\partial\Omega_{R,\epsilon}\times(-T,T)} \phi|\nabla_c\psi|^2\nabla_c\beta\cdot\nu_c + CS\lambda^4\int_Q \phi|\psi|^2 + CS^3\lambda^3\int_Q \phi^3|\psi|^2 \\ &\quad + CS\lambda\int_Q \phi|\nabla_c\beta\cdot\nabla_c\psi|^2 + C\int_Q |M\psi|^2 + C\int_Q V_\gamma^2|\psi|^2. \end{aligned} \tag{22}$$

Since V_γ is bounded on $\Omega_{R,\epsilon}$ and since ϕ is a positive continuous function there exists a positive constant depending upon T such that $V_\gamma \le cst\,\phi^3$. Choosing such S and λ sufficiently large, we deduce that there exists a positive constant C_1 such that (22) becomes

$$\begin{aligned}\int_Q [|M_1\psi|^2 + |M_2\psi|^2] + S\lambda^2 \int_Q \phi|\nabla_c\beta \cdot \nabla_c\psi|^2 \\ + S\lambda \operatorname{Re} \int_Q \phi D_c^2\beta(\nabla_c\overline{\psi}, \nabla_c\psi) + S^3\lambda^4 \int_Q \phi^3|\psi|^2 - S\lambda \int_Q \phi\frac{1}{c}\nabla_c c \cdot \nabla_c\beta|\sqrt{c}\partial_s\psi|^2 \\ \leq C_1 S\lambda \int_{\Gamma_{R,\epsilon}\times(-T,T)} \phi|\nabla_c\psi|^2\nabla_c\beta \cdot \nu_c + C_1 \int_Q |M\psi|^2.\end{aligned}$$

Finally, we come back to $q = e^{S\eta}\psi$. And this concludes the proof. □

§3. Inverse problem

First, using an idea developed in [10], we prove the following lemma:

Lemma 4. *Let z_0 be a real function in $C^2(\overline{\Omega_{R,\epsilon}})$ and define the following first order differential operator $P_0 g := \partial_s z_0 \partial_s g$. Let η_0 be a real function in $C^2(\overline{\Omega_{R,\epsilon}})$. Assume that for all $(s,u) \in \overline{\Omega_{R,\epsilon}}$, $(\partial_s z_0 \partial_s \eta_0)^2 \geq cst > 0$. Then there exists a positive constant C such that for S sufficiently large*

$$S^2 \int_{\Omega_{R,\epsilon}} e^{-2S\eta_0}|g|^2 \leq C \int_{\Omega_{R,\epsilon}} |P_0 g|^2 e^{-2S\eta_0} + CS \int_{\Gamma_\epsilon} e^{-2S\eta_0}|g|^2|\partial_s\eta_0\nu_s|$$

for any $g \in H^1(\Omega_{R,\epsilon})$.

Proof. Let $g \in H^1(\Omega_{R,\epsilon})$. Define $w = e^{-S\eta_0}g$ and $Q_0 w := e^{-S\eta_0}P_0(e^{S\eta_0}w)$. If we set $q_0 = \partial_s z_0 \partial_s \eta_0$, then we get $Q_0 w = Sq_0 w + P_0 w$. Therefore we have:

$$\begin{aligned}\int_{\Omega_{R,\epsilon}} |Q_0 w|^2 = \int_{\Omega_{R,\epsilon}} |P_0 g|^2 e^{-2S\eta_0} &= S^2 \int_{\Omega_{R,\epsilon}} q_0^2|w|^2 + \int_{\Omega_{R,\epsilon}} |P_0 w|^2 + 2S \operatorname{Re} \int_{\Omega_{R,\epsilon}} q_0 w\overline{P_0 w} \\ &\geq S^2 \int_{\Omega_{R,\epsilon}} q_0^2|w|^2 + S \int_{\Omega_{R,\epsilon}} q_0\partial_s z_0 \partial_s(|w|^2)\end{aligned}$$

and so, integrating by parts, since $\nu_s = 0$ on $\Sigma_{R,\epsilon}^+ \cup \Sigma_{R,\epsilon}^-$, we get

$$\begin{aligned}&\int_{\Omega_{R,\epsilon}} |P_0 g|^2 e^{-2S\eta_0} \\ &\quad\geq S^2 \int_{\Omega_{R,\epsilon}} q_0^2 e^{-2S\eta_0}|g|^2 + S(-\int_{\Omega_{R,\epsilon}} \partial_s(q_0\partial_s z_0)e^{-2S\eta_0}|g|^2 + \int_{\Gamma_\epsilon} e^{-2S\eta_0}|g|^2 q_0\partial_s z_0\nu_s).\end{aligned}$$

Since $\partial_s(q_0\partial_s z_0)$ is a bounded function in $\overline{\Omega_{R,\epsilon}}$ and $q_0\partial_s z_0\nu_s = (\partial_s z_0)^2\partial_s\eta_0\nu_s$, we can conclude. □

Then, we consider γ and $\widetilde{\gamma}$ two functions satisfying Assumption 1.

Let z be a solution of

$$\begin{cases} -i\partial_t z(s,u,t) - \partial_s(c_\gamma(s,u)\partial_s z(s,u,t)) - \partial_u^2 z(s,u,t) + V_\gamma(s,u)z(s,u,t) = 0, \\ (s,u,t) \in \Omega_{R,\epsilon} \times (0,T), \\ z(s,u,t) = l(s,u,t),\ (s,u) \in \partial\Omega_{R,\epsilon},\ t \in (0,T), \\ z(s,u,0) = z_0(s,u),\ (s,u) \in \Omega_{R,\epsilon}, \end{cases} \tag{23}$$

and let $\widetilde{z}$ be a solution of

$$\begin{cases} -i\partial_t \widetilde{z}(s,u,t) - \partial_x(c_{\widetilde{\gamma}}(s,u)\partial_s \widetilde{z}(s,u,t)) - \partial_u^2 \widetilde{z}(s,u,t) + V_{\widetilde{\gamma}}(s,u)\widetilde{z}(s,u,t) = 0, \\ (s,u,t) \in \Omega_{R,\epsilon} \times (0,T), \\ \widetilde{z}(s,u,t) = l(s,u,t),\ (s,u) \in \partial\Omega_{R,\epsilon},\ t \in (0,T), \\ \widetilde{z}(s,u,0) = z_0(s,u),\ (s,u) \in \Omega_{R,\epsilon}. \end{cases} \tag{24}$$

Let $\Lambda_N := \{f \in C^1([R_1,R_2]),\ |f'(s)| \le N|f(s)|$ and $|f(s| \le N$ for all $s \in [R_1,R_2]\}$ with N a positive real given. We obtain the following theorem:

Theorem 5. *Let γ and $\widetilde{\gamma}$ be functions both satisfying Assumption 1 and such that $(\gamma-\widetilde{\gamma})(s) \neq 0$ and $(\gamma' - \widetilde{\gamma}')(s) \neq 0$ for all $s \in [R_1,R_2]$. Assume that β is a function which satisfies Assumption 2 w.r.t. c_γ with c_γ defined by* (2). *Assume also that*

(i) *z_0 is a real function such that $z_0 \in C^2(\overline{\Omega_{R,\epsilon}})$.*

(ii) *For all $(s,u) \in \overline{\Omega_{R,\epsilon}}$, $(\partial_s z_0(s,u)\partial_s\eta(s,u,0))^2 \ge cst > 0$ (where η is defined by* (5)*).*

(iii) *$\partial_t\widetilde{z} \in L^\infty(\Omega_{R,\epsilon} \times (0,T))$, $\partial_s(\partial_t\widetilde{z}) \in L^\infty(\Omega_{R,\epsilon} \times (0,T))$, $\partial_s^2(\partial_t\widetilde{z}) \in L^\infty(\Omega_{R,\epsilon} \times (0,T))$, $\partial_\nu(\partial_t(z-\widetilde{z})) \in L^\infty(\Gamma_\epsilon \times (0,T))$ and the L^∞-norm of each of these functions is less than N.*

(iv) *$\gamma - \widetilde{\gamma} \in \Lambda_N$ and $\gamma' - \widetilde{\gamma}' \in \Lambda_N$.*

Then there exists a positive constant C, depending upon N, T, $\|\beta\|_{L^\infty}$, $\|\partial_s\beta\|_{L^\infty}$ such that, for S and λ sufficiently large, we have:

$$\int_{L^2(\Omega_{R,\epsilon})} e^{-2S\eta_0}|\gamma(s) - \widetilde{\gamma}(s)|^2\, dsdu \le C \int_{\Sigma^+_{R,\epsilon}\times(-T,T)} \phi e^{-2S\eta}|\partial_\nu(\partial_t(z-\widetilde{z}))|^2 + C\epsilon. \tag{25}$$

Note that $\partial_s z_0 \partial_s\eta := -\lambda\partial_s z_0(e^{\lambda\beta}/T^2)\,\partial_s\beta$ satisfies the above hypothesis (ii) for any function z_0 such that $\partial_s z_0$ is a continuous and non null function in $\overline{\Omega_{R,\epsilon}}$ (by assuming also that $\partial_s\beta$ is a non null function in $\overline{\Omega_{R,\epsilon}}$, which is true for $\beta(s,u) = e^s + e^u$ for example). Note that since $\gamma - \widetilde{\gamma}$ is assumed satisfying $(\gamma - \widetilde{\gamma})(s) \neq 0$ and $(\gamma' - \widetilde{\gamma}')(s) \neq 0$ for all $s \in [R_1,R_2]$, then $\frac{\gamma'-\widetilde{\gamma}'}{\gamma-\widetilde{\gamma}}$ and $\frac{\gamma''-\widetilde{\gamma}''}{\gamma'-\widetilde{\gamma}'}$ are bounded functions in $[R_1,R_2]$ and therefore the previous hypothesis iv) is verified for some N. Note also that the above hypothesis (iii) is satisfied for any function $\widetilde{z} \in C^3(\Omega_{R,\epsilon} \times (0,T))$.

Proof. Now, recall that z (resp. $\widetilde{z}$) is a solution of (23) (resp. (24)). If we set $w = z - \widetilde{z}$, $v = \partial_t w$, $g = c_\gamma - c_{\widetilde{\gamma}}$ and $h = V_\gamma - V_{\widetilde{\gamma}}$, we get

$$\begin{cases} -i\partial_t w - \partial_s(c_\gamma\partial_s w) - \partial_u^2 w + V_\gamma w = \partial_s(g\partial_s\widetilde{z}) - h\widetilde{z} \text{ in } \Omega_{R,\epsilon} \times (0,T), \\ w = 0 \text{ on } \partial\Omega_{R,\epsilon} \times (0,T), \\ w(s,u,0) = 0,\ (s,u) \in \Omega_{R,\epsilon}, \end{cases}$$

$$\begin{cases} -i\partial_t v - \partial_s(c_\gamma \partial_s v) - \partial_u^2 v + V_\gamma v = \partial_s(g\partial_s(\partial_t \widetilde{z})) - h\partial_t\widetilde{z} \text{ in } \Omega_{R,\epsilon} \times (0,T), \\ v = 0 \text{ on } \partial\Omega_{R,\epsilon} \times (0,T), \\ v(s,u,0) = i(\partial_s(g(s,u)\partial_s z_0(s,u)) - hz_0(s,u)), \ (s,u) \in \Omega_{R,\epsilon}. \end{cases}$$

As in [1] or [5], we extend the function v on $\Omega_{R,\epsilon}\times(-T,T)$ by the formula $v(s,u,t) = \overline{v}(s,u,-t)$ for every $(s,u,t) \in \Omega_{R,\epsilon} \times (-T,0)$. Note that this extension is available if the initial data is a real valued function. Note also that this extension satisfies the previous Carleman estimate. We set $\psi = e^{-S\eta} v$ with η defined by (5). We recall that $M_1\psi = -i\partial_t\psi - \Delta_c\psi - S^2\lambda^2\phi^2\psi|\nabla_c\beta|^2$ with $c = c_\gamma$.

In a first step, we define $I := \text{Im} \int_{\Omega_{R,\epsilon}\times(-T,0)} M_1\psi\overline{\psi}$. Then by integrations by parts, we obtain: $I = (-1/2)\int_{\Omega_{R,\epsilon}} |\psi(s,u,0)|^2 ds\, du$. If we denote by $\eta_0(s,u) := \eta(s,u,0)$ and by $\phi_0(s,u) := \phi(s,u,0)$, recalling that $\psi = e^{-S\eta}v = e^{-S\eta}\partial_t w$, we get:

$$I = -\frac{1}{2}\int_{\Omega_{R,\epsilon}} e^{-2S\eta_0(s,u)}|\partial_t w(s,u,0)|^2\, ds\, du. \tag{26}$$

Moreover, we have:

$$\begin{aligned} |I| &\leq S^{-3/4}\lambda^{-1}\left(\int_Q |M_1\psi|^2\right)^{1/2} S^{3/4}\lambda\left(\int_Q |\psi|^2\right)^{1/2} \\ &\leq \frac{S^{-3/2}\lambda^{-2}}{2}\left(\int_Q |M_1(e^{-S\eta}v)|^2 + S^3\lambda^4\int_Q e^{-2S\eta}|v|^2\right). \end{aligned}$$

Since $H_\gamma v = \partial_s(g\partial_s(\partial_t\widetilde{z})) - h\partial_t\widetilde{z}$, applying the Carleman inequality, we get:

$$|I| \leq CS^{-3/2}\lambda^{-2}\int_Q e^{-2S\eta}|\partial_s(g\partial_s(\partial_t\widetilde{z})) - h\partial_t\widetilde{z}|^2 + CS^{-1/2}\lambda^{-1}\int_{\Gamma_{R,\epsilon}\times(-T,T)} \phi e^{-2S\eta}|\partial_\nu v|^2,$$

with C a positive constant. Since $\partial_t\widetilde{z} \in L^\infty(\Omega_{R,\epsilon}\times(0,T))$, $\partial_s(\partial_t\widetilde{z}) \in L^\infty(\Omega_{R,\epsilon}\times(0,T))$, $\partial_s^2(\partial_t\widetilde{z}) \in L^\infty(\Omega_{R,\epsilon}\times(0,T))$ and $e^{-2S\eta(s,u,t)} \leq e^{-2S\eta(s,u,0)}$ we have:

$$|I| \leq CS^{-3/2}\lambda^{-2}\int_Q e^{-2S\eta_0}[|\partial_s g|^2 + |g|^2 + |h|^2] + CS^{-1/2}\lambda^{-1}\int_{\Gamma_{R,\epsilon}\times(-T,T)} \phi e^{-2S\eta}|\partial_\nu v|^2, \tag{27}$$

with C a positive constant depending on T. Moreover, from $-i\partial_t w(s,u,0) = \partial_s(g\partial_s z_0) - hz_0 = \partial_s g\partial_s z_0 + g\partial_s^2 z_0 - hz_0$, applying the Lemma 2 for the function $g = c_\gamma - c_{\widetilde{\gamma}}$ and $P_0 g = \partial_s z_0 \partial_s g = -i\partial_t w(s,u,0) - g\partial_s^2 z_0 + hz_0$, we obtain:

$$S^2\int_{\Omega_{R,\epsilon}} e^{-2S\eta_0}|g|^2 \leq C\int_{\Omega_{R,\epsilon}} |-i\partial_t w(s,u,0) - g\partial_s^2 z_0 + hz_0|^2 e^{-2S\eta_0} + CS\int_{\Gamma_\epsilon} e^{-2S\eta_0}|g|^2|\partial_s\eta_0\nu_s|.$$

And so

$$S^2\int_{\Omega_{R,\epsilon}} e^{-2S\eta_0}|g|^2 \leq C\int_{\Omega_{R,\epsilon}} [|\partial_t w(s,u,0)|^2 + |g|^2 + |h|^2]e^{-2S\eta_0} + CS\lambda\int_{\Gamma_\epsilon} e^{-2S\eta_0}|g|^2|\partial_s\beta\phi_0\nu_s|. \tag{28}$$

From (26)–(28) we get:

$$S^2 \int_{\Omega_{R,\epsilon}} e^{-2S\eta_0}|g|^2 \le CS^{-3/2}\lambda^{-2} \int_{\Omega_{R,\epsilon}} [|\partial_s g|^2 + |g|^2 + |h|^2]e^{-2S\eta_0} + CS\lambda \int_{\Gamma_\epsilon} e^{-2S\eta_0}|g|^2|\partial_s\beta\phi_0\nu_s|$$
$$+ CS^{-1/2}\lambda^{-1} \int_{\Gamma_{R,\epsilon}\times(-T,T)} \phi e^{-2S\eta}|\partial_\nu v|^2 + C\int_{\Omega_{R,\epsilon}} [|g|^2 + |h|^2]e^{-2S\eta_0}. \quad (29)$$

Finally note that

$$0 < cst|\gamma - \widetilde{\gamma}| \le |g| \le cst|\gamma - \widetilde{\gamma}|,\ |\partial_s g| \le cst[|\gamma - \widetilde{\gamma}| + |\gamma' - \widetilde{\gamma}'|],$$
$$|h| \le cst[|\gamma - \widetilde{\gamma}| + |\gamma' - \widetilde{\gamma}'| + |\gamma'' - \widetilde{\gamma}''|]. \quad (30)$$

Combining (29) and (30) we can conclude for S sufficiently large. □

Remark 1. Such result (25) can be generalized on the whole space Ω_1 ($\int_{\Omega_1} e^{-2S\eta_0}|\gamma - \widetilde{\gamma}|^2 \le C\int_{\Sigma_R^+\times(-T,T)} \phi e^{-2S\eta}|\partial_\nu(\partial_t(z - \widetilde{z}))|^2$) under the condition that there exists a function $\widetilde{\beta}$ which satisfies Assumption 2 on the whole Ω_1.

Acknowledgements

The idea of studying an inverse problem for a curved quantum guide was from Professor Pierre Duclos of the University of Toulon, France and I would like to dedicate this work to his memory.

References

[1] Baudouin, L., and Puel, J.-P. Uniqueness and stability in an inverse problem for the Schrödinger equation. *Inverse Problems 18* (2002), 1537–1554.

[2] Bukhgeim, A. L. *Volterra Equations and Inverse Problems*. Inverse and Ill-Posed Problems Series. Vsp, Utrecht, 1999.

[3] Bukhgeim, A. L., and Klibanov, M. V. Uniqueness in the large of a class of multidimensional inverse problems. *Soviet. Math. Dokl. 17* (2006), 244–247.

[4] Cardoulis, L., and Cristofol, M. Inverse problem for a curved quantum guide. Submitted.

[5] Cardoulis, L., Cristofol, M., and Gaitan, P. Inverse problem for the Schrödinger operator in an unbounded strip. *Journal of Inverse and Ill-Posed Problem 16*, 2 (2008), 127–6146.

[6] Cardoulis, L., and Gaitan, P. Simultaneous identification of the diffusion coefficient and the potential for the Schrödinger operator with one observation. *Inverse Problems 26* (2010), 035012.

[7] Chenaud, B., Duclos, P., Freitas, and Krejcirik, D. Geometrically induced discrete spectrum in curved tubes. *Diff. Geom. Appl. 23*, 2 (2005), 95–6105.

[8] Exner, P., and Seba, P. Bound states in curved quantum waveguides. *J. Math. Phys. 30*, 11 (1989), 2574–2580. Math. Meth. Appl. Sci. 27, (2004), 1–17.

[9] Goldstone, J., and Jaffe, R. L. Bound states in twisting tubes. *Phys. Rev. B 45* (1992), 14100–14107.

[10] Immanuvilov, O. Y., and Yamamoto, M. Carleman estimates for the non-stationary Lamé system with two sets of boundary data. *Cpam LVI* (2003), 1366–1382.

[11] Klibanov, M. V. Inverse problems in the large and Carleman bounds. *Differential Equations 20* (1984), 755–760.

[12] Klibanov, M. V. Inverse problems and Carleman estimates. *Inverse Problems 8* (1992), 575–596.

[13] Klibanov, M. V., and Timonov, A. *Carleman Estimates for Coefficient Inverse Problems and Numerical Applications*. Vsp, Utrecht, 2004.

[14] Klibanov, M. V., and Yamamoto, M. Lipschitz stability for an inverse problem for an acoustic equation. *Applicable Analysis 85* (2006), 515–538.

Laure Cardoulis
Université de Toulouse, UT1 Ceremath , CNRS, Intitut de Mathématiques de Toulouse, UMR 5219
21 Allées de Brienne, 31042 Toulouse, France
laure.cardoulis@univ-tlse1.fr

Monografías Matemáticas *García de Galdeano* **37**, 91–101 (2012)

Detecting an obstacle immersed in a fluid: the Stokes case

Fabien Caubet

Abstract. This paper presents a theoretical study of a detection of an object immersed in a fluid. The fluid motion is governed by the Stokes equations. We detail the Dirichlet case for which the results are just stated in [3]. We make a shape sensitivity analysis of order two in order to prove the existence of the first and the second orders shape derivatives. The strategy adopted to detect the object is to minimize a least-squares functional. We characterize the gradient of the functional using an adjoint problem. Finally, we study the stability of this setting. We give the expression of the shape Hessian at a critical point and the compactness of the Riesz operator corresponding to this shape Hessian is shown. The ill-posedness of the identification problem follows which explains the need of regularization to numerically solve this problem.

Keywords: Stationary Stokes problem, sensitivity with respect to the domain of order two, geometric inverse problem.

AMS classification: 35R30, 35Q30, 49Q10, 49Q12, (76D07).

§1. Introduction

Notations and references on the Stokes equations. For a domain Ω, $\langle\cdot\,,\,\cdot\rangle_\Omega$ and $\langle\cdot\,,\,\cdot\rangle_{\partial\Omega}$ will denote respectively the duality products $\langle\cdot\,,\,\cdot\rangle_{\mathbf{H}^{-1}(\Omega),\mathbf{H}^1_0(\Omega)}$ and $\langle\cdot\,,\,\cdot\rangle_{\mathbf{H}^{-1/2}(\partial\Omega),\mathbf{H}^{1/2}(\partial\Omega)}$. Moreover, $\mathbf{n}$ represents the external unit normal to $\partial\Omega$.

In this paper, we use some existence, uniqueness and regularity results concerning the Stokes equations: we refer for example to [9, Chapter 1]. Moreover, we also use some local regularity arguments: see [5, Theorem IV.5.1] for details.

Setting of the problem. Let Ω a bounded, connected open subset of $\mathbb{R}^N$ (with $N = 2$ or $N = 3$) with a $C^{1,1}$ boundary. Let $\delta > 0$ fixed (small). We define O_δ the set of all open subsets ω of Ω with a $C^{2,1}$ boundary such that $\mathrm{d}(x, \partial\Omega) > \delta$ for all $x \in \omega$ and such that $\Omega \setminus \overline{\omega}$ is connected. We also define Ω_δ an open set with a C^∞ boundary such that

$$\{x \in \Omega\,;\ \mathrm{d}(x, \partial\Omega) > \delta/2\} \subset \Omega_\delta \subset \{x \in \Omega\,;\ \mathrm{d}(x, \partial\Omega) > \delta/3\}\,.$$

Let $\boldsymbol{f_b}$ be an admissible boundary measurement. Let $\boldsymbol{g} \in \mathbf{H}^{3/2}(\partial\Omega)$ such that $\boldsymbol{g} \neq \mathbf{0}$ and satisfying the following condition:

$$\int_{\partial\Omega} \boldsymbol{g} \cdot \mathbf{n} = 0. \tag{1}$$

Let us consider, for $\omega \in \mathcal{O}_\delta$, the following overdetermined Stokes boundary values problem:

$$\begin{cases} -\operatorname{div}(\sigma(\boldsymbol{u},p)) = \mathbf{0} & \text{in } \Omega \setminus \overline{\omega}, \\ \operatorname{div}\boldsymbol{u} = 0 & \text{in } \Omega \setminus \overline{\omega}, \\ \boldsymbol{u} = \boldsymbol{g} & \text{on } \partial\Omega, \\ \boldsymbol{u} = \mathbf{0} & \text{on } \partial\omega, \\ \sigma(\boldsymbol{u},p)\mathbf{n} = \boldsymbol{f_b} & \text{on } \partial\Omega, \end{cases} \tag{2}$$

where $\sigma(\boldsymbol{u},p) = \nu(\nabla\boldsymbol{u} + {}^t\nabla\boldsymbol{u}) - p\,\mathrm{I}$ is the stress tensor and $\nu > 0$ is a given constant representing the kinematic viscosity of the liquid.

We assume there exists $\omega \in \mathcal{O}_\delta$ such that (2) has a solution. This means that the measurement $\boldsymbol{f_b}$ is perfect, *i.e.* without error. Thus, we consider the following geometric inverse problem:

$$\textit{find } \omega \in \mathcal{O}_\delta \textit{ and a pair } (\boldsymbol{u},p) \textit{ which satisfies the overdetermined system (2).} \tag{3}$$

To solve this inverse problem, we consider, for $\omega \in \mathcal{O}_\delta$, the least-squares functional

$$J(\omega) = \frac{1}{2}\int_{\partial\Omega} |\sigma(\boldsymbol{u}(\omega), p(\omega))\,\mathbf{n} - \boldsymbol{f_b}|^2,$$

where $(\boldsymbol{u}(\omega), p(\omega)) \in \mathbf{H}^2(\Omega \setminus \overline{\omega}) \times \mathrm{H}^1(\Omega \setminus \overline{\omega})$ is a solution of the Stokes problem

$$\begin{cases} -\operatorname{div}(\sigma(\boldsymbol{u},p)) = \mathbf{0} & \text{in } \Omega \setminus \overline{\omega}, \\ \operatorname{div}\boldsymbol{u} = 0 & \text{in } \Omega \setminus \overline{\omega}, \\ \boldsymbol{u} = \boldsymbol{g} & \text{on } \partial\Omega, \\ \boldsymbol{u} = \mathbf{0} & \text{on } \partial\omega. \end{cases} \tag{4}$$

Since we imposed the compatibility condition (1), problem (4) has a unique solution once a normalization condition on the pressure p is imposed (see for example [9, Chapter 1]). Such a solution $(\boldsymbol{u}, p)$ is called the state of the system. Here, we choose the normalization

$$\int_{\partial\Omega} (\sigma(\boldsymbol{u},p)\mathbf{n}) \cdot \mathbf{n} = \int_{\partial\Omega} \boldsymbol{f_b} \cdot \mathbf{n}. \tag{5}$$

Then, we try to minimize the least-squares criterion J:

$$\omega^* = \operatorname*{arg\,min}_{\omega \in \mathcal{O}_\delta} J(\omega). \tag{6}$$

Indeed, if ω^* is solution of the inverse problem (3), then $J(\omega^*) = 0$ and (6) holds. Conversely, if ω^* solves (6) with $J(\omega^*) = 0$, then this domain ω^* is a solution of the inverse problem.

Introduction of the needed functional tools. Let $\boldsymbol{U} = \{\boldsymbol{\theta} \in \mathbf{W}^{3,\infty}(\mathbb{R}^N);\ \operatorname{supp}\boldsymbol{\theta} \subset \overline{\Omega_\delta}\}$ and $\boldsymbol{\mathcal{U}} = \{\boldsymbol{\theta} \in \boldsymbol{U};\ \|\boldsymbol{\theta}\|_{3,\infty} < 1\}$ be the space of admissible deformations. Notice that if $\boldsymbol{\theta} \in \boldsymbol{\mathcal{U}}$ then $(\mathbf{I} + \boldsymbol{\theta})$ is a diffeomorphism. For such a $\boldsymbol{\theta} \in \boldsymbol{U}$ and $\omega \in \mathcal{O}_\delta$, we check $\Omega = (\mathbf{I} + \boldsymbol{\theta})(\Omega)$ and we define the perturbed domain $\omega_\theta = (\mathbf{I} + \boldsymbol{\theta})(\omega)$ which is so that $\Omega \setminus \overline{\omega_\theta} \in \mathcal{O}_\delta$.

Let $T > 0$, that we will have to fix small. We will use the shape calculus introduced in [7] by F. Murat and J. Simon. Thus, we consider the function

$$\phi : t \in [0, T) \mapsto \mathbf{I} + t\boldsymbol{V} \in \mathbf{W}^{3,\infty}(\mathbb{R}^N),$$

where $\boldsymbol{V} \in \boldsymbol{U}$. Note that for small t, $\phi(t)$ is a diffeomorphism of $\mathbb{R}^N$ and that $\phi'(0) = \boldsymbol{V}$ vanishes on $\partial\Omega$ and even on the tubular neighborhood $\Omega \setminus \overline{\Omega_\delta}$ of $\partial\Omega$. For $t \in [0, T)$, we define $\omega_t = \phi(t)(\omega)$ and $\mathbf{n}_t$ the external unit normal of $\Omega \setminus \overline{\omega_t}$.

Outlines of the paper. This paper is organized as follows. In Section 2, we state the main results of this work. We first mention an identifiability result proved by C. Alvarez *et al.* in [1]. We claim the existence of the first order shape derivative of the state and we characterize this derivative. We then give the expression of the gradient of the least-squares functional introducing an adjoint problem. Furthermore, we discuss higher order shape derivatives and we characterize the shape Hessian at a possible solution of the original inverse problem. Finally, we justify the instability of the problem: the Riesz operator corresponding to the shape Hessian at a critical shape is compact, which means that the functional is degenerate for the high frequencies. In Section 3, we present some preliminary results: we recall an extension of the usual implicit functions Theorem proved by J. Simon in [8] and we prove some results used in section 4 where the main results of this work are proved. In Section 5, we compare the Neumann case exposed in [3] and the Dirichlet case treated in this paper: we point out the difficulties and the mistakes made in the statement of the Dirichlet case in [3].

§2. Statement of the main results

Identifiability result. According to [1, Theorem 1.2] proved by C. Alvarez *et al.*, the inverse problem (3) is well posed, in the sense that the solution (which exists by assumption) is unique. Indeed, this identifiability result claims that given a fixed $\boldsymbol{g}$, two different geometries ω_0 and ω_1 in $\mathcal{O}_\delta$ yield two different measures $\boldsymbol{f_{b1}}$ and $\boldsymbol{f_{b2}}$.

Sensitivity with respect to the domain. Secondly, we aim to make a sensitivity (with respect to the shape) analysis. The Stokes problem on $\Omega \setminus \overline{\omega_t}$

$$\begin{cases} -\operatorname{div}(\sigma(\boldsymbol{u_t}, p_t)) = \mathbf{0} & \text{in } \Omega \setminus \overline{\omega_t}, \\ \operatorname{div} \boldsymbol{u_t} = 0 & \text{in } \Omega \setminus \overline{\omega_t}, \\ \boldsymbol{u_t} = \boldsymbol{g} & \text{on } \partial\Omega, \\ \boldsymbol{u_t} = \mathbf{0} & \text{on } \partial\omega_t, \end{cases} \tag{7}$$

admits a unique solution $(\boldsymbol{u_t}, p_t) \in \mathbf{H}^2(\Omega \setminus \overline{\omega_t}) \times \mathrm{H}^1(\Omega \setminus \overline{\omega_t})$ satisfying the normalization condition $\int_{\partial\Omega} (\sigma(\boldsymbol{u_t}, p_t)\mathbf{n}) \cdot \mathbf{n} = \int_{\partial\Omega} \boldsymbol{f_b} \cdot \mathbf{n}$.

Proposition 1 (First order shape derivatives of the state)**.** *The solution* $(\boldsymbol{u}, p)$ *is differentiable with respect to the domain and the derivatives* $(\boldsymbol{u'}, p') \in \mathbf{H}^2(\Omega \setminus \overline{\omega}) \times \mathrm{H}^1(\Omega \setminus \overline{\omega})$ *is the only* *solution of the following boundary values problem*

$$\begin{cases} -\operatorname{div}(\sigma(\boldsymbol{u}', p')) = \mathbf{0} & \text{in } \Omega \setminus \overline{\omega}, \\ \operatorname{div} \boldsymbol{u}' = 0 & \text{in } \Omega \setminus \overline{\omega}, \\ \boldsymbol{u}' = \mathbf{0} & \text{on } \partial\Omega, \\ \boldsymbol{u}' = -\partial_{\mathbf{n}} \boldsymbol{u} \, (\boldsymbol{V} \cdot \mathbf{n}) & \text{on } \partial\omega, \end{cases} \tag{8}$$

with the normalization condition $\int_{\partial\Omega} (\sigma(\boldsymbol{u}', p')\mathbf{n}) \cdot \mathbf{n} = 0$.

Proposition 2 (First order shape derivatives of the functional)**.** *For* $\boldsymbol{V}$ *in* $\boldsymbol{U}$*, the least-squares functional* J *is differentiable at* ω *in the direction* $\boldsymbol{V}$ *with*

$$\mathrm{D}\, J(\omega) \cdot \boldsymbol{V} = -\int_{\partial\omega} [(\sigma(\boldsymbol{w}, q)\mathbf{n}) \cdot \partial_{\mathbf{n}} \boldsymbol{u}] \, (\boldsymbol{V} \cdot \mathbf{n}),$$

where $(\boldsymbol{w}, q) \in \mathbf{H}^1(\Omega \setminus \overline{\omega}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega})$ *is the solution of the Stokes boundary values problem:*

$$\begin{cases} 2 - \operatorname{div}(\sigma(\boldsymbol{w}, q)) = \mathbf{0} & \text{in } \Omega \setminus \overline{\omega}, \\ \operatorname{div} \boldsymbol{w} = 0 & \text{in } \Omega \setminus \overline{\omega}, \\ \boldsymbol{w} = \sigma(\boldsymbol{u}, p)\mathbf{n} - \boldsymbol{f_b} & \text{on } \partial\Omega, \\ \boldsymbol{w} = \mathbf{0} & \text{on } \partial\omega, \end{cases} \tag{9}$$

with the normalization condition $\langle \sigma(\boldsymbol{w}, q)\mathbf{n} \, , \, \mathbf{n} \rangle_{\partial\Omega} = 0$.

Remark 1. Propositions 1 and 2 remain true under weaker assumptions. Indeed, the proofs are still valid if ω has a $C^{1,1}$ boundary and $\boldsymbol{V} \in \mathbf{W}^{2,\infty}(\mathbb{R}^N)$. However, in this case, the expression of $\mathrm{D}\, J(\omega) \cdot \boldsymbol{V}$ has to be understood as a duality product $\mathrm{H}^{-1/2} \times \mathrm{H}^{1/2}$ and $(\boldsymbol{u}', p')$ only belongs to $\mathbf{H}^1(\Omega \setminus \overline{\omega}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega})$. Moreover, we will prove Proposition 1 only assuming Ω is Lipschitz.

Second order analysis: justification of the instability. Finally, we want to study the stability of the optimization problem (6) at ω^*.

Proposition 3 (Characterization of the shape Hessian at a critical shape)**.** *The solution* $(\boldsymbol{u}, p)$ *is twice differentiable with respect to the domain. Moreover, for* $\boldsymbol{V} \in \boldsymbol{U}$*, we have*

$$\mathrm{D}^2 J(\omega^*) \cdot \boldsymbol{V} \cdot \boldsymbol{V} = -\int_{\partial\omega^*} [(\sigma(\boldsymbol{w}', q')\mathbf{n}) \cdot \partial_{\mathbf{n}} \boldsymbol{u}] \, (\boldsymbol{V} \cdot \mathbf{n}),$$

where $(\boldsymbol{w}', q') \in \mathbf{H}^1(\Omega \setminus \overline{\omega^*}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega^*})$ *is the solution of the following problem:*

$$\begin{cases} -\operatorname{div}(\sigma(\boldsymbol{w}', q')) = \mathbf{0} & \text{in } \Omega \setminus \overline{\omega^*}, \\ \operatorname{div} \boldsymbol{w}' = 0 & \text{in } \Omega \setminus \overline{\omega^*}, \\ \boldsymbol{w}' = \sigma(\boldsymbol{u}', p')\mathbf{n} & \text{on } \partial\Omega, \\ \boldsymbol{w}' = \mathbf{0} & \text{on } \partial\omega^*, \end{cases}$$

with the normalization condition $\langle \sigma(\boldsymbol{w}', q')\mathbf{n} \, , \, \mathbf{n} \rangle_{\partial\Omega} = 0$.

Proposition 4 (Compactness at a critical point)**.** *The Riesz operator corresponding to* $\mathrm{D}^2 J(\omega^*)$ *defined from* $\mathbf{H}^{1/2}(\partial\omega^*)$ *to* $\mathbf{H}^{-1/2}(\partial\omega^*)$ *is compact.*

This last statement points out the lack of stability of the optimization problem (6). This compactness result means, roughly speaking, that in a neighborhood of ω^* (*i.e.* for t small), J behaves as its second order approximation and one cannot expect an estimate of the kind $C\,t \leq \sqrt{J(\omega_t)}$ with a constant C uniform in $\boldsymbol{V}$. This proposition emphasizes that the gradient has not a uniform sensitivity with respect to the deformation directions: J is degenerate for the high frequencies. This explains the numerical difficulties encountered to solve numerically this problem. For more details, we refer to [3, §2.3].

§3. Differentiability results

To prove the existence of the shape derivatives of the state, we have to prove the existence of the total first variations. In order to prove it, we use a generalized implicit function theorem proved by J. Simon (see [8, Theorem 6]) that we recall the statement for the reader's convenience.

Theorem 5 (J. Simon [8])**.** *We give us*

- *an open set* $\mathcal{U}$ *in a Banach space* U*,* $u_0 \in \mathcal{U}$*, two reflexive Banach spaces* E_1 *and* E_2*,*
- *a map* $F : \mathcal{U} \times E_1 \to E_2$*, such that* $F(u,\cdot) \in \mathcal{L}(E_1, E_2)$ *for all* $u \in \mathcal{U}$*,*
- *a function* $m : \mathcal{U} \to E_1$ *and a function* $f : \mathcal{U} \to E_2$ *such that*

$$F(u, m(u)) = f(u) \quad \forall u \in \mathcal{U}.$$

(i) *Assume that*

- $u \mapsto F(u,\cdot)$ *is differentiable at* u_0 *into* $\mathcal{L}(E_1, E_2)$*,*
- f *is differentiable at* u_0*,*
- $\|F(u_0, x)\|_{E_2} \geq \alpha \|x\|_{E_1} \quad \forall x \in E_1$*, for some* $\alpha > 0$*.*

Then, the map $u \mapsto m(u)$ *is differentiable at* u_0*. Its derivative* $m'(u_0,\cdot)$ *is the unique solution of*

$$F(u_0, m'(u_0, v)) = f'(u_0, v) - \partial_u F(u_0, m(u_0), v) \quad \forall v \in U.$$

(ii) *In addition, assume that for some integer* $k \geq 1$*,* $u \mapsto F(u,\cdot)$ *and* f *are* k *times differentiable at* u_0*. Then, the map* $u \mapsto m(u)$ *is* k *times differentiable at* u_0*.*

Let $\boldsymbol{\theta} \in \mathcal{U}$. We set $(\boldsymbol{u_\theta}, p_\theta)$ the unique solution in $\mathbf{H}^1(\Omega \setminus \overline{\omega_\theta}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega_\theta})$ of

$$\begin{cases} -\operatorname{div}(\sigma(\boldsymbol{u_\theta}, p_\theta)) = \mathbf{0} & \text{in } \Omega \setminus \overline{\omega_\theta}, \\ \operatorname{div} \boldsymbol{u_\theta} = 0 & \text{in } \Omega \setminus \overline{\omega_\theta}, \\ \boldsymbol{u_\theta} = \boldsymbol{g} & \text{on } \partial\Omega, \\ \boldsymbol{u_\theta} = \mathbf{0} & \text{on } \partial\omega_\theta, \end{cases}$$

with $\langle (\sigma(\boldsymbol{u_\theta}, p_\theta)\mathbf{n}) \cdot \mathbf{n}\,,\ 1\rangle_{\partial\Omega} = \langle \boldsymbol{f_b} \cdot \mathbf{n}\,,\ 1\rangle_{\partial\Omega}$. Let us consider $\boldsymbol{G} \in \mathbf{H}^1(\Omega)$ such that

$$\boldsymbol{G} = \boldsymbol{g} \ \text{ on } \partial\Omega, \quad \operatorname{div} \boldsymbol{G} = 0 \ \text{ in } \Omega \quad \text{ and } \quad \boldsymbol{G} = \mathbf{0} \ \text{ in } \Omega_\delta.$$

Thus $(\boldsymbol{z_\theta} = \boldsymbol{u_\theta} - \boldsymbol{G}, p_\theta) \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega_\theta}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega_\theta})$ is such that

$$\begin{cases} \displaystyle\int_{\Omega\setminus\overline{\omega_\theta}} \sigma(\boldsymbol{z_\theta}, p_\theta) : \nabla\boldsymbol{\varphi}_\theta = -\int_{\Omega\setminus\overline{\omega_\theta}} \nu\nabla\boldsymbol{G} : \nabla\boldsymbol{\varphi}_\theta, & \forall\boldsymbol{\varphi_\theta} \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega_\theta}), \\ \displaystyle\int_{\Omega\setminus\overline{\omega_\theta}} \xi_\theta \operatorname{div} \boldsymbol{z_\theta} = 0, & \forall\xi_\theta \in \mathrm{L}^2(\Omega \setminus \overline{\omega_\theta}), \\ \langle(\sigma(\boldsymbol{z_\theta}, p_\theta)\mathbf{n}) \cdot \mathbf{n}\,,\ 1\rangle_{\partial\Omega} = \langle(\boldsymbol{f_b} - \sigma(\boldsymbol{G}, 0)\mathbf{n}) \cdot \mathbf{n}\,,\ 1\rangle_{\partial\Omega}\,. \end{cases} \tag{10}$$

Let us define the key objects of our differentiability proof:

$$\boldsymbol{v_\theta} = \boldsymbol{z_\theta} \circ (\mathbf{I} + \boldsymbol{\theta}) \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega}) \quad \text{and} \quad q_\theta = p_\theta \circ (\mathbf{I} + \boldsymbol{\theta}) \in \mathrm{L}^2(\Omega \setminus \overline{\omega}).$$

For $k \geq -1$ and $m \geq 0$ integers with $k < m$, we note $\mathrm{X}^{k,m}(\Omega \setminus \overline{\omega}, \Omega_\delta \setminus \overline{\omega})$ the space of functions in $\mathrm{H}^k(\Omega \setminus \overline{\omega})$ such that their restriction to $\Omega_\delta \setminus \overline{\omega}$ belongs to $\mathrm{H}^m(\Omega_\delta \setminus \overline{\omega})$. This space endowed with the norm $\|u\|_{\mathrm{X}^{k,m}(\Omega\setminus\overline{\omega},\Omega_\delta\setminus\overline{\omega})} = \left(\|u\|^2_{\mathrm{H}^k(\Omega\setminus\overline{\omega})} + \|u\|^2_{\mathrm{H}^m(\Omega_\delta\setminus\overline{\omega})}\right)^{1/2}$ is hilbertian.

First order differentiability. To prove the existence of the first order shape derivative, we first have to prove the following three lemmas:

Lemma 6 (Characterization of $(\boldsymbol{v_\theta}, q_\theta)$). *For $\boldsymbol{\theta} \in \boldsymbol{\mathcal{U}}$, the pair $(\boldsymbol{v_\theta}, q_\theta)$ satisfies for all test functions $\boldsymbol{\varphi} \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega})$ and $\xi \in \mathrm{L}^2(\Omega \setminus \overline{\omega})$*

$$\begin{cases} \displaystyle\int_{\Omega\setminus\overline{\omega}} [(\nu\nabla\boldsymbol{v_\theta} A(\boldsymbol{\theta})) : \nabla\boldsymbol{\varphi} - q_\theta B(\boldsymbol{\theta}) : \nabla\boldsymbol{\varphi}] = \int_{\Omega\setminus\overline{\omega}} -\nu\,\nabla\boldsymbol{G} : \nabla\boldsymbol{\varphi}, \\ \displaystyle\int_{\Omega\setminus\overline{\omega}} (\nabla\boldsymbol{v_\theta} : B(\boldsymbol{\theta}))\,\xi = 0, \\ \langle(\sigma(\boldsymbol{v}, q)\mathbf{n}) \cdot \mathbf{n}\,,\ 1\rangle_{\partial\Omega} = \langle(\boldsymbol{f_b} - \sigma(\boldsymbol{G}, 0)\mathbf{n}) \cdot \mathbf{n}\,,\ 1\rangle_{\partial\Omega}\,, \end{cases}$$

with

$$\begin{aligned} J_{\boldsymbol{\theta}} &= \det(\mathrm{I} + \nabla\boldsymbol{\theta}) \in \mathrm{W}^{2,\infty}\left(\overline{\Omega_\delta}\right), \\ A(\boldsymbol{\theta}) &= J_{\boldsymbol{\theta}}\,(\mathrm{I} + \nabla\boldsymbol{\theta})^{-1}(\mathrm{I} + {}^t\nabla\boldsymbol{\theta})^{-1} \in \mathrm{W}^{2,\infty}\left(\overline{\Omega_\delta}, \mathcal{M}_{N,N}\right), \\ B(\boldsymbol{\theta}) &= J_{\boldsymbol{\theta}}(\mathrm{I} + {}^t\nabla\boldsymbol{\theta})^{-1} \in \mathrm{W}^{2,\infty}\left(\overline{\Omega_\delta}, \mathcal{M}_{N,N}\right). \end{aligned}$$

Lemma 7 (Differentiability of $\boldsymbol{\theta} \mapsto (\boldsymbol{v_\theta}, q_\theta)$). *The function*

$$\boldsymbol{\theta} \in \boldsymbol{\mathcal{U}} \mapsto (\boldsymbol{v_\theta}, q_\theta) \in \mathbf{X}^{1,2}(\Omega \setminus \overline{\omega}, \Omega_\delta \setminus \overline{\omega}) \times \mathrm{X}^{0,1}(\Omega \setminus \overline{\omega}, \Omega_\delta \setminus \overline{\omega})$$

is differentiable in a neighborhood of $\mathbf{0}$.

Lemma 8 (Differentiability of $\boldsymbol{\theta} \mapsto (\boldsymbol{u_\theta}, p_\theta)$). *There exists $\widetilde{\boldsymbol{u}}_{\boldsymbol{\theta}}$, $\widetilde{p}_\theta$ some respective extensions of $\boldsymbol{u_\theta} \in \mathbf{H}^1(\Omega \setminus \overline{\omega})$, $p_\theta \in \mathrm{L}^2(\Omega \setminus \overline{\omega})$ such that the functions*

$$\boldsymbol{\theta} \in \boldsymbol{\mathcal{U}} \mapsto \widetilde{\boldsymbol{u}}_{\boldsymbol{\theta}} \in \mathbf{H}^1(\Omega) \quad \textit{and} \quad \boldsymbol{\theta} \in \boldsymbol{\mathcal{U}} \mapsto \widetilde{p}_\theta \in \mathrm{L}^2(\Omega)$$

are differentiable at $\mathbf{0}$.

Remark 2. We will prove this three lemmas under weaker assumptions: ω with a $C^{1,1}$ boundary, Ω with a Lipschitz boundary and $\boldsymbol{\theta} \in \mathbf{W}^{2,\infty}(\mathbb{R}^N)$.

Proof of Lemma 6: characterization of $(\boldsymbol{v}_{\boldsymbol{\theta}}, q_\theta)$. We make a change of variables in (10). First, notice that, since $\operatorname{div} \boldsymbol{z}_{\boldsymbol{\theta}} = 0$ in $\Omega \setminus \overline{\omega_\theta}$,

$$\int_{\Omega\setminus\overline{\omega_\theta}} \sigma(\boldsymbol{z}_{\boldsymbol{\theta}}, p_\theta) : \nabla \boldsymbol{\varphi}_\theta = \int_{\Omega\setminus\overline{\omega_\theta}} (\nu \nabla \boldsymbol{z}_{\boldsymbol{\theta}} : \nabla \boldsymbol{\varphi}_\theta - p_\theta \operatorname{div} \boldsymbol{\varphi}_\theta), \quad \forall \boldsymbol{\varphi}_{\boldsymbol{\theta}} \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega_\theta}).$$

Let $\boldsymbol{\varphi} \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega})$, $\xi \in \mathrm{L}^2(\Omega \setminus \overline{\omega})$ and $\boldsymbol{\theta} \in \boldsymbol{\mathcal{U}}$. Then we proceed in the same manner than the proof of Lemma 3.1 in [3]: we use the test functions $\boldsymbol{\varphi}_{\boldsymbol{\theta}} = \boldsymbol{\varphi} \circ (\mathbf{I} + \boldsymbol{\theta})^{-1} \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega_\theta})$ and $\xi_\theta = \xi \circ (\mathbf{I} + \boldsymbol{\theta})^{-1} \in \mathrm{L}^2(\Omega \setminus \overline{\omega_\theta})$ in the variational formulation (10) and we make the change of variables $x = (\mathbf{I} + \boldsymbol{\theta})y$. Noticing that $\boldsymbol{\theta} \equiv \mathbf{0}$ in $\Omega \setminus \overline{\Omega_\delta}$ (and therefore on $\partial\Omega$) and that $\boldsymbol{G} \equiv 0$ in Ω_δ, we obtain the result. □

The proof of Lemma 7 is based on Simon's Theorem: we adapt the method used in the proof of Lemma 3.2 in [3].

Proof of Lemma 7: differentiability of $\theta \mapsto (\boldsymbol{v}_{\boldsymbol{\theta}}, q_\theta)$. Let us check the assumptions of Simon's Theorem.

First step: notations. We need some additional tools: a third domain $\widetilde{\Omega}_\delta$ which is an open set with a C^∞ boundary such that $\Omega_\delta \subset\subset \widetilde{\Omega}_\delta \subset\subset \Omega$ and a truncation function $\Phi \in C^\infty_c(\widetilde{\Omega}_\delta)$ such that $\Phi \equiv 1$ in Ω_δ. Then, we define the spaces

$$\begin{aligned}
\boldsymbol{E_1} &= \left\{(\boldsymbol{v}, q) \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega})\,;\ (\Phi\boldsymbol{v}, \Phi q) \in \mathbf{H}^2(\Omega \setminus \overline{\omega}) \times \mathrm{H}^1(\Omega \setminus \overline{\omega})\right\},\\
\boldsymbol{E_2} &= \left\{(\boldsymbol{f}, g) \in \mathbf{H}^{-1}(\Omega \setminus \overline{\omega}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega});\ (\Phi\boldsymbol{f}, \Phi g) \in \mathbf{L}^2(\Omega \setminus \overline{\omega}) \times \mathrm{H}^1(\Omega \setminus \overline{\omega})\right\} \times \mathbb{R}.
\end{aligned}$$

Note that $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ are Hilbert spaces with respective norms

$$\begin{aligned}
\|(\boldsymbol{v}, q)\|^2_{\boldsymbol{E_1}} &= \|\boldsymbol{v}\|^2_{\mathbf{H}^1(\Omega\setminus\overline{\omega})} + \|q\|^2_{\mathrm{L}^2(\Omega\setminus\overline{\omega})} + \|\Phi\boldsymbol{v}\|^2_{\mathbf{H}^2(\Omega\setminus\overline{\omega})} + \|\Phi q\|^2_{\mathrm{H}^1(\Omega\setminus\overline{\omega})},\\
\|((\boldsymbol{f}, g), r)\|^2_{\boldsymbol{E_2}} &= \|\boldsymbol{f}\|^2_{\mathbf{H}^{-1}(\Omega\setminus\overline{\omega})} + \|g\|^2_{\mathrm{L}^2(\Omega\setminus\overline{\omega})} + \|\Phi\boldsymbol{f}\|^2_{\mathbf{L}^2(\Omega\setminus\overline{\omega})} + \|\Phi g\|^2_{\mathrm{H}^1(\Omega\setminus\overline{\omega})} + |r|^2.
\end{aligned}$$

Moreover, we can also notice that $\boldsymbol{E_1} \hookrightarrow \mathbf{X}^{1,2}(\Omega \setminus \overline{\omega}, \Omega_\delta \setminus \overline{\omega}) \times \mathrm{X}^{0,1}(\Omega \setminus \overline{\omega}, \Omega_\delta \setminus \overline{\omega})$ and that $\boldsymbol{E_2} \hookrightarrow \mathbf{X}^{-1,0}(\Omega \setminus \overline{\omega}, \Omega_\delta \setminus \overline{\omega}) \times \mathrm{X}^{0,1}(\Omega \setminus \overline{\omega}, \Omega_\delta \setminus \overline{\omega}) \times \mathbb{R}$. Using the notations introduced in Lemma 6, we also define, for $\boldsymbol{\theta} \in \mathcal{U}$ and $(\boldsymbol{v}, q) \in \boldsymbol{E_1}$, the following functions:

- $\boldsymbol{f_1}(\boldsymbol{\theta}) \in \mathbf{H}^{-1}(\Omega \setminus \overline{\omega})$ by $\forall \boldsymbol{\varphi} \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega})$,

$$\langle \boldsymbol{f_1}(\boldsymbol{\theta}), \boldsymbol{\varphi} \rangle_{\Omega\setminus\overline{\omega}} = -\int_{\Omega\setminus\overline{\omega}} \nu J_{\boldsymbol{\theta}} \nabla \boldsymbol{G} : \nabla \boldsymbol{\varphi} = -\int_{\Omega\setminus\overline{\Omega_\delta}} \nu \nabla \boldsymbol{G} : \nabla \boldsymbol{\varphi},$$

- $\boldsymbol{F_1}(\boldsymbol{\theta}, (\boldsymbol{v}, q)) \in \mathbf{H}^{-1}(\Omega \setminus \overline{\omega})$ by $\forall \boldsymbol{\varphi} \in \mathbf{H}^1_0(\Omega \setminus \overline{\omega})$,

$$\langle \boldsymbol{F_1}(\boldsymbol{\theta}, (\boldsymbol{v}, q)), \boldsymbol{\varphi} \rangle_{\Omega\setminus\overline{\omega}} = \int_{\Omega\setminus\overline{\omega}} \{[\nu \nabla \boldsymbol{v}\, A(\boldsymbol{\theta})] : \nabla \boldsymbol{\varphi} - q\, B(\boldsymbol{\theta}) : \nabla \boldsymbol{\varphi}\},$$

- $\boldsymbol{m}(\boldsymbol{\theta}) = (\boldsymbol{v}_{\boldsymbol{\theta}}, q_\theta)$ and $\boldsymbol{f}(\boldsymbol{\theta}) = (\boldsymbol{f_1}(\boldsymbol{\theta}), 0, \langle(\boldsymbol{f_b} - \sigma(\boldsymbol{G}, 0)\mathbf{n}) \cdot \mathbf{n}, 1\rangle_{\partial\Omega})$,

- $\boldsymbol{F}(\boldsymbol{\theta},(\boldsymbol{v},q)) = \big(\boldsymbol{F_1}(\boldsymbol{\theta},(\boldsymbol{v},q)),\, \nabla\boldsymbol{v} : B(\boldsymbol{\theta}),\, \langle(\sigma(\boldsymbol{v},q)\mathbf{n})\cdot\mathbf{n},\, 1\rangle_{\partial\Omega}\big)$.

By the characterization of $(\boldsymbol{v_\theta}, q_\theta)$ obtained in Lemma 6,

$$\boldsymbol{F}(\boldsymbol{\theta},\boldsymbol{m}(\boldsymbol{\theta})) = \boldsymbol{f}(\boldsymbol{\theta}) \quad \forall\boldsymbol{\theta}\in\boldsymbol{\mathcal{U}}.$$

Second step: differentiability of $\boldsymbol{F}$ *and* $\boldsymbol{f}$ *at* $\mathbf{0}$. In the same way as what is done in the proof of Lemma 3.2 in [3], we prove that $\boldsymbol{F}$ and $\boldsymbol{f}$ are C^∞ in a neighborhood of $\mathbf{0}$.

Third step: existence of $\alpha > 0$ *such that* $\|\boldsymbol{F}(\mathbf{0},(\boldsymbol{v},q))\|_{\boldsymbol{E_2}} \geq \alpha\|(\boldsymbol{v},q)\|_{\boldsymbol{E_1}}$. We consider a pair $(\boldsymbol{v},q)\in\boldsymbol{E_1}$ and we define $(\boldsymbol{\xi},\eta,\boldsymbol{r})\in\boldsymbol{E_2}$ by $\boldsymbol{F}(\mathbf{0},(\boldsymbol{v},q)) = (\boldsymbol{\xi},\eta,\boldsymbol{r})$. Then,

$$\begin{cases} \displaystyle\int_{\Omega\setminus\overline{\omega}} \{\nu\nabla\boldsymbol{v} : \nabla\boldsymbol{\varphi} - q\,\mathrm{div}\,\boldsymbol{\varphi}\} = \langle\boldsymbol{\xi},\,\boldsymbol{\varphi}\rangle_{\Omega\setminus\overline{\omega}} & \forall\boldsymbol{\varphi}\in\mathbf{H}^1_0(\Omega\setminus\overline{\omega}),\\ \displaystyle\int_{\Omega\setminus\overline{\omega}} \phi\,\mathrm{div}\,\boldsymbol{v} = \int_{\Omega\setminus\overline{\omega}} \phi\,\eta & \forall\phi\in\mathrm{L}^2(\Omega\setminus\overline{\omega}),\\ \langle(\sigma(\boldsymbol{v},q)\mathbf{n})\cdot\mathbf{n},\,1\rangle_{\partial\Omega} = r. \end{cases}$$

The compatibility condition of the previous problem is automatically satisfied because of $\int_{\Omega\setminus\overline{\omega}}\eta = \int_{\partial(\Omega\setminus\overline{\omega})}\boldsymbol{v}\cdot\mathbf{n} = 0$ since $\boldsymbol{v}\in\mathbf{H}^1_0(\Omega\setminus\overline{\omega})$. Thus, proceeding in the same manner than in the proof of Lemma 3.2 in [3], we check using a local regularity argument that there exists a constant $\alpha > 0$ such that

$$\|\boldsymbol{F}(\mathbf{0},(\boldsymbol{v},q))\|_{\boldsymbol{E_2}} \geq \alpha\|(\boldsymbol{v},q)\|_{\boldsymbol{E_1}}.$$

Fourth step: conclusion. By Simon's Theorem, the function $\boldsymbol{\theta}\in\boldsymbol{\mathcal{U}}\mapsto(\boldsymbol{v_\theta},q_\theta)\in\boldsymbol{E_1}$ is differentiable (and even C^∞) in a neighborhood of $\mathbf{0}$. We conclude using the fact that $\boldsymbol{E_1}$ is continuously embedded in $\mathbf{X}^{1,2}(\Omega\setminus\overline{\omega},\Omega_\delta\setminus\overline{\omega})\times\mathrm{X}^{0,1}(\Omega\setminus\overline{\omega},\Omega_\delta\setminus\overline{\omega})$. □

Proof of Lemma 8: differentiability of $\boldsymbol{\theta}\mapsto(\boldsymbol{u_\theta},p_\theta)$. This proof is exactly the same than the proof of Lemma 3.3 in [3]. We refer to this one for details. The idea is to use the differentiability result by composition by $(\mathbf{I}+\boldsymbol{\theta})^{-1}$ (see [6, Lemma 5.3.9]). □

Higher order differentiability. To prove the existence of the second total variations, we will proceed in the same way that what is done previously. We mimic the proof of Lemma 7, only increasing the local regularity in the used spaces to prove that the function

$$\boldsymbol{\theta}\in\boldsymbol{\mathcal{U}}\mapsto(\boldsymbol{v_\theta},q_\theta)\in\mathbf{X}^{1,3}(\Omega\setminus\overline{\omega},\Omega_\delta\setminus\overline{\omega})\times\mathrm{X}^{0,2}(\Omega\setminus\overline{\omega},\Omega_\delta\setminus\overline{\omega})$$

is twice differentiable in a neighborhood of $\mathbf{0}$. Then, proceeding in exactly the same way than in the proof of Lemma 3.5 in [3], we prove the following lemma:

Lemma 9 (Second order shape differentiability)**.** *The solution* $(\boldsymbol{u},p)$ *is twice differentiable with respect to the domain.*

§4. Proof of the main results

First order shape derivatives of the state. *Proof of Proposition 1.* The existence of the shape derivative $(\boldsymbol{u}',p')$ is proved using the Fréchet differentiability Lemma 8. Using the variational formulation of problem (7), we use classical shape derivatives calculus to characterize

$(\boldsymbol{u}', p')$ (see [6, proof of Theorem 5.3.1] concerning the Laplacian case for example). We just precise that, since $\boldsymbol{u} = \boldsymbol{0}$ on $\partial\omega$, $\nabla\boldsymbol{u} = \partial_{\mathbf{n}}\boldsymbol{u} \otimes \mathbf{n}$, where $\otimes$ is the tensorial product. Hence the classical boundary condition $\boldsymbol{u}' = -\nabla\boldsymbol{u}\,\boldsymbol{V}$ on $\partial\omega$ can be written $\boldsymbol{u}' = -\partial_{\mathbf{n}}\boldsymbol{u}\,(\boldsymbol{V}\cdot\mathbf{n})$. □

First order shape derivatives of the functional. For all $t \in [0, T)$, consider $(\boldsymbol{u_t}, p_t)$ solution of (7) and define,

$$J(\omega_t) = j(t) = \frac{1}{2}\int_{\partial\Omega} |\sigma(\boldsymbol{u_t}, p_t)\,\mathbf{n} - \boldsymbol{f_b}|^2.$$

Proof of Proposition 2. First step: derivative of j and adjoint problem. Noting $(\boldsymbol{u}', p')$ the shape derivative of $(\boldsymbol{u}, p)$, we differentiate j with respect to t at 0 to obtain

$$j'(0) = \nabla J(\omega)\cdot\boldsymbol{V} = \int_{\partial\Omega} (\sigma(\boldsymbol{u}', p')\,\mathbf{n})\cdot(\sigma(\boldsymbol{u}, p)\,\mathbf{n} - \boldsymbol{f_b})\,. \tag{11}$$

Then, we consider the adjoint problem (9). Since we choose the normalization condition (5), the compatibility condition of the adjoint problem is satisfied. Therefore it admits a unique solution $(\boldsymbol{w}, q) \in \mathbf{H}^1(\Omega\setminus\overline{\omega})\times\mathrm{L}^2(\Omega\setminus\overline{\omega})$ with $\langle\sigma(\boldsymbol{w}, q)\,\mathbf{n}\,,\,\mathbf{n}\rangle_{\Omega\setminus\overline{\omega}} = 0$.

Second step: writing of $j'(0)$ *as an integral on* $\partial\omega$*.* We proceed by successive integrations by parts. We multiply the first equation of the adjoint problem (9) by $\boldsymbol{u}'$ to get

$$\int_{\Omega\setminus\overline{\omega}} \nu\,\nabla\boldsymbol{w} : \nabla\boldsymbol{u}' = -\left\langle -\sigma(\boldsymbol{w}, q)\,\mathbf{n}\,,\,\boldsymbol{u}'\right\rangle_{\partial(\Omega\setminus\overline{\omega})}\,, \tag{12}$$

since $\operatorname{div}\boldsymbol{u}' = 0$ in $\Omega\setminus\overline{\omega}$ (see Proposition 1). Then, we multiply the first equation of the problem (8) by $\boldsymbol{w}$ to obtain

$$\int_{\Omega\setminus\overline{\omega}} \nu\,\nabla\boldsymbol{u}' : \nabla\boldsymbol{w} = -\left\langle -\sigma(\boldsymbol{u}', p')\,\mathbf{n}\,,\,\boldsymbol{w}\right\rangle_{\partial(\Omega\setminus\overline{\omega})}\,, \tag{13}$$

since $\operatorname{div}\boldsymbol{w} = 0$ in $\Omega\setminus\overline{\omega}$. Gathering (11), (12) and (13) and using the boundary conditions of $(\boldsymbol{u}', p')$ and $(\boldsymbol{w}, q)$ (see problems (8) and (9)), we obtain the announced result. □

Characterization of the shape Hessian at a critical point. We consider $\omega^* \in \mathcal{O}_\delta$ a critical shape of the functional J.

Proof of Proposition 3. First step: second order shape differentiability. By Lemma 9, the second order shape derivative exists which is noted $(\boldsymbol{u}'', p'')$.

Second step: second derivative of j and derivative of the adjoint problem. Let $\boldsymbol{V} \in \boldsymbol{U}$. We differentiate the function j twice with respect to t. At $t = 0$, it holds

$$j''(0) = \mathrm{D}^2J(\omega)\cdot\boldsymbol{V}\cdot\boldsymbol{V} = \int_{\partial\Omega}\left[(\sigma(\boldsymbol{u}'', p'')\,\mathbf{n})\cdot((\sigma(\boldsymbol{u}, p)\,\mathbf{n}) - \boldsymbol{f_b}) + |\sigma(\boldsymbol{u}', p')\,\mathbf{n}|^2\right].$$

Since ω^* solves the inverse problem, $\sigma(\boldsymbol{u}, p)\,\mathbf{n} = \boldsymbol{f_b}$ on $\partial\Omega$. Therefore

$$\mathrm{D}^2J(\omega^*)\cdot\boldsymbol{V}\cdot\boldsymbol{V} = 2\int_{\partial\Omega}|\sigma(\boldsymbol{u}', p')\,\mathbf{n}|^2. \tag{14}$$

We introduce $(\boldsymbol{w}, q) \in \mathbf{H}^1(\Omega \setminus \overline{\omega}) \times \mathrm{L}^2(\Omega \setminus \overline{\omega})$ with $\langle \sigma(\boldsymbol{w}, q)\,\mathbf{n}\,,\,\mathbf{n}\rangle_{\partial\Omega} = 0$ the solution of the adjoint system (9). Notice that, for $\omega = \omega^*$, $\sigma(\boldsymbol{u}, p)\,\mathbf{n} = \boldsymbol{f_b}$ on $\partial\Omega$. Hence, the uniqueness of the solution of the Stokes problem enforces that $\boldsymbol{w} = \mathbf{0}$ in $\Omega \setminus \overline{\omega^*}$. Therefore, characterizing $\boldsymbol{w'}$ and q', the shape derivatives of $\boldsymbol{w}$ and q, in the same manner that we characterized $\boldsymbol{u'}$ and p' (see Proposition 1), we obtain the system (3).

Third step: writing of $j''(0)$ as an integral on $\partial\omega$. We multiply the first equation of problem (3) by $\boldsymbol{u'}$ to get

$$\int_{\Omega\setminus\overline{\omega^*}} \nu\nabla\boldsymbol{w'} \,:\, \nabla\boldsymbol{u'} = -\left\langle -\sigma(\boldsymbol{w'}, q')\,\mathbf{n}\,,\,\boldsymbol{u'}\right\rangle_{\partial\omega^*}. \tag{15}$$

We multiply the first equation of problem (8) by $\boldsymbol{w'}$ to get

$$\int_{\Omega\setminus\overline{\omega^*}} \nu\nabla\boldsymbol{u'} \,:\, \nabla\boldsymbol{w'} = -\left\langle -\sigma(\boldsymbol{u'}, p')\,\mathbf{n}\,,\,\boldsymbol{w'}\right\rangle_{\partial\Omega}. \tag{16}$$

Therefore, gathering (14), (15) and (16), we obtain the announced result. □

Justifying the ill-posedness of the problem. *Proof of Proposition 4.* The proof is an adaptation of the proof of Proposition 2.8 in [3]. The idea is to decompose the shape Hessian as a composition of linear continuous operators and a compact operator. The compactness is proved using a local regularity argument. □

§5. Conclusion

The formal calculus of the shape derivative for the Stokes equations is easier in the Dirichlet case than in the Neumann case which is presented by M. Badra *et al.* in [3], particularly the characterization of $(\boldsymbol{u'}, p')$. However, an other difficulty arises here, due to the introduction of the adjoint problem (9). Indeed, the boundary condition $\sigma(\boldsymbol{u}, p)\,\mathbf{n} - \boldsymbol{f_b}$ on $\partial\Omega$ imposed in (9) has to belong in $\mathbf{H}^{1/2}(\partial\Omega)$. Thus, we have to assume that $\partial\Omega$ is $C^{1,1}$ while we can work with a Lipschitz domain in the Neumann case. Moreover, if we want to make the measurement on a part O of $\partial\Omega$ like what is done in [3], we are confronted to the same difficulty. Indeed, the boundary condition on $\partial\Omega$ of the adjoint problem (9) would be then $(\sigma(\boldsymbol{u}, p)\,\mathbf{n} - \boldsymbol{f_b})\mathbb{1}_O$ which does not belong to $\mathbf{H}^{1/2}(\partial\Omega)$, even if Ω is smooth. A solution could be to use the *very weak solutions* (see *e.g.* [2, §4.2, Definition 1]), even if this method need again that $\partial\Omega$ is $C^{1,1}$. Then, it would be necessary to prove the differentiability with respect to the domain of the *very weak solution* $(\boldsymbol{w}, q) \in \mathbf{L}^2(\Omega \setminus \overline{\omega}) \times \mathrm{H}^{-1}(\Omega \setminus \overline{\omega})/\mathbb{R}$ of the adjoint problem, which is not classical. An other solution is to use a smooth cut-off function as what is done in [4].

References

[1] Alvarez, C., Conca, C., Friz, L., Kavian, O., and Ortega, J. H. Identification of immersed obstacles via boundary measurements. *Inverse Problems 21*, 5 (2005).

[2] Amrouche, C., and Rodríguez-Bellido, M. Á. Stationary Stokes, Oseen and Navier-Stokes equations with singular data. *Arch. Ration. Mech. Anal. 199*, 2 (2011), 597–651.

[3] Badra, M., Caubet, F., and Dambrine, M. Detecting an obstacle immersed in a fluid by shape optimization methods. *M3AS 21* (2011), 2069–2101.

[4] Caubet, F. Shape sensitivity and instability of an inverse problem for the stationary navier-stokes equations. *SICON* (submitted).

[5] Galdi, G. P. *An introduction to the mathematical theory of the Navier-Stokes equations. Vol. I*, vol. 38 of *Springer Tracts in Natural Philosophy*. Springer-Verlag, New York, 1994. Linearized steady problems.

[6] Henrot, A., and Pierre, M. *Variation et optimisation de formes*, vol. 48 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Berlin, 2005. Une analyse géométrique. [A geometric analysis].

[7] Murat, F., and Simon, J. Sur le contrôle par un domaine géométrique. *Rapport du L.A. 189* (1976). Université de Paris VI, France.

[8] Simon, J. Domain variation for drag in Stokes flow. In *Control theory of distributed parameter systems and applications (Shanghai, 1990)*, vol. 159 of *Lecture Notes in Control and Inform. Sci.* Springer, Berlin, 1991, pp. 28–42.

[9] Temam, R. *Navier-Stokes equations*. AMS Chelsea Publishing, Providence, RI, 2001. Theory and numerical analysis, Reprint of the 1984 edition.

Fabien Caudet
UMR CNRS 5142 - LMA,
Université de Pau et des Pays de l'Adour
Postal address IPRA BP 64013 Pau Cedex (France)
fabien.caubet@univ-pau.fr

Monografías Matemáticas *García de Galdeano* **37**, 103–113 (2012)

SIMULTANEOUS TRIANGULATION OF COMMUTING FAMILIES OF MATRICES – WHY AND HOW PRECISELY?

Vanesa Cortés, Juan Manuel Peña and Tomas Sauer

Abstract. We present an algorithm that provides an extension of the QR method in order to compute the joint eigenvalues of a family of commuting real matrices.

Keywords: QR method, commuting matrices, simultaneous triangulation.

AMS classification: 65F15.

§1. Introduction

In this paper, we present the algorithm that describes an extension of the QR method to simultaneously compute the joint eigenvalues of a finite family of commuting matrices defined in [1]. This problem is motivated by the task of finding solutions of polynomial systems of equations of the form

$$F(X) = 0, \qquad F = (f_1, \dots, f_m), \qquad f_j \in \mathbb{C}[x_1, \dots, x_n].$$

The idea behind this approach is to extend the *Frobenius companion matrix* to the multivariate case. Recall that if $f(x) = a_0 + a_1 x + \cdots + a_n x^n$, $a_n \neq 0$, is a polynomial in one variable, then its zeros are the eigenvalues of the *companion matrix*

$$A := \begin{bmatrix} 0 & & & & -a_0/a_n \\ 1 & 0 & & & -a_1/a_n \\ & \ddots & \ddots & & \vdots \\ & & 1 & 0 & -a_{n-2}/a_n \\ & & & 1 & -a_{n-1}/a_n \end{bmatrix}.$$

Since this result can be proved, for example, by division with remainder and considering multiplication modulo the (principal) ideal generated by f, it allows for an extension via computational ideal theory, especially the concept of Gröbner bases or H–bases. However, in n variables one does not have to find the eigenvalues of a single matrix A but the *joint eigenvalues* of a system $\mathcal{A} = (A_1, \dots, A_n)$ of *commuting* matrices, i.e. $A_j A_k = A_k A_j$. A naive and direct approach would be to compute the eigenvalues and eigenvectors for each of the matrices separately and then connect the eigenvalues (which are the components of one of the zeros) by means of the associated eigenvectors. This approach, however, faces difficulties as soon as the coordinate projections of the solutions are not well separated as then the eigenvectors will usually not be unique any more, hence the connection can only be made via a numerically instable intersection of eigenspaces. And while multiple eigenvalues do not constitute a thread for the *QR* method of eigenvalue dermination, cf. [3], clustered eigenvalues are a numerical problem, hence should be avoided. To overcome these difficulties, we proposed a method that extends the QR method and the included concept of splitting matrices by treating the matrices as simultaneous as possible, making use of the fact that under certain circumstances a *split*, that is, a separation of eigenspaces, obtained for one of the matrices from the family carries over to the whole family. The eigenvalue approach applies only to the case when the set X of solutions to $F(X) = 0$ is *finite*, or, in algebraic terminology, if the ideal generated by F is zero dimensional. Then the size of the matrices is the number of zeros which is determined automatically by the algebraic reduction process.

The following lemma, proved in [1], provides the condition that will allow us to perform the simultaneous triangularization process of all commuting matrices.

Lemma 1 (Cf. [1], Lemma 2.1)**.** *Let A, B be two real $n \times n$ matrices such that $AB = BA$. If there exists a nonsingular matrix P such that*

$$P^{-1}AP = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix},$$

where the $p \times p$ ($1 \leq p \leq n-1$) matrix A_1 and the $(n-p) \times (n-p)$ matrix A_3 satisfy $Spec(A_1) \cap Spec(A_3) = \emptyset$, then

$$P^{-1}BP = \begin{bmatrix} B_1 & B_2 \\ 0 & B_3 \end{bmatrix} \tag{1}$$

and B_1 is a $p \times p$ matrix.

In Section 2, we comment the steps of the algorithm and present the core `MATLAB` code of each one. In Section 3, we add the subroutines used in the main program.

§2. Algorithm

Let $\mathcal{A}$ be a set of m real commuting matrices $n \times n$ such that $A, B \in \mathcal{A}$ implies $AB = BA$. Besides, let *threshold*, *toleranceEig* be variables holding (small) positive numbers used as thresholds in the algorithm. We apply the following algorithm in order to find the orthogonal matrix P that has to exist by Lemma 1.

Step 1. Storing and ordering matrices. The matrices of $\mathcal{A}$ which have only one eigenvalue or a unique pair of complex eigenvalues will be recognized by a routine `uniqueEigenvalue` and will not be considered for decomposition. The other ones will concatenated into a single matrix.

```
j=0;
for i=1:m
    [H]=hessenberg(𝒜(:,(i-1)*n+1:(i)*n));
    [only]=uniqueEigenvalue(H,n,threshold,toleranceEig);
    if(only==0)
        j=j+1;
        $\mathbf{A}$(:,(j-1)*n+1:(j)*n)=$\mathcal{A}$(:,(i-1)*n+1:(i)*n);
    end -end if
end -end for
```

The assumption that the ideal is zero dimensional and that there is more than one solution of $F(X) = 0$ ensures that $\mathcal{A}$ is nonempty.

We then compute quantities that will allow us to order the matrices of $\mathcal{A}$ according to their spread. The *spread* is an easily computed estimate how widely the eigenvalues of a given matrix vary. A matrix with large spread is more likely to have well–separated eigenvalues and hence the *QR* iterations will be expected to lead to a separation of eigenspaces after a smaller number of iterations. To estimate the spread of the eigenvalues of a matrix in $\mathcal{A}$, we use Gerschgorin circles as well as an estimate based on determinant and trace of the matrix, cf. [1].

```
for i=1:j
    A=$\mathbf{A}$(:,(i-1)*n+1:(i)*n);
    L(i)=Gerschgorin(A);
    F(i)=means(A);
    G(i)=max(L(i),F(i));
end -end for
```

The matrix **A** is then reordered according to the values in G. This is in fact crucial as a large spread makes an early split after only a few iterations more likely.

Step 2. Initializing variables for starting the process. We will initialize a variable `matrix` with the first unused matrix in **A** and other variables (some of them related to possible found blocks on the process) that will play an important role through the algorithm. Since we also consider parts of the matrices after a *QR*–split, the size of the theses blocks is stored in a variable `newBlockMatrix` which is initialized with the full matrix.

```
newBlockMatrix=[1,n,1];
matrix=$\mathbf{A}$(:,(i-1)*n+1:i*n); initialMatrix=$\mathbf{A}$(:,(i-1)*n+1:i*n);
auxStoringQ=eye(n); change=0;
```

Step 3. Selecting the matrix which starts the process with the process matrix. A variable `processMatrix` will be initialized with the first unused matrix in **A**, and transformed into Hessenberg form with the routine `hessenberg` based on Householder reflections and stored again in the variable `processMatrix`.

Step 4. Selecting the process matrix if it has not order *n*. If we have to consider `blocksNum` subblocks of `matrix`, we extract those into a variable `processMatrix` to be treated by a *QR* factorization thereafter.

```
for v=1:blocksNum
    p=blockMatrix(v,1);q=blockMatrix(v,2);
    processMatrix=matrix(p:q,p:q);
    processMatrix=hessenberg(processMatrix);
    storingQ=eye(q-p+1);
    auxQ=eye(length(processMatrix));
    Process the matrix ...
end
```

Step 5. Initial transformation of the process matrix when it has trace zero. If the matrix `processMatrix` has trace zero, then we first perform two consecutive steps of the shifted QR algorithm with shifts (using the routine `qrShift`) in order to make our spread estimation work. The tolerance of 10^{-10} is just chosen as an example and can be adapted if necessary.

```
if(abs(trace(processMatrix))<=1e-10)
    [processMatrix,storingQ1]=choosingShift(processMatrix);
    storingQ=auxQ*storingQ1;
    auxQ=storingQ;
end -end if
```

Step 6. Localizating an special subdiagonal element to find an eigenvalue. If some subdiagonal element in `processMatrix` has an absolute value less than a given `tolerance`, that is, the matrix is already "almost split", we will apply a step of the *QR* algorithm with a shift whose value will be given by the routine `choiceShift`. As is well–known, a properly chosen shift will significantly improve the performance of the *QR* iteration, cf. [2]. This process will be continued until either we get the absolute value of this subdiagonal element smaller than a positive number `threshold`, less than `tolerance`, of course, or until we arrive at a maximal number of iterations.

```
[subdiagonal]=subdiagonal(processMatrix,w);
[j,lessThanTolerance] = min(subdiagonal <= tolerance);
while(subdiagonal(j)>threshold and approxNum<=n)
    [shift]=choosingShift(processMatrix,j);
    [processMatrix,obtainedQ]=qrShift(processMatrix,shift,w);
    storingQ=auxQ*obtainedQ;
    auxQ=storingQ;
    matrix(p:q,p:q)=processMatrix;
    subdiagonal(j)=abs(processMatrix(j+1,j));
    approxNum = approxNum + 1;
end
if(subdiagonal(j)<=threshold)
    first=p; last=p+j-1; first1=p+j; last1=q;
    lessThanThreshold=1;
    foundBlock=1;
    matrix(p:q,p:q)=processMatrix;
else
    lessThanThreshold=0;
end -end if
```

In the first case, we have obtained a candidate for the split and we continue with the process described in Step 7, while in the second case we continue as described in Step 8. Keep in mind here that according to Lemma 1 a split is only useful as a common split if it decomposes the `processMatrix` in such a way that the spectra of the submatrices are disjoint.

If there is no such a subdiagonal element, on the other hand, we will apply one step of the *QR* algorithm (using the routine `qr` of `MATLAB`) and repeat the above process up to reach a maximal number of iterations (depending on the order of the matrix in `processMatrix`).

```
if(lessThanTolerance==0)
    [Q,R]=qr(processMatrix);
    storingQ=auxQ*Q;
    auxQ=storingQ;
    processMatrix=R*Q;
    matrix(p:q,p:q)=processMatrix;
end -end if
if(lessThanThreshold==0)
    [processMatrix,storingQ1]=choosingShift(processMatrix);
    storingQ=auxQ*storingQ1;
    auxQ=storingQ;
    matrix(p:q,p:q)=processMatrix;
end -end if
```

If after a certain number of iterations which depends on the order of the matrix in `processMatrix` we still did not obtain a split, we pick the next unused matrix from **A**, store it in `processMatrix` and restart the iteration.

Step 7. Storing the matrix obtained in the previous process. In a variable called `storingMatrix` of order n, we store the `processMatrix` at the same relative position with respect to the `initialMatrix`. Initially the variable `storingMatrix` is a zero matrix.

```
newstoringQ=eye(n);
newstoringQ([p:q],[p:q])=storingQ;
auxStoringQ=auxStoringQ*newstoringQ;
storingMatrix=auxStoringQ;
```

Then we split this `storingMatrix` into two blocks with “break” at the relative position of the subdiagonal element from Step 6 with absolute value less than the variable `threshold` in the `storingMatrix` and compute the spectra of these blocks:

```
totalBlocksNum = totalBlocksNum + 1;
newBlockMatrix=zeros(totalBlocksNum,3);
[takeFirstBlock]=consideredBlock(matrix, first, last);
[takeSecondBlock]=consideredBlock(matrix, first1,last1);
[newBlockMatrix]=doingNBM(v, first, last, first1, last1, takeFirstBlock,
                          takeSecondBlock, totalBlocksNum, oldBlockMatrix)
[endProcess]=spectraIntersection(matrix, last, first1, toleranceEig);
```

If the spectra are disjoint, then the joint triangularization process is essentially finished and we just have to perform some conclusive computations as described in Step 9. If the spectra are not disjoint and we can split the `storingMatrix` into blocks of order 1 or 2, these final blocks contain either a single real value to be checked by the routine `uniqueEigenvalue` or a pair of conjugate complex eigenvalues. Then we carry out an appropriate symmetric permutation in order to separate the spectra of the blocks. So, we are either in the situation of Step 3 again, now with the freshly computed `storingMatrix`, which will then be stored in the variable `processMatrix`, or the procedure pointed out in Step 4 will be applied to the `storingMatrix`, as the variable `processMatrix`, with properly chosen blocks of this matrix to be considered in Step 4. If one of these blocks has only a unique eigenvalue or a block of order two with a couple of conjugate eigenvalues, we do not consider this block in Step 4.

Step 8. Not all the eigenvalues of the process matrix have been located. If we have not found all the eigenvalues of the matrix in spite of performing the *QR* algorithm with shifts the maximal number of iterations given by the order of the variable `processMatrix`, then we apply a step of the *QR* algorithm with shift. If all the subdiagonal elements are greater than the `tolerance`, then we select a new matrix of the family **A**. If there exists a subdiagonal element with absolute value less than the `tolerance`, we switch to Step 6 which we can repeat until we arrive at a maximal number of iterations given by the order of the `processMatrix`. If this maximal number of iterations is exceeded, then we again choose a new matrix of the family **A**. This can be performed cyclically and after each cycle the maximal number of iterations can be raised, if necessary.

Step 9. Computing the ortogonal matrix *P*. In the end, after the iterations have been performed, we calculate the product of all the matrices corresponding to all the intermediate transformations carried out over the `initialMatrix` in order to obtain the orthogonal matrix P mentioned in Lemma 1.

§3. Subroutines

In this section we give more detailed descriptions of the subroutines used in our algorithm. As an easy starting point we list the very simple function that just checks whether a matrix is normal, that is, whether $A^T A = AA^T$.

The subroutine that chooses the shift for the QR step is already more complicated as the choice of the shift depends on whether the matrix is normal and the shift also has to be adapted to the eigenstructure of the lower left 2×2 block of B.

choosingShift

```
[processMatrix,storingQ]=choosingShift(B)
    N=length(B); storingQ=eye(N); a=eig(B([N-1:N],[N-1:N]));
    if( checkingNormalMatrix(B,N)==1 )
        shift1=sqrt(norm(B,’fro’)/N);
        [Q,R]=qr(B-shift1*eye(N));
        storingQ = storingQ * Q;
        processMatrix=R*Q+shift1*eye(N);
    else
        if(abs(imag(a(1)))<1e-10)    % The real case
            if( max( abs( real(a) ) < 1e-10 )
                newShift=norm(B,inf)/N;
                [Q,R]=qr(B-newShift*eye(N));
                storingQ = storingQ * Q;
                processMatrix=R*Q+newShift*eye(N);
            else
                [Q,R]=qr(B-maximum*eye(N));
                storingQ = storingQ * Q;
                [Q,R]=qr( R*Q );
                storingQ = storingQ * Q;
                processMatrix=R*Q+maximum*eye(N);
            end -end if
        else    % The complex case
            if(N =2)
                [Q,R]=qr(B-a(1)*eye(N));
                storingQ = storingQ * Q;
                [Q,R]=qr( R*Q + ( a(1) -a(2) ) * eye(N));
                processMatrix=R*Q+a(2)*eye(N);
                storingQ = storingQ * Q;
            else
                processMatrix=B;
            end -end if
        end -end if
    end -end if
```

The next subroutine, `consideredBlock`, checks whether a given block of a matrix needs to be considered for the *QR* method. This is the case if the block is not of size 1 or does not have a unique single eigenvalue, the latter being checked by the subroutine `uniqueEigenvalue` which does a quick check whether that is the case.

consideredBlock

```
[takeFirstBlock]=consideredBlock(matrix, first, last)
    firstBlock=matrix(first:last,first:last); takeFirstBlock=1;
    l1=length(firstBlock);
    if l1==1
        takeFirstBlock=0;
    elseif uniqueEigenvalue(firstBlock,last-first+1,threshold,toleranceEig) == 1
        takeFirstBlock=0;
    else
        if l1==2 && eig(firstBlock)  = 0
            takeFirstBlock=0;
        end -end if
    end -end elseif
```

In `doingNBM`, the block matrices are extracted and the so far unused ones “advance” one position in the queue.

doingNBM

```
[newBlockMatrix]=doingNBM(v,first, last, first1, last1, takeFirstBlock,
                        takeSecondBlock, totalBlocksNum, oldBlockMatrix)
    if(v==1)
        newBlockMatrix(1,:)=[first,last,takeFirstBlock];
        newBlockMatrix(2,:)=[first1,last1,takeSecondBlock];
        if(totalBlocksNum>2)
            newBlockMatrix(3:totalBlocksNum,:)=oldBlockMatrix(2:totalBlocksNum-1,:);
        end -end if
    else
        newBlockMatrix(1:v-1,:)=oldBlockMatrix(1:v-1,:);
        newBlockMatrix(v+2:totalBlocksNum,:)=oldBlockMatrix(v+1:totalBlocksNum-1,:);
        newBlockMatrix(v,:)=[first,last,takeFirstBlock];
        newBlockMatrix(v+1,:)=[first1,last1,takeSecondBlock];
    end -end if
```

The subroutine `Gerschgorin` estimates the spectrum of A by means of Gerschgorin circles, cf. [2].

Gerschgorin

```
[dif,maxi,mini]=Gerschgorin(A)
    n = length( A );
    r = abs( A ) * ones( n,1 ) - abs( diag( A ) );
    maxi = max( diag(A) + r );
    mini = min( diag(A) - r );
    dif = maxi - mini;
```

In `hessenberg` the matrix is transformed into Hessenberg form by means of Householder transforms, see again [2]; the extra part is to apply the same transformations to the variable `storedMatrix` that has to be processed in the rest of the method.

hessenberg

```
[H]=hessenberg(A)
    n=length(A);
    storedMatrix=A;
    for i=1:n-2
        v=zeros(n-i,1);
        if(sign(storedMatrix(i+1,i))>=0)
            v=norm(storedMatrix(i+1:n,i),2)*eye(n-i,1)+storedMatrix(i+1:n,i);
        else
            v=-norm(storedMatrix(i+1:n,i),2)*eye(n-i,1)+storedMatrix(i+1:n,i);
        end -end if
        v=v/(norm(v,2));
        P=eye(n-i)-2*v*v';
        U=eye(n);
        U([i+1:n],[i+1:n])=P;
        storedMatrix=U*storedMatrix*U';
    end -end for
    H=storedMatrix;
```

In `means` the difference between the algebraic and geometric means of a matrix is determined; in a really performant version of the algorithm (which will, of course, not use `matlab`), the computation of the determinant can be done more efficiently.

means

```
[dif]=means(A)
    n=length(A);
    mA=abs( trace(A)/n );
    mG=abs( det(A) )^(1/n);
    dif=abs( mA - mG );
```

The shifted *QR* method as in `qrShift` is simply standard.

qrShift

```
[B,E]=qrShift(A,a,n)
    [Q,R]=qr(A-a*eye(n));
    E=Q;
    B=R*Q+a*eye(n);
```

The subroutine `uniqueEigenvalue` is slightly more tricky as it tries to figure out whether a matrix A has only one, unique, real eigenvalue or, a single complex conjugate pair of eigenvalues. Also keep in mind that we assume that any input A will be in Hessenberg form so that all the computations below are relatively cheap. The first step it to consider the number $a = (1/n)\,\mathrm{trace}\,A$ which would be a guess for the single real eigenvalue and the real part of the complex eigenvalue. To check whether a is indeed a single real eigenvalue, we perform one *QR* step on a with shift a. If a were such an eigenvalue, the resulting matrix would be upper triangular with a on the triangle.

If that is not the case, we have to check whether there is a single complex conjugate pair $a \pm ib$ of eigenvalues which requires that the size n of the matrix is even. Then $\det A = (a^2+b^2)^{n/2}$ and we can guess b via $b^2 = (\det A)^{2/n} - a^2$, where for the determinant computation we can re–use the matrix R from the real *QR* decomposition. The we perform either a real *QR double step*, cf. [2], with the guessed eigenvalue $a + ib$ and check whether the result is block diagonal, or, for simplicity, we can do a *complex QR* step and again check for diagonality of the resulting complex matrix.

uniqueEigenvalue

```
[a]=uniqueEigenvalue(A,n,toleranceE)
    a=0;
    possibleEig=(trace(A))/n;
    [Q,R]=qr( A - possibleEig * eye(n) );
    auxMatrix = R * Q + possibleEig * eye(n);
    if ( norm( tril( auxMatrix - possibleEig * eye( n ) ) ) < toleranceE )
```

```
        a = 1;
    elseif rem( n,2 ) == 1
        a = 0;
    else   % Check for complex
        possibleIm = sqrt( prod( diag(R).^2 )^(1/n) - possibleEig^2 );
        possibleEig = possibleEig + I * possibleIm;
        [Q,R]=qr( A - possibleEig * eye(n) );
        auxMatrix= R*Q + possibleEig * eye(n);
        if ( norm( tril( auxMatrix - possibleEig * eye( n ) ) ) < toleranceE )
            a = 1;
        end
    end
```

§4. Acknowledgement

Work partially supported by the Spanish Research Grant MTM2009-07315 and by Gobierno de Aragón and Fondo Social Europeo.

References

[1] Cortés, V., Peña, J. M., and Sauer, T. Simultaneous triangularization of commuting matrices for the solution of polynomial equations. *Central European Journal of Mathematics 10*, 1 (2011), 277–291.

[2] Golub, G., and van Loan, C. F. *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[3] Wilkinson, J. H. *The Algebraic Eigenvalue Problem.* Oxford University Press, 1965.

Vanesa Cortés and Juan Manuel Peña
Departamento de Matemática Aplicada
Universidad de Zaragoza
E-50009 Zaragoza, Spain
`vcortes@unizar.es`, `jmpena@unizar.es`

Tomas Sauer
Lehrstuhl für Numerische Mathematik
Justus–Liebig–Universität Gießen
Heinrich–Buff–Ring 44
D-35392 Gießen, Germany
`tomas.sauer@math.uni-giessen.de`

Monografías Matemáticas *García de Galdeano* **37**, 115–121 (2012)

High-precision periodic orbit corrector

Ángeles Dena, Alberto Abad and Roberto Barrio

Abstract. An algorithm to compute periodic orbits of dynamical systems up to an arbitrary number of precision digits is presented. The algorithm is based on an optimized Newton-Raphson method combined with a new numerical ODE solver, TIDES that uses a Taylor series method. Finally, we present some numerical tests for the Lorenz model and the Hénon-Heiles Hamiltonian which show the quadratic convergence and the good behaviour of the proposed method.

Keywords: Periodic orbits, shooting method, Taylor series method, TIDES.

AMS classification: 37M20, 65P20.

§1. Introduction

Nowadays, more and more theoretical and applied problems need high-precision results. In Dynamical Systems we may find a large plethora of such problems, like studying the exponentially small splitting of separatrices, in the analysis of SNAs, in the study of complex singularities of systems like Lorenz model, and so on. Studying and locating the periodic orbits of dynamical systems give relevant information. So, the periodic orbits are an important topic in several physical applications and finding them accurately is of great importance in periodic orbit theory [2, 3, 4]. In this paper we propose a new algorithm to locate periodic orbits up to any arbitrary precision.

The only algorithm known on the literature capable of computing periodic orbits accurate and highly convergent is the method proposed by D. Viswanath [6] that is based on the Lindstedt-Poincaré technique. To introduce the problem and the new method here proposed, we describe briefly the Viswanath's technique. The problem is to find an isolated orbit of the dynamical system $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x})$, with $\mathbf{x} \in \mathbb{R}^n$. Rescaling time using $\tau = \omega t$, we have the following one $\omega\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x})$. The starting guesses, ω_0 and $\mathbf{x}_0(\tau)$ must be sufficiently close to the periodic orbit. The aim is to improve approximations for ω_i and $\mathbf{x}_i(\tau)$ and each iteration is made up of a sequence of steps. Let $\omega_0\dot{\mathbf{y}}(\tau) = A(\tau)\mathbf{y} + \mathbf{r}(\tau) - \delta\omega\dot{\mathbf{x}}_0(\tau)$ be the *correction equation*, then compute the Fourier series for all n^2 entries of $A(\tau)$, n Fourier series for the residual $\mathbf{r}(\tau)$ and another n Fourier series for $\dot{\mathbf{x}}_0(\tau)$. The general solution of the above equation is written as, $\mathbf{y}(\tau) = Y(\tau)\mathbf{y}(0) + f_1(\tau) - \delta\omega f_2(\tau)$, where $Y(\tau)$ is the Fundamental solution of $\omega_0\dot{\mathbf{y}}(\tau) = A(\tau)\mathbf{y}$. We take into consideration that $Y(\tau)$, $f_1(\tau)$ and $f_2(\tau)$ are computed by using an accurate ODE solver in double precision. So, to obtain an arbitrary precision periodic orbit this algorithm uses several numerical techniques in a sophisticated way to use just double precision in the numerical integration of the ODE system.

As remarked, the method of D. Viswanath avoids the use of the integration of ODEs in multiple precision, but at the price of using a complicated algorithm. Therefore, we have tried to develop a new algorithm for computing periodic orbits using a multiple precision ODE integrator. This method is described in the next section. Our algorithm is based on an optimized shooting method combined with TIDES (Taylor Integrator for Differential EquationS). This tool is an accurate numerical ODE integrator which allows us to integrate in multiple precision arithmetic. We remark that nowadays this method, the Taylor series method, is the only capable method to integrate and ODE system up to any desired precision level (any Runge-Kutta or similar numerical method for ODEs cannot be used for such a high-precision).

§2. The corrector algorithm

Let

$$\mathbf{x} = \mathbf{x}(t; \mathbf{y}), \quad t \in \mathbb{R}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \tag{1}$$

be the solution of the autonomous differential system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}); \quad \mathbf{x}(0) = \mathbf{y}, \quad \mathbf{x} \in \mathbb{R}^n, \tag{2}$$

where $\mathbf{y}$ represents the initial conditions.

The solution of (2) is periodic if it verifies the periodicity condition

$$\mathbf{x}(T, \mathbf{y}) - \mathbf{y} = 0. \tag{3}$$

The Newton method is a common procedure to find the roots of this equation. Our algorithm is an iterative scheme that begins with a set $(\mathbf{y}_0, T_0)$ of approximate initial conditions. At each iteration we update the initial conditions $(\mathbf{y}_i, T_i)$ by adding them the corrections ($\Delta\mathbf{y}_i$, ΔT_i) that are obtained by expanding

$$\mathbf{x}(T_i + \Delta T_i; \mathbf{y}_i + \Delta\mathbf{y}_i) - (\mathbf{y}_i + \Delta\mathbf{y}_i) = 0,$$

in a Taylor series up to the first order

$$\mathbf{x}(T_i; \mathbf{y}_i) - \mathbf{y}_i + \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} - I\right)\Delta\mathbf{y}_i + \left(\frac{\partial \mathbf{x}}{\partial t}\right)\Delta T_i = 0. \tag{4}$$

The $n \times n$ matrix $\partial\mathbf{x}/\partial\mathbf{y}$ is the fundamental matrix, i.e. the solution of the variational equations. This matrix evaluated at $(\mathbf{y}_i, T_i)$ is an approximation of the monodromy matrix M. I is the identity matrix of order n. The column vector $\partial\mathbf{x}/\partial\mathbf{t}$ represents the derivative of the solution with respect to the time, i.e., $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$. This vector, evaluated in $(\mathbf{y}_i, T_i)$ corresponds to the expression $\mathbf{f}(\mathbf{y}_{T_i})$, where $\mathbf{y}_{T_i} = \mathbf{x}(T_i, \mathbf{y}_i)$. To do that, we use the accurate numerical ODE integrator TIDES [1] that computes simultaneously both, the solution and the partial derivatives of the solutions of (2). So, the previous equation is equivalent to the next one

$$(M - I)\Delta\mathbf{y}_i + \mathbf{f}(\mathbf{y}_{T_i})\Delta T_i = -(\mathbf{y}_{T_i} - \mathbf{y}_i). \tag{5}$$

To solve this linear system, it must take into account that varying $\Delta\mathbf{y}_i$ along the periodic orbit gives different representations of the same periodic orbit. Therefore, we impose the additional requirement that $\Delta\mathbf{y}_i$ must be orthogonal to the vector field at $\mathbf{y}_i$; i.e.,

$$\langle \mathbf{f}(\mathbf{y}_i), \Delta\mathbf{y}_i \rangle = 0. \tag{6}$$

2.1. Dissipative case

Equations (5) and (6) are written in a matrix form of dimension $(n+1)\times(n+1)$,

$$\begin{pmatrix} M - I & \mathbf{f}(\mathbf{y}_{T_i}) \\ (\mathbf{f}(\mathbf{y}_i))^T & 0 \end{pmatrix} \begin{pmatrix} \Delta\mathbf{y}_i \\ \Delta T_i \end{pmatrix} = \begin{pmatrix} \mathbf{y}_i - \mathbf{y}_{T_i} \\ 0 \end{pmatrix}. \tag{7}$$

In order to obtain the corrections $\Delta\mathbf{y}_i$ and ΔT_i, we use an iterative scheme which solves the linear system by using the Singular Value Decomposition Algorithm (SVD) [5] although it may use any known solver method for linear systems as the matrix is a non-singular square matrix.

2.2. Hamiltonian case

When the differential system (2) admits one or more integrals, a new constrain or vector of constrains, respectively, must be added to the periodicity condition (3). To maintain the new constrain, $\mathbf{G}(t;\mathbf{x}) = \mathbf{g}$, we impose the condition

$$\mathbf{G}(T_i + \Delta T_i; \mathbf{y}_i + \Delta\mathbf{y}_i) - \mathbf{g} \approx \mathbf{G}(T_i;\mathbf{y}_i) - \mathbf{g} + \left.\frac{\partial\mathbf{G}}{\partial\mathbf{x}}\right|_{(T_i;\mathbf{y}_i)} \Delta\mathbf{y}_i + \left.\frac{\partial\mathbf{G}}{\partial t}\right|_{(T_i;\mathbf{y}_i)} \Delta T_i = 0.$$

In a Hamiltonian problem we have the integral of energy $\mathcal{H}(\mathbf{x}) = H$. So, in this case, we add the following condition to the above linear system,

$$(\nabla_{\mathbf{x}}\mathcal{H})|_{(T_i;\mathbf{y}_i)}\, \Delta\mathbf{y}_i + (\mathcal{H}_t)|_{(T_i;\mathbf{y}_i)}\, \Delta T_i = H - H_{T_i}.$$

Taking into account that the Hamiltonian does not depend on the time, the second term of the addition is cancelled. So, the constrain condition has the form

$$(\nabla_{\mathbf{x}}\mathcal{H})|_{(T_i;\mathbf{y}_i)}\, \Delta\mathbf{y}_i = H - H_{T_i}. \tag{8}$$

Hamiltonian condition (8) is computed using TIDES and Mathematica's operator gradient, $\nabla_{\mathbf{x}}\mathcal{H}$. The matrix of the new linear system has dimension $(n+2)\times(n+1)$. So, we wish to find the least-norm solution to an overdetermined set of linear equations and for this, we use the SVD Algorithm for constructing the singular value decomposition of the matrix. Here, we have for the Hamiltonian case the matrix form,

$$\begin{pmatrix} M - I & \mathbf{f}(\mathbf{y}_{T_i}) \\ (\mathbf{f}(\mathbf{y}_i))^T & 0 \\ (\nabla_{\mathbf{x}}\mathcal{H})|_{(T_i;\mathbf{y}_i)} & 0 \end{pmatrix} \begin{pmatrix} \Delta\mathbf{y}_i \\ \Delta T_i \end{pmatrix} = \begin{pmatrix} \mathbf{y}_i - \mathbf{y}_{T_i} \\ 0 \\ H - H_{T_i} \end{pmatrix}. \tag{9}$$

§3. ODE's, partial derivatives and multiple precision with TIDES

To compute the correction, as well as to solve the linear system (7) and (9), we have to compute the matrix of the systems. For that, we need to integrate the ODE (2) and to compute the partial derivatives of its solution (1) with respect to the initial condition $\mathbf{y}$. To do that we

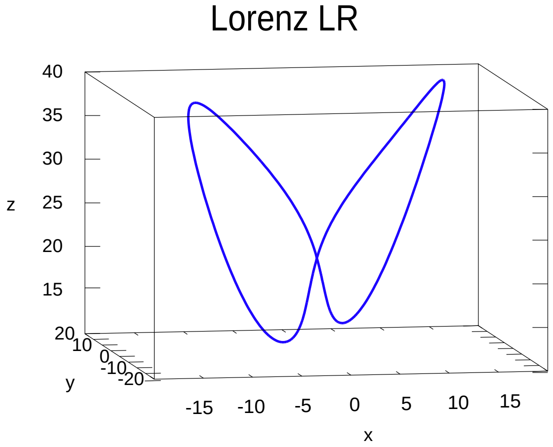


Figure 1: The periodic orbit LR of the Lorenz model.

use the software TIDES [1], that consists of a C library and a Mathematica precompiler that writes a C program which permits to compute simultaneously both, the solution and the partial derivatives of the solution of (2), in double or multiple precision, by using the Taylor Series Method (TSM).

Usually, the matrix of partial derivatives $\mathbf{\Phi} = \partial\mathbf{x}/\partial\mathbf{y}$ of the solution with respect to the initial condition is computed by using the variational equations $\dot{\mathbf{\Phi}} = (\partial\mathbf{f}/\partial\mathbf{x}) \cdot \mathbf{\Phi}$, that are different for each problem and sometimes very difficult to formulate. In TIDES, instead of formulate the variational equations, we use the Taylor series expression

$$\mathbf{x}(t) = \sum_i \mathbf{x}^{[i]}\, h^i, \qquad h = t - t_0, \qquad \mathbf{x}^{[i]} = \frac{1}{i!}\frac{d\mathbf{x}^{(i)}(t_0)}{dt^i},$$

to create iterative formulas to compute simultaneously both, the solution and the partial derivatives. This simplifies the process and permits to extend it to any differential equation and work with any precision without difficulties. Obviously, to use the Taylor series method the second member of the differential equations has to be a smooth enough function.

§4. Tests

This method has proved its applicability with two paradigmatic examples, Lorenz model and Hénon-Heiles Hamiltonian. The classical Lorenz model is given by the ordinary differential equation

$$\dot{x} = \sigma(y - x), \qquad \dot{y} = -xz + rx - y, \qquad \dot{z} = xy - bz. \tag{10}$$

In this work, we will take the classical Saltzman values of the parameters $b = 8/3$, $\sigma = 10$ and $r = 28$ and the initial conditions $(x, y, z) = (-13.764, -19.579, 27)$ and a period $T = 1.5586$ (with just five correct digits). So, we have computed the LR periodic orbit up to one hundred

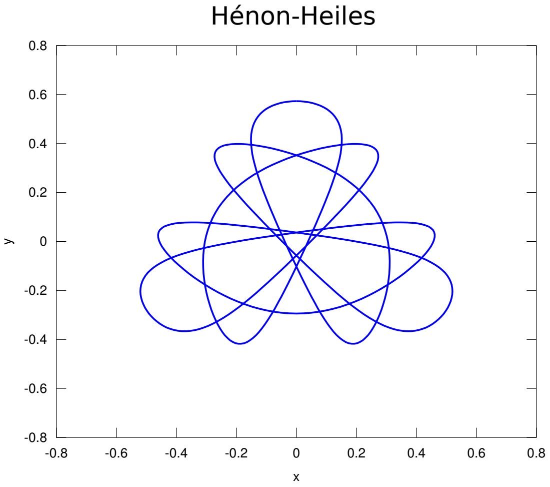


Figure 2: Stable orbit of the Hénon-Heiles problem.

digits of precision and we have obtained the next corrected initial conditions:

$$
\begin{aligned}
x &= -13.763809685186058958073230618459671664631238848297 7\\
&\quad 26222500121342876008079691601274879478926826271846,\\
y &= -19.578732026230613926743618660803430026955625649678 3\\
&\quad 66597735394648946838029436937301740808647462 61638,\\
z &= 27.000675803239826810615080341095213706029740774444 11\\
&\quad 86706712936762835283686545722164080192144099 6386,\\
T &= 1.558652210716174727567870209212696070528480548997 2433\\
&\quad 935889521578319019875625888085435558510826601 4236.
\end{aligned}
$$

It is well known that the chaotic attractor of the Lorenz model presents the shape of the wings of a butterfly. There are infinite unstable periodic orbits foilated to this attractor. Emphasize that the periodic orbit is labelled LR (see Figure 1) to indicate the sequence in which it moves, so it does one loop on the left and another one on the right.

On the other hand, the Hénon-Heiles problem is given by the Hamiltonian:

$$\mathcal{H}((x,y),(X,Y)) = \frac{1}{2}(X^2+Y^2) + \frac{1}{2}(x^2+y^2) + x^2y - \frac{1}{3}y^3. \tag{11}$$

where (x,y) and (X,Y) represent the position and velocity vectors, respectively. In Figure 2, we show a stable periodic orbit with initial conditions $(x, y, X, Y) = (0, 0.5729, 0.2171, 0)$ and period $T = 32.378$ (with just five correct digits). Therefore, we have computed this stable periodic orbit of the Hénon-Heiles problem up to one hundred digits of precision and we have

Lorenz		Hénon-Heiles	
No. of iterations	$-\log_{10}\lvert Error\rvert$	No. of iterations	$-\log_{10}\lvert Error\rvert$
1	2.7520	1	3.6921
2	5.1912	2	6.8717
3	11.8648	3	12.6562
4	23.4265	4	25.3239
5	48.3370	5	49.9962
6	96.2892	6	98.9237

Table 1: Error estimates in each iteration of the algorithm.

obtained the next corrected initial conditions:

$$\begin{aligned}
x &= -0.00015085289594894024496799412284793291748395269910754456347530550064296127893491578491858310406310932 37,\\
y &= 0.57295301370224350348504778098660473159863485033171053200607809433680990659867827743286884431646312 43,\\
X &= 0.21706126541161424223171335987599024998866524355885647295531449799555593630239513190046828312316602 122,\\
Y &= 0.00017004577430097839491728371273170373793851553739810493718722386095558824928331758251476500580851 36020,\\
T &= 32.37774034214117077101749261854234714537204730508816304777025017758227599170926401377549088558254881.
\end{aligned}$$

The computational complexity of the numerical solution of an ODE system using a TSM as TIDES with $D = -\log_{10}(TOL)$ number of digits is $O(D^4)$, using variable-precision arithmetic up to one hundred digits of precision, variable-order and variable-stepsize. Moreover, it is well known that the Newton method has quadratic convergence, so the previous algorithm which has been presented in the second section, is quadratically convergent too. We achieve the preset tolerance in six iterations with about one hundred digits of fixed precision arithmetic for both, the Lorenz model and the Hénon-Heiles Hamiltonian. As we can see in the Table 1, the number of digits of precision in the initial conditions of the periodic orbits is doubled at each iteration.

Acknowledgements

The authors have been supported for this research by the Spanish Research Grant MTM2009-10767.

References

[1] Abad, A., Barrio, R., Blesa, F., and Rodríguez, M. TIDES: a Taylor series Integrator of Differential EquationS. http://gme.unizar.es/software/tides (2010).

[2] Barrio, R., Blesa, F., and Serrano, S. Fractal structures in the Hénon-Heiles Hamiltonian. *Europhysics Letters 82* (2008), 10003.

[3] Barrio, R., Blesa, F., and Serrano, S. Bifurcations and safe regions in open Hamiltonians. *New Journal of Physics 11* (2009), 053004.

[4] Barrio, R., and Serrano, S. Bounds for the chaotic region in the Lorenz model. *Phys. D 238* (2009), 1615–1624.

[5] Dieci, L., Gasparo, M., and Papini, A. Path following by SVD. *Computational Science* (2006), 677–684.

[6] Viswanath, D. The Lindstedt-Poincaré technique as an algorithm for computing periodic orbits. *SIAM Rev. 43* (2001), 478–495.

A. Dena and R. Barrio
Department of Applied Mathematics and IUMA
University of Zaragoza
adena@unizar.es and rbarrio@unizar.es

A. Abad
Department of Theoretical Physics and IUMA
University of Zaragoza
abad@unizar.es

Monografías Matemáticas *García de Galdeano* **37**, 123–132 (2012)

Construction of majorizing sequences for operators with unbounded second derivative

J. A. Ezquerro, D. González and M. A. Hernández

Abstract. The aim of this paper is to construct majorizing sequences for Newton's method in Banach spaces, when the second Fréchet derivative of the operator involved is unbounded, and prove then the semilocal convergence of the method. The new results are illustrated with a nonlinear integral equation of mixed Hammerstein type.

Keywords: Newton's method, semilocal convergence, majorizing sequence, Hammerstein's integral equation.

AMS classification: 45G10, 47H99, 65J15.

§1. Introduction

We present a study for approximating a solution x^* of the equation

$$F(x) = 0, \tag{1}$$

where F is a nonlinear operator defined on a non-empty open convex subset Ω of a Banach space X with values in a Banach space Y, by the most famous iterative method, Newton's method, whose algorithm is:

$$x_{n+1} = x_n - [F'(x_n)]^{-1}F(x_n), \quad n = 0, 1, 2, \dots, \tag{2}$$

where the starting point x_0 is given.

The generalization of Newton's method to Banach spaces is due to the Russian mathematician L. V. Kantorovich, who publishes several papers at the mid-twentieth century. Initially, see [3] , Kantorovich proves the semilocal convergence of Newton's method under the conditions: $\|\Gamma_0\| \leq \beta$, $\|\Gamma_0 F(x_0)\| \leq \eta$ and

$$\|F''(x)\| \leq K, \quad x \in \Omega, \tag{3}$$

where it is supposed that the operator $\Gamma_0 = [F'(x_0)]^{-1} \in \mathcal{L}(Y, X)$ exists at some $x_0 \in \Omega$, where $\mathcal{L}(Y, X)$ is the set of bounded linear operators from Y into X. The great majority of the results appearing in the literature are concerning with the need for the operator F'' to be bounded in the domain Ω, where the solution x^* must exist. According to this, the number of equations that can be solved by Newton's method is limited. For instance, we cannot analyse the convergence of Newton's method to a solution of an equation where the second derivative of the operator involved is not bounded in a domain, what usually happens in some nonlinear integral equations of mixed Hammerstein type [2]; i.e.:

$$x(s) = u(s) + \sum_{i=1}^{m} \int_a^b G_i(s,t)H_i(x(t))\,dt, \quad s \in [a,b], \tag{4}$$

where $-\infty < a < b < \infty$, G_i, H_i $(i = 1,2,\dots,m)$ and u are known functions and x is a continuous function (solution) to be determined. In particular, for nonlinear integral equations of the form

$$x(s) = u(s) + \int_a^b G(s,t)[x(t)^{2+p} + \frac{1}{2}x(t)^2]\,dt, \quad s \in [a,b], \tag{5}$$

with $p \in [0, 1]$, where u is a continuous function and the kernel G is the Green function

$$G(s,t) = \begin{cases} \dfrac{(b-s)(t-a)}{b-a}, & t \le s, \\ \dfrac{(s-a)(b-t)}{b-a}, & s \le t. \end{cases}$$

Integral equations of this type can be found in the dynamic model of a chemical reactor, which is governed by a control equation and justify the analysis and computation of mixed Hammerstein equations [1].

Solving nonlinear integral equation (5) is equivalent to solve (1), where

$$F : \Omega \subseteq C[a,b] \longrightarrow C[a,b], \quad \Omega = \{x \in C[a,b] :\ x(s) > 0,\ s \in [a,b]\},$$

$$[F(x)](s) = x(s) - u(s) - \int_a^b G(s,t)[x(t)^{2+p} + \frac{1}{2}x(t)^2]\,dt, \quad p \in (0,1].$$

Taking into account the expression of F, it follows

$$[F'(x)y](s) = y(s) - \int_a^b G(s,t)[(2+p)x(t)^{1+p} + x(t)]y(t)\,dt,$$

$$[F''(x)(yz)](s) = -\int_a^b G(s,t)[(2+p)(1+p)x(t)^p + 1]z(t)y(t)\,dt. \tag{6}$$

Notice that condition (3) is not satisfied since $\|F''(x)\|$ is not bounded in all Ω. To see this, we use *reductio ad absurdum*. We suppose $\|F''(x)\| \le K$ in Ω for the max-norm and denote $M = \max_{[a,b]} \int_a^b |G(s,t)|\,dt$. Then, if $x(t) = \big((K - M + \epsilon)/(M(2+p)(1+p))\big)^{1/p}$, with $\epsilon \in (M-K, +\infty)$ if $M > K$ or $\epsilon \in (0, +\infty)$ if $M \le K$, and $y(t) = z(t) = 1$, it follows that

$$\begin{aligned} \|[F''(x)(yz)](s)\| &= \left\| \int_a^b G(s,t)[(2+p)(1+p)x(t)^p + 1]\,dt \right\| \\ &= \left\| \frac{K+\epsilon}{M} \int_a^b G(s,t)\,dt \right\| = K + \epsilon > K. \end{aligned}$$

Thus, the last is contradictory to the given statement, since there does not exist a constant K such that $\|F''(x)\| \leq K$ in all Ω. To solve the last, we can use an elegant alternative which consists of relaxing condition (3) by the following one:

$$\|F''(x)\| \leq \omega(\|x\|), \quad x \in \Omega, \tag{7}$$

where $\omega : \mathbb{R}_+ \cup \{0\} \longrightarrow \mathbb{R}$ is a continuous non-decreasing real function.

In this paper, we prove the semilocal convergence of Newton's method under condition (7) instead of condition (3) and illustrate the new result with a nonlinear integral equation of mixed Hammerstein type. The results and their proofs are given in Banach spaces and based on the concept of majorizing sequence:

> Let $\{x_n\}$ be a sequence in a Banach space X and $\{t_n\}$ a scalar sequence. The sequence $\{t_n\}$ majorizes to the sequence $\{x_n\}$ if
>
> $$\|x_{n+1} - x_n\| \leq t_{n+1} - t_n, \quad n = 0, 1, 2, \ldots$$

Emphasize that the interest of majorizing sequences is that the convergence of the sequence in Banach spaces is deduced from the convergence of the scalar sequence, as we can see in the following result [3]:

> Let $\{x_n\}$ be a sequence in a Banach space X and $\{t_n\}$ a scalar majorizing sequence of $\{x_n\}$. If $\{t_n\}$ converges to $t^* < \infty$, there exists $x^* \in X$ such that $x^* = \lim_n x_n$ and $\|x^* - x_n\| \leq t^* - t_n$, for $n \geq 0$.

Throughout the paper we denote $\overline{B(x,\rho)} = \{y \in X : \|y - x\| \leq \rho\}$ and $B(x,\rho) = \{y \in X : \|y - x\| < \rho\}$.

§2. Semilocal convergence

Once the definition of majorizing sequence is introduced, Kantorovich establishes the semilocal convergence of Newton's method under the conditions $\|\Gamma_0\| \leq \beta$, $\|\Gamma_0 F(x_0)\| \leq \eta$ and (3), so that the semilocal convergence of Newton's method is then guaranteed from the quadratic polynomial (see [3])

$$f(t) = \frac{K}{2}(t - t_0)^2 - \frac{t - t_0}{\beta} + \frac{\eta}{\beta}$$

and the scalar sequence $\{t_n\}$,

$$t_{n+1} = t_n - \frac{f(t_n)}{f'(t_n)}, \quad n = 0, 1, 2, \ldots, \tag{8}$$

which majorizes sequence (2).

The main aim of this paper is to present a new version of the Kantorovich study, where condition (3) is relaxed by condition (7). Specifically, we suppose

(C_1) There exists $x_0 \in \Omega$ such that the operator $\Gamma_0 = [F'(x_0)]^{-1}$ is well-defined and $\|\Gamma_0\| \leq \beta$,

(C_2) $\|\Gamma_0 F(x_0)\| \leq \eta$,

(C_3) $\|F''(x)\| \leq \omega(\|x\|)$, $x \in \Omega$, where $\omega : \mathbb{R}_+ \cup \{0\} \to \mathbb{R}$ is a continuous real non-decreasing function.

If we follow a similar way to Kantorovich, we cannot consider a quadratic polynomial to define the scalar majorizing sequence, since condition (C_3) does not permit it. So, from (C_1)–(C_3), we can construct the function

$$f(t) = \int_{t_0}^{t} \int_{t_0}^{\theta} \omega(\xi)\, d\xi\, d\theta - \frac{t - t_0}{\beta} + \frac{\eta}{\beta}, \quad t_0 \geq 0, \tag{9}$$

where ω is the function defined in (7).

Before establishing the new semilocal convergence of Newton's method, we give some previous results that are needed. Lemmas 1 and 2 are technical and the proofs follow immediately.

Lemma 1. *Let ω and f be the real functions defined in (7) and (9), respectively. Then:*

a) If there exists a solution $\alpha > 0$ of the equation

$$W(t) - W(t_0) - \frac{1}{\beta} = 0, \tag{10}$$

where W is a primitive for ω in $\mathbb{R}_+$, then α is the unique minimum of f in $\mathbb{R}_+$.

b) The function f is non-increasing in (t_0, α),

c) If $f(\alpha) \leq 0$, then equation $f(t) = 0$ has at least one solution in $\mathbb{R}_+$. Moreover, if we denote the smallest positive root of $f(t) = 0$ by t^*, we have $t^* \in (t_0, \alpha]$.

Lemma 2. *Let (8) with $f(t)$ defined in (9). Suppose that there exists a positive root α of (10) such that $f(\alpha) \leq 0$. Then, $\{t_n\}$ is a non-decreasing sequence that converges to t^*.*

Next, we prove that sequence $\{x_n\}$ is well-defined. To do this, firstly, we see that $x_1 \in B(x_0, t^* - t_0)$; and secondly, if we assume that $B(x_0, t^* - t_0) \subseteq \Omega$, it follows that $x_n \in B(x_0, t^* - t_0)$, for all $n = 2, 3, 4, \ldots$

To see that x_1 is well-defined, we take into account that $\Gamma_0 = [F'(x_0)]^{-1}$ and $\|\Gamma_0\| \leq -1/f'(t_0) = \beta$ and $\|x_1 - x_0\| \leq \eta = t_1 - t_0$, so that $x_1 \in B(x_0, t^* - t_0)$. In the following result we see that $x_n \in B(x_0, t^* - t_0)$ and $\{t_n\}$ is a majorizing sequence.

Lemma 3. *Let F be a nonlinear twice continuously differentiable operator defined on a non-empty open convex domain Ω of a Banach space X with values in a Banach space Y. We suppose that conditions (C_1)–(C_3) hold and $f(\alpha) \leq 0$, where $f(t)$ is defined in (9) and α is a solution of (10), $\|x_0\| \leq t_0$ and $B(x_0, t^* - t_0) \subseteq \Omega$. Then, $x_n \in B(x_0, t^* - t_0)$, for all $n \in \mathbb{N}$. Moreover, the sequence $\{t_n\}$ defined in (8) majorizes to the sequence $\{x_n\}$ defined in (2); i.e: $\|x_{n+1} - x_n\| \leq t_{n+1} - t_n$ with $n = 0, 1, 2, \ldots$*

Proof. Firstly, by the Banach lemma, observe that there exists $\Gamma_1 = [F'(x_1)]^{-1}$ and $\|\Gamma_1\| \leq$

$-1/f'(t_1)$, since $\|I - \Gamma_0 F'(x_1)\| < 1$. Indeed,

$$\begin{aligned}
\|I - \Gamma_0 F'(x_1)\| &= \left\| \int_{x_0}^{x_1} \Gamma_0 F''(x)\, dx \right\| = \left\| \int_0^1 \Gamma_0 F''(x_0 + t(x_1 - x_0))(x_1 - x_0)\, dt \right\| \\
&\le \|\Gamma_0\| \int_0^1 \|F''(x_0 + t(x_1 - x_0))\|\, \|x_1 - x_0\|\, dt \le \beta(t_1 - t_0) \int_0^1 \omega(\|x_0 + t(x_1 - x_0)\|)\, dt \\
&\le \beta(t_1 - t_0) \int_0^1 \omega(t_0 + t(t_1 - t_0))\, dt = 1 - \frac{f'(t_1)}{f'(t_0)} < 1,
\end{aligned}$$

since $\omega(t) = f''(t)$ and ω is a non-decreasing function. Therefore,

$$\| [\Gamma_0 F'(x_1)]^{-1} \| \le \frac{f'(t_0)}{f'(t_1)} \qquad \text{and} \qquad \|\Gamma_1\| \le \| [\Gamma_0 F'(x_1)]^{-1} \|\, \|\Gamma_0\| \le -\frac{1}{f'(t_1)}.$$

Secondly, since $\|x_0\| \le t_0$, then $\|x_1\| \le \|x_1 - x_0\| + \|x_0\| \le t_1$, then $\|x_1\| \le t_1$.

Thirdly, the Taylor series expansion of $F(x)$ about x_0 is

$$\begin{aligned}
F(x_1) = F(x_0) + F'(x_0)(x_1 - x_0) + \int_{x_0}^{x_1} F''(x)(x_1 - x)\, dx \\
= \int_0^1 F''(x_0 + \tau(x_1 - x_0))(1 - \tau)(x_1 - x_0)^2\, d\tau,
\end{aligned}$$

so that

$$\begin{aligned}
\|F(x_1)\| &\le \int_0^1 \omega(\|x_0 + \tau(x_1 - x_0)\|)(1 - \tau)\|x_1 - x_0\|^2\, d\tau \\
&\le \int_0^1 \omega(\|x_0\| + \tau\|x_1 - x_0\|)(1 - \tau)\|x_1 - x_0\|^2\, d\tau \\
&\le \int_0^1 \omega(t_0 + \tau(t_1 - t_0))(1 - \tau)(t_1 - t_0)^2\, d\tau = f(t_1),
\end{aligned}$$

since

$$f(t_1) = \int_0^1 f''(t_0 + \tau(t_1 - t_0))(1 - \tau)(t_1 - t_0)^2\, d\tau = \int_0^1 \omega(t_0 + \tau(t_1 - t_0))(1 - \tau)(t_1 - t_0)^2\, d\tau.$$

Fourthly, from $\|\Gamma_1\| \le -1/f'(t_1)$ and $\|F(x_1)\| \le f(t_1)$, it follows that

$$\|x_2 - x_1\| \le \|\Gamma_1 F(x_1)\| \le \|\Gamma_1\|\, \|F(x_1)\| \le -\frac{f(t_1)}{f'(t_1)} = t_2 - t_1.$$

Fifthly, we see that $\|x_2 - x_0\| \le \|x_2 - x_1\| + \|x_1 - x_0\| \le t_2 - t_0$, so that $x_2 \in B(x_0, t^* - t_0)$.

Finally, if we assume, for $n \in \mathbb{N}$, that

$[I_n]$ there exists $\Gamma_n = [F'(x_n)]^{-1}$ and $\|\Gamma_n\| \le -\frac{1}{f'(t_n)}$,

$[II_n]$ $\|x_n\| \le t_n$,

$[III_n]$ $\|F(x_n)\| \leq f(t_n)$,

$[IV_n]$ $\|x_{n+1} - x_n\| \leq t_{n+1} - t_n$,

$[V_n]$ $\|x_{n+1} - x_0\| \leq t^* - t_0$,

it follows in the same way that $[I_{n+1}]$–$[V_{n+1}]$ hold, so that $[I_n]$–$[V_n]$ are true for all positive integers n by mathematical induction. Consequently, (8) is a majorizing sequence of (2). □

We are now ready to prove in the next theorem the semilocal convergence of Newton's method when the operator F satisfies (C_1)–(C_3). The proof of the theorem follows from the previous lemmas.

Theorem 4. *Let F be a nonlinear twice continuously differentiable operator defined on a non-empty open convex domain Ω of a Banach space X with values in a Banach space Y. Suppose that conditions (C_1)–(C_3) are satisfied. If $f(\alpha) \leq 0$, where $f(t)$ is defined in (9), $\|x_0\| \leq t_0$ and $B(x_0, R) \subseteq \Omega$ with $R = t^* - t_0$, then Newton's method (2) converges to a solution x^* of (1). Moreover, $x_n, x^* \in \overline{B(x_0, R)}$, for all $n \in \mathbb{N}$ and $\|x^* - x_n\| \leq t^* - t_n$, $n \geq 0$. If r is the biggest positive root of the equation*

$$\int_R^t \int_{t_0}^{t_0+u} \omega(z)\,dz\,du = \frac{t-R}{\beta}, \tag{11}$$

the solution x^* is unique in $B(x_0, r) \cap \Omega$ if $r > R$ or in $\overline{B(x_0, R)}$ if $r = R$.

Proof. On the one hand, from Lemma 3 and the fact that the scalar sequence $\{t_n\}$ is convergent, it follows that there exists x^* such that $x^* = \lim_n x_n$, since $\{t_n\}$ is a majorizing sequence of $\{x_n\}$, and $x_n, x^* \in \overline{B(x_0, R)}$, for all $n \in \mathbb{N}$.

On the other hand, as

$$\|F(x_n)\| = \|F'(x_n)(x_{n+1} - x_n)\| \leq \|F'(x_n)\|\,\|x_{n+1} - x_n\|$$

and

$$\begin{aligned}\left\|F'(x_n) - F'(x_0)\right\| &= \left\|\int_{x_0}^{x_n} F''(x)dx\right\| \\ &= \left\|\int_0^1 F''(x_0 + t(x_n - x_0))(x_n - x_0)\,dt\right\| \leq \int_0^1 \|F''(x_0 + t(x_n - x_0))\|\,\|x_n - x_0\|\,dt \\ &\leq \int_0^1 \omega(\|x_0 + t(x_n - x_0)\|)\,\|x_n - x_0\|\,dt \leq \omega(t_0 + R)R,\end{aligned}$$

we have,

$$\|F'(x_n)\| \leq \|F'(x_n) - F'(x_0)\| + \|F'(x_0)\| \leq \omega(t_0 + R)R + \|F'(x_0)\|,$$

and consequently $\{\|F'(x_n)\|\}$ is bounded and $\lim_n \|F(x_n)\| = 0$. Now, by the continuity of F, it is clear that x^* is a solution of $F(x) = 0$.

To see the unicity of x^*, when $r > R$, we suppose that y^* is another solution of $F(x) = 0$ in $B(x_0, r) \cap \Omega$. Since

$$0 = F(y^*) - F(x^*) = \int_{x^*}^{y^*} F'(x)dx = \int_0^1 F'(x^* + t(y^* - x^*))(y^* - x^*)dt,$$

it suffices to see that there exists the operator

$$\left[\Gamma_0 \int_0^1 F'(x^* + t(y^* - x^*))dt\right]^{-1}. \tag{12}$$

Indeed, from

$$\begin{aligned} I - \Gamma_0 \int_0^1 F'(x^* + t(y^* - x^*))\,dt &= \Gamma_0 \left[\int_0^1 F'(x_0)dt - \int_0^1 F'(x^* + t(y^* - x^*))\,dt\right] \\ &= -\Gamma_0 \int_0^1 \left(\int_{x_0}^{x^*+t(y^*-x^*)} F''(z)\,dz\right) dt, \end{aligned}$$

if we take norms, we have

$$\begin{aligned} &\left\|I - \Gamma_0 \int_0^1 F'(x^* + t(y^* - x^*))dt\right\| \leq \|\Gamma_0\| \left\|\int_0^1 \int_{x_0}^{x^*+t(y^*-x^*)} F''(z)\,dz\,dt\right\| \\ &\leq \beta \int_0^1 \int_0^1 \left\|F''(x_0 + v((x^* - x_0) + t(y^* - x^*)))((x^* - x_0) + t(y^* - x^*))\right\| \,dv\,dt \\ &\leq \beta \int_0^1 \int_0^1 \left\|F''(x_0 + v((x^* - x_0) + t(y^* - x^*)))\right\| \left\|(x^* - x_0) + t(y^* - x^*)\right\| dv\,dt \\ &\leq \beta \int_0^1 \|(x^* - x_0) + t(y^* - x^*)\| \left(\int_0^1 \left\|F''(x_0 + v((x^* - x_0) + t(y^* - x^*)))\right\| \,dv\right) dt \\ &\leq \beta \int_0^1 ((1-t)\|x^* - x_0\| + t\|y^* - x_0\|) \left(\int_0^1 \omega(\|x_0 + v((x^* - x_0) + t(y^* - x^*))\|)\,dv\right) dt \\ &< \beta \int_0^1 ((1-t)R + tr) \left(\int_0^1 \omega(\|x_0\| + \|v((x^* - x_0) + t(y^* + x_0 - x_0 - x^*))\|)\,dv\right) dt \\ &\leq \beta \int_0^1 ((1-t)R + tr) \left(\int_0^1 \omega(t_0 + v(R + t(r-R)))\,dv\right) dt \end{aligned}$$

and, since

$$\beta \int_0^1 ((1-t)R + tr) \left(\int_0^1 \omega(t_0 + v(R + t(r-R)))\,dv\right) dt = \frac{\beta}{r-R} \int_R^r \int_{t_0}^{t_0+u} \omega(z)dzdu = 1,$$

by the Banach lemma, operator (12) exists.

If $r = R$, we suppose that y^* is another solution of $F(x) = 0$ in $\overline{B(x_0, R)}$. Since $\|y^* - x_0\| \leq t^* - t_0$, by mathematical induction we suppose that $\|y^* - x_k\| \leq t^* - t_k$ for $k = 0, 1, \ldots, n$. Then, having into account that $F(y^*) = 0$ and $x_{n+1} = x_n - \Gamma_n F(x_n)$ we can write

$$y^* - x_{n+1} = -\Gamma_n \int_0^1 F''(x_n + t(y^* - x_n))(1-t)(y^* - x_n)^2\,dt,$$

as $\|x_n + t(y^* - x_n)\| \leq t_n + t(y^* - t_n)$, we obtain

$$\|y^* - x_{n+1}\| \leq -\frac{M}{f'(t_n)}\|y^* - x_n\|^2, \tag{13}$$

being $M = \int_0^1 \omega(t_n + t(y^* - t_n))(1-t)\,dt$.

In the same way for f function, we have

$$t^* - t_{n+1} = -\frac{1}{f'(t_n)} \int_0^1 f''(t_n + t(t^* - t_n))(1-t)(t^* - t_n)^2\,dt,$$

and therefore we obtain

$$t^* - t_{n+1} = -\frac{M}{f'(t_n)}(t^* - t_n)^2. \tag{14}$$

Then, from (13) and (14) we prove that $\|y^* - x_{n+1}\| \leq t^* - t_{n+1}$. So $\|y^* - x_n\| \leq t^* - t_n$ for all n, therefore as $\lim_n t_n = t^*$ and $\lim_n x_n = x^*$, it follows that $y^* = x^*$. □

§3. Application to a particular equation (5)

We have seen in the introduction that second derivative (6) is not bounded in all $\Omega = \{x \in C[a,b] : \ x(s) > 0,\ s \in [a,b]\}$. On the contrary, we see in the following that the alternative condition given by (C_3) in Theorem 4 holds, and consequently the convergence of Newton's method to a solution of (5) is then guaranteed from Theorem 4. From (C_3) we deduce

$$\omega(z) = M\,(1 + (2+p)(1+p)z^p)\,. \tag{15}$$

Moreover, for a fixed $x_0(s)$, we have

$$\|I - F'(x_0)\| \leq M\Big((2+p)\|x_0^{1+p}\| + \|x_0\|\Big),$$

and by the Banach lemma, we obtain

$$\|\Gamma_0\| \leq \frac{1}{1 - M\Big((2+p)\|x_0^{1+p}\| + \|x_0\|\Big)} = \beta,$$

provided that $M\Big((2+p)\|x_0^{1+p}\| + \|x_0\|\Big) < 1$. Furthermore, since $\|F(x_0)\| \leq \|x_0 - u\| + M\Big(\|x_0^{2+p}\| + \frac{1}{2}\|x_0^2\|\Big)$, it follows that

$$\|\Gamma_0 F(x_0)\| \leq \|\Gamma_0\|\,\|F(x_0)\| \leq \frac{\|x_0 - u\| + M\Big(\|x_0^{2+p}\| + \frac{1}{2}\|x_0^2\|\Big)}{1 - M\Big((2+p)\|x_0^{1+p}\| + \|x_0\|\Big)} = \eta.$$

Once the parameters β and η are calculated and function (15) is known, we use Theorem 4 to prove the existence of solution of equation (5) and guarantee the convergence of Newton's method.

To determine the domain of existence of solution, we consider the following particular equation (5):

$$x(s) = 1 + \int_0^1 G(s,t)\left(x(t)^{5/2} + \frac{1}{2}x(t)^2\right)dt, \quad s \in [0,1], \tag{16}$$

where the kernel G is the Green function.

If we repeat what is done for (5) with $u(s) = 1$, $p = 1/2$, $[a, b] = [0, 1]$ and choose $x_0(s) = 1/2$, we can guarantee by the Banach lemma that the operator Γ_0 exists and $\|\Gamma_0\| \leq 32(12 + \sqrt{2})/355$, since

$$\|[(I - F'(x_0))y](s)\| \leq \frac{1}{64}\left(4 + 5\sqrt{2}\right)\|y\| \qquad \text{and} \qquad \|I - F'(x_0)\| < 1.$$

Moreover, $\|F(x_0)\| \leq (33 + \sqrt{2})/64$ and

$$\beta = 1.2091\ldots, \qquad \eta = 0.6501\ldots, \qquad \omega(z) = \frac{1}{8} + \frac{\sqrt{z}}{32}.$$

Since $t_0 \geq \|x_0\| = 1/2$ in Theorem 4, we take $t_0 = 1/2$, so that the equation

$$W(t) - W(t_0) - \frac{1}{\beta} = \frac{1}{96}(2t\sqrt{t} + 12t - 7\sqrt{2} - 96) = 0,$$

has only one root: $\alpha = 5.1992\ldots$

If we now construct the function $f(t)$ of theorem 4, we obtain

$$f(t) = (0.0083\ldots)t^2\sqrt{t} + (0.0625\ldots)t^2 - (0.8968\ldots)t + (0.9690\ldots),$$

so that $f(\alpha) = -1.4908\ldots < 0$. The smallest positive root of $f(t) = 0$ is $t^* = 1.1943\ldots$ and $t^* - \|x_0\| = 0.6943\ldots = R$, so that the domain of existence of solution is

$$\{\varphi \in C[0, 1];\ \|\varphi - \frac{1}{2}\| \leq 0.6943\ldots\}.$$

Moreover, as the biggest positive root of the corresponding equation (11) is $r = 8.5193\ldots$, then the domain of uniqueness of solution is

$$\{\varphi \in C[0, 1];\ \|\varphi - \frac{1}{2}\| < 8.5193\ldots\} \cap \Omega.$$

Note that in practice we can observe that the domain of existence of solution is optimum when $t_0 = \|x_0\|$.

Acknowledgements

This paper was partly supported by the Spanish Ministry of Science and Innovation (MTM 2008-01952/MTM) and the Government of Riojan Autonomous Community (Colabora 2009/04).

References

[1] Bruns, D. D., and Bailey, J. E. Nonlinear feedback control for operating a nonisothermal cstr near an unstable steady state. *Chem. Eng. Sci. 32* (1997), 257–264.

[2] Ganesh, M., and Joshi, M. C. Numerical solvability of hammerstein integral equations of mixed type. *IMA J. Numer. Anal. 11* (1991), 21–31.

[3] Kantorovich, L. V., and Akilov, G. P. *Functional Analysis*. Pergamon Press, Oxford, 1982.

José Antonio Ezquerro, Daniel González and Miguel Ángel Hernández
Department of Mathematics and Computation
University of La Rioja
C/ Luis de Ulloa, s/n
26004 Logroño, Spain
`jezquer@unirioja.es`, `daniel.gonzalez@unirioja.es`, `mahernan@unirioja.es`

Monografías Matemáticas *García de Galdeano* **37**, 133–137 (2012)

FREENESS OF LINE ARRANGEMENTS WITH MANY CONCURRENT LINES

Daniele Faenzi and Jean Vallès

Abstract. We propose here a new approach in order to study line arrangements on the projective plane. We use this approach to prove Terao's conjecture when many lines of the arrangement are concurrent.

Keywords: Line arrangements, free arrangements, Terao's conjecture.

AMS classification: 52C35, 14F05, 32S22.

A line arrangement in $\mathbb{P}^2 = \mathbb{P}(\mathbb{C}[x_0, x_1, x_2])$ is a finite collection of lines, say $\{l_1, \ldots, l_s\}$. The union of these lines is a reduced divisor denoted by $D = \{f = 0\}$, where f is the product of the s linear forms defining the l_i's. Saito (see [3]) associates to D the bundle $T(\log D)$ of vector fields with logarithmic poles along D. This is a vector bundle of rank 2, defined by the following exact sequence of sheaves:

$$0 \longrightarrow T(\log D) \longrightarrow \mathcal{O}^3_{\mathbb{P}^2} \xrightarrow{(\frac{\partial f}{\partial x_0}, \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2})} \mathcal{O}_{\mathbb{P}^2}(s-1). \tag{1}$$

We say that the arrangement is *free* when $T(\log D)$ splits as a sum of two line bundles and more precisely we will say that it is *free of type* (a, b), with $0 \leq a \leq b$, when $T(\log D) \simeq \mathcal{O}_{\mathbb{P}^2}(-a) \oplus \mathcal{O}_{\mathbb{P}^2}(-b)$.

The main open question about these bundles (also valid on $\mathbb{P}^n$, for $n \geq 2$) is the so-called Terao's conjecture (see [2]):

Conjecture 1 (Terao)**.** *Freeness of D depends only on its combinatorial type.*

By combinatorial type here we mean the intersection lattice associated to the arrangement. Its incidence graph has one vertex v_i for each line l_i of the arrangement and one vertex $u_{i,j}$ an intersection point $l_i \cap l_j$. The vertex $u_{i,j}$ is linked by an edge to all vertices v_k such that $l_i \cap l_j$ lies in l_k. So two line arrangements are said to have the same combinatorial type if these graphs are isomorphic.

We propose a new approach to Terao's conjecture, based on projective duality. Any line of the divisor D corresponds to a point in $\mathbb{P}^{2\vee}$. This way we associate to D a finite set Z of points in $\mathbb{P}^{2\vee}$. From now, in order to insist on the correspondence, we will denote by $Z \subset \mathbb{P}^{2\vee}$ the finite set of points and by $D_Z \subset \mathbb{P}^2$ the corresponding divisor.

Let us now introduce the variety $\mathbb{F} \subset \mathbb{P}^2 \times \mathbb{P}^{2\vee}$. which is the incidence variety *point-line* in $\mathbb{P}^2$, and the projections p and q on $\mathbb{P}^2$ and $\mathbb{P}^{2\vee}$.

$$\begin{tikzcd}
\mathbb{F} \arrow[r, "q"] \arrow[d, "p"'] & \mathbb{P}^{2\vee} \\
\mathbb{P}^2 &
\end{tikzcd}$$

Let $\mathcal{J}_Z$ be the ideal sheaf of Z in $\mathbb{P}^{2\vee}$. We show first that the Saito vector bundle $T(\log D)$ is obtained by looking at $\mathcal{J}_Z(1)$ on $\mathbb{P}^2$. More precisely, we prove:

Theorem 2. $p_*q^*\mathcal{J}_Z(1) \simeq T(\log D)$.

Proof. Let us consider the canonical exact sequence:

$$0 \longrightarrow \mathcal{J}_Z(1) \longrightarrow \mathcal{O}_{\mathbb{P}^{2\vee}}(1) \longrightarrow \mathcal{O}_Z(1) \longrightarrow 0.$$

Looking at the above exact sequence over $\mathbb{P}^2$ means applying the functor p_*q^*. Then, denoting by $T_{\mathbb{P}^2}$ the tangent bundle to $\mathbb{P}^2$, we have:

$$0 \longrightarrow p_*q^*\mathcal{J}_Z(1) \longrightarrow T_{\mathbb{P}^2}(-1) \longrightarrow \oplus_{l\in Z}\mathcal{O}_l.$$

Now, the equation f of D_Z provides a non zero global section of the ideal sheaf generated by the partial derivatives of f, namely the Jacobian ideal $\mathcal{J}_f(s)$. This amounts to an injective morphism of sheaves of the form $\mathcal{O}_{\mathbb{P}^2} \to \mathcal{J}_f(s)$. This morphism induces a commutative diagram:

$$\begin{tikzcd}
 & & \mathcal{O}_{\mathbb{P}^2}(-1) \arrow[r, equal] \arrow[d] & \mathcal{O}_{\mathbb{P}^2}(-1) \arrow[d, "f"'] & \\
0 \arrow[r] & T(\log D) \arrow[r] \arrow[d, equal] & \mathcal{O}^3_{\mathbb{P}^2} \arrow[r] \arrow[d] & \mathcal{J}_f(s-1) \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & T(\log D) \arrow[r] & T_{\mathbb{P}^2}(-1) \arrow[r] & C \arrow[r] & 0,
\end{tikzcd}$$

where the middle row is the exact sequence (1) defining $T(\log D)$. The sheaf C is the ideal sheaf of the singular locus of the hypersurface $\{f = 0\}$ considered on the hypersurface. We have a natural inclusion $C \subset \oplus_{l\in Z}\mathcal{O}_l$ by desingularization. Then, since the homomorphism $T_{\mathbb{P}^2}(-1) \to \oplus_{l\in Z}\mathcal{O}_l$ is essentially unique (see [4]) this proves that both kernels $p_*q^*\mathcal{J}_Z(1)$ and $T(\log D)$ coincide. □

In order to show that this approach is relevant we prove here a special case of Terao's conjecture, without using any further material.

Theorem 3. *Terao's conjecture is true for a free divisor D_Z of type $(n, n + r)$, with $r \geq 0$, as soon as $(n + 2)$ points of Z are collinear.*

Saying that $(n + 2)$ points of Z are collinear amounts to require that $(n + 2)$ lines of D_Z are concurrent, hence we may say that freeness of arrangements with many concurrent lines is combinatorial.

The first step to prove the theorem is the following lemma relating sections on one side to decomposition over lines on the dual side.

Lemma 4. *Let $Z \subset \mathbb{P}^{2\vee}$ be a set of points and x be a general point in $\mathbb{P}^{2\vee}$. Assume that $T(\log D_Z) \otimes \mathcal{O}_{x^\vee} = \mathcal{O}_{x^\vee}(-n) \oplus \mathcal{O}_{x^\vee}(-n - r)$ with $r \geq 0$. Then $H^0((\mathcal{J}_Z \otimes \mathcal{J}_x^n)(n + 1)) \neq 0$.*

Proof. Let us denote by $\widehat{\mathbb{P}}$ the blowing up of $\mathbb{P}^{2\vee}$ along the point x. We recall that $\widehat{\mathbb{P}} \simeq p^{-1}(x^\vee) \subset \mathbb{F}$ and we consider the induced incidence diagram:

$$\begin{tikzcd}
\widehat{\mathbb{P}} \arrow[r, "\widehat{q}"] \arrow[d, "\widehat{p}"'] & \mathbb{P}^{2\vee} \\
x^\vee &
\end{tikzcd}$$

Moreover we have the following resolution of $\widehat{\mathbb{P}}$ in $\mathbb{F}$:

$$0 \longrightarrow p^*\mathcal{O}_{\mathbb{P}^2}(-1) \longrightarrow \mathcal{O}_{\mathbb{F}} \longrightarrow \mathcal{O}_{\widehat{\mathbb{P}}} \longrightarrow 0.$$

Tensoring the exact sequence above by $q^*\mathcal{J}_Z(1)$ we get:

$$0 \to q^*(\mathcal{J}_Z(1)) \otimes p^*\mathcal{O}_{\mathbb{P}^2}(-1) \to q^*(\mathcal{J}_Z(1)) \to \hat{q}^*(\mathcal{J}_Z(1)) \to 0.$$

Now we apply the functors R^ip_* to the above sequence (see for instance [1, Chapter III]). Let us describe the effect of applying p_* (i.e. R^ip_* for $i = 0$) to the above sequence. For the middle term, the result is computed by Theorem 1 and agrees with $T(\log D)$. For the leftmost term, we get $T(\log D)(-1)$ by Theorem 1 and projection formula (see again [1, Chapter III]). For the rightmost term, we get $\widehat{p}_*\widehat{q}^*\mathcal{J}_Z(1)$ for $\widehat{p}$ and $\widehat{q}$ are the restrictions of p and q to $\widehat{\mathbb{P}}$. Denote by $R^1T(\log D_Z)$ the sheaf $R^1p_*q^*\mathcal{J}_Z(1)$. We can now write the long exact sequence obtained applying R^ip_* for $i = 0, 1$ the exact sequence above.

$$0 \to T(\log D_Z)(-1) \xrightarrow{x^\vee} T(\log D_Z) \longrightarrow \widehat{p}_*\widehat{q}^*\mathcal{J}_Z(1) \to$$

$$\to R^1T(\log D_Z)(-1) \xrightarrow{x^\vee} R^1T(\log D_Z) \longrightarrow R^1\widehat{p}_*\widehat{q}^*\mathcal{J}_Z(1) \to 0.$$

Since x is general, any line through x is at most 1-secant to Z. Then the support of the sheaf $R^1\widehat{p}_*\widehat{q}^*\mathcal{J}_Z(1)$, which is the locus of 3-secant lines to Z through x, is empty. We have proved $\widehat{p}_*\widehat{q}^*\mathcal{J}_Z(1) = T(\log D_Z) \otimes \mathcal{O}_{x^\vee}$.

Then the decomposition $T(\log D_Z) \otimes \mathcal{O}_{x^\vee} = \mathcal{O}_{x^\vee}(-n) \oplus \mathcal{O}_{x^\vee}(-n-r)$ gives an injective homomorphism:

$$\mathcal{O}_{x^\vee}(-n) \hookrightarrow \widehat{p}_*\widehat{q}^*\mathcal{J}_Z(1).$$

This means that we have a non zero map on $\widehat{\mathbb{P}}$:

$$\widehat{p}^*\mathcal{O}_{x^\vee}(-n) \hookrightarrow \widehat{q}^*\mathcal{J}_Z(1),$$

that we can write also as:

$$\mathcal{O}_{\widehat{\mathbb{P}}} \hookrightarrow \widehat{q}^*\mathcal{J}_Z(1) \otimes \widehat{p}^*\mathcal{O}_{x^\vee}(n).$$

This last map is equivalent to a non zero map on $\mathbb{P}^{2\vee}$:

$$\mathcal{O}_{\mathbb{P}^{2\vee}} \hookrightarrow \mathcal{J}_Z(1) \otimes \mathcal{J}_x^n(n) = (\mathcal{J}_Z \otimes \mathcal{J}_x^n)(n+1).$$

□

Proof of the main theorem. We first describe the combinatorial type according to the given data. By hypothesis, there exists a line L such that $|L \cap Z| \geq n+2$. This line is a fixed component in the linear system of curves of degree $n+1$ passing through Z. Since x is general, a curve of degree $n+1$ passing through Z and having multiplicity n at x is the union of L and of n lines through x. Then there are at most n points of Z that do not lie in L. Since the length of Z is $2n+r+1$, there are at least $n+r+1$ points on L. In fact, according to the decomposition, L is exactly $(n+r+1)$ secant to Z. Indeed, if there were strictly more than $n+r+1$ points on L, then one could find, for a general x, a curve of degree n passing through Z and having multiplicity $n-1$ at x (take the union of L and of the $n-1$ lines through x and the remaining points) and this contradicts the decomposition.

Assume now that Z_0 has the same combinatorial type than Z. Then, according to Yoshinaga ([5, Thm. 2.2]) the splitting of $T(\log D_{Z_0})$ on the general line $l = x^\vee$ (where x is a general point) is of type $\mathcal{O}_{x^\vee}(-n+t) \oplus \mathcal{O}_{x^\vee}(-n-r-t)$ with $t \geq 0$. This means that there is a curve of degree $n-t+1$ passing through Z_0 and having multiplicity $n-t$ at x. Then this curve is the union of L and $n-t$ lines through x. But since there are n points outside L the number t cannot be positive. So the arrangement D_{Z_0} is free of type $(n, n+r)$. □

Remark 1. We can say more about the combinatorial type of Z, assuming that it is free of type $(n, n+r)$, and that it admits a $(n+r+1)$-secant line. Let us write the reduction exact sequence. Set $Z_1 = Z \setminus Z \cap L$. Then we have:

$$0 \longrightarrow \mathcal{J}_{Z_1} \longrightarrow \mathcal{J}_Z(1) \longrightarrow \mathcal{O}_L(-n-r) \longrightarrow 0.$$

We apply the functor p_*q^* to obtain the following long exact sequence:

$$0 \to \mathcal{O}_{\mathbb{P}^2}(-n) \longrightarrow p_*q^*\mathcal{J}_Z(1) \longrightarrow \mathcal{O}_{\mathbb{P}^2}(-n-r) \to$$

$$\to R^1p_*q^*\mathcal{J}_{Z_1} \longrightarrow R^1p_*q^*\mathcal{J}_Z(1) \longrightarrow R^1p_*q^*\mathcal{O}_L(-n-r) \to 0.$$

Since $p_*q^*\mathcal{J}_Z(1) \cong \mathcal{O}_{\mathbb{P}^2}(-n) \oplus \mathcal{O}_{\mathbb{P}^2}(-n-r)$, we have a short exact sequence relating the locus of 2-secant lines to Z_1 to the locus of 3-secant lines to Z:

$$0 \to R^1p_*q^*\mathcal{J}_{Z_1} \longrightarrow R^1p_*q^*\mathcal{J}_Z(1) \longrightarrow R^1p_*q^*\mathcal{O}_L(-n-r) \to 0.$$

The last sheaf is the structure sheaf of the fat point of length $\binom{n+r}{2}$ supported on $L^\vee$. So any 2-secant line to Z_1 must correspond to a 3-secant line to Z, i.e. any line passing through $r+2$ points ($r \geq 0$) of Z_1 must further pass through a point of Z lying on L.

References

[1] Hartshorne, R. *Algebraic geometry*. Springer-Verlag, New York, 1977. Graduate Texts in Mathematics, No. 52.

[2] Orlik, P., and Terao, H. *Arrangements of hyperplanes*, vol. 300 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1992.

[3] Saito, K. Theory of logarithmic differential forms and logarithmic vector fields. *J. Fac. Sci. Univ. Tokyo Sect. IA Math. 27*, 2 (1980), 265–291.

[4] VALLÈS, J. Fibrés logarithmiques sur le plan projectif. *Ann. Fac. Sci. Toulouse Math. (6) 16*, 2 (2007), 385–395. Available from: `http://afst.cedram.org/item?id=AFST_2007_6_16_2_385_0`.

[5] YOSHINAGA, M. Characterization of a free arrangement and conjecture of Edelman and Reiner. *Invent. Math. 157*, 2 (2004), 449–454. `doi:10.1007/s00222-004-0359-2`.

Daniele Faenzi and Jean Vallès
Université de Pau et des Pays de l'Adour
Avenue de l'Université - BP 576
64012 PAU Cedex
France
`daniele.faenzi@univ-pau.fr` and `jean.valles@univ-pau.fr`

Monografías Matemáticas *García de Galdeano* **37**, 139–147 (2012)

Lyapunov stability for a Rydberg atom in a circularly polarized microwave field and a static magnetic field

Manuel Iñarrea, Víctor Lanchares, Ana Isabel Pascual and José Pablo Salas

Abstract. The interaction of a Rydberg atom with a circularly polarized microwave field leads, with finely tuned parameters, to the creation of stable equilibrium positions similar to the well known Lagrangian equilibrium points in celestial mechanics (cf. [6]). Besides, the addition of a static magnetic field, perpendicular to the plane of polarization, can be used to manipulate the stability properties of the equilibria (cf. [9] and [10]).

The aim of this communication is the characterization of nonlinear stability properties for equilibrium points by making use of appropriate results, based on KAM theory. Special attention will be paid when resonance conditions take place between the fundamental frequencies of the system.

Keywords: Circularly microwave field, static magnetic field, equilibria, stability, resonances.

AMS classification: 70H08, 70H14, 37N05.

§1. Introduction

The dynamics of a hydrogen atom in the presence of a circularly polarized (CP) microwave field crossed with a magnetic field **B**, denoted hereafter by CP × **B**, is known to give rise to two different behaviors (cf. [9]). On the one hand, the electron can follow a perturbed Keplerian orbit, which can be studied under the point of view of classical mechanics, by means of perturbation methods (cf. [8]). On the other hand, the electron can be trapped in a region beyond the Stark saddle point, by properly tuning the external parameters of the problem.

The last case is specially interesting, because of the appearance of equilibrium points similar to the Lagrangian points in the restricted three body problem. These points are directly linked with the ionization threshold for the electron (cf. [6]). Besides this remarkable connection, the stability properties of these points are also of interest as they are the key to have a real trapping region. In this way, the main goal of this work is the characterization of nonlinear stability properties of these points as a function of the external parameters. We will not pay attention to the size of the region of stability, a question that deserves a further analysis and that is relevant to get an effective trapping region.

The problem we deal with has also a Celestial Mechanics counterpart, as it describes the dynamics of a dust particle subject to radiation pressure, the sun magnetic field and orbiting an idealized planet that revolves around the sun in circular orbit. In this context, the existence of stable trapping regions associated with stable equilibrium points may account for dust clouds responsible for the phenomenon known as zodiacal light (cf. [12]).

§2. The problem

In atomic units, the Hamiltonian of the problem CP × **B**, in the dipole aproximation, is given by

$$\mathcal{H} = \frac{1}{2}(P_x^2+P_y^2+P_y^2)-\frac{1}{\sqrt{x^2+y^2+z^2}}+\frac{\omega_c}{2}(xP_y-yP_x)+\frac{\omega_c^2(x^2+y^2)}{8}\pm f(x\cos\omega_f t+y\sin\omega_f t),$$

where the magnetic field is supposed to be parallel to the direction of the z-axis, ω_c is the cyclotron frequency, ω_f is the CP field frequency and f the electric field strength ($f > 0$). The plus or minus sign depends on the polarization direction of the microwave field.

The explicit time dependence in the Hamiltonian can be removed by going to a sinodic reference frame that rotates at the constant angular velocity ω_f, in such a way that the moving x-axis is aligned with the direction of the electric field. The new Hamiltonian becomes

$$\mathcal{H} = \frac{1}{2}(P_x^2+P_y^2+P_z^2)-\frac{1}{\sqrt{x^2+y^2+z^2}}-\left(\omega_f\pm\frac{\omega_c}{2}\right)(xP_y-yP_x)+\frac{\omega_c^2(x^2+y^2)}{8}\pm fx. \quad (1)$$

In this work, we study the planar model, that is the model restricted to the invariant set $z = P_z = 0$. In this way, by setting z and P_z to zero in (1), we obtain the Hamiltonian corresponding to the plane case:

$$\mathcal{H} = \frac{1}{2}(P_x^2+P_y^2)-\frac{1}{\sqrt{x^2+y^2}}-\left(\omega_f\pm\frac{\omega_c}{2}\right)(xP_y-yP_x)+\frac{\omega_c^2(x^2+y^2)}{8}\pm fx. \quad (2)$$

As we are interested in the equilibria of the system given by the Hamiltonian (2) we have to solve the corresponding Hamilton equations equated to zero. These are

$$\begin{cases} \dot{x} = P_x + \omega y, \\ \dot{y} = P_y - \omega x, \\ \dot{P}_x = -\dfrac{x}{r^3} + \omega P_y \mp f - \dfrac{\omega_c^2}{4}x, \\ \dot{P}_y = -\dfrac{y}{r^3} - \omega P_x - \dfrac{\omega_c^2}{4}y, \end{cases}$$

where $\omega = \omega_f \pm \omega_c/2$ and $r = \sqrt{x^2+y^2}$.

From the system above it follows that an equilibrium point (x, y, P_x, P_y) must verify $P_x = y = 0$ and $P_y = \omega x$. Moreover, x must be a positive root of the equation

$$\omega_f(\omega_f \pm \omega_c)x^3 \mp fx^2 - 1 = 0,$$

or a negative root of the equation

$$\omega_f(\omega_f \pm \omega_c)x^3 \mp f x^2 + 1 = 0.$$

The discussion on the number of equilibria is summarized in the following proposition

Proposition 1. *For the positive sign in (2), there are two equilibrium points, one of them with $x < 0$ and the other one with $x > 0$. For the minus sign in (2), the number of equilibrium points depends on the sign of $\omega_f(\omega_f - \omega_c)$:*

- *If $\omega_f(\omega_f - \omega_c) > 0$, then there are 2 equilibria, one with $x > 0$ and another one with $x < 0$.*
- *If $\omega_f(\omega_f - \omega_c) = 0$, there is one equilibrium point with $x > 0$.*
- *If $\omega_f(\omega_f - \omega_c) < 0$, there are two equilibria if $f > F_c$, where $F_c = \sqrt[3]{\frac{27}{4}\omega_f^2(\omega_f - \omega_c)^2}$. In this case, the two of them verify $x > 0$. If $f \leq F_c$, no equilibria exist.*

To study the stability properties of the equilibrium points we introduce a function of the coordinates x and y, usually called the effective potential, in such a way that linear stable points correspond to relative maxima and minima of this function. In the positive case, the effective potential is given by

$$EP_p = f x - \frac{1}{2}\omega_f(\omega_c + \omega_f)(x^2 + y^2) - \frac{1}{\sqrt{x^2 + y^2}}.$$

As a result, the equilibrium with $x > 0$ is a maximum and the equilibrium with $x < 0$ is a saddle. In the negative case, the effective potential reads as

$$EP_n = -f x - \frac{1}{2}\omega_f(\omega_c - \omega_f)(x^2 + y^2) - \frac{1}{\sqrt{x^2 + y^2}},$$

and the character of the equilibria depends on the sign of $\omega_f(\omega_f - \omega_c)$. In this sense, if $\omega_f(\omega_f - \omega_c) > 0$, the equilibrium with $x > 0$ is a saddle and the equilibrium with $x < 0$ is a maximum. If $\omega_f(\omega_f - \omega_c) < 0$ and $f > F_c$, one of the positive equilibria is a saddle (we call x_s) and the other is a minimum (we call x_m).

The previous analysis shows that there are two different configurations for the effective potential, maximum-saddle or minimum-saddle (cf. [7]). It is known that saddle points correspond to unstable points and a minimum give rise to a nonlinear stable point (cf. [13]), as it follows from Dirichlet's theorem (cf. [14]). On the other hand, a maximum is a linear stable point but its character from the point of view of Lyapunov is not decided. In this way, there are well known counterexamples where a linear stable point of a Hamiltonian system is unstable in the Lyapunov sense (cf. [3]). To solve the question of nonlinear stability we will make use of KAM theory and the next sections are devoted to introduce the results we will use.

§3. Lyapunov stability

One of the results from KAM theory is Arnold's theorem (cf. [1]) that guarantees nonlinear stability of a maximum for almost all set of admissible external parameters. Here we reproduce the version of this theorem given by Meyer and Schmidt in (cf. [11]).

Theorem 2. *Let be a two degrees of freedom Hamiltonian system expressed in variables* $(\Psi_1, \Psi_2, \psi_1, \psi_2)$ *as*

$$\mathcal{H} = \mathcal{H}_2(\Psi_1, \Psi_2) + \mathcal{H}_4(\Psi_1, \Psi_2) + \cdots + \mathcal{H}_{2N}(\Psi_1, \Psi_2) + \overline{\mathcal{H}}(\Psi_1, \Psi_2, \psi_1, \psi_2)$$

where it is verified that:

1. $\mathcal{H}$ *is a real analytic function in a neighborhood of the origin.*
2. *Each* $\mathcal{H}_{2k}$, $1 \leq k \leq N$ *is a homogeneous polynomial of degree* k *in* Ψ_1, Ψ_2 *with real coefficients independent of the angles. In particular,*

$$\mathcal{H}_2 = \omega_1\Psi_1 - \omega_2\Psi_2, \quad \omega_1, \omega_2 > 0, \tag{3}$$

$$\mathcal{H}_4 = \frac{1}{2}(A\Psi_1^2 - 2B\Psi_1\Psi_2 + C\Psi_2^2),$$

   where A, B *and* C *depend on the parameters of the Hamiltonian.*

3. $\overline{\mathcal{H}} = \overline{\mathcal{H}}(\Psi_1, \Psi_2, \psi_1, \psi_2) = O((\Psi_1 + \Psi_2)^{2N+1})$.

With these conditions, the origin is a stable equilibrium if exists $2 \leq k \leq N$ *such as*

$$\mathcal{D}_{2k} = \mathcal{H}_{2k}(\omega_2, \omega_1) \neq 0,$$
$$\mathcal{D}_{2j} = \mathcal{H}_{2j}(\omega_2, \omega_1) = 0, \quad 2 \leq j < k.$$

The practical implementation of this theorem yields a great amount of computation work. First, the Hamiltonian must be expressed in action–angle variables, in such a way the quadratic part reduces to (3).

This is achieved by means of a linear transformation. Next, the most tricky part of the process, the Hamiltonian must be brought to the so-called Birkhoff normal form (cf. [2] and [15]) up to a certain order through the application of successive canonical transformations near to the identity. This process can be made in a recursive manner using the algorithm of the Lie–Deprit perturbation method (cf. [4]). This process simplifies if it is carried out in complex variables, returning to Poincare, or action-angle, variables at the end. Once the normalization has been completed Arnold's theorem can be applied.

We note that in the statements of theorem 2 there are some implicit assumptions. The first one is that the Hamiltonian is written in normal form, that is, the computation work is already supposed done. The second one is that the frequencies of the system ω_1 and ω_2 are not in resonance of order less or equal than $2N$, because the terms $\mathcal{H}_{2k}$, $1 \leq k \leq N$, only depend on the momenta Ψ_1 and Ψ_2. When the frequencies of the system verify a resonance condition, the normal form is no longer as those presented in the theorem and terms depending on the angles ψ_1 and ψ_2 appear.

To handle resonant cases we need a more general result. In this sense we will make use of a geometric criterion (cf. [5] and [13]), that extends Arnold's theorem to the resonant cases.

§4. The geometric criterium

Let us suppose that $\mathcal{H}_2$ can be written as in (3). Then, the Hamiltonian $\mathcal{H}$ can be brought to normal form, in such a way that $\mathcal{H}_2$ becomes a formal integral. Also let us assume that ω_1 and

ω_2 satisfy a $n{:}m$ resonant condition of order greater or equal than two, that is, $n\omega_1 - m\omega_2 = 0$ with $n + m \geq 2$. Under these assumptions we introduce a set of action–angle variables named Lissajous variables, with a twofold goal. On the one hand, the formal integral depends on one of the actions and the normalization procedure can be viewed as the elimination of a fast variable by means of an averaging process.

Lissajous variables $(\Phi_1, \Phi_2, \phi_1, \phi_2)$ are specifically designed for each particular value of the resonance $n{:}m$ and they are related with the Poincaré variables through the formulae

$$\begin{aligned} \Psi_1 &= \frac{\Phi_1 + \Phi_2}{2m}, & \psi_1 &= m(\phi_1 + \phi_2), \\ \Psi_2 &= \frac{\Phi_1 - \Phi_2}{2n}, & \psi_2 &= n(\phi_1 - \phi_2). \end{aligned}$$

Now, $\mathcal{H}_2$ turns to be simply $\nu\,\Phi_2$, being

$$\nu = \frac{\omega_1}{m} = \frac{\omega_2}{n}.$$

Besides, the Poisson's bracket $(\mathcal{H}_2, \mathcal{H}_j)$, needed to compute the normal form, is just

$$(\mathcal{H}_2, \mathcal{H}_j) = \nu \frac{\partial \mathcal{H}_j}{\partial \phi_2},$$

and the process of normalization is no more than an averaging over the ϕ_2 angle.

Moreover, the normal form is generated by the invariants (cf. [5]) M_1, M_2, C and S that, as functions of Lissajous variables, are given by

$$\begin{aligned} M_1 &= \frac{1}{2}\Phi_1, & C &= 2^{-(m+n)/2}(\Phi_1 - \Phi_2)^{m/2}(\Phi_1 + \Phi_2)^{n/2}\cos 2nm\phi_1, \\ M_2 &= \frac{1}{2}\Phi_2, & S &= 2^{-(m+n)/2}(\Phi_1 - \Phi_2)^{m/2}(\Phi_1 + \Phi_2)^{n/2}\sin 2nm\phi_1. \end{aligned} \tag{4}$$

In this way, the normal form up to order N is written as

$$\mathcal{H} = \mathcal{H}_2 + \sum_{j=3}^{N} \mathcal{H}_j,$$

where $\mathcal{H}_2 = 2\omega M_2$, and

$$\mathcal{H}_j = \sum_{2(\gamma_1+\gamma_2)+(n+m)(\gamma_3+\gamma_4)=j} a_{\gamma_1\gamma_2\gamma_3\gamma_4} M_1^{\gamma_1} M_2^{\gamma_2} C^{\gamma_3} S^{\gamma_4}, \qquad 3 \leq j \leq N,$$

with $a_{\gamma_1\gamma_2\gamma_3\gamma_4} \in \mathbb{R}$.

The invariants are not independent and they satisfy the equation

$$C^2 + S^2 = (M_1 + M_2)^n (M_1 - M_2)^m, \tag{5}$$

together with the restriction

$$M_1 \geq |M_2|\,. \tag{6}$$

Note that the reduced phase space is given by the equation (5) and the restriction (6). It is a set of surfaces of revolution, one for each constant value of M_2. Fixed a value for M_2, (5) is a surface of revolution with a vertex in the point $M_1 = |M_2|$, $C = S = 0$.

Once the reduced phase space is determined, it is possible to know the flow of the normalized system, when it is truncated to a prescribed order. Indeed, the flow results as the intersection of the normalized Hamiltonian function with the surface defined by (5). Based on this idea, in (cf. [5] and [13]), Elipe et al. established the following results, the first is the geometric criterion and the second one is a derived result from it.

Theorem 3. *Let us assume that the Hamiltonian is normalized up to a certain order $N \geq s$, being $\mathcal{H}_N$ the first term that does not vanish for $M_2 = 0$. Let us consider the two surfaces*

$$\mathcal{G}_1 = \{(C, S, M_1) \in \mathbb{R}^3; \quad \mathcal{H}_N(C, S, M_1, 0) = 0\},$$

and

$$\mathcal{G}_2 = \{(C, S, M_1) \in \mathbb{R}^3; \quad C^2 + S^2 = M_1^s\}.$$

If the origin is an isolated point of intersection, then it is stable. In other case, if the two surfaces are not tangent, the origin is unstable.

Theorem 4. *Let us assume that $\mathcal{H}_s$ (s is the order of resonance) is the first term in the normal form that does not vanish for $M_2 = 0$. If s is odd ($s \geq 3$), then $\mathcal{H}_s(C, S, M_1, 0) = \gamma C + \eta S$ with $\gamma^2 + \eta^2 \neq 0$ and the origin is an unstable equilibrium. If s is even ($s \geq 4$), then $\mathcal{H}_s(C, S, M_1, 0) = a_s M_1^{s/2} + \gamma C + \eta S$ with $a_s^2 + \gamma^2 + \eta^2 \neq 0$ and the stability of the origin depends on the relative values of a_s^2 and $\gamma^2 + \eta^2$: if $a_s^2 > \gamma^2 + \eta^2$, the origin is a stable equilibrium, whereas if $a_s^2 < \gamma^2 + \eta^2$, the origin is unstable.*

§5. Resonant cases

Using the previous results, we study the Lyapunov stability of the maximum when it is verified a resonant condition. We start with the resonance of order three to be followed by the fourth order resonance.

5.1. 1:2 resonance

For a 1:2 resonance ($\omega_1 = 2\omega_2$), the term of order 3 in the normal form can be expressed in complex variables as

$$\mathcal{H}_3 = a_{1002} u V^2 + a_{0120} U v^2,$$

where $a_{1002} = i a_{0120}$ and $a_{0120} \in \mathbb{C}$. Therefore

$$\mathcal{H}_3 = a_{0120}(U v^2 - i u V^2).$$

Expressed in Lissajous invariants, $\mathcal{H}_3$ can be written as

$$\mathcal{H}_3 = a_s S,$$

with $a_s = -a_{0120}\sqrt{2}$.

Then, the two surfaces, from the geometric criterion, are given by

$$\mathcal{G}_1 = \{(C, S, M_1) \in \mathbb{R}^3 \,|\, a_s S = 0\},$$
$$\mathcal{G}_2 = \{(C, S, M_1) \in \mathbb{R}^3 \,|\, C^2 + S^2 = M_1^3, \quad M_1 \geq 0\}.$$

It is clear that if $a_s \neq 0$, then $\mathcal{G}_1 \cap \mathcal{G}_2 = \{(C, S, M_1) \in \mathbb{R}^3 \,|\, S = 0, C = \pm M_1^{3/2}\}$ and therefore, the equilibrium is unstable. By the contrary, if $a_s = 0$, then $\mathcal{H}_3(M_2 = 0) = 0$ and also $\mathcal{H}_3 \equiv 0$, and, in consequence, we need more terms of the normal form to decide about the stability. We have to compute the next term in the normal form, that is, $\mathcal{H}_4$. In this way the next nonzero term in the normal form $\mathcal{H}_4$ can be written, in complex variables, as

$$\mathcal{H}_4 = a_{1200} u U^2 + a_{1111} u U v V + a_{0012} v V^2.$$

Once expressed in terms of the real invariants, we have that

$$\mathcal{H}_4(M_2 = 0) = \frac{29.877}{\omega^4 x_0^8} M_1^2,$$

where x_0 is the x coordinate of the equilibrium point and $\omega = \omega_c + \frac{\omega_f}{2}$. Therefore the origin is the only point in the intersection $\mathcal{G}_1 \cap \mathcal{G}_2$ and, as a consequence of the geometric criterion, the equilibrium is stable.

5.2. 1:3 resonance

In presence of a 1:3 resonance, the term of fourth order in the normal form $\mathcal{H}_4$ evaluated at $M_2 = 0$ is given, in complex variables, by

$$\mathcal{H}_4 = a_{2200} u^2 U^2 + a_{1111} u U v V + a_{0022} v^2 V^2 + a_{1003} u V^3 + a_{0130} U v^3.$$

Expressed in Lissajous invariants, $\mathcal{H}_4$ can be written as

$$\mathcal{H}_4(M_2 = 0) = a_m M_1^2 + a_c C + a_s S,$$

being a_m, a_c, a_s dependent on the parameters of the problem and the coordinates of the equilibrium point.

In our problem it is always verified that $a_m^2 > a_c^2 + a_s^2$. Therefore, as a consequence of Theorem 4, the equilibrium is always stable.

Acknowledgements

We thank financial support from Ministerio de Ciencia e Innovación (Project #MTM2008–03818/MTM) from Spain.

References

[1] Arnold, V. I. The stability of the equilibrium position of a Hamiltonian system of ordinary differential equations in the general elliptic case. *Soviet Math. Dokl. 2* (1961), 247–249.

[2] Birkhoff, G. D. *Dynamical Systems*, vol. 9. American Mathematical Society, Providence, Rhode Island, 1991.

[3] Cherry, T. M. On periodic solutions of Hamiltonian systems of differential equations. *Phil. Trans. Roy. Soc. 227* (1928), 137–221.

[4] Deprit, A. Canonical transformations depending on a small parameter. *Celestial Mechanics and Dynamical Astronomy* (1969), 12–30. `doi:10.1007/BF01230629`.

[5] Elipe, A., Lanchares, V., and Pascual, A. I. On the stability of equilibria in two degrees of freedom Hamiltonian systems under resonances. *J. Nonlinear Sci. 15* (2005), 305–319. `doi:10.1007/s00332-004-0674-1`.

[6] Farrelly, D., and Uzer, T. Ionization mechanism of Rydberg atoms in a circularly polarized microwave field. *Phys. Rev. Lett. 74* (1995), 1720–1723. `doi:10.1103/PhysRevLett.74.1720`.

[7] Iñarrea, M., Lanchares, V., Pascual, A. I., and Salas, J. P. Electronic traps in a circularly polarized microwave field and a static magnetic field: Stability analysis. *Monografías De La Real Academia De Ciencias De Zaragoza 14* (1999), 114–118.

[8] Lanchares, V., Iñarrea, M., and Salas, J. P. Bifurcations in the hydrogen atom in the presence of a circularly polarized microwave field and a static magnetic field. *Phys. Rev. A. 56*, 3 (1997), 1839–1843. `doi:10.1103/PhysRevA.56.1839`.

[9] Lee, E., Brunello, F., and Farrelly, D. Single atom quasi–penning trap. *Phys. Rev. Lett. 75* (1995), 3461–3643. `doi:10.1103/PhysRevLett.75.3641`.

[10] Lee, E., Brunello, F., and Farrelly, D. Coherent states in a Rydberg atom: Classical mechanics. *Phys. Rev. A. 55* (1997), 2203–2221. `doi:10.1103/PhysRevA.55.2203`.

[11] Meyer, K. R., and Schmidt, D. S. The Stability of the Lagrange Triangular Point and a Theorem of Arnold. *Journal of Differential Equations 62* (1986), 222–236. `doi:http://math.uc.edu/~meyer/jde86.pdf`.

[12] Moulton, F. R. *An introduction to Celestial Mechanics.* Dover, New York, 1970. 2nd edition.

[13] Pascual, A. I. *Doctoral Thesis: Sobre la estabilidad de sistemas hamiltonianos de dos grados de libertad bajo resonancias.* Servicio de Publicaciones de la Universidad de La Rioja, Logroño, La Rioja, Spain, 2005. Available from: `http://www.unirioja.es/servicios/sp/tesis/119.shtml`.

[14] Siegel, C. L., and Moser, L. K. *Lectures on Celestial Mechanics*, vol. 187. Springer–Verlag, New York, 1971.

[15] Verhulst, F. *Nonlinear Differential Equations and Dynamical Systems.* Springer–Verlag, New York, 1991.

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

Monografías Matemáticas *García de Galdeano* **37**, 149–159 (2012)

Reduction of Gibbs phenomenon for 1D RBF interpolation

Diego Izquierdo, María Cruz López de Silanes
and María Cruz Parra

Abstract. The Gibbs phenomenon can be observed in different interpolation methods. Radial basis functions (RBF) is a modern meshfree interpolation technique in any number of dimensions. Here we investigate the Gibbs phenomenon for 1D RBF interpolation numerically, and propose a procedure to reduce Gibbs oscillations using nonsmooth basis functions locally. The accuracy in the smooth region is enhanced by applying piecewise linear basis functions in the proximity of discontinuity.

Keywords: Radial basis functions, RBF, Gibbs phenomenon, interpolation.

AMS classification: 65D05, 41A05.

§1. Introduction

Radial basis functions interpolation is a modern meshfree technique in any number of dimensions collected in [7] and introduced by Hardy using multiquadrics [5].

Gibbs phenomenon is the peculiar manner in which the Fourier series of a function f behaves at a jump discontinuity. The overshoot does not die out as the frequency increases, but approaches a finite limit. Gibbs phenomenon can also be observed in different interpolation methods. Fornberg and Flyer [3] perform cardinal interpolation for discontinuous functions with centers $x_j = j \in \mathbb{Z}$ and study expansion coefficients for some RBFs. Guessab, Moncayo and Schmeisser [4] define a class of nonlinear four point subdivision schemes. These schemes include as a particular case the PPH scheme (or power-2 scheme) previously studied by Amat, Donat, Liandrat and Trillo [1]. The general schemes, by using generalized harmonic means, reduce the Gibbs phenomenon around jump discontinuities, as occurs with power-2 scheme. Their properties (e.g. stability, convexity preservation, approximation order) are more balanced than those of the power-p schemes.

Jung [6] makes a complete study of RBF interpolation on $\mathbb{R}$ of step function with uniformly distributed centers in $[-1, 1]$ and uses multiquadric with shape parameter, γ, $\Phi(x) = \sqrt{|x|^2 + \gamma^2}$. Jung proposes a method to reduce Gibbs phenomenon adapting shape parameter, i.e. to define $\gamma = 0$ at centers next to discontinuity. Actually, multiquadric is changed by linear RBF at these centers. Here, our aim is to describe a similar interpolation technique that eliminates oscillations next to discontinuity, using different RBFs.

This paper is divided into the following sections. In Section 2, we establish the necessary notations and preliminaries for RBF interpolation on $\mathbb{R}^d$, a technique described in [7]. In Section 3, we consider an interpolation example of the discontinuous function studied in [6]. First, we study local performance of interpolation with two centers and then interpolation with N centers uniformly distributed in $[-1, 1]$. We use RBFs of [2, Appendix D] to obtain interpolant features for different step functions and two other functions. Finally, in Section 4, we develop a local piecewise linear interpolation for discontinuous functions using different RBFs to reduce the Gibbs oscillations in the vicinity of the discontinuity. This technique adapts and expands the method described in [6] to most RBF of [2, Appendix D]. Then, this technique is applied to some examples presented in the previous section. We finish the section obtaining some errors for an example in [4] and compare these results with those given there. The numerical and graphical examples presented in this paper have been executed using Mathematica 8.0.

§2. RBF interpolation

Definition 1. A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is said to be radial if there exists a continuous function $\phi : [0, +\infty) \to \mathbb{R}$ such that $\Phi(x) = \phi(\|x\|_2)$ for all $x \in \mathbb{R}^d$.

Let $N \in \mathbb{N}$. We interpolate an unknown function $f : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$, with data values $F = (f_1, \ldots, f_N)^\top$ at given data sites $X = \{x_1, \ldots, x_N\} \subseteq \Omega$, the set of centers, so that we look for an interpolant as

$$s_{f,X}(x) = \sum_{j=1}^{N} \alpha_j \Phi(x - x_j), \qquad x \in \mathbb{R}^d,$$

with expansion coefficients vector, $\alpha = (\alpha_1, \ldots, \alpha_N)^\top$, so that the interpolation conditions are verified

$$s_{f,X}(x_j) = f_j, \quad 1 \leq j \leq N. \tag{1}$$

Let $A_{\Phi,X} = (\Phi(x_j - x_k))_{1 \leq j,k \leq N}$ be the interpolation matrix. If there exists a unique solution of the system

$$A_{\Phi,X}\alpha = F,$$

then $s_{f,X}$ will be defined.

Definition 2. A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is positive definite on $\mathbb{R}^d$ if, for all $N \in \mathbb{N}$, all pairwise distinct $x_1, \ldots, x_N \in \mathbb{R}^d$ and all $\alpha \in \mathbb{R}^N \setminus \{0\}$, the quadratic form

$$\sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_j \alpha_k \Phi(x_j - x_k)$$

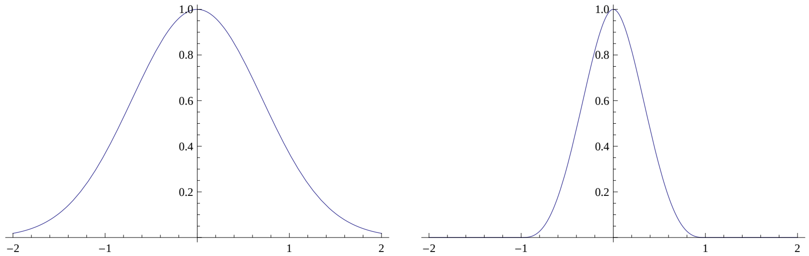
is positive.

By definition $A_{\Phi,X}$ is symmetric. If it is positive definite, then the interpolant will be defined. In this way, we can also say that Φ is positive definite when the interpolation matrix $A_{\Phi,X}$ is positive definite.

Not every RBF used for interpolation is a positive definite function, although the corresponding quadratic form is positive for some expansion coefficients. In general, RBF interpolation uses a conditionally positive definite function of some order.

Definition 3. Let $m \in \mathbb{N}$. A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is conditionally positive definite of order m on $\mathbb{R}^d$ if, for all $N \in \mathbb{N}$, all pairwise distinct $x_1, \ldots, x_N \in \mathbb{R}^d$ and all $\alpha \in \mathbb{R}^N \setminus \{0\}$ satisfying

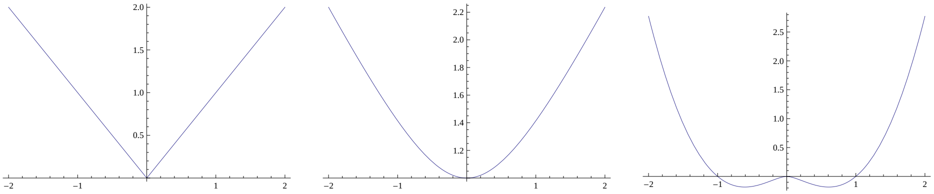
$$\sum_{j=1}^{N} \alpha_j p(x_j) = 0$$



$\phi_1(|x|) = e^{-|x|^2}$
Gaussian RBF

$\phi_2(|x|) = (1 - |x|)_+^3(3|x| + 1)$
C^2-Wendland RBF

Figure 1: Two positive definite functions on $\mathbb{R}$.



$\phi_3(|x|) = |x|$
Linear RBF
$(m = 1)$

$\phi_4(|x|) = \sqrt{1 + |x|^2}$
Multiquadric RBF
$(m = 1)$

$\phi_5(|x|) = |x|^2 \log(|x|)$
Thin Plate Spline RBF
$(m = 2)$

Figure 2: Three conditionally positive definite functions on $\mathbb{R}$.

for all real-valued polynomials of degree less than m, the quadratic form

$$\sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_j \alpha_k \Phi(x_j - x_k)$$

is positive.

For any $m \in \mathbb{N}$, we denote by $\pi_{m-1}(\mathbb{R}^d)$ the space of polynomial functions defined over $\mathbb{R}^d$ of degree $\leq m - 1$ with respect to the set of variables. If we want to interpolate f using a conditionally positive definite function of order m, we will look for an interpolant of the form

$$s_{f,X}(x) = \sum_{j=1}^{N} \alpha_j \Phi(x - x_j) + \sum_{k=1}^{Q} \beta_k p_k, \qquad x \in \mathbb{R}^d, \tag{2}$$

where $\{p_1, \ldots, p_Q\}$ is a basis of the polynomial space $\pi_{m-1}(\mathbb{R}^d)$.

The coefficients $\alpha = (\alpha_1, \ldots, \alpha_N)^\top$ and $\beta = (\beta_1, \ldots, \beta_Q)^\top$ in (2) are uniquely determined by (1) and the additional conditions

$$\sum_{j=1}^{N} \alpha_j p_k(x_j) = 0, \qquad 1 \leq k \leq Q.$$

If we define the matrix $P = (p_k(x_j)) \in \mathbb{R}^{N\times Q}$, α and β will be the solution of the system

$$\begin{pmatrix} A_{\Phi,X} & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} F \\ 0 \end{pmatrix}.$$

In this paper, we consider the RBF interpolation on $\mathbb{R}$, i.e. for the case $d = 1$. For more details and proofs, revise [7].

§3. Interpolation of a discontinuous function

In this section, we study features of an interpolant $s_{f,X}$ for a piecewise function

$$f(x) = \begin{cases} f_1(x), & -1 \le x < 0, \\ f_2(x), & 0 < x \le 1, \end{cases} \tag{3}$$

with f_1 and f_2 continuous, and such that it has a finite jump discontinuity at $x_c = 0$, i.e. $|f^+ - f^-| \neq 0$, where $f^- = \lim_{x\to 0^-} f(x)$ and $f^+ = \lim_{x\to 0^+} f(x)$.

First we present a study of RBF interpolation with two centers near discontinuity and then we make a general study with N centers in $[-1, 1]$.

3.1. Local performance of interpolation

We now select two centers in a small neighbourhood of the discontinuity. Let $X = \{-\delta/2, \delta/2\}$ for $\delta > 0$. Most RBFs produce a strictly monotone interpolant $s_\delta(x)$ defined in $[-\delta/2, \delta/2]$. By definition, $s_\delta(x)$ is continuous, so we can then evaluate it at $x_c = 0$:

- If Φ is positive definite, we will get as interpolant

  $$s_\delta(x) = \alpha_1 \Phi(x + \delta/2) + \alpha_2 \Phi(x - \delta/2),$$

  where

  $$\alpha_1 = \frac{f(\delta/2)\Phi(\delta) - f(-\delta/2)\Phi(0)}{\Phi^2(\delta) - \Phi^2(0)} \quad \text{and} \quad \alpha_2 = \frac{f(-\delta/2)\Phi(\delta) - f(\delta/2)\Phi(0)}{\Phi^2(\delta) - \Phi^2(0)}.$$

  Then

  $$s_\delta(0) = (f(\delta/2) + f(-\delta/2))\frac{\Phi(\delta/2)}{\Phi(\delta) + \Phi(0)}.$$

- If Φ is conditionally positive definite of order one, we will get as interpolant

  $$s_\delta(x) = \alpha_1 \Phi(x + \delta/2) + \alpha_2 \Phi(x - \delta/2) + \beta_1,$$

  where

  $$\alpha_1 = \frac{f(\delta/2) - f(-\delta/2)}{2(\Phi(\delta) - \Phi(0))} = -\alpha_2 \quad \text{and} \quad \beta_1 = \frac{f(-\delta/2) + f(\delta/2)}{2}.$$

  Then

  $$s_\delta(0) = \frac{f(\delta/2) + f(-\delta/2)}{2}.$$

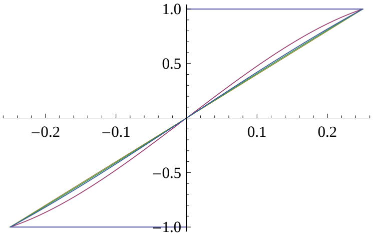


Figure 3: $s_\delta(x)$, with $\delta = 1/2$, for $f_1(x) = -1$ and $f_2(x) = 1$.

- If Φ is any conditionally positive definite of order two, we will get as interpolant

  $$s_\delta(x) = \alpha_1\Phi(x+\delta/2) + \alpha_2\Phi(x-\delta/2) + \beta_2 x + \beta_1,$$

  where

  $$\alpha_1 = 0 = \alpha_2, \quad \beta_1 = \frac{f(\delta/2)+f(-\delta/2)}{2} \quad \text{and} \quad \beta_2 = \frac{f(\delta/2)-f(-\delta/2)}{\delta}.$$

  Then

  $$s_\delta(0) = \frac{f(\delta/2)+f(-\delta/2)}{2}.$$

Let us observe that, if Φ is any conditionally positive definite function of a higher order, we will not get a unique interpolant. In Figure 3, we show interpolants $s_\delta(x)$, with $\delta = 1/2$, for the fuctions ϕ_1, ϕ_2, ϕ_3 and ϕ_4 defined in Figures 1 and 2, with $f_1(x) = -1$ and $f_2(x) = 1$. The graphic shows that interpolants are strictly increasing and $s_\delta(0) = 0$.

3.2. Interpolation with N centers

We reduce interpolation study to an even number N of centers $X = \{x_1, \ldots, x_N\}$, but the same results are obtained for an odd N.

We consider that centers are uniformly distributed in $[-1,1]$, that is, for $j = 1, \ldots, N$, $x_j = -1 + 2(j-1)/(N-1)$. Discontinuity exists at $x_c = (x_{N/2} + x_{N/2+1})/2 = 0$. Any RBF used to interpolate produces a continuous interpolant $s_{f,X}$, defined in Section 2. For most RBFs of [2, Appendix D], $s_{f,X}$ has the same features. We have obtained lots of examples, using the mentioned RBFs, for different step functions and functions in Example 2. The next two examples show the interpolant features.

Example 1. Let f be given by (3) with $f_1(x) = -1$ and $f_2(x) = 1$. We interpolate it with $N =$ 4, 16, 32, 64 and 128, using the RBFs ϕ_2, ϕ_3, $\widetilde{\phi}(r) = \phi_4(\sqrt{50}r)$ and ϕ_5 (see Figure 4). We have modified ϕ_4 to get good interpolation matrices in the sense that Mathematica is able to solve the associated systems. We observe that $s_{f,X}$ is strictly increasing in $(x_{N/2}, x_{N/2+1})$. In addtion, ϕ_3-interpolants do not present oscillations near the discontinuity. In fact, by definition of f in this example, Jung [6] shows that any ϕ_3-interpolant is

$$s_{f,X}(x) = \begin{cases} -1, & x < x_{N/2}, \\ (N-1)x, & x_{N/2} \le x \le x_{N/2+1}, \\ 1, & x > x_{N/2+1}. \end{cases}$$

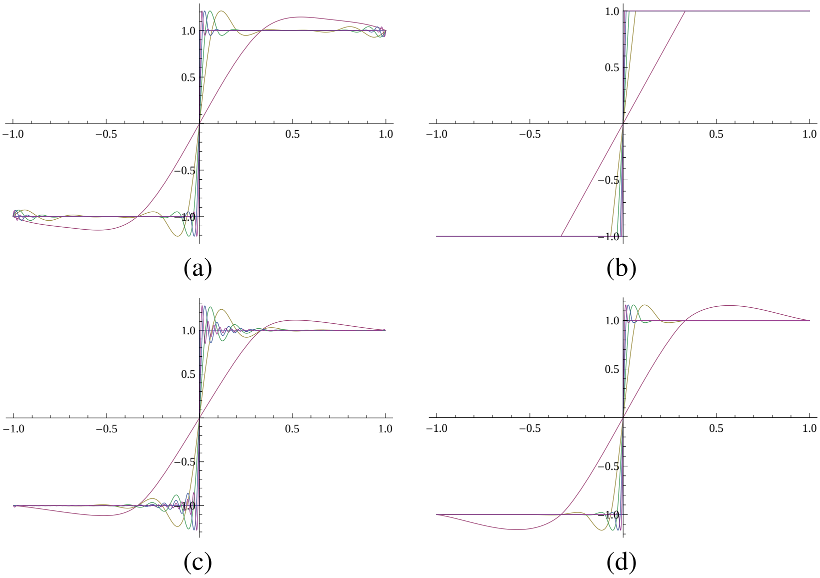


Figure 4: Interpolants of the function f given in Example 1. The used RBFs are ϕ_2 in (a), ϕ_3 in (b), $\widetilde{\phi}$ in (c) and ϕ_5 in (d).

Example 2. We consider the two non-step functions

$$g_1(x) = \begin{cases} \sin x, & x < 0, \\ \cos x, & x > 0, \end{cases} \qquad \text{and} \qquad g_2(x) = \begin{cases} \log(1-x), & x < 0, \\ 0.5(x-0.5)^3, & x > 0, \end{cases}$$

and we interpolate them with $N = 4, 16, 32, 64$, and 128 centers. The function g_1 has also been considered in [6].

Figure 5 shows several interpolants of g_1 and g_2, using ϕ_3 as RBF. These interpolants do not present oscillations. They are polygonal functions with vertices at $(x_i, f(x_i))$ for $i = 1, \ldots, N$, and so they are not differential functions at vertices. Therefore, ϕ_3-interpolants are not good approximations of functions. In Figure 6, we show interpolants of g_1 on top and of g_2 on the bottom. We use $\widetilde{\phi}$ at (a) and (d), ϕ_2 at (b) and (e), and ϕ_5 at (c) and (f).

Numerical experiments for not oscillatory differentiable RBFs of [2, Appendix D] yield interpolants with the same features:

- The interpolant of f has oscillations near x_c. Oscillations do not disappear even for high values of N, Gibbs phenomenon, but increase up to a limit. Maximum oscillations are located in $(x_{N/2-1}, x_{N/2})$ and $(x_{N/2+1}, x_{N/2+2})$.
- $s_{f,X}$ is a strictly increasing monotone function in $(x_{N/2}, x_{N/2+1})$ if $f(x_{N/2}) < f(x_{N/2+1})$ and strictly decreasing if $f(x_{N/2}) > f(x_{N/2+1})$.
- The expansion coefficient α_i is related to the center x_i, for $i = 1, \ldots, N$. Taking centers each time close to x_c the absolute values of associated expansion coefficients become much bigger than at the boundary.

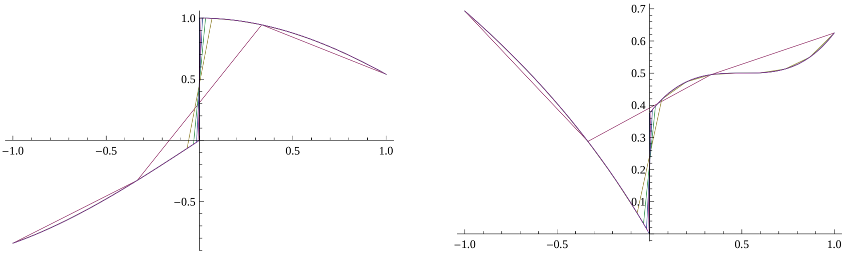


Figure 5: ϕ_3-interpolants of the functions g_1 (left) and g_2 (right) given in Example 2.

Now we define a rate R to measure maximum oscillation on the right of discontinuity. Let $s^\star$ be the value of the interpolant at maximum oscillation located in $(x_{N/2+1}, x_{N/2+2})$. For step functions f with $f_1(x) = f^-$ and $f_2(x) = f^+$, we define the ratio R between the maximum absolute value of over/under-shoots and the jump discontinuity by

$$R = \frac{|s^\star - f^+|}{|f^+ - f^-|}. \tag{4}$$

We consider different step functions and compute R, i.e. oscillations performance, with different RBFs, number of centers and jump discontinuities. Table 1 collects this information and shows that maximum oscillation limit depends on the discontinuity jump and the RBF used, for a given N. Values of Table 1 point out that R is a relative measure of the maximum oscillation since R is invariant for fixed N and RBF. This means that R does not depend on the jump discontinuity for fixed N and RBF. Looking through Table 1, we can affirm that the interpolation using Φ_5 produces a maximum oscillation limit about 8% of jump.

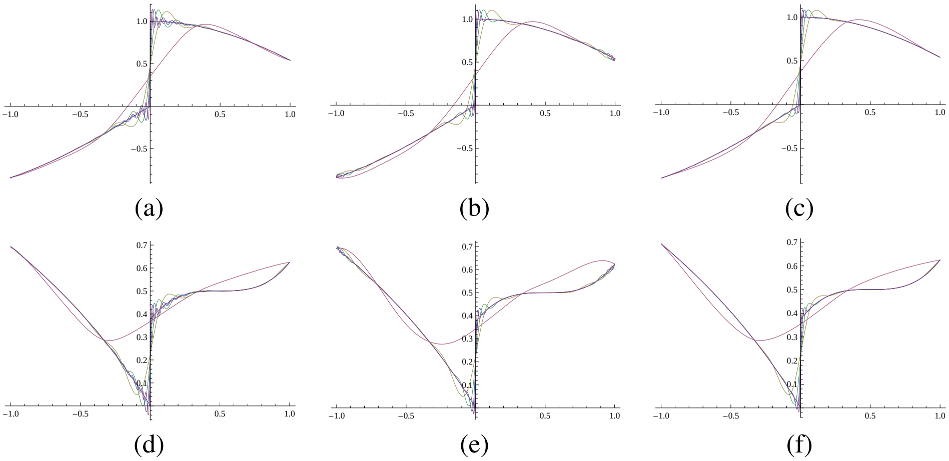
Remark 1. All results in this section could also be obtained for any interval and with a discontinuity at another point.

§4. Local piecewise linear interpolation

In the previous section, we have described the behaviour of the interpolant $s_{f,X}$ of a function with a discontinuity for N centers uniformly distributed. The interpolant does not reproduce the discontinuity of function and the Gibbs phenomenon appears.

Anyway, we observe a special performance of interpolant using ϕ_3 as RBF: $s_{f,X}$ has no oscillation because it is a piecewise linear function.

Looking through Fornberg's paper [3], we confirm that RBF expansion coefficients are bigger near discontinuity. Moreover, Jung [6] gives a method to eliminate oscillations of interpolant using multiquadrics. Jung's paper adapts the interpolation by changing the shape parameter of multiquadrics, $\gamma = 0$, at centers with expansion coefficients in absolute value bigger than that at the boundary. This is changing multiquadric by linear RBF, ϕ_3. We realize that it is enough to change RBF at centers next to the discontinuity: $x_{N/2}$ and $x_{N/2+1}$. We can eliminate oscillations using ϕ_3 only at those centers and most RBFs of [2, Appendix D] at the other centers.



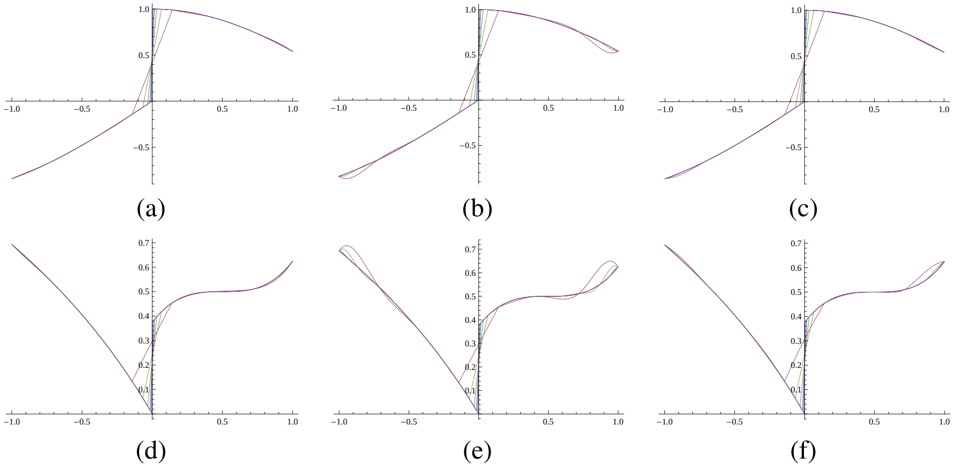
(a) (b) (c)

(d) (e) (f)

Figure 6: Interpolants $s_{g_1,X}$ (top row) and $s_{g_2,X}$ (bottom row). The used RBFs are $\widetilde{\phi}$ in (a) and (d), ϕ_2 in (b) and (e), and ϕ_5 in (c) and (f).

RBF	(f^-, f^+)	$N = 4$	$N = 16$	$N = 32$	$N = 64$	$N = 128$
Φ_2	$(-1, 1)$	0.07269	0.10546	0.10538	0.10540	0.10545
	$(0, 1)$	0.07269	0.10546	0.10538	0.10540	0.10545
	$(-1.5, 1.5)$	0.07269	0.10546	0.10538	0.10540	0.10545
	$(-0.4, 0.4)$	0.07269	0.10546	0.10538	0.10540	0.10545
$\widetilde{\Phi}$	$(-1, 1)$	0.05727	0.11899	0.13324	0.13877	0.14041
	$(0, 1)$	0.05727	0.11899	0.13324	0.13877	0.14036
	$(-1.5, 1.5)$	0.05727	0.11899	0.13324	0.13877	0.14055
	$(-0.4, 0.4)$	0.05727	0.11899	0.13324	0.13877	0.14029
Φ_5	$(-1, 1)$	0.07740	0.08046	0.08046	0.08046	0.08046
	$(0, 1)$	0.07741	0.08046	0.08046	0.08046	0.08046
	$(-1.5, 1.5)$	0.07740	0.08046	0.08046	0.08046	0.08046
	$(-0.4, 0.4)$	0.07740	0.08046	0.08046	0.08046	0.08046

Table 1: Values of R for different RBF, (f^-, f^+) and N



(a) (b) (c)

(d) (e) (f)

Figure 7: Interpolants $\widetilde{s}_{g_1,X}$ (top row) and $\widetilde{s}_{g_2,X}$ (bottom row). The used RBFs are $\widetilde{\phi}$ in (a) and (d), ϕ_2 in (b) and (e), and ϕ_5 in (c) and (f).

In the conditions described in Section 3, we seek an interpolant $\widetilde{s}_{f,X}$, using ϕ_3 at centers next to discontinuity, of the form

$$\widetilde{s}_{f,X}(x) = \sum_{\substack{j=1 \\ j\neq N/2, N/2+1}}^{N} \widetilde{\alpha}_j\, \phi(|x - x_j|) + \sum_{j=1}^{2} \widetilde{\alpha}_{N/2-1+j}\, |x - x_{N/2-1+j}| + \sum_{k=1}^{\widetilde{m}} \lambda_k p_k, \qquad x \in \mathbb{R},$$

where $\{p_1, \dots, p_{\widetilde{m}}\}$ is a basis of the polynomial space $\pi_{\widetilde{m}-1}(\mathbb{R})$. The coefficients $\widetilde{\alpha}_1, \dots, \widetilde{\alpha}_N$ and $\lambda_1, \dots, \lambda_{\widetilde{m}}$ are determined by (1) and the additional conditions

$$\sum_{j=1}^{N} \widetilde{\alpha}_j p_k(x_j) = 0, \qquad 1 \le k \le \widetilde{m}.$$

We use $\widetilde{m} = 1$ for Φ positive definite and $\widetilde{m} = m$ for Φ conditionally positive definite of order m. Finally we add the constant needed by the linear RBF ϕ_3.

Next, we present two examples. Example 3 shows graphical behaviour of this method for two functions studied in the previous section. Example 4 provides some errors at some distance from discontinuity to show the fitting of the new interpolant.

Example 3. We apply this technique to Example 2 to eliminate oscillations of the interpolants in Figure 6. In Figure 7, we observe that the oscillations are eliminated and interpolants fit better to the function at $[-1, x_{N/2}] \cup [x_{N/2+1}, 1]$. This technique eliminates oscillations because we get an interpolant that is a straight line by $(x_{N/2}, f(x_{N/2}))$ and $(x_{N/2+1}, f(x_{N/2+1}))$ in $[x_{N/2}, x_{N/2+1}]$.

RBF	$x = \frac{-41}{46}$	$x = \frac{-24}{46}$	$x = \frac{-8}{46}$	$x = \frac{-7}{46}$	$x = \frac{-5}{46}$	$x = \frac{-4}{46}$	$x = \frac{-3}{46}$	$x = 0$
$\widetilde{\phi_1}(r) = \phi_1(4.1r)$	1.95e−4	2.00e−5	7.21e−6	3.16e−6	5.07e−6	1.97e−5	3.02e−5	8.16e−5
$\widetilde{\phi_2}(r) = \phi_2(1.3r)$	1.12e−3	3.41e−6	2.30e−4	1.79e−4	1.16e−4	1.49e−4	1.12e−4	2.20e−4
$\widetilde{\phi_4}(r) = \phi_4(2r)$	7.31e−8	1.46e−6	1.41e−5	1.31e−5	2.47e−5	5.12e−5	5.60e−5	3.60e−5

Table 2: Values of E for different points and RBFs.

Example 4. Let

$$g_3(x) = \begin{cases} \exp(x), & x \in [-1, 0), \\ 3, & x = 0, \\ 5 + \sin x, & x \in (0, 1], \end{cases}$$

be a function given in [4]. We apply the described technique with $N = 24$ centers for different RBFs. Let $E(x) = \left|f(x) - \widetilde{s}_{g_3,X}(x)\right|$ be the error function. It is obvious that $E(x_i) = 0$ for $i = 1, \ldots, N$. Errors close to 1 occur at next to discontinuity due to the approximation of the technique near to discontinuity. Table 2 shows the values of E for different points and RBFs. We observe that these errors are similar to the ones obtained in [4] for the same example.

Finally, as conclusions, we have investigated the Gibbs phenomenon for 1D RBF interpolation numerically, and proposed a procedure to reduce oscillations using nonsmooth basis functions locally. This technique is the first step of an approximation method of discontinuous functions which we plan to develop in the future.

Acknowledgements

This work has been partially supported by the Ministry of Science and Innovation (Spain), through the Research Project MTM2009-07315 and the Government of Aragón through two Research groups, E65 and T68, and Research Network CTPR04/10. The authors want to thank Juan José Torrens for his invaluable help. They also thank María Moncayo for suggestions and support and appreciate the comments of the anonymous referee.

References

[1] Amat, S., Donat, R., Liandrat, J., and Trillo, C. Analysis of a new nonlinear subdivision scheme. Applications in image processing. *Found. Comput. Math. 6* (2006), 193–225.

[2] Fasshauer, G. E. *Meshfree Approximation Methods with Matlab*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007.

[3] Fornberg, B., and Flyer, N. The Gibbs phenomenon for radial basis functions. In *The Gibbs Phenomenon in Various Representations and Applications*. Sampling Publishing, Potsdam (New York), 2005, pp. 201–224.

[4] Guessab, A., Moncayo, M., and Schmeisser, G. A class of nonlinear four-point subdivision schemes properties in terms of conditions. *Adv. Comput. Math.* (In press).

[5] Hardy, R. L. Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res. 76* (1971), 1905–1915.

[6] Jung, J.-H. A note on the gibbs phenomenon with multiquadric radial basis functions. *Appl. Numer. Math. 57* (2007), 213–229.

[7] Wendland, H. *Scattered Data Approximation*. Cambridge Univ. Press, Cambridge, 2005.

Diego Izquierdo, María Cruz López de Silanes and María Cruz Parra
Departamento de Matemática Aplicada
Universidad de Zaragoza
C/ María de Luna 3. 50018 Zaragoza. Spain
`dizquier@unizar.es`, `mcruz@unizar.es` , `cparra@unizar.es`

Monografías Matemáticas *García de Galdeano* **37**, 161–170 (2012)

1D NUMERICAL SIMULATION FOR NONLINEAR PSEUDOPARABOLIC PROBLEMS

Robert Luce, Ngonn Seam and Guy Vallet

Abstract. In this paper, we are interested in the numerical simulation of a pseudo-parabolic fully-nonlinear equation with a nonlinear term of Barenblatt's type. We are exactly interested in the illustrations of the solution of the boundary-value problem: find u such that

$$f(u_t) - \operatorname{div}\{a(u)\nabla u + b(u)\nabla u_t\} = g.$$

The mathematical analysis of a close problem and its simulation have recently been studied by S. N. Antontsev *et al.* [3] when $f = Id_R$ and the existence result has been generalized by N. Seam and G. Vallet in [8]. We propose in particular simulations of the nonlinear problem of the Barenblatt's type: $f(u_t) - \Delta u - \epsilon\Delta u_t = g$ (see [1]).

Keywords: Pseudoparabolic problems, numerical simulations, Barenblatt's problem.

AMS classification: 35K65, 35K70.

§1. Introduction

In this paper, we deal with the 1D numerical simulation to the fully-nonlinear pseudoparabolic problem:

$$f\left(\frac{\partial u}{\partial t}\right) - \frac{\partial}{\partial x}\left\{a(u)\frac{\partial u}{\partial x} + b(u)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right)\right\} = g \text{ in } Q,\ u|_\Gamma = 0 \text{ and } u(0,\cdot) = u_0, \tag{1}$$

where f is a Lipschitz-continuous and increasing function, a is Lipschitz-continuous and bounded and b is a positive Lipschitz-continuous and bounded function.

Problems close to that one have been previously studied by S. N. Antontsev *et al.* [3] for stratigraphic models by the way of an implicit time-discretization, and has recently been generalized by N. Seam and G. Vallet in [8] by the same way (see [6, 7] too). The existence of the solution at each step of the discretized scheme is based on Schauder-Tikhonov's fixed point theorem and the convergence of the scheme on an adapted compactness argument.

Our aim is then to illustrate the solution of the above problem by a standard P_1-finite element method in space and an implicit time discretization. In particular, we are interested in the pseudoparabolic singular perturbation when the molecular diffusion changes sign. To do this, we have modified the codes developed by Alberty [2] for the diffusion-reaction problem.

Let us denote by $\Omega =]x_l, x_r[$ a bounded interval of $\mathbb{R}$, T a positive number and assume the following assumptions:

(H$_1$) a and b are Lipschitz continuous functions over $\mathbb{R}$ such that

$$\exists\beta, M > 0,\ \forall u \in \mathbb{R},\ |a(u)| \le M,\ \beta \le b(u) \le M.$$

(H$_2$) f is a Lipschitz continuous and nondecreasing function over $\mathbb{R}$.

(H$_3$) $g \in L^2(Q)$ and $u_0 \in H_0^1(\Omega)$.

Then, one would say that

Definition 1. A solution of the problem (1) is $u \in H^1\left(0, T; H_0^1(\Omega)\right)$ such that for all $v \in H_0^1(\Omega)$ and $t \in]0, T[$ *a.e.*,

$$\int_{x_l}^{x_r} \left\{ f\left(\frac{\partial u}{\partial t}\right) v + a(u)\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + b(u)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right)\frac{\partial u}{\partial x} \right\} dx = \int_{x_l}^{x_r} gv\, dx$$

with the initial condition $u(0, \cdot) = u_0$.

Let us recall a theorem concerning the existence and uniqueness:

Theorem 1 (N. Seam and G. Vallet [8]). *Under hypotheses (H$_1$), (H$_2$) and (H$_3$), there exists u in $H^1\left(0, T, H_0^1(\Omega)\right)$ such that for all v in $H_0^1(\Omega)$ and t almost everywhere in* $]0, T[$,

$$\int_{x_l}^{x_r} \left\{ f\left(\frac{\partial u}{\partial t}\right) v + a(u)\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + b(u)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right)\frac{\partial u}{\partial x} \right\} dx = \int_{x_l}^{x_r} gv\, dx \;\; \text{with} \;\; u(0,\cdot) = u_0 \qquad (2)$$

§2. 1D finite elements formulation

Let us remark that the problem can be strongly non linear and generally the explicit formulation fails because of very restrictive conditions of C.F.L type. So, an implicit formulation has been chosen to obtain solutions with reasonable time steps.

For any $N_t \in \mathbb{N}^*$ and all $k \in [0, N_t]$, let us denote by $\Delta t = T/N_t$ and $t_k = k\Delta t$. Thus, the implicit time discretization of the problem (2) is: find u^{k+1} in $H_0^1(\Omega)$ for a given u^k in $H_0^1(\Omega)$ such that, for any $v \in H_0^1(\Omega)$,

$$\int_{x_l}^{x_r} f\left(\frac{u^{k+1} - u^k}{\Delta t}\right) v\, dx + \int_{x_l}^{x_r} a\left(u^{k+1}\right)\frac{\partial u^{k+1}}{\partial x}\frac{\partial v}{\partial x}\, dx$$
$$+ \int_{x_l}^{x_r} b\left(u^{k+1}\right)\frac{\partial}{\partial x}\left(\frac{u^{k+1} - u^k}{\Delta t}\right)\frac{\partial v}{\partial x}\, dx = \int_{x_l}^{x_r} g^{k+1} v dx, \quad k \in [0, N_t - 1],$$

where g^{k+1} is an approximation of g at time t_{k+1}.
The formulation can be written

$$\int_{x_l}^{x_r} f\left(\frac{u^{k+1} - u^k}{\Delta t}\right) v\, dx + \int_{x_l}^{x_r} \left[a\left(u^{k+1}\right) + \frac{1}{\Delta t} b\left(u^{k+1}\right)\right]\frac{\partial u^{k+1}}{\partial x}\frac{\partial v}{\partial x}\, dx$$
$$- \frac{1}{\Delta t}\int_{x_l}^{x_r} b\left(u^{k+1}\right)\frac{\partial u^k}{\partial x}\frac{\partial v}{\partial x}\, dx - \int_{x_l}^{x_r} g^{k+1} v\, dx = 0, \quad k \in [0, N_t - 1]. \qquad (3)$$

Now, for any $N_x \in \mathbb{N}$, denote by $h = \Delta x = (x_r - x_l)/(N_x + 1)$ for a uniform mesh with $x_0 = x_l$, and $x_{N_x+1} = x_r$. Thus $x_i = x_0 + ih$ for $i \in [0, N_x + 1]$. We construct the finite dimensional space V_h formed of linear piecewise polynomials:

$$V_h = \left\{ v_h \in H_0^1(\Omega);\; v_h|_{[x_i, x_{i+1}]} \in \mathbb{P}_1,\; 0 \le i \le N_x;\; v_h(x_l) = v_h(x_r) = 0 \right\}.$$

Clearly, $V_h = \operatorname{span}\{\phi_1, \phi_2, \ldots, \phi_{N_x}\}$, where the ϕ_i's are the hat functions, and $\dim V_h = N_x$. By using V_h in place of $H_0^1(\Omega)$, the approximation by the finite element of the problem (3) can be written: find $u_h^{k+1} \in V_h$ for a giving $u_h^k \in V_h$ such that for all $v_h \in V_h$

$$\int_{x_l}^{x_r} f\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}\right) v_h \, dx + \int_{x_l}^{x_r} \left[a\left(u_h^{k+1}\right) + \frac{b\left(u_h^{k+1}\right)}{\Delta t}\right] \frac{\partial u_h^{k+1}}{\partial x} \frac{\partial v_h}{\partial x} \, dx$$
$$- \int_{x_l}^{x_r} \frac{b\left(u_h^{k+1}\right)}{\Delta t} \frac{\partial u_h^k}{\partial x} \frac{\partial v_h}{\partial x} \, dx - \int_{x_l}^{x_r} g^{k+1} v_h \, dx = 0, \quad k = 0, 1, \ldots, N_t - 1,$$

For $k \in [0, N_t]$, inserting $u_h^{k+1} = \sum_{j=1}^{N_x} u_j^{k+1} \phi_j$ with a given approximation $u_h^0 = \sum_{j=1}^{N_x} u_j^0 \phi_j$ of u_0 and using, for $i \in [1, N_x]$, that ϕ_i as an admissible test function, we get the nonlinear system

$$\int_{x_l}^{x_r} f\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}\right) \phi_i \, dx - \frac{1}{\Delta t} \int_{x_l}^{x_r} b\left(u_h^{k+1}\right) \left(u_h^k\right)' \phi_i' \, dx$$
$$+ \int_{x_l}^{x_r} \left[a\left(u_h^{k+1}\right) + \frac{b\left(u_h^{k+1}\right)}{\Delta t}\right] \left(u_h^{k+1}\right)' \phi_i' \, dx - \int_{x_l}^{x_r} g^{k+1} \phi_i \, dx = 0, \quad k \in [0, N_t], \ i \in [1, N_x].$$

The nonlinear system can be usually solved by the Newton Raphson method (cf. [4, 9]).
In this case, for $k \in [0, N_t]$, we denote by $U_h^{k+1} = \left(u_1^{k+1}, u_2^{k+1}, \ldots, u_{N_x}^{k+1}\right)^T$ and we introduce the function $F : \mathbb{R}^{N_x} \to \mathbb{R}^{N_x}$, $U_h^{k+1} \mapsto F_i\left(U_h^{k+1}\right)$ for $[1, N_x]$, defined by the formula

$$F_i\left(U_h^{k+1}\right) = \int_{x_l}^{x_r} f\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}\right) \phi_i \, dx - \frac{1}{\Delta t} \int_{x_l}^{x_r} b\left(u_h^{k+1}\right) \left(u_h^k\right)' \phi_i' \, dx$$
$$+ \int_{x_l}^{x_r} \left[a\left(u_h^{k+1}\right) + \frac{b\left(u_h^{k+1}\right)}{\Delta t}\right] \left(u_h^{k+1}\right)' \phi_i' \, dx - \int_{x_l}^{x_r} g^{k+1} \phi_i \, dx.$$

Thus, we have to solve the nonlinear system $F\left(U_h^{k+1}\right) = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^{N_x}$, by the Newton Raphson algorithm (see [4], [5] and [9] for the details):

1. For $k \in [0, N_t]$, we initialize the vector U_h^k,
2. then, we compute U_h^{k+1}, solution to the linear system in the Newton method,
3. we give a initial estimation $U_h^{k+1,0}$ of U_h^{k+1},
4. for $\ell = 0, 1, 2, \ldots, \ell_{\max}$, we compute $\Delta U_h^{k+1,\ell}$, solution to the linear system

   $$F'\left(U_h^{k+1,\ell}\right) \Delta U_h^{k+1,\ell} = -F\left(U_h^{k+1,\ell}\right),$$

   where $F'\left(U_h^{k+1,\ell}\right)$ is the Jacobian of F at point $U_h^{k+1,\ell}$,
5. we finally let $U_h^{k+1,\ell+1} = U_h^{k+1,\ell} + \Delta U_h^{k+1,\ell}$.

By definition of the Jacobian,

$$F'_{ij}\left(U_h^{k+1,\ell}\right) = \frac{\partial F_i}{\partial u_j^{k+1,\ell}}\left(U_h^{k+1,\ell}\right)$$

and we get that

$$\begin{aligned}F'_{ij}\left(U_h^{k+1,\ell}\right) = \frac{1}{\Delta t}\int_{x_l}^{x_r} f'\left(\frac{u_h^{k+1,\ell} - u_h^{k,\ell}}{\Delta t}\right)\phi_i\phi_j\,dx - \frac{1}{\Delta t}\int_{x_l}^{x_r} b'\left(u^{k+1,\ell}\right)\left(u_h^{k,\ell}\right)'\phi'_i\phi_j\,dx \\ + \int_{x_l}^{x_r}\left[\phi_j a'\left(u_h^{k+1,\ell}\right)\left(u_h^{k+1,\ell}\right)' + a\left(u_h^{k+1}\right)\phi'_j\right]\phi'_i\,dx \\ + \frac{1}{\Delta t}\int_{x_l}^{x_r}\left[\phi_j b'\left(u_h^{k+1,\ell}\right)\left(u_h^{k+1,\ell}\right)' + b\left(u_h^{k+1,\ell}\right)\phi'_j\right]\phi'_i\,dx.\end{aligned}$$

Thus, we can compute the coefficient matrix and the right-hand side matrix.

§3. Numerical simulations

In this section, we illustrate the solution to the problem (1) with different given data. In the following examples, $]x_l, x_r[\, = \,]-\pi, \pi[$ and

- either $u_0 = 0$ and $g(t,x) = 1$ if $x \in [\pi/4, \pi/2]$, $g(t,x) = -1$ if $x \in [-\pi/2, -\pi/4]$, $g(t,x) = 0$ otherwise (configuration 1),
- or $u_0(x) = 4x/\pi$ if $x \in [-\pi/4, \pi/4]$, $u_0(x) = 2-4x/\pi$ if $x \in\,]\pi/4, \pi/2]$, $u_0(x) = -2-4x/\pi$ if $x \in [-\pi/2, -\pi/4[$, $u_0(x) = 0$ otherwise, and $g(t,x) = 0$ (configuration 2).

3.1. Linear pseudoparabolic equation or Sobolev' equation

Here, $f(r) = r$, $a(r) = 1$ and $b(r) = \tau$ with $\tau = 0, 1/2, 1$ and 5. We present the simulation of configuration 1 (*i.e.* $u_0 = 0$) in Figure 1 and that of Configuration 2 (*i.e.* $u_0 \neq 0$) in Figure 2.

Remark first that the pseudoparabolic perturbation slows down the evolution of the system. The second remark concerns the space regularity of the solution for $t > 0$: in the pseudoparabolic case, the initial condition fixes the regularity of the solution. Indeed, the first step in the time-iteration solves the elliptic problem: $u - (\Delta t + \tau)\,\Delta u = \Delta_t g + u_0 - \tau\Delta u_0$. Consequently, if $\tau > 0$ and if u_0 is in $H_0^1(\Omega)$, it will be the same for the solution u.

In Figures 3 and 4, we illustrate the same problem unless b where $b(r) = 0.1$ if $r < 0$, $b(r) = k$ else. We can see the dissymmetry of the solution.

3.2. Nonlinear pseudoparalolic equation

In Figures 5 and 6, $f(r) = r$, $a(r) = \arctan(r)$ and $b(r) = \tau$ where $\tau = 0.1, 0.2, 0.5, 1$. Since the sign of a changes, we observe diffusive and anti-diffusive effects illustrated by a convergence to a Dirac mass, especially for small ϵ.

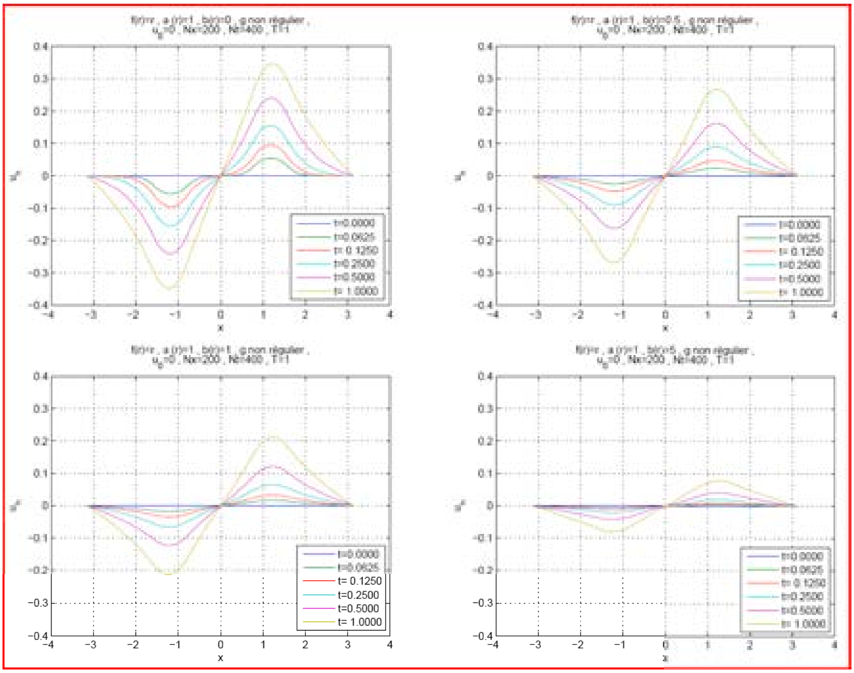


Figure 1: $\partial_t u - \partial^2_{xx} u - \tau \partial^3_{xxt} u = g$ with $u_0(x) = 0$, $\tau = 0, 1/2, 1, 5$.

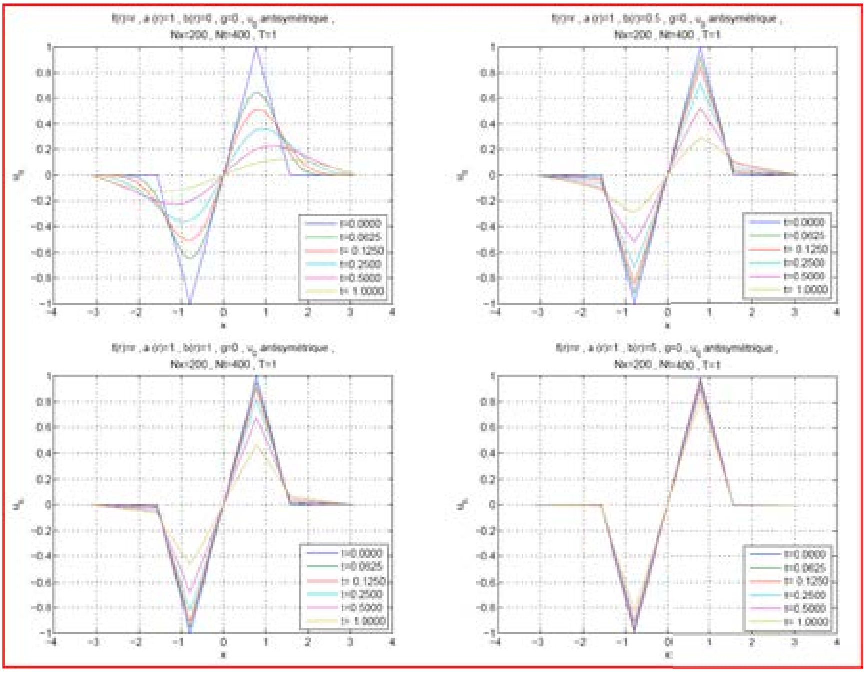


Figure 2: $\partial_t u - \partial^2_{xx} u - \tau \partial^3_{xxt} u = 0$ with $u_0(x) \neq 0$, $\tau = 0, 1/2, 1, 5$.

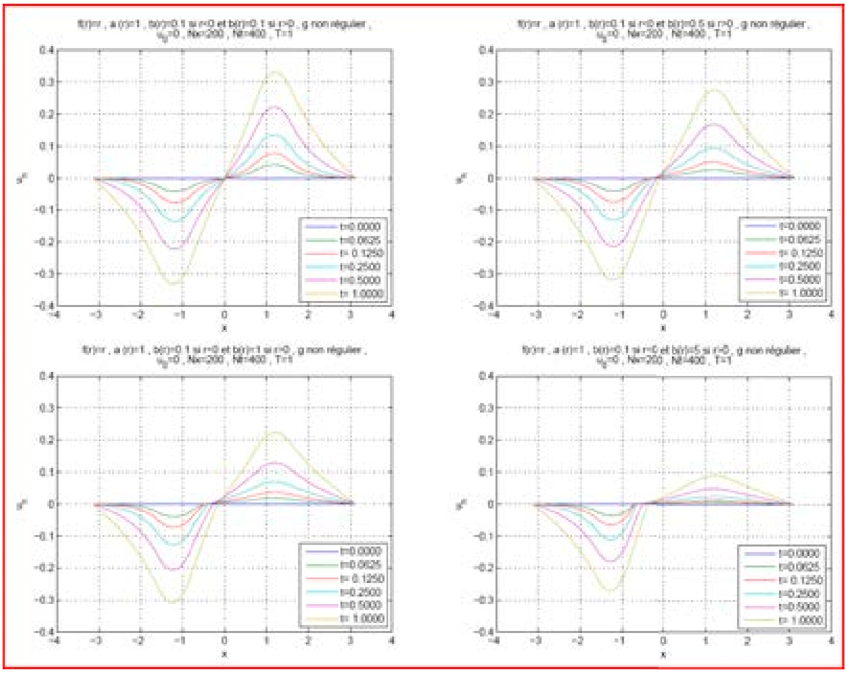


Figure 3: $\partial_t u - \partial^2_{xx} u - \tau \partial^3_{xxt} u = g$ with $u_0(x) = 0$, $b(r) = \tau r^+ - 0.1 r^-$, $\tau = 0, 1/2, 1, 5$.

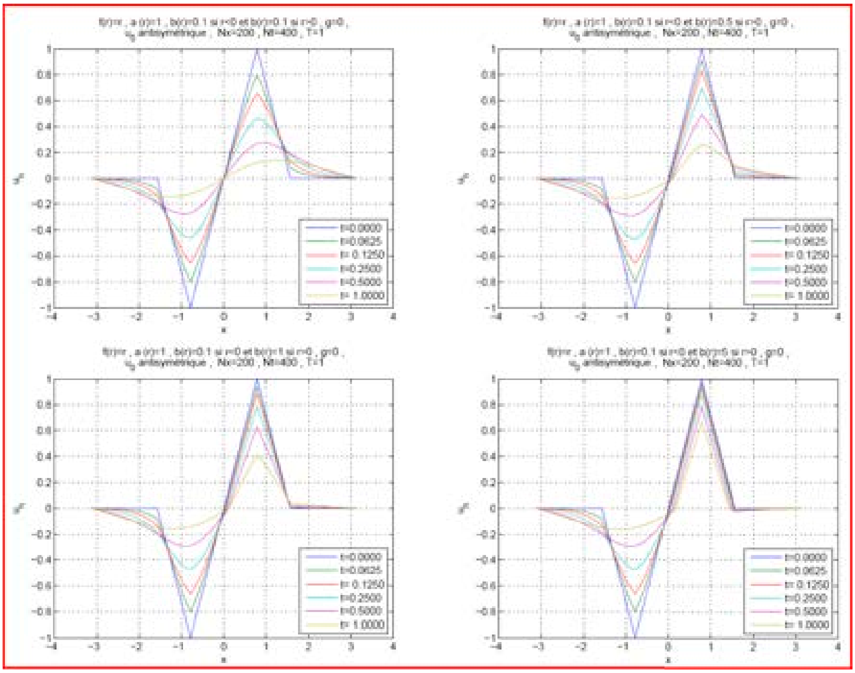


Figure 4: $\partial_t u - \partial^2_{xx} u - \tau \partial^3_{xxt} u = 0$ with $u_0(x) \neq 0$, $b(r) = \tau r^+ - 0.1 r^-$, $\tau = 0, 1/2, 1, 5$.

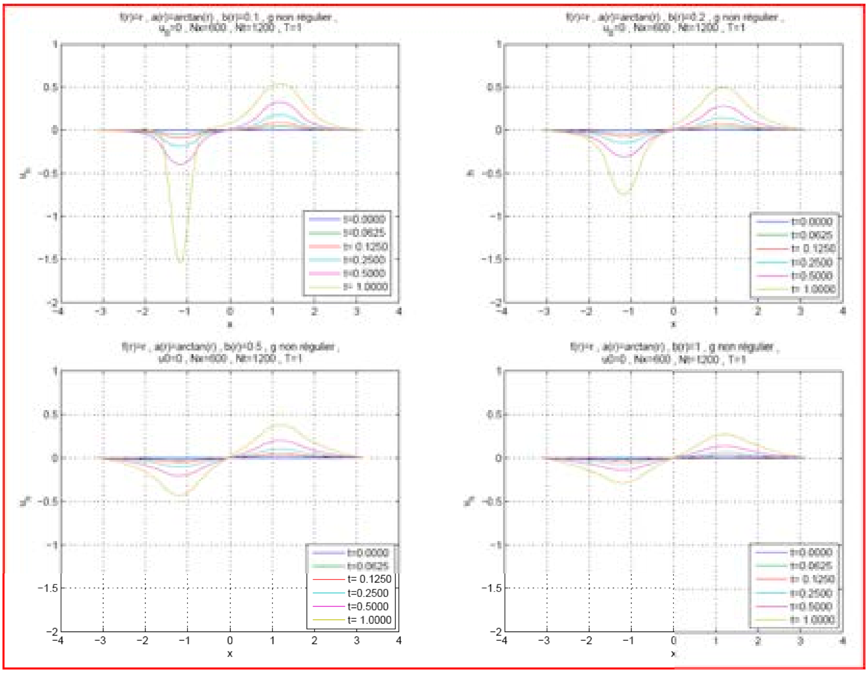


Figure 5: $\partial_t u - \partial_x[\arctan(u)\partial_x u] - \tau\partial^3_{xxt}u = g$ with $u_0 = 0$, $\tau = 0.1, 0.2, 0.5, 1$.

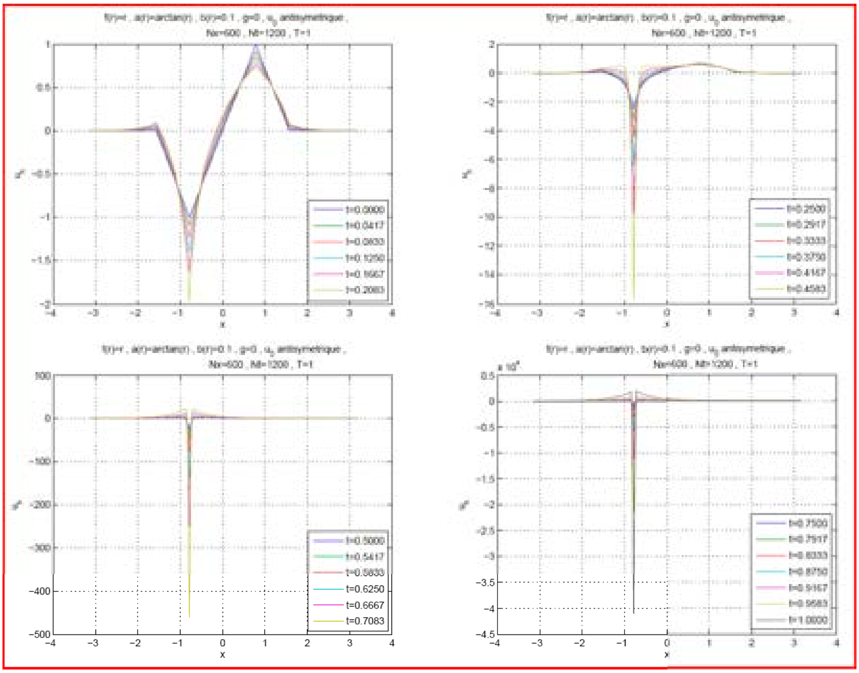


Figure 6: $\partial_t u - \partial_x[\arctan(u)\partial_x u] - \tau\partial^3_{xxt}u = 0$ with $u_0 \neq 0$, $\tau = 0.1$ and small times.

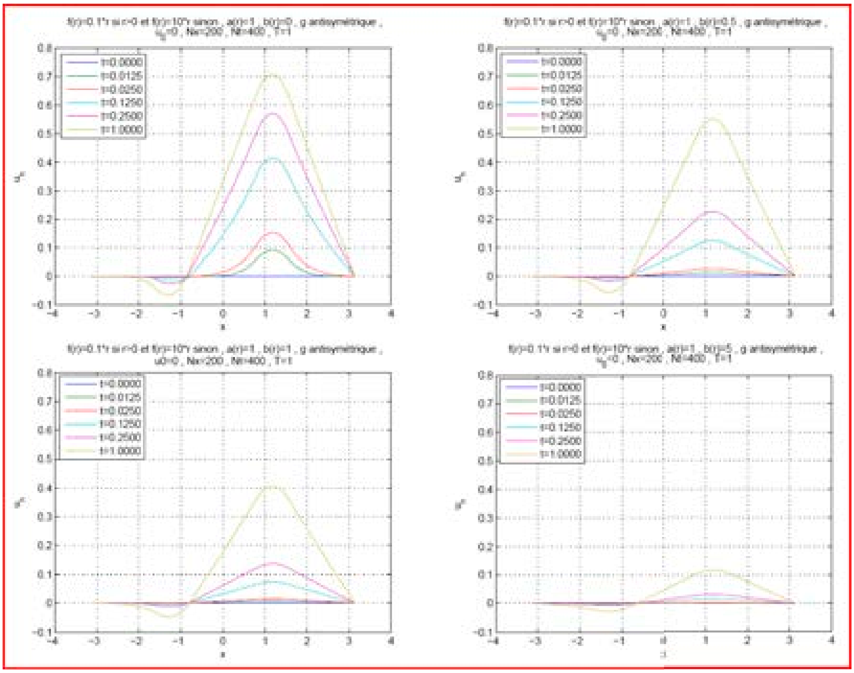


Figure 7: $f(\partial_t u) - \partial^2_{xx} u - \tau \partial^3_{xxt} u = g$ with $u_0(x) = 0$, $\tau = 0, 1/2, 1, 5$.

3.3. Barenblatt's Equation

In Figures 7 to 10, $f(r) = r/10$ if $r > 0$ and $f(r) = 10r$ otherwise, $a(r) = 1$ and $b(r) = \tau$ with different values of $\tau = 0.1, 0.2, 0.5, 1$. The two configurations are illustrated, as well as the asymptotic behaviour.

Note that, in spite of odd data, $x \mapsto u(t, x)$ is not a odd function any more if $t > 0$. Indeed, for negative x, $t \mapsto u(t, x)$ is an increasing function. Thus, the equation is formally $\partial_t u - 10\,\Delta u - 10\epsilon \Delta \partial_t u = 10g$. Else, for positive x, $t \mapsto u(t, x)$ is a decreasing function. Thus, the equation is formally $\partial_t u - \frac{1}{10}\Delta u - \frac{\epsilon}{10}\Delta \partial_t u = \frac{g}{10}$.

References

[1] Adimurthi, Seam, N., and G., V. On the equation of Barenblatt-Sobolev. *Communications in Contemporary Mathematics* (submitted).

[2] Alberty, J., Carstensen, C., and Funken, S. A. Remarks around 50 lines of Matlab: short finite element implementation. *Numerical Algorithms 20*, 2-3 (1999), 117–137.

[3] 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.

[4] Dennis, Jr., J. E., and Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial & Applied Mathematics, 1996.

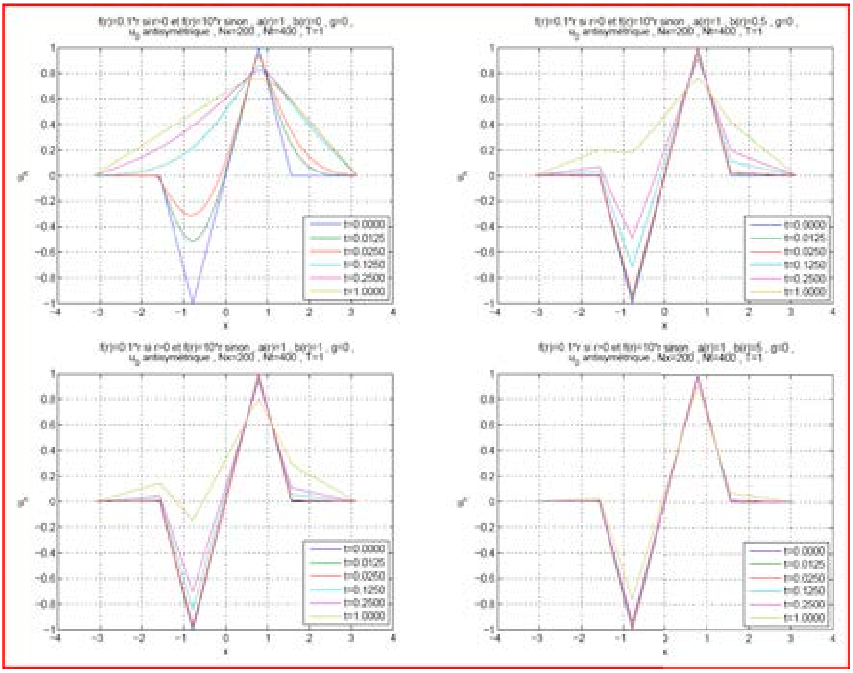


Figure 8: $f(\partial_t u) - \partial^2_{xx} u - \tau \partial^3_{xxt} u = 0$ with $u_0(x) \neq 0$, $\tau = 0, 1/2, 1, 5$.

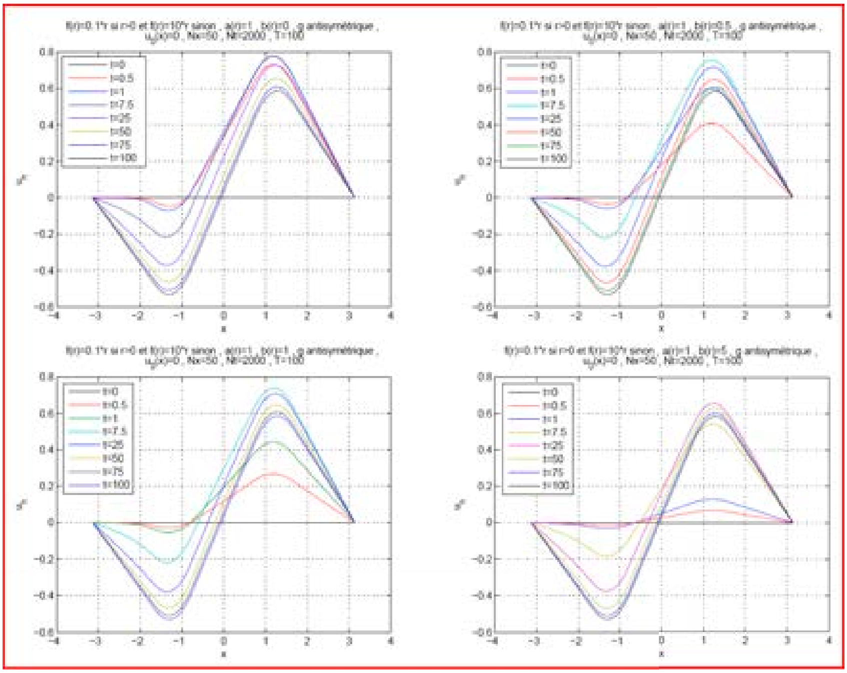


Figure 9: $f(\partial_t u) - \partial^2_{xx} u - \tau \partial^3_{xxt} u = g$ with $u_0(x) = 0$, $\tau = 0, 1/2, 1, 5$, $t \to \infty$.

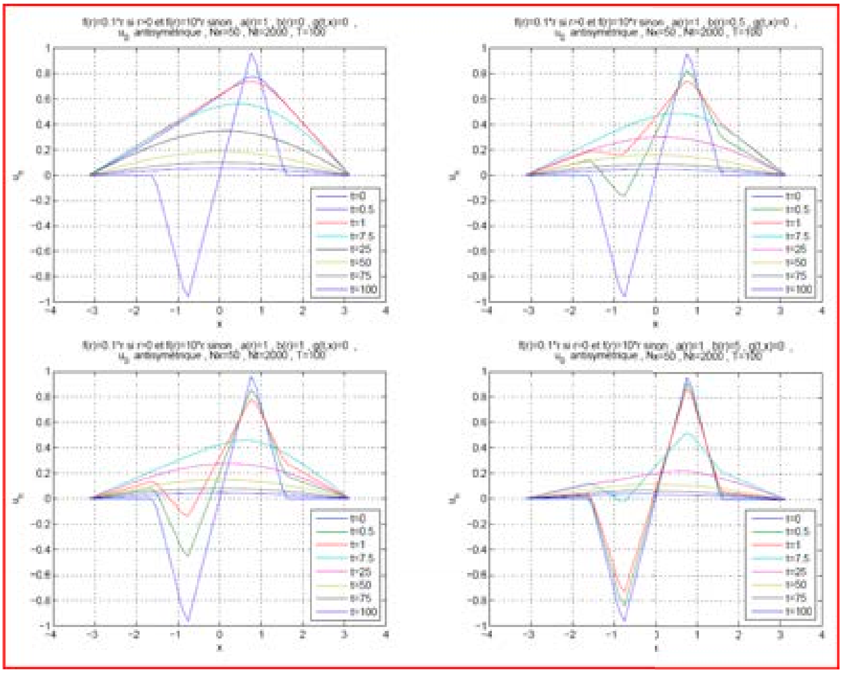


Figure 10: $f(\partial_t u) - \partial^2_{xx} u - \tau \partial^3_{xxt} u = 0$ with $u_0(x) \neq 0$, $\tau = 0, 1/2, 1, 5$, $t \to \infty$.

[5] Quarteroni, A., Sacco, R., and Saleri, F. *Numerical methods. Algorithms, analysis and applications*. Springer, 2007.

[6] Seam, N., and Vallet, G. Study of a nonlinear problem of pseudoparabolic type. In *Proc. of the Fifth Asian Mathematical Conference*, vol. I. Universiti Sains Malaysia, Malaysia, 2009, pp. 9–15.

[7] Seam, N., and Vallet, G. Existence of a solution to a class of pseudoparabolic problems. In *Tenth International Conference Zaragoza-Pau on Applied Mathematics and Statistics*, vol. 35 of *Monogr. Semin. Mat. García Galdeano*. Prensas Univ. Zaragoza, Zaragoza, 2010, pp. 237–244.

[8] Seam, N., and Vallet, G. Existence results for nonlinear pseudoparabolic problems. *Nonlinear Analysis* (submitted).

[9] Sibony, M., and Mardon, J.-C. *Analyse numérique I: Systèmes linéaires et non linéaires*, vol. 1405 of *Actualités Scientifiques et Industrielles*. Hermann, 1984.

LMA, University of Pau - Postal address IPRA BP 1155 Pau Cedex (France)
robert.luce@univ-pau.fr - guy.vallet@univ-pau.fr

Departement of Mathematics, Royal University of Phnom Penh
Russian Federation Boulevard, Toul Kork, Phnom Penh (Cambodia).
seamngonn@yahoo.fr

Monografías Matemáticas *García de Galdeano* **37**, 171–180 (2012)

Uniqueness of strong solutions to doubly nonlinear evolution equations

Jochen Merker

Abstract. In this article uniqueness of strong solutions to the abstract doubly nonlinear evolution equation

$$\frac{\partial Bu}{\partial t} + Au = f$$

is discussed under the main assumptions that B^{-1} is strongly monotone and there is a $C < \infty$ such that $\Phi_A + C\Phi_B$ is convex for the potentials Φ_A resp. Φ_B of A resp. B.

Keywords: Doubly nonlinear evolution equations, strong solutions, uniqueness.

AMS classification: 35K90, 47J35, 34G20, 35A02.

§1. Introduction

The aim of this article is to discuss strong solutions of abstract doubly nonlinear evolution equations

$$\frac{\partial Bu}{\partial t} + Au = f\,, \tag{1}$$

and especially the uniqueness of strong solutions to an initial value. Hereby, $A : X \to X^*$ resp. $B : Y \to Y^*$ are operators on Banach spaces X resp. Y with a dense and separable intersection, and f is an inhomogeneity or nonlinearity.

Uniqueness of weak solutions to initial data with finite energy has been established for the concrete case of a degenerate elliptic-parabolic equation

$$\frac{\partial b(u)}{\partial t} + \operatorname{div}(a(b(u), \nabla u)) = f \tag{2}$$

by [11] via an L^1-contraction principle for $b(u)$. Uniqueness of entropy solutions to L^1-initial data has been shown by [3] (even in presence of transport terms and therefore for degenerate elliptic-parabolic-hyperbolic equations), and uniqueness of renormalized solutions has been proved by [4]. In literature uniqueness is also discussed for several variants of (2) like the anisotropic case ([9]), the so-called triply nonlinear case ([1]) or the case of variable exponents ([2]). All these articles have in common that uniqueness is proved via Kruzhkov's method of doubling the variables.

In this article, an elementary proof of the uniqueness of strong solutions to the abstract problem (1) along the lines of [7, 5, 6] is given, see also [12, Section 8.5 and 11.2.3]. While a discussion of the abstract problem is more general than a discussion of the concrete equation (2) (e.g. parts of B could be fractional derivatives or general convolution operators), it is a major restriction to prove uniqueness only for strong solutions and not for weak, entropy or renormalized solutions, because in general strong solutions may not exist. However, this is the price to pay for applying an elementary method instead of a more sophisticated method like Kruzhkov's doubling of variables.

1.1. Outline

In Section 2 existence of strong solutions to (1) is established for initial values $u_0 \in X \cap Y$ under the main additional assumption that

$$\langle v^*, dB^{-1}(u^*)v^* \rangle \geq c\|v^*\|^2_{H^*} \tag{3}$$

holds for all $u^*, v^* \in Y^*$ with a constant $c > 0$, where $X \cap Y \subset H \subset Y$ is an interpolation triple with a Hilbert space H. This assumption is equivalent to strong monotonicity of B^{-1} as an operator $B^{-1} : Y^* \subset H^* \to H$. Note that there are also other situations which allow to prove the existence of certain types of strong solutions (see [10]), but here we concentrate on this situation.

For the concrete equation (2) existence of strong solutions can be guaranteed for regular initial data and potential $a = d\phi_a$, if b is not only assumed to be nondecreasing, but additionally b^{-1} is assumed to be differentiable with a nonvanishing derivative at 0. Thus, b must not be degenerate or singular at 0, but is still allowed to grow nonlinearly.

Uniqueness of strong solutions is shown in section 3 under the convexity assumption that there is a $C < \infty$ such that $\Phi_A + C\Phi_B$ is convex for the potentials Φ_A resp. Φ_B of A resp. B. Further, continuous dependence of strong solutions on the initial value and on the right hand side is established within this abstract framework. However, before we start our discussion let us mention two examples which illustrate that in general neither u nor Bu need to be unique.

1.2. Examples for non-uniqueness

The following examples illustrate in which way weak solutions of a doubly nonlinear reaction diffusion equation (2) to an initial value may not be unique.

Example 1. Let $A : W^{1,2}(\Omega) \to (W^{1,2}(\Omega))^*$ be the negative of the one-dimensional Laplacian on the interval $\Omega := (0, 1)$ under Neumann-boundary conditions $\partial u/\partial x = 0$ on $\partial\Omega$, and let $B : L^2(\Omega) \to L^2(\Omega)$ be the superposition operator $(Bu)(x) := b(u(x))$ induced by

$$b(u) := \begin{cases} u + 1, & \text{if } u \leq -1, \\ 0, & \text{if } -1 \leq u \leq 1, \\ u - 1, & \text{if } u \geq 1. \end{cases}$$

Obviously, B is a monotone potential operator, which is coercive, bounded and continuous. However, the equation $\partial Bu/\partial t + Au = 0$ does not have a unique solution u to the zero function as initial value of Bu. In fact, if $u(t, x)$ is an arbitrary continuous function independent of x which attains values between -1 and 1, then $Au(t) = 0$ and $Bu(t) = 0$ for every t. Thus, there are many weak solution u to the the initial value 0 of Bu.

Non-uniqueness of u may not be considered as a problem if at least Bu is unique. However, in general it may even happen that Bu is not unique, as the following example shows, where B is multivalued (so that B^{-1} is not strictly monotone), see also [8, Remark 4].

Example 2. Let $\Omega := (0,1)$, let $B : L^2(\Omega) \to L^2(\Omega)$ be the superposition operator induced by the multivalued mapping

$$b(u) := \begin{cases} u-1, & \text{if } u<0, \\ [-1,1], & \text{if } u=0, \\ u+1, & \text{if } u>0, \end{cases}$$

and let $A : W_0^{1,2}(\Omega) \to (W_0^{1,2}(\Omega))^*$ be the operator $\langle Au, w\rangle = \int_\Omega u_x w_x + b(u) w_x \, dx$, i.e. $Au = -u_{xx} - v_x$ on smooth functions with $v(t,x) \in b(u(t,x))$ under Dirichlet conditions $u = 0$ on $\partial\Omega$. Then one solution to the initial value 1 of Bu is given by $u := 0$ and $v := 1 \in Bu$, but for every $C \geq 1$ and every C^1-function h on $[0,\infty)$ with values between 0 and 2 also $u := 0$ and

$$v(t,x) := \begin{cases} 1, & \text{if } 0 \leq t+x \leq C, \\ 1 - h(t+x-C), & \text{if } t+x \geq C, \end{cases}$$

define a solution with $v(0) = 1$ due to $v_t - v_x = 0$ (and $u_{xx} = 0$).

§2. Existence of strong solutions

In this section the existence of strong solutions to the abstract equation (1) is discussed by energy methods for the case that B^{-1} exists and is strongly monotone as an operator on some intermediate Hilbert space. However, first let us formulate standard structural assumptions which allow to prove existence of weak solutions to (1) :

(A1) X and Y are reflexive Banach spaces with a dense and separable intersection $X \cap Y$ [1], which is compactly embedded into Y.

(A2) $B : Y \to Y^*$ is a continuous strictly monotone potential operator, which is coercive and satisfies the growth condition $\|Bu\|_{Y^*} \leq C(1 + \|u\|_Y^{m-1})$ with a constant $C < \infty$ and a parameter $1 < m < \infty$.

(A3) $A : X \to X^*$ is a pseudomonotone operator, which satisfies the semicoercivity condition $\langle Au, u\rangle \geq c_1\|u\|_X^p - c_2\|u\|_X - c_3\|Bu\|_{Y^*}^{m'}$ and has growth $\|Au\|_{X^*} \leq C(\|u\|_Y)(1 + \|u\|_X^{p-1})$ for a parameter $1 < p < \infty$ with constants $c_1 > 0$, c_2, c_3 and an increasing function $C : \mathbb{R}_0^+ \to \mathbb{R}_0^+$.

If $f \in L^{p'}(0,T;X^*) + L^1(0,T;Y^*)$ is an inhomogeneity, then under the assumptions (A1)-(A3) a weak solution u exists to an initial value $u_0 \in Y$ in the sense that $u \in L^p(0,T;X) \cap L^\infty(0,T;Y)$ is such that $Bu \in L^\infty(0,T;Y^*)$ has the initial value $Bu_0 \in Y^*$ and a weak derivative $\partial Bu/\partial t \in L^{p'}(0,T;X^*) + L^1(0,T;Y^*)$ satisfying (1) as an equation in $(X \cap Y)^*$ for a.e. $t \in (0,T)$, or equivalently as an equation in $L^{p'}(0,T;X^*) + L^1(0,T;Y^*)$.

[1] i.e. there are continuous linear embeddings of X and Y into a complete locally convex space Z such that the intersection $X \cap Y$ within Z is dense in X resp. Y w.r.t. the norms $\|\cdot\|_X$ resp. $\|\cdot\|_Y$, and that $X \cap Y$ is separable w.r.t. the norm $\|\cdot\|_X + \|\cdot\|_Y$.

Here we are interested in a slightly different case, where the inhomogeneity f satisfies $f \in L^{p'}(0,T;X^*)+L^2(0,T;H^*)$ for an intermediate Hilbert space H of the inclusion $X \cap Y \subset Y$ given by (A1). More precisley, we require that $X \cap Y \subset H \subset Y$ is an interpolation triple, i.e. there is a $\theta \in [0,1]$ and a constant $C < \infty$ such that $\|u\|_H \leq C\|u\|_X^\theta \|u\|_Y^{1-\theta}$ for every $u \in X \cap Y$. In this case, under the additional assumptions that B satisfies the coercivity condition $\|u\|_Y \leq C(1+\|Bu\|_{Y^*}^{m'})$ with a constant $C < \infty$ and $p \geq 2$ or $1/2 \leq \theta \leq p/2$ hold, there exists a weak solution of (1) in the following sense:

Definition 1. A function $u \in L^p(0,T;X) \cap L^\infty(0,T;Y)$ is called a weak solution of equation (1) to the initial value $u_0 \in Y$, if $Bu \in L^\infty(0,T;Y^*)$ has the initial value $Bu_0 \in Y^*$ and a weak derivative $\partial Bu/\partial t \in L^{p'}(0,T;X^*) + L^2(0,T;H^*)$ satisfying equation (1) as an equation in $(X \cap H)^*$ for a.e. $t \in (0,T)$, or equivalently as an equation in $L^{p'}(0,T;X^*) + L^2(0,T;H^*)$.

The existence of weak solutions in the sense of Definition 1 can even be generalised to the case where $f = f(t,u)$ is a nonlinearity. In fact, if B satisfies the stronger coercivity condition $\|u\|_Y \leq C(1+\|Bu\|_{Y^*}^{m'-1})$ with a constant $C < \infty$ and $f = f(t,u)$ is a nonlinearity which satisfies the growth condition $\|f(t,u)\|_{H*} \leq C(\gamma(t) + \|u\|_Y^{(m-1)(1-\theta)})$ with a constant $C < \infty$ and a function $\gamma \in L^2(0,T)$, then there still exist weak solutions in the sense of Definition 1.

Now we are interested in assumptions, which guarantee that weak solutions even have better properties than those mentioned in Definition 1. The following theorem formulates such assumptions in the special case that $B^{-1} : Y^* \subset H^* \to H$ is strongly monotone, see [10].

Theorem 1. *Additionally to the structural assumptions (A1)-(A3) assume that H is a Hilbert space such that* $X \cap Y \subset H \subset Y$ *is an interpolation triple and* $p \geq 2$ *or* $1/2 \leq \theta \leq p/2$ *hold. Further, assume that*

- $B^{-1} : Y^* \to Y$ *is* C^1*, satisfies* $\|u\|_Y \leq C(1+\|Bu\|_{Y^*}^{m'-1})$ *with a constant* $C < \infty$*, and is strongly monotone in the sense that* $\langle v^*, dB^{-1}(u^*)v^*\rangle \geq c\|v^*\|_{H^*}^2$ *for all* $u^*, v^* \in Y^*$ *with a constant* $c > 0$ [2]*,*
- $A : X \to X^*$ *is a potential operator such that the intersection of Y and the domain* $D(A) := \{u \in X \,|\, Au \in H^*\}$ *of A w.r.t.* H^* *is dense in* $X \cap Y$*,*
- f *is an inhomogeneity in* $L^2(0,T;H^*)$ *or a nonlinearity* $f = f(t,u)$ *such that* $g(t,u) := dB^{-1}(Bu)^* f(t,u)$ *satisfies the growth condition* $\|g(t,u)\|_H \leq C\left(\gamma(t) + \|u\|_Y^{(m-1)(1-\theta)}\right)$ *with a constant* $C < \infty$ *and a function* $\gamma \in L^2(0,T)$*.*

Then there exists to every initial value $u_0 \in X \cap Y$ *a strong solution u of equation* (1) *in the sense that u is a weak solution which additionally satisfies* $u \in L^\infty(0,T;X)$*, and* $Bu \in L^\infty(0,T;Y^*)$ *and a weak derivative* $\partial Bu/\partial t \in L^2(0,T;H^*)$*.*

Let us shortly sketch the proof of this theorem given in [10].

Proof. Use a Faedo-Galerkin method and consider the restrictions

$$\frac{\partial B_k u_k}{\partial t} + A_k u_k = f_k \tag{4}$$

of equation (1) to an increasing sequence of finite-dimensional subspaces $W_k \subset D(A) \cap Y \subset X \cap Y$, where A_k, B_k are the restrictions of A, B to W_k and f_k is a continuous approximation

[2] This condition is equivalent to strong monotonicity of B^{-1} as an operator $B^{-1} : Y^* \subset H^* \to H \subset Y$, i.e. to $\langle u^* - v^*, B^{-1}u^* - B^{-1}v^*\rangle \geq c\|u^* - v^*\|_{H^*}^2$ for arbitrary $u^*, v^* \in Y^*$ with a constant $c > 0$.

of f with values in W_k. Due to (A1)-(A3) short-time existence of solutions u_k of this ODE to initial values $u_{0k} \in W_k$ can be guaranteed. Test (4) by u_k to obtain from the semicoercivity condition on A the a priori estimate

$$\begin{aligned}\hat{\Phi}_B(u_k(t)) + \left(c_1 - \frac{\epsilon^p}{p}\right)\int_0^T \|u_k(s)\|_X^p\, ds \\ \leq \hat{\Phi}_B(u_{0k}) + \frac{1}{p'\epsilon^{p'}}|c_2|^{p'}T + \int_0^t c_3\|B_k u_k(s)\|_{Y^*}^{m'}\, ds + \int_0^t \|f_k(s)\|_{H^*}\|u\|_H\, ds\,,\end{aligned}$$

where $\hat{\Phi}_B(u) = \Phi_B^*(Bu)$ denotes the Legendre transform of the convex potential Φ_B of B in dependence of Bu, $\epsilon > 0$ is sufficiently small and the energy identity $\frac{d}{dt}\hat{\Phi}_B(u) = \langle \frac{\partial Bu}{\partial t}, u\rangle$ was used. As a consequence of the growth condition $\|Bu\|_{Y^*} \leq C(1 + \|u\|_Y^{m-1})$ we have $\|Bu\|_{Y^*}^{m'} \leq C(1 + \hat{\Phi}_B(u))$, as a consequence of the coercivity condition $\|u\|_Y \leq C(1 + \|Bu\|_{Y^*}^{m'})$ we have $\|u\|_Y \leq C(1 + \hat{\Phi}_B(u))$, and the assumptions $p \geq 2$ or $1/2 \leq \theta \leq p/2$ allow to estimate the last term by $C\int_0^t(1 + \hat{\Phi}_B(u))\, ds$ in the case that $f \in L^2(0, T; H^*)$ is an inhomogeneity. In the case that $f = f(t, u)$ is a nonlinearity apply inequality (3) to $u^* := Bu$, $v^* := f(t, u)$, to obtain

$$c\|f(t,u)\|_{H^*}^2 \leq \langle f(t,u), dB^{-1}(Bu)f(t,u)\rangle = \langle f(t,u), g(t,u)\rangle \leq \|f(t,u)\|_{H^*}\|g(t,u)\|_H$$

so that by the assumptions on g the growth condition

$$\|f(t,u)\|_{H^*} \leq \frac{1}{c}\|g(t,u)\|_H \leq \frac{C}{c}\left(\gamma(t) + \|u\|_Y^{(m-1)(1-\theta)}\right)$$

is valid and the last term can again be estimated by $C\int_0^t(1 + \hat{\Phi}_B(u))\, ds$. Thus, Gronwall's lemma allows to deduce uniform bounds w.r.t. k of u_k in $L^\infty(0, T; Y) \cap L^p(0, T; X)$, Bu_k in $L^\infty(0, T; Y^*)$ and Au_k in $L^{p'}(0, T; X^*)$. Due to these bounds a weakly convergent subsequence $u_k \rightharpoonup u$ can be extracted. Finally, time-compactness and pseudomonotonicity allow to conclude that u is a weak solution of (1).

To obtain a strong solution we would like to test the approximate equation (4) by $\partial u_k/\partial t$, but (4) only guarantees the existence of $\partial B_k u_k/\partial t \in C(0, T; W^*)$ and not the existence of $\partial u_k/\partial t$. However, as B^{-1} is assumed to be continuously differentiable, the chain rule implies the existence of

$$\frac{\partial u}{\partial t} = dB^{-1}(Bu)\frac{\partial Bu}{\partial t}\,. \tag{5}$$

Due to $W_k \subset D(A)$ and $f_k \in L^2(0, T; H^*)$ a solution $u_k \in C^1(0, T; W_k)$ of the approximate equation (4) satisfies $\partial B_k u_k/\partial t \in H^*$ for a.e. t. Especially, inequality (3) can be applied to $u^* := Bu_k(t)$, $v^* = \partial Bu_k(t)/\partial t$, to obtain

$$\left\langle \frac{\partial B_k u_k}{\partial t}, \frac{\partial u_k}{\partial t}\right\rangle \geq c\left\|\frac{\partial Bu_k}{\partial t}\right\|_{H^*}^2\,.$$

Further, as $g(t, u) := dB^{-1}(Bu)^* f(t, u)$ satisfies $\|g(t, u)\|_H \leq C(\gamma(t) + \|u\|_Y^{(m-1)(1-\theta)})$ with a constant $C < \infty$ and a function $\gamma \in L^2(0, T)$, and as a uniform bound of u_k in $L^\infty(0, T; Y)$

w.r.t. k has already been established, we can conclude that $g(\cdot, u_k(\cdot))$ is uniformly bounded in $L^2(0,T;H)$. Thus, a test of (4) by $\partial u_k/\partial t$ yields

$$\left(c - \frac{\epsilon^2}{2}\right)\left\|\frac{\partial Bu_k}{\partial t}\right\|_{H^*}^2 + \frac{d}{dt}\Phi_A(u_k) \leq \frac{1}{2\epsilon^2}\|g(\cdot, u_k(\cdot))\|_H^2 \leq C$$

with a constant $C < \infty$ for sufficiently small $\epsilon > 0$. Using this differential inequality uniform a priori estimates w.r.t. k of $\partial Bu_k/\partial t$ in $L^2(0,T;H^*)$ and u_k in $L^\infty(0,T;X)$ can be established. Therefore, additionally we are able to guarantee weak* convergence of a subsequence of the approximate solutions u_k in $L^\infty(0,T;X)$ and weak convergence of $\partial Bu_k/\partial t$ in $L^2(0,T;H^*)$, It is simple to verify that the weak limits of these sequences are identical with their expected values u and $\partial Bu/\partial t$, hence the proof of Theorem 1 is finished. □

As a consequence of Theorem 1 we have $Bu \in W^{1,2}(0,T;H^*) \subset C(0,T;H^*)$ for strong solutions due to $Bu \in L^\infty(0,T;Y^*) \subset L^2(0,T;H^*)$ and $\partial Bu/\partial t \in L^2(0,T;H^*)$. Further, as $\partial Bu/\partial t$ and f lie in $L^2(0,T;H^*)$, also $Au = f - \partial Bu/\partial t$ lies in $L^2(0,T;H^*)$. Therefore, equation (1) holds as an equation in H^* for a.e. $t \in [0,T]$, and thus $u(t) \in D(A)$ for a.e. $t \in [0,T]$. Let us explicitly mention this observation as a corollary.

Corollary 2. *Under the assumptions of Theorem 1 the relation $Au \in L^2(0,T;H^*)$ holds for a strong solution u, and equation* (1) *is valid as an equation in H^* for a.e. $t \in (0,T)$.*

The following example shows how Theorem 1 can be applied to the concrete problem (2).

Example 3. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and let $1 < m < 2$. Consider the space $Y := L^m(\Omega)$ so that $H := L^2(\Omega)$ is continuously embedded into Y. Assume that $\phi_b : \mathbb{R} \to \mathbb{R}$ is a convex function which behaves like $(C_1/2)|u|^2+o(|u|^2)$ as $|u| \to 0$ and like $(C_2/m)|u|^m+\omega(|u|^m)$ as $|u| \to \infty$. Denote by $b := d\phi_b$ the derivative of ϕ_b and by $B : Y \to Y^*$ the corresponding superposition operator. Then $b^{-1}(u)$ behaves like ${C_1}^{-1}u$ as $|u| \to 0$ and like $C_2^{1-m'}|u|^{m'-2}u$ as $|u| \to \infty$, so that $(b^{-1})'(u)$ behaves like ${C_1}^{-1}$ as $|u| \to 0$ and like $(m'-1)C_2^{1-m'}|u|^{((2-m)/(m-1))}$ as $|u| \to \infty$. Especially, pointwisely $(b^{-1})'(u) \geq c$ for a constant $c > 0$ so that

$$c\|v^*\|_2^2 = \int_\Omega c|v^*|^2\,dx \leq \int_\Omega (b^{-1})'(u^*)|v^*|^2\,dx$$

and as a consequence

$$\langle v^*, dB^{-1}(u^*)v^*\rangle \geq c\|v^*\|_2^2$$

for all $u^*, v^* \in Y^*$, i.e. inequality (3) is valid. Note that although b is not degenerate or singular at $u = 0$, the operator B can not be realized as an operator on H as b grows like $C|u|^{m-1}$ as $|u| \to \infty$. Thus (2) is not degenerate or singular at $u = 0$, but still should be considered as an equation for $u \in Y$ and not for $u \in H$.

Finally, assume that a has a p-coercive potential, $1 < p < \infty$, e.g. $a(\nabla u) = |\nabla u|^{p-2}\nabla u$, and consider the corresponding operator $A : W_0^{1,p}(\Omega) \to (W_0^{1,p}(\Omega))^*$, $\langle Au, v\rangle := \int_\Omega a(\nabla u)\cdot\nabla u\,dx$, so that (2) is solved under Dirichlet boundary conditions. For this choice $X := W_0^{1,p}(\Omega)$, and $m < p^*$ has to be required to have a compact embedding $X \cap Y \subset Y$. Now Gagliardo-Nirenberg inequalities

$$\|u\|_{L^2} \leq \|\nabla u\|_{L^{p^*}}^\theta \|u\|_{L^m}^{1-\theta}$$

are valid for $1/2 = \theta/p^* + (1-\theta)/m$, $1/p^* \leq 1/2$, where the parameter θ of the interpolation triple $X \cap Y \subset H \subset Y$ is given by $\theta = ((2-m)p^*)/(2(p^*-m))$. Especially, in the case $p < 2$ the inequality $1/2 \leq \theta \leq p/2$ is valid iff $((2-p)p^*)/(p^*-p) \leq m \leq p^*/(p^*-1)$. For example, if $n = 3$ and p is slightly smaller than 2, then already $m < 6/5$ has to be required.

Further, the right hand side f of (2) and $g := dB^{-1}(Bu)^* f(u)$ are related by

$$f(t,x,u) = \frac{g(t,x,u)}{(b^{-1})'(u)}, \tag{6}$$

where $(b^{-1})'$ is bounded away from zero. Thus, if $g(t,x,u)$ is a pregiven nonlinearity such that $|g(t,x,u)| \leq C\left(\gamma(t,x) + |u|^{(m-1)(1-\theta)}\right)$, then by (6) a corresponding right hand side f can be defined such that the assumptions of Theorem 1 are satisfied. Thus, under the former conditions there exists a strong solution of (2) to initial values $u_0 \in W_0^{1,p}(\Omega) \cap L^m(\Omega)$.

Finally, it can be shown that inhomogeneities

$$f \in L^{2\left(\left(m(m-1)^2(p^*-2)\right)/\left(2(2-m)(p^*-m)\right)\right)'}\left(0,T; L^{\left(2m(m-1)\right)/\left((m+1/2)^2-17/4\right)}(\Omega)\right)$$

can be represented via (6) by a function $g(t,x,u) = f(t,x)(b^{-1})'(u)$ satisfying the growth condition provided that

$$\frac{\sqrt{17}-1}{2} < m \leq 2 \quad \text{and} \quad p^* > \frac{2m(m^2-m-1)}{m^3-2m^2+3m-4}$$

in the case $p < n$.

§3. Uniqueness of strong solutions

Under the assumptions of Theorem 1 equation (1) admits a strong solution to an initial value $u_0 \in X \cap Y$ in the sense that $u \in L^\infty(0,T; X \cap Y)$ is a weak solution such that $Bu \in L^\infty(0,T;Y^*)$ has a weak derivative $\partial Bu/\partial t \in L^2(0,T;H^*)$, and especially $Au \in L^2(0,T;H^*)$. The following theorem guarantees uniqueness of strong solutions and continuous dependence on the initial value and the right hand side.

Theorem 3. *Additionally to the assumptions of Theorem 1 suppose that there is a constant* $C < \infty$ *such that*

$$\langle Au - Av, u - v\rangle + C\langle Bu - Bv, u - v\rangle \geq 0 \, \textit{for all } u, v \in X \cap Y \textit{ and} \tag{7}$$

$$\langle Bu - Bv, dB^{-1}(Bu)Au - dB^{-1}(Bv)Av\rangle + C\langle Bu - Bv, u - v\rangle \geq 0 \ \textit{ for all } u, v \in D(A) \cap Y\,, \tag{8}$$

where $D(A) = \{u \in X \,|\, Au \in H^*\}$ *denotes the domain of* A *w.r.t.* H^*. *Then the following statements are valid:*

- *If* $f = 0$, *then strong solutions of equation* (1) *are unique.*
- *If* $f \in L^1(0,T;Y^*)$ *and* $dB^{-1} : Y^* \subset H \to L(Y^*, H)$ *is Lipschitz continuous, then strong solutions of equation* (1) *are unique and* $Y \ni u_0 \mapsto Bu \in C(0,T;H^*)$ *is continuous.*
- *If* dB^{-1} *and* B^{-1} *are Lipschitz continuous, then* $Y \times L^1(0,T;Y^*) \ni (u_0, f) \mapsto Bu \in C(0,T;H^*)$ *is continuous.*

Remark 1. Note that inequality (7) is equivalent to the convexity of $\Phi_A + C\Phi_B$ on $X \cap Y$, while inequality (8) is equivalent to the convexity of $\Phi_A \circ B^{-1} + C\Phi_B^*$ on $B(D(A) \cap Y)$, where Φ_B^* is the Legendre transform of Φ_B and hence a potential of B^{-1}.

Proof. Assume that u, v are strong solutions of

$$\frac{\partial Bu}{\partial t} + Au = f_1 \quad \text{resp.} \quad \frac{\partial Bv}{\partial t} + Av = f_2 \,.$$

To prove uniqueness, test the difference of these equations by $u - v$ and integrate the resulting equation over $[0, t]$ to obtain

$$\int_0^t \left\langle \frac{\partial}{\partial s}(Bu - Bv), u - v \right\rangle ds + \int_0^t \langle Au - Av, u - v \rangle \, ds = \int_0^t \langle f_1 - f_2, u - v \rangle \, ds \,.$$

Now

$$\left\langle \frac{\partial}{\partial s}(Bu - Bv), u - v \right\rangle = \frac{d}{dt}\langle Bu - Bv, u - v \rangle - \left\langle Bu - Bv, \frac{\partial}{\partial s}(u - v) \right\rangle$$

and thus

$$\begin{aligned}\int_0^t \left\langle \frac{\partial}{\partial s}(Bu - Bv), u - v \right\rangle ds &= (\langle Bu - Bv, u - v \rangle)(t) - (\langle Bu - Bv, u - v \rangle)(0) \\ &\quad - \int_0^t \langle Bu - Bv, dB^{-1}(Bu)(f_1 - Au) - dB^{-1}(Bv)(f_2 - Av) \rangle \, ds\end{aligned}$$

due to $\partial u/\partial t = dB^{-1}(Bu)\partial Bu/\partial t = dB^{-1}(Bu)(f_1 - Au)$ and similar for v. Hence, if $f_1 = 0 = f_2$, then

$$\begin{aligned}&(\langle Bu - Bv, u - v \rangle)(t) \\ &= (\langle Bu - Bv, u - v \rangle)(0) - \int_0^t \langle Au - Av, u - v \rangle \, ds - \int_0^t \langle Bu - Bv, dB^{-1}(u)Au - dB^{-1}(v)Av \rangle \, ds \\ &\leq (\langle Bu - Bv, u - v \rangle)(0) + 2C \int_0^t \langle Bu - Bv, u - v \rangle \, ds\end{aligned}$$

due to the assumptions (7) and (8). By Gronwall's lemma

$$(\langle Bu - Bv, u - v \rangle)(t) \leq (\langle Bu - Bv, u - v \rangle)(0) \exp(2Ct) \,,$$

so that $u(0) = v(0)$ implies $(\langle Bu - Bv, u - v \rangle)(t) = 0$ for a.e. $t \in [0, T]$ and hence $u = v$ by strict monotonicity of B.

If $f_1 = f_2 =: f \in L^1(0, T; Y^*)$, then

$$\begin{aligned}&(\langle Bu - Bv, u - v \rangle)(t) \\ &= (\langle Bu - Bv, u - v \rangle)(0) - \int_0^t \langle Au - Av, u - v \rangle \, ds \\ &\quad - \int_0^t \langle Bu - Bv, dB^{-1}(u)Au - dB^{-1}(v)Av \rangle \, ds + \int_0^t \langle Bu - Bv, (dB^{-1}(Bu) - dB^{-1}(Bv))f \rangle \, ds \\ &\leq (\langle Bu - Bv, u - v \rangle)(0) + 2C \int_0^t \langle Bu - Bv, u - v \rangle \, ds + M \int_0^t \|Bu - Bv\|_{H^*}^2 \|f\|_{Y^*} \, ds\end{aligned}$$

with the Lipschitz constant M of $dB^{-1} : Y^* \subset H^* \to L(Y^*, H)$. By strong monotonicity of $B^{-1} : Y^* \subset H^* \to H$ the inequality $\|Bu - Bv\|_{H^*}^2 \leq c^{-1}\langle Bu - Bv, u - v\rangle$ is valid, hence by Gronwall's lemma

$$(\langle Bu - Bv, u - v\rangle)(t) \leq (\langle Bu - Bv, u - v\rangle)(0) \exp(2CT + \frac{M}{c}\int_0^T \|f\|_{Y^*}\, ds)\,.$$

Especially, again by strong monotonicity of B^{-1}

$$c\|Bu(t) - Bv(t)\|_{H^*}^2 \leq \|Bu(0) - Bv(0)\|_{Y^*}\|u(0) - v(0)\|_Y \exp\Big(2CT + \frac{M}{c}\int_0^T \|f\|_{Y^*}\, ds\Big),$$

so that $Y \ni u(0) \mapsto Bu \in C(0, T; H^*)$ is continuous.

Finally, if $f_1, f_2 \in L^1(0, T; Y^*)$, then the additional terms may be estimated by

$$\int_0^t \langle f_1 - f_2, u - v\rangle\, ds \leq \frac{L}{2}\int_0^t \|f_1 - f_2\|_{Y^*}(1 + \|Bu - Bv\|_{H^*}^2)\, ds\,,$$

with the Lipschitz constant L of $B^{-1} : Y^* \subset H^* \to Y$ and by

$$\begin{aligned}\int_0^t &\langle Bu - Bv, dB^{-1}(Bu)f_1 - dB^{-1}(Bv)f_2\rangle\, ds \\ &\leq M\int_0^t \|Bu - Bv\|_{H^*}^2\|f_1\|_{Y^*}\, ds + \frac{MK}{2}\int_0^t (1 + \|Bu - Bv\|_{H^*}^2)\|f_1 - f_2\|_{Y^*}\, ds\,,\end{aligned}$$

with a bound K of $dB^{-1}(Bv)$ in $C(0, T; L(Y^*, H))$. Thus,

$$\begin{aligned}(\langle Bu - Bv, u - v\rangle)(t) \leq &\left((\langle Bu - Bv, u - v\rangle)(0) + \frac{MK + L}{2}\int_0^T \|f_1 - f_2\|_{Y^*}\, ds\right) \\ &\exp\left(2CT + \frac{M}{c}\int_0^T \|f_1\|_{Y^*}\, ds + \frac{MK + L}{2c}\int_0^T \|f_1 - f_2\|_{Y^*}\, ds\right),\end{aligned}$$

and especially

$$\begin{aligned}c^2\|Bu(t) - Bv(t)\|_{H^*}^2 \leq &\left(\|Bu(0) - Bv(0)\|_{Y^*}\|u(0) - v(0)\|_Y + \frac{MK + L}{2}\int_0^T \|f_1 - f_2\|_{Y^*}\, ds\right) \\ &\exp\left(2CT + \frac{M}{c}\int_0^T \|f_1\|_{Y^*}\, ds + \frac{MK + L}{2c}\int_0^T \|f_1 - f_2\|_{Y^*}\, ds\right),\end{aligned}$$

so that $Y \times L^1(0, T; Y^*) \ni (u(0), f) \mapsto Bu \in C(0, T; H^*)$ is continuous. □

§4. Conclusion

In this article strong solutions to abstract doubly nonlinear evolution equations were discussed under the assumption that B^{-1} is strongly monotone on some intermediate Hilbert space. In this case, strong solutions behave similar as strong solutions to nonlinear evolution equations $\partial u/\partial t + Au = f$. Particularly, under the two convexity conditions (7) and (8) it is possible to give an elementary proof of uniqueness and to obtain continuous dependence on the data. However, for degenerate resp. singular problems where merely weak solutions exist it does not seem possible to avoid more sophisticated methods like Kruzhkov's doubling of variables.

References

[1] ANDREIANOV, B., BENDAHMANE, M., KARLSEN, K. H., , AND OUARO, S. Well-posedness results for triply nonlinear degenerate parabolic equations. *Journal of Differential Equations 247* (2002), 277–302. doi:10.1016/j.jde.2009.03.001.

[2] BENDAHMANE, M., WITTBOLD, P., AND ZIMMERMANN, A. Renormalized solutions for a nonlinear parabolic equation with variable exponent and L^1-data. *Journal of Differential Equations 249* (2010), 1483–1515. doi:10.1016/j.jde.2010.05.011.

[3] CARRILLO, J. Entropy solutions for nonlinear degenerate problems. *Archive for Rational Mechanics and Analysis 147* (1999), 269–361. doi:10.1007/s002050050152.

[4] CARRILLO, J., AND WITTBOLD, P. Uniqueness of renormalized solutions of degenerate elliptic-parabolic problems. *Journal of Differential Equations 156* (1999), 93–121. doi:10.1006/jdeq.1998.3597.

[5] GAJEWSKI, H. On a variant of monotonicity and its application to differential equations. *Nonlinear Analysis: Theory, Methods & Applications 22* (1994), 73–80.

[6] GAJEWSKI, H., AND SKRYPNIK, I. V. On the uniqueness of solutions for nonlinear elliptic-parabolic equations. *Journal of Evolution Equations 3* (2003), 247–281. doi:10.1007/s00028-003-0094-y.

[7] GRÖGER, K., AND NEČAS, J. On a class of nonlinear initial-value problems in hilbert spaces. *Mathematische Nachrichten 93* (1979), 21–31.

[8] IGBIDA, N., AND URBANO, J. M. Uniqueness for nonlinear degenerate problems. *Nonlinear Differential Equations and Applications 10* (2003), 287–307. doi:10.1007/s00030-003-1030-0.

[9] KARLSEN, K. H., AND OHLBERGER, M. A note on the uniqueness of entropy solutions of nonlinear degenerate parabolic equations. *Journal of Mathematical Analysis and Applications 275* (2002), 439–458.

[10] MATAS, A., AND MERKER, J. Strong solutions of doubly nonlinear parabolic equations. *Journal for Analysis and its Applications* ((accepted)).

[11] OTTO, F. L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. *Journal of Differential Equations 131* (1996), 20–38. doi:10.1006/jdeq.1996.0155.

[12] ROUBÍČEK, T. *Nonlinear Partial Differential Equations with Applications*, vol. 153 of *International Series of Numerical Mathematics*. Birkhäuser, Basel, 2005.

Jochen Merker
University of Rostock - Institute of Mathematics
Ulmenstr. 69 (Haus 3)
18057 Rostock
jochen.merker@uni-rostock.de

Monografías Matemáticas *García de Galdeano* **37**, 181–188 (2012)

Legendre Transform of sampled signals by fractal methods

María Antonia Navascués and María Victoria Sebastián

Abstract. The fractal interpolation functions provide an alternative to the classical methods of study of experimental variables. They have been proved useful in many applications, from image compression to signal processing.

The spectral methods (in terms of trigonometric polynomials) are suitable to model periodic or near periodic phenomena. However some experimental variables are far from periodicity. In this paper we present a method to compute Legendre Transform and series expansions for sampled signals by means of fractal methods.

The periodic Fourier case is generalized considering polynomial orthogonal series. Pointwise, uniform and mean-square convergences of the sums are studied and weak sufficient conditions for these types of approximation are found. The procedures ensure a good approach whenever the sampling frequency and the order of the sums are properly chosen.

Keywords: Fractal interpolation functions, orthogonal expansions, Legendre series.

AMS classification: 28A80, 65D05, 41A10, 58C05.

§1. Introduction

We present a method of computing a Legendre expansion for a sampled signal, with the single hypothesis of continuity. The calculus is made via an affine fractal interpolation of the experimental variable. For a suitable election of the scale vector, the pointwise, uniform and mean-square convergences of the expansion obtained are proved. The Legendre Transform provides a formula for the power of the signal, where the hypothesis of periodicity is not needed.

§2. Affine fractal interpolation functions

Let $t_0 < t_1 < \cdots < t_N$ be real numbers, and $I = [t_0, t_N]$ the closed interval that contains them. Let a set of data points $\{(t_n, x_n) \in I \times \mathbb{R} : \; n = 0, 1, 2, \ldots, N\}$ be given. Set $I_n = [t_{n-1}, t_n]$ and let $L_n : I \to I_n, \; n \in \{1, 2, \ldots, N\}$ be contractive homeomorphisms such that:

$$L_n(t_0) = t_{n-1}, \; L_n(t_N) = t_n \tag{1}$$

$$|L_n(c_1) - L_n(c_2)| \leq l \, |c_1 - c_2| \quad \forall \, c_1, c_2 \; \in I \tag{2}$$

for some $0 \leq l < 1$.

Let $-1 < \alpha_n < 1$, for $n = 1, 2, ..., N$, $F = I \times \mathbb{R}$ and N continuous mappings $F_n : F \to \mathbb{R}$ be given satisfying:

$$F_n(t_0, x_0) = x_{n-1}, \quad F_n(t_N, x_N) = x_n, \tag{3}$$

where $n = 1, 2, ..., N$ and

$$|F_n(t,x) - F_n(t,y)| \leq |\alpha_n|\,|x-y| \tag{4}$$

with $t \in I$, $x, y \in \mathbb{R}$.

Now define functions

$$w_n(t,x) = (L_n(t), F_n(t,x))$$

for $n = 1, 2, \ldots, N$.

Theorem 1 (Cf. [1]). *The iterated function system (IFS) $\{F,\ w_n : n = 1, 2, ..., N\}$ defined above admits a unique attractor G. G is the graph of a continuous function $f : I \to \mathbb{R}$ which obeys $f(t_n) = x_n$ for $n = 0, 1, 2, \ldots, N$.*

The previous function is called a fractal interpolation function (FIF) corresponding to $\{(L_n(t), F_n(t,x))\}_{n=1}^{N}$ and it is unique satisfying the functional equation [1]:

$$f(t) = F_n(L_n^{-1}(t), f \circ L_n^{-1}(t)) \tag{5}$$

for $n = 1, 2, ..., N$, $t \in I_n = [t_{n-1}, t_n]$.

The most widely studied fractal interpolation functions so far are defined by the IFS

$$\begin{cases} L_n(t) = a_n t + b_n, \\ F_n(t,x) = \alpha_n x + q_n(t), \end{cases} \tag{6}$$

where

$$a_n = \frac{t_n - t_{n-1}}{t_N - t_0} \quad \text{and} \quad b_n = \frac{t_N t_{n-1} - t_0 t_n}{t_N - t_0}; \tag{7}$$

α_n is called a vertical scaling factor of the transformation w_n and $\bar{\alpha}$ is the scale vector of the IFS, $\bar{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$. In this case, the equation (5) becomes

$$f(t) = \alpha_n\, f \circ L_n^{-1}(t) + q_n \circ L_n^{-1}(t) \tag{8}$$

for $n = 1, 2, \ldots, N$, $t \in I_n = [t_{n-1}, t_n]$.

If $q_n(t)$ is a line, the FIF is termed affine (AFIF). In this case, by Eq. (3), $q_n(t) = q_{n1}t + q_{n0}$, where

$$q_{n1} = \frac{x_n - x_{n-1}}{t_N - t_0} - \alpha_n \frac{x_N - x_0}{t_N - t_0}, \tag{9}$$

$$q_{n0} = \frac{t_N x_{n-1} - t_0 x_n}{t_N - t_0} - \alpha_n \frac{t_N x_0 - t_0 x_N}{t_N - t_0}. \tag{10}$$

These approximants are discussed in the references [4], [5], [6] and [7]. In [4] and [7], several ways of obtaining the scaling factors from the data are presented.

2.1. Rate of approximation

We consider the following notation, for a continuous function g defined on a compact interval I,

$$\|g\|_\infty = \max\{|g(t)| : t \in I\}$$

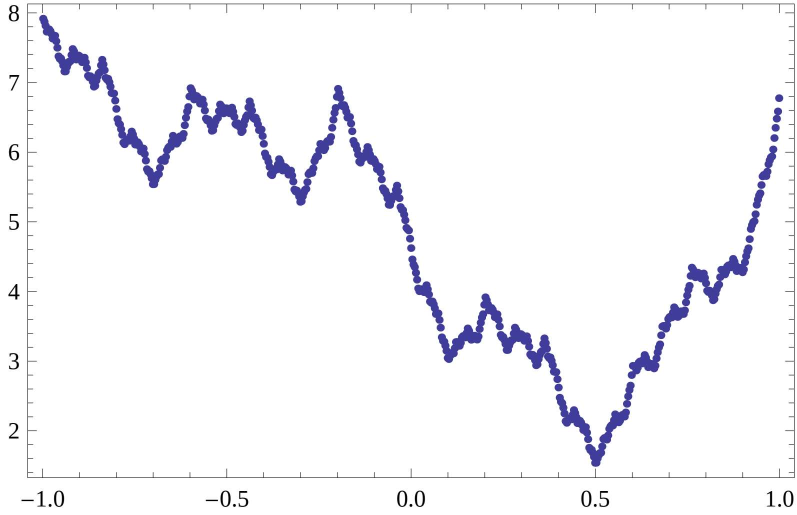


Figure 1: Graph of an affine fractal interpolation function for the set of data points $\{(-1,8),(-3/5,7),(-1/5,7),(1/5,4),(3/5,3),(1,7)\}$ and scale factors $\alpha_n = 0.3$ for $n = 1,2,\ldots,5$

The modulus of continuity of g is defined as

$$\omega_g(\delta) = \sup\{|g(t) - g(t')|\,;\ |t - t'| \leq \delta, t, t' \in I\}$$

By $g \in Lip\,\beta$ (g is Hölder-continuous with exponent β) we mean that there exists $M \geq 0$ such that, for all $t, t' \in I$,

$$|g(t) - g(t')| \leq M|t - t'|^{\beta}.$$

Lemma 2. *$g \in Lip\,\beta$ if and only if $\omega_g(\delta) \leq K\delta^{\beta}$.*

Proof. See [3]. □

Proposition 3. *If x is a continuous function providing the data $\{(t_n, x_n)\}_{n=0}^{N}$ with a constant step $h = t_n - t_{n-1}$, and f is the corresponding AFIF with scale vector $\bar{\alpha}$,*

$$\|x - f\|_\infty \leq\ w_x(h) + \frac{2|\bar{\alpha}|_\infty}{1 - |\bar{\alpha}|_\infty}\,\|x\|_\infty, \tag{11}$$

where $w_x(h)$ is the modulus of continuity of $x(t)$.

Proof. Let g_0 be the polygonal with vertices $\{(t_n, x_n)\}_{n=0}^{N}$. One has

$$\|x - f\|_\infty \leq \|x - g_0\|_\infty + \|g_0 - f\|_\infty.$$

The first term is bounded in Lemma 3.9 of [7] and the second in Proposition 5.1 of [6]. Thus

$$\|x - g_0\|_\infty \leq w_x(h), \tag{12}$$

$$\|g_0 - f\|_\infty \leq \frac{2|\bar{\alpha}|_\infty}{1 - |\bar{\alpha}|_\infty} X_{\max},$$

where $X_{\max} = \max_{0 \leq n \leq N}\{|x_n|\}$ and the result is deduced. □

§3. Legendre Transform

In the article [1], a recurrence formula for the computation of the moments M_m,

$$M_m = \int_I t^m f(t)\, dt \tag{13}$$

was given, for a function f defined by the general iterated function system (6). The formula is expressed as

$$M_m = \frac{1}{(1 - \sum_{n=1}^N a_n^{m+1} \alpha_n)} \left(\sum_{k=0}^{m-1} \binom{m}{k} M_k \sum_{n=1}^{N} a_n^{k+1} \alpha_n b_n^{m-k} + Q_m \right) \tag{14}$$

where

$$Q_m = \int_I t^m Q(t)\, dt \tag{15}$$

and

$$Q(t) = q_n \circ L_n^{-1}(t) \quad \text{if} \quad t \in I_n \tag{16}$$

Without loss of generality, we consider here the interval $I = [-1, 1]$. Let $\{p_n\}_{n=0}^{\infty}$ be the system of normalized polynomials of Legendre. These functions are orthonormal with respect to the inner product

$$(f, g) = \int_I f(t) g(t)\, dt. \tag{17}$$

To compute the Fourier-Legendre coefficients of a FIF f with respect to this complete system, we can proceed in the following way; if the n-th Legendre polynomial p_n is

$$p_n(t) = \sum_{m=0}^{n} d_m t^m,$$

the coefficients of f are

$$c_n = (f, p_n) = \int_I f(t) p_n(t)\, dt = \sum_{m=0}^{n} d_m \int_I t^m f(t)\, dt = \sum_{m=0}^{n} d_m M_m, \tag{18}$$

where M_m are the moments defined in (13). The expansion of f in terms of Legendre polynomials is

$$\sum_{n=0}^{+\infty} c_n p_n$$

and the sequence (c_n) is the Legendre transform of f.

§4. Power of the signal

The scalars c_n enable the construction of the expansion

$$f \sim \sum_{n=0}^{\infty} c_n p_n,$$

which is convergent in quadratic mean to f, that is to say, it is convergent with respect to the $\mathcal{L}^2$-norm:

$$\|f\|_2 = \left(\int_I |f(t)|^2 \, dt \right)^{1/2}$$

To compute the convolution (in a wide sense) of two FIFs, we may use the Parseval's identity:

$$(f, g) = \int_I f(t) g(t) \, dt = \sum_{n=0}^{\infty} c_n^f \, c_n^g, \tag{19}$$

where c_n^f and c_n^g are the Fourier coefficients of f and g respect to Legendre polynomials. The power (or energy) of a signal is given by the Parseval's equality as

$$P = (f, f) = \int_I |f(t)|^2 dt = \sum_{n=0}^{+\infty} |c_n|^2,$$

where c_n are the coefficients of f.

Proposition 4. *The error in the computation of the square root of the power is bounded by the expression*

$$\left| P_x^{1/2} - P_f^{1/2} \right| \leq \left(w_x(h) + \frac{2|\bar{\alpha}|_\infty}{1 - |\bar{\alpha}|_\infty} \|x\|_\infty \right) \left(\text{length}(I) \right)^{1/2},$$

where P_x is the power of the original continuous function $x(t)$, P_f is the power computed by means of an AFIF f with scale vector $\overline{\alpha}$, and length$(I) = (b - a)$ *if $I = [a, b]$.*

Proof. The error in the square root of the power is given by

$$\left| P_x^{1/2} - P_f^{1/2} \right| = \left| \|x\|_2 - \|f\|_2 \right| \leq \|x - f\|_2,$$

where $\|g\|_2 = \left(\int_I |g(t)|^2 \, dt \right)^{1/2}$. Moreover,

$$\|x - f\|_2 = \left(\int_I |x(t) - f(t)|^2 \, dt \right)^{1/2} \leq \|x - f\|_\infty \left(\text{length}(I) \right)^{1/2}. \tag{20}$$

Proposition 3 provides then the estimation of the statement. □

§5. Convergence of the Legendre expansion

The next result proves the validity of using AFIFs to construct Legendre series expansions of a real sampled signal, according to the procedure described in the Section 3.

Theorem 5. *Let $x \in C(I)$ be the original function providing the data. If we choose a fractal f with scale vector $\bar{\alpha}_h$ tending to zero as $h \to 0$, then the Legendre expansion defined by means of f converges in quadratic mean to x as $m \to \infty$ and $h \to 0$.*

Proof. Let $S_m f$ be the m-th partial sum of the Legendre series of f. Let us consider

$$\|x - S_m f\|_2 \leq \|x - f\|_2 + \|f - S_m f\|_2. \tag{21}$$

By (20),

$$\|x - S_m f\|_2 \leq \|x - f\|_\infty (\text{length}(I))^{1/2} + \|f - S_m f\|_2$$

and, by (11),

$$\|x - S_m f\|_2 \leq (\text{length}(I))^{1/2} \left(w_x(h) + \frac{2|\bar{\alpha}_h|_\infty}{1 - |\bar{\alpha}_h|_\infty} \|x\|_\infty \right) + \|f - S_m f\|_2.$$

The uniform continuity of x on I implies that $\lim \omega_x(h) = 0$ as h tends to zero ([3]).

The second adding of (21) goes to zero as m tends to infinity due to the convergence in quadratic mean of the Legendre series of f. □

Remark 1. The former theorem ensures the goodness of the procedure to obtain the power whenever the step and the expansion order are suitably chosen.

In the following we study the pointwise and uniform convergence of the Legendre series. We need two previous lemmas.

Lemma 6. *Let f be a FIF defined by (6) with equally spaced t_n and q_n arbitrary satisfying $q_n(t) \in Lip\, \delta_n$, $0 < \delta_n \leq 1$. Let $\delta = \min\{\delta_n : n = 1, 2, \ldots, N\}$. Then, if $|\bar{\alpha}|_\infty < h^\delta$, $f(t) \in Lip\, \delta$.*

Proof. ([2]) □

Lemma 7. *If $f \in C^p[-1, 1]$ is such that $f^{(p)} \in Lip\, \delta$, then the m-th Legendre sum of f satisfies the inequality*

$$\|f - \sum_{n=0}^{m} c_n p_n\|_\infty \leq \frac{K \ln m}{m^{p+\delta-1/2}} \tag{22}$$

for $p + \delta \geq 1/2$.

Proof. ([9]) □

Theorem 8. *The Legendre expansion of any affine fractal interpolation function f converges pointwisely to f almost everywhere. If the scale vector of f is such that $|\bar{\alpha}|_\infty < h$ then the Legendre expansion of f converges pointwise and uniformly to f on the interval $I = [-1, 1]$.*

Proof. In the reference [8], the author proves that the Legendre series of any function $f \in \mathcal{L}^p(I)$ such that $p > 4/3$ converges pointwisely to f almost everywhere. This fact assures the pointwise convergence for any AFIF a.e. (due to its continuity on I).

The mappings q_n defined in the Section 2 are linear and, consequently, $q_n \in Lip\,1$. If $|\bar{\alpha}|_\infty < h$ according to the Lemma 6, $f(t) \in Lip\,1$. Now, we apply the Lemma 7 for $p = 0$ and $\delta = 1$ obtaining

$$\left\| f - \sum_{n=0}^{m} c_n p_n \right\|_\infty \leq \frac{K \ln m}{m^{1/2}}. \tag{23}$$

As m tends to infinity the Legendre sum tends to f and the uniform convergence is satisfied on the interval $I = [-1, 1]$. □

Remark 2. This result is true for any step h.

Theorem 9. *Let $x(t) \in C(I)$ be the original function providing the data. If we choose $|\bar{\alpha}|_\infty < h$, then the Legendre expansion defined by means of an AFIF converges uniformly to x as $m \to \infty$ and $h \to 0$.*

Proof. The uniform continuity of $x(t)$ on I implies that $\lim \omega_x(h) = 0$ as h tends to zero ([3]). Let $S_m f$ be the m-th partial sum of the Legendre series of f. Let us consider

$$\|x - S_m f\|_\infty \leq \|x - f\|_\infty + \|f - S_m f\|_\infty.$$

The first term goes to zero if $h \to 0$ due to Proposition 3. The second term goes to zero as well when $m \to \infty$ according to the previous theorem,

$$\lim_{m\to\infty} \|f - S_m f\|_\infty = 0$$

and the result is obtained. □

Remark 3. The former theorem ensures the goodness of the procedure to represent and evaluate the signal whenever the step and the expansion order are suitably chosen.

References

[1] Barnsley, M. F. Fractal functions and interpolation. *Constr. Approx. 2*, 4 (1986), 303–329.

[2] Chen, G. The smoothness and dimension of fractal interpolation functions. *Appl. Math-JCU 11* (1996), 409–418.

[3] Cheney, E. *Approximation Theory*. AMS Chelsea Publ., 1966.

[4] 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.

[5] Navascués, M. A., and Sebastián, M. V. Error bounds for affine fractal interpolation functions. *Mathematical Inequalities Applications 9*, 2 (2006), 273–288.

[6] Navascués, M. A., and Sebastián, M. V. Spectral and affine fractal methods in signal processing. *Int. Math. Forum 1*, 29 (2006), 1405–1422.

[7] Navascués, M. A., and Sebastián, M. V. Construction of affine fractal functions close to classical interpolants. *J. of Comp. Anal. Appl. 9*, 3 (2007), 271–285.

[8] Pollard, H. The convergence almost everywhere of legendre series. *Proc. Am. Math. Soc. 35*, 2 (1972), 442–444.

[9] Suetin, P. Representation of continuous and differentiable functions by fourier series of legendre polynomials. *Soviet Math. Dokl. 5* (1964), 1408–1410.

María Antonia Navascués
Departamento de Matemática Aplicada
Universidad de Zaragoza
Campus Río Ebro
50018 Zaragoza, Spain
manavas@unizar.es

María Victoria Sebastián
Centro Universitario de la Defensa
Academia General Militar
Ctra. de Huesca s/n
50090 Zaragoza, Spain
msebasti@unizar.es

Monografías Matemáticas *García de Galdeano* **37**, 189–196 (2012)

Symmetry breaking bifurcations in a D_4 symmetric Hamiltonian system

Sławomir Piasecki, Roberto Barrio and Fernando Blesa

Abstract. In this work we investigate a numerical method to locate periodic orbits in Hamiltonian systems of two degrees of freedom in a D_4 and time reversal symmetric Hamiltonian. The procedure to obtain the "skeleton" of periodic orbits is a combination of several methods such as continuation theory, systematic search algorithm, Poincaré surface of section and a fast chaos indicator, OFLI2. Those techniques are used to provide a complete study of symmetry breaking bifurcations in a particular Hamiltonian system. Moreover, we show in detail the evolution of some families of periodic orbits and an analysis of new bifurcations.

Keywords: skeleton of periodic orbits, bifurcations, Poincaré surfaces of section, OFLI2.

AMS classification: 37G15, 37G25.

§1. Introduction

Periodic orbits (PO) and their stabilities are powerful tool in understanding of dynamical systems. The studies of changes in the behavior of PO a can provide essential insights into nature of simple integrable dynamics and complicated, chaotic dynamics. These knowledge have considerably broad applications in physics (quantum eigen state studies [11]) and in astrophysics (numerous problems of stellar and celestial dynamics, e.g., satellite orbits stabilities, etc. Cf. [13]).

Bifurcation is nothing more than qualitative changes in the system's asymptotic behavior and the points where those changes appear are called Bifurcation Points (BP). Whereas, a bifurcation of PO is when those changes affects on the stability of a equilibria or a PO. For better understanding of the bifurcation we have to concentrate on the study of Periodic Orbits. For instance at the period-doubling bifurcation a PO of period T jumps from stable to unstable branch and simultaneously a new stable PO of period $2T$ is created.

A symmetry breaking bifurcation, appears when some perturbation with less symmetry is added to symmetric system. In this note we consider the quartic homogeneous potential system having a general form

$$\mathcal{H} = \frac{1}{2}(X^2 + Y^2) + \frac{1}{4}(x^4 + y^4) + \alpha x^2 y^2 + \beta(x^2 + y^2), \tag{1}$$

in terms of the Cartesian coordinates x, y and their conjugate momenta X, Y, $\alpha, \beta \in \mathcal{R}$. This system is characterized by discrete symmetries and it is invariant under a rotation by $\pi/4$ (Fig. 1). This system was studied e.g. in [10] to find soliton solution in three space dimensions and also in [7], where a direct method to identify integrable N-degree of freedom Hamiltonian systems was described. The existence of large regions of chaotic orbits in parameter space in the neighborhood of the degenerate bifurcation point was reported in [1].

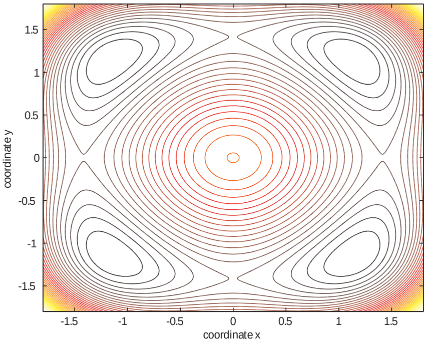


Figure 1: Contour plot of the potential

We choose this Hamiltonian system to create skeleton of periodic orbits and investigate connection between them. Therefore, we set $\alpha = 1/4, \beta = -1$, hence

$$\mathcal{H} = \frac{1}{2}(X^2 + Y^2) + \frac{1}{4}(x^2 + y^2)^2 - x^2 - y^2 - \frac{1}{4}x^2y^2. \tag{2}$$

The dynamics of the Takens-Bogdanov bifurcation with D$_4$ symmetry was studied by Rucklidge [15], and he founded that a symmetry-breaking, period-doubling bifurcation and chaotic sets with five symmetry types allows a quantitative description of the bifurcation sequence were stability is assigned from one subspace to the another.

This Hamiltonian system (2) can be explored for the largest number of orbits. For instance, at Poincaré surface of section our computations include up to $70 \times 70 = 4900$ orbits, however with chaos indicator we compute even with $700 \times 700 = 562500$ orbits for different energy E.

§2. Numerical techniques

Our goal was to find families of periodic orbits and to create the skeleton of periodic orbits in the Hamiltonian system with D_4, and time-reversal symmetries. For that we use set of numerical techniques that are introduced in this section.

First tool is based on continuation theory implemented in the software AUTO created by [9], that handles continuation and bifurcation problems in ordinary, differential equation [14]. Not only it prevents the continuation of the solution curve irrespective of the direction of this curve, but also it allows to detect and follow vertical solution branches. A disadvantage of this technique is that initial computation requires a well defined periodic orbit, without it we are not able to obtain the complete family of periodic orbits nor bifurcations points on it. For further studies two families were chosen (Fig. 5).

To define initial condition systematic search algorithms were used [5]. This technique was developed based on the Brent's method and the Taylor series method that permits to compute the orbits using extended precision. This technique contains several steps, starting with computing of the Poincaré map. The manifold was chosen to be transverse to all orbits, therefore we choose $y = 0, \dot{x} = 0$ and $\dot{y}$ was obtained from Hamiltonian constant. Considering an orbit which starts at position perpendicular to the x-axis

$$(x(0), y(0), \dot{x}(0), \dot{y}(0)) = (x_0, 0, 0, \dot{y}(0)), \tag{3}$$

and crosses the x-axis again perpendicularly, then the orbit is closed and symmetric. We define a new cross at the half period T of the orbit which is perpendicular to the x-axis. Next step in the method is giving a mesh in the parameter and variable space ($x - \mathcal{H}$ plane). Complete set of initial conditions is specified by a value of x and $\mathcal{H}$. By integrating numerically each set of initial conditions we obtain Poincaré map for a given multiplicity (for more details see, [5]).

Next technique that was used in this work is chaos indicator OFLI2, that is an interesting alternative to the standard Poincaré sections, to distinguish among periodic, regular and chaotic orbits [4]. With the second order variational equations, numerical ODE integrator and a specially developed Taylor method [3] gives a fast and accurate numerical integration. The OFLI2 is looking for a set of initial conditions where we may expect strong dependence on initial conditions. The OFLI2 indicator at the final time t_f is given by

$$\text{OFLI2} := \sup_{0<t<t_f} \log \left\| \left\{ \delta \boldsymbol{y}(t) + \frac{1}{2}\, \delta^2 \boldsymbol{y}(t) \right\}^{\perp} \right\|, \tag{4}$$

where $\delta \boldsymbol{y}(t)$ and $\delta^2 \boldsymbol{y}(t)$ are the first and second order sensitivities with respect to carefully chosen initial vectors and $\boldsymbol{y}^{\perp}$ stands for the component of $\boldsymbol{y}$ orthogonal to the flow [4]. The above description gives us the value of the OFLI2 for a particular orbit for a given set of initial conditions. The OFLI2 picture is describing the global dynamical properties of the system when Poincaré section does only for local multiplicity.

In Fig. 3 we compare the evolution of the OFLI2 for the system with energy $E = 2.0$ and $E = 2.5$ on the surface $y = 0$, with Poincaré section. Note that OFLI2 gives much more information than the Poincaré section and locates the periodic orbits and the chain of regular islands inside the chaotic area (see magnification), where the Poincaré maps instead gives a cloud of points.

§3. Bifurcation

In this section we present a study of the bifurcation points of the dynamical system using the Monodromy Method ([2, 8]). 4×4 matrix (M) provides full information about periodic trajectories and can be represented as a first order variational equation

$$\dot{M} = K \cdot \text{Hess}(\mathcal{H}(\boldsymbol{q}, \boldsymbol{p})) \cdot M. \tag{5}$$

For $T = 0$ we can simplify (5) to $M(0) = I_4$, which is four dimensional identity matrix, K is canonical sympletic matrix and $\text{Hess}(\mathcal{H}(\boldsymbol{u}))$ is the Hessian matrix of $\mathcal{H}$ with respect to $\boldsymbol{u}$. Characteristic multipliers of the fixed point (eigenvalues of M), can be use to study linear stability of the system. From now on, the multipliers will be denoted by λ_i ($i = 1 \dots 4$) and are in reciprocal pairs

$$\lambda_1 \lambda_2 = 1, \quad \lambda_3 \lambda_4 = 1. \tag{6}$$

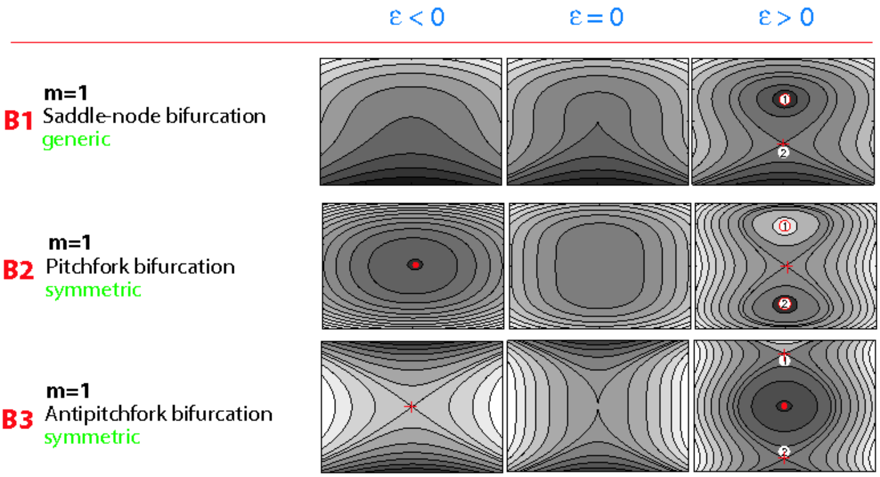


Figure 2: Typical bifurcations for local multiplicity $m = 1$.

That is possible, because system is Hamiltonian and the monodromy matrix M is a real symplectic matrix. Also, complex eigenvalues are in conjugated pairs. In the work we are using definition of stability index introduced by [12] in a form

$$\kappa := \kappa(M(T)) = \text{Tr } (M(T)) - 2, \tag{7}$$

were three cases can be distinguished:

- $|\kappa| < 0$, periodic orbit is stable,
- $|\kappa| > 0$, periodic orbit is unstable (λ_3, λ_4 are real),
- $|\kappa| = 2$, appear special point where stability may change.

The bifurcation point among the family of periodic orbits appears when $\kappa = \lambda_3 + \lambda_4 = 2\ \text{Re}(\lambda_{3,4}) = 2$.

The most typical bifurcation called *saddle-node* bifurcation (Fig. 2) is an example of creating new families of periodic orbits, (apart from the boundaries of the domain of definition of the Poincaré map). This special point is a place where two branches (stable and unstable) met and annihilate (or create).

Since our system have symmetries two more types of the bifurcation points can be detected *pitchfork* and *antipitchfork* (Fig. 2). The former appears when stable family changes to unstable branch and in the same point two new stable branches are created. For antipitchfork is opposite, basic family is unstable and jumps into stable branch and two unstable families are created. In all the cases of the bifurcated families a symmetry lost compare to main family.

A 4-islands chain of isochronous bifurcation was also detected in the system. In this case, the main family after bifurcation point remains in the stable branch and four new stable families are created (see [6]).

3.1. Bifurcation on the system

The focus of this study is pitchfork and 4-islands chain of isochronous bifurcation points. We compare two maps with different energy value $E_1 = 2.5$ (before BP, left) and $E_2 = 2.0$ (after BP, right); Fig. 3. From the maps we know that we start with a stable family and after bifurcation point the main family jumps into unstable branch and two new stable families are created, this can be seen on top of figure, where OFLI2 results are plotted. Those families are also in the skeleton of periodic orbits obtained from AUTO (fig. 4a). If we compare projected orbits from the main family (Fig. 4b) with orbits from the new families we can see that orbits projected into the xy plane, lose one symmetry with respect to y-axis.

The 4-islands chain of isochronous bifurcation is special bifurcation that appears in the symmetric systems Fig. 4c. We project five different periodic orbits from each family. One orbit represents orbit at BP and we see that it is symmetric with respect to x-axis and y-axis, and other orbits loose one of the symmetry. Plot on Fig. 4c contains two bifurcated families, each one consists of two branches. Notice that a orbits from opposite branches have the same shape and are shifted by 180° relative to each other.

§4. Connection symmetric and asymmetric families of periodic orbits

To study the evolution of a periodic orbits along a family we choose two different families (symmetric and asymmetric). In the symmetric family (Fig. 5), we start the evolution from a point close to extreme ($x = 0$, $E = 0$) and moving clockwise. The family that starts with highly eccentricity decreases until reaching the highest energy where eccentricity is the lowest. This family was found to have only one perpendicular intersection with y-axis, so the evolution runs symmetrically. Moreover, along the family we can see that stability changes several times, at those points we have bifurcation points (Fig. 5c). The plot presenting how the orbits change along the families are in the figures (5a, 5d).

From the main symmetric family we choose two bifurcation points and we found two new asymmetric families of periodic orbits (Fig. 5e). Those branches finished at the extreme ($x = 0$, $E = 0$) and are symmetric with respect to x-axis and y-axis. The study of the evolution of this family we start from BP and we decrease value of parameter y. The orbit begins with symmetry with respect to both axis, but the farther we are from bifurcation point, the more significant asymmetry is (Fig. 5f).

In conclusion, we have shown a procedure to obtain skeleton of PO. We have started with creation of a initial conditions using the systematic search with fixed multiplicity. Then those results were used to create the skeleton, which consists of symmetric and asymmetric families of PO. We found also some special bifurcations (the 4-islands chain of isochronous bifurcation) and shown in details the evolution of some families of PO.

Acknowledgements

The authors F.B. and S.P. acknowledge support from the Spanish Research project AYA2008-05572 and R.B. from the Spanish Research project MTM2009-10767.

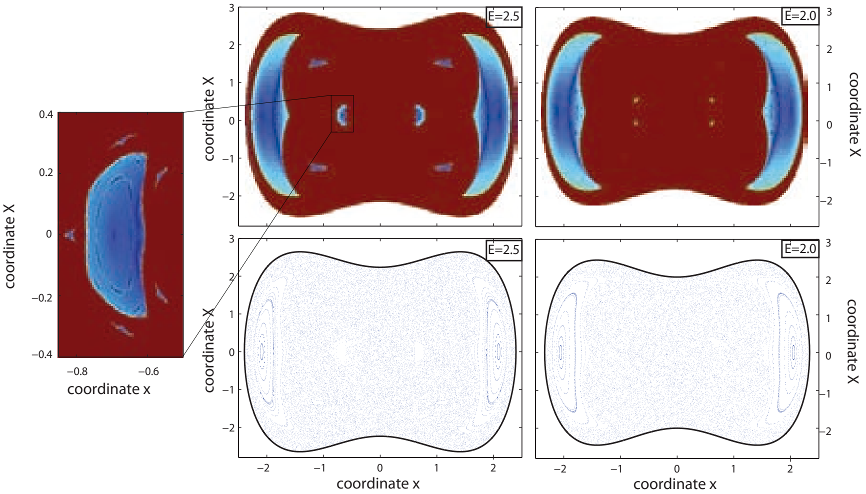


Figure 3: OFLI2 (top) and Poincaré surface of section (bottom), before (left) and after (right) bifurcation point (pitchfork bifurcation), projected on xX plane.

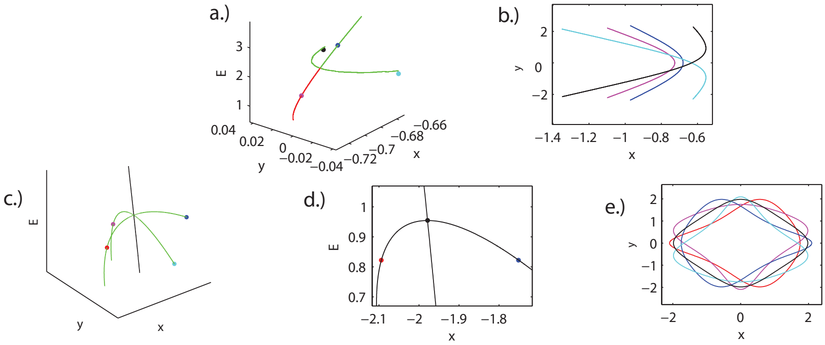


Figure 4: Skeleton of periodic orbits close to pitchfork bifurcation ((a) green and red correspond respectively to stable and unstable family), next to it there are orbits projected on plane xy (b). Outline of 4-islands chain of isochronous bifurcation (c), dots represents orbits projected on xy plane (e). In the middle (d) we got skeleton obtained from AUTO.

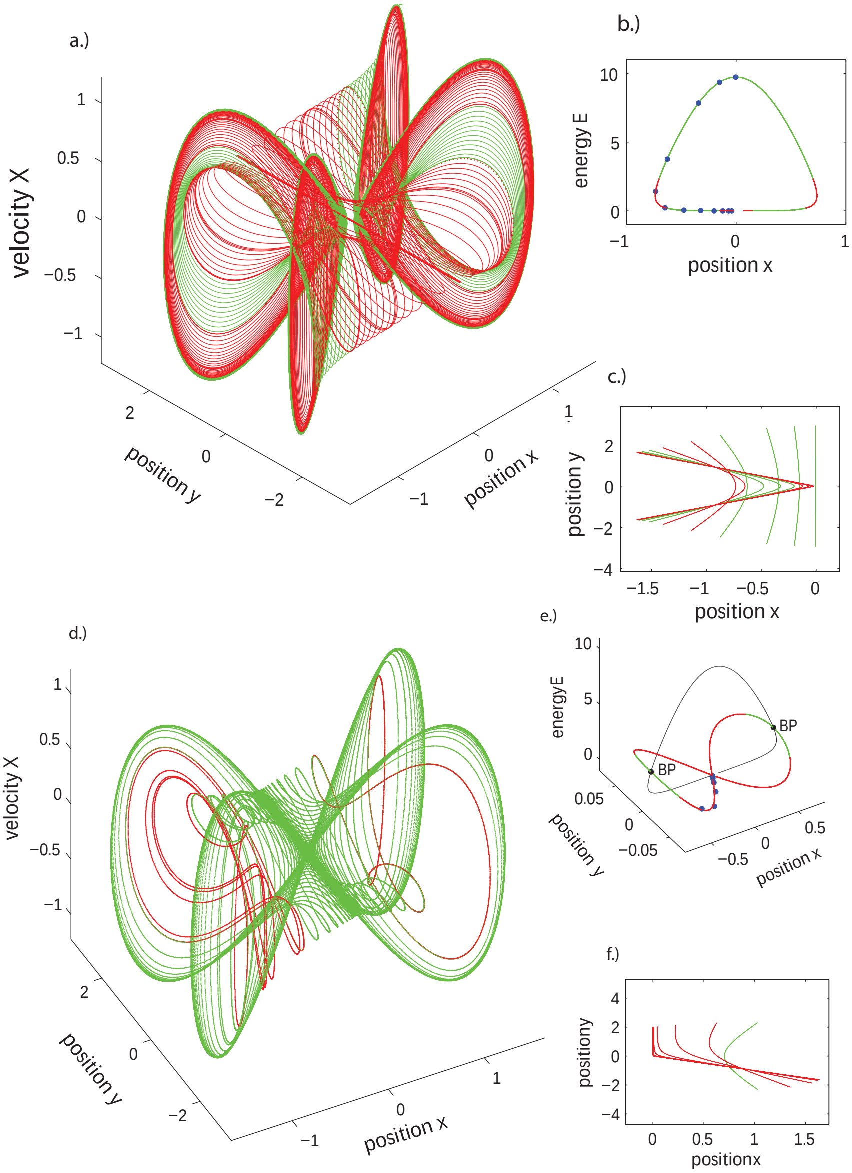


Figure 5: Evolution of symmetric (*a,b,c*) and asymmetric (*d,e,f*) periodic orbits. Main graphs shows (*a,d*), evolution of periodic orbits along the family in three dimension. The main family is plotted on figures *b* and *e* (black). Blue dots represent chosen orbits projected on the plane *xy* (*c, f*). Colors on plots corresponds to stability of orbits, red unstable and green stable.

References

[1] Armbruster, D., and Guckenheimer, J. Chaotic dynamics in systems with square symmetry. *Phys Let A 140* (1989), 416–420.

[2] Baranger, M. Davies, K., and Mahoney, J. The calculation for computing periodics orbits. *Ann Phys 186(1)* (1988), 95–110.

[3] Barrio, R. Blesa, F., and Lara, M. VSVO formulation of the taylor method for the numerical solution of ODEs. *Comput Math Appl 50* (2005), 93–111.

[4] Barrio, R. Sensitivity tools vs. Poincaré sections. *Chaos Soliton Fract. 25* (2005), 711–726.

[5] Barrio, R., and Blesa, F. Systematic search of symmetric periodic orbits in 2DOF Hamiltonian systems. *Chaos, Solit and Fract 41* (2009), 560–582.

[6] Barrio, R., Blesa, F., and Piasecki, S. Connecting symmetric ans asymmetric families of periodic orbits in squared symmetric Hamiltonians. Preprint, 2011.

[7] Bountis, T. Segur, H., and Vivaldi, F. Integrable Hamiltonian system and the Painlevé property. *Phys. Rev. A 13* (1982), 1257–1264.

[8] Davies, KTR. Huston, T., and Baranger, M. Calculation of periodic trajectories for the Hénon Helies Hamiltonian using the monodromy method. *Chaos 2(2)* (1992), 215–224.

[9] Doedel, E. J., Paffenroth, R. C., Champneys, A., and et al. *AUTO 07P-Continuation and bifurcation software for ordinary differential equations.*, 2007. Available from: `http://cmvl.cs.concordia.ca/auto/`.

[10] Friedberg, R. Lee, T., and Sirlin, A. Class of scalar-field soliton solutions in three space dimensions. *Phys. Rev. D 13*, 10 (1976), 2739–2761.

[11] Heller, E. J. Bound-State Eigenfunctions of Classically Chaotic Hamiltonian Systems: Scars of Periodic Orbits. *Phys. Rev. Lett. 53*, 16 (1984), 1515–1518. `doi:10.1103/PhysRevLett.53.1515`.

[12] Henon, M. Numerical exploration of the restricted problem. *A&A 1* (1969), 223,238.

[13] Markellos, V. V. Numerical investigation of the planar restricted three-bodyproblem. *Celestial Mechanics and Dynamical Astronomy 10* (1974), 87–134.

[14] Muñoz-Almaraz, F. J., Freire, E., Galán, J., Doedel, E., and Vanderbauwhede, A. Continuation of periodic orbits in conservative and Hamiltonian systems. *Physica D: Nonlinear Phenomena 181*, 1-2 (2003), 1–38.

[15] Rucklidge, A. M. Global bifurcations in the Takens-Bogdanov normal form with D_4 symmetry near the O(2) limit. *Physics Letter A 284* (2001), 99–111.

Sławomir Piasecki, Roberto Barrio and Fernando Blesa
Department of Applied Mathematics and IUMA
University of Zaragoza
`piasek@unizar.es, rbarrio@unizar.es and fblesa@unizar.es`

Monografías Matemáticas *García de Galdeano* **37**, 197–206 (2012)

From the heat equation to the Sobolev equation

Guy Vallet

Abstract. In this paper, we consider the theorem of Lions-Tartar in $W(0,T,V,V')$ with different “pivot-spaces” H. In a first part, depending on H, we have a look at the corresponding solved problem. Then, the second energy equality set forth in a second part.

Keywords: Lions-Tartar, pivot space, second energy.

AMS classification: 35K05, 58D25, 35B65.

§1. Introduction

Considering two separable Hilbert spaces V and H, V being continuously embedded in H and dense in H, the interpretation of the equation $du/dt + Au = f$ in V', with initial condition u_0 in H, is under discussion in the situation where one changes the pivot space H in the usual Gelfand-Lions framework.

In [5], J. Simon warns us against the use of the common identification of H with its dual space in the functional frame $V \hookrightarrow H \equiv H' \hookrightarrow V'$. In particular, it is mentioned that if $D(\Omega)$[†] is not dense in V, the study is incompatible with the distributional frame for some standard PDE's. Remaining with $D(\Omega)$ dense V, we present in this paper the different type of solved problems by the theorem of Lions-Tartar when one changes the pivot's space. We systematically illustrate our remarks with the rigged Hilbert space $(H^s(\mathbb{R}^d), H^1(\mathbb{R}^d))$ when $s \in [0,1]$. For example, the equation $du/dt - \Delta u = f$ would correspond to the heat equation $\partial u/\partial t - \Delta u = f$ if $s = 0$. Here, $\partial u/\partial t$ denotes the time derivative of u in the sense of the distribution of $D'(Q)$. It would correspond to the Sobolev equation $(I - \Delta)\partial u/\partial t - \Delta u = f$ if $s = 1$.

In a last part of the paper, we will be interested in the “second energy equality” for the solution to the lemma of Lions-Tartar. More precisely, Theorem 4 asserts that if $u_0 \in V$, $g \in L^2(0,T,H)$ and assuming that the bilinear form a is independent of time, symmetric and coercive, then the corresponding solution u to the lemma of Lions-Tartar belongs to $C([0,T],V)$ and for any $t \in [0,T]$,

$$\int_{]0,t[} \left|\frac{du}{dt}\right|^2 d\sigma + \frac{1}{2}a(u(t),u(t)) = \frac{1}{2}a(u(0),u(0)) + \int_{]0,t[} \left(g(\sigma), \frac{du}{dt}(\sigma)\right) d\sigma.$$

Outlines of the paper

One presents in Section 2 some notations, then, in Section 3, one reminds the reader of the embedding of V in V' when the Riesz-identification $H \equiv H'$ is assumed. In particular, what is

[†] The space of infinitely differentiable functions with a compact support.

the characterization of the image of $H^1(\mathbb{R}^d)$ when $H = H^s(\mathbb{R}^d)$. Then, thanks to this, we will be interested in the sense given to the space $W(0,T) = \{u \in L^2(0,T,V),\ du/dt \in L^2(0,T,V')\}$. We will look more closely to the case $V = H^1(\mathbb{R}^d)$ and $H = H^s(\mathbb{R}^d)$ when $s \in [0,1]$ and to the link with fractional operators.

Section 4 will be devoted to the lemma of Lions-Tartar and Section 5 to the second energy equality. Then, we end this paper with an annex that precise the regularization of Landes, used in the proof of the result of Section 5.

§2. Notations

Let V and H be two separable Hilbert spaces, with norm $\|\cdot\|$ for V, associated with the scalar product $((\cdot,\cdot))$, and norm $|\cdot|$ for H, associated with the scalar product $(\cdot,\cdot)$. Assume moreover that V is continuously embedded in H with a dense injection. Then, the dual space H' is continuously embedded in V' and dense. The norm in V' is denoted by $\|\cdot\|_*$.

$\Omega \subset \mathbb{R}^d$ denotes a regular open set and for any positive T, $Q =]0,T[\times\Omega$.

As usual, $D(A)$ denotes the class of C^∞-derivable functions in a given open set A, with compact support in A and its dual space $D'(A)$ denotes the space of distributions in A.

$\mathcal{S}$ denotes the Schwartz space in $\mathbb{R}^d$ and $\mathcal{S}'$ the tempered distributions.

For any $s \in [0,1]$, $H^s(\mathbb{R}^d)$ denotes the fractional Sobolev space defined, for any s, by $H^s(\mathbb{R}^d) = \{u \in L^2(\mathbb{R}^d),\ |\xi|^s\mathcal{F}_x(u) \in L^2(\mathbb{R}^d)\}$, where $\mathcal{F}_x$ is the Fourier transform of variable $x \in \mathbb{R}^d$.

Given $s \in]-d/2,1]$ and $f \in \mathcal{S}$, we recall the fractional operators $(-\Delta)^s f$ as $(-\Delta)^s f = \mathcal{F}_x^{-1}[|\xi|^{2s}\mathcal{F}_x(f)]$ and $(I-\Delta)^s f$ as $(I-\Delta)^s f = \mathcal{F}_x^{-1}[(1+|\xi|^2)^s\mathcal{F}_x(f)]$.

Then, one denotes by $W_{(H,V)}(0,T) = \{u \in L^2(0,T,V),\ du/dt \in L^2(0,T;V')\}$.

§3. The space $W_{(H,V)}(0,T)$

3.1. How to embed V in V'?

In this section, we lay stress on the question: how to embed V in V'? Since V is not *a priori* a space with a finite dimension, there exist many possibilities to identify V with its image in V' when one says that $V \hookrightarrow V'$?

Classically, the rigged Hilbert space (H,V) is considered (or Gelfand-Lions triple):

1. Either $H = V$. Then, thanks to the theorem of Riesz, V is identified with its dual V'. Indeed,
$$J : V \to V',\ u \mapsto Ju \quad \text{such that} \quad Ju : v \in V \mapsto ((u,v))$$
is an isometric mapping.
2. Or, $H \subsetneq V$. Then, H is identified with its dual H' (Riesz's theorem) and V is embedded in V' by "passing through $H \equiv H'$". H is called the pivot-space, or intermediate space. Then,
$$J_H : V \to V',\ u \mapsto J_H u \quad \text{such that} \quad J_H u : v \in V \mapsto (u,v)$$
is an injective mapping.

Remark 1.

1. Note that if $H = V$, then $J_H = J$.
2. If H and $\widetilde{H}$ are two pivot-spaces with $H \subsetneq \widetilde{H}$ and V is densely embedded in H and $\widetilde{H}$, then we get that $J_{\widetilde{H}}(V) \subsetneq J_H(V)$.
3. If $V = H^1(\mathbb{R}^d)$ and $H = L^2(\mathbb{R}^d)$ then, for any $u \in V$, we get that $J^{-1} \circ J_H u = w$ where w is the unique solution in $H^1(\mathbb{R}^d)$ of the problem: $w - \Delta w = (u, \cdot)_{L^2(\mathbb{R}^d)}$.

3.1.1. Fractional Sobolev spaces

Let us recall some basics about $H^s(\mathbb{R}^d)$ from J.-L. Lions *et al.* [2] and L. Tartar [7]. Remind that $\mathcal{S}$ denotes the Schwartz space and $\mathcal{S}'$ the tempered distributions.

Definition 1. Let us denote by $\mathcal{F}_x$ the Fourier transform of variable $x \in \mathbb{R}^d$. Then, for a real number $s \geq 0$, $H^s(\mathbb{R}^d) = \{u \in L^2(\mathbb{R}^d),\ |\xi|^s \mathcal{F}_x(u) \in L^2(\mathbb{R}^d)\}$, and, for a real number s, $H^s(\mathbb{R}^d) = \{u \in \mathcal{S}'(\mathbb{R}^d),\ (1 + |\xi|^2)^{s/2}\mathcal{F}_x(u) \in L^2(\mathbb{R}^d)\}$.

Then,

Lemma 1.

1. *When $s \in \mathbb{N}$, $H^s(\mathbb{R}^d)$ denotes the classical Sobolev space (with $H^0(\mathbb{R}^d) = L^2(\mathbb{R}^d)$).*
2. *$D(\mathbb{R}^d)$ is dense in the Hilbert space $H^s(\mathbb{R}^d)$ for the norm $u \mapsto \left\|[1 + |\xi|^2]^{s/2}\mathcal{F}_x(u)\right\|_{L^2(\mathbb{R}^d)}$.*
3. *If $s \in]0, 1[$, then $u \in H^s(\mathbb{R}^d)$ if and only if $u \in L^2(\mathbb{R}^d)$ and*

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{|u(x) - u(y)|^2}{|x - y|^{d+2s}}\, dx\, dy < \infty.$$

For an open set Ω, one could define $H^s(\Omega)$ for $0 < s < 1$ in (at least) three different ways:

1. $u \in L^2(\Omega)$ and $\displaystyle\int_{\Omega\times\Omega} \frac{|u(x) - u(y)|^2}{|x - y|^{d+2s}}\, dx\, dy < \infty$.
2. u is the restriction to Ω of an element U in $H^s(\mathbb{R}^d)$.
3. One may define $H^s(\Omega)$ by interpolation $H^s(\Omega) = [H^1(\Omega), L^2(\Omega)]_{1-s,2}$.

For a bounded open set with a Lipschitz boundary, the three definitions give the same space with equivalent norms.

3.1.2. Fractional Laplace operator

Let us now remind some basics on the fractional operators (cf. L. E. Silvestre [3]):

Definition 2. Given $s \in]-d/2, 1]$, and $f \in \mathcal{S}$, we define:

1. $(-\Delta)^s f$ as $(-\Delta)^s f = \mathcal{F}_x^{-1}[|\xi|^{2s}\mathcal{F}_x(f)]$.
2. $(I - \Delta)^s f$ as $(I - \Delta)^s f = \mathcal{F}_x^{-1}[(1 + |\xi|^2)^s \mathcal{F}_x(f)]$.

Clearly, if $s = 1$, then $(-\Delta)^s = -\Delta$; if $s = 0$, then $(-\Delta)^s = Id$; and $(-\Delta)^{s_1} \circ (-\Delta)^{s_2} = (-\Delta)^{s_1+s_2}$, respectively with $I - \Delta$ instead of $-\Delta$.

When $f \in \mathcal{S}$, we can also compute the same operator by using the singular integral

$$(-\Delta)^s f(x) = c_{n,s} PV \int_{\mathbb{R}^d} \frac{f(x) - f(y)}{|x-y|^{d+2s}}\, dy.$$

Let us remark also that for any $f, g \in \mathcal{S}$,

$$\begin{aligned}\int_{\mathbb{R}^d} [(Id - \Delta)^s] f\, g\, dx &= \int_{\mathbb{R}^d} (1 + |\xi|^2)^s \mathcal{F}_x(f)\mathcal{F}_x(g)\, d\xi \\ &= \int_{\mathbb{R}^d} (1 + |\xi|^2)^{s/2} \mathcal{F}_x(f)(1 + |\xi|^2)^{s/2} \mathcal{F}_x(g)\, d\xi \\ &= \int_{\mathbb{R}^d} (I - \Delta)^{s/2} f (I - \Delta)^{s/2} g\, dx,\end{aligned}$$

which is the scalar product of $H^s(\mathbb{R}^d)$.

3.1.3. Intermediate spaces

If one assumes that $V \hookrightarrow H \equiv H' \hookrightarrow V'$, then (J.-L. Lions *et al.* [2]) there exists an unbounded operator A on V' such that $[D(A), H]_{1/2} = V$, $[V, V']_{1/2} = H$ and $D(A^{1/2}) = V$.

Classically, when $V = H^1(\mathbb{R}^d)$, we consider that $H = L^2(\mathbb{R}^d)$. Therefore, the image of the dual of $H^1(\mathbb{R}^d)$ by the identification $L^2(\mathbb{R}^d)' \equiv L^2(\mathbb{R}^d)$ is $H^{-1}(\mathbb{R}^d)$, the space of "derivatives of order less than one of elements of L^2", and $[H^1(\mathbb{R}^d), H^{-1}(\mathbb{R}^d)]_{1/2} = H^0(\mathbb{R}^d) = L^2(\mathbb{R}^d)$. Moreover, since we have

$$D(\mathbb{R}^d) \hookrightarrow H^1(\mathbb{R}^d) \hookrightarrow L^2(\mathbb{R}^d) \equiv L^2(\mathbb{R}^d)' \hookrightarrow H^{-1}(\mathbb{R}^d) \hookrightarrow D'(\mathbb{R}^d),$$

any element u of $H^1(\mathbb{R}^d)$ is a distribution *via* the identification $L^2(\mathbb{R}^d) \equiv L^2(\mathbb{R}^d)'$, *i.e.*, u is identifiable with the distribution: $\varphi \in D(\mathbb{R}^d) \mapsto \int_{\mathbb{R}^d} u\varphi\, dx$.

Consider now that the pivot-space is $H^s(\mathbb{R}^d)$, for a given $s \in [0, 1]$. Then,

$$D(\mathbb{R}^d) \hookrightarrow H^1(\mathbb{R}^d) \hookrightarrow H^s(\mathbb{R}^d) \equiv H^s(\mathbb{R}^d)' \hookrightarrow H^1(\mathbb{R}^d)' \hookrightarrow D'(\mathbb{R}^d),$$

and an element u of $H^1(\mathbb{R}^d)$ is a distribution *via* the identification $H^s(\mathbb{R}^d) \equiv H^s(\mathbb{R}^d)'$, *i.e.*, u is identifiable with the distribution: $\varphi \in D(\mathbb{R}^d) \mapsto (u, \varphi)_{H^s}$.

Now, the question is: since in this case $[H^1(\mathbb{R}^d), H^1(\mathbb{R}^d)']_{1/2} = H^s(\mathbb{R}^d)$, what is the image in the dual of $H^1(\mathbb{R}^d)$ by the identification $H^s(\mathbb{R}^d)' \equiv H^s(\mathbb{R}^d)$? More precisely, since we have to obtain $[H^1(\mathbb{R}^d), H^1(\mathbb{R}^d)']_{1/2} = H^s(\mathbb{R}^d)$, why can we identify $H^1(\mathbb{R}^d)'$ with $H^{2s-1}(\mathbb{R}^d)$? Indeed, let us denote by

$$\Phi : H^{2s-1}(\mathbb{R}^d) \to H^1(\mathbb{R}^d)';\ w \mapsto \Phi_w$$

where

$$\Phi_w : H^1(\mathbb{R}^d) \to \mathbb{R};\ u \mapsto \int_{\mathbb{R}^d} (1 + |\xi|^2)^s \mathcal{F}_x w \mathcal{F}_x u\, d\xi.$$

Clearly, Φ exists and is an injection.

Consider $T \in H^1(\mathbb{R}^d)'$. Since $H^s(\mathbb{R}^d)'$ is dense in $H^1(\mathbb{R}^d)'$, T is the limit of a sequence $(T_n) \subset H^s(\mathbb{R}^d)'$ in $H^1(\mathbb{R}^d)'$. Since $H^s(\mathbb{R}^d)$ is the pivot space, there exists $w_n \in H^s(\mathbb{R}^d)$ such that

$$\forall v \in H^1(\mathbb{R}^d),\ \langle T_n, v\rangle = (w_n, v)_{H^s} = \int_{\mathbb{R}^d} (1+|\xi|^2)^s \mathcal{F}_x w_n \mathcal{F}_x v\, d\xi.$$

Since $H^1(\mathbb{R}^d) = \left\{u \in \mathcal{S}',\ \int_{\mathbb{R}^d}(1+|\xi|^2)|\mathcal{F}_x v|^2\, d\xi < +\infty\right\}$ and $\|v\|^2_{H^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d}(1+|\xi|^2)|\mathcal{F}_x|^2 v\, d\xi$,

$$\|T_n\|_{H^1(\mathbb{R}^d)'} = \sup_{v \in H^1(\mathbb{R}^d)\setminus\{0\}} \frac{\int_{\mathbb{R}^d}(1+|\xi|^2)^{s-1/2}\mathcal{F}_x w_n (1+|\xi|^2)^{1/2}\mathcal{F}_x v\, d\xi}{\|v\|_{H^1(\mathbb{R}^d)}} = \|w_n\|_{H^{2s-1}(\mathbb{R}^d)}.$$

Then, the result holds by passing to the limit and Φ is an isometry.

3.2. Time derivation

Consider a positive real number T and assume that $u \in L^2(0, T, V)$. Then, u is said to belong to $W_{(H,V)}(0,T)$ if $u \in L^2(0,T,V)$, $du/dt \in L^2(0,T;V')$ and $V \hookrightarrow H \equiv H' \hookrightarrow V'$. Then, in this section, we wish to discuss about the sense given to $du/dt \in L^2(0,T;V')$ (cf. J. Simon [4]).

1. On the one hand, one can consider u as an element of $D'(0,T;V)$, the V-valued distributions. Thus, du/dt, the time derivative of u in the sense of $D'(0,T;V)$, exists and

   $$\forall \varphi \in D(0,T),\ \frac{du}{dt}(\varphi) = -\int_0^T u(t)\varphi'(t)\, dt \text{ in } V.$$

   Then, by using $J_H : V \hookrightarrow H \equiv H' \hookrightarrow V'$, we have that

   $$\forall \varphi \in D(0,T),\ \forall v \in V,\ \left\langle J_H\left[\frac{du}{dt}(\varphi)\right], v\right\rangle = -\left(\int_0^T u(t)\varphi'(t)\, dt, v\right).$$

   Since $u\varphi' \in L^1(0,T,V)$ and $T : V \to \mathbb{R}, u \mapsto (u,v)$ is a linear and continuous mapping, we get that $\left(\int_0^T u(t)\varphi'(t)\, dt, v\right) = \int_0^T \varphi'(t)(u(t),v)\, dt$. Note that this result is an obvious fact for simple functions u, then for any u by passing to the limit. Therefore, for all $\varphi \in D(0,T)$ and $v \in V$,

   $$\left\langle J_H\left[\frac{du}{dt}(\varphi)\right], v\right\rangle = -\int_0^T \varphi'(t)(u(t),v)\, dt = \left\langle \frac{d}{dt}(u(t),v), \varphi\right\rangle_{D'(0,T),D(0,T)}.$$

2. On the other hand, one can consider that $J_H(u)$ is then an element of $L^2(0,T,V')$, thus an element of $D'(0,T;V')$, the V'-valued distributions. Therefore, dJ_Hu/dt, the time derivative of J_Hu in the sense of $D'(0,T;V')$, exists and

   $$\forall \varphi \in D(0,T),\ \frac{dJ_Hu}{dt}(\varphi) = -\int_0^T J_Hu(t)\varphi'(t)\, dt \text{ in } V',$$

   i.e.

   $$\forall \varphi \in D(0,T),\ \forall v \in V,\ \left\langle \frac{dJ_Hu}{dt}(\varphi), v\right\rangle = -\left\langle \int_0^T J_Hu(t)\varphi'(t)\, dt, v\right\rangle.$$

Since $J_H u\varphi' \in L^1(0,T,V')$ and $T : V' \to \mathbb{R}$, $f \mapsto \langle f, v\rangle$ is a linear and continuous mapping, we get also that $\left\langle \int_0^T J_H u(t)\varphi'(t)\,dt, v\right\rangle = \int_0^T \varphi'(t)\,\langle u(t), v\rangle\,dt$. Therefore, for all $\varphi \in D(0,T)$ and $v \in V$,

$$\begin{aligned}\left\langle \frac{dJ_H u}{dt}(\varphi), v\right\rangle &= -\int_0^T \varphi'(t)\,\langle J_H u(t), v\rangle\,dt \\ &= \left\langle \frac{d}{dt}\,\langle J_H u(t), v\rangle\,, \varphi\right\rangle_{D'(0,T),D(0,T)} = \left\langle \frac{d}{dt}(u(t), v), \varphi\right\rangle_{D'(0,T),D(0,T)}.\end{aligned}$$

Thus, $J_H \circ \dfrac{du}{dt} = \dfrac{dJ_H u}{dt}$.

Assume, for example, that $V = H^1(\mathbb{R}^d)$ and $H = H^s(\mathbb{R}^d)$ with $s \in [0,1]$. Then, for any $v \in D(\mathbb{R}^d)$ and any $\varphi \in D(0,T)$,

$$\left\langle \frac{dJ_H u}{dt}(\varphi), v\right\rangle = -\int_0^T \varphi'(t)(u(t), v)_{H^s(\mathbb{R}^d)}\,dt = \left\langle \frac{d}{dt}(u(t), v)_{H^s(\mathbb{R}^d)}, \varphi\right\rangle_{D'(0,T),D(0,T)}.$$

1. Assume that $s = 0$. Then,

   $$\left\langle \frac{dJ_{L^2(\mathbb{R}^d)} u}{dt}(\varphi), v\right\rangle = -\int_0^T \varphi'(t)\int_{\mathbb{R}^d} uv\,dx\,dt = \left\langle \frac{\partial u}{\partial t}, \varphi \otimes v\right\rangle_{D'(Q),D(Q)},$$

   where $\partial u/\partial t$ denotes the time derivative of u in the sense of the distribution of $D'(Q)$ where $Q =]0,T[\times \mathbb{R}^d$ with the classical identification $L^2 \equiv (L^2)'$.

2. Assume that $s = 1$. Then, up eventually to a constant due to the Fourier transform,

   $$\left\langle \frac{dJ_{H^1(\mathbb{R}^d)} u}{dt}(\varphi), v\right\rangle = -\int_0^T \varphi'(t)\int_{\mathbb{R}^d}(uv + \nabla u \nabla v)\,dx\,dt = \left\langle \frac{\partial u}{\partial t} - \Delta\frac{\partial u}{\partial t}, \varphi \otimes v\right\rangle_{D'(Q),D(Q)},$$

   where the derivations are in the sense of the distribution of $D'(Q)$ with the classical identification $L^2 \equiv (L^2)'$.

3. Assume that $s \in]0,1[$. Then,

   $$\left\langle \frac{dJ_{H^s(\mathbb{R}^d)} u}{dt}(\varphi), v\right\rangle = -\int_0^T \varphi'(t)\,\langle (I-\Delta)^s u, v\rangle_{(H^s)',H^s}\,dt$$

   and

   $$\frac{dJ_{H^s(\mathbb{R}^d)} u}{dt} = (I-\Delta)^s\frac{du}{dt},$$

   where du/dt is understood in the sense of $D'(0,T;H^s(\mathbb{R}^d))$.

§4. Lemma of Lions-Tartar

Lemma 2 (J.-L. Lions [1], J. Simon [4, 5] and L. Tartar [6]). *Let $a \in L^\infty(0,T,\mathcal{L}(V,V'))$ such that*

$$\exists\alpha > 0,\ \beta \in \mathbb{R},\ \textit{for which},\ \forall u \in V,\ a(u,u) \geq \alpha\|u\|^2 - \beta|u|^2.$$

Given $u_0 \in H$, $f_1 \in L^1(0,T;H')$ and $f_2 \in L^2(0,T;V')$, there exists a unique $u \in C([0,T];H) \cap L^2(0,T;V)$, solution, for any $v \in V$ and t a.e. in $]0,T[$, *of*

$$\begin{cases} \dfrac{d}{dt}(u,v) + \langle a(\cdot,u),v\rangle_{V',V} = \langle f_1,v\rangle_{H',H} + \langle f_2,v\rangle_{V',V}\,, \\ u(0) = u_0, \end{cases} \tag{1}$$

and the bilinear application $(f_1+f_2,u_0) \mapsto u$ is continuous from $(L^2(0,T;V')+L^1(0,T;H')) \times H$ to $L^2(0,T;V) \cap C([0,T];H)$. Moreover,

$$\frac{dJ_H u}{dt} \in L^1(0,T;H') + L^2(0,T;V')$$

and the first energy equality holds

$$\frac{1}{2}\frac{d}{dt}|u|^2 + \langle a(\cdot,u),u\rangle_{V',V} = \langle f_1,u\rangle_{H',H} + \langle f_2,u\rangle_{V',V}\,.$$

Lemma 3 (J. Simon [4, 5])**.** *With the same hypothesis than the previous lemma, unless $a \in L^2(0,T,\mathcal{L}(V,V'))$ (instead of L^∞), there exists a unique u in $L^2(0,T;V) \cap L^\infty(0,T;H) \cap C_w([0,T];H)$ solution of* (1)*. Moreover,*

$$\frac{dJ_H u}{dt} \in L^1(0,T;V').$$

Remark 2.

1. J.-L. Lions considered $f \in L^2(0,T;V')$ which gives $dJ_H u/dt \in L^2(0,T;V')$, *i.e.*, $u \in W_{(H,V)(0,T)}$.
2. Assume for example that $V = H^1(\mathbb{R}^d)$, $H = H^s(\mathbb{R}^d)$ with $s \in [0,1]$, that $\langle a(\cdot,u),v\rangle_{V',V} = \int_{\mathbb{R}^d} \nabla u.\nabla v\,dx$ and denote by u_s the solution of Lions-Tartar's lemma. Then, if $s = 0$, u_s is the solution of the heat equation; if $s = 1$, u_s is the solution of the pseudoparabolic Sobolev equation; else, u_s is the solution of intermediate evolution problems, hard to characterize in term of PDE's since $(I-\Delta)^s$ is a non local fractional operator.

§5. Second energy equality

Theorem 4. *Consider $T > 0$, $Q =]0,T[\times \Omega$, $u_0 \in V$, $g \in L^2(0,T,H)$ and u the solution of the lemma of Lions-Tartar. If a is independent of time, symmetric and coercive (i.e. $\beta = 0$) bilinear form, then $u \in H^1(0,T;H) \cap C_w([0,T],V)$. Moreover, $u \in C([0,T],V)$ and for any $t \in [0,T]$,*

$$\int_{]0,t[} \left|\frac{du}{dt}\right|^2 d\sigma + \frac{1}{2}a\,(u(t),u(t)) = \frac{1}{2}a\,(u(0),u(0)) + \int_{]0,t[} \Big(g(\sigma), \frac{du}{dt}(\sigma)\Big)\,d\sigma. \tag{2}$$

Proof. Since u is a mild solution, *i.e.* obtained by an implicit time-discretization scheme, it is a classic exercise to prove that $u \in H^1(0,T;H) \cap L^\infty(0,T;V)$. Then,

$$u \in C([0,T];H) \cap L^\infty(0,T;V) = C_w([0,T];V)$$

(cf. [2]). Moreover, since $u \in H^1(0,T;H)$, the time differentiation is understood in the space H, without any embeddings. Then, we will denote it by du/dt.

Let us fix $s \in [0,T[$ and for any positive ϵ, denote by v_ϵ the solution of the differential equation (see section 6 for further informations)

$$\epsilon\frac{dv_\epsilon}{dt} + v_\epsilon = u, \text{ for } t > s, \quad \text{with } v_\epsilon(s,.) = u(s).$$

Then, testing the evolution equation with $u - v_\epsilon$ leads us to

$$\epsilon\int_{]s,t[}\Big(\frac{du}{dt},\frac{dv_\epsilon}{dt}\Big)d\sigma + \int_{]s,t[} a(u,u-v_\epsilon)\,d\sigma = \epsilon\int_{]s,t[}\Big(g,\frac{dv_\epsilon}{dt}\Big)d\sigma.$$

Thus, by monotonicity of a,

$$\epsilon\int_{]s,t[}\Big(\frac{du}{dt},\frac{dv_\epsilon}{dt}\Big)d\sigma + \int_{]s,t[} a(v_\epsilon,u-v_\epsilon)\,d\sigma \leq \epsilon\int_{]s,t[}\Big(g,\frac{dv_\epsilon}{dt}\Big)d\sigma,$$

i.e., by using the differential equation, we get

$$\epsilon\int_{]s,t[}\Big(\frac{du}{dt},\frac{dv_\epsilon}{dt}\Big)d\sigma + \epsilon\int_{]s,t[} a\Big(v_\epsilon,\frac{dv_\epsilon}{dt}\Big)d\sigma \leq \epsilon\int_{]s,t[}\Big(g,\frac{dv_\epsilon}{dt}\Big)d\sigma,$$

and, by integration,

$$\int_{]s,t[}\Big(\frac{du}{dt},\frac{dv_\epsilon}{dt}\Big)d\sigma + \frac{1}{2}a\left(v_\epsilon(t),v_\epsilon(t)\right) \leq \int_{]s,t[}\Big(g,\frac{dv_\epsilon}{dt}\Big)d\sigma + \frac{1}{2}a(u(s),u(s)). \tag{3}$$

Since by construction (see annex) v_ϵ converges to u in $H^1(s,T;H) \cap L^2(s,T;V)$ and, for any t, $v_\epsilon(t)$ converges weakly to $u(t)$ in V,

$$\int_{]s,t[}\left|\frac{du}{dt}\right|^2 d\sigma + \frac{1}{2}a(u(t),u(t)) \leq \int_{]s,t[}\Big(g,\frac{du}{dt}\Big)d\sigma + \frac{1}{2}a(u(s),u(s)).$$

Moreover, $u \in C_w([0,T],V)$ and $\limsup\limits_{t\to s^+} a(u(t),u(t)) \leq a(u(s),u(s))$. Then, u is continuous from the right from $[0,T[$ to V.

Consider now $0 < t < t+\Delta t \leq T$. Then,

$$\int_0^t\Big(\frac{du}{dt},\frac{u(\sigma+\Delta t)-u(\sigma)}{\Delta t}\Big)d\sigma + \int_0^t a\Big(u(\sigma),\frac{u(\sigma+\Delta t)-u(\sigma)}{\Delta t}\Big)d\sigma$$
$$= \int_0^t\Big(g(\sigma),\frac{u(\sigma+\Delta t)-u(\sigma)}{\Delta t}\Big)d\sigma.$$

Thus,

$$\int_0^t\Big(\frac{du}{dt}(\sigma),\frac{u(\sigma+\Delta t)-u(\sigma)}{\Delta t}\Big)d\sigma + \frac{1}{2\Delta t}\int_t^{t+\Delta t} a(u(\sigma),u(\sigma))\,d\sigma$$
$$\geq \frac{1}{2\Delta t}\int_0^{\Delta t} a(u(\sigma),u(\sigma))\,d\sigma + \int_0^t\Big(g(\sigma),\frac{u(\sigma+\Delta t)-u(\sigma)}{\Delta t}\Big)d\sigma.$$

Therefore, the above remark yields

$$\int_0^t \left|\frac{du}{dt}(\sigma)\right|^2 d\sigma + \frac{1}{2}a(u(t),u(t)) \geq \frac{1}{2}a(u_0,u_0) + \int_0^t \left(g(\sigma), \frac{du}{dt}\right) d\sigma.$$

Adding this to (3) with $s = 0$, we get (2) for any $t \in [0,T[$. Then, $u \in C_w([0,T],V)$ and $\lim_{t\to s} a(u(t),u(t)) = a(u(s),u(s))$ yield $u \in C([0,T[,V)$. We conclude the proof by remarking that the same result holds for time $T+1$ instead of T. □

Corollary 5. *The same result holds if $\beta \neq 0$.*

Proof. If u is a solution, then it is also the solution, for any $v \in V$ and t a.e. in $]0,T[$, of

$$\frac{d}{dt}(u,v) + a(u,v) + \beta(u,v) = (g+\beta u, v), \quad \text{with } u(0) = u_0. \tag{4}$$

Then, the result is just a consequence of the theorem. □

§6. Annex

Let us fix $s \in [0,T[$ and, for any positive ϵ, denote by v_ϵ the solution of the differential equation

$$\epsilon\frac{dv_\epsilon}{dt} + v_\epsilon = u, \text{ for } t > s, \quad \text{with } v_\epsilon(s,\cdot) = u(s), \tag{5}$$

where $u \in H^1(s,T,H) \cap C_w([s,T],V)$.

Lemma 6. *As ϵ goes to 0^+, v_ϵ converges to u in $H^1(s,T;H) \cap L^2(s,T;V)$ and $v_\epsilon(t)$ converges weakly to $u(t)$ in V, for any t.*

Proof. If v_ϵ is the solution of (5), then,

$$v_\epsilon(t) = u(s)e^{(s-t)/\epsilon} + \int_s^t \frac{u(\sigma)}{\epsilon} e^{(\sigma-t)/\epsilon} d\sigma$$

and $v_\epsilon(t)$ is bounded in V, independently of t. Thus, by "multiplying in V" equation (5) by v_ϵ, we get that

$$\epsilon\frac{d}{dt}\|v_\epsilon\|^2 + \|v_\epsilon\|^2 \leq \|u\|^2,$$

i.e.

$$\epsilon\|v_\epsilon(t)\|^2 + \int_s^t \|v_\epsilon\|^2 d\sigma \leq \int_s^t \|u\|^2 d\sigma + \epsilon\|u(s)\|^2. \tag{6}$$

Moreover, dv_ϵ/dt satisfies

$$\epsilon\frac{d^2v_\epsilon}{dt^2} + \frac{dv_\epsilon}{dt} = \frac{du}{dt}, \text{ for } t > s, \quad \text{with } \frac{dv_\epsilon}{dt}(s) = 0, \tag{7}$$

where $du/dt \in L^2(s,T,H)$. Thus, by "multiplying in H" the above equation by dv_ϵ/dt, we get that

$$\epsilon\frac{d}{dt}\left|\frac{dv_\epsilon}{dt}\right|^2 + \left|\frac{dv_\epsilon}{dt}\right|^2 \leq \left|\frac{du}{dt}\right|^2,$$

i.e.

$$\epsilon\left|\frac{dv_\epsilon}{dt}(t)\right|^2 + \int_s^t \left|\frac{dv_\epsilon}{dt}\right|^2 d\sigma \le \int_s^t \left|\frac{du}{dt}\right|^2 d\sigma. \tag{8}$$

As a first conclusion, there exists a positive constant C such that

$$\left|\frac{dv_\epsilon}{dt}\right|_{L^2(s,T,H)} \le C; \quad \forall t,\ \sqrt{\epsilon}\left|\frac{dv_\epsilon}{dt}(t)\right| \le C, \quad |v_\epsilon(t) - u(t)| \le C\sqrt{\epsilon},$$

and v_ϵ converges weakly to u in $H^1(s,T,H)$ and strongly in $C([s,T],H)$.

Adding that $v_\epsilon(t)$ is bounded in V for any t, $v_\epsilon(t)$ converges weakly to $u(t)$ in V for any t and v_ϵ converges weakly to u in $L^2(s,T,V)$ (note that u is the only possible limit-point).

Then, on the one hand, (6) yields

$$\limsup_{\epsilon\to 0^+} \int_s^t \|v_\epsilon\|^2\, d\sigma \le \int_s^t \|u\|^2\, d\sigma$$

and v_ϵ converges to u in $L^2(s,T,V)$. On the other hand, (8) yields

$$\limsup_{\epsilon\to 0^+} \int_s^t \left|\frac{dv_\epsilon}{dt}\right|^2 d\sigma \le \int_s^t \left|\frac{du}{dt}\right|^2 d\sigma$$

and v_ϵ converges to u in $H^1(s,T,H)$. □

References

[1] Dautray, R., and Lions, J.-L. *Analyse mathématique et calcul numérique pour les sciences et les techniques.* Masson, 1988.

[2] Lions, J. L., and Magenes, E. *Problèmes aux limites non homogenes et applications. Vol. 1.* Paris: Dunod 1: XIX, 372 p.; 2: XV, 251 p., 1968.

[3] Silvestre, L. E. *Regularity of the obstacle problem for a fractional power of the Laplace operator.* ProQuest LLC, Ann Arbor, MI, 2005. Thesis (Ph.D.)–The University of Texas at Austin. Available from: `http://proquest.umi.com/pqdlink?did=954063741&Fmt=7&clientId=79356&RQT=309&VName=PQD`.

[4] Simon, J. Una generalización del teorema de Lions-Tartar. *Boletín SeMA 40* (2007), 43–69.

[5] Simon, J. On the identification $H = H'$ in the Lions theorem and a related inaccuracy. *Ric. Mat.* (to appear).

[6] Tartar, L. *An introduction to Navier-Stokes equation and oceanography*, vol. 1 of *Lecture Notes of the Unione Matematica Italiana.* Springer-Verlag, Berlin, 2006.

[7] Tartar, L. *An introduction to Sobolev spaces and interpolation spaces*, vol. 3 of *Lecture Notes of the Unione Matematica Italiana.* Springer, Berlin, 2007.

Guy Vallet
UMR CNRS 5142 - LMA, University of Pau
Postal address IPRA BP 1155 Pau Cedex (France)
guy.vallet@univ-pau.fr

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.

35. M. C. López de Silanes, M. Palacios, G. Sanz, J. J. Torrens, M. Madaune-Tort, C. Paroissin y D. Trujillo, (eds.): *Tenth International Conference Zaragoza-Pau on Applied Mathematics and Statistics*, 2010, xxx + 302 pp., ISBN: 978-84-15031-53-6.

36. L. M. Esteban, B. Lacruz, F. J. López, P. M. Mateo, A. Pérez-Palomares, G. Sanz y C. Paroissin, (eds.): *The Pyrenees International Workshop on Statistics, Probability and Operations Research: SPO 2009*, 2011, 164 pp., ISBN: 978-84-15031-92-5.

ISBN 978-84-15538-15-8

9 788415 538158

1542

Prensas Universitarias
Universidad Zaragoza

monografías
garcía de galdeano

Instituto Universitario de Investigación
de Matemáticas
y Aplicaciones
Universidad Zaragoza