Fifteenth International Conference Zaragoza-Pau on Mathematics and its Applications

Etienne Ahusborde, Gilles Carbou, Chérif Amrouche, José Luis Gracia, María Cruz Lopez de Silanes, Manuel Palacios

MONOGRAFÍAS MATEMÁTICAS
GARCÍA DE GALDEANO

Número **42**, 2019

Fifteenth International Conference Zaragoza-Pau on Mathematics and its Applications

Jaca (Spain), September 10–12, 2018

Editors

É. Ahusborde
C. Amrouche
G. Carbou
Université de Pau et des Pays de l'Adour, France

J. L. Gracia
M. C. López de Silanes
M. Palacios
Universidad de Zaragoza, Spain

JORNADAS ZARAGOZA–PAU DE MATEMÁTICAS (15.ª. Jaca 2018. Jaca)
Fifteenth International Conference Zaragoza–Pau on Mathematics and its Applications : Jaca (Spain), September 10 – 12, 2018 / editors É. Ahusborde… [et al.]. — Zaragoza : Prensas de la Universidad de Zaragoza : Instituto Universitario de Investigación de Matemáticas y Aplicaciones, Universidad de Zaragoza, 2019
XXVIII, 306 p. ; 24 cm. — (Monografías Matemáticas García de Galdeano ; 42)
ISBN [redacted]
Matemáticas-Congresos y asambleas
Ahusborde, É.
51(063)

Monografías Matemáticas García de Galdeano n.º 42
Noviembre 2019
Universidad de Zaragoza

Imprime: Servicio de Publicaciones. Universidad de Zaragoza

D.L.: [redacted]
ISBN: [redacted]

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

XV Journées Jornadas ZARAGOZA-PAU

Fifteenth International Conference Zaragoza-Pau on Mathematics and its Applications

Jaca

September 10-12
Residencia Universitaria de Jaca
2018

PLENARY SPEAKERS

Martin BUHMANN, Universität Giessen, Germany
Thierry GALLOUET, Aix Marseille Université, France
Raphaëlle HERBIN, Aix Marseille Université, France
Radu IGNAT, Université Paul Sabatier, Toulouse, France
Mihai MARIS, Université Paul Sabatier, Toulouse, France
Peter MASSOPUST, Technische Universität München, Germany
Eugene O'RIORDAN, Dublin City University, Ireland
Tomas SAUER, Universität Passau, Germany
Martin STYNES, Beijing Computational Science Research Center, China

http://pcmap.unizar.es/~jaca2018

cnrs · INSTITUT CARNOT ISIFoR · Instituto Universitario de Investigación de Matemáticas y Aplicaciones Universidad Zaragoza · Université de Pau et des Pays de l'Adour · Universidad Zaragoza

Contents

Preface **xiii**

Contributors **xv**

List of participants **xvii**

Other communications **xxv**

Published articles

Generalized resolvent of the Stokes problem with Navier-type boundary conditions
H. Al Baba and A. Jabbour ***1***

Beyond Wentzell-Freidlin: semi-deterministic approximations for diffusions with small noise and a repulsive critical boundary point
F. Avram and J. Cresson ***13***

Adaptive augmented mixed FEM for the Oseen problem with mixed boundary conditions
T. P. Barrios, J. M. Cascón and M. González ***25***

A triaxial model for the roto-orbital coupling in a binary system
A. Cantero, F. Crespo and S. Ferrer ***35***

Stability of domain walls in ferromagnetic rings
G. Carbou, M. Moussaoui and R. Rachi ***45***

Generalized fractional differential equations with order varying in time in complex Banach spaces: analytic and numerical asymptotic behavior
E. Cuesta and R. Ponce ***57***

A profile decomposition for the limiting Sobolev embedding
G. Devillanova and C. Tintarev ***65***

Mathematical aspects of computerized tomography: compression and compressed computing
B. Diederichs, T. Sauer and A. M. Stock ***79***

Convergence and error estimates for the compressible Navier-Stokes equations
T. Gallouët ***95***

A collocation method for a two-point boundary value problem with a Riemann-Liouville-Caputo fractional derivative
J. L. Gracia, E. O'Riordan and M. Stynes ***111***

A decoupled staggered scheme for the shallow water equations
R. Herbin, J. C. Latché, Y. Nasseri and N. Therme ***127***

Stabilized virtual element method for the incompressible Navier-Stokes equations
D. Irissari and G. Hauke ***143***

Analysis of the equilibria and limit cycle oscillations of flight dynamics and airfoil aeroelasticity
S. Kolb ***153***

Periodic solutions in the Hénon-Heiles rotating system
V. Lanchares, M. Iñarrea, J. Palacián, A. I. Pascual, J. P. Salas and P. Yanguas .. ***165***

Lipschitz spaces associated to the harmonic oscillator
M. de León-Contreras and J. L. Torrea ***173***

Accurate least squares fitting with a general class of shape preserving bases
E. Mainar, J. M. Peña and B. Rubio ***183***

On the Laplacian flow and coflow of G_2-structures
V. Manero, A. Otal and R. Villacampa ***193***

On some generalizations of B-splines
P. Massopust ***203***

Fractal Jackson approximation on the torus
M. A. Navascués, S. Jha, M. V. Sebastián and A. K. B. Chand ***219***

Non-associative algebraic hyperstructures and its applications to biological inheritance
M. A. Oyebola and T. G. Jaiyeola ***229***

Best regularity for a Schrödinger type equation with non smooth data and interpolation spaces
J. M. Rakotoson ***243***

Renormalized solutions for a stochastic p-Laplace equation with L^1 initial data
N. Sapountzoglou and A. Zimmermann ***253***

The matroid structure of vectors of the Mordell-Weil lattice and the topology of plane quartics and bitangent lines
R. Sato and S. Bannai ***265***

Sparse polynomial surrogates for uncertainty quantification in computational fluid dynamics
E. Savin ***275***

Tools to prove a parabolic Lewy-Stampacchia's inequality
Y. Tahraoui ***285***

Periodic solutions for impulsive differential equations
J. M. Uzal ***297***

PREFACE

The *International Conference Zaragoza-Pau on Mathematics and its Applications* was organized by the *Departamento de Matemática Aplicada*, the *Departamento de Métodos Estadísticos* and the *Departamento de Matemáticas*, all of them from the *Universidad de Zaragoza* (Spain), and the *Laboratoire de Mathématiques et de leurs Applications*, from the *Université de Pau et des Pays de l'Adour* (France). This conference has been held every two years since 1989. The aim of this conference is to present recent advances in Applied Mathematics, Statistics and Pure Mathematics, putting special emphasis on subjects linked to petroleum engineering and environmental problems.

The Fifteenth Conference took place in Jaca (Spain) from 10th to 12nd September 2018. The official opening ceremony was graced by the presence of the Chancellor of the University of Zaragoza, Rector Mgfco. D. José Antonio Mayoral Murillo, and the Chancellor of the University of Pau, M. le Président Mohamed Amara. During those three days, 99 mathematicians, coming from different universities, research institutes or the industrial sector, attended 9 plenary lectures, 63 contributed talks and a poster session with 7 posters. We note that in this edition there were 10 mini-symposia, two of them co-organized by colleagues from the *Universidad de Zaragoza* and the *Université de Pau et des Pays de l'Adour*.

The principal talks were about theoretical and numerical analysis of deterministic models described by partial differential equations, statistics and stochastics processes, surface approximation and image analysis. At the same time, there was also a discussion session about problems in Algebra and Geometry. These proceedings contain 26 refereed research papers, 25 of them based on the corresponding contributions and one paper by E. Savin, which was mislaid and not included in the monograph of the previous conference.

We would like to thank the following institutions for their regular financial and material support in our cooperation programmes: Université de Pau et des Pays de l'Adour, Universidad de Zaragoza and Gobierno de Aragón. Thanks are also due to the Institut Carnot ISIFoR, the Centre National de la Recherche Scientifique (CNRS), Common Funds Aquitaine-Aragón and European Social Fund (ESF), Instituto Universitario de Matemáticas y Aplicaciones (IUMA) and the Fédération IPRA of Pau (Institut Pluridisciplinaire de Recherche Appliquée) for the grants specially allotted at the time of the Fifteenth Conference.

We wish to express our gratitude to Alberto Abad (U. Zaragoza), Enrique Artal (U. Zaragoza), Carmelo Clavero (U. Zaragoza), Jacky Cresson (U. Pau), Marc Dambrine (U. Pau), Jacqueline Fleckinger (U. Toulouse I), Vincent Florens (U. Pau), Jacques Giacomoni (U. Pau), Pedro Jodrá (U. Zaragoza), Sophie Mercier (U. Pau), Pedro J. Miana (U. Zaragoza), Philippe Poncet (U. Pau), Carmen Sangüesa (U. Zaragoza), Peter Takáč (U. Rostock), Guy Vallet (U. Pau), who, together with us, formed the Scientific Committee. We would like also to express our special thanks to Pedro Mateo (U. Zaragoza) and to Juan José Torrens (U. Pública de Navarra), for their invaluable help in organizing the web and editing these proceedings, respectively. We are also indebted to all the others who helped in the organization of the Conference, in particular, María Luisa Gómez, Marta Gómez, María del Carmen Izaguerri and Beatriz Malo.

We finally acknowledge the kind cooperation of the referees, as well as the assistance provided for the realization of the proceedings by the Servicio de Publicaciones of the University of Zaragoza.

The next Conference Zaragoza-Pau will be held in Jaca from 9th to 11st September 2020. All of you are cordially invited to participate in this event.

Pau and Zaragoza, November, 2019
The Editors

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

Étienne Ahusborde
Chérif Amrouche
Gilles Carbou
Laboratoire de Mathématiques et de leurs Applications
Université de Pau et des Pays de l'Adour

Contributors

Al Baba, H., 1
Avram, F., 13

Bannai, S., 265
Barrios, T. P., 25

Cantero, A., 35
Carbou, G., 45
Cascón, J. M., 25
Chand, A. K. B., 219
Crespo, F., 35
Cresson, J., 13
Cuesta, E., 57

Devillanova, G., 65
Diederichs, B., 79

Ferrer, S., 35

Gallouët, T., 95
González, M., 25
Gracia, J. L., 111

Hauke, G., 143
Herbin, R., 127

Iñarrea, M., 165
Irissari, D., 143

Jabbour, A., 1
Jaiyeola, T. G., 229
Jha, S., 219

Kolb, S., 153

Lanchares, V., 165
Latché, J. C., 127
León-Contreras, M. de, 173

Mainar, E., 183
Manero, V., 193
Massopust, P., 203
Moussaoui, M., 45

Nasseri, Y., 127
Navascués, M. A., 219

O'Riordan, E., 111
Otal, A., 193
Oyebola, M. A., 229

Palacián, J., 165
Pascual, A. I., 165
Peña, J. M., 183
Ponce, R., 57

Rachi, R., 45
Rakotoson, J. M., 243
Rubio, B., 183

Salas, J. P., 165
Sapountzoglou, N., 253
Sato, R., 265
Sauer, T., 79
Savin, E., 275
Sebastián, M. V., 219
Stock, A. M., 79
Stynes, M., 111

Tahraoui, Y., 285
Therme, N., 127
Tintarev, C., 65
Torrea, J. L., 173

Uzal, J. M., 297

Villacampa, R., 193

Yanguas, P., 165

Zimmermann, A., 253

List of participants

ABAD, Alberto
Grupo de Mecánica espacial & IUMA
Facultad de Ciencias,
Universidad de Zaragoza,
Edificio de Matemáticas,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain. abad@unizar.es

AHUSBORDE, Étienne
Laboratoire de Mathématiques et de leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
etienne.ahusborde@univ-pau.fr

AL BABA, Hind
Laboratoire de Mathématiques et de leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
albaba@math.cas.cz

AL SAYED, Abdel Kader
École des Mines de Nancy,
Université de Lorraine
abdelkader.alsayed@univ-pau.fr

ALZIARY, Bénédicte
Université Toulouse I Capitole,
21 allée de Brienne,
31015 Toulouse Cedex 06, France.
alziary@ut-capitole.fr

AMROUCHE, Chérif
Laboratoire de Mathématiques et de 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

ARORA, Rakesh
Laboratoire de Mathématiques et de leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
rakesh.arora@univ-pau.fr

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

AVRAM, Florin
Laboratoire de Mathématiques et de leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
florin.avram@univ-pau.fr

BADÍA, Francisco Germán
Departamento de Métodos Estadísticos,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain.
gbadia@unizar.es

BANNAI Shinzo
National Institute of Technology,
Ibaraki College,
866 Nakane,Hitachinaka-shi,
Ibaraki-ken 312-8508 Japan.
sbannai@ge.ibaraki-ct.ac.jp

BUHMANN, Martin
Justus–Liebig University,
Heinrich–Buff–Ring 44,
35392 Giessen, Germany.
buhmann@math.uni-giessen.de

CANTERO, Antonio
DITEC, Facultad de Informática,
Universidad de Murcia,
30071 Espinardo, Murcia, Spain.
antonio.cantero@um.es

CARBOU, Gilles
Laboratoire de Mathématiques et de leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
gilles.carbou@univ-pau.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

CHAINAIS–HILLAIRET, Claire
Laboratoire Painlevé,
Université de Lille,
59650 Villeneuve d'Ascq, France.
claire.chainais@math.univ-lille1.fr

CLAVERO, Carmelo
Departamento de Matemática Aplicada &
IUMA,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain.
clavero@unizar.es

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

CRESSON, Jacky
Laboratoire de Mathématiques et de 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.f

CUESTA, Eduardo
Departamento de Matemática Aplicada,
E.T.S.I. de Telecomunicación,
Campus Miguel Delibes,
Universidad de Valladolid,
Paseo Belén 15,
47011 Valladolid, Spain.
eduardo@mat.uva.es

DENA, Ángeles
Centro Universitario de la Defensa,
Academia General Militar,
Ctra. de Huesca s/n,
50090 Zaragoza, Spain.
adena@unizar.es

DRÁBEK, Pavel
Department of Mathematics and NTIS,
University of West Bohemia in Pilsen,
FAV ZCU Univerzitni 8,
306 14 Pilsen, Czech Republic.
pdrabek@kma.zcu.cz

DUBOIS, François
Département de Mathématiques,
Bâtiment 307,
Faculté des Sciences d'Orsay,
Université Paris-Sud,
F-91405 Orsay Cedex, France.
francois.dubois@u-psud.fr

EL SAADI, Nadjia
LAMOPS, ENSSEA,
Algiers, Algeria.
enadjia@gmail.com

ELIPE, Antonio
Grupo de Mecánica Espacial & IUMA,
Centro Universitario de la Defensa,
Academia General Militar,
Ctra. de Huesca s/n,
50090 Zaragoza, Spain.
elipe@unizar.es

ETANCELIN, JEAN–MATTHIEU
Laboratoire de Mathématiques et de leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
jean-matthieu.etancelin@univ-pau.fr

FERNÁNDEZ–PATO, JAVIER
Computational Hydraulics Group,
Escuela de Ingeniería y Arquitectura,
Universidad de Zaragoza,
Campus Río Ebro,
c/ María de Luna 3,
50018 Zaragoza, Spain.
jfpato@unizar.es

FERREIRA, CHELO
Departamento de Matemática Aplicada & IUMA,
Facultad de Veterinaria,
Universidad de Zaragoza,
c/ Miguel Servet 117,
50013 Zaragoza, Spain.
cferrei@unizar.es

FLORENS, VINCENT
Laboratoire de Mathématiques et de 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

FLORÍA, LUIS
Departamento de Física Teórica, GME & IUMA,
Facultad de Ciencias, Edificio B,
Universidad de Zaragoza,
c/ Pedro Cerbuna 12,
50009 Zaragoza, Spain.
lfloria@unizar.es

GALLOUET, THIERRY
Université de Marseille,
CMI, 39 rue Joliot Curie,
F-13453 Marseille Cedex 13, France.
thierry@gallouet.fr

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

GHOSH, AMRITA
Laboratoire de Mathématiques et de leurs Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
amrita.ghosh@univ-pau.fr

GÓMEZ, MARTA
Escuela Politécnica Superior,
Universidad San Jorge,
Campus Universitario Villanueva de Gállego,
Autov. A-23 Zaragoza-Huesca, km. 299,
50830 Villanueva de Gállego, Zaragoza, Spain.
mgunizar@gmail.com

GONZÁLEZ TABOADA, MARÍA
Departamento de Matemáticas,
Universidade da Coruña,
Facultad de Informática,
Campus de Elviña, s/n,
15071 A Coruña, Spain.
maria.gonzalez.taboada@udc.es

GORDILLO, GEOVANNY
Departamento de Mecánica de Fluidos,
Universidad de Zaragoza,
LIFTEC-CSIC,
Escuela de Ingeniería y Arquitectura,
Universidad de Zaragoza,
Campus Río Ebro,
c/ María de Luna 3,
50018 Zaragoza, Spain.
ggordillo@unizar.es

GOSSEZ, JEAN-PIERRE
Département de Mathématique, CP214,
Université Libre de Bruxelles,
1050 Bruselles, Belgium.
gossez@ulb.ac.be

GRACIA, José Luis
Departamento de Matemática Aplicada &
IUMA,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain.
jlgracia@unizar.es

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

HABIBI, Noora
Faculty of Mathematical Sciences,
Shahrood University of Technology,
Shahrood, Semnan, Iran.
habibi85nu@gmail.com

HAUKE, Guillermo
Área de Mecánica de Fluidos,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain. ghauke@unizar.es

HERBIN, Raphaèle
Institut de Mathématiques de Marseille,
Université d'Aix–Marseille,
39 rue Joliot Curie,
13453 Marseille 13, France.
raphaele.herbin@univ-amu.fr

HERNÁNDEZ, Jesús
Instituto de Matemática Interdisciplinar,
Facultad de Ciencias Matemáticas,
Universidad Complutense de Madrid,
Plaza de Ciencias 3, Ciudad Universitaria,
28040 Madrid, Spain.
jesus.hernande@telefonica.net

HUME, Laurène
Laboratoire de Mathématiques et de leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
laurene.hume@univ-pau.fr

JÓDAR, Joaquín
Departamento de Matemáticas,
Universidad de Jaén,
Edificio B–3,
Campus de las Lagunillas.
23071 Jaén, Spain.
jjodar@ujaen.es

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

JORGE, Juan Carlos
Departamento de Estadística, Informática y
Matemáticas,
Universidad Pública de Navarra,
Campus de Arrosadía,
31006 Pamplona, Spain.
jcjorge@unavarra.es

KOLB, Sébastien
CReA, French Air Force Research Centre,
BA 701,
13661 Salon Air, France.
sebastien.kolb@ecole-air.fr

KUMAR, Prashant
CWI Amsterdam,
Science Park 123,
1098 XG Amsterdam, Netherlands.
pkumar@cwi.nl

LANCHARES, Víctor
Departamento de Matemáticas y Computación,
Universidad de La Rioja,
Edificio Científico Tecnológico - CCT,
c/ Madre de Dios 53,
26006 Logroño, Spain.
vlancha@unirioja.es

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

MALO, Beatriz
Universidad de Zaragoza,
Zaragoza, Spain.
bea.malo.91@gmail.com

MARIŞ, Mihai
Institut de Mathématiques de Toulouse,
Université Paul Sabatier,
118 Route de Narbonne,
31062 Toulouse, France.
mihai.maris@math.univ-toulouse.fr

MARTÍNEZ ARANDA, Sergio
Departamento de Ciencia y Tecnología de Materiales y Fluidos,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain. sermar@unizar.es

MASSOPUST, Peter
Technical University of Munich,
Center of Mathematics,
Boltzmannstrasse 3,
85748 Garching b. Munich, Germany.
massopust@ma.tum.de

MERCIER, Sophie
Laboratoire de Mathématiques et de 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

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

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

NAZAR, Mudassar
School of Mathematical Sciences, University of Science and Technology of China,
Hefei, Anhui, China &
Centre for Advanced Studies in Pure and Applied Mathematics,
Bahauddin Zakariya University,
Multan, Pakistan.
mudassar_666@yahoo.com

NOVO, Julia
Departamento de Matemáticas,
Universidad Autónoma de Madrid,
Campus de Cantoblanco,
28049 Madrid, Spain.
julia.novo@uam.es

OMNES, Pascal
CEA Saclay,
DM2S–STMF, Bât 451,
91191 Gif-sur-Yvette Cedex, France.
pascal.omnes@cea.fr

O'RIORDAN, Eugene
School of Mathematical Sciences,
Dublin City University,
Dublin 9, Ireland.
eugene.oriordan@dcu.ie

OYEBOLA, Oyeyemi Oluwaseyi
Department of Mathematics,
Federal University of Agriculture Abeokuta,
Ogun State, Nigeria.
oyebolaoo@funaab.edu.ng

PAGOLA, Pedro
Departamento de Estadística, Informática y Matemáticas,
Universidad Pública de Navarra,
Campus de Arrosadía,
31006 Pamplona, Spain.
pedro.pagola@unavarra.es

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

PALACIOS, Pablo
Universidad de Zaragoza,
Zaragoza, Spain.
681090@celes.unizar.es

PARRA, María Cruz
Departamento de Matemática Aplicada,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain.
cparra@unizar.es

PASADAS, Miguel
Departamento de Matemática Aplicada,
ETSI Caminos Canales y Puertos,
Universidad de Granada,
Campus Fuentenueva,
Avda. Severo Ochoa s/n,
18071 Granada, Spain.
mpasadas@ugr.es

PE, Álvaro
Departamento de Matemática Aplicada,
Universidad de Zaragoza,
Zaragoza, Spain.
apedelariva@gmail.com

PÉREZ, Ester
Departamento de Matemática Aplicada &
IUMA,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain.
ester.perez@unizar.es

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

PONCET, Philippe
Laboratoire de Mathématiques et de leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
philippe.poncet@univ-pau.fr

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

RADU, Ignat
Institut de Mathématiques de Toulouse,
Université Paul Sabatier,
bât. 1R3,
118, route de Narbonne,
31062 Toulouse, France
radu.ignat@math.univ-toulouse.fr

RAKOTOSON, Jean Michel
Laboratoire de Mathématiques et Applications,
UFR Sciences fondamentales et appliquées,
Université de Poitiers,
15, rue de l'Hôtel Dieu,
TSA 71117,
86073 POITIERS Cedex 9, France.
jean.michel.rakotoson@univpoitiers.fr

RIAGUAS, Andrés
Departamento de Matemática Aplicada,
Universidad de Valladolid,
Campus "Duques de Soria",
42004 Soria, Spain.
andresrg@mac.uva.es

RODRIGO, Carmen
Departamento de Matemática Aplicada &
IUMA,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain.
carmenr@unizar.es

RODRÍGUEZ, Miguel L.
Departamento de Matemática Aplicada,
Facultad de Ciencias,
Universidad de Granada,
Campus Fuentenueva,
18071 Granada miguelrg@ugr.es

RUBIO, Beatriz
Departamento de Matemática Aplicada &
IUMA,
EINA, Universidad de Zaragoza,
Edificio Torres Quevedo,
c/ María de Luna 3,
50018 Zaragoza, Spain.
brubio@unizar.es

SALLES, Gabriel
Laboratoire de Mathématiques et de leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
g.salles@etud.univ-pau.fr

SÁNDEZ, Carlos
Universidad de Zaragoza,
Zaragoza, Spain.
carsandez@gmail.com

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

SAUER, Tomas
Lehrstuhl für Mathematik mit
Schwerpunkt Digitale Bildverarbeitung,
Universität Passau,
Innstr. 43,
94032 Passau, Germany.
tomas.sauer@uni-passau.de

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

SEBASTIÁN, María Victoria
Centro Universitario de la Defensa,
Academia General Militar,
Ctra. de Huesca s/n,
50090 Zaragoza, Spain.
msebasti@unizar.es

SHIRANE, Taketo
Tokushima University,
2-1 Minamijousanjima-cho,
Tokushima-shi,
Tokushima-ken 770-8501, Japan.
shirane@tokushima-u.ac.jp

STYNES, Martin
Division of Applied and Computational
Mathematics,
Beijing Computational Science Research Center,
China.
m.stynes@csrc.ac.cn

TAHRAOUI, Yassine
Laboratoire de Mathématiques et de leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
tahraouiyacine@yahoo.fr

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

TINTAREV, Cyril
University of Upsala,
Uppsala, Sweden.
tammouz@gmail.com

TOKUNAGA, Hiroo
Department of Mathematical Sciences,
Graduate School of Sciences,
Tokyo Metropolitan University,
Minamiohsawa 1-1,
Hachoji Tokyo 192-0397, Japan.
tokunaga@tmu.ac.jp

TORREA, José Luis
Departamento de Matemáticas,
Facultad de Ciencias,
Universidad Autónoma de Madrid,
Ciudad Universitaria de Cantoblanco,
28049 Madrid, Spain.
joseluis.torrea@uam.es

TRESACO, Eva
Centro Universitario de la Defensa,
Academia General Militar,
Ctra. de Huesca s/n,
50090 Zaragoza, Spain.
etresaco@unizar.es

TURPAULT, Rodolphe
Institut de Mathématiques de Bordeaux,
351 Cours de la libération,
33405 Talence Cedex, France.
rodolphe.turpault@u-bordeaux.fr

UZAL, José Manuel
Departamento de Estatística, Análise
Matemática e Optimización,
Facultade de Matemáticas,
Universidade de Santiago de Compostela,
Rúa Lope Gómez de Marzoa s/n, 15782
Santiago de Compostela, Spain.
josemanuel.uzal@rai.usc.es

VALLET, Guy
Laboratoire de Mathématiques et de 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

VILLACAMPA, Raquel
Centro Universitario de la Defensa & IUMA,
Academia General Militar,
Ctra. de Huesca s/n,
50090 Zaragoza, Spain.
raquelvg@unizar.es

WARNAULT, Guillaume
Laboratoire de Mathématiques et de leurs
Applications,
Université de Pau et des Pays de l'Adour,
IPRA - UMR CNRS 5142,
BP 1155, 64013 Pau Cedex, France.
guillaume.warnault@iniv-pau.fr

ZIMMERMANN, Aleksandra
Fakultät für Mathematik,
Universität Duisburg-Essen,
Campus Essen,
45117 Essen, Germany.
aleksandra.zimmermann@uni-due.de

Other communications

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.

Optimization methods for computing periodic orbits
A. Abad

A finite volume method for fully coupled multiphase flow and chemical processes in porous media–application to CO_2 storage
É. Ahusborde, M. El Ossmani and M. Id Moulay

Walls in a junction of three ferromagnetic nanowires
A. K. Al Sayed, G. Carbou and S. Labbé

Semigroup theory for the Stokes operator with Navier boundary condition on L^p spaces
C. Amrouche, M. Escobedo and A. Ghosh

Elliptic problems in smooth and non smooth domains
C. Amrouche, M. Moussaoui and H. H. Nguyen

Triangular curves
E. Artal

Approximations of diffusions with small noise near a critical point
F. Avram

Optimal replacement policy under a general failure and repair model: minimal versus worse than old repair
F. G. Badía, M.D. Berrade, J. H. Cha and H. Lee

Stochastic comparisons and multivariate dependence for the epoch times of trend renewal processes
F. G. Badía, C. Sangüesa and J. H. Cha

Quasi–interpolation and applications to PDEs with radial basis functions
M. D. Buhmann

Multidimensional discrete PDE splines using radial basis functions
M. D. Buhmann, J. Jódar and M. L. Rodríguez

An evolutionary algorithm for the multi–period facility location problem
H. I. Calvete, P. M. Mateo and C. Sández

A free energy diminishing DDFV scheme for convection–diffusion equations
C. Cancès, C. Chainais–Hillairet and S. Krell

Walker regime for walls in ferromagnetic nanotubes
G. Carbou

An efficient uniformly convergent method for solving singularly perturbed semilinear reaction–diffusion systems
C. Clavero and J. C. Jorge

CoPO, the corrector of periodic orbits algorithm with high precision
A. Dena

Travelling waves in the Fisher–KPP equation with nonlinear diffusion and a non–Lipschitzian reaction term
P. Drábek and P. Takáč

Raviart–Thomas finite elements of Petrov–Galerkin type
F. Dubois, I. Greff and Ch. Pierre

On the existence of solutions for a nonlinear stochastic partial differential equation arising as a model of phytoplankton aggregation
N. El Saadi and Z. Benbaziz

Symmetric periodic orbits in a four–body problem
A. Elipe

Reactive flows at pore scale with hybrid computing
J.-M. Etancelin and Ph. Poncet

On the application of novel 2D techniques to model streamflow generation in response to rainfall
J. Fernández-Pato, J. L. Gracia and P. García-Navarro

Asymptotic behaviour of the swallowtail catastrophe
C. Ferreira, J. L. López and E. Pérez Sinusía

Two–point Taylor expansions in singular one-dimensional boundary value problems: application to the spheroidal wave equation
C. Ferreira, J. L. López and E. Pérez Sinusía

Slopes of colored links
V. Florens

Canonical Constants in a problem of Radzievskij
L. Floría

Filling holes using a mesh of filled curves
M. A. Fortes, P. González, A. Palomares and M. L. Rodríguez

Error analysis of non inf–sup stable discretizations of the time–dependent Navier–Stokes equations with local projection stabilization
J. de Frutos, B. García–Archilla, V. John and J. Novo

Development of a control tool for releases of pollutants in rivers
G. Gordillo, M. Morales–Henández and P. García–Navarro

Elliptic problems involving a gradient term with natural growth
J. P. Gossez

Multigrid Waveform relaxation based on finite element discretisation
N. Habibi

Flat and compact support solutions to some semilinear elliptic problems with non–Lipschitz nonlinearities
J. Hernández

Vortex–based penalized method for permeability estimation of real samples
L. Hume and Ph. Poncet

On the uniqueness of minimisers of Ginzburg–Landau functionals
R. Ignat

On a bounded distribution derived from the shifted Gompertz law
P. Jodrá

A multigrid multilevel Monte Carlo method for transport in the Darcy–Stokes system
P. Kumar

Uniformly convergent expansions of the Struve functions in terms of elementary functions
J.L. López, P.J. Pagola and P. Palacios

On some minimization problems in $\mathbb{R}^N$
M. Maris

Equilibrium and non-equilibrium models applied to unsteady sediment transport
S. Martínez–Aranda, J. Murillo and P. García–Navarro

Rate of numerical difusion of finite volume schemes
P. Omnes

Numerical analysis and thin layers
E. O'Riordan

Uniformly convergent expansions of the generalized hypergeometric function pFq in terms of elementary functions
P. J. Pagola and J. L. López

Filling holes of generalized offset surfaces by biquadratic splines
M. Pasadas

Geometric multilevel methods for isogeometric analysis
A. Pe de la Riva, C. Rodrigo and F. J. Gaspar

Statistical splicing of economic series by smoothing quadratic splines
L. Pedauga, E. Delgado—Márquez, M. Márquez and M. Pasadas

Reactive flows at the pore-scale of porous materials
Ph. Poncet

Numerical continuation of one–parameter families of periodic orbits
A. Riaguas and E. Tresaco

Parametric inference for two imperfect repair models for gamma deteriorating systems
G. Salles, S. Mercier and L. Bordes

On compactness properties and ground states of an affine Laplacian
I. Schindler

Galois covers of graphs and embedded topology of plane curves
T. Shirane

The origin of the p–Laplacian and A. Missbach
P. Takáč, J. Benedikt, P. Girg and L. Kotrla

Topology of plane curves and "arithmetic" of double covers of $\mathbb{P}^2$
H. Tokunaga

A domain decomposition strategy for a very high–order finite volumes scheme applied to cardiac electrophysiology
R. Turpault and Y. Coudière

On a nonlocal stochastic PDE
G. Vallet and U. Koley

Monografías Matemáticas *García de Galdeano* **42**, 1–11 (2019)

GENERALIZED RESOLVENT OF THE STOKES PROBLEM WITH NAVIER-TYPE BOUNDARY CONDITIONS

Hind Al Baba and Antonia Jabbour

Abstract. We study in this paper the generalized resolvent of the Stokes problem with Navier-type boundary conditions.

Keywords: Generalized resolvent, Stokes Problem, Navier-type boundary conditions.
AMS classification: 35Q30, 76D05, 76D07, 35K20, 35K22, 76N10, 35A20, 35Q40.

§1. Introduction

This paper is devoted to the existence and uniqueness of weak and strong and very weak solutions to the problem

$$\begin{cases} \lambda \boldsymbol{u} - \Delta \boldsymbol{u} + \nabla \pi = \boldsymbol{f}, & \operatorname{div} \boldsymbol{u} = \chi & \text{in} & \Omega \times (0,T), \\ \boldsymbol{u} \cdot \boldsymbol{n} = g, & \mathbf{curl}\, \boldsymbol{u} \times \boldsymbol{n} = \boldsymbol{h} \times \boldsymbol{n} & \text{on} & \Gamma \times (0,T), \end{cases} \tag{1}$$

where we study the generalized resolvent of the Stokes operator with nonstandard Navier-type boundary conditions. Up to now most research concerns the homogeneous boundary conditions, and the case χ=0. Although the case $\chi \neq 0$ has many important applications, specially in treating more general boundary value problems and using cut-off procedure.

There exists several references on (1) when $\chi = 0$ in Ω. This question was already studied by Solonnikov in [12] for the homogeneous Dirichlet boundary condition (*i.e.* $\boldsymbol{u} = \mathbf{0}$ on Γ). In that work, the author considered the resolvent Problem when $|\arg \lambda| \leq \delta + \pi/2$ where $\delta \geq 0$ is small. Later on, the resolvent of the Stokes operator with Dirichlet boundary condition in bounded domains has been studied by Giga in [6] using the theory of pseudo-differential operators. The results in [6] extends those in [12] in two directions. First, he consider larger set of values of λ. More precisely λ in the sector $|\arg \lambda| \leq \pi - \varepsilon$, for any $\varepsilon > 0$. Second, the resolvent of the Stokes operator is obtained explicitly and this enables him to describe the domains of fractional powers of the Stokes operator with Dirichlet boundary condition.

In exterior domains, Giga and Sohr [7] approximate the resolvent of the Stokes operator with Dirichlet boundary condition with the resolvent of the Stokes operator in the entire space.

Farwig and Sohr [5] investigate the Problem (1) when $\operatorname{div} \boldsymbol{u} \neq 0$ in Ω and $\boldsymbol{u} = \mathbf{0}$ on Γ. Their results include bounded and unbounded domains, for the whole and the half space the proof relies on multiplier technique. The problem is also investigated for bended half spaces and for cones by using perturbation criterion and referring to the half space problem.

The Problem (1) is also studied with Robin boundary conditions by Saal [10], Shibata and Shimada [11]. In [10], Saal proves that the Stokes operator with homogeneous Robin

boundary conditions is sectorial and admits an $\boldsymbol{H}^\infty$-calculus on L^p-spaces. Shibata and Shimada proved in [11] a generalized resolvent estimate for the Stokes equations with non-homogeneous Robin boundary conditions and divergence condition in $\boldsymbol{L}^p$-framework in a bounded or exterior domain by extending the argument of Farwig and Shor [5].

Concerning the Navier-type boundary conditions, Miyakawa [9] shows that the Laplacian operator with homogeneous Navier-type boundary conditions generates a holomorphic semi-group on L^p-spaces when the domain Ω is of class C^∞. Mitrea and Monniaux [8] consider the resolvent of the Stokes operator with homogeneous Navier-type boundary conditions in Lipschitz domains using differential forms on Lipschitz sub-domains of a smooth compact Riemannian manifold. In [1] and [2] Al Baba et al. consider the Problem (1) when $\chi = 0$ in Ω and $g = 0$, $\boldsymbol{h} = \boldsymbol{0}$ on Γ and prove the existence of weak, strong and very weak solutions to this problem.

This paper is organized as follows. In Section 2 we give the functional framework and some preliminary results at the basis of our proofs. In Section 3 we prove our main results on the existence of weak, strong and very weak solutions to Problem (1).

§2. Preliminaries

In this subsection we review some basic notations, definitions and functional framework which are essential in our work.

In what follows, if we do not state otherwise, Ω will be considered as an open bounded domain of $\mathbb{R}^3$ of class $C^{2,1}$. Then a unit normal vector to the boundary can be defined almost everywhere it will be denoted by $\boldsymbol{n}$, $\boldsymbol{n}$ is defined everywhere because $\boldsymbol{n}$ is $C^{1,1}$. The generic point in Ω is denoted by $\boldsymbol{x} = (x_1, x_2, x_3)$. The domain Ω is not necessarily simply-connected and the boundary Γ is not necessarily connected.

Let us introduce some functional spaces.
Let $\boldsymbol{L}^p(\Omega)$ denote the usual vector valued $\boldsymbol{L}^p$-space over Ω. Let us define the spaces:

$$\begin{aligned}
\boldsymbol{H}^p(\mathbf{curl}, \Omega) &= \{\boldsymbol{v} \in \boldsymbol{L}^p(\Omega);\ \mathbf{curl}\, \boldsymbol{v} \in \boldsymbol{L}^p(\Omega)\},\\
\boldsymbol{H}^p(\mathrm{div}, \Omega) &= \{\boldsymbol{v} \in \boldsymbol{L}^p(\Omega);\ \mathrm{div}\ \boldsymbol{v} \in \boldsymbol{L}^p(\Omega)\},\\
\boldsymbol{X}^p(\Omega) &= \boldsymbol{H}^p(\mathbf{curl}, \Omega) \cap \boldsymbol{H}^p(\mathrm{div}, \Omega),
\end{aligned}$$

equipped with their graph norms. Thanks to [4] and [3] we know that $D(\overline{\Omega})$ is dense in $\boldsymbol{H}^p(\mathbf{curl}, \Omega)$, $\boldsymbol{H}^p(\mathrm{div}, \Omega)$ and $\boldsymbol{X}^p(\Omega)$. We also define the subspaces:

$$\begin{aligned}
\boldsymbol{H}^p_0(\mathbf{curl}, \Omega) &= \{\boldsymbol{v} \in \boldsymbol{H}^p(\mathbf{curl}, \Omega);\quad \boldsymbol{v} \times \boldsymbol{n} = \boldsymbol{0} \text{ on } \Gamma\},\\
\boldsymbol{H}^p_0(\mathrm{div}, \Omega) &= \{\boldsymbol{v} \in \boldsymbol{H}^p(\mathrm{div}, \Omega);\quad \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma\},\\
\boldsymbol{X}^p_N(\Omega) &= \{\boldsymbol{v} \in \boldsymbol{X}^p(\Omega);\quad \boldsymbol{v} \times \boldsymbol{n} = \boldsymbol{0} \text{ on } \Gamma\},\\
\boldsymbol{X}^p_T(\Omega) &= \{\boldsymbol{v} \in \boldsymbol{X}^p(\Omega);\quad \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma\}.
\end{aligned}$$

We recall that for all function $\boldsymbol{v} \in \boldsymbol{H}^p(\mathbf{curl}, \Omega)$ (respectively $\boldsymbol{v} \in \boldsymbol{H}^p(\mathrm{div}, \Omega)$), the tangential trace $\boldsymbol{v} \times \boldsymbol{n}$ (respectively the normal trace $\boldsymbol{v} \cdot \boldsymbol{n}$) exists and belongs to $\boldsymbol{W}^{-1/p,p}(\Gamma)$ (respectively to $W^{-1/p,p}(\Gamma)$). Thanks to [4] we know that $D(\Omega)$ is dense in $\boldsymbol{H}^p_0(\mathbf{curl}, \Omega)$ and in $\boldsymbol{H}^p_0(\mathrm{div}, \Omega)$. Finally we denote by $[\boldsymbol{H}^p_0(\mathbf{curl}, \Omega)]'$ and $[\boldsymbol{H}^p_0(\mathrm{div}, \Omega)]'$ the dual spaces of $\boldsymbol{H}^p_0(\mathbf{curl}, \Omega)$ and $\boldsymbol{H}^p_0(\mathrm{div}, \Omega)$ respectively.

Next, we review some known results which are essential in our work. First, we recall that the vector-valued Laplace operator of a vector field $\boldsymbol{v}= (v_1, v_2, v_3)$ is equivalently defined by

$$\Delta\, \boldsymbol{v} = \mathbf{grad}\ (\mathrm{div}\ \boldsymbol{v})\ \text{-}\ \mathbf{curl}\ \mathbf{curl}\ \boldsymbol{v}.$$

We have the following lemmas [4]

Lemma 1. *The spaces $\boldsymbol{X}^p_N(\Omega)$ and $\boldsymbol{X}^p_T(\Omega)$ defined above are continuously embedded in $\boldsymbol{W}^{1,p}(\Omega)$.*

In order to consider the case of nonhomogeneous boundary conditions, we introduce the following spaces:

$$\boldsymbol{X}^{1,p}(\Omega) = \{\boldsymbol{v} \in \boldsymbol{L}^p(\Omega);\ \mathrm{div}\boldsymbol{v} \in L^p(\Omega),\ \mathbf{curl}\,\boldsymbol{v} \in \boldsymbol{L}^p(\Omega)\,\text{and}\,\boldsymbol{v}\cdot\boldsymbol{n} \in W^{1-1/p,p}(\Gamma)\},$$

$$\boldsymbol{Y}^{1,p}(\Omega) = \{\boldsymbol{v} \in \boldsymbol{L}^p(\Omega);\ \mathrm{div}\boldsymbol{v} \in L^p(\Omega), \mathbf{curl}\,\boldsymbol{v} \in \boldsymbol{L}^p(\Omega)\,\text{and}\,\boldsymbol{v}\times\boldsymbol{n} \in W^{1-1/p,p}(\Gamma)\}.$$

Lemma 2. *The spaces $\boldsymbol{X}^{1,p}(\Omega)$ and $\boldsymbol{Y}^{1,p}(\Omega)$ are continuously embedded in $\boldsymbol{W}^{1,p}(\Omega)$.*

Consider as well the spaces:

$$\boldsymbol{X}^{2,p}(\Omega) = \{\boldsymbol{v} \in \boldsymbol{L}^p(\Omega);\ \mathrm{div}\boldsymbol{v} \in W^{1,p}(\Omega), \mathbf{curl}\,\boldsymbol{v} \in \boldsymbol{W}^{1,p}(\Omega)\,\text{and}\,\boldsymbol{v}\cdot\boldsymbol{n} \in W^{2-1/p,p}(\Gamma)\},$$

$$\boldsymbol{Y}^{2,p}(\Omega) = \{\boldsymbol{v} \in \boldsymbol{L}^p(\Omega);\ \mathrm{div}\boldsymbol{v} \in W^{1,p}(\Omega), \mathbf{curl}\,\boldsymbol{v} \in \boldsymbol{W}^{1,p}(\Omega)\,\text{and}\,\boldsymbol{v}\times\boldsymbol{n} \in W^{2-1/p,p}(\Gamma)\}.$$

Theorem 3. *Assume that Ω is of class $C^{2,1}$, then the spaces $\boldsymbol{X}^{2,p}(\Omega)$ and $\boldsymbol{Y}^{2,p}(\Omega)$ are continuously embedded in $\boldsymbol{W}^{2,p}(\Omega)$.*

Consider now the space

$$\boldsymbol{E}^p(\Omega) = \{\boldsymbol{v} \in \boldsymbol{W}^{1,p}(\Omega);\ \ \Delta\boldsymbol{v} \in [\boldsymbol{H}^{p'}_0(\mathrm{div},\Omega)]'\},$$

which is a Banach space for the norm $\|\boldsymbol{v}\|_{\boldsymbol{E}^p(\Omega)} = \|\boldsymbol{v}\|_{\boldsymbol{W}^{1,p}(\Omega)} + \|\Delta\boldsymbol{v}\|_{[\boldsymbol{H}^{p'}_o(\mathrm{div},\Omega)]'}$. Thanks to [3, Lemma 4.1] we know that $D(\overline{\Omega})$ is dense in $\boldsymbol{E}^p(\Omega)$. Moreover, (see [3, Corollary 4.2]), the linear mapping $\gamma : \boldsymbol{v} \longmapsto \mathbf{curl}\boldsymbol{v}\times\boldsymbol{n}$ defined on $D(\overline{\Omega})$ can be extended to a linear and continuous mapping $\gamma : \boldsymbol{E}^p(\Omega) \longmapsto \boldsymbol{W}^{-1/p,p}(\Omega)$. Moreover, we have the Green formula: for any $\boldsymbol{v} \in \boldsymbol{E}^p(\Omega)$ and $\boldsymbol{\varphi} \in \boldsymbol{X}^{p'}_\tau(\Omega)$ such that div $\boldsymbol{\varphi}$= 0 in Ω,

$$-\langle\Delta\boldsymbol{v},\boldsymbol{\varphi}\rangle_{[\boldsymbol{H}^{p'}_0(\mathrm{div},\Omega)]'\times\boldsymbol{H}^{p'}_0(\mathrm{div},\Omega)} = \int_\Omega \mathbf{curl}\,\boldsymbol{v}\cdot\mathbf{curl}\,\overline{\boldsymbol{\varphi}}\ \mathrm{d}x - \langle\mathbf{curl}\,\boldsymbol{v}\times\boldsymbol{n},\boldsymbol{\varphi}\rangle_\Gamma,$$

where $\langle\cdot,\cdot\rangle_\Gamma = \langle\cdot,\cdot\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma)\times\boldsymbol{W}^{1/p,p'}(\Gamma)}$.

Next, we introduce the following space

$$\boldsymbol{T}^p(\Omega) = \{\boldsymbol{\phi} \in \boldsymbol{H}^p_0(\mathrm{div},\Omega);\ \mathrm{div}\boldsymbol{\phi} \in W^{1,p}_0(\Omega)\}.$$

The space $\mathcal{D}(\Omega)$ is dense in $\boldsymbol{T}^p(\Omega)$ and for all $\chi \in W^{-1,p}(\Omega)$ and $\boldsymbol{\phi} \in \boldsymbol{T}^{p'}(\Omega)$, we have:

$$\langle\boldsymbol{\nabla}\chi,\boldsymbol{\phi}\rangle_{(\boldsymbol{T}^{p'}(\Omega))'\times\boldsymbol{T}^{p'}(\Omega)} = -\langle\chi,\mathrm{div}\boldsymbol{\phi}\rangle_{W^{-1,p}(\Omega)\times W^{1,p'}_0(\Omega)}. \tag{2}$$

A distribution $\boldsymbol{f}$ belongs to $(\boldsymbol{T}^p(\Omega))'$ if and only if there exist $\boldsymbol{\psi} \in \boldsymbol{L}^{p'}(\Omega)$ and $f_0 \in W^{-1,p'}(\Omega)$, such that $\boldsymbol{f} = \boldsymbol{\psi} + \boldsymbol{\nabla} f_0$. Moreover, we have the estimate

$$\|\boldsymbol{\psi}\|_{\boldsymbol{L}^{p'}(\Omega)} + \|f_0\|_{W^{-1,p'}(\Omega)} \le C\|\boldsymbol{f}\|_{(\boldsymbol{T}^p(\Omega))'}.$$

We will need also the following space

$$\boldsymbol{H}_p(\Delta;\Omega) = \{\boldsymbol{v} \in \boldsymbol{L}^p(\Omega);\ \Delta\boldsymbol{v} \in (\boldsymbol{T}^{p'}(\Omega))'\},$$

which is a Banach space for the norm $\|\boldsymbol{v}\|_{\boldsymbol{H}_p(\Delta;\Omega)} = \|\boldsymbol{v}\|_{\boldsymbol{L}^p(\Omega)} + \|\Delta\boldsymbol{v}\|_{(\boldsymbol{T}^{p'}(\Omega))'}$. The space $\mathcal{D}(\overline{\Omega})$ is dense in $\boldsymbol{H}_p(\Delta;\Omega)$ and The mapping γ: $\boldsymbol{v} \longmapsto \mathbf{curl}\,\boldsymbol{v} \times \boldsymbol{n}$ defined on $D(\overline{\Omega})$ can be extended by continuity to a linear and continuous mapping $\gamma : \boldsymbol{H}_p(\Delta;\Omega) \longmapsto \boldsymbol{W}^{-1-1/p,p}(\Omega)$. Moreover, we have the Green formula: for any $\boldsymbol{v} \in \boldsymbol{H}_p(\Delta;\Omega)$ and $\boldsymbol{\phi} \in \boldsymbol{Y}_\tau^{p'}(\Omega)$,

$$\langle \Delta\boldsymbol{v}, \boldsymbol{\phi}\rangle_{(\boldsymbol{T}^{p'}(\Omega))'\times\boldsymbol{T}^{p'}(\Omega)} = \int_\Omega \boldsymbol{v}\cdot\Delta\overline{\boldsymbol{\phi}}\,\mathrm{d}x + \langle \mathbf{curl}\boldsymbol{v}\times\boldsymbol{n}, \boldsymbol{\phi}\rangle_\Gamma, \tag{3}$$

where $\langle\cdot,\cdot\rangle_\Gamma = \langle\cdot,\cdot\rangle_{\boldsymbol{W}^{-1-1/p,p}(\Gamma)\times\boldsymbol{W}^{1+1/p,p'}(\Gamma)}$ and

$$\boldsymbol{Y}_\tau^p(\Omega) = \{\boldsymbol{\phi} \in \boldsymbol{W}^{2,p}(\Omega);\ \boldsymbol{\phi}\cdot n = 0,\ \mathrm{div}\boldsymbol{\phi} = 0,\ \mathbf{curl}\boldsymbol{\phi}\times n = 0 \text{ on } \Gamma\}.$$

§3. Generalized resolvent problem

In this section we consider the generalized resolvent Problem (1) and we prove the existence and uniqueness of weak, strong and very weak solution to this problem.

3.1. Weak solution

Consider the problem

$$\begin{cases} \lambda\boldsymbol{u} - \Delta\boldsymbol{u} + \nabla\pi = \boldsymbol{f}, & \mathrm{div}\,\boldsymbol{u} = 0 & \text{in} & \Omega\times(0,T),\\ \boldsymbol{u}\cdot\boldsymbol{n} = 0, & \mathbf{curl}\,\boldsymbol{u}\times\boldsymbol{n} = \boldsymbol{h}\times\boldsymbol{n} & \text{on} & \Gamma\times(0,T), \end{cases} \tag{4}$$

We start by the existence and uniqueness of weak solution to (4).

Theorem 4. *Let $\varepsilon \in]0,\pi[$ be fixed and $\lambda \in \Sigma_\varepsilon$. Let $p \geq 2$, $\boldsymbol{f} \in (\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'$ and $\boldsymbol{h}\times\boldsymbol{n} \in \boldsymbol{W}^{-1/p,p}(\Gamma)$. Then the problem* (4) *has a unique solution $(\boldsymbol{u},\pi) \in \boldsymbol{W}^{1,p}(\Omega)\times L^p(\Omega)/\mathbb{R}$ satisfying the following estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{1,p}(\Omega)} \leq C(\Omega,p)\left(\|\boldsymbol{f}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'} + \|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)}\right). \tag{5}$$

Proof. **Step 1 : Existence and uniqueness.** We can easily verify that problem (4) is equivalent to the variational problem: Find $\boldsymbol{u} \in \boldsymbol{V}_\tau^p(\Omega)$ such that for all $\boldsymbol{v} \in \boldsymbol{V}_\tau^{p'}(\Omega)$

$$\lambda\int_\Omega \boldsymbol{u}\cdot\overline{\boldsymbol{v}}\,\mathrm{d}x + \int_\Omega \mathbf{curl}\,\boldsymbol{u}\cdot\mathbf{curl}\,\overline{\boldsymbol{v}}\,\mathrm{d}x = \langle\boldsymbol{f},\boldsymbol{v}\rangle_\Omega + \langle\boldsymbol{h}\times\boldsymbol{n},\boldsymbol{v}\rangle_\Gamma, \tag{6}$$

where $\langle\cdot,\cdot\rangle_\Omega = \langle\cdot,\cdot\rangle_{[\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)]'\times\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)}$ and $\langle\cdot,\cdot\rangle_\Gamma = \langle\cdot,\cdot\rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma)\times\boldsymbol{W}^{-1/p,p'}(\Gamma)}$.

The proof is done in two steps:

i) Case $2 \le p \le 6$. The case $p = 2$ can be directly obtained using Lax-Milgram theorem. Suppose that $2 < p \le 6$, then Problem (4) has a unique solution $(\boldsymbol{u}, \pi) \in \boldsymbol{H}^1(\Omega) \times L^2(\Omega)/\mathbb{R}$. We write (4) in the form:

$$\left\{ \begin{array}{rrcl} -\Delta \boldsymbol{u} + \nabla \pi = \boldsymbol{f} - \lambda \boldsymbol{u} = \boldsymbol{F}, & \operatorname{div} \boldsymbol{u} = 0 & \text{in} & \Omega \\ \boldsymbol{u} \cdot \boldsymbol{n} = 0, & \mathbf{curl}\, \boldsymbol{u} \times \boldsymbol{n} = \boldsymbol{h} \times \boldsymbol{n} & \text{on} & \Gamma \end{array} \right. . \tag{7}$$

As $\boldsymbol{H}^1(\Omega) \hookrightarrow \boldsymbol{L}^p(\Omega)$, we have $\boldsymbol{F} \in (\boldsymbol{H}_0^{p'}(\operatorname{div}; \Omega))'$ and

$$\forall\, \boldsymbol{v} \in \boldsymbol{K}_\tau^{p'}(\Omega), \quad \langle \boldsymbol{F}, \boldsymbol{v} \rangle_\Omega + \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{v} \rangle_\Gamma = 0. \tag{8}$$

Theorem 4.4 of [3] implies that $\boldsymbol{u} \in \boldsymbol{W}^{1,p}(\Omega)$ and $\pi \in L^p(\Omega)$.

Let $\boldsymbol{v} \in \boldsymbol{K}_\tau^{p'}(\Omega)$, using the variational formulation we have

$$\langle \boldsymbol{F}, \boldsymbol{v} \rangle_\Omega + \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{v} \rangle_\Gamma = 0.$$

Then our solution $(\boldsymbol{u}, \pi)$ belongs to $\boldsymbol{W}^{1,p}(\Omega) \times L^p(\Omega)/\mathbb{R}$.

ii) Case $p \ge 6$. Observe that $(\boldsymbol{H}_0^{p'}(\operatorname{div}, \Omega))' \hookrightarrow (\boldsymbol{H}_0^{6/5}(\operatorname{div}, \Omega))'$ and $\boldsymbol{W}^{-1/p,p}(\Gamma) \hookrightarrow \boldsymbol{W}^{-1/6,6}(\Gamma)$. Then Problem (7) has a unique solution $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{1,6}(\Omega) \times L^6(\Omega)/\mathbb{R}$. Thanks to the embedding $\boldsymbol{W}^{1,6}(\Omega) \hookrightarrow \boldsymbol{L}^\infty(\Omega)$ we deduce that $\boldsymbol{F} = \boldsymbol{f} - \lambda \boldsymbol{u} \in (\boldsymbol{H}_0^{p'}(\operatorname{div}, \Omega))'$. Moreover, $\boldsymbol{F}$ satisfies the compatibility condition (8), then we conclude that $(\boldsymbol{u}, \pi)$ belongs to $\boldsymbol{W}^{1,p}(\Omega) \times L^p(\Omega)/\mathbb{R}$.

Step 2: Estimate. Let $B \in \mathcal{L}(\boldsymbol{V}_\tau^p(\Omega), (\boldsymbol{V}_\tau^{p'}(\Omega))')$ be the operator defined by

$$\forall\, \boldsymbol{u} \in \boldsymbol{V}_\tau^p(\Omega), \forall\, \boldsymbol{v} \in \boldsymbol{V}_\tau^{p'}(\Omega), \quad \langle B\boldsymbol{u}, \boldsymbol{v} \rangle_{(\boldsymbol{V}_\tau^{p'}(\Omega))' \times \boldsymbol{V}_\tau^p(\Omega)} = \lambda \int_\Omega \boldsymbol{u} \cdot \overline{\boldsymbol{v}}\, \mathrm{d}x + \int_\Omega \mathbf{curl}\, \boldsymbol{u} \cdot \mathbf{curl}\, \overline{\boldsymbol{v}}\, \mathrm{d}x.$$

For all $p \ge 2$, the operator B is an isomorphism from $\boldsymbol{V}_\tau^p(\Omega)$ into $(\boldsymbol{V}_\tau^{p'}(\Omega))'$ and $\|\boldsymbol{u}\|_{\boldsymbol{X}_\tau^p} \approx \|B\boldsymbol{u}\|_{(\boldsymbol{V}_\tau^{p'}(\Omega))'}$ for all $\boldsymbol{u} \in \boldsymbol{V}_\tau^p(\Omega)$. Moreover using the continuous embedding $\boldsymbol{X}_\tau^p(\Omega) \hookrightarrow \boldsymbol{W}^{1,p}(\Omega)$ we have for every $\boldsymbol{u} \in \boldsymbol{V}_\tau^p(\Omega)$ solution of problem (6),

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{1,p}(\Omega)} \le C(\Omega, p) \|\boldsymbol{u}\|_{\boldsymbol{X}_\tau^p(\Omega)} \le C(\Omega, p) \|B\boldsymbol{u}\|_{(\boldsymbol{V}_\tau^{p'}(\Omega))'}$$

and

$$\begin{aligned} \|B\boldsymbol{u}\|_{(\boldsymbol{V}_\tau^{p'}(\Omega))'} = \sup_{\substack{\boldsymbol{v} \in \boldsymbol{V}_\tau^{p'}(\Omega) \\ \boldsymbol{v} \neq 0}} \frac{|\langle B\boldsymbol{u}, \boldsymbol{v} \rangle|}{\|\boldsymbol{v}\|_{\boldsymbol{X}_\tau^{p'}(\Omega)}} &= \sup_{\substack{\boldsymbol{v} \in \boldsymbol{V}_\tau^{p'}(\Omega) \\ \boldsymbol{v} \neq 0}} \frac{|\langle \boldsymbol{f}, \boldsymbol{v} \rangle_\Omega + \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{v} \rangle_\Gamma|}{\|\boldsymbol{v}\|_{\boldsymbol{X}_\tau^{p'}(\Omega)}} \\ &\le C(\Omega, p) \left(\|\boldsymbol{f}\|_{(\boldsymbol{H}_0^{p'}(\operatorname{div}, \Omega))'} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)} \right), \end{aligned}$$

which is estimate (5). □

Theorem 5. *Let $\lambda \in \Sigma_\varepsilon$. Let $p \ge 2$. Let $\boldsymbol{f} \in (\boldsymbol{H}_0^{p'}(\operatorname{div}, \Omega))'$, $\boldsymbol{h} \times \boldsymbol{n} \in \boldsymbol{W}^{-1/p,p}(\Gamma)$, $g \in W^{1-1/p,p}(\Gamma)$ and $\chi \in L^p(\Omega)$ verifying the following compatibility condition*

$$\int_\Omega \chi\, \mathrm{d}x = \int_\Gamma g\, \mathrm{d}\sigma. \tag{9}$$

Then problem (1) *has a unique solution* $(\boldsymbol{u},\pi)\in \boldsymbol{W}^{1,p}(\Omega)\times L^p(\Omega)/\mathbb{R}$ *satisfying the following estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{1,p}(\Omega)}+\|\pi\|_{L^p(\Omega)/\mathbb{R}}\leq C(\Omega,p,\lambda)(\|\boldsymbol{f}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}+\|\chi\|_{L^p(\Omega)}+\|g\|_{W^{1-1/p,p}(\Gamma)}+\|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)}). \quad (10)$$

Proof. **i) Existence and uniqueness.** Consider the following Neumann problem

$$\Delta\boldsymbol{\theta}=\chi \quad \text{in} \quad \Omega \qquad \text{and} \qquad \frac{\partial\theta}{\partial n}=g \quad \text{on} \quad \Gamma. \quad (11)$$

Since $g\in W^{1-1/p,p}(\Gamma)$ and $\chi\in L^p(\Omega)$ verifying the compatibility condition (9) this problem has a unique solution $\theta\in W^{2,p}(\Omega)/\mathbb{R}$ such that

$$\|\theta\|_{W^{2,p}(\Omega)/\mathbb{R}}\leq C\left(\|g\|_{W^{1-1/p,p}(\Gamma)}+\|\chi\|_{L^p(\Omega)}\right). \quad (12)$$

Set $\boldsymbol{F}=\boldsymbol{f}-\lambda\nabla\theta+\nabla\chi$ and observe that $\boldsymbol{F}\in(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'$. Then using Theorem 4 we deduce that the problem

$$\begin{cases} \lambda z-\Delta z+\nabla\pi=\boldsymbol{F}, & \mathrm{div}\, z=0 \quad \text{in} \quad \Omega \\ z\cdot\boldsymbol{n}=0, & \mathbf{curl}\, z\times\boldsymbol{n}=\boldsymbol{h}\times\boldsymbol{n} \quad \text{on} \quad \Gamma \end{cases} \quad (13)$$

has a unique solution $(z,\pi)\in\boldsymbol{W}^{1,p}(\Omega)\times L^p(\Omega)/\mathbb{R}$ satisfying the following estimate

$$\|z\|_{\boldsymbol{W}^{1,p}(\Omega)}\leq C(\Omega,p)\left(\|\boldsymbol{F}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}+\|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)}\right). \quad (14)$$

Set $\boldsymbol{u}=z+\nabla\theta$. Then $(\boldsymbol{u},\pi)$ solve (1).

ii) Estimate. Observe that

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{1,p}(\Omega)}\leq C(\Omega,p)(\|\boldsymbol{f}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}+|\lambda|\|\nabla\theta\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}+\|\nabla\chi\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}+\|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)})+\|\nabla\theta\|_{\boldsymbol{W}^{1,p}(\Omega)}.$$

Then using estimate (12) one gets

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{1,p}(\Omega)}\leq C(\Omega,p,\lambda)(\|\boldsymbol{f}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}+\|\chi\|_{L^p(\Omega)}+\|g\|_{W^{1-1/p,p}(\Gamma)}+\|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)}). \quad (15)$$

Moreover $\|\pi\|_{L^p(\Omega)/\mathbb{R}}\leq C(\Omega,p)\,\|\nabla\pi\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}=\|\boldsymbol{f}-\lambda\boldsymbol{u}+\Delta\boldsymbol{u}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}$. Thus

$$\|\pi\|_{L^p(\Omega)/\mathbb{R}}\leq C(\Omega,p,\lambda)(\|\boldsymbol{f}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'}+\|\chi\|_{L^p(\Omega)}+\|g\|_{W^{1-1/p,p}(\Gamma)}+\|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)}). \quad (16)$$

Combining (15) together with (16) we obtain estimate (10).

□

Theorem 6. *Let* $1 < p < 2$, $\boldsymbol{f} \in (\boldsymbol{H}_0^{p'}(\mathrm{div}, \Omega))'$ *and* $\boldsymbol{h} \times \boldsymbol{n} \in \boldsymbol{W}^{-1/p,p}(\Gamma)$, $g \in W^{1-1/p,p}(\Gamma)$ *and* $\chi \in L^p(\Omega)$ *verifying the following compatibility condition* (9). *Then Problem* (1) *has a unique solution* $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{1,p}(\Omega) \times L^p(\Omega)/\mathbb{R}$.

Proof. **Step 1:** We suppose that $g = 0$. The problem

$$\left\{ \begin{array}{rll} \lambda \boldsymbol{u} - \Delta \boldsymbol{u} + \nabla \pi = \boldsymbol{f}, \quad \mathrm{div}\, \boldsymbol{u} = \chi, & \text{in} & \Omega, \\ \boldsymbol{u} \cdot \boldsymbol{n} = 0, \quad \mathbf{curl}\, \boldsymbol{u} \times \boldsymbol{n} = \boldsymbol{h} \times \boldsymbol{n}, & \text{on} & \Gamma, \end{array} \right. \tag{17}$$

has the following equivalent variational formulation: Find $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{1,p}(\Omega) \times L^p(\Omega)/\mathbb{R}$ satisfying $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ on Γ, such that $\forall\ \boldsymbol{w} \in \boldsymbol{W}^{1,p'}$ satisfying $\boldsymbol{w} \cdot \boldsymbol{n} = 0$ and $\mathbf{curl}\, \boldsymbol{w} \times \boldsymbol{n}$=0 on Γ

$$\lambda \int_\Omega \boldsymbol{u} \cdot \overline{\boldsymbol{w}}\, \mathrm{d}x + \int_\Omega \mathbf{curl}\, \boldsymbol{u} \cdot \mathbf{curl}\, \overline{\boldsymbol{w}}\, \mathrm{d}x - \int_\Omega \pi \cdot \mathrm{div}\, \overline{\boldsymbol{w}}\, \mathrm{d}x = \langle \boldsymbol{f}, \boldsymbol{w} \rangle_{[\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)]' \times \boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)} + \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{w} \rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma) \times \boldsymbol{W}^{-1/p,p'}(\Gamma)} - \int_\Omega \chi \cdot \mathrm{div}\, \overline{\boldsymbol{w}}\, \mathrm{d}x.$$

According to theorem 5, for any $(\boldsymbol{F}, \varphi)$ in $(\boldsymbol{H}_0^p(\mathrm{div}, \Omega))' \times L_0^{p'}(\Omega)$ there exists a unique solution $(\boldsymbol{w}, \eta) \in \boldsymbol{W}^{1,p'}(\Omega) \times L^{p'}(\Omega)/\mathbb{R}$ solution to

$$\left\{ \begin{array}{rll} \lambda \boldsymbol{w} - \Delta \boldsymbol{w} + \nabla \eta = \boldsymbol{F}, \quad \mathrm{div}\, \boldsymbol{w} = \varphi, & \text{in} & \Omega, \\ \boldsymbol{w} \cdot \boldsymbol{n} = 0, \quad \mathbf{curl}\, \boldsymbol{w} \times \boldsymbol{n} = 0, & \text{on} & \Gamma, \end{array} \right. \tag{18}$$

and satisfying

$$\|\boldsymbol{w}\|_{\boldsymbol{W}^{1,p'}(\Omega)} + \|\eta\|_{L^{p'}(\Omega)/\mathbb{R}} \leq C(\Omega, p', \lambda)(\|\boldsymbol{F}\|_{(\boldsymbol{H}_0^p(\mathrm{div},\Omega))'} + \|\varphi\|_{\boldsymbol{L}^{p'}(\Omega)}).$$

Let T be a linear form defined from $(\boldsymbol{H}_0^p(\mathrm{div}, \Omega))' \times L_0^{p'}(\Omega)$ onto $\mathbb{C}$ by

$$T : (\boldsymbol{F}, \varphi) \longmapsto \langle \boldsymbol{f}, \boldsymbol{w} \rangle_{[\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)]' \times \boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)} + \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{w} \rangle_\Gamma - \int_\Omega \chi \cdot \overline{\eta}\, \mathrm{d}x.$$

Observe that

$$|T(\boldsymbol{F}, \varphi)| \leq \|\boldsymbol{f}\|_{(\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'} \|\boldsymbol{w}\|_{\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega))'} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{-1/p,p}(\Gamma)} \|\boldsymbol{w}\|_{\boldsymbol{W}^{1/p,p'}(\Gamma)} + \|\varphi\|_{L^{p'}(\Omega)}.$$

Then T is continuous on $(\boldsymbol{H}_0^p(\mathrm{div}, \Omega))' \times \boldsymbol{L}^{p'}(\Omega)$ and we deduce that there exists a unique $(\boldsymbol{u}, \pi) \in \boldsymbol{H}_0^p(\mathrm{div}, \Omega) \times L^p(\Omega)/\mathbb{R}$ such that

$$T(\boldsymbol{F}, \varphi) = \langle \boldsymbol{u}, \boldsymbol{F} \rangle_{\boldsymbol{H}_0^p(\mathrm{div},\Omega) \times (\boldsymbol{H}_0^p(\mathrm{div},\Omega))'} - \int_\Omega \pi \cdot \overline{\varphi}\, \mathrm{d}x.$$

As a result

$$\lambda \int_\Omega \boldsymbol{u} \cdot \overline{\boldsymbol{w}}\, \mathrm{d}x + \int_\Omega \mathbf{curl}\, \boldsymbol{u} \cdot \mathbf{curl}\, \overline{\boldsymbol{w}}\, \mathrm{d}x - \int_\Omega \pi \cdot \mathrm{div}\, \overline{\boldsymbol{w}}\, \mathrm{d}x = \langle \boldsymbol{f}, \boldsymbol{w} \rangle_{[\boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)]' \times \boldsymbol{H}_0^{p'}(\mathrm{div},\Omega)} + \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{w} \rangle_{\boldsymbol{W}^{-1/p,p}(\Gamma) \times \boldsymbol{W}^{-1/p,p'}(\Gamma)} - \int_\Omega \chi \cdot \mathrm{div}\, \overline{\boldsymbol{w}}\, \mathrm{d}x.$$

To finish, we shall prove that $\boldsymbol{u}$ belongs to $\boldsymbol{W}^{1,p}(\Omega)$. To this end we write our problem in the form (7) where $\boldsymbol{F} = \boldsymbol{f} - \lambda\boldsymbol{u}$ belongs to $(\boldsymbol{H}_0^{p'}(\mathrm{div}, \Omega))'$ and satisfies (8). Then using [3, Remark 4.6] our solution $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{1,p}(\Omega) \times L^p(\Omega)$.
Step 2 : $g \neq 0$. Let $\theta \in W^{2,p}(\Omega)/\mathbb{R}$ be the unique solution of the Neumann problem (11) with $\chi \in L^p(\Omega)$ and $g \in \boldsymbol{W}^{1-1/p,p}(\Gamma)$ satisfying (9). Let $\boldsymbol{F} = \boldsymbol{f} + \nabla\chi - \lambda\nabla\theta \in (\boldsymbol{H}_0^{p'}(\mathrm{div}, \Omega))'$. Then there exists $(\boldsymbol{z}, \pi) \in \boldsymbol{W}^{1,p}(\Omega) \times L^p(\Omega)/\mathbb{R}$ solution of (13). Set $\boldsymbol{u}$=$\boldsymbol{z} + \nabla\theta$. We can easily verify that $(\boldsymbol{u}, \pi)$ solves (1). □

3.2. Strong solution

Theorem 7. *Let $1 < p < \infty$. Let $\boldsymbol{f} \in L^p(\Omega)$ and $\boldsymbol{h} \times \boldsymbol{n} \in \boldsymbol{W}^{1-1/p,p}(\Gamma)$. Then the problem* (4) *has a unique solution $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{2,p}(\Omega) \times W^{1,p}(\Omega)/\mathbb{R}$ satisfying the following estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{2,p}(\Omega)} + \|\pi\|_{W^{1,p}(\Omega)/\mathbb{R}} \leq C(\lambda, p, \Omega)(\|\boldsymbol{f}\|_{\boldsymbol{L}^p(\Omega)} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{1-1/p,p}(\Gamma)}). \tag{19}$$

Proof. We know that problem (4) has a unique solution $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{1,p}(\Omega) \times L^p(\Omega)/\mathbb{R}$. Moreover π satisfies

$$\mathrm{div}(\nabla\pi - \boldsymbol{f}) = 0 \text{ in } \Omega, \qquad (\nabla\pi - \boldsymbol{f}) \cdot \boldsymbol{n} = -\mathrm{div}_\Gamma(\boldsymbol{h} \times \boldsymbol{n}) \text{ on } \Gamma.$$

Since $\boldsymbol{h} \times \boldsymbol{n} \in \boldsymbol{W}^{1-1/p,p}(\Gamma)$ we deduce that that $\pi \in W^{1,p}(\Omega)$.

Set $\boldsymbol{z} = \mathbf{curl}\,\boldsymbol{u}$. Notice that $\boldsymbol{z}$ verify the following problem:

$$\begin{cases} \lambda\boldsymbol{z} - \Delta\boldsymbol{z} = \mathbf{curl}\boldsymbol{f}, \quad \mathrm{div}\,\boldsymbol{z} = 0, & \text{in} \quad \Omega, \\ \boldsymbol{z} \times \boldsymbol{n} = \boldsymbol{h} \times \boldsymbol{n}, & \text{on} \quad \Gamma, \end{cases} \tag{20}$$

where $\mathbf{curl}\,\boldsymbol{f} \in (\boldsymbol{H}_0^{p'}(\mathbf{curl}, \Omega))'$ and $\boldsymbol{h} \times \boldsymbol{n} \in \boldsymbol{W}^{1-1/p,p}(\Gamma)$. Then $\boldsymbol{z} \in \boldsymbol{W}^{1,p}(\Omega)$ and satisfies

$$\|\boldsymbol{z}\|_{\boldsymbol{W}^{1,p}(\Omega)} \leq C(\Omega)(\|\boldsymbol{f}\|_{\boldsymbol{L}^p(\Omega)} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{1-1/p,p}(\Gamma)}).$$

Thus $\boldsymbol{u} \in L^p(\Omega)$, $\mathrm{div}\,\boldsymbol{u} = 0 \in W^{1,p}(\Omega)$, $\mathbf{curl}\,\boldsymbol{u} = \boldsymbol{z} \in \boldsymbol{W}^{1,p}(\Omega)$ and $\boldsymbol{u} \cdot \boldsymbol{n} = 0 \in W^{1-1/p,p}(\Gamma)$. Then $\boldsymbol{u} \in \boldsymbol{W}^{2,p}(\Omega)$ and

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{2,p}(\Omega)} \leq C(\lambda, p, \Omega)(\|\boldsymbol{f}\|_{\boldsymbol{L}^p(\Omega)} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{1-1/p,p}(\Gamma)}).$$

Finally proceeding as in step 2 of the proof of theorem 5 , we obtain that the solution $(\boldsymbol{u}, \pi)$ satisfies the estimation (19) which ends the proof. □

Corollary 8. *Let $1 < p < \infty$. Let $\boldsymbol{f} \in \boldsymbol{L}^p(\Omega)$, $\boldsymbol{h} \times \boldsymbol{n} \in \boldsymbol{W}^{1-1/p,p}(\Gamma)$, $g \in W^{2-1/p,p}(\Gamma)$ and $\chi \in W^{1,p}(\Omega)$ verifying the following compatibility condition* (9)*. Then problem* (1) *has a unique solution $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{2,p}(\Omega) \times W^{1,p}(\Omega)/\mathbb{R}$ satisfying*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{2,p}(\Omega)} + \|\pi\|_{W^{1,p}(\Omega)/\mathbb{R}} \leq C(\Omega, p, \lambda)(\|\boldsymbol{f}\|_{\boldsymbol{L}^p(\Omega)} + \|\chi\|_{W^{1,p}(\Omega)} + \|g\|_{W^{2-1/p,p}(\Gamma)} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{1-1/p,p}(\Gamma)}). \tag{21}$$

Proof. Let $\theta \in W^{2,p}(\Omega)$ be the unique solution of the Neumann problem (11).
Set $\boldsymbol{F} = \boldsymbol{f} - \lambda\nabla\theta + \nabla\chi$ and observe that $\boldsymbol{F} \in \boldsymbol{L}^p(\Omega)$. Thanks to Theorem 7, the Problem (13) has a unique solution $(\boldsymbol{z}, \pi) \in \boldsymbol{W}^{2,p}(\Omega) \times W^{1,p}(\Omega)/\mathbb{R}$ satisfying

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{2,p}(\Omega)} + \|\pi\|_{W^{1,p}(\Omega)/\mathbb{R}} \leq C(\Omega, p, \lambda)(\|\boldsymbol{f}\|_{\boldsymbol{L}^p(\Omega)} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{1-1/p,p}(\Gamma)}).$$

By setting $\boldsymbol{u} = \boldsymbol{z} + \nabla\theta$, we can easily verify that $(\boldsymbol{u}, \pi)$ solves (1) and verifies (21). □

3.3. Very weak solution

In this subsection we prove the existence of very week solution to Problem (1).

Theorem 9. *Let $\boldsymbol{f} \in (\boldsymbol{T}^{p'}(\Omega))'$, $\chi \in L^p(\Omega)$, $g \in W^{-1/p,p}(\Gamma)$ and $\boldsymbol{h} \times \boldsymbol{n} \in \boldsymbol{W}^{-1-1/p,p}(\Gamma)$ verifying the compatibility condition* (9)*. Then problem* (1) *has a unique solution $(\boldsymbol{u}, \pi) \in \boldsymbol{L}^p(\Omega) \times W^{-1,p}(\Omega)/\mathbb{R}$. Moreover the following estimate holds*

$$\|\boldsymbol{u}\|_{\boldsymbol{L}^p(\Omega)} + \|\pi\|_{W^{-1,p}(\Omega)/\mathbb{R}} \leq C(\Omega, p, \lambda)(\|\boldsymbol{f}\|_{(\boldsymbol{T}^{p'}(\Omega))'} + \|\chi\|_{L^p(\Omega)} + \|g\|_{W^{-1/p,p}(\Gamma)} + \|\boldsymbol{h} \times \boldsymbol{n}\|_{\boldsymbol{W}^{-1-1/p,p}(\Gamma)}). \quad (22)$$

Proof. **Step 1.** Problem (1) is equivalent to the variational formulation: find $(\boldsymbol{u}, \pi) \in \boldsymbol{L}^p(\Omega) \times W^{-1,p}(\Omega)/\mathbb{R}$ such that for any $\boldsymbol{\phi} \in \boldsymbol{Y}_\tau^{p'}(\Omega)$, and for any $q \in W^{1,p'}(\Omega)$,

$$\lambda \int_\Omega \boldsymbol{u} \cdot \overline{\boldsymbol{\phi}}\, \mathrm{d}x - \int_\Omega \boldsymbol{u} \cdot \Delta\overline{\boldsymbol{\phi}}\, \mathrm{d}x - \langle \pi, \mathrm{div}\boldsymbol{\phi} \rangle_{W^{-1,p}(\Omega) \times W_0^{1,p'}(\Omega)} = \langle \boldsymbol{f}, \boldsymbol{\phi} \rangle_\Omega + \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{\phi} \rangle_\Gamma \quad (23)$$

$$\int_\Omega \boldsymbol{u} \cdot \nabla \overline{q}\, \mathrm{d}x = -\int_\Omega \chi \overline{q}\, \mathrm{d}x + \langle g, q \rangle_{W^{-1/p,p}(\Gamma) \times W^{1/p,p'}(\Gamma)}, \quad (24)$$

where $\langle \cdot, \cdot \rangle_\Omega = \langle \cdot, \cdot \rangle_{(\boldsymbol{T}^{p'}(\Omega))' \times \boldsymbol{T}^{p'}(\Omega)}$ and $\langle \cdot, \cdot \rangle_\Gamma = \langle \cdot, \cdot \rangle_{\boldsymbol{W}^{-1-1/p,p}(\Gamma) \times \boldsymbol{W}^{1+1/p,p'}(\Gamma)}$.
Indeed, using the Green formula (3), we can verify that every $(\boldsymbol{u}, \pi) \in \boldsymbol{L}^p(\Omega) \times W^{-1,p}(\Omega)$ solution to (1) solves (23)-(24). Conversely, let $(\boldsymbol{u}, \pi) \in \boldsymbol{L}^p(\Omega) \times W^{-1,p}(\Omega)$ be a solution to (23)-(24). Clearly, $-\Delta\boldsymbol{u} + \nabla\pi = \boldsymbol{f}$ and $\mathrm{div}\,\boldsymbol{u} = \chi$ in Ω.
Consequently, $\boldsymbol{u} \in \boldsymbol{L}^p(\Omega)$ and since $\nabla\pi \in (\boldsymbol{T}^{p'}(\Omega))'$, we have $\Delta\boldsymbol{u} = -\boldsymbol{f} + \lambda\boldsymbol{u} + \nabla\pi \in (\boldsymbol{T}^{p'}(\Omega))'$. Then $\boldsymbol{u} \in \boldsymbol{H}_p(\Delta, \Omega)$. Using (2) and (3), we obtain that for any $\boldsymbol{\phi} \in \boldsymbol{Y}_\tau^{p'}(\Omega)$:

$$\lambda \int_\Omega \boldsymbol{u} \cdot \overline{\boldsymbol{\phi}}\, \mathrm{d}x - \int_\Omega \boldsymbol{u} \cdot \Delta\overline{\boldsymbol{\phi}}\, \mathrm{d}x - \langle \mathbf{curl}\boldsymbol{u} \times \boldsymbol{n}, \boldsymbol{\phi} \rangle_\Gamma - \langle \pi, \mathrm{div}\boldsymbol{\phi} \rangle_{W^{-1,p}(\Omega) \times W_0^{1,p'}(\Omega)} = \langle \boldsymbol{f}, \boldsymbol{\phi} \rangle_\Omega.$$

Thus $\langle \mathbf{curl}\boldsymbol{u} \times \boldsymbol{n}, \boldsymbol{\phi} \rangle_\Gamma = \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{\phi} \rangle_\Gamma$. Let $\boldsymbol{\mu} \in \boldsymbol{W}^{1+1/p,p'}(\Gamma)$, there exists a function $\boldsymbol{\phi} \in \boldsymbol{W}^{2,p}(\Omega)$ satisfying

$$\boldsymbol{\phi}_\tau = \boldsymbol{\mu}_\tau \quad \text{and} \quad \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{n}} = -\boldsymbol{n}\,\mathrm{div}_\Gamma \boldsymbol{\mu}_\tau + \sum_{j=1}^{2} \Big(\frac{\partial \boldsymbol{\mu}_\tau}{\partial s_j} \times \boldsymbol{T}_j \Big) \times \boldsymbol{n} \ \text{ on } \ \Gamma.$$

It is clear that $\boldsymbol{\phi} \in \boldsymbol{Y}_\tau^{p'}(\Omega)$ and

$$\langle \mathbf{curl}\boldsymbol{u} \times \boldsymbol{n}, \boldsymbol{\mu} \rangle_\Gamma - \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{\mu} \rangle_\Gamma = \langle \mathbf{curl}\boldsymbol{u} \times \boldsymbol{n}, \boldsymbol{\phi}_\tau \rangle_\Gamma - \langle \boldsymbol{h} \times \boldsymbol{n}, \boldsymbol{\phi}_\tau \rangle_\Gamma = 0.$$

Thus $\mathbf{curl}\,\boldsymbol{u} \times \boldsymbol{n} = \boldsymbol{h} \times \boldsymbol{n}$ on Γ. Next using that $\mathrm{div}\,\boldsymbol{u} = \chi$ in Ω, we deduce that for any $q \in W^{1,p'}(\Omega)$, we have
$\langle \boldsymbol{u} \cdot \boldsymbol{n}, q \rangle_{W^{-1/p,p}(\Gamma) \times W^{1/p,p'}(\Gamma)} = \langle g, q \rangle_{W^{-1/p,p}(\Gamma) \times W^{1/p,p'}(\Gamma)}$. Consequently, $\boldsymbol{u} \cdot \boldsymbol{n} = g \in W^{-1/p,p}(\Gamma)$.

Step 2. Let us now solve Problem (23)-(24). We suppose that

$$g = 0 \quad \text{on} \quad \Gamma \qquad \text{and} \qquad \int_\Omega \chi\, \mathrm{d}x = 0.$$

Thanks to Theorem 8, for any pair $(\boldsymbol{F},\xi)\in \boldsymbol{L}^{p'}(\Omega)\times(W_0^{1,p'}(\Omega)\cap L_0^{p'}(\Omega))$ there exists a unique $(\boldsymbol{\phi},q)\in \boldsymbol{W}^{2,p'}(\Omega)\times W^{1,p'}(\Omega)/\mathbb{R}$ satisfying:

$$\left\{\begin{array}{rl} \lambda\boldsymbol{\phi}-\Delta\boldsymbol{\phi}+\nabla q=\boldsymbol{F}, & \operatorname{div}\boldsymbol{\phi}=\xi, \qquad \text{in} \quad \Omega,\\ \boldsymbol{\phi}\cdot\boldsymbol{n}=0, & \mathbf{curl}\,\boldsymbol{\phi}\times\boldsymbol{n}=0, \quad \text{on} \quad \Gamma, \end{array}\right. \tag{25}$$

with the estimate

$$\|\boldsymbol{\phi}\|_{\boldsymbol{W}^{2,p'}(\Omega)}+\|q\|_{\boldsymbol{W}^{1,p'}(\Omega)/\mathbb{R}}\leq C(\lambda,\Omega,p')(\|\boldsymbol{F}\|_{\boldsymbol{L}^{p'}(\Omega)}+\|\xi\|_{W^{1,p'}(\Omega)}).$$

Let T be a linear form defined from $\boldsymbol{L}^{p'}(\Omega)\times(W_0^{1,p'}(\Omega)\cap L_0^{p'}(\Omega))$ onto $\mathbb{C}$ by

$$T:(\boldsymbol{F},\xi)\longmapsto\langle\boldsymbol{f},\boldsymbol{\phi}\rangle_\Omega+\langle\boldsymbol{h}\times\boldsymbol{n},\boldsymbol{\phi}\rangle_\Gamma-\int_\Omega\chi q\,\mathrm{d}x.$$

An easy computation shows that

$$|T(\boldsymbol{F},\xi)|\leq C(\Omega,p',\lambda)(\|\boldsymbol{f}\|_{(\boldsymbol{T}^{p'}(\Omega))'}+\|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1-1/p,p}(\Gamma)}+\|\chi\|_{L^p(\Omega)})(\|\boldsymbol{F}\|_{\boldsymbol{L}^{p'}(\Omega)}+\|\xi\|_{W^{1,p'}(\Omega)}).$$

This means that T defines an element of the dual space of $\boldsymbol{L}^{p'}(\Omega)\times(W_0^{1,p'}(\Omega)\cap L_0^{p'}(\Omega))$ and according to the Riesz's representation theorem, there exists a unique $(\boldsymbol{u},\pi)\in\boldsymbol{L}^p(\Omega)\times W^{-1,p}(\Omega)/\mathbb{R}$ such that

$$T(\boldsymbol{F},\xi)=\langle\boldsymbol{u},\boldsymbol{F}\rangle_{\boldsymbol{T}^{p'}(\Omega)\times(\boldsymbol{T}^{p'}(\Omega))'}-\int_\Omega\pi\xi\,\mathrm{d}x.$$

Then $(\boldsymbol{u},\pi)$ is a solution to (23)-(24) and satisfies (22).

Step 3. Suppose that $g\neq 0$ and the compatibility condition (9) holds. The Neumann problem (11) has a unique solution $\theta\in W^{1,p}(\Omega)/\mathbb{R}$ satisfying the estimate:

$$\|\theta\|_{W^{1,p}(\Omega)/\mathbb{R}}\leq C(\|\chi\|_{L^p(\Omega)}+\|g\|_{W^{-1/p,p}(\Gamma)}).$$

Set $\boldsymbol{F}=\boldsymbol{f}-\lambda\nabla\theta+\nabla\chi$. Then $\boldsymbol{F}\in(\boldsymbol{T}^{p'}(\Omega))'$ and the Problem (13) has a unique solution (z,π) $\in\boldsymbol{L}^p(\Omega)\times W^{-1,p}(\Omega)/\mathbb{R}$ satisfying the following estimate

$$\|z\|_{\boldsymbol{L}^p(\Omega)}+\|\pi\|_{W^{-1,p}(\Omega)/\mathbb{R}}\leq C(\lambda,\Omega,p)\left(\|\boldsymbol{F}\|_{(\boldsymbol{T}^{p'}(\Omega))'}+\|\boldsymbol{h}\times\boldsymbol{n}\|_{\boldsymbol{W}^{-1-1/p,p}(\Gamma)}\right). \tag{26}$$

Then $(\boldsymbol{u},\pi)$ with $\boldsymbol{u}=z+\nabla\theta$ solves (1) and satisfies (22). □

Remark 1. **i)** Consider the Problem (1) with $\chi\in W^{1,p}(\Omega)$ such that $\int_\Omega\chi\,\mathrm{d}\,x=0$, $g=0$ and $\boldsymbol{h}=\boldsymbol{0}$ on Γ. As in [7] we can prove that the solution $(\boldsymbol{u},\pi)$ satisfies the following estimate

$$|\lambda|\,\|\boldsymbol{u}\|_{\boldsymbol{L}^p(\Omega)}+\|\nabla\pi\|_{\boldsymbol{L}^p(\Omega)}\leq\ C\,(\|\boldsymbol{f}\|_{\boldsymbol{L}^p(\Omega)}\ +\ \|\nabla\chi\|_{\boldsymbol{L}^p(\Omega)}\ +\ |\lambda|\,\|\chi\|_{W^{-1,p}(\Omega)}). \tag{27}$$

Indeed, let $\theta\in W^{2,p}(\Omega)/\mathbb{R}$ solution to $\Delta\,\theta=\chi$ in Ω, $\frac{\partial\theta}{\partial\boldsymbol{n}}=0$ on Γ and satisfying $\|\theta\|_{W^{2,p}(\Omega)}\leq C\,\|\chi\|_{W^{1,p}(\Omega)}$. Set $\boldsymbol{F}=\boldsymbol{f}-\lambda\nabla\theta+\nabla\chi$, then $\boldsymbol{F}\in\boldsymbol{L}^p(\Omega)$) and the problem

$$\left\{\begin{array}{rl} \lambda z-\Delta z+\nabla\pi=\boldsymbol{F}, & \operatorname{div}z=0 \qquad \text{in} \quad \Omega\\ z\cdot\boldsymbol{n}=0, & \mathbf{curl}\,z\times\boldsymbol{n}=\boldsymbol{0} \quad \text{on} \quad \Gamma \end{array}\right.$$

has a unique solution $(z,\pi)\in\boldsymbol{W}^{1,p}(\Omega)\times L^p(\Omega)/\mathbb{R}$ satisfying the following estimate

$$|\lambda|\,\|z\|_{\boldsymbol{W}^{1,p}(\Omega)}\ +\ \|\nabla\pi\|_{\boldsymbol{L}^p(\Omega)}\leq C(\Omega,p)\left(\|\boldsymbol{f}\|_{\boldsymbol{L}^p(\Omega))}+\|\nabla\chi\|_{\boldsymbol{L}^p(\Omega))}+|\lambda|\|\nabla\theta\|_{\boldsymbol{L}^p(\Omega))}\right)$$

Set $\boldsymbol{u}=z+\nabla\theta$. Then $(\boldsymbol{u},\pi)$ is a solution to (1) and satisfies (27).

ii) Notice that when $\chi=0$ we recover the resolvent estimate established in [1] and [2].

References

[1] Al Baba, H., Amrouche, C., and Escobedo, M. Analyticity of the semi-group generated by the Stokes operator with Navier-type boundary conditions on L_p-spaces. In *Recent advances in partial differential equations and applications*, vol. 666 of *Contemp. Math.* Amer. Math. Soc., Providence, RI, 2016, pp. 23–40.

[2] Al Baba, H., Amrouche, C., and Escobedo, M. Semi-group theory for the Stokes operator with Navier-type boundary conditions on L^p-spaces. *Arch. Ration. Mech. Anal. 223*, 2 (2017), 881–940.

[3] Amrouche, C., and Seloula, N. E. H. On the Stokes equations with the Navier-type boundary conditions. *Differ. Equ. Appl. 3*, 4 (2011), 581–607.

[4] Amrouche, C., and Seloula, N. E. H. L^p-theory for vector potentials and Sobolev's inequalities for vector fields: application to the Stokes equations with pressure boundary conditions. *Math. Models Methods Appl. Sci. 23*, 1 (2013), 37–92.

[5] Farwig, R., and Sohr, H. Generalized resolvent estimates for the Stokes system in bounded and unbounded domains. *J. Math. Soc. Japan 46*, 4 (1994), 607–643.

[6] Giga, Y. Analyticity of the semigroup generated by the Stokes operator in L_r spaces. *Math. Z. 178*, 3 (1981), 297–329.

[7] Giga, Y., and Sohr, H. On the Stokes operator in exterior domains. *J. Fac. Sci. Univ. Tokyo Sect. IA Math. 36*, 1 (1989), 103–130.

[8] Mitrea, M., and Monniaux, S. On the analyticity of the semigroup generated by the Stokes operator with Neumann-type boundary conditions on Lipschitz subdomains of Riemannian manifolds. *Trans. Amer. Math. Soc. 361*, 6 (2009), 3125–3157.

[9] Miyakawa, T. The L^p approach to the Navier-Stokes equations with the Neumann boundary condition. *Hiroshima Math. J. 10*, 3 (1980), 517–537.

[10] Saal, J. Stokes and Navier-Stokes equations with Robin boundary conditions in a half-space. *J. Math. Fluid Mech. 8*, 2 (2006), 211–241.

[11] Shibata, Y., and Shimada, R. On a generalized resolvent estimate for the Stokes system with Robin boundary condition. *J. Math. Soc. Japan 59*, 2 (2007), 469–519.

[12] Solonnikov, V. A. Estimates of the solution of model evolution generalized Stokes problem in weighted Hölder spaces. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 336*, Kraev. Zadachi Mat. Fiz. i Smezh. Vopr. Teor. Funkts. 37 (2006), 211–238, 277.

Hind Al Baba and Antonia Jabbour
LAMA-Liban, Laboratoire de Mathématiques
et Appliquations, Lebanese University
Beirut, Lebanon
hind.albaba@ul.edu.lb
antoniajabbour95@hotmail.com

Monografías Matemáticas *García de Galdeano* **42**, 13–23 (2019)

BEYOND WENTZELL-FREIDLIN: SEMI-DETERMINISTIC APPROXIMATIONS FOR DIFFUSIONS WITH SMALL NOISE AND A REPULSIVE CRITICAL BOUNDARY POINT

Florin Avram and Jacky Cresson

Abstract. We extend below a limit theorem [2] for diffusion models used in population theory.

Keywords: dynamical systems, small noise, linearization, semi-deterministic fluid approximation.

AMS classification: AMS 60J60.

§1. Introduction

A diffusion with small noise is defined as the solution of a stochastic differential equation (SDE) driven by standard Brownian motion $B_t(\cdot)$ (defined on a probability space and progressively measurable with respect to an increasing filtration)

$$\begin{cases} dX_t^\varepsilon = \mu(X_t^\varepsilon)dt + \sqrt{\varepsilon}\sigma(X_t^\varepsilon)dB_t, \quad t \geq 0, \\ X_0^\varepsilon = x_0 = \varepsilon, X_t^\varepsilon \in \mathcal{I} := (0, r) \end{cases} \tag{1}$$

where $0 < r \leq +\infty$, $\varepsilon > 0, \mu : \mathcal{I} \mapsto \mathbb{R}$, $\sigma : \mathcal{I} \mapsto \mathbb{R}_{>0}$ and μ, σ satisfy conditions ensuring that (1) has a strong unique solution (for example, μ is locally Lifshitz and σ satisfies the Yamada-Watanabe conditions [14, (2.13), Ch.5.2.C]).[§]

When $\varepsilon \to 0$, (1) is a small perturbation of the dynamical system/ordinary differential equation (ODE):

$$\frac{dx_t}{dt} = \mu(x_t), \quad t \geq 0, \tag{2}$$

which will also be supposed to admit a unique continuous solution $x_t, t \in \mathbb{R}_+$ subject to any $x_0 \in (0, r)$, and the flow of which will be denoted by $\phi_t(x)$.

A basic result in the field is the "fluid limit", which states that when (1) admits a strong unique solution, the effect of noise is negligible as $\varepsilon \to 0$, on any **fixed time interval** $[0, T]$:

[§]For reviews discussing the existence of strong and weak solutions, see for example [5, 13, 8].

Theorem 1. *[Freidlin and Wentzell] [11, Thm 1.2, Ch. 2.1] Let X_t^ε satisfy* (1)*, assume μ, σ satisfy the Lifshitz condition, and that $X_0^\varepsilon \xrightarrow[\varepsilon\to 0]{¶} x_0 \in \mathbb{R}_+$, where $\xrightarrow[\varepsilon\to 0]{¶}$ denotes convergence in probability. Then, for any fixed T*

$$\sup_{t\le T} |X_t^\varepsilon - x_t| \xrightarrow[\varepsilon\to 0]{¶} 0,$$

where x_t is the solution of (2) *subject to the initial condition x_0.*[¶]

Although interesting, this result does not give any understanding of the asymptotic behavior of the diffusion process for times converging to infinity; in particular, it does not tell us how the diffusion travels between equilibrium points (which requires times converging to infinity). Following [4, 2], we go here beyond Theorem 1, by analyzing the way a diffusion process leaves an unstable equilibrium point. Precisely, we make the following assumptions:

Assumption 1. Suppose from now on that $l = 0, \mu(0) = 0, \mu'(0) > 0$, which makes zero an **unstable equilibrium point of** (2) and of (1).

Note that under Assumption 1, the Freidlin-Wentzell theorem 1 implies that the solution of (1) started from a small positive initial condition $X_0^\varepsilon = \varepsilon > 0$ converges to zero on any fixed bounded interval

$$\sup_{t\le T} \left|X_t^\varepsilon\right| \xrightarrow[\varepsilon\to 0]{¶} 0, \qquad \forall T \ge 0.$$

Assumption 2. Put now $a(x) = \sigma^2(x)$, and assume that $a(0) = \sigma(0) = 0, a'(0) > 0$, which makes 0 a singular point of the diffusion (1)– see for example [8].

Remark 1. Note that $a'(0) > 0$ rules out important population theory models like the linear Gilpin Ayala diffusion [17] with

$$\mu(x) = \gamma x\Big(1 - (\frac{x}{x_c})^\alpha\Big), \sigma(x) = \sqrt{\varepsilon} x \Leftrightarrow a(x) = \varepsilon x^2, \gamma > 0, x_c > 0, \alpha > 0, \tag{3}$$

which includes by setting $\alpha = 1$ another favorite, the logistic-type Verlhurst-Pearl diffusion [12, 9, 1].

Recently, a new type of limit theorem [2] was discovered when $T \to \infty$ under Assumptions 1 and 2, when x_0^ε converges to the unstable equilibrium point of (2). Following [2], let

$$T^\varepsilon := \frac{1}{\mu'(0)} \log \frac{1}{\varepsilon} \tag{4}$$

denote the solution of the equation $\phi_{t,lin}(x_0) = x_0 e^{\mu'(0)t} = 1$ where $\phi_{t,lin}(x_0)$ is the flow of the linearized system of (2) in 0, and divide the evolution of the process in three time-intervals:

$$[0, t_c := cT^\varepsilon], [t_c, t_1 := T^\varepsilon], [t_1, \infty), c \in (1/2, 1) \tag{5}$$

(the restriction $c > 1/2$ is used in (15)).

It turns out that this partition allows separating the life-time of diffusions with small noise, exiting an unstable point of the fluid limit, into three periods with distinct behaviors:

[¶]For other deterministic limit theorems for one-dimensional diffusions, see also Gikhman and Skorokhod [19], Freidlin and Wentzell [11], Keller et al. [16], and Buldygin et al. [6].

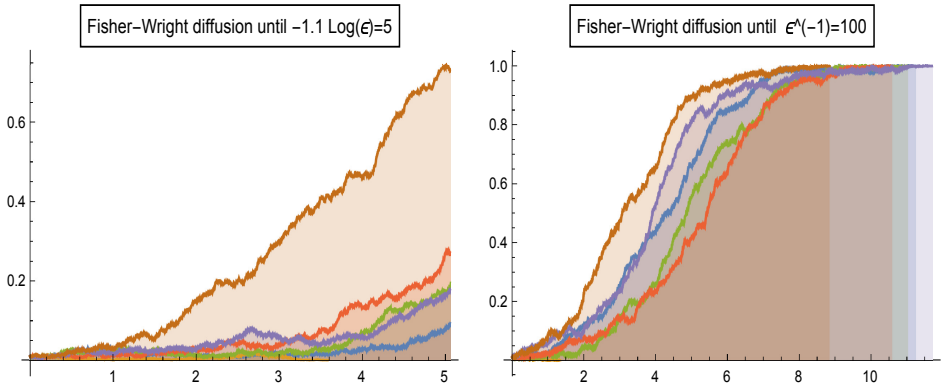


Figure 1: 6 paths of the Kimura-Fisher-Wright diffusion $dX_t = \gamma X_t(1 - X_t)dt + \sqrt{\varepsilon X_t(1 - X_t)}dB_t$, where $x_c = 1$ is an exit boundary, with $\varepsilon = .01$. On the right, three stages of evolution may be discerned

1. In the first stage, the process leaves the neighborhood of the unstable point. The linearization of the SDE implies that here a Feller branching approximation may be used, and this produces a certain exit law W which will be carried over to the next stage as a (random) initial condition.
2. In the second "semi-deterministic stage" (meaning that paths cross very rarely here), the system moves towards its first stable critical point x_c, following the trajectories of its fluid limit (2), again over a time whose length converges to ∞. A further renormalization produces here the main result, the limit exit law (7).
3. In the third stage, after the SDE has approaches the stable critical point of the fluid limit, "randomness is regained" – see crossings of paths in figures 1 and 2); (if the process may reach and overshoot the stable critical point, convergence towards a stationary distribution may occur).

The following result was obtained first in [2], for the "Kimura-Fisher-Wright" diffusion, and extended subsequently to diffusions with bounded volatility.

Theorem 2. Fluid limit with random initial conditions *[2]. Let X_t^ε satisfy Assumption 1,* (1)*, and $X_0^\varepsilon = \varepsilon > 0$. Suppose in addition that the diffusion coefficient $\sigma(\cdot)$ is continuous and* **bounded***, as well as its first derivative, and that $\mu(\cdot)$ satisfies the following drift condition:*

$$\left|\mu(y) - \mu(x)\right| \leq \mu'(0)|y - x|, \quad x, y \in \mathbb{R}_+.$$

Let Y_t denote the solution to the **scaled linearized equation**

$$dY_t = \mu'(0)Y_t dt + \sqrt{a'(0)Y_t}dB_t, Y_0 = 1 \Longrightarrow Y_t = 1 + \int_0^t \mu'(0)Y_s ds + \int_0^t \sqrt{a'(0)Y_s}dB_s, \quad (6)$$

known as **Feller branching diffusion**.

Then, it holds that :

(A)

$$X^{\varepsilon}_{T^{\varepsilon}} \xrightarrow[\varepsilon\to 0]{\P} \widetilde{\phi}(W), \tag{7}$$

where

(i) the random variable W is the a.s. martingale limit

$$W := \lim_{t\to\infty} e^{-\mu'(0)t} Y_t = 1 + \int_0^{\infty} e^{-\frac{\mu'(0)}{2}s} \sqrt{a'(0) Y_s} dB_s \tag{8}$$

(ii) $\widetilde{\phi}(x)$ *denotes the* **limit of the deterministic flow pushed first backward in time by the linearized deterministic flow** $\phi_{t,lin}(x) = xe^{\mu'(0)t}$ *near the unstable critical point* 0

$$\widetilde{\phi}(x) = \lim_{t\to\infty} \phi_t(\phi_{-t,lin}(x)) = \lim_{t\to\infty} \phi_t(xe^{-\mu'(0)t}), \quad x \geq 0. \tag{9}$$

(B) Also, for any $T > 0$*,*

$$\sup_{t\in[0,T]} \left| X^{\varepsilon}_{T_{\varepsilon}+t} - x_t \right| \xrightarrow[\varepsilon\to 0]{\P} 0, \tag{10}$$

where x_t *is the solution of* (2) *subject to the initial condition* $X_0 = \widetilde{\phi}(W)$.

Remark 2. Note that W depends only on the local parameters $\mu'(0), a'(0)$ of the diffusion at the critical point. Assume from now on, without loss of generality that $a'(0) = 1$ (recalling however that this is the only part of the stochastic perturbation that survives in the limiting regime), and put

$$\gamma := \mu'(0) > 0. \tag{11}$$

The Laplace transform of W is well known [18] and easy to compute.

$$Ee^{-\lambda W} = \lim_{t\to\infty} Ee^{-\lambda W_t} = \exp\left(-\frac{2\gamma\lambda}{2\gamma+\lambda}\right) = E\exp\Big(-\lambda \sum_{j=0}^{\Pi} \tau_j\Big), \tag{12}$$

which is the Laplace transform of a Poisson $\Pi \sim \mathrm{Poi}(2\gamma)$ sum of independent random variables $\tau_j \sim \mathrm{Exp}(2\gamma)$.

Remark 3. The main part of Theorem 2 is the equation (7) which identifies the limit after the second stage

$$X^{\varepsilon}_{T^{\varepsilon}} = \Phi^{\varepsilon}_{T^{\varepsilon}}(\varepsilon) \xrightarrow[\varepsilon\to 0]{\P} \lim_{t\to\infty} \phi_t(\phi_{-t,lin}(W)) = \widetilde{\phi}(W), \tag{13}$$

$\Phi^{\varepsilon}_t(x)$ denotes the flow generated by the SDE (1).

Note that $\widetilde{\phi}$ depends only on the dynamical system μ, and that by [2, Prop. 4.1], it is a nontrivial solution of the ODE $\gamma x\widetilde{\phi}'(x) = \mu(\widetilde{\phi}(x)), \quad x > 0, \widetilde{\phi}(0) = 0$.

(13) suggests possible generalizations to multidimensional diffusions and possibly to jump-diffusions (where a CBI might replace the Feller diffusion in the limit).

Remark 4. Part 2. of Theorem 2 follows immediately by a simple change of time: letting $\widetilde{X}^{\varepsilon}_t = X^{\varepsilon}_{T^{\varepsilon}+t}$, and $\widetilde{B}_t = B_{T^{\varepsilon}+t} - B_{T^{\varepsilon}}$ one obtains from (1)

$$\widetilde{X}^{\varepsilon}_t = \widetilde{X}^{\varepsilon}_0 + \int_0^t f(\widetilde{X}^{\varepsilon}_s) ds + \int_0^t \sqrt{\varepsilon\sigma(\widetilde{X}^{\varepsilon}_s)} d\widetilde{B}_s,$$

and the result follows from (7) by the fluid convergence Theorem 1. This part may be viewed as describing "short transitions" (invisible on a long time scale) between the second and third stages.

Remark 5. The limit (7) describing the position after the second stage has been established in [2] for one dimensional distributions with bounded $\sigma(x)$. This assumption seems however restrictive, since for typical diffusions whose fluid limit $\phi_t(x)$ admits a stable critical point x_c, the probability of leaving the neighborhood of the stable point x_c is very small as $\varepsilon \to 0$. This intuition is confirmed by simulations –see Figure 2.

The remark 5 suggests the relation of our problem to that of studying the maximum of X_t. More precisely, we would like to establish and exploit the plausible fact that $\forall \theta > 1$

$$\lim_{\varepsilon\to 0} P[T_{\theta x_c} < T^\varepsilon | X_0 = \varepsilon] = \lim_{\varepsilon\to 0} P[\sup_{0\le t\le T^\varepsilon} X_t^\varepsilon > \theta x_c | X_0 = \varepsilon] = 0, \tag{14}$$

where x_c is the closest critical point towards which the diffusion is attracted, and $T_{\theta x_c}$ is the hitting time of θx_c; clearly, (14) renders unnecessary the assumption that the diffusion coefficient $\sigma(\cdot)$ be bounded.

A weaker statement than (14), but still sufficient for a slight extension, is provided in the elementary Lemma (3) below.

Contents. The paper is organized as follows. In Section 2 we offer, based on Lemma 3, a slight extension of Theorem 2 of [2]. A conjecture (see Problem 1 is presented here as well. We illustrate our new result with the example of the logistic Feller diffusion in Section 3. We include for convenience in Section 4 an outline of the remarkable paper [2].

§2. An extension of Theorem 2 [2]

Recall now from [2] that the restrictive condition $\|\sigma\|_\infty < \infty$ is used for proving that[‡]

$$\|\sigma\|_\infty < \infty, c \in (1/2, 1) \Longrightarrow \Phi_{t_c,t_1}(X^\varepsilon_{t_c}) - \phi_{t_c,t_1}(X^\varepsilon_{t_c}) \xrightarrow[\varepsilon\to 0]{L^2} 0, \tag{16}$$

where $t_c = cT^\varepsilon$.

We will show now that it is possible to remove the condition $\|\sigma\|_\infty < \infty$ in (16), if only convergence in probability is needed, by assuming rather weak and natural conditions on the scale function $s(\cdot)$. Recall that the scale function s is defined (up to two integration constants) as an arbitrary increasing solution of the equation $\mathcal{L}s(x) = 0$, where $\mathcal{L}$ is the generator

[‡]Let us recall the proof of this important piece of the puzzle. Let $\Phi_{s,t}(x)$, $\phi_{s,t}(x)$ denote the stochastic and deterministic flows generated respectively by the SDE (1) and ODE (2), put $\Phi^\varepsilon_t := \Phi_{t_c,t_c+t}(X^\varepsilon_{t_c})$, $\phi_t := \phi_{t_c,t_c+t}(X^\varepsilon_{t_c})$ for brevity, and define $\delta^\varepsilon_t = \Phi^\varepsilon_t - \phi_t$. Subtracting equations (1) and (2) and applying the Itô formula:

$$E(\delta^\varepsilon_t)^2 = E\int_0^t 2\delta_s(\mu(\Phi^\varepsilon_s) - \mu(\phi_s))ds + \int_0^t \varepsilon E\sigma(\Phi^\varepsilon_s)ds \le \int_0^t 2\gamma E(\delta_s)^2 ds + \varepsilon t\|\sigma\|_\infty, t \in \mathbb{R}_+$$

where assumption 2 was used. By Grönwall's inequality

$$E\Big(\Phi_{t_c,t_1}(X^\varepsilon_{t_c}) - \phi_{t_c,t_1}(X^\varepsilon_{t_c})\Big)^2 = E(\delta^\varepsilon_{t_1-t_c})^2 \le C_1\varepsilon t_1 e^{2\gamma(t_1-t_c)} \le C_2\varepsilon^{2c-1}\log\frac{1}{\varepsilon} \xrightarrow[\varepsilon\to 0]{} 0 \tag{15}$$

where the convergence holds since $c \in (\frac{1}{2}, 1)$.

operator of the diffusion, and that this function is continuous – see [15, Ch. 15, (3.5), (3.6)] (noting that [15] denote the scale function by $S(\cdot)$).

Lemma 3. *Assume that* 0 *is an attracting boundary and that r is an unattracting boundary, i.e. that* , $s(0_+) > -\infty, s(r-) = \infty$. *Put*

$$\overline{X}^{\varepsilon} = \sup_{0 \le t < \infty} X_t^{\varepsilon}, \tag{17}$$

where X^{ε} *is defined in* (1). *Then:*

$$(A) \quad \forall \varepsilon, \lim_{M \to r} P_{\varepsilon}[\overline{X}^{\varepsilon} > M] = \lim_{M \to r} \frac{s(\varepsilon) - s(0)}{s(M) - s(0)} = (s(\varepsilon) - s(0)) \lim_{M \to r} \frac{1}{s(M) - s(0)} = 0, \tag{18}$$

and

$$(B) \quad c \in (1/2, 1) \Longrightarrow \Phi_{t_c,t_1}(X^{\varepsilon}_{t_c}) - \phi_{t_c,t_1}(X^{\varepsilon}_{t_c}) \xrightarrow[\varepsilon \to 0]{P} 0. \tag{19}$$

Proof. (18) is straightforward. Indeed, recall that the boundary 0 is attracting. Then,

$$P_{\varepsilon}[\overline{X}^{\varepsilon} > M] = P_{\varepsilon}[T_M < T_0] = \frac{s(\varepsilon) - s(0)}{s(M) - s(0)} \tag{20}$$

where T_0, T_M are the hitting times of X_t^{ϵ} at 0 and M – see [15, Ch. 15, (3.1), (3.10)]. Using now the continuity of the scale function $s(\cdot)$ [15, Ch. 15, (3.5), (3.6)] (note that [15] denote the scale function by $S(\cdot)$) yields $\lim_{M \to r} s(M) = s(r_-) = \infty$ and the result.

(19) follows by a similar argument. Indeed, denote the deterministic and stochastic flows generated by the ODE (2) and SDE (1) (i.e. the solutions of these equations at time t that start at x at time s) by $\phi_{s,t}(x)$ and $\Phi_{s,t}(x)$, respectively, and put $\Phi^{\varepsilon} := \Phi_{t_c,t_1}(X^{\varepsilon}_{t_c})$ and $\phi^{\varepsilon} := \phi_{t_c,t_1}(X^{\varepsilon}_{t_c})$ for brevity and define $\delta^{\varepsilon} = \Phi^{\varepsilon} - \phi^{\varepsilon}$. For fixed ε and M, it holds that

$$\begin{aligned} \forall \delta > 0, P_{\varepsilon}[|\delta^{\varepsilon}| > \delta] &\le P_{\varepsilon}[\overline{X}^{\varepsilon}_{T^{\varepsilon}} \le M] P_{\varepsilon}[|\delta^{\varepsilon}| > \delta | \overline{X}^{\varepsilon}_{T^{\varepsilon}} \le M] + P_{\varepsilon}[\overline{X}^{\varepsilon}_{T^{\varepsilon}} > M] \\ &\le P_{\varepsilon}[\overline{X}^{\varepsilon}_{T^{\varepsilon}} \le M] P_{\varepsilon}[|\delta^{\varepsilon}| > \delta | \overline{X}^{\varepsilon}_{T^{\varepsilon}} \le M] + P_{\varepsilon}[\overline{X}^{\varepsilon} > M]. \end{aligned}$$

Letting now ε to 0 makes the first term go to 0 by (16), yielding

$$\forall M < r, \forall \delta > 0, \limsup_{\varepsilon \to 0} P_{\varepsilon}[|\delta^{\varepsilon}| > \delta] \le \lim_{\varepsilon \to 0} \frac{s(\varepsilon) - s(0)}{s(M) - s(0)} = 0$$

where we have used again the continuity of the scale function. □

Theorem 4. *The conclusions of Theorem 2 still hold under the assumptions of Lemma 3.*

Proof. Theorem 2 of [2] only uses the assumption $\|\sigma\|_{\infty} < \infty$ in establishing the unnecessarily strong result (16). Providing weaker conditions for the weaker but still sufficient result (19) establishes therefore our claim. □

Problem 1. Note that essential use of $s(0) > -\infty$ was made in (18). We conjecture however that a finer analysis will reveal that the result of Theorem 4 still holds whenever r is "repelling/unattracting", more precisely when it is natural unattracting or entrance, cf. Feller's classification of boundary points [15, Ch. XV].

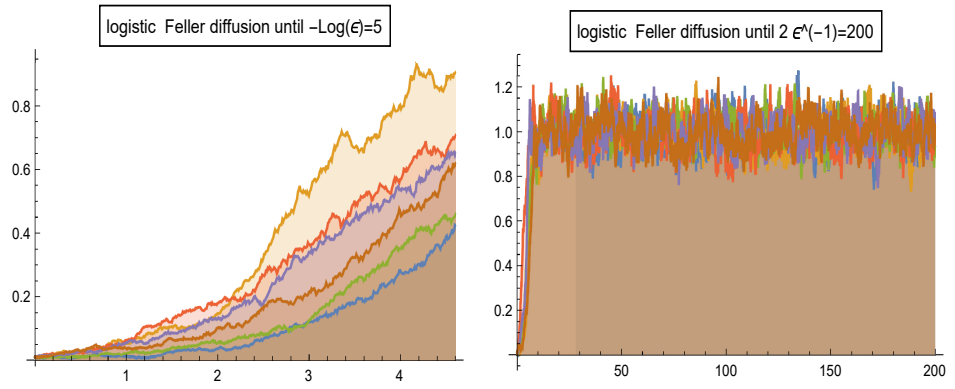


Figure 2: 6 paths of the logistic Feller diffusion ($x_c = 1$ is regular) with $\varepsilon = .01$, until T_ε and after

§3. Examples with $\lim_{t\to\infty} X_t/x_t = 0$: The logistic Feller and Gilpin-Ayala diffusions

We recall now some famous examples for which the conditions of our Lemma 3 hold. The logistic Feller diffusion is defined by

$$dX_t = \gamma X_t\left(1 - \frac{X_t}{x_c}\right)dt + \sqrt{\varepsilon X_t}dB_t,\ X_t \in (0,\infty).$$

The limit point x_c of x_t is a regular point for the diffusion; w.l.o.g. we will take it equal to 1. The scale density $s'(x) = e^{-\frac{2\gamma}{\varepsilon}(x-\frac{x^2}{2})}$ is integrable at 0, but not at ∞, and the speed density [15] $m'(x) = \frac{e^{\frac{2\gamma}{\varepsilon}(x-\frac{x^2}{2})}}{\varepsilon x}$ is integrable at ∞, but not at 0, so that the conditions of Lemma 3 hold.[§]

Therefore, fluid convergence with random initial point before T_ε [2] still holds, with the same deterministic flow and random initial condition as for the Kimura-Fisher Wright diffusion studied in [2]

$$\phi_t(x) = \frac{xe^{\gamma t}}{1 - x + xe^{\gamma t}},\ \widetilde{\phi}(x) = \frac{x}{1+x}, X_0 = \frac{W}{W+1}$$

(since $\mu(.), a'(0)$ did not change)–see Figure 2.

In fact, the paths of the logistic Feller and Kimura-Fisher-Wright diffusions are almost indistinguishable up to T^ε of each other –see Figure 3. After reaching the neighborhood of x_c however, the paths split, reflecting the different natures (regular and exit) of x_c for these two stochastic processes.

[§]Furthermore, conform Feller's boundary classification [15], 0 is an exit boundary since $s'(x)m[x,1]$ is integrable at 0, and absorbtion in 0 occurs with probability 1, and ∞ is an entrance (nonattracting) boundary, since $m'(x)s[1,x]$ is integrable at ∞–see also [7, 3] and [10] for the generalization to continuous-state branching processes with competition.

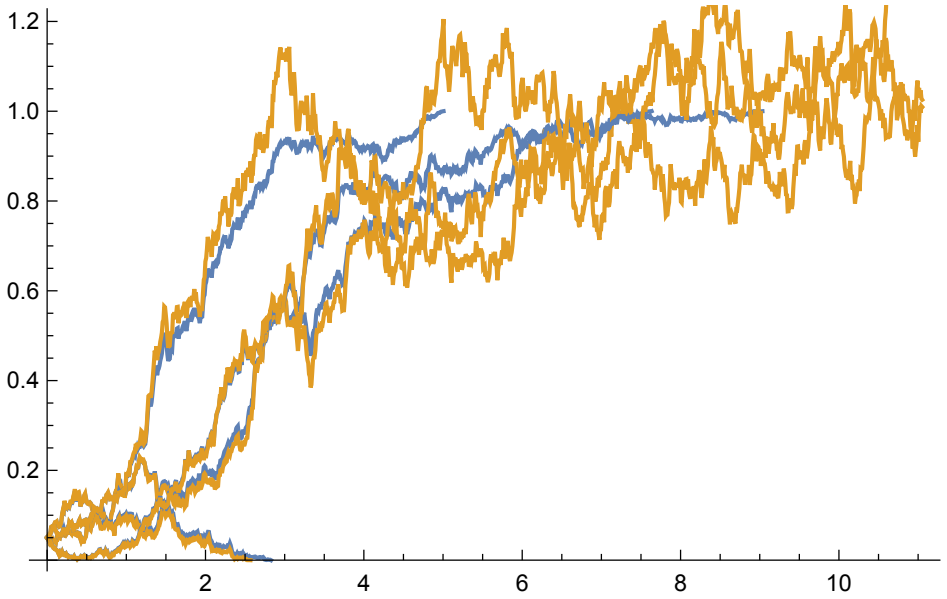


Figure 3: 6 paths of the logistic Feller and Kimura-Fisher-Wright diffusions with $\varepsilon = 1/20$, before and after T_ε

Some other examples of interest in population theory are the diffusion processes defined by the SDEs

$$dX_t = \gamma X_t\Big(1 - \big(\frac{X_t}{x_c}\big)^\theta\Big)dt + \sigma\sqrt{X_t}dB_t, \sigma >, \theta > 0,$$
$$dX_t = \Big[\gamma X_t\Big(1 - \frac{X_t}{x_c} - \beta\frac{X_t^{n-1}}{1+X_t^n}\Big)\Big]dt + \sigma\sqrt{X_t}dB_t, \beta \geq 0, n \geq 1,$$

which are stochastic extensions with square root volatility of deterministic population models introduced by Gilpin and Ayala and Holling respectively.

It is easy to check that adding the exponents θ and n does not affect integrability of the scale and speed densities of these diffusions, so that our extension applies. Furthermore, the rescaled flow $\widetilde{\phi}$ may be computed numerically by [2, Prop. 4.1] (and even symbolically for small integer values of θ, n).

Moving away from the square root volatility case, an interesting, still open question is to investigate whether analogues of the [2] result are available for the processes satisfying $dX_t = \gamma X_t\Big(1 - \big(\frac{X_t}{x_c}\big)^\theta\Big)dt + \sqrt{\varepsilon}(X_t)^\alpha dB_t, \quad \alpha > 0.$[§]

[§]The particular case $\alpha = \theta = 1$ is the famous Verlhurst-Pearl diffusion (VP)– see for example [17].

§4. Sketch of the proof of Theorem 2 [2]

Recall that $t_c = ct_1$ with $c \in (1/2, 1)$, arbitrary, and note that $X^\varepsilon_{T^\varepsilon} = \Phi_{t_c,t_1}(X^\varepsilon_{t_c}) = \Phi_{t_c,t_1}(\Phi_{t_c}(\varepsilon))$. The idea of the proof is to approximate this random variable by

$$X^\varepsilon_{T^\varepsilon} \approx \phi_{t_c,t_1}(\Phi_{t_c}(\varepsilon)) \xrightarrow{\varepsilon\to 0} \widetilde{\phi}(W), \tag{21}$$

with the random variable W from (8).

The proof of [2] involves several steps

1. The first idea for establishing the approximation $\widetilde{\phi}(W)$ of $X^\varepsilon_{T^\varepsilon}$ is to **blow-up** the process near the boundary 0
$$\widetilde{X}^\varepsilon_t := \varepsilon^{-1} X^\varepsilon_t,$$
which fixes the initial condition to 1 and changes the SDE to
$$d\widetilde{X}^\varepsilon_t = \varepsilon^{-1}\mu(\varepsilon\widetilde{X}^\varepsilon_t)dt + \sqrt{\frac{a(\varepsilon\widetilde{X}^\varepsilon_t)}{\varepsilon}}dB_t, \quad t \geq 0, \tag{22}$$
it is easy to check that a subsequent **linearization of the SDE** yields
$$\widetilde{X}^\varepsilon_t \approx Y_t$$
where Y_t is a **Feller branching diffusion** started from 1, defined by
$$Y_t = 1 + \int_0^t \mu'(0)Y_s ds + \int_0^t \sqrt{Y_s}dB_s, \quad t \geq 0. \tag{23}$$
One may take advantage then of the well-known nonnegative martingale convergence theorem for the "scaled final position" of the branching process Y_t
$$W := \lim_{t\to\infty} e^{-\mu'(0)t}Y_t. \tag{24}$$
Remark 6. Let us note that the linearization for processes satisfying $a(x) = O(x^2)$ and failing Assumption 2, like the linear Gilpin-Ayala (3), leads to geometric Brownian motion. In this case, (24) holds with $W = 0$, and a different approach seems necessary.
2. After "blowing up" the beginning of the path, the second idea is to **"look from far away"**. We want to break the trajectory at a suitably chosen time point
$$t_c < t_1 = T^\varepsilon = \frac{1}{\gamma}\log\frac{1}{\varepsilon} \tag{25}$$
such that before t_c, the original process is close to Feller's branching diffusion (23), and convergence to the limit W of the Feller diffusion occurs, i.e.
$$X^\varepsilon_{t_c} = \varepsilon\widetilde{X}^\varepsilon_{t_c} = e^{-\gamma t_1}\widetilde{X}^\varepsilon_{t_c} \approx e^{-\gamma t_1}Y_{t_c} = e^{-\gamma(t_1-\tau_c)}e^{-\gamma t_c}Y_{t_c} \approx e^{-\gamma(t_1-\tau_c)}W. \tag{26}$$
The first approximation $e^{-\gamma t_c}\widetilde{X}^\varepsilon_{t_c} \xrightarrow[\varepsilon\to 0]{L^1} Y_{t_c}$ follows from the following lemma [2] showing that the solution of (1) converges, under appropriate scaling, to the Feller branching diffusion (23).

Lemma 5. *Let* $\widetilde{X}_t^\varepsilon := \varepsilon^{-1} X_t^\varepsilon$, *where* X_t^ε *is the solution of* (1) *subject to* $X_0^\varepsilon = \varepsilon$. *Then*

$$\widetilde{X}_t^\varepsilon \xrightarrow[\varepsilon\to 0]{L^1} Y_t, \quad \forall\, t \geq 0,$$

where Y_t *is the solution of* (23).

Putting these together yields $\phi_{t_c,t_1}(X_{t_c}^\varepsilon) \xrightarrow[\varepsilon\to 0]{\P} \widetilde{\phi}(W)$.

3. The hardest part is proving that in the second portion $[t_c, t_1]$, the influence of the stochasticity is negligible, for example that $\Phi_{t_c,t_1}(X_{t_c}^\varepsilon) - \phi_{t_c,t_1}(X_{t_c}^\varepsilon) \xrightarrow[\varepsilon\to 0]{L^2} 0$, as proved in [2] under the restrictive assumption $\|\sigma\|_\infty < \infty$.

Putting it all together in one line, one must prove that

$$X_{t_1}^\varepsilon = \Phi_{t_c,t_1}(X_{t_c}^\varepsilon) \approx \Phi_{t_c,t_1}(We^{-\gamma(t_1-\tau_c)}) \approx \phi_{t_c,t_1}(We^{-\gamma(t_1-\tau_c)}) \xrightarrow[\varepsilon\to 0]{\P} \widetilde{\phi}(W). \tag{27}$$

To extend [2], it is sufficient to improve the third approximation step above.

Acknowledgements

We thank J.L. Perez for useful remarks and the referee for the help in improving the exposition.

References

[1] Alvarez, L., and Hening, A. Optimal sustainable harvesting of populations in random environments. *arXiv preprint arXiv:1807.02464* (2018).

[2] Baker, J., Chigansky, P., Hamza, K., and Klebaner, F. Persistence of small noise and random initial conditions in the wright-fisher model. *arXiv preprint arXiv:1802.06231* (2018).

[3] Bansaye, V., Collet, P., Martinez, S., Méléard, S., and Martin, J. S. Diffusions from infinity. *arXiv preprint arXiv:1711.08603* (2017).

[4] Barbour, A. D., Hamza, K., Kaspi, H., and Klebaner, F. C. Escape from the boundary in markov population processes. *Advances in Applied Probability 47*, 4 (2015), 1190–1211.

[5] Breiman, L. Probability, volume 7 of classics in applied mathematics. *Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA* (1992).

[6] Buldygin, V., Klesov, O., and Steinebach, J. Prv property and the ϕ-asymptotic behavior of solutions of stochastic differential equations. *Lithuanian Mathematical Journal 47*, 4 (2007), 361–378.

[7] Cattiaux, P., Collet, P., Lambert, A., Martínez, S., Méléard, S., and San Martín, J. Quasi-stationary distributions and diffusion models in population dynamics. *The Annals of Probability 37*, 5 (2009), 1926–1969.

[8] Cherny, A. S., and Engelbert, H.-J. *Singular stochastic differential equations*. No. 1858. Springer Science & Business Media, 2005.

[9] Evans, S. N., Hening, A., and Schreiber, S. J. Protected polymorphisms and evolutionary stability of patch-selection strategies in stochastic environments. *Journal of mathematical biology 71*, 2 (2015), 325–359.

[10] Foucart, C. Continuous-state branching processes with competition: duality and reflection at infinity. *arXiv preprint arXiv:1711.06827* (2017).

[11] Freidlin, M. I., and Wentzell, A. D. *Random perturbations of dynamical systems*. Springer, Heidelberg, third edition, translated from 1979 Russian original, 2012.

[12] Giet, J.-S., Vallois, P., and Wantz-Mézières, S. The logistic sde. *Theory of Stochastic Processes 20*, 1 (2015), 28–62.

[13] Helland, I. One-dimensional diffusion processes and their boundaries. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420* (1996).

[14] Karatzas, I., and Shreve, S. *Brownian motion and stochastic calculus*, vol. 113. Springer Science & Business Media, 2012.

[15] Karlin, S., and Tavarè, S. Linear birth and death processes with killing. *Journal of Applied Probability* (1982), 477–487.

[16] Keller, G., Kersting, G., and Rösler, U. On the asymptotic behaviour of first passage times for discussions. *Probability theory and related fields 77*, 3 (1988), 379–395.

[17] Liu, L., and Shen, Y. Sufficient and necessary conditions on the existence of stationary distribution and extinction for stochastic generalized logistic system. *Advances in Difference Equations 2015*, 1 (2015), 10.

[18] Pardoux, É. Probabilistic models of population evolution. *Mathematical Biosciences Institute Lecture Series. Stochastics in Biological Systems. Springer, Berlin* (2016).

[19] Skorokhod, A. V. *Asymptotic methods in the theory of stochastic differential equations*, vol. 78. American Mathematical Soc., 2009.

Florin Avram and Jacky Cresson
CNRS / UNIV PAU & PAYS ADOUR/LMAP - IPRA, UMR5142
64000, PAU, FRANCE
`Florin.Avram@orange.fr` and `jacky.cresson@univ-pau.fr`

Monografías Matemáticas *García de Galdeano* **42**, 25–34 (2019)

ADAPTIVE AUGMENTED MIXED FEM FOR THE OSEEN PROBLEM WITH MIXED BOUNDARY CONDITIONS

Tomás P. Barrios, José Manuel Cascón and María González

Abstract. We present an adaptive augmented dual-mixed method for the Oseen problem with mixed boundary conditions in the pseudostress-velocity variables. The new variational formulation and the corresponding Galerkin scheme are well-posed for appropriate values of the stabilization parameters. We provide the rate of convergence when each row of the pseudostress is approximated by Raviart-Thomas elements and the velocity is approximated by continuous piecewise polynomials. Moreover, we give an a posteriori error indicator and show the performance of the corresponding adaptive algorithm through a numerical example.

Keywords: Incompressible flow, Oseen, mixed finite element, stabilization, a posteriori error estimates.

AMS classification: 65N30, 65N12, 65N15.

§1. Introduction

The problem of computing the flow of a viscous and incompressible fluid at small Reynolds numbers is described by the Oseen equations. In the recent paper [4], we introduced a new augmented variational formulation for this problem in the pseudostress-velocity variables under homogeneous Dirichlet boundary conditions for the velocity, and developed a simple a posteriori error analysis.

Now, we propose a related method for the case when mixed boundary conditions are considered. We remark that the new method is not an extension of the one proposed in [4] since here the Dirichlet boundary condition is imposed weakly.

The paper is organized as follows. In Section 2 we describe a new augmented dual-mixed variational formulation for the Oseen problem in the pseudostress-velocity variables with mixed boundary conditions. Then, in Section 3 we analyze the stabilized mixed finite element method. In Section 4 we present an a posteriori error indicator that is reliable and locally efficient. Finally, numerical experiments are reported in Section 5.

§2. The augmented dual-mixed variational formulation

Assume that the fluid at hand occupies the region Ω, a polygonal domain in $\mathbb{R}^2$ with boundary Γ. We assume that $\Gamma = \Gamma_D \cup \Gamma_N$, where Γ_D is a closed part of Γ with positive measure and $\Gamma_N = \Gamma \setminus \Gamma_D$. Let $\nu > 0$ be the kinematic viscosity of the fluid, and let $\mathbf{a} \neq \mathbf{0}$ denote the advective velocity. We assume that $\mathbf{a}$ is solenoidal in Ω. Let $\mathbf{f}$ be an external body force, and denote by $\mathbf{u}_D$ a prescribed velocity on Γ_D and by $\mathbf{g}$ the Neumann data.

We consider the following Oseen problem: find the velocity field $\mathbf{u}$ and the pressure p such that

$$\begin{cases} -\nu\Delta\mathbf{u} + \mathbf{a}\cdot\nabla\mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega, \\ \operatorname{div}(\mathbf{u}) = 0 & \text{in } \Omega, \\ \mathbf{u} = \mathbf{u}_D & \text{on } \Gamma_D, \\ -p\,\mathbf{n} + \nu\dfrac{\partial\mathbf{u}}{\partial\mathbf{n}} = \mathbf{g} & \text{on } \Gamma_N, \end{cases} \tag{1}$$

where $\mathbf{n}$ is the unit outward normal to Γ_N.

Let $\mathbf{I}$ be the identity matrix in $\mathbb{R}^{2\times 2}$ and denote by $\boldsymbol{\sigma} := \nu\nabla\mathbf{u} - p\,\mathbf{I}$ the pseudostress. Proceeding similarly as in [4], problem (1) can be stated equivalently in terms of $\boldsymbol{\sigma}$ and $\mathbf{u}$, and the pressure can be recovered as $p = -\frac{1}{2}\operatorname{tr}(\boldsymbol{\sigma})$.

For simplicity, we consider the following decomposition of $\boldsymbol{\sigma}$: $\boldsymbol{\sigma} = \boldsymbol{\sigma_0} + \boldsymbol{\sigma_g}$, with $\boldsymbol{\sigma_0}\mathbf{n} = \mathbf{0}$ and $\boldsymbol{\sigma_g}\mathbf{n} = \mathbf{g}$ on Γ_N. Moreover, given a tensor $\boldsymbol{\tau}$, we denote by $\boldsymbol{\tau}^{\mathrm{d}} := \boldsymbol{\tau} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\tau})\,\mathbf{I}$ the *deviator* of $\boldsymbol{\tau}$. Then, problem (1) is equivalent to the following problem:

$$\begin{cases} -\mathbf{div}(\boldsymbol{\sigma_0}) + \mathbf{a}\cdot\nabla\mathbf{u} = \tilde{\mathbf{f}} & \text{in } \Omega, \\ \dfrac{1}{\nu}\boldsymbol{\sigma}_{\mathbf{0}}^{\mathrm{d}} = \nabla\mathbf{u} + \boldsymbol{\zeta} & \text{in } \Omega, \\ \mathbf{u} = \mathbf{u}_D & \text{on } \Gamma_D, \\ \boldsymbol{\sigma_0}\mathbf{n} = \mathbf{0} & \text{on } \Gamma_N, \end{cases} \tag{2}$$

where $\tilde{\mathbf{f}} := \mathbf{f} + \mathbf{div}(\boldsymbol{\sigma_g})$ and $\boldsymbol{\zeta} := -\frac{1}{\nu}\boldsymbol{\sigma}_{\mathbf{g}}^{\mathrm{d}}$.

Throughout this paper, we will use the standard notations for Sobolev spaces and norms. In particular, we denote by $H(\mathbf{div},\Omega) := \{\boldsymbol{\tau} \in [L^2(\Omega)]^{2\times 2} : \mathbf{div}(\boldsymbol{\tau}) \in [L^2(\Omega)]^2\}$ and $\mathbf{H_0} := \{\boldsymbol{\tau} \in H(\mathbf{div},\Omega) : \boldsymbol{\tau}\mathbf{n} = \mathbf{0} \quad \text{on } \Gamma_N\}$.

Let us define now the bilinear forms $a : \mathbf{H_0}\times\mathbf{H_0} \to \mathbb{R}$, $b : [H^1(\Omega)]^2 \times \mathbf{H_0} \to \mathbb{R}$ and $c : [H^1(\Omega)]^2 \times [H^1(\Omega)]^2 \to \mathbb{R}$ as follows:

$$a(\boldsymbol{\sigma},\boldsymbol{\tau}) := \frac{1}{\nu}\int_\Omega \boldsymbol{\sigma}^{\mathrm{d}} : \boldsymbol{\tau}^{\mathrm{d}}, \qquad b(\mathbf{u},\boldsymbol{\tau}) := \int_\Omega \mathbf{u}\cdot\mathbf{div}(\boldsymbol{\tau}), \qquad c(\mathbf{u},\mathbf{v}) := \int_\Omega (\mathbf{a}\cdot\nabla\mathbf{u})\cdot\mathbf{v},$$

for any $\boldsymbol{\sigma},\boldsymbol{\tau} \in \mathbf{H}_0$ and $\mathbf{u},\mathbf{v} \in [H^1(\Omega)]^2$.

We also define the linear functionals $l : [L^2(\Omega)]^2 \to \mathbb{R}$ and $m : \mathbf{H_0} \to \mathbb{R}$ as follows:

$$l(\mathbf{v}) := -\int_\Omega \tilde{\mathbf{f}}\cdot\mathbf{v}, \quad \forall\,\mathbf{v} \in [L^2(\Omega)]^2,$$

$$m(\boldsymbol{\tau}) := \int_\Omega \boldsymbol{\zeta} : \boldsymbol{\tau} + \int_\Gamma \mathbf{u}_D\cdot\boldsymbol{\tau}\mathbf{n}, \quad \forall\,\boldsymbol{\tau} \in \mathbf{H_0}.$$

Then, we have the following dual-mixed variational formulation of problem (2): find $(\boldsymbol{\sigma}_0,\mathbf{u}) \in \mathbf{H_0}\times[H^1(\Omega)]^2$ such that

$$\begin{cases} a(\boldsymbol{\sigma}_0,\boldsymbol{\tau}) + b(\mathbf{u},\boldsymbol{\tau}) = m(\boldsymbol{\tau}), & \forall\,\boldsymbol{\tau} \in \mathbf{H_0}, \\ b(\mathbf{v},\boldsymbol{\sigma}_0) - c(\mathbf{u},\mathbf{v}) = l(\mathbf{v}), & \forall\,\mathbf{v} \in [H^1(\Omega)]^2. \end{cases} \tag{3}$$

We remark that the variational formulation (3) exhibits a generalized saddle-point structure, with a non-symmetric bilinear form $c(\cdot,\cdot)$. According to [5], to ensure that problem (3) has a unique solution, we require, among other conditions, that the bilinear form $a(\cdot,\cdot)$ be coercive on $\mathbf{H_0}$. However, it is well-known that $a(\cdot,\cdot)$ is coercive in the divergence free subspace of $\mathbf{H_0}$ (see, for instance, the proof of Theorem 2.3 in [6]) but not on $\mathbf{H_0}$. We also require that the bilinear form $b(\cdot,\cdot)$ satisfies an inf-sup condition in $\mathbf{H_0}\times[H^1(\Omega)]^2$. These facts motivated us to consider an augmented formulation of problem (2).

Combining ideas from [4] and [9], we subtract the second equation in (3) from the first one and then, add the following least-squares type terms, that arise from the equilibrium and constitutive equations in (2) and from the Dirichlet boundary condition:

$$\kappa_1\int_\Omega(\mathbf{div}(\boldsymbol{\sigma_0})-\mathbf{a}\cdot\nabla\mathbf{u})\cdot(\mathbf{div}(\boldsymbol{\tau})+\mathbf{a}\cdot\nabla\mathbf{v}) \;=\; -\kappa_1\int_\Omega\tilde{\mathbf{f}}\cdot(\mathbf{div}(\boldsymbol{\tau})+\mathbf{a}\cdot\nabla\mathbf{v})$$

$$\kappa_2\int_\Omega(\nabla\mathbf{u}-\frac{1}{\nu}\boldsymbol{\sigma}_{\mathbf{0}}^{\mathrm{d}}) \,:\, (\nabla\mathbf{v}+\frac{1}{\nu}\boldsymbol{\tau}^{\mathrm{d}}) \;=\; -\kappa_2\int_\Omega\boldsymbol{\zeta}:(\nabla\mathbf{v}+\frac{1}{\nu}\boldsymbol{\tau}^{\mathrm{d}}),$$

and

$$\kappa_3\int_{\Gamma_D}\mathbf{u}\cdot\mathbf{v}=\kappa_3\int_{\Gamma_D}\mathbf{u}_D\cdot\mathbf{v}$$

where $(\boldsymbol{\sigma_0},\mathbf{u})\in\mathbf{H_0}\times[H^1(\Omega)]^2$ is a solution of (2) and $(\boldsymbol{\tau},\mathbf{v})\in\mathbf{H_0}\times[H^1(\Omega)]^2$ is a test function. The parameters κ_1, κ_2 and κ_3 are positive constants to be chosen so that the augmented bilinear form

$$A((\boldsymbol{\sigma},\mathbf{u}),(\boldsymbol{\tau},\mathbf{v})) := \frac{1}{\nu}\int_\Omega\boldsymbol{\sigma}^{\mathrm{d}} \,:\, \boldsymbol{\tau}^{\mathrm{d}}+\int_\Omega\mathbf{u}\cdot\mathbf{div}(\boldsymbol{\tau})-\int_\Omega\mathbf{div}(\boldsymbol{\sigma})\cdot\mathbf{v}+\int_\Omega(\mathbf{a}\cdot\nabla\mathbf{u})\cdot\mathbf{v}$$

$$+\kappa_1\int_\Omega(\mathbf{div}(\boldsymbol{\sigma})-\mathbf{a}\cdot\nabla\mathbf{u})\cdot(\mathbf{div}(\boldsymbol{\tau})+\mathbf{a}\cdot\nabla\mathbf{v})+\kappa_2\int_\Omega(\nabla\mathbf{u}-\frac{1}{\nu}\boldsymbol{\sigma}^{\mathrm{d}}):(\nabla\mathbf{v}+\frac{1}{\nu}\boldsymbol{\tau}^{\mathrm{d}})+\kappa_3\int_{\Gamma_D}\mathbf{u}\cdot\mathbf{v}$$

be coercive in the whole space $\mathbf{H_0}\times[H^1(\Omega)]^2$.

Let us define the linear functional $F:\mathbf{H_0}\times[H^1(\Omega)]^2\to\mathbb{R}$ by

$$F(\boldsymbol{\tau},\mathbf{v}) := \int_\Omega\boldsymbol{\zeta}:\boldsymbol{\tau}+\int_{\Gamma_D}\mathbf{u}_D\cdot\boldsymbol{\tau}\mathbf{n}+\int_\Omega\tilde{\mathbf{f}}\cdot\mathbf{v}-\kappa_1\int_\Omega\tilde{\mathbf{f}}\cdot(\mathbf{div}(\boldsymbol{\tau})+\mathbf{a}\cdot\nabla\mathbf{v})$$

$$-\kappa_2\int_\Omega\boldsymbol{\zeta}:(\nabla\mathbf{v}+\frac{1}{\nu}\boldsymbol{\tau}^{\mathrm{d}})+\kappa_3\int_{\Gamma_D}\mathbf{u}_D\cdot\mathbf{v},\qquad\forall\,(\boldsymbol{\tau},\mathbf{v})\in\mathbf{H_0}\times[H^1(\Omega)]^2.$$

Then, the augmented variational formulation of problem (2) reads: find $(\boldsymbol{\sigma_0},\mathbf{u})\in\mathbf{H_0}\times[H^1(\Omega)]^2$ such that

$$A((\boldsymbol{\sigma_0},\mathbf{u}),(\boldsymbol{\tau},\mathbf{v}))=F(\boldsymbol{\tau},\mathbf{v}),\qquad\forall\,(\boldsymbol{\tau},\mathbf{v})\in\mathbf{H_0}\times[H^1(\Omega)]^2. \tag{4}$$

Remark 1. In case of homogeneous Dirichlet boundary conditions, that is, when $\Gamma_D=\Gamma$, $\Gamma_N=\emptyset$ and $\mathbf{u}_D=\mathbf{0}$ on Γ, we obtain the same linear functional F as in [4]. However, the variational formulation is not equivalent, since here we look for $\mathbf{u}\in[H^1(\Omega)]^2$.

In what follows, we assume that $\mathbf{a} \in [L^\infty(\Omega)]^2$, $\mathbf{a}\cdot\mathbf{n} \geq 0$ on Γ_N, and $\mathbf{f} \in [L^2(\Omega)]^2$. Moreover, we assume that

$$0 < \kappa_1 < \frac{\kappa_2}{2\,\|\mathbf{a}\|^2_{[L^\infty(\Omega)]^2}}\,, \qquad 0 < \kappa_2 < \nu\,, \qquad \text{and} \qquad \kappa_3 > \frac{1}{2}\,\|\mathbf{a}\cdot\mathbf{n}\|_{L^\infty(\Omega)}\,.$$

Then, there exists $C_{\mathtt{ell}} > 0$ such that

$$A((\boldsymbol{\tau},\mathbf{v}),(\boldsymbol{\tau},\mathbf{v})) \geq C_{\mathtt{ell}}\,\|(\boldsymbol{\tau},\mathbf{v})\|^2_{\mathbf{H_0}\times[H^1(\Omega)]^2}\,, \quad \forall\,(\boldsymbol{\tau},\mathbf{v}) \in \mathbf{H_0}\times[H^1(\Omega)]^2\,,$$

with

$$C_{\mathtt{ell}} = \min\Big(\frac{1}{\nu}\,(1-\frac{\kappa_2}{\nu})c_1, \frac{\kappa_1}{2}c_1, \frac{\kappa_1}{2}, (\kappa_2 - 2\,\kappa_1\,\|\mathbf{a}\|^2_{[L^\infty(\Omega)]^2})\,c_2, (\kappa_3 - \frac{1}{2}\|\mathbf{a}\cdot\mathbf{n}\|_{L^\infty(\Omega)})\,c_2\Big)\,,$$

where c_1 and c_2 are the positive constants in Lemma 3.1 in [2] and in Lemma 3.3 in [8], respectively.

Theorem 1. *Under the previous hypotheses, problem* (4) *has a unique solution* $(\boldsymbol{\sigma_0},\mathbf{u}) \in \mathbf{H_0}\times[H^1(\Omega)]^2$ *and*

$$\|(\boldsymbol{\sigma_0},\mathbf{u})\|_{\mathbf{H_0}\times[H^1(\Omega)]^2} \leq C^{-1}_{\mathtt{ell}}\,M\,(\|\mathbf{f}\|_{[L^2(\Omega)]^2} + \|\mathbf{u}_D\|_{[H^{1/2}(\Gamma_D)]^2} + \|\boldsymbol{\sigma}\mathbf{g}\|_{H(\mathbf{div};\Omega)})\,,$$

where $M := \max(1 + \kappa_1\,(1 + \sqrt{2}\,\|\mathbf{a}\|_{[L^\infty(\Omega)]^2}), \frac{1}{\nu}(1+\kappa_2(1+\frac{1}{\nu})), 1+\kappa_3)$.

Proof. It follows from the Lax-Milgram Lemma. □

§3. Augmented mixed finite element method

Let $\{\mathcal{T}_h\}_{h>0}$ be a family of shape-regular meshes of $\bar\Omega$ made up of triangles. We denote by h_T the diameter of an element $T \in \mathcal{T}_h$ and define $h := \max_{T\in\mathcal{T}_h} h_T$.

Let $\mathbf{H}_{\mathbf{0},h}$ and V_h be any finite element subspaces of $\mathbf{H_0}$ and $[H^1(\Omega)]^2$, respectively. Then, the Galerkin scheme associated to problem (4) reads: find $(\boldsymbol{\sigma}_{\mathbf{0},h},\mathbf{u}_h) \in \mathbf{H}_{\mathbf{0},h}\times V_h$ such that

$$A((\boldsymbol{\sigma}_{\mathbf{0},h},\mathbf{u}_h),(\boldsymbol{\tau}_h,\mathbf{v}_h)) = F(\boldsymbol{\tau}_h,\mathbf{v}_h)\,, \qquad \forall\,(\boldsymbol{\tau}_h,\mathbf{v}_h) \in \mathbf{H}_{\mathbf{0},h}\times V_h\,. \tag{5}$$

Under the same hypotheses as for the continuous problem (4), problem (5) has a unique solution $(\boldsymbol{\sigma}_{\mathbf{0},h},\mathbf{u}_h) \in \mathbf{H}_{\mathbf{0},h}\times V_h$. Moreover, there exists a constant $C_{\mathtt{Cea}} > 0$, independent of h, such that

$$\|(\boldsymbol{\sigma_0}-\boldsymbol{\sigma}_{\mathbf{0},h},\mathbf{u}-\mathbf{u}_h)\|_{\mathbf{H_0}\times[H^1_0(\Omega)]^2} \leq C_{\mathtt{Cea}} \inf_{(\boldsymbol{\tau}_h,\mathbf{v}_h)\in\mathbf{H}_{\mathbf{0},h}\times V_h} \|(\boldsymbol{\sigma_0}-\boldsymbol{\tau}_h,\mathbf{u}-\mathbf{v}_h)\|_{\mathbf{H_0}\times[H^1_0(\Omega)]^2}\,. \tag{6}$$

In order to establish a rate of convergence result, we consider specific finite element subspaces $\mathbf{H}_{\mathbf{0},h}$ and V_h. Hereafter, given $T \in \mathcal{T}_h$ and an integer $l \geq 0$, we denote by $\mathcal{P}_l(T)$ the space of polynomials of total degree at most l on T and, given an integer $r \geq 0$, we denote by $\mathcal{RT}_r(T)$ the local Raviart-Thomas space of order $r+1$ (cf. [12]),

$$\mathcal{RT}_r(T) := [\mathcal{P}_r(T)]^2 \oplus [\mathbf{x}]\mathcal{P}_r(T) \subset [\mathcal{P}_{r+1}(T)]^2\,,$$

where $\mathbf{x}$ is a generic vector of $\mathbb{R}^2$.

Let $r \geq 0$ and $m \geq 1$. Then, we let $\mathbf{H}_{\mathbf{0},h}$ be

$$\mathbf{H}_{\mathbf{0},h} := [\mathcal{RT}_r^{\mathtt{t}}]^2 = \Big\{\boldsymbol{\tau}_h \in \mathbf{H}_0 \,:\, \boldsymbol{\tau}_h\big|_T \in [\mathcal{RT}_r(T)^{\mathtt{t}}]^2, \quad \forall\, T \in \mathcal{T}_h\Big\},$$

and define

$$V_h := [\mathcal{L}_m]^2 = \Big\{\mathbf{v}_h \in [C(\overline{\Omega})]^2 \,:\, \mathbf{v}_h\big|_T \in [\mathcal{P}_m(T)]^2, \quad \forall\, T \in \mathcal{T}_h\Big\}.$$

The corresponding rate of convergence is given in the next theorem.

Theorem 2. *Assume $\boldsymbol{\sigma}_{\mathbf{0}} \in [H^t(\Omega)]^{2\times 2}$, $\mathbf{div}(\boldsymbol{\sigma}_{\mathbf{0}}) \in [H^t(\Omega)]^2$ and $\mathbf{u} \in [H^{t+1}(\Omega)]^2$. Then, there exists $C = O(C_{\mathtt{Cea}}) > 0$, independent of h, such that*

$$\|(\boldsymbol{\sigma}_{\mathbf{0}} - \boldsymbol{\sigma}_{\mathbf{0},h}, \mathbf{u} - \mathbf{u}_h)\|_{\mathbf{H}_{\mathbf{0}}\times[H^1(\Omega)]^2} \;\leq\; C\, h^{\alpha}\Big(\|\boldsymbol{\sigma}_{\mathbf{0}}\|_{[H^t(\Omega)]^{d\times d}} + \|\mathbf{div}(\boldsymbol{\sigma}_{\mathbf{0}})\|_{[H^t(\Omega)]^2} + \|\mathbf{u}\|_{[H^{t+1}(\Omega)]^2}\Big), \tag{7}$$

where $\alpha := \min\{t, m, r+1\}$.

Proof. It follows straightforwardly from inequality (6) and the approximation properties of the corresponding finite element subspaces. □

§4. A posteriori error analysis

The a posteriori error analysis of the Oseen equations is very important for the numerical solution of the stationary incompressible Navier-Stokes equations. The incompressibility condition and the presence of a non-selfadjoint operator in the momentum equations are the main difficulties to obtain a posteriori error estimates for the Oseen problem.

We let E_h be the set of all the edges induced by the triangulation $\mathcal{T}_h$ and write $E_h = E_I \cup E_{\Gamma_D} \cup E_{\Gamma_N}$, where $E_I := \{e \in E_h \,:\, e \subseteq \Omega\}$, $E_{\Gamma_D} := \{e \in E_h \,:\, e \subseteq \Gamma_D\}$ and $E_{\Gamma_N} := \{e \in E_h \,:\, e \subseteq \Gamma_N\}$. Also, for each edge $e \in E_h$, we denote by h_e the length of edge e and fix a unit normal vector $\mathbf{n}_e := (n_1, n_2)^{\mathtt{t}}$; finally, we let $\mathbf{t}_e := (-n_2, n_1)^{\mathtt{t}}$ be the corresponding fixed unit tangential vector along e.

We define the local a posteriori error indicator

$$\theta_T^2 := \|\tilde{\mathbf{f}} + \mathbf{div}(\boldsymbol{\sigma}_{0,h}) - \mathbf{a}\cdot\nabla\mathbf{u}_h\|^2_{[L^2(T)]^2} + \|\boldsymbol{\zeta} + \nabla\mathbf{u}_h - \frac{1}{\nu}\,\boldsymbol{\sigma}^{\mathtt{d}}_{0,h}\|^2_{[L^2(T)]^{d\times d}}$$

$$+ \sum_{e\in E_{\Gamma_D}\cap\partial T} h_e\Big(\|\mathbf{u}_D - \mathbf{u}_h\|^2_{[L^2(e)]^2} \;+\; \|\nabla(\mathbf{u}_D - \mathbf{u}_h)\mathbf{t}_e\|^2_{[L^2(e)]^2}\Big)$$

and the global a posteriori error indicator

$$\theta := \Big(\sum_{T\in\mathcal{T}_h} \theta_T^2\Big)^{1/2}$$

The following theorem establishes the reliability of the a posteriori error indicator.

Theorem 3. *Assume* $\mathbf{u}_D \in [H^1(\Gamma_D)]^2$. *Then, there exists* $C_{\mathtt{rel}} > 0$, *independent of h, such that*

$$\|(\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_{0,h}, \mathbf{u} - \mathbf{u}_h)\|_{\mathbf{H}_0 \times [H^1(\Omega)]^2} \leq C_{\mathtt{rel}}\, \theta$$

Proof. We proceed as in [4] to bound the error in terms of residuals, but use a quasi-Helmholtz decomposition [7] instead of the usual Helmholtz decomposition. □

The next theorem establishes the local efficiency of the a posteriori error indicator.

Theorem 4. *Assume* $\mathbf{u}_D \in [H^1(\Gamma)]^2$ *is component-piecewise polynomial on* Γ_D. *Then, there exists* $C_{\mathtt{eff}} = C(\nu, \kappa_1, \kappa_2, , \kappa_3, \mathbf{a}) > 0$, *independent of h, such that for all* $T \in \mathcal{T}_h$ *we have*

$$C_{\mathtt{eff}}\, \theta_T \;\leq\; \|(\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_{0,h}, \mathbf{u} - \mathbf{u}_h)\|_{\mathbf{H}_0(T) \times [H^1(T)]^2} \qquad \forall\, T \in \mathcal{T}_h$$

Proof. We proceed with the first two terms of θ_T as usual. The second term is bounded using a discrete trace inequality [1, Theorem 3.10]. Finally, the last term is bounded similarly as in Lemma 3.9 in [3]. □

§5. Numerical experiments

We performed numerical experiments with the finite spaces $\mathbf{H}_{0,h}$ and V_h defined in Section 3, with $r = 0$ and $m = 1$. We implemented the standard adaptive finite element method (AFEM) based on the loop

$$\text{SOLVE} \;\rightarrow\; \text{ESTIMATE} \;\rightarrow\; \text{MARK} \;\rightarrow\; \text{REFINE}$$

(see, for instance, [11]). For the numerical experiments, we used the finite element toolbox ALBERTA [13]. This toolbox employs the Kossaczky refinement algoritm, that uses recursive bisection [10]. The corresponding linear systems are solved using MATLAB (UMFPACK).

We consider an example in which $\Omega = (0,1) \times (0,1)$ is the unit square, $\Gamma_N = \{0\} \times [0,1]$ and $\Gamma_D = \Gamma \setminus \Gamma_N$. We take the kinematic viscosity $\nu = 1$ and the advective velocity $\mathbf{a} = (1,0)$. Then, we let

$$\phi(x,y) = 10x^2y^2(1-y)^2 \tanh\left(100(x - \frac{1}{2})\right),$$

and choose $\mathbf{f}$ and $\mathbf{u}_D$ so that the exact solution is

$$\mathbf{u} = \mathbf{curl}\ \phi = \left(\frac{\partial \phi}{\partial y}, -\frac{\partial \phi}{\partial x}\right), \qquad p(x,y) = \exp\left(-(x - \frac{1}{2})^2\right).$$

We remark that the velocity $\mathbf{u}$ exhibits an inner layer around the line $x = \frac{1}{2}$.

In Figure 1 we show the individual errors in the velocity and the pseudostress for the uniform (U) and adaptive (A) refinements with respect to the number of degrees of freedom (DOFs). We can observe that the adaptive refinement performs better than the uniform refinement. In Figure 2 we show the total error and the estimator vs. the DOFs for the uniform and adaptive refinements. In this case, we can observe that the estimator fits the total error. Accordingly, in this example the efficiency indices are almost one for both refinements (see Figure 3).

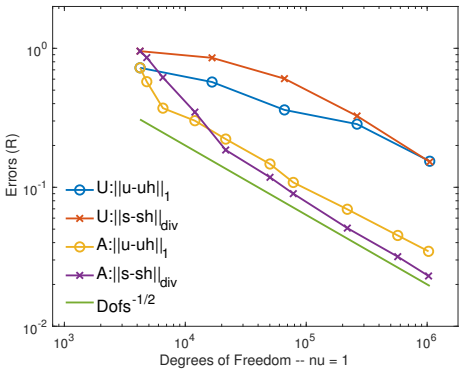


Figure 1: Individual errors in the velocity and the pseudostress.

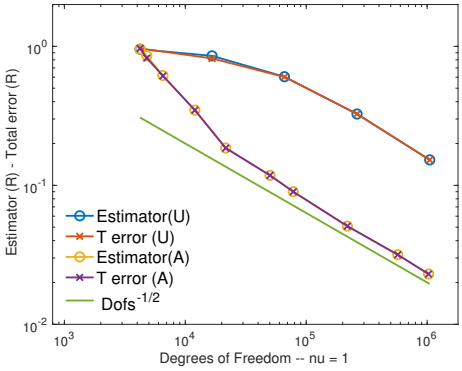


Figure 2: Total error and estimator vs. the DOFs.

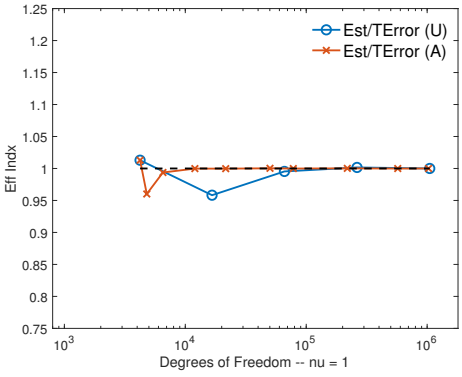


Figure 3: Efficiency indices for the uniform and adaptive refinements.

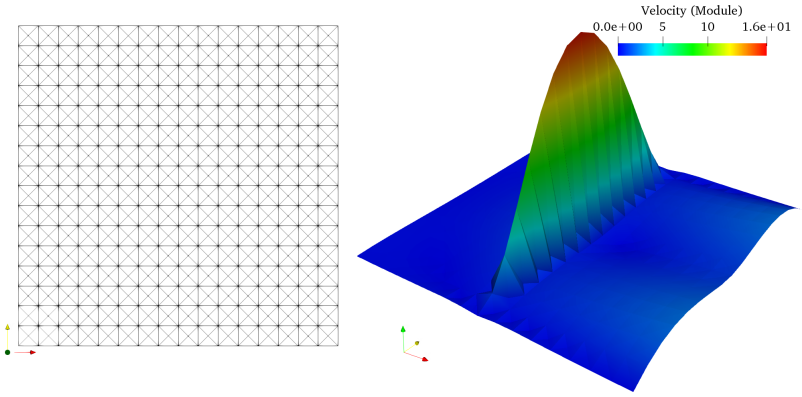


Figure 4: Initial mesh and corresponding velocity module.

In Figures 4-6 we show, respectively, the initial mesh, an intermediary mesh and the final mesh (after 8 iterations) obtained with the AFEM algorithm, together with the corresponding velocity modules. We can observe that the AFEM algorithm is able to locate the inner layer of the solution, since the refinement is essentially concentrated around the line $x = \frac{1}{2}$.

Acknowledgements

The research of T.P. Barrios is partially supported by Dirección de Investigación of the Universidad Católica de la Santísima Concepción (Chile) and through CONICYT-Chile FONDECYT project 1160578. The research of J.M. Cascón is partially supported by the Conserjería de Educación of the Junta de Castilla y León, Grant SA020U16. The research of M. González is partially supported by the Spanish Ministerio de Economía y Competitividad Grant MTM2016-76497-R.

References

[1] Agmon, S. *Lectures on Elliptic Boundary Value Problems*. Princeton, 1965.

[2] Arnold, D., Douglas, J., and Gupta, C. A family of higher order mixed finite element methods for plane elasticity. *Numer. Math. 45* (1984), 1–22.

[3] Barrios, T., Bustinza, R., García, G., and González, M. A posteriori error analyses of a velocity-pseudostress formulation of the generalized Stokes problem. *Journal of Computational and Applied Mathematics 357* (2019), 349–365.

[4] Barrios, T., Cascón, J., and González, M. Augmented mixed finite element method for the Oseen problem: A priori and a posteriori error analyses. *Comput. Methods Appl. Mech. Engrg. 313* (2017), 216–238.

[5] Brezzi, F., and Fortin, M. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, 1991.

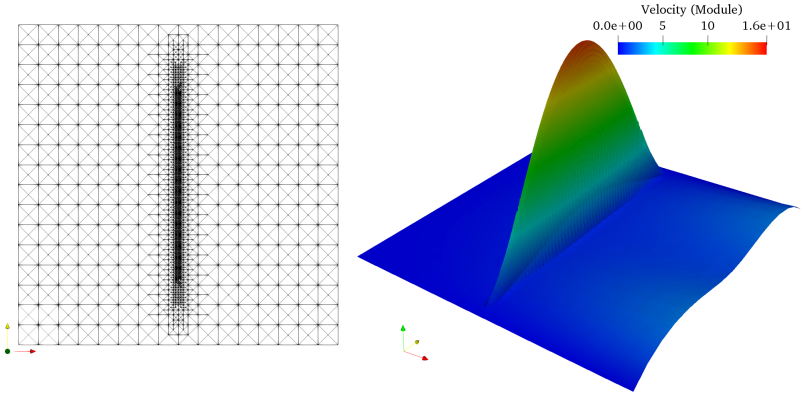


Figure 5: Mesh generated by the adaptive algorithm and velocity module at iteration 4.

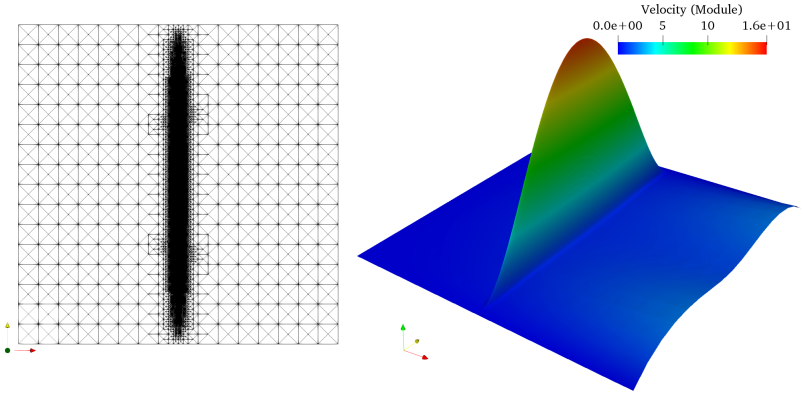


Figure 6: Mesh generated by the adaptive algorithm and velocity module at iteration 8.

[6] Cai, Z., Tong, C., Vassilevski, P., and Wang, C. Mixed finite element methods for incompressible flow: stationary Stokes equations. *Numer. Methods PDEs 26* (2010), 957–978.

[7] Cascón, J. M., Nochetto, R. H., and Siebert, K. G. Design and convergence of AFEM in H(div). *Mathematical Models and Methods in Applied Sciences 17* (2007), 1849–1881.

[8] Figueroa, L., Gatica, G. N., and Márquez, A. Augmented mixed finite element methods for the stationary Stokes equations. *SIAM Journal on Scientific Computing 31* (2008), 1082–1119.

[9] González, M. Stabilized dual-mixed method for the problem of linear elasticity with mixed boundary conditions. *Applied Mathematics Letters 30* (2014), 1–5.

[10] Kossaczký, I. A recursive approach to local mesh refinement in two and three dimensions. *J. Comput. Appl. Math. 55* (1994), 275–288.

[11] Morin, P., Nochetto, R., and Siebert, K. Convergence of adaptive finite element methods. *SIAM Review 44* (2002), 631–658.

[12] Roberts, J., and Thomas, J.-M. Mixed and hybrid methods. In *Handbook of Numerical Analysis* (Amsterdam, 1991), vol. II, North-Holland.

[13] Schmidt, A., and Siebert, K. G. Design of adaptive finite element software: The finite element toolboox ALBERTA. In *Lecture Notes in Computer Science and Engineering* (2005), vol. 42, Springer.

T. P. Barrios
Departamento de Matemática y Física Aplicadas
Universidad Católica de la Santísima Concepción
Casilla 297, Concepción, Chile
tomas@ucsc.cl

J. M. Cascón
Departamento de Economía e Historia Económica
Universidad de Salamanca
Salamanca, 37008, Spain
casbar@usal.es

M. González
Departamento de Matemáticas
Universidade da Coruña
Campus de Elviña s/n, 15071, A Coruña, Spain
maria.gonzalez.taboada@udc.es

Monografías Matemáticas *García de Galdeano* **42**, 35–44 (2019)

A TRIAXIAL MODEL FOR THE ROTO-ORBITAL COUPLING IN A BINARY SYSTEM

Antonio Cantero, Francisco Crespo and Sebastian Ferrer

Abstract. We study the roto-orbital dynamics of a uniform sphere and a triaxial body by means of a model which defines a 2-DOF Hamiltonian system using variables referred to the total angular momentum. The validity and applicability of our model is been assessed numerically. We present a classification of some relative equilibria, finding constant radius solutions filling 4-D and lower dimensional tori. These families of relative equilibria include some of the classical ones reported in the literature and some new types showing the triaxiality influence on both. For a number of scenarios the relation between the triaxiality and the inclination connected with relative equilibria are discussed and a full analysis in in progress [2].

Keywords: Roto-orbital dynamics, rigid body, relative equilibria, triaxiality..

§1. Introduction

We study a 2-DOF Hamiltonian model for the roto-orbital dynamics of a general binary system made of two rigid bodies $\mathcal{B}_1$ and $\mathcal{B}_2$, with masses m_1 and m_2 respectively. This problem is known as the full gravitational 2-body problem (FG2BP)[10] and usually is approximated by means of the MacCullagh's truncation [8], which is the second non vanishing term of the expansion of the potential energy. That is to say, the associated Hamiltonian with the FG2BP is obtained out of the sum of the rotational and orbital kinetic energies plus the potential energy, which is computed as a series expansion in Legendre polynomials. The first step in this expansion leaves us with a maximally super-integrable model, the Kepler plus the free rigid body. Nevertheless, accuracy increasing demands ask for a more realistic model. With this purpose, the usual procedure is adding the following term (MacCullagh's term) of the potential expansion, leading us to a non-integrable system in many degrees of freedom, which involves an extraordinary complexity. The main idea of this communication is to present a halfway model between these two extremes. The interest of our model is twofold. On the one hand, it allows us to identify special solutions that could become nominal trajectories in missions design whereas it alleviates usual heavy computations. On the other hand, it can be used to build a perturbation theory based on a new unperturbed part avoiding the degenerate character inherent to the classical superintegrable models. In other words, a first order perturbed solution based on this model might be accurate enough for tracking purposes. The benefits of a similar approach are now seen in areas such as the relative motion in formation flights [7].

In a series of previous works and with the same idea in mind, the authors have presented and analysed 1-DOF models [4, 3, 1]. In this work, we consider a 2-DOF model.

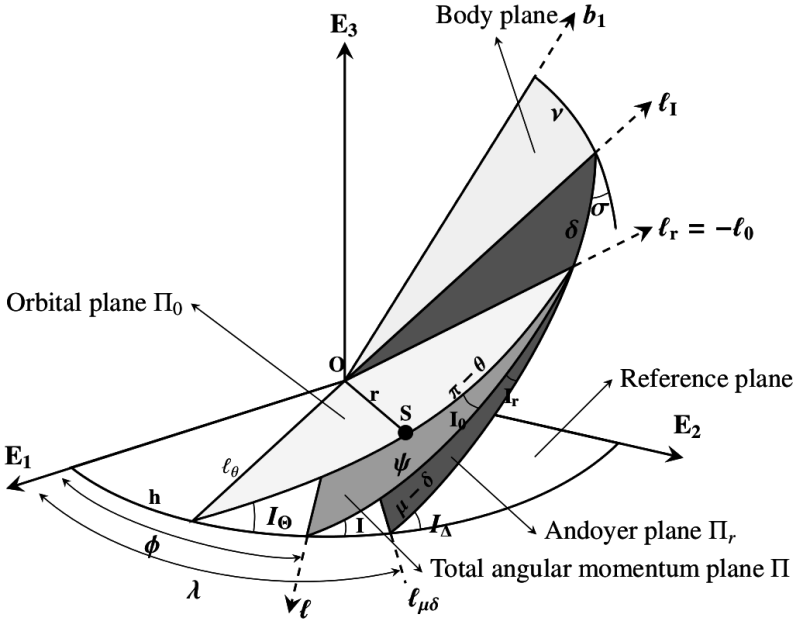


Figure 1: Geometry of the variables $(r, \phi, \psi, \theta, \delta, \nu, R, \Phi, \Psi, \Theta, \Delta, N)$. The variable r and the angles are explicitly given in the figure, while the associated momenta are included implicitly through the inclinations of the planes. The conjugate variable R remains unrepresented because of its pure dynamical sense. Note that this figure appeared first in [6]

.

§2. Variables

The variables in which the problem is posed may have a significant impact on its treatment. Our choice is the use of the total angular momentum as the key object to define them, which application for the roto-translatory problem was first introduced in [6] as a result of the application of the elimination of the nodes in the n-body problem [5] to the roto-translatory model. Nevertheless, quoting Meyer [9] "there is a saying in celestial mechanics that no set of coordinates is good enough". This claim highlights that in every choice of variables, a sacrifice must be done. More precisely, Cartesian variables have a simple formulation, but they do not take advance of the presence of symmetries. Conversely, by using variables referred to the total angular momentum, we incorporate the angles associated to the symmetries allowing for compact expressions and intuitive geometric insight of the relative equilibria. However, this is done at the expenses of having singularities, *i. e.* a global study of the system requires the use of several charts.

A complete set of canonical variables related with the angular momentum planes are used here denoted by $(r, \phi, \psi, \theta, \delta, \nu, R, \Phi, \Psi, \Theta, \Delta, N)$. We are not going to provide a complete derivation of them, which may be found in [6]. Instead and with the aim of fixing notation,

we provide the geometric meaning of the angles by means of Figure 1 and briefly recall the definition of the canonical angles by following [4]: Let us consider the reference frame $S^* = (\boldsymbol{\ell},\ \mathbf{n}\times\boldsymbol{\ell},\ \mathbf{n})$, where $\boldsymbol{\ell}$ is the unitary vector defining node of the total angular momentum plane with the horizontal spatial plane and $\mathbf{n}$ is the unitary vector pointing in the direction of the total angular momentum. In addition, $S^E = \{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3\}$ and $S^b = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ are the spatial and body frames respectively, where $\mathbf{b}_i$ corresponds with the principal moment of inertia of $\mathcal{B}_1$. The orientation and center of mass of the body are referred to the new frame by means of $(r, \phi, \psi, \theta, \delta, \nu)$, see Figure 1. These angles are determined by the nodes; $\boldsymbol{\ell}_{\mu\delta}$ defined by the rotational angular momentum and the spatial plane, $\boldsymbol{\ell}_r = -\boldsymbol{\ell}_o$ given by the intersection of the total, rotational and orbital angular momentum planes and $\boldsymbol{\ell}_\theta$ generated by the orbital and spatial planes intersection. Precisely, we have that $\phi = (\widehat{\mathbf{E}_1, \boldsymbol{\ell}})$, $\psi = (\widehat{\boldsymbol{\ell}, \boldsymbol{\ell}_I})$, $\theta = (\widehat{\boldsymbol{\ell}_o, \mathbf{r}})$, $\delta = (\widehat{\boldsymbol{\ell}_r, \boldsymbol{\ell}_I})$ and $\nu = (\widehat{\boldsymbol{\ell}_I, \mathbf{b}_1})$. Moreover, there are three more auxiliary angles which are not among the canonical variables but we will use them later on; $\lambda = (\widehat{\mathbf{E}_1, \boldsymbol{\ell}})$, $\mu = (\widehat{\boldsymbol{\ell}_{\mu\delta}, \boldsymbol{\ell}_I})$, $\sigma = (\widehat{\Pi_r, \Pi_b})$ and $h = (\widehat{\mathbf{E}_1, \boldsymbol{\ell}_\theta})$. In addition, the conjugate momenta of the variables read as follows

$$R, \quad \Phi = \mathbf{G}\cdot\mathbf{E}_3, \quad \Psi = \mathbf{G}\cdot\mathbf{n} = G, \quad \Theta = G_o, \quad \Delta = G_r, \quad N = \mathbf{G}_r\cdot\mathbf{b}_3,$$

where $\mathbf{G}$ is the total angular momentum vector, $\mathbf{G}_r$ is the angular momentum of the secondary body in the body frame and $\mathbf{G}_o$ is the orbital angular momentum. Thus, we have the following interpretation of the momenta: (R) Radial velocity of the center of mass. (Φ) Third component of the total angular momentum in space frame. (Ψ) Magnitude of the total angular momentum. (Θ) Magnitude of the angular momentum of the center of mass. (Δ) Magnitude of the angular momentum of the rigid body. (N) Third component of the angular momentum of the rigid body in the body frame (principal axes of inertia).

§3. Hamiltonian formulation of the triaxial model

The formulation of the triaxial model follows the same derivation as the one made in Crespo *et al.* [4], which is based in six simplifying assumptions. More specifically, the following set of simplifications are assumed in order to define our modelization: *(i) Barycentric coordinates.* The inertial frame is chosen to be moving with the total center of mass. *(ii) Shape and mass distribution of* $\mathcal{B}_2$. The main body $\mathcal{B}_2$ (mass m_2) is endowed with spherical symmetry. *(iii) Size ratios.* Dimensions of the secondary body $\mathcal{B}_1$ are small when compared to the distance between the centers of mass of the two bodies. *(iv) Shape and mass distribution of* $\mathcal{B}_1$. The secondary body may be approximated by an homogeneous triaxial ellipsoid with total mass m_1. *(v) Eccentricity.* Only small eccentricity orbits are considered. *(vi) Resonances.* The case of spin-orbit resonances is not considered.

The Hamiltonian of the roto-orbital model is obtained from the mechanic energy function. Thus, denoting T_O, T_R the orbital and rotational kinetic energies and $\mathcal{P}$ the potential, the Hamiltonian function is defined in the cotangent bundle of the special Euclidean group $T^*SE(3)$

$$\mathcal{H} = T_O + T_R + \mathcal{P} = T_O + T_R - \frac{\mathcal{G}M}{r} + \mathcal{V} = \mathcal{H}_K + \mathcal{H}_R + \mathcal{V},$$

in other words, the potential is usually split in two parts: a term which depends only on $1/r$ and $\mathcal{V}$, called the perturbing potential, depending on the rest of the variables of the problem. As a result, we have that $\mathcal{H}_K = T_O - \mathcal{G}M/r$ is the Keplerian part of the system, where $\mathcal{G}$ is the gravitational constant and $\mathcal{H}_R = T_R$ is referred as the Euler system (or the free rigid body).More explicitly, we obtain the following expression for $\mathcal{H}$ in the $\mathcal{B}_1$-body frame

$$\mathcal{H}(\mathbf{r}, \mathbf{A}, \mathbf{p}, \mathbf{\Pi}) = \frac{|\mathbf{p}|^2}{2m} + \frac{1}{2}\,\mathbf{\Pi}\cdot\mathbf{I}^{-1}\cdot\mathbf{\Pi} - \mathcal{G}\,m_2 \int_{\mathcal{B}_1} \frac{dm_1(\mathbf{x_1})}{|\mathbf{r} - \mathbf{x_1}|},$$

where $m = m_1 m_2/(m_1 + m_2)$, $\mathbf{r}$ is the vector joining the center of mass of the bodies, $\mathbf{A}$ is the rotation matrix transforming a vector in the body-fixed frame into the inertial frame and $\mathbf{p}$ and $\mathbf{\Pi}$ are the linear and angular momenta. In addition, the assumption *(iii)* allows us to consider the approximation of the gravitational potential $\mathcal{P}$ given by $-\mathcal{G}M/r$ and the MacCullagh's term [8]

$$U = -\frac{\kappa\, m}{2\, m_1\, r^3}\left[(A_3 - A_2)(1 - 3\gamma_3^2) - (A_2 - A_1)(1 - 3\gamma_1^2)\right], \tag{1}$$

where $\kappa = \mathcal{G}M$, $M = m_1 + m_2$ is the total mass of the system, $A_1 \le A_2 \le A_3$ are the principal moments of inertia associated to the secondary body and $(\gamma_1, \gamma_2, \gamma_3)$ are the director cosines of $\mathbf{r}$.

The direction cosines appearing in (1) may be expressed in the body frame by means of the following composition of rotations:

$$\gamma = R_3(\nu)\, R_1(\sigma)\, R_3(\delta)\, R_1(\iota)\, R_3(\pi - \theta)\ \mathbf{e}_1$$

where $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ and $\mathbf{e}_1 = (1, 0, 0)$. Finally, taking into account that $\gamma_1^2 + \gamma_2^2 + \gamma_3^2 = 1$ and after some calculations, we are allowed to express the MacCullagh's term (1) as follows

$$U = \frac{\kappa m}{32 m_1 r^3}\left[(2A_3 - A_2 - A_1)V_1 + \frac{3}{2}(A_2 - A_1)V_2\right], \tag{2}$$

where

$$\begin{aligned} V_1 \;=\; & -2(1 - 3c_\iota^2)(1 - 3c_\sigma^2) \\ & -3s_\sigma^2\Big[(1 - c_\iota)^2 C_{2,2,0} + (1 + c_\iota)^2 C_{-2,2,0}\Big] \\ & -6s_\iota^2\Big[s_\sigma^2 C_{0,2,0} - (1 - 3c_\sigma^2)C_{2,0,0}\Big] \\ & +12c_\sigma s_\iota s_\sigma\Big[(1 - c_\iota)C_{2,1,0} + 2c_\iota C_{0,1,0} - (1 + c_\iota)C_{-2,1,0}\Big] \end{aligned} \tag{3}$$

which is independent of ν, and V_2, the "triaxiality part" given by

$$\begin{aligned} V_2 \;=\; & -(1 - c_\sigma)^2\Big[(1 - c_\iota)^2 C_{2,2,-2} + (1 + c_\iota)^2 C_{-2,2,-2} + 2s_\iota^2 C_{0,2,-2}\Big] \\ & -(1 + c_\sigma)^2\Big[(1 - c_\iota)^2 C_{2,2,2} + (1 + c_\iota)^2 C_{-2,2,2} + 2s_\iota^2 C_{0,2,2}\Big] \\ & -6s_\iota^2 s_\sigma^2\Big[C_{2,0,2} + C_{2,0,-2}\Big] + 4s_\sigma^2(1 - 3c_\iota^2)C_{0,0,2} \\ & +4s_\iota s_\sigma(1 - c_\sigma)\Big[(1 - c_\iota)C_{2,1,-2} + 2c_\iota C_{0,1,-2} - (1 + c_\iota)C_{-2,1,-2}\Big] \\ & +4s_\iota s_\sigma(1 + c_\sigma)\Big[-(1 - c_\iota)C_{2,1,2} - 2c_\iota C_{0,1,2} + (1 + c_\iota)C_{-2,1,2}\Big], \end{aligned} \tag{4}$$

and the notation has been abbreviated by writing $C_{i,j,k} \equiv \cos(i\theta + j\delta + k\nu)$ and $c_x \equiv \cos x$ and $s_x \equiv \sin x$.

3.1. A model for roto-orbital dynamics.

Facing a non-integrable Hamiltonian system of 4-DOF requires the development of a perturbation theory. A usual way to proceed is to expand the Hamiltonian function in power series and truncate it at a certain order; this procedure gives in general an approximation. However a different approach to the problem is based in a simplification of the original Hamiltonian considering a related Hamiltonian of less degrees of freedom. In fact in this search for a simplified model a radial of 2 separable DOF has been proposed. Indeed, in [4], the authors proposed an axis-symmetric integrable model, whose accuracy was tested by comparing with the MacCullagh's truncation and showing a good performance in the numerical experiments. Here, we continue this previous study by investigating a triaxial case. One of our aims is to analyze the physical-parametric families of relative equilibria asociated. Keeping this motivation in mind, we propose our model following exactly the same procedure than in [4], except for the triaxial parameter. That is to say, we only take into account the first line of V_1 in (3) and in V_2 (4) the only terms that depends exclusively on ν . Then, the perturbing potential of the model is given by

$$\mathcal{V} = \frac{\kappa m}{32 m_1 r^3}\left[-2(2A_3 - A_2 - A_1)(1 - 3c_\iota^2)(1 - 3c_\sigma^2) + \frac{3}{2}(A_2 - A_1)4s_\sigma^2(1 - 3c_\iota^2)\cos(2\nu)\right],$$

which leads us to the final expression of the model Hamiltonian

$$\begin{aligned}\mathcal{H} = {} & \frac{1}{2}\left(R^2 + \frac{\Theta^2}{r^2}\right) - \frac{\kappa}{r} + \frac{q}{2}\left[\left(\frac{\sin^2(\nu)}{A_1} + \frac{\cos^2(\nu)}{A_2}\right)(\Delta^2 - N^2) + \frac{1}{A_3}N^2\right] \\ & - \frac{\kappa(1 - 3c_\iota^2)}{16 r^3}\left[(2A_3 - A_2 - A_1)(1 - 3c_\sigma^2) + 3(A_1 - A_2)s_\sigma^2\cos(2\nu)\right],\end{aligned} \tag{5}$$

where $q = m/m_1$. Furthermore, with the aim of alleviate formulas, we have considered the Hamiltonian per unit of mass by scaling the system and inertia momenta as follows:

$$\begin{aligned}&\mathcal{H}' = \mathcal{H}/m; \quad R' = R/m; \quad \Theta' = \Theta/m; \quad \Delta' = \Delta/m; \quad N' = N/m; \quad \Psi' = \Psi/m; \\ &\Phi' = \Phi/m; \quad A_1' = A_1/m_1; \quad A_2' = A_2/m_1; \quad A_3' = A_3/m_1.\end{aligned} \tag{6}$$

Nevertheless, for the sake of simplicity, we keep the original notation without primes on the variables. Then, the 2-DOF Hamiltonian system of differential equations associated with (5) is given by the following expressions:

$$\begin{aligned}\dot{r} &= R \\ \dot{R} &= \frac{\Theta^2}{r^3} - \frac{\kappa}{r^2} - \frac{3\kappa\left(1 - 3c_\iota^2\right)}{16 r^4}\left[\alpha\left(1 - 3c_\sigma^2\right) - 3(A_2 - A_1)(1 - c_\sigma^2)\cos(2\nu)\right] \\ \dot{\nu} &= q\left[\frac{1}{A_3} - \left(\frac{\sin^2(\nu)}{A_1} + \frac{\cos^2(\nu)}{A_2}\right) + \frac{3\kappa}{8\Delta^2 q r^3}\left(1 - 3c_\iota^2\right)(\alpha - (A_2 - A_1)\cos(2\nu))\right]N \\ \dot{N} &= q(A_1 - A_2)\left(1 - c_\sigma^2\right)\Delta^2\left[\frac{1}{2A_1A_2} - \frac{3\kappa\left(1 - 3c\iota^2\right)}{8\Delta^2 q r^3}\right]\sin(2\nu) \\ \dot{\theta} &= \frac{\Theta}{r^2} - \frac{3\kappa}{8r^3}\left(\frac{c_\iota}{\Delta} + \frac{c_\iota^2}{\Theta}\right)\left[\alpha\left(1 - 3c_\sigma^2\right) - 3(A_2 - A_1)(1 - c_\sigma^2)\cos(2\nu)\right]\end{aligned}$$

$$\dot{\psi} = \frac{3\kappa\Psi}{8r^3\Delta\Theta}\, c_\iota \left[\alpha\left(1-3c_\sigma^2\right) - 3(A_2-A_1)\left(1-c_\sigma^2\right)\cos(2\nu)\right]$$

$$\dot{\delta} = q\left[\left(\frac{\sin^2(\nu)}{A_1} + \frac{\cos^2(\nu)}{A_2}\right) - \frac{3\kappa}{8\Delta^2 qr^3}\left(1-3c_\iota^2\right)c_\sigma^2\left(\alpha-(A_2-A_1)\cos(2\nu)\right)\right.$$
$$\left. -\frac{3\kappa}{8\Delta^2 qr^3}\, c_\iota\left(c_\iota + \frac{\Delta}{\Theta}\right)\left(\alpha\left(1-3c_\sigma^2\right) - 3(A_2-A_1)\left(1-c_\sigma^2\right)\cos(2\nu)\right)\right]\Delta$$

where $\alpha = 2A_3 - A_2 - A_1$ together with the integrals $\dot{\phi} = \dot{\Phi} = \dot{\Theta} = \dot{\Psi} = \dot{\Delta} = 0$. In other words the 2-DOF system is made of the (r, ν) subsystem and three quadrature associated to θ, ψ and δ.

Note that, in general, a 2-DOF system is not integrable. Thus, in the triaxial case, the analytical integration is not provided and the integrability of the system remains as an open question, which is not in the scope of the present paper.

§4. Numerical assessment of our model.

We assess the validity of our model by carrying out a simulation comparing our model versus the MacCullagh's approximation [8]. The expansion of the gravitational potential truncated to the third term known as the MacCullagh's term is commonly used as a good approximation to the potential because considering the next term lead to expressions with r^5 in the denominator. For situations where the term with r^5 is required a new model should be provided. However this is out the scope of this paper.

Numerical simulations have been carried out by using the Mathematica 11 software [11] running on the platform macOS Sierra, 3.1 GHz Intel Core i5 (64-bit), 8 GB RAM.

There are several details to bear in mind through this section in order to proceed with the numerical experiment. Firstly, in what follows it is convenient to use the triaxiality parameter defined in [3] $\rho = (A_2-A_1)/(2A_3-A_2-A_1)$, noticing that due to the constrains of the principal moments of inertia $\rho \in (0,1)$. Secondly, we have considered the Hamiltonian per unit of mass and the canonical and inertia momenta have been scaled, see (6). Furthermore, we have changed internally the units for longitudes by choosing the radius of the spherical body R_p as the new one. However, we set these units back to Km when we present our results. Regarding the initial conditions, the radius and angles (radians) are given directly. In our simulations we consider the scenario of a massive spherical primary body and an arbitrary triaxial secondary body. More precisely, the two bodies are described as follows. Main body $\mathcal{B}_2$: a sphere with radius 500 Km and mean density $d = 2.8\, g/\mathrm{cm}^3$, and mass $m_2 = 1.47 \cdot 10^{21}$ Kg. Secondary body $\mathcal{B}_1$: an ellipsoid with mean density $d = 1.4\, g/\mathrm{cm}^3$ while the principal axes and the triaxiality parameter are: $A_1 = 1.069 \cdot 10^{21}$, $A_2 = 1.18 \cdot 10^{21}$, $A_3 = 1.28 \cdot 10^{21}$, $\rho = 0.353$. Initial distance between the center of masses is 2060 Km and we also assume the secondary body in a slow rotating regime. Solutions are evaluated for three orbital periods, see Figure 2, where we show the evolution of variables which are not constant for the model we are presenting.

We would like to highlight that, after three orbital periods, the differences between the slow variables r, R, ψ, θ, N are always in the order of thousandth or less. For the case of the fast variables δ, ν, we have a competitive performance for 4 hours, which represent 1/4 orbital periods, see Figure 3.

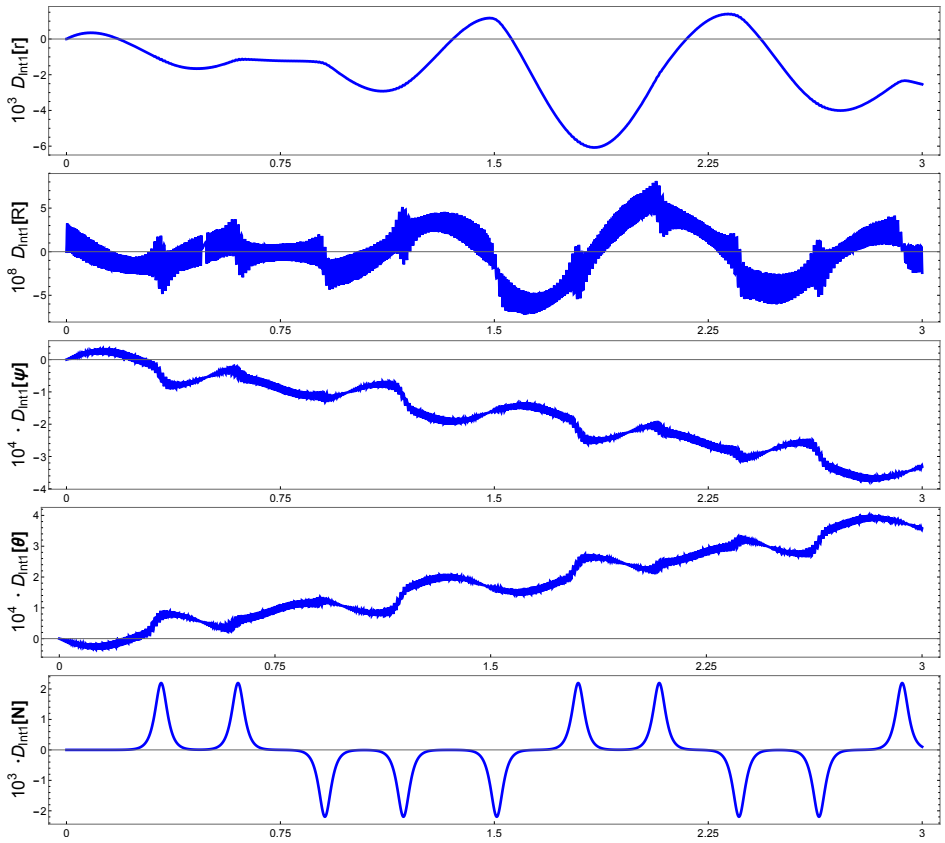


Figure 2: **Slow variables:** Differences between the 2-DOF model versus the MacCullagh's approximation. Abscissas are orbital periods and angles are given in radians. The orbital period is 16.5 hours and the rotation regime for each orbital period is 1-100.

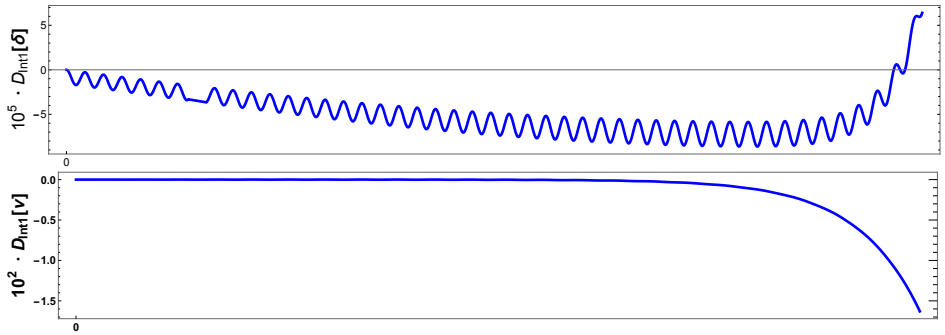


Figure 3: **Fast variables:** Differences between the 2-DOF model versus the MacCullagh's approximation. Abscissas represent 1/4 of the orbital period and angles are given in radians.

§5. Constant radius solutions. Some relative equilibria.

The system of differential equations defined by the Hamiltonian (5) is endowed with several distinguish and physical parameters. Thus, bifurcations occur in several directions in the parametric space [2].

With the aim of simplifying this scenario and provide a geometric interpretation of our equilibria, we organize our families of relative equilibria according to the inclinations of pairs of fundamental planes (orbital, rotational and body planes) due to the fact that the associated momenta of (θ), (ψ), (δ) and (ν) are included through the inclinations of the planes. More precisely, we consider the relative inclination between orbital and rotational planes (ι) and the one determined by the rotational and body planes (σ). For that reason $\cos\iota$ and $\cos\sigma$ are the key objects to present the analysis of the relative equilibria and allowed us to classify the relative equilibria on the following families: critical inclination equilibria when $(1-3c_\iota^2)=0$, body-inclined equilibria when $c_\sigma \neq 0$ and body-perpendicular equilibria when $c_\sigma = 0$. Each of these families of relative equilibria contains different orbits of constant radius filling different tori depending on the fixed angles. Note that $\rho = 1/3$ is equivalent to $A_3 - A_2 = A_2 - A_1$ leading to the maximum triaxiality case. Below we show two particular cases of these families of relative equilibria found in this problem [2].

Case 1: Body-Inclined equilibria $c_\sigma \neq 0$ with ν and ψ fixed.

This particular case shows a family of relative equilibria filling a 2-tori manifold $\mathbb{T}^2(\theta,\delta)$. On one hand the orbital variables behave as a keplerian "circular" orbit, however on the other hand the rotational part shows the triaxiality influence and introduce several novelties with respect to the classical scheme of the free rigid body. More precisely imposing the following initial conditions and relations between the momenta and physical parameters:

$$r=\frac{\Theta^2}{\kappa},\qquad R=0,\qquad c_\iota^2=\frac{1}{3}-\frac{4q\Theta^6\Delta^2}{9\kappa^4A_2A_3},\qquad \nu=0,\pi,\qquad c_\sigma^2=\frac{A_1-2A_2+A_3}{3(A_3-A_2)}$$

we get a relative equilibria with the following mean motions:

$$\dot\theta=\frac{\kappa^2}{\Theta^3}\qquad \dot\delta=q\Delta\left(\frac{-A_1+2(A_2+A_3)}{3A_2A_3}\right)$$

Note that the values $\nu=0,\pi$ are related to well-known equilibria of the free rigid body. It is worth noticing that in general $c_\sigma\neq 0$ which is a notorious difference from the classical case. Nevertheless for the particular value $\rho=1/3$ we get $c_\sigma=0$ and therefore we recover the Euler equilibria and obtain a simplified form of the mean motion $\dot\delta = q\Delta/A_2$

Case 2: Body-perpendicular equilibria $c_\sigma = 0$ with ν and ψ fixed.

This case shows also a relative equilibria filling a 2-tori manifold $\mathbb{T}^2(\theta,\delta)$ where the orbital variables behave as a keplerian "circular". As it happens with the previous case the rotational part shows a triaxiality influence and introduce several novelties with respect to the classical

scheme of the free rigid body. In particular imposing the following initial conditions and relations between the momenta and physical parameters:

$$r = \frac{\Theta^2}{\kappa}, \quad R = 0, \quad c_\iota^2 = \frac{1}{3} - \frac{4q\Theta^6\Delta^2}{9\kappa^4 A_1 A_2}, \quad \cos(2\nu) = \frac{2A_3 - A_1 - A_2}{3(A_2 - A_1)}, \quad c_\sigma = 0,$$

we get that a relative equilibria with the following mean motions:

$$\dot{\theta} = \frac{\kappa^2}{\Theta^3} \qquad \dot{\delta} = \frac{q\Delta}{3}\left(\frac{2A_1 + 2A_2 - A_3}{A_1 A_2}\right)$$

Note that for this relative equilibria being $c_\sigma = 0$ we get $\cos(2\nu) \neq 0$ which is also a difference from the classical. It is worth mentioning that for the particular value $\rho = 1/3$ we get $\cos(2\nu) = 1$ and $\dot{\delta} = \dfrac{q\Delta}{A_2}$ which is a particular relative equilibria of our model and it is work in progress [2].

Observe, on both cases shown, that conditions for periodic orbits are easily obtained since expression for mean motions are explicitly given.The reader should also take into account that bounds among the integrals and physical parameters have to be added to the formulas given above.

Acknowledgements

Support from Research Agencies of Spain and Chile is acknowledged. They came in the form of research projects MTM2015-64095-P and ESP2017-87271-P, of the Ministry of Science of Spain and from the project 11160224 of the Chilean national agency FONDECYT.

References

[1] A. Cantero, F. C., and Ferrer, S. The triaxiality role in the spin-orbit dynamics of a rigid body. *Applied Mathematics and Nonlinear Sciences 3*, 1 (May 2018), 187–208. doi:10.21042/AMNS.2018.1.00015.

[2] Cantero, A. *Mathematical Models for the Full Gravitational 2-Body Problem. A Perturbative Scheme by Stages.* PhD thesis, Universidad de Murcia, In progress.

[3] Crespo, F., and Ferrer, S. Roto-orbital dynamics of a triaxial rigid body around a sphere. Relative equilibria and stability. *Advances in Space Research 61*, 11 (2018), 2725 – 2739.

[4] Crespo, F., Molero, F. J., Ferrer, S., and Scheeres, D. J. A radial axial-symmetric intermediary model for the roto-orbital motion. *The Journal of the Astronautical Sciences 65*, 1 (Mar 2018), 1–28. Available from: https://doi.org/10.1007/s40295-017-0121-9, doi:10.1007/s40295-017-0121-9.

[5] Deprit, A. Elimination of the nodes in problem of N bodies. *Celestial Mechanics 30*, 2 (1983), 181–195.

[6] Ferrándiz, J. M., and Sansaturio, M. E. Elimination of the nodes when the satellite is a non spherical rigid body. *Celestial Mechanics 46* (1989), 307–320.

[7] Lara, M., and Gurfil, P. Integrable approximation of j_2-perturbed relative orbits. *Celestial Mechanics and Dynamical Astronomy 114* (2012), 229–254.

[8] MacCullagh, J. On the rotation of a solid body. *Proceedings of the Royal Irish Academy 2* (1840), 520–545.

[9] Meyer, K., Hall, G., and Offin, D. *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*. Springer, New York, 2009.

[10] Scheeres, D. *Orbital Motion in Strongly Perturbed Environments: Application to Asteroid, Comet and Planetary Satellite Orbiters*. Jointly published with Praxis Publishing, UK, 2012.

[11] Wolfram, S. *Mathematica 10.3.0.0, student edition*. Wolfram Research Inc./Cambridge Univ. Press, 2015.

A. Cantero and S. Ferrer
Universidad de Murcia
DITEC
Facultad de Informática
antonio.cantero@um.es and s.ferrer@um.es.

F. Crespo
GISDA
Dept. de Matemática
Facultad de Ciencias, Universidad del Bío-Bío
fcrespo@ubiobio.cl.

Monografías Matemáticas *García de Galdeano* **42**, 45–55 (2019)

STABILITY OF DOMAIN WALLS IN FERROMAGNETIC RINGS

Gilles Carbou, Mohand Moussaoui and Romeissa Rachi

Abstract. In this work we consider a one-dimensional model of ferromagnetic ring taking into account curvature and anisotropy effects. We describe all the planar static solutions representing domain walls and we study their stability.

Keywords: ferromagnetism, Landau-Lifshitz equation, stability, domain walls,... .
AMS classification: 35K55, 35Q60.

§1. Introduction

Ferromagnetic materials are permanent magnets characterized by a spontaneous magnetization [1, 4]. In ferromagnetic nanowires, the wire axis is a preferential axis of magnetization, and one observes formation of domains (zone in which the magnetization is oriented along the wire) separated by domain walls (zones of magnetization switching). This property plays an important role for applications in data storage or logic devices (see [10] and [2]).

In this paper, we deal with ferromagnetic rings and we study the influence of their shape on their performances for data storage. In particular, the main criterium is the number of stable configurations possibly stored by the device.

First we recall the three-dimensional model of the ferromagnetic materials (see [7, 9]): we denote by $\Omega \subset \mathbb{R}^3$ the ferromagnetic domain and by $\mathbf{m}(\mathbf{t}, \mathbf{x})$ the distribution of the magnetization at time t and at point $x \in \Omega$. We suppose that the material is saturated, *i.e.* the norm of $\mathbf{m}$ constant equals to $\mathbf{m}_s$. The magnetic induction $\mathbf{b}$ and the magnetic field $\mathbf{h}$ are linked by the constitutive relation $\mathbf{b} = \mathbf{h} + \overline{\mathbf{m}}$, where $\overline{\mathbf{m}}$ is the extension of $\mathbf{m}$ by zero outside Ω. The variation of $\mathbf{m}$ satisfies the following Landau-Lifshitz equation:

$$\frac{\partial \mathbf{m}}{\partial \mathbf{t}} = -\gamma \mathbf{m} \times \mathbf{h}_{\mathbf{eff}} - \frac{\alpha\gamma}{\mathbf{m}_s} \mathbf{m} \times (\mathbf{m} \times \mathbf{h}_{\mathbf{eff}}),$$

where γ is the gyromagnetic ratio, α is the damping coefficient, and $\mathbf{h}_{\mathbf{eff}}$ is the effective field given by:

$$\mathbf{h}_{\mathbf{eff}}(\mathbf{m}) = \frac{A}{\mu_0 \mathbf{m}_s^2} \Delta \mathbf{m} + \mathbf{h}_{\mathbf{d}}(\mathbf{m}),$$

where A is the exchange constant, μ_0 the permeability of the vacuum and $\mathbf{h}_{\mathbf{d}}$ the demagnetizing field obtained by solving Maxwell-Faraday equation:

$$\operatorname{curl} \mathbf{h}_{\mathbf{d}}(\mathbf{m}) = 0, \qquad \operatorname{div}(\mathbf{h}_{\mathbf{d}}(\mathbf{m}) + \overline{\mathbf{m}}) = 0, \quad \text{in } \mathbb{R}^3.$$

We consider a ferromagnetic ring $\Omega_\eta \subset \mathbb{R}^3$ obtained by rotation around the z axis of the ellipse contained in the plane $x = 0$ of equation $\dfrac{(y-R)^2}{a^2} + \dfrac{z^2}{b^2} < \eta^2$, where η is a small parameter.

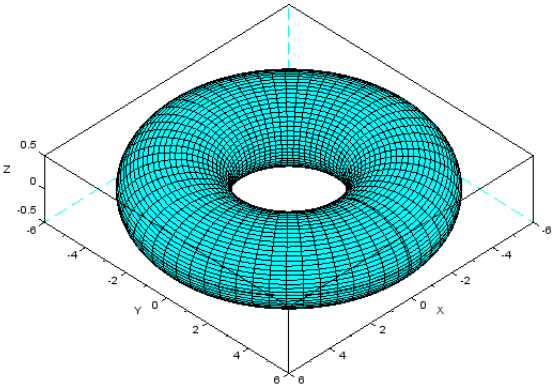


Figure 1: Ferromagnetic Ring.

When η tends to zero, the ring tends to the circle contained in the plane $z = 0$ of equation $x^2 + y^2 = R^2$. We parametrize this circle by $\theta \mapsto (R\cos\theta, R\sin\theta, 0)$. As it is established in [3] by asymptotic process, we obtain the following one-dimensional limit model: writing the magnetic moment as $\mathbf{m}(\mathbf{t}, R\cos\theta, R\sin\theta, 0) = m_s\, \mathcal{M}(\frac{\gamma \mathbf{m}_s}{\lambda}\mathbf{t}, \theta)$, where the parameter λ is given by $\lambda = \dfrac{A}{\mu_0 R^2 m_s^2}$, the new unknown $\mathcal{M}$ satisfies the renormalized saturation constraint $|\mathcal{M}| = 1$ and verifies:

$$\begin{cases} \mathcal{M} : (t,\theta) \mapsto \mathcal{M}(t,\theta) \in S^2 \subset \mathbb{R}^3, \quad 2\pi\text{-periodic in the variable } \theta, \\ \dfrac{\partial \mathcal{M}}{\partial t} = -\mathcal{M} \times \mathcal{H}_{\mathbf{eff}}(\mathcal{M}) - \alpha \mathcal{M} \times (\mathcal{M} \times \mathcal{H}_{\mathbf{eff}}(\mathcal{M})), \\ \mathcal{H}_{\mathbf{eff}}(\mathcal{M}) = \partial_{\theta\theta}\mathcal{M} + \dfrac{1}{\lambda}\mathcal{H}_{\mathbf{d}}(\mathcal{M}), \\ \mathcal{H}_{\mathbf{d}}(\mathcal{M})(\theta) = -\dfrac{b}{a+b}\,\langle \mathcal{M}(\theta)|\mathbf{e_r}(\theta)\rangle\, \mathbf{e_r}(\theta) - \dfrac{a}{a+b}\mathcal{M}_3(\theta)\mathbf{e_3}, \end{cases} \tag{1}$$

with

$$\mathbf{e_r} = \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix}, \quad \mathbf{e}_\theta = \begin{pmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{pmatrix}, \quad \mathbf{e_3} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

We remark in particular that the limit demagnetizing operator $\mathcal{H}_d$ is local in the one-dimen-

sional model. We introduce the rotation R_σ given by

$$R_\sigma = \begin{pmatrix} \cos\sigma & -\sin\sigma & 0 \\ \sin\sigma & \cos\sigma & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Remark 1. Equation (1) is invariant by translation-rotation: if we denote by $\mathcal{M}$ a solution of (1), we consider $\mathcal{M}^\sigma$ defined by:

$$\mathcal{M}^\sigma(t,\theta) = R_\sigma(\mathcal{M}(t,\theta-\sigma)).$$

Since $\mathbf{e}_r(\theta) = R_\sigma(\mathbf{e}_r(\theta-\sigma))$, we have:

$$\mathcal{H}_d(\mathcal{M}^\sigma)(t,\theta) = R_\sigma(\mathcal{H}_d(\mathcal{M})(t,\theta-\sigma)).$$

In addition,

$$\partial_{\theta\theta}\mathcal{M}^\sigma(t,\theta) = R_\sigma(\partial_{\theta\theta}\mathcal{M}(t,\theta-\sigma)) \text{ and } \frac{\partial\mathcal{M}^\sigma}{\partial t}(t,\theta) = R_\sigma(\frac{\partial\mathcal{M}}{\partial t}(t,\theta-\sigma)).$$

Therefore $\mathcal{M}^\sigma$ is also solution for (1).

We denote by $\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \\ \mathbf{M}_3 \end{pmatrix}$ the vector of the coordinates of $\mathcal{M}(\mathbf{t},\theta)$ in the frame $(\mathbf{e_r}, \mathbf{e}_\theta, \mathbf{e_3})$:

$$\mathcal{M}(\mathbf{t},\theta) = \mathbf{M}_1(t,\theta)\mathbf{e_r}(\theta) + \mathbf{M}_2(t,\theta)\mathbf{e}_\theta(\theta) + \mathbf{M}_3(t,\theta)\mathbf{e_3}.$$

Rewriting equation (1) in the mobile frame $(\mathbf{e_r}(\theta), \mathbf{e}_\theta(\theta), \mathbf{e_3})$ with these coordinates, we obtain the following model:

$$\begin{cases} \mathbf{M} : (t,\theta) \mapsto \mathbf{M}(\mathbf{t},\theta) \in S^2 \quad 2\pi - \text{periodic in } \theta, \\ \dfrac{\partial\mathbf{M}}{\partial t} = -\mathbf{M}\times\mathbf{H_{eff}}(\mathbf{M}) - \alpha\mathbf{M}\times(\mathbf{M}\times\mathbf{H_{eff}}(\mathbf{M})), \\ \mathbf{H_{eff}}(\mathbf{M}) = \partial_{\theta\theta}\mathbf{M} + 2\mathbf{e_3}\times\partial_\theta\mathbf{M} - \mathbf{M}_1\mathbf{e_1} - \mathbf{M}_2\mathbf{e_2} + \dfrac{1}{\lambda}\mathbf{H_d}(\mathbf{M}), \\ \mathbf{H_d}(\mathbf{M}) = -\dfrac{1}{a+b}(b\mathbf{M}_1\mathbf{e_1} + a\mathbf{M}_3\mathbf{e_3}). \end{cases} \tag{2}$$

Remark 2. From the invariance by rotation-translation of (1), we obtain that (2) is invariant by translation, *i.e.* if $\mathbf{M}$ satisfies (2), then for all $\sigma\in\mathbb{R}$, $(t,\theta)\mapsto\mathbf{M}(t,\theta-\sigma)$ is also solution for (2).

We focus on static planar solutions $\mathbf{M}^0$ for Equation (2) taking their values in the plane $z=0$, that is on the form

$$\mathbf{M^0} = (\cos u(\theta), \sin u(\theta), 0), \tag{3}$$

where $u \in H^1_{loc}(\mathbb{R};\mathbb{R})$ satisfies:

$$\exists k \in \mathbb{Z},\ \forall \theta \in \mathbb{R}, u(\theta + 2\pi) = u(\theta) + 2k\pi, \tag{4}$$

in order to ensure that $\mathbf{M}^0$ is 2π-periodic. We denote by $\mathcal{M}^0$ the corresponding solution for Equation (1):

$$\mathcal{M}^0(\theta) = \cos u(\theta)\mathbf{e_r}(\theta) + \sin u(\theta)\mathbf{e}_\theta(\theta).$$

We remark that $k+1$ is the winding number of $\mathcal{M}^0$ as a function from the unit circle S^1 into itself. As already said, we take care about Domain Walls. Since the wire direction is an easy axis of magnetization, we call domain a point in which the magnetization $\mathcal{M}^0$ is tangent to the ring, and we call Domain Wall (or magnetization switching) a point separating two consecutive domains in which the magnetization is orthogonal to the ring. We remark that by periodicity argument, the number of switchings in even. The key point for applications is to address the stability of the configurations in order to fix the number of switchings. As we will see after, the number of switching for a configuration u is equal to $2|k|$ in Formula (4). In the following section, we will describe all the static planar configurations, and we will study their stability in Section 3.

Remark 3. We can construct static solutions M^0 of (2) taking their values in the plane $y = 0$, that is on the form $M^0 = (\cos u(\theta), 0, \sin u(\theta))$ (for example $M^0 \equiv \mathbf{e}_3$). From the physical point of view, since $a > b$, we can prove that these solutions are unstable. The existence of static solutions of (2) which do not take their values either in the plane $z = 0$ or in the plane $y = 0$ remains an open problem.

§2. Construction of static profiles

By a straightforward calculation, $\mathbf{M}^0$ is a static solution of (2) is and only if u satisfies (4) and the pendulum equation:

$$u'' + \frac{b}{\lambda(a+b)} \cos u \sin u = 0. \tag{5}$$

By multiplying the pendulum equation by u' and integration, we obtain that there exists a constant ρ such that for all θ,

$$(u'(\theta))^2 + \frac{b}{\lambda(a+b)} \sin^2 u(\theta) = \rho^2. \tag{6}$$

2.1. Case $k = 0$

First we look for planar static solutions $\mathcal{M}^0$ for (1) of winding number equal to one, *i.e.* we look for the solutions u of (5) such that $u(\theta + 2\pi) = u(\theta)$ (*i.e.* with $k = 0$). The periodic solutions of (5) are either the constant solutions equal to 0 modulo $\frac{\pi}{2}$ or are the non constant trajectories between the separatrix, which are the lines $p = \pm\sqrt{\frac{b}{\lambda(a+b)}}\cos u$ in the phase portrait (where we denote by (u, p) the coordinate in the phase plane, see Figure 2). By

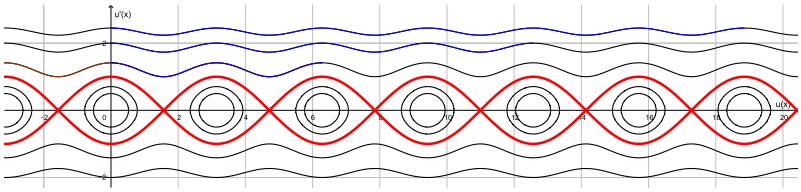


Figure 2: Phase portrait $(u(\theta), u'(\theta))$ for (5).

classical arguments, such a solution $\theta \mapsto (u(\theta), u'(\theta))$ remains in one cell C_n between the separatrix, where:

$$C_n = \left\{ (u,p) \in \mathbb{R}^2, -\pi/2 + n\pi < u < \pi/2 + n\pi \text{ with } |p| < \sqrt{\frac{b}{\lambda(a+b)}} |\cos u| \right\}.$$

We first look for the 2π-periodic solutions in C_0. By translation in the variable θ, we can assume that $u(0) \in]0, \frac{\pi}{2}[$ and $u'(0) = 0$. For $\gamma \in]0, \frac{\pi}{2}[$, we denote by u_γ the solution of (5) such that $u_\gamma(0) = \gamma$ and $u'(0) = 0$. We have:

$$\forall \theta \in \mathbb{R}, \quad (u'_\gamma(\theta))^2 + \frac{b}{\lambda(a+b)} \sin^2 u_\gamma(\theta) = \frac{b}{\lambda(a+b)} \sin^2 \gamma.$$

By classical calculation, the period $L(\gamma)$ of this solution is given by:

$$L(\gamma) = 4\sqrt{\frac{\lambda(a+b)}{b}} \int_0^\gamma \frac{du}{\sqrt{\sin^2\gamma - \sin^2 u}}.$$

The function u_γ satisfies $u_\gamma(0) = u_\gamma(2\pi)$ if and only if there exists $n \in \mathbb{N}^*$ such that $nL(\gamma) = 2\pi$. The function L is continuous and non decreasing. In addition, we have $\lim_{\gamma \to \frac{\pi}{2}} L(\gamma) = +\infty$, and $\lim_{\gamma \to 0} L(\gamma) = 2\pi\sqrt{\frac{\lambda(a+b)}{b}}$.

Therefore, if $\frac{b}{\lambda(a+b)} \leq 1$, for all $\gamma \in]0, \frac{\pi}{2}[$, $L(\gamma) > 2\pi$, so there is no 2π-periodic solution of this type.

If $\frac{b}{\lambda(a+b)} > 1$, let $l \in \mathbb{N}^*$ such that $l + 1 \geq \sqrt{\frac{b}{\lambda(a+b)}} > l$. Then, $\frac{2\pi}{l+1} \leq \lim_{\gamma \to 0} L(\gamma) < \frac{2\pi}{l}$.
So on the one hand, by monotonicity argument, for all $n \in \{1, \ldots, l\}$, there exists only one $\gamma_n \in]0, \frac{\pi}{2}[$ such that $L(\gamma_n) = \frac{2\pi}{n}$. On the other hand, for all $\gamma \in]0, \frac{\pi}{2}[$, $L(\gamma) > \frac{2\pi}{l+1}$, so the minimal possible period of such solutions is $\frac{2\pi}{l}$. Therefore, there are exactly l 2π-periodic solutions (modulo translation in θ) in the cell C_0.
By the same arguments, we find exactly l 2π-periodic solutions in the cell C_1.

So, in the case $k = 0$, we have the following theorem:

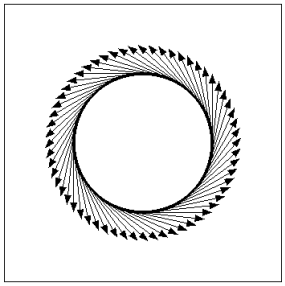


Figure 3: Profile of e_θ.

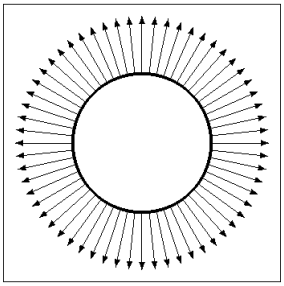


Figure 4: Profile of e_r.

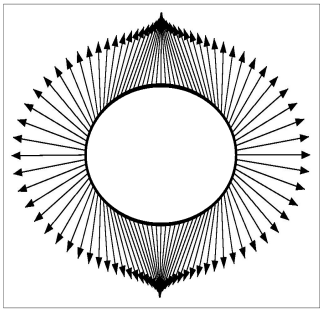


Figure 5: Solution with $l = 2$.

Theorem 1. *Let $\lambda > 0$, $a > 0$, $b > 0$. Let $l \in \mathbb{N}$ such that $l < \sqrt{\frac{b}{\lambda(a+b)}} \leq l + 1$. In addition to the solutions $\pm e_r$ and $\pm e_\theta$, Equation (1) admits $2l$ other degree-one planar static solutions modulo rotation-translation.*

2.2. Case $k \neq 0$

Now we look for planar static solutions of (1) of degree $k+1$, $k \neq 0$, *i.e.* we look for solutions u for (5) such that $u(\theta + 2\pi) = u(\theta) + 2k\pi$, with $k \neq 0$. These solutions are outside the separatrix, since the solutions inside the separatrix remain in intervals which sizes are less than π. These solutions satisfy (6) with $|\rho|^2 > \dfrac{b}{\lambda(a+b)}$.

For $k \geq 1$, we consider, for $\rho > \sqrt{\frac{b}{\lambda(a+b)}}$, the solution v_ρ of (5) such that $v_\rho(0) = 0$ and $v'_\rho(0) = \rho$. Writing (6), we obtain that v_ρ reaches the value $2k\pi$ at the point θ_ρ given by:

$$\theta_\rho = \int_0^{2k\pi} \frac{dv}{\sqrt{\rho^2 - \dfrac{b}{\lambda(a+b)} \sin^2 v}} = 4k \int_0^{\frac{\pi}{2}} \frac{dv}{\sqrt{\rho^2 - \dfrac{b}{\lambda(a+b)} \sin^2 v}}.$$

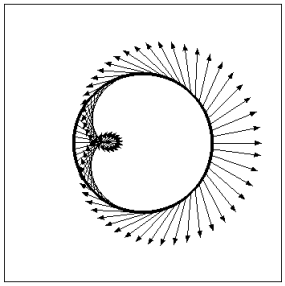


Figure 6: Solution with 2 walls (k=1).

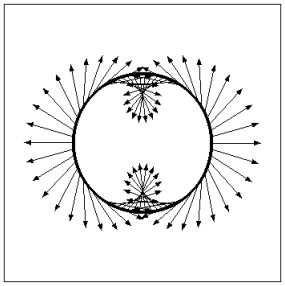


Figure 7: Solution with 4 walls (k=2).

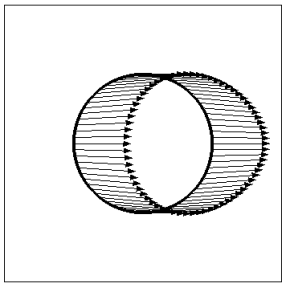


Figure 8: Solution with 2 walls (k=-1).

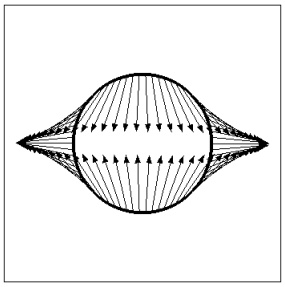


Figure 9: Solution with 4 walls (k=-2).

We remark that $\rho \mapsto \theta_\rho$ is continuous and non increasing. In addition, we have:

$$\lim_{\rho \to \sqrt{\frac{b}{\lambda(a+b)}}} \theta_k(\rho) = +\infty \text{ and } \lim_{\rho \to +\infty} \theta_k(\rho) = 0.$$

Then we deduce that for all fixed $k \geq 1$ there exist an unique $\rho \in \left]\sqrt{\frac{b}{\lambda(a+b)}}, +\infty\right[$ such that $\theta_k(\rho) = 2\pi$.

By the same way we find the same result for $k \leq -1$ with $\rho < -\sqrt{\dfrac{b}{\lambda(a+b)}}$. So, in the case $k \in \mathbb{Z}^*$ we have the following theorem:

Theorem 2. *For any fixed $k \in \mathbb{Z}^*$, Equation (1) admits a planar static solution of degree $k+1$. This solution is unique modulo translation-rotations and presents $2|k|$ walls.*

§3. Stability of wall profiles

In this part we address the stability of the solutions given in the previous part. The first difficultly comes from the saturation constraint: we must consider only perturbations satisfying this constraint. To solve this problem we use the mobile frame technique developed in [5].

3.1. Mobile frame technique

We address the stability of a static solution $\mathbf{M^0} = \begin{pmatrix} \cos u \\ \sin u \\ 0 \end{pmatrix}$ for Equation (2), obtained either in Theorem 1 or in Theorem 2. We denote by ρ^2 the conserved quantity $(u')^2 + \frac{b}{\lambda(a+b)} \sin^2 u$ in (6). We introduce the mobile frame $(\mathbf{M}^0(\theta), \mathbf{M}^1(\theta), \mathbf{M}^2)$, where

$$\mathbf{M}^1(\theta) = \begin{pmatrix} -\sin u(\theta) \\ \cos u(\theta) \\ 0 \end{pmatrix} \text{ and } \mathbf{M}^2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

We describe the perturbations of $\mathbf{M}^0$ as follows:

$$\mathbf{M}(t,\theta) = r_1(t,\theta)\mathbf{M^1}(\theta) + r_2(t,\theta)\mathbf{M^2} + (1 + \nu(r(t,\theta)))\mathbf{M^0}(\theta),$$

with $\nu(r) = \sqrt{1 - r_1^2 - r_2^2} - 1$, so that the saturation constraint is satisfied. We write the Landau-Lifshitz equation (2) with this new unknown $r : \mathbb{R}^+ \times [0, 2\pi] \to \mathbb{R}^2$, and by projection onto $\mathbf{M^1}$ and $\mathbf{M^2}$, we establish as in [5] that $\mathbf{M}$ satisfies (2) if and only if r satisfied the equation:

$$\partial_t r = \begin{pmatrix} -\alpha & -1 \\ 1 & -\alpha \end{pmatrix} Lr + F(\theta, r, \partial_\theta r, \partial_{\theta\theta} r), \tag{7}$$

where $F(\theta, r, \partial_\theta r, \partial_{\theta\theta} r)$ is the non linear part, and with:

$$Lr = \begin{pmatrix} \mathcal{L}_1 r_1 \\ \mathcal{L}_2 r_2 \end{pmatrix}$$

with

$$\begin{aligned} \mathcal{L}_1 &= -\partial_{\theta\theta} + \frac{b}{\lambda(a+b)}(\sin^2 u - \cos^2 u), \\ \mathcal{L}_2 &= \mathcal{L}_1 + (\frac{a}{\lambda(a+b)} - \rho^2 - 2u' - 1). \end{aligned} \tag{8}$$

In addition, $\mathbf{M}$ is stable for (2) if and only if 0 is stable for (7). The positivity of $\mathcal{L}_1$ and $\mathcal{L}_2$ is crucial for the stability (see [6]). Let us study the different cases.

3.2. Stability of e_θ

The static planar solution $\mathbf{e}_\theta$ for Equation (1) corresponds to the static planar solution $\mathbf{M^0} = (0, 1, 0)$ for Equation (2) with $u = \frac{\pi}{2}$. The obtained linearization is L given by:

$$L = \begin{pmatrix} -\partial_{\theta\theta} + \dfrac{b}{\lambda(a+b)} \\ -\partial_{\theta\theta} + \left(\dfrac{a}{\lambda(a+b)} - 1\right) \end{pmatrix}.$$

As already said, we prove in [6] that if $L > 0$ then 0 is asymptoticly stable for (7). The operator $\mathcal{L}_1$ is positive. Concerning $\mathcal{L}_2$, its positiveness is related to the sign of $\frac{a}{\lambda(a+b)} - 1$, so we obtain the following theorem:

Theorem 3. *If $\lambda < \frac{a}{a+b}$, then $\mathbf{e}_\theta$ is asymptoticly stable. If $\lambda > \frac{a}{a+b}$ then $\mathbf{e}_\theta$ is linearly unstable.*

Remark 4. In the previous theorem, if $\lambda > \frac{a}{a+b}$, i.e. if the radius of the ring is sufficiently small, then the exchange energy of $\mathbf{e}_\theta$ becomes large and creates instability.

3.3. Instability of e_r

We study the static planar solution $\mathbf{e_r}$ of the equation (1), which corresponds to the static planar solution $\mathbf{M^0} = (1, 0, 0)$ for Equation (2), *i. e.* with $u = 0$. The obtained linearization is given by:

$$L = \begin{pmatrix} \mathcal{L}_1 \\ \mathcal{L}_2 \end{pmatrix},$$

where

$$\mathcal{L}_1 = -\partial_{\theta\theta} - \frac{b}{\lambda(a+b)} \text{ and } \mathcal{L}_2 = -\partial_{\theta\theta} + \frac{a-b}{\lambda(a+b)} - 1.$$

In particular, $\mathcal{L}_1$ admits negative eigenvalues so we have the following theorem:

Theorem 4. *Whatever $\lambda > 0$, $a > 0$ and $b > 0$, e_r is linearly unstable for Equation (1)..*

3.4. Stability of the non constant solutions

We address the stability of a non constant solution $\mathbf{M^0} = \begin{pmatrix} \cos u \\ \sin u \\ 0 \end{pmatrix}$ for Equation (2), obtained either in Theorem 1 in the case $\frac{b}{\lambda(a+b)} > 1$, or in Theorem 2. We denote by ρ^2 the conserved quantity $(u')^2 + \frac{b}{\lambda(a+b)} \sin^2 u$ in (6). We recall that the stability for $\mathbf{M}^0$ is related to the positivity of the operators $\mathcal{L}_1$ and $\mathcal{L}_2$ given by (8).

3.4.1. Linear instability for the non constant solutions given by Theorem 1

We assume that $\rho^2 < \frac{b}{\lambda(a+b)}$. In this case, the trajectories $\theta \mapsto (u(\theta), u'(\theta))$ are between the separatrix. We remark that $\mathcal{L}_1 \cos u = (\rho^2 - \frac{b}{\lambda(a+b)}) \cos u$. So $\rho^2 - \frac{b}{\lambda(a+b)}$ is a negative eigenvalue associated to the eigenvector $\cos u$. Thus, $\mathcal{L}_1$ is not positive. Therefore we have the following Theorem:

Theorem 5. *In the case $\rho^2 < \frac{b}{\lambda(a+b)}$, the static solution M^0 is linearly unstable for (1).*

3.4.2. Linear stability for the non constant solutions given by Theorem 2

We assume now that $\rho^2 > \frac{b}{\lambda(a+b)}$. We have the following proposition:

Proposition 6. *$\mathcal{L}_1$ is a linear non negative operator. In addition $Ker\mathcal{L}_1 = \mathbb{R}u'$.*

Proof. We set $\ell_1 = \partial_\theta + \dfrac{b}{\lambda(a+b)}\dfrac{\sin u \cos u}{u'}$, then $\ell_1^* = -\partial_\theta + \dfrac{b}{\lambda(a+b)}\dfrac{\sin u \cos u}{u'}$ and we have the factorization:

$$\ell_1^* \circ \ell_1 = \mathcal{L}_1.$$

So $\mathcal{L}_1$ is a positive operator. We have also

$$\mathcal{L}_1 u' = -\left(u'' + \frac{b}{\lambda(a+b)} \cos u \sin u\right)' = 0,$$

so $u' \in Ker\mathcal{L}_1$. □

Therefore in this case, $\mathcal{L}_1$ is non negative. The existence of an order-one vanishing eigenvalue is an additional difficulty to obtain the nonlinear stability. This is due to the invariance of (2) by translation (see Remark 2), so that there exists a one-parameter family of constant solutions for (2): $\theta \mapsto \mathbf{M}^0(\theta - \sigma)$ depending of the parameter σ. By projection on the mobile frame, we obtain the existence of a one-parameter family of constant solutions for (7): $\theta \mapsto R(\sigma)(\theta)$.

In order to take into account the zero eigenvalue of L, as in [5] or [8], we rewrite r in the following new system of coordinates:

$$r(t,\theta) = R(\sigma(t))(\theta) + w(t,\theta),$$

where now the parameter σ depends on the time variable: $\sigma \in C^1(\mathbb{R}^+;\mathbb{R})$, and $w \in C^1(\mathbb{R}^+; H^2_{per})$ such that the first component w_1 of w satisfies the orthogonality condition:

$$\forall t > 0, \quad \int_0^{2\pi} w_1(t,\theta) u'(\theta) d\theta = 0.$$

In this new unknown (σ, w), we are able to separate the dynamics of w and the dynamics of σ. In particular, if $\mathcal{L}_2$ is positive, we can prove by variational estimates that $w(t)$ tends to zero in H^1 and that $\sigma(t)$ tends to a finite limit when t tends to $+\infty$. This means that $\mathbf{M}(t)$ tends to a translation of $\mathbf{M}^0$ when t tends to $+\infty$ (asymptotic stability modulo translation in the variable θ).

Now, the difficulty is to prove the study of $\mathcal{L}_2$. We prove in [6] the following Theorems:

Theorem 7. *We consider the solutions of (1) given by Theorem 2 in the case $k \leq 1$.*
If $a \leq b$, these solutions are unstable.
If $a > b$, if λ is large enough, these solutions are unstable.
If $a > b$, there exists $\lambda_0 > 0$ such that if $0 < \lambda < \lambda_0$ then there exists $k_0 > 0$ such that the solutions with $k \leq k_0$ are stable and the solutions with $k > k_0$ are unstable.

Theorem 8. *We consider the solutions of (1) given by Theorem 2 in the case $k \geq -1$. If $a > b$, there exists $\lambda_0 > 0$ such that if $0 < \lambda < \lambda_0$ then there exists $k_0 < 0$ such that the solutions with $k_0 \leq k \leq -1$ are stable and the solutions with $k < k_0$ are unstable.*

Remark 5. We remark in particular that, if $a > b$, the solution with $k = -1$ is stable whatever $\lambda > 0$. In addition, we establish that the larger the diameter of the ring, the more information it can store.

References

[1] Aharoni, A. *Introduction of the Theory of Ferromagnetism*. 2000. Oxford University Press, 109.

[2] Allwood, D. A., Xiong, G., Faulkner, C., Atkinson, D., Petit, D., and Cowburn, R. P. Magnetic domain-wall logic. *Science* (2005), 1688–1692.

[3] AlSayed, A., Carbou, G., and Labbé, S. Asymptotic model for twisted bent ferromagnetic wires with electric current. *Z. Angew. Math. Phys. 70*, 1 (2019).

[4] Brown, W. F. *Micromagnetics*, vol. 40 of *Classics in Applied Mathematics*. Wiley, Philadelphia, 1963. Firstly published by North-Holland, Amsterdam, 1978.

[5] Carbou, G., and Labbé, S. Stability for static walls in ferromagnetic nanowires. *Discrete Contin. Dyn. Sys. B 6*, 2 (2006), 273–290.

[6] Carbou, G., Moussaoui, M., and Rachi, R. Stability of static magnetization in ferromagnetic rings. *In preparation*.

[7] Halpern, L., and Labbé, S. Modélisation et simulation du comportement des matériaux ferromagnétiques. *Matapli 66* (2001), 70–86.

[8] Kapitula, T. Multidimensional stability of planar traveling waves. *Trans. Amer. Math. Soc. 349*, 1 (1997), 257–269.

[9] Landau, L., and Lifschitz, E. *Electrodynamique des milieux continus*. 1969. Moscou, vol 8.

[10] Parkin, S. S., Hayashi, M., and Thomas, L. Magnetic domain-wall racetrack. *Science 320* (2008), 190–194.

G. Carbou and R. Rachi
LMAP - UMR CNRS 5142 - ES2 UPPA - IPRA,
Université de Pau et des Pays de l'Adour
Avenue de l'Université - BP 1155, 64013 PAU
CEDEX FRANCE
gcarbou@univ-pau.fr and
rrachi@univ-pau.fr

M. Moussaoui
Laboratoire d'Analyse Non Linéaire et Histoire des Maths
E.N.S, B.P. 92 Vieux Kouba16050 Algiers, Algeria
mmohand47@gmail.com

Monografías Matemáticas *García de Galdeano* **42**, 57–64 (2019)

GENERALIZED FRACTIONAL DIFFERENTIAL EQUATIONS WITH ORDER VARYING IN TIME IN COMPLEX BANACH SPACES: ANALYTIC AND NUMERICAL ASYMPTOTIC BEHAVIOR

Eduardo Cuesta and Rodrigo Ponce

Abstract. The asymptotic behavior of the solution of generalized fractional order integral equations with order varying in time arising in image processing is investigated in this work. It is shown here that the asymptotic behavior is extended from the corresponding property for the *scalar* abstract equation $u(t) = \partial_t^{-\alpha(t)} Au(t) + f(t), \quad 0 \le t \le T$, for a given $\alpha : [0, T] \to (1, 2)$, f defined in $0 \le t \le T$, $A : \mathcal{D}(A) \subset X \to X$ a bounded operator, and X a Banach space. It is also proved that a first order time discretization inherits the behavior of the continuous solution.

Keywords: fractional integrals, variable order, Banach spaces, convolution quadrature.
AMS classification: 45D05,65J08,65R20.

§1. Introduction

One of the most interesting properties in time dependent partial differential equations based models for image processing is the asymptotic behavior of the analytic solution as time goes to infinity, but an even more important issue is if the time discretization inherits this behavior. In fact, the asymptotic behavior allows us to predict the diffusion level of the solution as the scale parameter t grows up, or in image processing terminology, this allows one to predict the degree of blurring acting on the image as time tends to infinity.

The asymptotic behavior of most of local models related to image processing has been extensively investigated, on the contrary what happens with nonlocal models. The memory effect in nonlocal equations makes in many cases the study of the asymptotic behavior more difficult if compared to the local models, but in spite of this the study has been carried out for general Volterra equations [8], and in a particular and very well known kind of nonlocal models as they are the linear integro–differential equations of fractional order [1]. This behavior has been already experienced in practical instances related to image processing (see e.g. the pioneer work [3]).

Recently an extension of the integro–differential equations of fractional order in [3] consisting in replacing the constant fractional order by fractional order varying in time has been successfully applied in the framework of image filtering [2]. To the best of our knowledge up to now there was no particular results on the asymptotic behavior adapted to fractional equations with order varying in time, however a recent work solves this issue. In fact, in [5] the

authors study the asymptotic behavior of such a kind of equations, and they extend the result to its time discretization. The well–posedness, and the regularity of the solution is studied in [5] as well, everything done in the abstract framework of complex Banach spaces.

The main contribution of this work is the extension of these results to the case of generalized fractional equations in the sense of [2], whose main difference is that this approach involves severals varying in time integration orders in a matrix–form.

The paper is organized as follows, Section 2 is devoted to mathematical background and model formulation, in Section 3 and 4 we present the main result of this work related to the continuous and discrete solutions respectively, and finally in Section 5 we presents some observations and final conclusions.

§2. Mathematical background

The present work is motivated by the nonlocal in time evolution partial differential equations based approach to image processing introduced in [2], whose formulation is given in terms of time fractional integrals with orders varying in time. In fact, let $\mathbf{u}_0$ be an initial data, standing for a $J \times J$, perturbed sampled image, $J > 0$, vector–arranged as $J^2 \times 1$ vector, and intended to be restored. The nonlocal evolutionary model proposed in [2] reads

$$\mathbf{u}(t) = \mathbf{u}_0 + \int_0^t A_h \mathbf{D}(t-s)\mathbf{u}(s)\,\mathrm{d}s, \quad t > 0, \tag{1}$$

where $\mathbf{u} : [0,T] \to \mathcal{M}_{J^2\times 1}(\mathbb{R})$, stands for the original image evolved up to the time level $t > 0$, which has been vector–arranged as a column vector with J^2 entries, i.e. $\mathbf{u} = (u_j)_{1\le j\le J^2}$. Moreover, $A_h \in \mathcal{M}_{J^2\times J^2}(\mathbb{R})$ is a symmetric and negative semi–definite matrix. An example of matrix A_h is the one corresponding to the discrete Laplacian based on second order finite difference scheme, including discrete and homogeneous Newman boundary conditions. Notice that most of classical spatial discretizations of the Laplacian give rise to one of these matrices. Finally, $\mathbf{D} : [0,T] \to \mathcal{M}_{J^2\times J^2}(\mathbb{R})$ stands for a diagonal matrix, $\mathbf{D} = \mathrm{diag}_{1\le i\le J^2}(k_j)$, where the entries $k_j(t)$, $1 \le j \le J^2$, coincide with the convolution kernels those define the fractional integral with order varying in time $\alpha_j(t)$, for each $1 \le j \le J^2$.

Recall that several definitions for non integer integrals (or derivatives) with order varying in time can be found in the literature, and the convenience of using one vs. the others has been largely discussed, and basically depends on the purposes of the model. For the shortness of the presentation, we do not include such a discussion here, we just adopt the same definition as in [5] and we refer there the reader for a more precise motivation of this choice. Before recalling this definition let us denote $\mathcal{L}$ and $\mathcal{L}^{-1}$ the Laplace transform operator and the inverse Laplace transform operator, respectively. In that manner, let $\alpha : [0,T] \to (1,2)$ be a piecewise continuous function then, for $f \in L^1(0,+\infty)$, the fractional integral of order $\alpha(t)$ is defined as

$$\partial_t^{-\alpha(t)} f(t) = \int_0^t k(t-s) f(s)\,\mathrm{d}s, \quad t > 0, \tag{2}$$

where,

$$k(t) := \mathcal{L}^{-1}(K)(t), \quad \text{and} \quad K(z) := \frac{1}{z^{z\tilde{\alpha}(z)}}, \tag{3}$$

and

$$\tilde{\alpha}(z) = \mathcal{L}(\alpha)(z), \tag{4}$$

for $z \in \mathcal{D}(K) \subset \mathbb{C}$. Simply observe that, if the fractional order turns out to be constant, then $\tilde{\alpha}(z) = \alpha/z$ for certain constant α, and the definition (2)–(4) coincides with the very well known Riemann–Liouville one [9]. We refer the reader to [5] for a deeper discussion on this matter.

The underlying idea behind the use of this model in image filtering is that the diffusion in the original image $\mathbf{u}_0$ applies pixel–by–pixel by setting different viscosity parameters (or diffusion coefficients) $\alpha_j(t)$ for each single pixel, which evolves in time according to some criteria (edge–preserving, texture–preserving, among others). This fact gives rise to the convolution kernels $k_j(t)$, $1 \leq j \leq J^2$ of the type mentioned above. This approach extends many other previous fractional approaches whose diffusion orders keep constant along the whole time interval.

§3. Main result

In this section we present the main theorem of the paper related to the continuous solution, but we previously recall the result on which this is based on.

Let $(Y, \|\cdot\|)$ be a complex Banach space, $\alpha : [0, T] \to (1, 2)$ a piecewise continuous function, and consider the abstract integral equation

$$u(t) = u_0 + \partial_t^{-\alpha(t)}(Au)(t), \quad t > 0, \tag{5}$$

where $A : \mathcal{D}(A) \subset Y \to Y$ is a linear, closed, and θ–sectorial operator in Y, $0 < \theta < \pi/2$, $u_0 \in Y$ stands for the initial data, and $\partial_t^{-\alpha(t)}$ defines the fractional integral according the definition (2)–(4).

Recall that a linear and closed operator is θ–sectorial, $0 < \theta < \pi/2$, if there exist $w \in \mathbb{R}$ and $L > 0$ such that

- The resolvent $(zI - A)^{-1}$ is analytic, and
- It satisfies

$$\|(zI - A)^{-1}\|_{Y \to Y} \leq \frac{L}{|z - w|},$$

for $z \in \mathbb{C}$, with $\mathrm{Arg}(z - w) > \pi - \theta$.

Notice that, since we are assuming that $\alpha(t)$ is piecewise continuous in $[0, T]$, $\alpha(t)$ admits Laplace transform in a complex domain $\mathrm{Re}(z) \geq C_\alpha$, for some $C_\alpha > 0$. In addition assume that there exist $1 < m < M < 2$, $C > 0$, and $0 < \varepsilon < 1$, such that, for $z \in \mathbb{C}$, $\mathrm{Re}(z) \geq C_\alpha$,

(A1) $m \leq \mathrm{Re}(z\tilde{\alpha}(z)) \leq M$, and $\dfrac{M\pi}{2} < \varepsilon(\pi - \theta)$.

(A2) $|\mathrm{Im}(z\tilde{\alpha}(z))| \leq C$, and

$$\left|\log\left(|z|\mathrm{Im}(z\tilde{\alpha}(z))\right)\right| < (1 - \varepsilon)(\pi - \theta),$$

where ε is expected to be close to 1.

Assume also that

$$0 < \theta < \pi - \frac{M\pi}{2} - \max_{r\geq R} \frac{\log(r)}{r^{\varepsilon}},$$

for $R > 0$ large enough.

Under these assumptions, equation (5) can be written in terms of the Laplace transform as

$$U(z) = \frac{H(z)}{z}(H(z)I - A)^{-1}u_0, \tag{6}$$

where

$$H(z) := z^{z\tilde{\alpha}(z)}, \quad \text{and} \quad U(z) = \mathcal{L}(u)(z),$$

for $\mathrm{Re}(z) \geq C_\alpha$. Therefore there exists an evolution operator $E(t)$, $t > 0$, such that the mild solution of (5) can be written as

$$u(t) = E(t)u_0, \quad t > 0. \tag{7}$$

In addition the evolution operator $E(t)$ can be expressed by means of the Bromwich formula as

$$E(t) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \mathrm{e}^{tz} \frac{H(z)}{z}(H(z)I - A)^{-1}\, \mathrm{d}z, \tag{8}$$

where Γ is a convenient complex path running from $-\mathrm{i}\infty$ to $+\mathrm{i}\infty$ within the analyticity domain of the resolvent of A, and positively oriented, i.e. with increasing imaginary part (see [5] for more details).

The asymptotic behavior of the solution of (5) is stated in [5, Theorem 5.1], in fact it is proved that there exists $C > 0$ such that

$$\|E(t)\|_{Y\to Y} \leq \frac{CL}{1 + |w|t^m}, \quad \text{as} \quad t \to +\infty. \tag{9}$$

The first contribution of this work consists of extending the asymptotic behavior of the solution of (5) to the solution of (1). To this end denote the Banach space $Y = L^1((0,T),\mathbb{R})$ normed as usual by $\|\cdot\|_{L^1}$ and denoted by simplicity as $\|\cdot\|$.

Let $(X, \|\cdot\|_X)$ be the Banach space defined by

$$X := \prod_{j=1}^{J^2} Y \quad \text{normed by} \quad \|\mathbf{v}\|_X := \sup_{1\leq j\leq J^2} \|v_j\|, \tag{10}$$

for $\mathbf{v} = (v_j)_{1\leq j\leq J^2} \in X$.

It is straightforward to prove that the operator $A_h\mathbf{D}(t)$ in (1), and described in Section 2, is on the one hand commutative, i.e. $A_h\mathbf{D}(t) = \mathbf{D}(t)A_h$, and on the other hand θ_0–sectorial for certain $0 < \theta_0 < \pi/2$, and $w \in \mathbb{R}^-$.

Assume that the diffusion coefficients $\alpha_j(t)$ involved in the definition of kernels in the matrix $\mathbf{D}$, admit Laplace transform in a complex domain $\mathrm{Re}(z) \geq C_\alpha$, for some $C_\alpha > 0$, and in addition we assume (A1) and (A2) for each one. In fact assume that there exist $1 < m_j < M_j < 2$, $C_j > 0$ and $0 < \varepsilon_j < 1$, for $1 \leq j \leq J^2$, such that, for $z \in \mathbb{C}$, $\mathrm{Re}(z) \geq C_\alpha$,

(B1) $m_j \leq \mathrm{Re}(z\tilde{\alpha}_j(z)) \leq M_j$, and $\dfrac{M_j\pi}{2} < \varepsilon_j(\pi - \theta_0)$.

(B2) $|\mathrm{Im}(z\tilde{\alpha}_j(z))| \leq C$, and

$$|\log\left(|z|\mathrm{Im}(z\tilde{\alpha}_j(z))\right)| < (1-\varepsilon_j)(\pi-\theta_0)$$

where all ε_j are expected to be close to 1.

Assume also that

$$\text{(C)}\quad 0 < \theta_0 < \pi - \frac{\max\limits_{1\leq j\leq J^2}\{M_j\}\cdot\pi}{2} - \max_{r\geq R}\frac{\log(r)}{r^{\varepsilon}},$$

for $R > 0$ large enough, and

$$\varepsilon = \max_{1\leq j\leq J^2}\varepsilon_j. \tag{11}$$

The well–posedness, and the regularity of the solution stated in [5] can be straightforwardly extended to (1) under the assumptions (B1), (B2), and (C). Therefore, in order to not extend unnecessarily this work we will focus solely on the asymptotic behavior of the solution of (1). On the other hand, the mild solution $\mathbf{u}(t)$ of (1) can be writen as $\mathbf{u}(t) = \mathbf{E}(t)\mathbf{u}_0$ where the evolution operator $\mathbf{E} : X \to X$ admits the expression

$$\mathbf{E}(t) := \frac{1}{2\pi\,\mathrm{i}}\int_{\Gamma_0}\frac{\mathrm{e}^{tz}}{z}(I-\widetilde{\mathbf{D}}(z)A_h)^{-1}\,\mathrm{d}z, \quad t>0, \tag{12}$$

where $\widetilde{\mathbf{D}}(z)$ stands for the componentwise Laplace transform of $\mathbf{D}(t)$, and Γ_0 is once again a convenient complex path connecting $-\mathrm{i}\infty$ and $+\mathrm{i}\infty$ with increasing imaginary part.

The theorem below represents the main contribution of this section.

Theorem 1. *Let* $\mathbf{E}(t)$ *be the evolution operator* (12) *corresponding to the mild solution of* (1) *under assumptions (B1), (B2), and (C).*

If zero does not belong to the spectrum of A_h*, then there exists* $C > 0$ *independent on* t*, such that*

$$\|\mathbf{E}(t)\|_{X\to X} \leq \frac{C}{1+|\lambda|t^m}, \quad \textit{as} \quad t\to+\infty, \tag{13}$$

where $m = \min\limits_{1\leq j\leq J^2}\{m_j\}$*, and* λ *is the spectral value of* A_h *corresponding to same index as* m*.*

If zero belongs to the spectrum on A_h*, then* $\mathbf{E}(t)$ *is merely bounded, i.e. there exists* $C > 0$ *independent on* t*, such that*

$$\|\mathbf{E}(t)\|_{X\to X} \leq C, \quad t>0.$$

Proof. Since A_h stands for a symmetric and negative semi–definite matrix, there exists an orthogonal matrix P, and a diagonal matrix D_A with non positive diagonal entries such that $A_h = PD_AP^T$.

On the one hand, we can write $\mathbf{E}(t)$ as follows

$$\mathbf{E}(t) = \frac{1}{2\pi\,\mathrm{i}}\int_{\Gamma_0}\mathrm{e}^{tz}P^T\frac{\mathbf{H}(z)}{z}(\mathbf{H}(z)-D_A)^{-1}P\,\mathrm{d}z, \quad t>0,$$

where $\mathbf{H}(z) = P\widetilde{\mathbf{D}}^{-1}(z)P^T$ is a bounded operator in X along the complex path Γ_0. On the other hand

$$\|\mathbf{E}(t)\mathbf{u}_0\|_X \leq \frac{1}{2\pi}\int_{\Gamma_0}\left|\frac{\mathrm{e}^{tz}}{z}\right|\|\mathbf{H}(z)\|_{X\to X}\|(\mathbf{H}(z)-D_A)^{-1}\mathbf{u}_0\|_X\,\mathrm{d}z,$$

for $t > 0$.

Moreover, $\|\mathbf{H}(z)\|_{X\to X} = \|\widetilde{\mathbf{D}}^{-1}(z)\|_{X\to X}$, for $z \in \Gamma_0$, and the resolvent $(\mathbf{H}(z) - D_A)^{-1}$ corresponds to a system of scalar equations where the diagonal matrix D_A plays the role of the operator A in (5), and the convolution kernel associated in (5) is here replaced by a linear combination of kernels of the same type.

First of all recall that the spectrum of D_A is located in the negative real line. Therefore, in order to accomplish the bounds of the resolvent $(\mathbf{H}(z) - D_A)^{-1}$ and the term $\mathbf{H}(z)$, we make use, as in [5], of a suitable choice of the complex path Γ_0 in (12), now under the restrictions imposed by (B1), (B2), and (C). In particular, define the complex paths $\Gamma_0^{(1)}$ and $\Gamma_0^{(2)}$ respectively by

$$\gamma_0^{(1)}(\phi) := \frac{1}{t^m} + \rho_0\, \mathrm{e}^{\,\mathrm{i}\phi}, \qquad -\varepsilon(\pi-\theta) \le \phi \le \varepsilon(\pi-\theta),$$

and

$$\gamma_0^{(2)}(\rho) := \rho\, \mathrm{e}^{\pm \mathrm{i}\varepsilon(\pi-\theta)}, \qquad \rho \ge \rho_0,$$

where ε is defined in (11), $\pm$ in $\gamma_0^{(2)}$ represents the upper and lower branches (positive and negative imaginary parts respectively), and ρ_0 stands for the distance from the origin to the intersection point of $\gamma_0^{(1)}$ and $\gamma_0^{(2)}$. Therefore,

$$\Gamma_0 := \Gamma_0^{(1,1/m)} \cup \Gamma_0^{(2,1/m)}, \tag{14}$$

where $\Gamma_0^{(1,1/m)}$ and $\Gamma_0^{(2,1/m)}$ come parametrized by $(\gamma_0^{(1)}(\phi))^{1/m}$ and $(\gamma_0^{(2)}(\rho))^{1/m}$ respectively.

So, from the bounds along Γ_0 of all terms involved in the integral (12) the proof follows. □

§4. Time discretization

The time discretization considered in [5] is based on the backward Euler convolution quadrature (see [4, 6, 7]). Now we extend the formulation to the non–scalar case,

$$\mathbf{U}_n = \mathbf{u}_0 + \sum_{j=0}^{n} \mathbf{Q}_{n-j} A_h \mathbf{U}_j, \quad 0 \le n \le N, \tag{15}$$

where the $J^2 \times J^2$ quadrature weights $\{\mathbf{Q}_n\}_{n\ge 0}$, come out from the evaluation

$$\widetilde{\mathbf{D}}\left(\frac{1-\zeta}{\tau}\right) = \sum_{n=0}^{+\infty} \mathbf{Q}_n \zeta,$$

$\tau = T/N$, and $\mathbf{D}(t)$ is the matrix–valued function in (1). In order to not extend unnecessarily this presentation we refer again the reader for more details on the convolution quadratures to [6, 7], and in fact for the one based on the backward Euler method see [4].

The key point here is that the numerical solution can be written in terms of discrete evolution operators $\{\mathbf{E}_n\}_{n\ge 0}$.

$$\mathbf{U}_n = \mathbf{E}_n \mathbf{u}_0, \quad 0 \le n \le N, \tag{16}$$

and that the Bromwich formula in vectorial form allows us to write

$$\mathbf{E}_n = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_0} \frac{r_n(tz)}{z} (I - \widetilde{\mathbf{D}}(z) A_h)^{-1} \, \mathrm{d}z, \quad n \ge 0,$$

where Γ_0 is the complex path stated in (14), and $r_n(z) := 1/(1-z)^n$. Notice that $r_n(z)$ stands for the characteristic function of the backward Euler method.

What follows is the main result of this section.

Theorem 2. *Let $\{\mathbf{E}_n\}_{n\ge 0}$ be the discrete evolution operators* (16) *associated to the numerical solution* (15) *under assumptions (B1), (B2), and (C).*

If zero does not belong to the spectrum of A_h, then there exists $C > 0$, independent on t, such that

$$\|\mathbf{E}_n\|_{X\to X} \le \frac{C}{1 + |\lambda| t_n^m}, \quad \text{as} \quad t \to +\infty, \tag{17}$$

where $m = \min\limits_{1\le j\le J^2} \{m_j\}$, and λ is the spectral value of A_h corresponding to same index as m.

If zero belongs to the spectrum on A_h, then $\mathbf{E}$ is merely bounded, i.e. there exists $C > 0$ independent on t, such that

$$\|\mathbf{E}_n\|_{X\to X} \le C, \quad t > 0.$$

The proof of Theorem 2 follows the same steps as the one of the Theorem 1, now replacing the exponential e^{tz} by the rational function $r_n(z)$.

§5. Observations and final conclusions

The first to be observed is that the numerical solution inherits the behavior as t goes to infinity of the analytic solution. Observe also that the asymptotic behavior turns out to be independent of the initial data $\mathbf{u}_0$ and its regularity since the proofs of both, Theorem 1 and 2, are done merely for the continuous and discrete evolution operators respectively. In other words, the regularity of the initial data does not affect the asymptotic behavior nor of the analytic solution neither the numerical one. This fact is a crucial issue specially in the context of image processing because this proves that the blurring is the same whatever the original image one has.

Observe also that if the matrix D_A has a null eigenvalue, the evolution operator is merely bounded and the decrease is not longer guaranteed. This confirms what happens in the case of abstract infinitesimal semigroup generators when $w = 0$ (according the notation in Section 2). The reason is that the evolution operators do not longer admit analytic extension to the left hand side complex plane.

Moreover, if $\lambda \neq 0$ in Theorems 1 and 2, then the decrease of $\mathbf{u}$ is limited by the slowest decrease along all components, or in other words it is limited by the lowest diffusion along every single pixels of the image represented by $\mathbf{u}$.

Acknowledgements

The second author was partially supported by the program *Beca Interamericana para Jóvenes Profesores e Investigadores, Banco de Santander*, and also by the Fog Research Institute under contract no. FRI–454.

References

[1] Cuesta, E. Asymptotic behaviour of the solutions of fractional integro–differential equations and some time discretizations. *Discrete Contin. Dyn. Syst., Proceedings of the 6th AIMS International Conference, suppl. (2007)* (2007), 277–285.

[2] Cuesta, E., Durán, A., and Kirane, M. On evolutionary integral models for image restoration. In *Developments in Medical Image Processing and Computational Vision* (2015), N. J. R. e. Tavares J., Ed., vol. 19 of *Lecture Notes in Computational Vision and Biomechanics*, Springer, Cham, pp. 241–260.

[3] Cuesta, E., and Finat, J. Image processing by means of a linear integro–differential equation. *3rd IASTED Int. Conf. Visualization, Imaging and Image Processing 1* (2003), 438–442.

[4] Cuesta, E., and Palencia, C. A numerical method for an integro–differential equation with memory in banach spaces: Qualitative properties. *SIAM J. Numer. Anal. 41* (2003), 1232–1241.

[5] Cuesta, E., and Ponce, R. Well–posedness, regularity, and asymptotic behavior of the continuous and discrete solutions of linear fractional integro–differential equations with order varying in time. *Electron. J. Differ. Eq. 173* (2018), 1–27.

[6] Lubich, C. Convolution quadrature and discretized operational calculus i. *Numer. Math. 52* (1988), 129–145.

[7] Lubich, C. Convolution quadrature and discretized operational calculus ii. *Numer, Math. 52* (1988), 413–425.

[8] Prüss, J. *Evolutionary Integral Equations and Applications*. Series Modern Birkhüser Classics. Birkhäuser Basel, 2012.

[9] Trujillo, A. K. H. M. S. J. *Theory and Applications of Fractional Differential Equations*, 1st ed., vol. 204 of *North-Holland Mathematics Studies*. Elsevier Science, 2006.

E. Cuesta
Department of Applied Mathematic, E.T.S.I. of Telecomunication
Campus Miguel Delibes, University of Valladolid
Paseo Belén 15, 47011 Valladolid, Spain.
eduardo@mat.uva.es

R. Ponce
Instituto de Matemática y Física
Casilla 747, Universidad de Talca
Talca, Chile.
rponce@inst-mat.utalca.cl

Monografías Matemáticas *García de Galdeano* **42**, 65–78 (2019)

A PROFILE DECOMPOSITION FOR THE LIMITING SOBOLEV EMBEDDING

Giuseppe Devillanova and Cyril Tintarev

Abstract. For many known non-compact embeddings of two Banach spaces $E \hookrightarrow F$, every bounded sequence in E has a subsequence that takes form of a *profile decomposition* - a sum of clearly structured terms with asymptotically disjoint supports plus a remainder that vanishes in the norm of F. In this note we construct a profile decomposition for arbitrary sequences in the Sobolev space $H^{1,2}(M)$ of a compact Riemannian manifold, relative to the embedding of $H^{1,2}(M)$ into $L^{2^*}(M)$, generalizing the well-known profile decomposition of Struwe [12, Proposition 2.1] to the case of arbitrary bounded sequences.

Keywords: concentration compactness, profile decompositions, multiscale analysis.
AMS classification: 46E35, 46B50, 58J99, 35B44, 35A25.

§1. Introduction

When the embedding of two Banach spaces $E \hookrightarrow F$ is continuous and not compact, the lack of compactness can be manifested by the (behavior in F of the) difference $u_k - u$ between the elements of a weakly convergent sequence $(u_k)_{k\in\mathbb{N}} \subset E$ and its weak limit u. Therefore one may call *defect of compactness* of $(u_k)_{k\in\mathbb{N}}$ the (sequences of) differences $u_k - u$ taken up to a suitable remainder that vanishes in the norm of F. (Note that, if the embedding is compact and E is reflexive, the defect of compactness is itself infinitesimal and so it can be identified with zero). For many embeddings there exist well-structured representations of the defect of compactness, known as *profile decompositions*. Best studied are profile decompositions relative to Sobolev embeddings, which are sums of terms with asymptotically disjoint supports, called *elementary concentrations* or *bubbles*. Profile decompositions were originally motivated by studies of concentration phenomena in PDE in the early 1980's by Uhlenbeck, Brezis, Coron, Nirenberg, Aubin and Lions, and they play a significant role in the verification process of the convergence of sequences of functions in applied analysis, particularly when the information available via the classical concentration-compactness method is not enough detailed.

Profile decompositions are known to exist when the embedding $E \hookrightarrow F$ is *cocompact* relative to some group $\mathcal{G}$ of isometries on E, see [11]. We recall that an embedding $E \hookrightarrow F$ is called cocompact relative to a group $\mathcal{G}$ of isometries ($\mathcal{G}$-cocompact for short) if any sequence $(u_k)_{k\in\mathbb{N}} \subset E$ such that $g_k(u_k) \rightharpoonup 0$ for any sequence of operators $(g_k)_{k\in\mathbb{N}} \subset \mathcal{G}$ turns out to be infinitesimal in the norm of F. (An elementary example due to Jaffard [7], which is easy to verify, is cocompactness of embedding of $\ell^\infty(\mathbb{Z})$ into itself relative to the group of shifts $\mathcal{G} := \{g_m := (a_n)_{n\in\mathbb{N}} \mapsto (a_{n+m})_{n\in\mathbb{N}} \mid m \in \mathbb{Z}\}$.) Up to the authors knowledge the first cocompactness result for functional spaces is [8, Lemma 6] by E. Lieb which expresses (using different terminology than the present note) that the nonhomogeneous Sobolev space $H^{1,p}(\mathbb{R}^N)$ is cocompactly embedded into $L^q(\mathbb{R}^N)$, when $N > p$ and $q \in (p, p^*)$ (where $p^* =$

$\frac{Np}{N-p}$), relative to the group of shifts $u \mapsto u(\cdot - y)$, $y \in \mathbb{R}^N$. A profile decomposition relative to a group $\mathcal{G}$ of bijective isometries on a Banach space E represents defect of compactness $u_k - u$ as a sum of *elementary concentrations*, or *bubbles*, namely $\sum_{n\in\mathbb{N}\setminus\{0\}} g_k^{(n)} w^{(n)}$ with some $g_k^{(n)} \in \mathcal{G}$ and $w^{(n)} \in E$. Note that in the above sum the index $n = 0$ is not allowed since, in the existing literature, usually $w^{(0)}$ represents the weak-limit u of the sequence and $(g_k^{(0)})_{k\in\mathbb{N}}$ is the constant sequence of constant value the identity map of the space. So, by using this convention, we can use defect of compactness to represent the sequence $(u_k)_{k\in\mathbb{N}}$ as a sum of $\sum_{n\in\mathbb{N}} g_k^{(n)} w^{(n)}$ and a remainder vanishing in F. In the above sums each of the elements $w^{(n)}$ (for $n \geq 1$), called *concentration profiles*, is obtained as the weak-limit (as $k \to \infty$) of the "deflated" sequence $((g_k^{(n)})^{-1}(u_k))_{k\in\mathbb{N}}$.

Typical examples of isometries groups $\mathcal{G}$, involved in profile decompositions, are the above mentioned group of shifts $u \mapsto u(\cdot - y)$ and the rescaling group, which is a product group of shifts and dilations $u \mapsto t^r u(t\cdot)$, $t > 0$, where, for instance, when u belongs to the homogeneous Sobolev space $\dot{H}^{s,p}(\mathbb{R}^N)$ ($N/s > p \geq 1$, $s > 0$), $r = r(p, s) = \frac{N-ps}{p}$.

Existence of profile decompositions for general bounded sequences in $\dot{H}^{1,p}(\mathbb{R}^N)$ (relative to the rescaling group) was proved by Solimini, see [10, Theorem 2], and independently, but with a weaker form of remainder, by Gérard in [6], with an extension to the case of fractional Sobolev spaces by Jaffard in [7]. Only in [9], for the first time, the authors observed that profile decomposition (and thus concentration phenomena in general) can be understood in functional-analytic terms, rather than in specific function spaces. Actually the results in [9] where extended in [11] to uniformly convex Banach spaces with the Opial condition (without the Opial condition profile decomposition still exists but weak convergence must be replaced by (a less-known) Delta convergence, see [4]). Finally the result has been extended up to a suitable class of metric spaces, see [5] and [3]. Despite the character of the statement in [11] is rather general, profile decompositions are still true, for instance, when the space E is not reflexive (e.g. [2]), or when one only has a semigroup of isometries (e.g. [1]), or when the profile decomposition can be expressed without the explicit use of a group (e.g. Struwe [12]) and so when [11, Theorem 2.10] does not apply.

The present paper generalizes, in the spirit of [10, Theorem 2], Struwe's result [12, Proposition 2.1] (which provides a profile decomposition for Palais-Smale sequences of particular functionals) to the case of general bounded sequences in $\dot{H}^{1,2}(M)$, where M is a smooth compact manifold in dimension $N \geq 3$.

The paper is organized as follows. In Section 2 we introduce some notation and state the main theorem of the paper and the result on which the related proof is based. In Section 3 we prove that the embedding $H^{1,2}(M) \hookrightarrow L^{2^*}(M)$ is cocompact with respect to a group of suitable transformations which are depending on the Atlas associated to the manifold. Section 4 is devoted to the proof of (the main) Theorem 1.

§2. Statement of the main result

Let $N \geq 3$ and let (M, g) be a compact smooth Riemannian N-dimensional manifold. We consider the Sobolev space $H^{1,2}(M)$ equipped with the norm defined by the quadratic form

of the Laplace-Beltrami operator,

$$\|u\|^2 = \int_M (|\mathrm{d}u|^2 + u^2)\mathrm{d}v_g, \tag{1}$$

(v_g denotes the Riemannian measure of the manifold). For every $y \in M$ we shall denote by $T_y(M)$ the tangent space in y to M, and by $\exp_y$ the exponential (local) map at the point y (defined on a suitable set $U_y \subset T_y(M)$ by setting, for all $v \in U_y$, $\exp_y(v) := \gamma_v(1)$ where γ_v is the unique geodesic, contained in M, such that $\gamma_v(0) = y$ and $\gamma_v'(0) = v$ and extended to the case $v = 0$ by setting $\exp_y(0) = y$). Since we will not use here any property of tangent bundles we will identify tangent spaces of M at different points with $\mathbb{R}^N$ and, for any $\rho > 0$, we shall denote by $B_\rho(0)$ the Euclidean N-dimensional ball centered at the origin with radius ρ. On the other hand, we shall denote by $\mathcal{B}_\rho(y)$ the open coordinate ball (i.e. the subset in M such that $\exp_y^{-1}(\mathcal{B}_\rho(y)) = B_\rho(0)$) with center y and radius $\rho > 0$. For the reader's convenience we recall that the injectivity radius ρ_y of a point $y \in M$ is the radius of the largest ball about the origin in $T_y(M)$ that can be mapped diffeomorfically via the map $\exp_y$, and that, the injectivity radius of the mainfold M, $\rho_M := \inf_{y\in M} \rho_y$. Since M is compact, ρ_M is strictly positive, so we can fix $0 < \rho < \frac{\rho_M}{3}$, moreover, there exists a finite set of points $(z_i)_{i\in I} \subset M$ such that $\left(\mathcal{B}_\rho(z_i), \exp_{z_i}^{-1}\right)_{i\in I}$ is a finite smooth atlas of M.

In what follows we shall fix $\chi \in C_0^\infty(B_\rho(0))$ so that, set for $i \in I$

$$\hat{\chi}_i := \hat{\chi}_{z_i} = \chi \circ \exp_{z_i}^{-1} \qquad \text{and} \qquad \chi_i := \frac{\hat{\chi}_i}{\sum_{j\in I} \hat{\chi}_j}, \tag{2}$$

$(\chi_i)_{i\in I}$ is a smooth partition of unity on M subordinated to the covering $(\mathcal{B}_\rho(z_i))_{i\in I}$. Then, since $\|u \circ \exp_{z_i}\|_{L^{2^*}(B_\rho(0))}$ is bounded by the $H^{1,2}(B_\rho(0))$-norm of $u \circ \exp_{z_i}$, the Sobolev embedding $H^{1,2}(M) \hookrightarrow L^{2^*}(M)$ can be deduced from the corresponding one on the Euclidean space (by the use of the fixed partition of unity $(\chi_i)_{i\in I}$). In fact, Theorem 1 below will provide a profile decomposition for bounded sequences in $H^{1,2}(M)$.

Finally we recall that the scalar product associated with (1) can be written with help of the partition of unity $(\chi_s)_{s\in I}$ in the following coordinate form:

$$\begin{aligned}\langle \Phi, \Psi \rangle := & \sum_{s\in I} \int_{B_\rho(0)} \sum_{i,j=1}^N g_{i,j}^{z_s} \partial_i((\chi_s\Phi)(\exp_{z_s}(\xi)))\partial_j(\Psi(\exp_{z_s}(\xi))) \sqrt{\det(g_{i,j}^{z_s})}\mathrm{d}\xi \\ & + \sum_{s\in I} \int_{B_\rho(0)} (\chi_s\Phi)(\exp_{z_s}(\xi))\Psi(\exp_{z_s}(\xi)) \sqrt{\det(g_{i,j}^{z_s})}\mathrm{d}\xi.\end{aligned} \tag{3}$$

Before stating the theorem, we warn the reader that, given a bounded sequence $(v_k)_{k\in\mathbb{N}} \subset H^{1,2}(B_\rho(0))$ and a vanishing sequence of positive numbers $(t_k)_{k\in\mathbb{N}}$, and setting $r = r(2) = \frac{N}{2^*} = \frac{N-2}{2}$, we will say (with a slight abuse on the definition of weak convergence) that the sequence $(t_k^r v_k(t_k\cdot))_{k\in\mathbb{N}}$ weakly converges to $v \in \dot{H}^{1,2}(\mathbb{R}^N)$ if for any $\varphi \in C_0^\infty(\mathbb{R}^N)$ such that $\operatorname{supp}\varphi \subset B_\rho(0)$

$$\int \varphi(x)\, t_k^r v_k(t_k x)\,\mathrm{d}x \longrightarrow \int \varphi(x)\, v(x)\,\mathrm{d}x \text{ as } k \to \infty.$$

Theorem 1. *Let M be a compact smooth Riemannian N-dimensional manifold ($N \geq 3$). Let $\rho \in (0, \frac{\rho_M}{3})$, let $\chi \in C_0^\infty(B_\rho(0))$, $\chi = 1$ on $B_{\frac{\rho}{2}}(0)$, and let $(\chi_i)_{i\in I}$, defined by (2), be a smooth partition of unity on M subordinated to the covering $(\mathcal{B}_\rho(z_i))_{i\in I}$. Then, given a bounded sequence $(u_k)_{k\in\mathbb{N}}$ in $H^{1,2}(M)$ and, with $r = \frac{N}{2^*} = \frac{N-2}{2}$, there exist:*

- *a sequence $\left(Y^{(n)}\right)_{n\in\mathbb{N}\setminus\{0\}}$ of sequences $Y^{(n)} := \left(y_k^{(n)}\right)_{k\in\mathbb{N}} \subset M$, $y_k^{(n)} \to \bar{y}^{(n)} \in M$,*
- *a sequence $\left(J^{(n)}\right)_{n\in\mathbb{N}\setminus\{0\}}$ of sequences $J^{(n)} := \left(j_k^{(n)}\right)_{k\in\mathbb{N}} \subset \mathbb{R}_+$,*
- *a sequence $\left(w^{(n)}\right)_{n\in\mathbb{N}\setminus\{0\}}$ of functions (profiles) $w^{(n)} \in \dot{H}^{1,2}(\mathbb{R}^N)$,*

such that, modulo subsequences,

$$j_k^{(n)} \longrightarrow \infty \text{ as } k \to \infty \quad \forall n \in \mathbb{N}\setminus\{0\}, \tag{4}$$

$$|j_k^{(n)} - j_k^{(m)}| + 2^{j_k^{(n)}} d(y_k^{(n)}, y_k^{(m)}) \to \infty \text{ whenever } m \neq n, \tag{5}$$

$$2^{-j_k^{(n)} r} u_k \circ \exp_{y_k^{(n)}}(2^{-j_k^{(n)}} \cdot) \rightharpoonup w^{(n)} \text{ in } \dot{H}^{1,2}(\mathbb{R}^N) \text{ as } k \to \infty. \tag{6}$$

Moreover, setting for all $k \in \mathbb{N}$

$$\mathcal{S}_k(x) := \sum_{n\in\mathbb{N}\setminus\{0\}} 2^{j_k^{(n)} r}\, \chi \circ \exp_{y_k^{(n)}}^{-1}(x)\, w^{(n)}\left(2^{j_k^{(n)}} \exp_{y_k^{(n)}}^{-1}(x)\right),\ x \in M, \tag{7}$$

the series $\mathcal{S}_k \in \dot{H}^{1,2}(M)$ are unconditionally convergent (with respect to n) and the sequence $(\mathcal{S}_k)_{k\in\mathbb{N}}$ is uniformly convergent (with respect to k) in $\dot{H}^{1,2}(M)$, in addition

$$u_k - u - \mathcal{S}_k \to 0 \text{ in } L^{2^*}(M)\,. \tag{8}$$

Finally the following energy bound holds

$$\sum_{n\in\mathbb{N}\setminus\{0\}} \left\|\nabla w^{(n)}\right\|^2_{L^2(\mathbb{R}^N)} + \|u\|^2_{H^{1,2}(M)} \leq \liminf_{k\to\infty} \|u_k\|^2_{H^{1,2}(M)}\,. \tag{9}$$

We want to emphasize that (8) states that, modulo subsequence, the defect of compactness $u_k - u$ of the bounded sequence $(u_k)_{k\in\mathbb{N}}$ (which, modulo subsequence, weakly converges to u) has a representation given (up to a remainder which vanishes in the norm of $L^{2^*}(M)$) by the clearly structured terms in $\mathcal{S}_k$.

The proof of this theorem is based on the following easy corollary to Solimini's profile decomposition [10, Theorem 2].

Theorem 2. *Given $m \in \mathbb{N}\setminus\{0\}$ and $1 < p < \frac{N}{m}$ let $r = \frac{N}{p^*(m)} = \frac{N-mp}{p}$. Let $(v_k)_{k\in\mathbb{N}}$ be a bounded sequence in the homogeneous Sobolev space $\dot{H}^{m,p}(\mathbb{R}^N)$ supported on a compact set $K \subset \mathbb{R}^N$. Then, there exists a (renamed) subsequence (s.t. $v_k \rightharpoonup v$) whose defect of compactness $v_k - v$ has the form*

$$S_k = \sum_{n\in\mathbb{N}\setminus\{0\}} 2^{j_k^{(n)} r} w^{(n)}(2^{j_k^{(n)}}(\cdot - \xi_k^{(n)})), \tag{10}$$

where, for any $n \in \mathbb{N}\setminus\{0\}$, $\Xi^{(n)} := (\xi_k^{(n)})_{k\in\mathbb{N}} \subset K$, and $J^{(n)} := (j_k^{(n)})_{k\in\mathbb{N}} \subset \mathbb{R}$ are such that $j_k^{(n)} \to +\infty$ as $k \to \infty$ and $w^{(n)}$ is the weak limit of the sequence $\left(2^{-j_k^{(n)} r} v_k(2^{-j_k^{(n)}} \cdot + \xi_k^{(n)})\right)_{k\in\mathbb{N}}$. Moreover the addenda are asymptotically mutually orthogonal, i.e.

$$|j_k^{(n)} - j_k^{(m)}| + 2^{j_k^{(n)}} |\xi_k^{(n)} - \xi_k^{(m)}| \to \infty \text{ whenever } m \neq n. \tag{11}$$

Proof. We shall assume, without restrictions, that $u_k \rightharpoonup 0$. According to the profile decomposition result [10, Theorem 2], modulo the extraction of a subsequence, each term v_k has concentration terms (depending on n) of the following shape

$$c_k^n := 2^{j_k^{(n)} r} w^{(n)}(2^{j_k^{(n)}}(\cdot - \xi_k^{(n)})) \tag{12}$$

for some $\xi_k^{(n)} \in \mathbb{R}^N$, $j_k^{(n)} \in \mathbb{R}$ where $w^{(n)}$ is obtained as the weak limit of the sequence $\left(2^{-j_k^{(n)} r} v_k(2^{-j_k^{(n)}} \cdot + \xi_k^{(n)})\right)_{k\in\mathbb{N}}$. We claim that the sequence $J^{(n)}$ is bounded from below. Indeed, on the contrary, the assumption $j_k^{(n)} \to -\infty$ as $k \to \infty$ would imply, since v_k has a bounded support, that

$$\left\| 2^{-j_k^{(n)} r} v_k \left(2^{-j_k^{(n)}} \cdot + \xi_k^{(n)}\right) \right\|_p \to 0 \text{ as } k \to \infty,$$

and so that $w^{(n)} = 0$.

As a consequence $\xi_k^{(n)} \in K$ for k large enough. Note also that $J^{(n)}$ cannot have any bounded subsequence, since otherwise $(v_k)_{k\in\mathbb{N}}$ should have a nonzero weak limit, in contradiction to our assumptions.

Finally, condition (11) is the condition of asymptotic orthogonality (decoupling) of bubbles from [10]. □

§3. Cocompactness in Sobolev spaces of compact manifolds

The Sobolev embedding $H^{1,2}(M) \hookrightarrow L^{2^*}(M)$ has the following property of cocompactness type.

Theorem 3. *Let M be a compact smooth Riemannian N-dimensional manifold ($N \geq 3$), and $0 < \rho < \frac{\rho_M}{3}$. Let $\left(\mathcal{B}_\rho(z_i), \exp_{z_i}^{-1}\right)_{i\in I}$ be a finite smooth atlas of M and let $\boldsymbol{\chi} \in C_0^\infty(B_\rho(0))$ so that $(\chi_i)_{i\in I}$, defined by* (2)*, is a smooth partition of unity on M subordinated to the covering $(\mathcal{B}_\rho(z_i))_{i\in I}$. Set $r = r(2) = \frac{N}{2^*} = \frac{N-2}{2}$. If $(u_k)_{k\in\mathbb{N}}$ is any bounded sequence in $H^{1,2}(M)$ such that for every $i \in I$, $(y_k)_{k\in\mathbb{N}} \subset \mathcal{B}_\rho(z_i)$, and $(j_k)_{k\in\mathbb{N}} \subset \mathbb{N}$ such that $j_k \to +\infty$*

$$2^{-j_k r}(\chi_i u_k) \circ \exp_{y_k}(2^{-j_k} \cdot) \rightharpoonup 0 \text{ as } k \to \infty, \tag{13}$$

then $u_k \rightharpoonup 0$ in $L^{2^*}(M)$.

Proof. We claim that for all sequences $(\xi_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ and $(j_k)_{k\in\mathbb{N}} \subset \mathbb{N}$ such that $j_k \to +\infty$ and for every $i \in I$ we have

$$2^{-j_k r}(\chi_i u_k) \circ \exp_{z_i}(2^{-j_k} \cdot + \xi_k) \rightharpoonup 0 \quad \text{as } k \to \infty. \tag{14}$$

Since (14) is obviously true when $|\xi_k| \geq \rho$, (indeed the terms in (14) are identically zero for k large enough), we shall assume $\xi_k \in B_\rho(0)$ for all $k \in \mathbb{N}$. Given $i \in I$, we set $y_k := \exp_{z_i}(\xi_k) \in M$ and denote by ψ_k the transition map between the charts $(\mathcal{B}_\rho(z_i), \exp_{z_i}^{-1})$ and $(\mathcal{B}_\rho(y_k), \exp_{y_k}^{-1})$ i.e. we set $\psi_k := \exp_{y_k}^{-1} \circ \exp_{z_i}$ (so that $\exp_{z_i} = \exp_{y_k} \circ \psi_k$ and $\psi_k(\xi_k) = 0$). Therefore, for k large enough, by using Taylor expansion of the first order at ξ_k (where, for a lighter notation, we denote by $\psi_k'(\xi_k)$ the Jacobi matrix of ψ_k at ξ_k $(\psi_k'(\xi_k))^{-1}$ its inverse and by $|(\psi_k'(\xi_k))^{-1}|$ the

corresponding Jacobian, and drop the dot symbol for the rows-by-columns product) we get, since $j_k \to +\infty$, that

$$\begin{aligned} 2^{-j_k r}(\chi_i u_k)(\exp_{z_i}(2^{-j_k}\xi + \xi_k)) &= 2^{-j_k r}(\chi_i u_k)(\exp_{y_k} \circ \psi_k)(2^{-j_k}\xi + \xi_k) \\ &= 2^{-j_k r}(\chi_i u_k)(\exp_{y_k}(2^{-j_k}(\psi_k'(\xi_k) + o(1))\xi)). \end{aligned} \tag{15}$$

(we are using the Landau symbol $o(1)$ to denote any (matrix valued) function uniformly convergent to zero). In correspondence to any test function $\varphi \in C_0^\infty(\mathbb{R}^N)$,

$$\begin{aligned} &\int_{B_{2\rho}(0)} \varphi(\xi) 2^{-j_k r}\left[(\chi_i u_k) \circ \exp_{z_i}(2^{-j_k}\xi + \xi_k) - (\chi_i u_k) \circ \exp_{y_k}(2^{-j_k}\psi_k'(\xi_k)\xi)\right] \mathrm{d}\xi \\ &= \int_{B_{2\rho}(0)} \varphi(\xi) 2^{-j_k r}\left[(\chi_i u_k) \circ \exp_{y_k} \circ \psi_k(2^{-j_k}\xi + \xi_k) - (\chi_i u_k) \circ \exp_{y_k}(2^{-j_k}\psi_k'(\xi_k)\xi)\right] \mathrm{d}\xi \\ &= |(\psi_k'(\xi_k))^{-1}| 2^{j_k \frac{N+2}{2}} \int_{|\eta|<C2^{-j_k}} \varphi(2^{j_k}(\psi_k'(\xi_k))^{-1}\eta) \\ &\qquad \times \left[(\chi_i u_k) \circ \exp_{y_k}\left(\psi_k((\psi_k'(\xi_k))^{-1}\eta + \xi_k\right) - (\chi_i u_k) \circ \exp_y(\eta)\right] \mathrm{d}\eta \\ &= |(\psi_k'(\xi_k))^{-1}| 2^{j_k \frac{N+2}{2}} \int_0^1 \mathrm{d}s \int_{|\eta|<C2^{-j_k}} \varphi(2^{j_k}(\psi_k'(\xi_k))^{-1}\eta) \\ &\qquad \times \nabla\Big((\chi_i u_k) \circ \exp_{y_k}(s\psi_k\left((\psi_k'(\xi_k))^{-1}\eta + \xi_k\right) + (1-s)\eta)\Big) \cdot (\psi_k((\psi_k'(\xi_k))^{-1}\eta + \xi_k) - \eta) \mathrm{d}\eta, \end{aligned}$$

(the second equality holds by integrating with respect to the variable $\eta = 2^{-j_k}\psi_k'(\xi_k)\xi$). Set, for each $s \in [0,1]$, $\zeta := s\psi_k\left((\psi_k'(\xi_k))^{-1}\eta + \xi_k\right) + (1-s)\eta$, since for $\eta \to 0$, $\zeta = \eta + O(|\eta|^2)$ and since the Jacobian of the transformation is close to 1 in the domain of integration, the modulus of the last expression is bounded by the following one, which, in turn, can be estimated by Cauchy inequality. So, we have

$$\begin{aligned} &C2^{j_k \frac{N+2}{2}} \int_{|\zeta|<C2^{-j_k}} \varphi(2^{j_k}(\psi_k'(\xi_k))^{-1}\eta(\zeta))|\nabla(\chi_i u_k) \circ \exp_{y_k}(\zeta)||\zeta|^2 \mathrm{d}\zeta \\ &\le C2^{j_k \frac{N+2}{2}} \|\nabla(\chi_i u_k) \circ \exp_{y_k}\|_2 \left(\int_{|\zeta|<C2^{-j_k}} |\varphi(2^{j_k}(\psi_k'(\xi_k))^{-1}\eta(\zeta))|^2 |\zeta|^4 \mathrm{d}\zeta\right)^{\frac{1}{2}} \\ &\le C2^{j_k \frac{N+2}{2}} \|u_k\|_{H^{1,2}(M)} \left(\int_{|\xi|<C} |\varphi(\xi)|^2 2^{-4j_k} |\xi|^4 2^{-j_k N} \mathrm{d}\xi\right)^{\frac{1}{2}} \le C2^{-j_k} \longrightarrow 0. \end{aligned}$$

Therefore, by taking into account (15), we deduce that both sequences $\left(2^{-j_k r}(\chi_i u_k)(\exp_{y_k}(2^{-j_k}\cdot))\right)_{k\in\mathbb{N}}$ and $\left(2^{-j_k r}(\chi_i u_k)(\exp_{z_i}(2^{-j_k}\ \cdot + \xi_k))\right)_{k\in\mathbb{N}}$ have the same weak limit and, since (13) holds true, (14) holds too.

Consequently, from the cocompactness of the embedding $\dot{H}^{1,2}(\mathbb{R}^N) \hookrightarrow L^{2^*}(\mathbb{R}^N)$ ([10, Theorem 1]), it follows that for every $i \in I$,

$$(\chi_i u_k) \circ \exp_{z_i} \to 0 \quad \text{in } L^{2^*}(\mathbb{R}^N) \text{ as } k \to \infty, \tag{16}$$

and therefore, since $(\chi_i)_{i\in I}$ is a partition of unity subordinated to the atlas $\left(\mathcal{B}_\rho(z_i), \exp_{z_i}^{-1}\right)_{i\in I}$, we deduce that

$$\int_M |u_k|^{2^*} \mathrm{d}v_g = \int_M \left|\sum_{i\in I} \chi_i u_k\right|^{2^*} \mathrm{d}v_g \le C \sum_{i\in I} \int_{\mathcal{B}_\rho(z_i)} |\chi_i u_k|^{2^*} \mathrm{d}v_g$$
$$\le C \sum_{i\in I} \int_{B_\rho(0)} |u_k \circ \exp_{z_i}(\xi)|^{2^*} \mathrm{d}\xi \to 0,$$

which proves the statement of the theorem. □

§4. Proof of Theorem 1 (profile decomposition)

1. Without loss of generality we may assume (by replacing u_k with $u_k - u$) that $u_k \rightharpoonup 0$.
Then, setting for all $i \in I$

$$v_{k,i} := (\chi_i u_k) \circ \exp_{z_i} \tag{17}$$

we get that the sequence $(v_{k,i})_{k\in\mathbb{N}}$ is bounded in $H_0^{1,2}(B_\rho(0))$ (and weakly converges to zero), and so we can consider a profile decomposition of $(v_{k,i})_{k\in\mathbb{N}}$ given by Theorem 2 when $m = 1$ and $r = \frac{N-2}{2}$. An iterated extraction allows to find a subsequence which has a profile decomposition for every $i \in I$ i.e. such that for all $i \in I$ the defect of compactness of $v_{k,i}$ has the following form

$$S_{k,i} = \sum_{n\in\mathbb{N}\setminus\{0\}} 2^{j_{k,i}^{(n)} r} w_i^{(n)} \left(2^{j_{k,i}^{(n)}} \left(\cdot - \xi_{k,i}^{(n)}\right)\right) =: \sum_{n\in\mathbb{N}\setminus\{0\}} c_{k,i}^{(n)}. \tag{18}$$

By taking into account (17) we will be able to get concentration terms of $\chi_i u_k$ by composing each concentration term $c_{k,i}^{(n)}$ of $v_{k,i}$ with $\exp_{z_i}^{-1}$. More in detail we consider for all $i \in I$ the term, defined on $\mathcal{B}_\rho(z_i)$,

$$C_{k,i}^{(n)} := c_{k,i}^{(n)} \circ \exp_{z_i}^{-1} = 2^{j_{k,i}^{(n)} r} w_i^{(n)} \left(2^{j_{k,i}^{(n)}} \left(\exp_{z_i}^{-1}(\cdot) - \xi_{k,i}^{(n)}\right)\right). \tag{19}$$

Setting

$$y_{k,i}^{(n)} := \exp_{z_i}(\xi_{k,i}^{(n)}) \tag{20}$$

we have that

$$C_{k,i}^{(n)} = 2^{j_{k,i}^{(n)} r} w_i^{(n)} \left(2^{j_{k,i}^{(n)}} \left(\exp_{z_i}^{-1}(\cdot) - \exp_{z_i}^{-1}(y_{k,i}^{(n)})\right)\right). \tag{21}$$

Since for all $i \in I$ and $n \in \mathbb{N} \setminus \{0\}$

$$w_i^{(n)} := \operatorname*{w\text{-}lim}_{k\to\infty} 2^{-j_{k,i}^{(n)} r} (\chi_i u_k) \circ \exp_{z_i}\left(2^{-j_{k,i}^{(n)}} \cdot + \xi_{k,i}^{(n)}\right), \tag{22}$$

we can see that $w_i^{(n)}$ "evaluates" $\chi_i u_k$ on points belonging to $\mathcal{B}_\rho(z_i)$ which are mapped by $\exp_{z_i}^{-1}$ in subsets of $B_\rho(0)$ which are (for large k) concentrated around the points $\xi_{k,i}^{(n)}$. So, due to (20), it is sufficient to evaluate $w_i^{(n)}$ on points which belong also to $\mathcal{B}_\rho(y_{k,i}^{(n)})$. So, setting

$$B_{i,k,n} := \exp_{y_{k,i}^{(n)}}^{-1}(\mathcal{B}_\rho(y_{k,i}^{(n)}) \cap \mathcal{B}_\rho(z_i)) \subset B_\rho(0), \tag{23}$$

we shall consider the transition map between the charts $(\mathcal{B}_\rho(y_{k,i}^{(n)}), \exp_{y_{k,i}^{(n)}}^{-1})$ and $(\mathcal{B}_\rho(z_i), \exp_{z_i}^{-1})$, i.e. the map

$$\psi_{i,k,n} := \exp_{z_i}^{-1} \circ \exp_{y_{k,i}^{(n)}} \tag{24}$$

defined on $B_{i,k,n}$. Note that $\psi_{i,k,n}(0) = \xi_{k,i}^{(n)}$, moreover, by setting for any $x \in B_{i,k,n}$

$$\eta := 2^{j_{k,i}^{(n)}} \exp_{y_{k,i}^{(n)}}^{-1}(x), \tag{25}$$

we have $\exp_{z_i}^{-1}(x) = \psi_{i,k,n}(2^{-j_{k,i}^{(n)}}\eta)$ for all $x \in B_{i,k,n}$. Therefore (by using Taylor expansion of the first order of the transition map $\psi_{i,k,n}$ at 0, where, to use a lighter notation we denote by $\psi'_{i,k,n}(0)$ the Jacobi matrix of $\psi_{i,k,n}$ at zero, $(\psi'_{i,k,n}(0))^{-1}$ its inverse and omit the dot symbol for the rows-by-columns product) we deduce

$$\begin{aligned} 2^{j_{k,i}^{(n)}}\left(\exp_{z_i}^{-1}(x) - \xi_{k,i}^{(n)}\right) &= 2^{j_{k,i}^{(n)}}\left(\psi_{i,k,n}(2^{-j_{k,i}^{(n)}}\eta) - \xi_{k,i}^{(n)}\right) = 2^{j_{k,i}^{(n)}}\left(\psi_{i,k,n}(2^{-j_{k,i}^{(n)}}\eta) - \psi_{i,k,n}(0)\right) \\ &= \psi'_{i,k,n}(0)\eta + O(2^{-j_{k,i}^{(n)}}\eta^2) = 2^{j_{k,i}^{(n)}}\psi'_{i,k,n}(0)\exp_{y_{k,i}^{(n)}}^{-1}(x) + O\left(2^{j_{k,i}^{(n)}}\left(\exp_{y_{k,i}^{(n)}}^{-1}(x)\right)^2\right). \end{aligned} \tag{26}$$

Without loss of generality, applying Arzelà-Ascoli theorem and passing to a suitable subsequence, we can assume that $(\psi_{i,k,n})_{k\in\mathbb{N}}$ converges in the norm of $C^1(\mathbb{R}^N)$ as $k \to \infty$ to some function $\psi_{i,n}$. We claim that, under a suitable renaming of the profile $w_i^{(n)}$, namely by renaming $w_i^{(n)}(\psi'_{i,n}(0)\ \cdot)$ as $w_i^{(n)}$, concentration terms $C_{k,i}^{(n)}$ (of $\chi_i u_k$) in (19) take the following form:

$$\tilde{C}_{k,i}^{(n)} := 2^{j_{k,i}^{(n)} r} w_i^{(n)}\left(2^{j_{k,i}^{(n)}}\exp_{y_{k,i}^{(n)}}^{-1}(\cdot)\right). \tag{27}$$

For this purpose we show that, as $k \to \infty$,

$$\int_{\mathcal{B}_\rho(y_{k,i}^{(n)})\cap\mathcal{B}_\rho(z_i)} \left|2^{j_{k,i}^{(n)} r}\mathrm{d}\left(w_i^{(n)}\left(2^{j_{k,i}^{(n)}}(\exp_{z_i}^{-1}(x) - \xi_{k,i}^{(n)})\right) - w_i^{(n)}\left(2^{j_{k,i}^{(n)}}\psi'_{i,n}(0)\exp_{y_{k,i}^{(n)}}^{-1}(x)\right)\right)\right|^2 \mathrm{d}v_g \to 0. \tag{28}$$

Indeed, the previous relation written under the coordinate map $\exp_{y_{k,i}^{(n)}}$, i.e. by setting $\xi = \exp_{y_{k,i}^{(n)}}^{-1}(x)$ becomes (by taking into account (24) and (23))

$$\int_{B_{i,k,n}} \left|2^{j_{k,i}^{(n)} r}\nabla\left(w_i^{(n)}\left(2^{j_{k,i}^{(n)}}(\psi_{i,k,n}(\xi) - \xi_{k,i}^{(n)})\right) - w_i^{(n)}\left(2^{j_{k,i}^{(n)}}\psi'_{i,n}(0)\xi\right)\right)\right|^2 \mathrm{d}\xi \to 0 \text{ as } k \to \infty,$$

and, by taking into account (25) (and by a null extension to whole of $\mathbb{R}^N$ of the involved functions), the claim will follow if, as $k \to \infty$,

$$2^{-j_{k,i}^{(n)}\frac{N+2}{2}} \int_{\mathbb{R}^N} \left|\psi'_{i,k,n}(2^{-j_{k,i}^{(n)}}\eta)\nabla w_i^{(n)}\left(2^{j_{k,i}^{(n)}}(\psi_{i,k,n}(2^{-j_{k,i}^{(n)}}\eta) - \xi_{k,i}^{(n)})\right) - \psi'_{i,n}(0)\nabla w_i^{(n)}(\psi'_{i,n}(0)\eta)\right|^2 \mathrm{d}\eta \to 0.$$

This last convergence easily follows by Lebesgue dominated convergence theorem, indeed (for all n and for all i) $\nabla w_i^{(n)} \in L^2(\mathbb{R}^N)$, and when $k \to \infty$, we have $j_{k,i}^{(n)} \to +\infty$, and (by taking

into account that convergence of $(\psi_{i,k,n})_{k\in\mathbb{N}}$ and $\left(\psi'_{i,k,n}\right)_{k\in\mathbb{N}}$ to $\psi_{i,n}$ and $\psi'_{i,n}$ respectively is uniform) the pointwise convergence of $\psi'_{i,k,n}(2^{-j^{(n)}_{k,i}}\eta)\to\psi'_{i,n}(0)$, $2^{j^{(n)}_{k,i}}\left(\psi_{i,k,n}(2^{-j^{(n)}_{k,i}}\eta)-\xi^{(n)}_i\right)\to\psi'_{i,n}(0)\eta$ (as easily follows by (26) and (25)).

It is easy to see now that the renamed profiles $w^{(n)}_i$ are obtained as pointwise limits (and thus also as weak limits)

$$w^{(n)}_i(\xi)=\lim_{k\to\infty}2^{-j^{(n)}_{k,i}r}(\chi_i u_k)\circ\exp_{y^{(n)}_{k,i}}\left(2^{-j^{(n)}_{k,i}}\xi\right),\ \text{for a.e. } \xi\in\mathbb{R}^N. \tag{29}$$

2. Since each $\overline{\mathcal{B}}_\rho(z_i)\subset\mathcal{B}_{2\rho}(z_i)\subset M$ and M is compact, we may assume that for all $n\in\mathbb{N}\setminus\{0\}$ and for all $i\in I$, there exist, up to subsequences, points of concentration

$$\overline{y}^{(n)}_i:=\lim_{k\to\infty}y^{(n)}_{k,i}. \tag{30}$$

In order to achieve the orthogonality relation (5) we shall introduce the following equivalence relation on the set of sequences in $M\times\mathbb{R}$. Namely given $(y_k,j_k)_{k\in\mathbb{N}}$ and $\left(y'_k,j'_k\right)_{k\in\mathbb{N}}$ in $M\times\mathbb{Z}$ we shall write

$$(y_k,j_k)_{k\in\mathbb{N}}\simeq\left(y'_k,j'_k\right)_{k\in\mathbb{N}}\ \text{ when }\ \left(|j_k-j'_k|+2^{j_k}d(y_k,y'_k)\right)_{k\in\mathbb{N}}\text{ is a bounded sequence.}\qquad(\mathcal{R})$$

Since the set I is a finite set, the number of sequences $\left(y^{(n)}_{k,i},j^{(n)}_{k,i}\right)_{k\in\mathbb{N}}$ which can be equivalent to a fixed sequence $\left(y^{(\bar n)}_{k,\bar\imath},j^{(\bar n)}_{k,\bar\imath}\right)_{k\in\mathbb{N}}$ is finite. Therefore we can exploit the unconditional convergence with respect to the indexes (n) of the series $S_{k,i}$ and synchronize them by replacing $\bar n$ and all the indexes m in the finite set

$$\mathcal{N}_{\bar n}:=\left\{m\in\mathbb{N}\setminus\{0\}\mid\exists i\in I\text{ s.t. }\left(y^{(n)}_{k,i},j^{(n)}_{k,i}\right)_{k\in\mathbb{N}}\simeq\left(y^{(\bar n)}_{k,\bar\imath},j^{(\bar n)}_{k,\bar\imath}\right)_{k\in\mathbb{N}}\right\} \tag{31}$$

with, say, the smallest integer in $\mathcal{N}_{\bar n}$.

Thanks to this synchronization procedure the following property

$$\left(y^{(n)}_{k,i_1},j^{(n)}_{k,i_1}\right)_{k\in\mathbb{N}}\simeq\left(y^{(m)}_{k,i_2},j^{(m)}_{k,i_2}\right)_{k\in\mathbb{N}}\quad\Longleftrightarrow\quad m=n, \tag{32}$$

holds true for all $i_1,i_2\in I$ and $m,n\in\mathbb{N}\setminus\{0\}$.

Note also that when $\left(y^{(n)}_{k,i_1},j^{(n)}_{k,i_1}\right)_{k\in\mathbb{N}}\simeq\left(y^{(n)}_{k,i_2},j^{(n)}_{k,i_2}\right)$, since $\left(\left|j^{(n)}_{k,i_2}-j^{(n)}_{k,i_1}\right|\right)_{k\in\mathbb{N}}$ is bounded, we can set, modulo subsequences

$$j(i_1,i_2,n):=\lim_{k\to+\infty}j^{(n)}_{k,i_2}-j^{(n)}_{k,i_1}\in\mathbb{R}, \tag{33}$$

so that, by redefining $w^{(n)}_{i_2}(2^{-j(i_1,i_2,n)}\cdot)$ as (the corresponding profile) $w^{(n)}_{i_2}$, we can assume that $\left(j^{(n)}_{k,i_2}\right)_{k\in\mathbb{N}}=\left(j^{(n)}_{k,i_1}\right)_{k\in\mathbb{N}}$. Moreover, since also $\left(2^{j^{(n)}_{k,i_1}}d\left(y^{(n)}_{k,i_1},y^{(n)}_{k,i_2}\right)\right)_{k\in\mathbb{N}}$ is bounded, we get (by (4)) that (see (30))

$$\bar y^{(n)}_{i_1}=\bar y^{(n)}_{i_2}\quad\text{ for all }\left(y^{(n)}_{k,i_1},j^{(n)}_{k,i_1}\right)_{k\in\mathbb{N}}\simeq\left(y^{(n)}_{k,i_2},j^{(n)}_{k,i_2}\right)_{k\in\mathbb{N}}. \tag{34}$$

Finally, we show that the elementary concentrations terms $C^{(n)}_{k,i}$ do not change (up to a vanishing term) by varying $\left(y^{(n)}_{k,i}, j^{(n)}_{k,i}\right)_{k\in\mathbb{N}}$ in the same equivalence class. Namely the following property holds true

$$\left(y^{(n)}_{k,i_1}, j^{(n)}_{k,i_1}\right)_{k\in\mathbb{N}} \simeq \left(y^{(n)}_{k,i_2}, j^{(n)}_{k,i_2}\right)_{k\in\mathbb{N}} \quad \Rightarrow \quad \left\|C^{(n)}_{k,i_1} - C^{(n)}_{k,i_2}\right\| \to 0, \tag{35}$$

for all $i_1, i_2 \in I$. Since, as shown above, we can assume, without restrictions, that $\left(j^{(n)}_{k,i_1}\right)_{k\in\mathbb{N}} = \left(j^{(n)}_{k,i_2}\right)_{k\in\mathbb{N}}$ (and we shall denote, to shorten notation, their common value as $\left(j^{(n)}_k\right)_{k\in\mathbb{N}}$) it will suffice to prove that, set $\bar\xi^{(n)}_{k,i_1} = \exp^{-1}_{z_{i_1}} y^{(n)}_{k,i_1}$ and $\bar\xi^{(n)}_{k,i_2} = \exp^{-1}_{z_{i_1}} y^{(n)}_{k,i_2}$, we have

$$\int_{\mathcal{B}_\rho(z_{i_1})} \left|2^{j^{(n)}_k r} \mathrm{d}\left(w^{(n)}_{i_1}\left(2^{j^{(n)}_k}(\exp^{-1}_{z_{i_1}}(x) - \bar\xi^{(n)}_{k,i_2})\right) - w^{(n)}_{i_1}\left(2^{j^{(n)}_k}(\exp^{-1}_{z_{i_1}}(x) - \bar\xi^{(n)}_{k,i_1})\right)\right)\right|^2 \mathrm{d}v_g \to 0 \text{ as } k\to\infty. \tag{36}$$

Indeed, by (20), we get, modulo subsequences, that

$$\begin{aligned} 2^{j^{(n)}_k}|\bar\xi^{(n)}_{k,i_2} - \bar\xi^{(n)}_{k,i_1}| &= 2^{j^{(n)}_k}|\exp^{-1}_{z_{i_1}} y^{(n)}_{k,i_2} - \exp^{-1}_{z_{i_1}} y^{(n)}_{k,i_1}| \\ &= 2^{j^{(n)}_k}|d(y^{(n)}_{k,i_2}, z_{i_1}) - (y^{(n)}_{k,i_1}, z_{i_1})| \le 2^{j^{(n)}_k} d(y^{(n)}_{k,i_2}, y^{(n)}_{k,i_1}) \to 0. \end{aligned}$$

Then, (5) follows directly from (34).

3. Consider now the sum $\sum_{n\in\mathbb{N}\setminus\{0\}}\sum_{i\in I}\tilde C^{(n)}_{k,i}$, with the sequences $y^{(n)}_{k,i}$ and $j^{(n)}_{k,i}$, which are synchronized at the Step 2 as $y^{(n)}_k$ and $j^{(n)}_k$, while $y^{(n)}_k \to \bar y^{(n)}$ and (29) takes form

$$w^{(n)}_i(\xi) = \lim_{k\to\infty} 2^{-j^{(n)}_k r}(\chi_i u_k) \circ \exp_{y^{(n)}_k}\left(2^{-j^{(n)}_k}\xi\right), \text{ for a.e. } \xi\in\mathbb{R}^N. \tag{37}$$

The latter yields for a.e. $\xi\in\mathbb{R}^N$, since $j^{(n)}_k \to \infty$ implies $\exp_{y^{(n)}_k}\left(2^{-j^{(n)}_k}\xi\right) \to \bar y^{(n)}$ in M,

$$w^{(n)}_i(\xi) = \chi_i(\bar y^{(n)}) \lim_{k\to\infty} 2^{-j^{(n)}_k r} u_k \circ \exp_{y^{(n)}_k}\left(2^{-j^{(n)}_k}\xi\right), \text{ for a.e. } \xi\in\mathbb{R}^N, \tag{38}$$

taking into account that for each $\xi\in\mathbb{R}^N$ the limit is evaluated with $k \ge k(\xi)$ with some $k(\xi)$ sufficiently large. Set

$$w^{(n)} := \sum_{i\in I} w^{(n)}_i. \tag{39}$$

Then relation (6) immediately follows from (38), $w^{(n)}_i = \chi_i(\bar y^{(n)})w^{(n)}$, and since, by Step 1, defect of compactness of $\chi_i u_k$ is a unconditionally convergent series, we have

$$\begin{aligned} \sum_{i\in I}\sum_{n\in\mathbb{N}\setminus\{0\}} \tilde C^{(n)}_{k,i}(x) &= \sum_{n\in\mathbb{N}\setminus\{0\}}\sum_{i\in I} \tilde C^{(n)}_{k,i}(x) \\ &= \sum_{n\in\mathbb{N}\setminus\{0\}}\sum_{i\in I} 2^{j^{(n)}_k r} w^{(n)}_i\left(2^{j^{(n)}_k}\exp^{-1}_{y^{(n)}_k}(x)\right) = \sum_{n\in\mathbb{N}\setminus\{0\}} w^{(n)}\left(2^{j^{(n)}_k}\exp^{-1}_{y^{(n)}_k}(x)\right), \ x\in M. \end{aligned}$$

Not now, which gives (7).

4. In order to prove the "energy" estimate (9), assume, without loss of generality, that the sum in (7) is finite and that all $w^{(n)}$ have compact support, and expand by bilinearity the trivial inequality $\|u-u_k+\mathcal{S}_k\|^2_{H^{1,2}(M)}\geq 0$. Then, by using the norm (1) and the representation (3) of the scalar product in $H^{1,2}(M)$, we have

$$\begin{aligned}0\leq{}&\|u_k\|^2+\|u\|^2-2\langle u_k,u\rangle+2\langle u-u_k,\mathcal{S}_k\rangle\\&+\sum_n\|2^{j_k^{(n)}r}\,\chi\circ\exp^{-1}_{y_k^{(n)}}\,w^{(n)}\left(2^{j_k^{(n)}}\exp^{-1}_{y_k^{(n)}}(\cdot)\right)\|^2\\&-\sum_{m\neq n}\left\langle 2^{j_k^{(m)}r}\,\chi\circ\exp^{-1}_{y_k^{(m)}}\,w^{(m)}\left(2^{j_k^{(m)}}\exp^{-1}_{y_k^{(m)}}(\cdot)\right),2^{j_k^{(n)}r}\,\chi\circ\exp^{-1}_{y_k^{(n)}}\,w^{(n)}\left(2^{j_k^{(n)}}\exp^{-1}_{y_k^{(n)}}(\cdot)\right)\right\rangle.\end{aligned}\tag{40}$$

The first line of (40) can be evaluated taking into account that $u_k\rightharpoonup u$, $\mathcal{S}_k\rightharpoonup 0$, that the definition of profiles $w^{(n)}$ given by (6) and that $r=\frac{N-2}{2}$:

$$\begin{aligned}&\|u_k\|^2+\|u\|^2-2\langle u_k,u\rangle+2\langle u-u_k,\mathcal{S}_k\rangle\\&\quad=\|u_k^2\|+\|u^2\|-2\|u\|^2+o(1)-2\sum_n\left\langle u_k,2^{j_k^{(n)}r}\,\chi\circ\exp^{-1}_{y_k^{(n)}}\,w^{(n)}\left(2^{j_k^{(n)}}\exp^{-1}_{y_k^{(n)}}(\cdot)\right)\right\rangle\\&\quad=\|u_k\|^2-\|u\|^2+o(1)\\&\qquad-2\sum_n 2^{j_k^{(n)}r}\int_{|\xi|<\rho}\sum_{i,j=1}^N g_{ij}^{y_k^{(n)}}\,\partial_i\left(u_k(\exp_{y_k^{(n)}}(\xi))\right)\partial_j\left(\chi(\xi)\,w^{(n)}(2^{j_k^{(n)}}\xi)\right)\sqrt{\det g_{i,j}^{y_k^{(n)}}(\xi)}\mathrm{d}\xi\\&\qquad-2\sum_n 2^{j_k^{(n)}r}\int_{|\xi|<\rho}u_k(\exp_{y_k^{(n)}}(\xi))\chi(\xi)\,w^{(n)}(2^{j_k^{(n)}}\xi)\sqrt{\det g_{i,j}^{y_k^{(n)}}(\xi)}\mathrm{d}\xi\\&\quad=\|u_k\|^2-\|u\|^2+o(1)\\&\qquad-2\sum_n\int_{|\eta|<\rho 2^{j_k^{(n)}}}\sum_{i,j=1}^N g_{ij}^{y_k^{(n)}}\,\partial_i\left(2^{-j_k^{(n)}r}u_k\circ\exp_{y_k^{(n)}}(2^{-j_k^{(n)}}\eta)\right)\partial_j\left(\chi(2^{-j_k^{(n)}}\eta)\,w^{(n)}(\eta)\right)\\&\qquad\qquad\cdot\sqrt{\det g_{i,j}^{y_k^{(n)}}(2^{-j_k^{(n)}}\eta)}\,\mathrm{d}\eta\\&\qquad-2\sum_n 2^{-2j_k^{(n)}}\int_{|\eta|<\rho 2^{j_k^{(n)}}}2^{-j_k^{(n)}r}u_k\circ\exp_{y_k^{(n)}}(2^{-j_k^{(n)}}\eta)\chi(2^{-j_k^{(n)}}\eta)w^{(n)}(\eta)\sqrt{\det g_{i,j}^{y_k^{(n)}}(2^{-j_k^{(n)}}\eta)}\mathrm{d}\eta\\&\quad=\|u_k\|^2-\|u\|^2+o(1)-2\sum_n\int_{\mathbb{R}^N}\sum_i^N|\partial_i w^{(n)}(\eta)|^2\mathrm{d}\eta-2\sum_n 2^{-2j_k^{(n)}}\int_{\mathbb{R}^N}|w^{(n)}(\eta)|^2\mathrm{d}\eta\\&\quad=\|u_k\|^2-\|u\|^2-2\sum_n\|\nabla w^{(n)}\|_2^2+o(1).\end{aligned}$$

(In the third equality we have set $\eta=2^{j_k^{(n)}}\xi$, while in the fourth we have used the fact, due to (6) that $2^{-j_k^{(n)}}\chi(2^{-j_k^{(n)}}\cdot)(u_k\circ\exp_{y_k^{(n)}})(2^{-j_k^{(n)}}\cdot)\rightharpoonup\chi(0)w^{(n)}=w^{(n)}$ as $k\to\infty$ (in our slightly modified sense of weak convergence). Note also we have still denoted by ∂_i (resp. ∂_j) the derivative with respect to the i^{th} (resp j^{th}) component of $\eta=2^{j_k^{(n)}}\xi$. Finally in the last equality we have used (1)).

In order to estimate the second line of (40) we shall split (according to (1)) the $H^{1,2}(M)$-norm into the L^2-norm of the gradient (gradient part) and the L^2-norm of the function (L^2 part) and consider first the latter. Since

$$\begin{aligned}
&\sum_n \left\| 2^{j_k^{(n)}\frac{N-2}{2}} \chi \circ \exp^{-1}_{y_k^{(n)}} w^{(n)}(2^{j_k^{(n)}} \exp^{-1}_{y_k^{(n)}}(\cdot))\right\|_2^2 \\
&= \sum_n 2^{j_k^{(n)}(N-2)} \int_{\mathcal{B}_\rho(y_n)} |\chi \circ \exp^{-1}_{y_k^{(n)}}(x)\, w^{(n)}(2^{j_k^{(n)}} \exp^{-1}_{y_k^{(n)}}(x))|^2 \mathrm{d}v_g \\
&= \sum_n 2^{j_k^{(n)}(N-2)} \int_{|\xi|<\rho} |\chi(\xi)(w^{(n)}(2^{j_k^{(n)}}\xi)|^2 \sqrt{\det g_{i,j}^{y_k^{(n)}}(\xi)}\mathrm{d}\xi \\
&= \sum_n 2^{-2j_k^{(n)}} \int_{|\eta|<\rho 2^{j_k^{(n)}}} |\chi(2^{-j_k^{(n)}}\eta) w^{(n)}(\eta)|^2 \sqrt{\det g_{i,j}^{y_k^{(n)}}(2^{-j_k^{(n)}}\eta)}\mathrm{d}\eta \to 0 \text{ as } k\to\infty,
\end{aligned}$$

(since $j_k^{(n)}\to\infty$) as $k\to\infty$, the second line of (40) is evaluated in the limit by the sum of the gradient terms as follows:

$$\begin{aligned}
&\sum_n 2^{j_k^{(n)}(N-2)} \int_{\mathcal{B}_\rho(y_k^{(n)})} \left|\mathrm{d}\left(\chi \circ \exp^{-1}_{y_k^{(n)}}(x)\, w^{(n)}(2^{j_k^{(n)}} \exp^{-1}_{y_k^{(n)}}(x))\right)\right|^2 \mathrm{d}v_g \\
&= \sum_n 2^{j_k^{(n)}(N-2)} \int_{|\xi|<\rho} \sum_{i,j=1}^N g_{ij}^{y_k^{(n)}}(\xi)\partial_i\left(\chi(\xi)\, w^{(n)}(2^{j_k^{(n)}}\xi)\right)\partial_j\left(\chi(\xi)\, w^{(n)}(2^{j_k^{(n)}}\xi)\right)\sqrt{\det g_{i,j}^{y_k^{(n)}}(\xi)}\mathrm{d}\xi \\
&= \sum_n \int_{|\eta|<\rho 2^{j_k^{(n)}}} \sum_{i,j=1}^N g_{ij}^{y_k^{(n)}}\partial_i\left(\chi(2^{-j_k^{(n)}}\eta)\, w^{(n)}(\eta)\right)\partial_j\left(\chi(2^{-j_k^{(n)}}\eta)\, w^{(n)}(\eta)\right)\sqrt{\det g_{i,j}^{y_k^{(n)}}(2^{-j_k^{(n)}}\eta)}\ \mathrm{d}\eta \\
&\longrightarrow \sum_n \int_{\mathbb{R}^N} |\nabla w^{(n)}(\eta)|^2\ \mathrm{d}\eta = \sum_n \left\|\nabla w^{(n)}\right\|^2 \text{ as } k\to\infty.
\end{aligned}$$

Consider now the terms in the sum in third line of (40). Note that the L^2-part of the scalar product converges to zero by Cauchy inequality and by the calculations for the first line of (40). At the light of the orthogonality condition (5) we have to face two cases.

Case 1: The sequence $(j_k^{(n)} - j_k^{(m)})_{k\in\mathbb{N}}$ is unbounded. Assume without loss of generality that $j_k^{(n)} - j_k^{(m)} \to +\infty$ as $k\to\infty$. Then, using changes of variables $\xi = \exp^{-1}_{y_k^{(n)}}(x)$ and $\eta = 2^{j_k^{(n)}}\xi$,

$$\begin{aligned}
&\left\langle 2^{j_k^{(m)}r}\chi \circ \exp^{-1}_{y_k^{(m)}}(x)\, w^{(m)}\left(2^{j_k^{(m)}} \exp^{-1}_{y_k^{(m)}}(\cdot)\right), 2^{j_k^{(n)}r}\chi \circ \exp^{-1}_{y_k^{(n)}}(x)\, w^{(n)}\left(2^{j_k^{(n)}} \exp^{-1}_{y_k^{(n)}}(\cdot)\right)\right\rangle \\
&= 2^{j_k^{(n)}r}2^{j_k^{(m)}r} \int_{\mathcal{B}_\rho(y_k^{(m)})\cap\mathcal{B}_\rho(y_k^{(n)})} \mathrm{d}\left(\chi \circ \exp^{-1}_{y_k^{(m)}}(x)\, w^{(m)}\left(2^{j_k^{(m)}} \exp^{-1}_{y_k^{(m)}}(x)\right)\right) \\
&\qquad\qquad \cdot \mathrm{d}\left(\chi \circ \exp^{-1}_{y_k^{(n)}}(x)\, w^{(n)}\left(2^{j_k^{(n)}} \exp^{-1}_{y_k^{(n)}}(x)\right)\right)\mathrm{d}v_g + o(1)
\end{aligned}$$

$$= 2^{j_k^{(n)} r} 2^{j_k^{(m)} r} \int_{|\xi|<\rho} \sum_{i,j=1}^{N} g_{ij}^{y_k^{(n)}}(\xi) \partial_i \left(\chi(\xi)\, w^{(n)}(2^{j_k^{(n)}} \xi) \right)$$

$$\cdot \partial_j \left(\chi(\exp^{-1}_{y_k^{(n)}}(\exp_{y_k^{(n)}}(\xi)))\, w^{(m)}(2^{j_k^{(m)}} \exp^{-1}_{y_k^{(m)}}(\exp_{y_k^{(n)}}(\xi))) \right) \sqrt{\det g_{i,j}^{y_k^{(n)}}(\xi)} \mathrm{d}\xi$$

$$= 2^{-j_k^{(n)} r} 2^{j_k^{(m)} r} \int_{|\eta|<\rho 2^{j_k^{(n)}}} \sum_{i,j=1}^{N} g_{ij}^{y_k^{(n)}}(2^{-j_k^{(n)}} \eta) \partial_i \left((1+o(1))\, w^{(n)}(\eta) \right)$$

$$\cdot \partial_j \left((1+o(1))\, w^{(m)}(2^{j_k^{(m)}} \exp^{-1}_{y_k^{(m)}}(\exp_{y_k^{(n)}}(2^{-j_k^{(n)}} \eta))) \right)$$

since, by (6),

$$\underset{k\to\infty}{\text{w-lim}}\, 2^{-j_k^{(n)} r} 2^{j_k^{(m)} r} w^{(m)}(2^{j_k^{(m)}} (\exp^{-1}_{y_k^{(m)}} \circ \exp_{y_k^{(n)}})(2^{-j_k^{(n)}} \cdot)) = \underset{k\to\infty}{\text{w-lim}}\, 2^{-j_k^{(n)} r} u_k(\cdot) = 0.$$

Case 2: $2^{j_k^{(n)}} d(y_k^{(n)}, y_k^{(m)}) \to \infty$ as $k \to \infty$. Since case 1 has been ruled out, we can assume without restrictions that the sequence $j_k^{(m)} - j_k^{(n)} = j \in \mathbb{R}$ for all large k. Then, by arguing as above (and in particular by taking into account that the L^2-part of the scalar product is negligible), we get that, as $k \to \infty$,

$$\left\langle 2^{j_k^{(m)} r}\, \chi \circ \exp^{-1}_{y_k^{(m)}}\, w^{(m)}(2^{j_k^{(m)}} \exp^{-1}_{y_k^{(m)}}(\cdot)), 2^{j_k^{(n)} r}\, \chi \circ \exp^{-1}_{y_k^{(n)}}\, w^{(n)}(2^{j_k^{(n)}} \exp^{-1}_{y_k^{(n)}}(\cdot)) \right\rangle \to 0,$$

since the values of $w^{(m)}$ and of $w^{(n)}$ are set to concentrate at sufficiently separated points, indeed $d(2^{j_k^{(n)}} y_k^{(n)}, 2^{j_k^{(m)}} y_k^{(m)}) = 2^{j_k^{(n)}} d(y_k^{(n)}, 2^{j} y_k^{(m)}) \geq 2^{j_k^{(n)}} d(y_k^{(n)}, y_k^{(m)}) \to \infty$.

Then, by applying the estimates obtained for the three lines of inequality (40) we finally deduce (9) concluding the proof of Theorem 1.

Acknowledgements

The first author is supported by GNAMPA of the "Istituto Nazionale di Alta Matematica (INdAM)" and by MIUR - FFABR - 2017 research grant. http://dx.doi.org/10.13039/501100003407

The second author had no academic affiliation when working on this paper.

References

[1] Adimurthi, A., and Tintarev, C. On compactness in the trudinger-moser inequality. *Ann. Sc. Norm. Sup. Pisa Cl. Sci. 5* (2014), 1–18.

[2] Adimurthi, A., and Tintarev, C. Defect of compactness in spaces of bounded variation. *J. Func. Anal.*, 271 (2016), 37–48.

[3] Devillanova, G. Multiscale weak compactness in metric spaces. *J. Elliptic Parabol. Eq. 2*, 131 (2016). doi:10.1007/BF03377397.

[4] Devillanova, G., Solimini, S., and Tintarev, C. On weak convergence in metric spaces. *Nonl. Anal. Opt., Contemp. Math. 659* (2016), 43–63.

[5] Devillanova, G., Solimini, S., and Tintarev, C. Profile decomposition in metric spaces. *Pure Appl. Funct. Anal. 2*, 4 (2017), 599–628.

[6] Gtextbackslash'erard, P. Description de compacittextbackslash'e de l'injection de sobolev. *ESAIM Control Optim. Calc. Var.*, 3 (1988), 213–233.

[7] Jaffard, S. Analysis of the lack of compactness in the critical sobolev embeddings. *J. Funct. Analysis*, 161 (1999), 384–396.

[8] Lieb, E. On the lowest eigenvalue of the laplacian for the intersection of two domains. *Invent. Math.*, 74 (1983), 441–448.

[9] Schindler, I., and Tintarev, K. An abstract version of the concentration compactness principle. *Revista Mat. Complutense*, 15 (2002), 1–20.

[10] Solimini, S. A note on compactness-type properties with respect to lorentz norms of bounded subsets of a sobolev space. *Ann. Inst. H. Poincartextbackslash'e Anal. Non Lintextbackslash'eaire*, 12 (1995), 319–337.

[11] Solimini, S., and Tintarev, C. Analysis of concentration in the banach space. *Comm. Contemp. Math.*, 18 (2016). doi:10.1142/S0219199715500388.

[12] Struwe, M. A global compactness result for elliptic boundary value problems involving limiting nonlinearities. *Math. Z.*, 187 (1984), 511–517.

G. Devillanova
Politecnico di Bari,
via Re David, 70125 Bari, Italy
giuseppe.devillanova@poliba.it

C. Tintarev
Sankt Olofsgatan 66B, 75330 Uppsala, Sweden
tammouz@gmail.com

Monografías Matemáticas *García de Galdeano* **42**, 79–93 (2019)

MATHEMATICAL ASPECTS OF COMPUTERIZED TOMOGRAPHY: COMPRESSION AND COMPRESSED COMPUTING

Benedikt Diederichs, Tomas Sauer and A. Michael Stock

Abstract. Modern industrial tomography can produce such huge amounts of data that they cannot be handled any more by normal computers. To overcome this problem, the data can be represented and even further compressed by means of a sparse representation with thresholding, as, for example, a three dimensional tensor product wavelet representation. This approach, on the other hand, requires that all operations are realized in the sparse basis. After introducing the basic concepts behind this approach, we show one explicit example, namely how to compute the correlation of two objects by means of sparse representations.

Keywords: Tomography, wavelets, compression.

AMS classification: AMS classification codes.

§1. Introduction

While computerized tomography (CT) is a standard method in medical diagnosis, it is less widely known that tomography is also applied quite frequently in industrial applications. These applications comprise metrology and reverse engineering, documentation and digitalization, for example of cultural heritage, as well as nondestructive testing in manufacturing processes. In contrast to medical applications, these scans are not restricted in size, materials and nature of the objects. A so-called XXL–CT can scan even a full size car with the help of a particle accelerator, a micro or nano CT may scan a "normal sized" object with an extremely high resolution and an inline CT may scan one object per second.

What all these applications have in common is the fact that they produce a huge amount of data: large objects and high-resolution scans can easily reach one Terabyte and more of voxel data after reconstruction, and even if the individual inline scans are usually of moderate size, they come in a large number, typically hundreds of scans every day. Clearly, these circumstances provide new challenges for image processing. The large variety of objects and tasks that occur in industrial CT require advanced and extremely flexible algorithms for segmentation, object separation and information extraction. In medical applications a lot of a priori knowledge can be applied: for most organs, for example, location, size and shape are roughly known and can often been modeled quite efficiently by combining geometric primitives like ellipsoids, cylinders and cones. In industrial CT, the effort of these methods is usually too large for complex separation tasks and methods from machine learning have to be applied so that the system automatically extracts the relevant aspects of different parts.

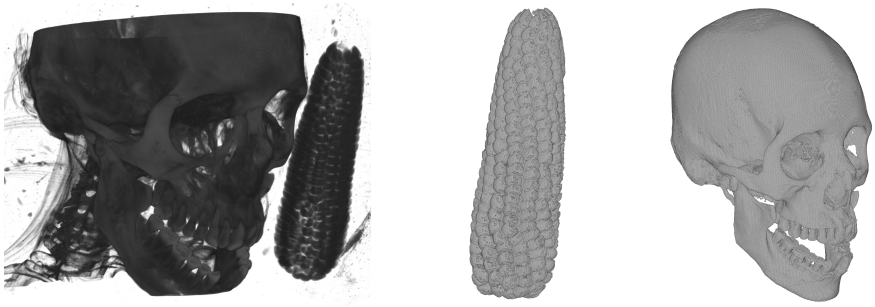


Figure 1: Cultural heritage CT bigture (Peruvian *mummy*, about 6th century AD, courtesy of *Lindenmuseum, Stuttgart*): *(Left)* region of interest containing skull and corncob, *(middle)* segmented corncob, *(right)* segmented skull.

An example for the complexity of such segmentation tasks in cultural heritage can be seen in Fig. 1.

But the main obstacle, of course, is the sheer size of data which leads us to an important concept.

Definition 1. An image is called a *bigture* (with respect to a certain representation) if it cannot be handled in the main memory of a computer any more.

The image on the left hand side of Fig. 1 is an example of a bigture: the size of the original image is about 170GB.

Whether an image is a bigture or not depends on two aspects: the size of the computer memory and the representation of the image. The trivial solution of the *bigture problem* would be to increase the computer memory and to rely on out of core memory techniques; while this is possible to a certain extent, it is not really practical since even if we take the very optimistic and somewhat unrealistic point of view that loading and access times only scale linearly with the amount of data, there is a significant slowdown, especially if we take into account that the amount of data scales *cubically* with the resolution: if we double the resolution of the image, the data increases by a factor of eight.

The more promising approach is to another *representation* of the image which is *sparse*. This means that we represent the same image, i.e., the same *information*, by a significantly smaller amount of *data*. Fortunately, such bases are known in many instances and as a general purpose tool, *wavelets* are still one of the best bases for sparse representation of discrete data on a rectangular grid, especially when this data is locally constant. This is one of the reasons why they were integrated into the *JPEG2000* standard. The drawback of sparse representations, however, is that now all operations have to be implemented in terms of the sparse basis as reconstruction of the full image would result in a bigture and thus render the algorithm useless. This paper will deal with some special case, namely computing the *correlation* of two bigtures using only their sparse representation. Correlations are important to register images and thus a fundamental operation in image processing.

The paper is organized as follows: we first give a short recapitulation of some of the basic concepts of computerized tomography, then recall the basics of the discrete wavelet

transformations, based on which we derive a method to compute the correlations depending on different shifts.

§2. Tomography basics

Tomography reconstructs objects from lower dimensional projections. Here, we focus on X-ray tomography which is based on the *attenuation* of X-rays by different materials. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the function that describes the spatial distribution of this attenuation, where the standard cases are $d = 2$ and $d = 3$. Moreover, let L denote the *straight line* that connects the radiation source and the detector pixel. Then the energy I_D arriving at the detector pixel has the form

$$\frac{I_D}{I_S} = \int_L f(x)\, dx. \tag{1}$$

Note that (1) is only a first order model of the physical process that does not take into account effects like scattering and beam hardening. On the other hand, provided that the intersection of the support of f with L is contained in the line segment between source and detector, X-ray attenuation measures, in $\mathbb{R}^2$, the value of the *Radon transform*

$$L \mapsto Rf(L) := \int_L f(x)\, dx, \qquad L = L(v, s) = \left\{x \in \mathbb{R}^d : v^T x = s\right\} \in \mathcal{L}, \tag{2}$$

where $\mathcal{L}$ denotes the set of all lines in $\mathbb{R}^2$. In $\mathbb{R}^3$ the situation is more intricate as the Radon transform is then defined for planes and in the general s-dimensional case for hyperplanes.

In 2D slices the classical reconstruction is based on the *filtered backprojection* formula

$$(R^* g) * f = R^*(g * Rf), \qquad f \in L_1(\mathbb{R}^2), \quad g \in \mathcal{S}(\mathcal{L}), \tag{3}$$

where $\mathcal{S}(\mathcal{L})$ is the Schwartz space and

$$R^* g(x) = \int_{\|v\|=1} g(v, v^T x) dv, \qquad x \in \mathbb{R}^2, \quad g \in \mathcal{S}(\mathcal{L}) \subset \mathcal{S}(\mathbb{S}^1 \times \mathbb{R}),$$

stands for the *dual Radon transform*, where we cover $\mathcal{L}$ by $\mathbb{S}^1 \times \mathbb{R}$. For the application of (3), one chooses g such that $R^* g$ is close to the Dirac delta functional, resulting in $(R^* g) * f \approx f$. Typically, g is a radial function, i.e., $g(v, v^T x) = g(|v^T x|)$ and constructed such that convolving with g acts as a low-pass filter, hence the name filtered back projection. This formula is then discretized.

Similar inversion formulas exist in the three dimensional case. However, while in the two dimensional case typically a rather dense sampling of all lines passing through the support of f is available, the scanning geometries used in practice for the three dimensional case are more limited. Specific approximate inversion formulas, tailored to different geometries, are available. For example in the important case of the cone-beam geometry, the classical *Feldkamp algorithm* is widely used. For details on analytical methods see [10, 11].

A different approach is to see (1) as a system of integral equations and to discretize those directly. This could be done in a function as in Galerkin methods, but the standard technique

is to discretize the region of interest Ω into a *voxel* grid

$$\Omega = \bigcup_{j=1}^{N} V_j, \qquad V_j = [a_{j1}, b_{j1}] \times \cdots \times [a_{jd}, b_{jd}],$$

to assume the function f to have the constant value f_j on V_j and to rewrite (1) as

$$\frac{I_D}{I_S} =: y_L = \sum_{j \in J_L} \lambda_{L,j} f_j, \qquad J_L := \left\{ 1 \le j \le n : V_j \cap L \neq \emptyset \right\}. \tag{4}$$

The values λ_{Lj} is normally chosen as the length of the intersection $V_j \cap L$. Using a finite set of measurement rays, $L_1, \ldots, L_M$, which describe the scanning geometry, we end up with the linear system

$$y = \Lambda f, \qquad \Lambda = \left[\lambda_{L_j,k} : \begin{array}{c} j = 1, \ldots, M \\ k = 1, \ldots, N \end{array} \right], \qquad y = \begin{bmatrix} y_{L_1} \\ \vdots \\ y_{L_M} \end{bmatrix} \in \mathbb{R}^M, \; f = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} \in \mathbb{R}^N. \tag{5}$$

This is the simple concept of the *algebraic reconstruction technique* (ART), but solving the linear system (5) is far from trivial since it is usually huge. Nevertheless ART has some advantages:

1. The method works in any dimensionality and with any measurement geometry. Whether cone beam or parallel beam is used, only results in a different geometry matrix Λ.
2. The method works independently of dimension and the matrix is modestly sparse: if the voxels are arranged on an $n \times \cdots \times n$ grid, i.e., $N = n^d$, then any equation still involves only $O(n)$ variables.
3. A priori knowledge like obstructions or side conditions like positivity can be easily integrated into the approach as well as regularizers.

For more details see, for example [2, 6] and, of course, [10].

§3. Wavelet basics and definitions

Next, we briefly fix the notation for a wavelet *multiresolution analysis* (MRA). The starting point is a *refinable function* ϕ, i.e., a solution of the *refinement equation*

$$\phi = \sum_{k \in \mathbb{Z}} a_k \, \phi(2 \cdot - k), \qquad a \in \ell_0(\mathbb{Z}), \tag{6}$$

where $\ell_0(\mathbb{Z})$ stands for the space of all bi–infinite sequences with finite support:

$$\ell_0(\mathbb{Z}) = \{ c = (c_k : k \in \mathbb{Z}) : \|c\|_0 < \infty \}, \qquad \|c\|_0 := \#\{k : c_k \neq 0\}.$$

In addition, ϕ is called an *orthonormal scaling function* if its integer translates are mutually orthonormal, that is,

$$\delta_{k0} = \langle \phi, \phi(\cdot - k) \rangle = \int_{\mathbb{R}} \phi(x) \, \phi(x - k) \, dx, \qquad k \in \mathbb{Z}.$$

Substituting the refinement equation (6) into this requirement, it is easily seen that the sequence a, called *mask* in the subdivision literature [1], has to satisfy the *quadrature mirror filter equation*

$$\delta_{k0} = \frac{1}{2} \sum_{j \in \mathbb{Z}} a_{k-2j} a_j, \qquad k \in \mathbb{Z}. \tag{7}$$

A generic construction for finitely supported masks that satisfy (7) and give rise to L_2–solutions of (6) has first been given by Daubechies [4], see also [5]. This was the starting point for a multitude of different wavelet constructions, orthogonal as well as biorthogonal ones, and eventually the inclusion of wavelets into the JPEG2000 standard.

Based on an orthogonal scaling function, we define the MRA as the sequence of spaces

$$V_j := \operatorname{span} \left\{ \phi(2^j \cdot -k) : k \in \mathbb{Z} \right\}, \qquad j \in \mathbb{N}_0;$$

by (6), these spaces are *nested* in the sense that $V_0 \subset V_1 \subset \cdots$ and the associated *wavelet*, defined as

$$\psi = \sum_{k \in \mathbb{Z}} b_k \, \phi(2 \cdot -k), \qquad b_k := (-1)^k a_{1-k}, \quad k \in \mathbb{Z}, \tag{8}$$

belongs to V_1 and satisfies $\langle \phi, \psi(\cdot - k) \rangle = 0$, $k \in \mathbb{Z}$, so that

$$V_{j+1} = V_j \oplus W_j, \qquad W_j := \operatorname{span} \left\{ \psi \left(2^j \cdot -k \right) : k \in \mathbb{Z} \right\}, \qquad j \in \mathbb{N}_0. \tag{9}$$

Moreover, the integer shifts of the wavelet ϕ form an orthonormal basis of W_0 and, accordingly, the functions $\psi_{j,k} := 2^{j/2} \psi(2^j \cdot -k)$, $k \in \mathbb{Z}$, are an orthonormal basis of W_j. Hence, any function $f \in V_n$ can be written as

$$f = \sum_{k \in \mathbb{Z}} c_k^n(f) \, \phi \left(2^n \cdot -k \right) = \sum_{k \in \mathbb{Z}} c_k(f) \, \phi(\cdot - k) + \sum_{j=0}^{n-1} \sum_{k \in \mathbb{Z}} d_k^j(f) \, 2^{j/2} \psi \left(2^j \cdot -k \right), \tag{10}$$

where

$$d_k^j(f) = 2^{j/2} \int_{\mathbb{R}} f(x) \, \psi(2^j x - k) \, dx.$$

The main point in favor of wavelets, however, is that the conversion from c^n to $c, d^0, \ldots, d^{n-1}$, i.e., the transmission between the two representations of f in (10) can be performed very efficiently by means of discrete *filterbank* operations; this is Mallat's *discrete wavelet transform* [8, 9], see also [13].

Wavelets are naturally related to *subdivision schemes*. This is an immediate consequence of the refinement equation (6) which yields that for any $f \in V_0$ we have

$$f = \sum_{j \in \mathbb{Z}} c_j(f) \, \phi(\cdot - j) = \sum_{j,k \in \mathbb{Z}} c_j(f) \, a_k \, \phi(2 \cdot -2j - k) = \sum_{k \in \mathbb{Z}} \left(\sum_{j \in \mathbb{Z}} a_{k-2j} \, c_j(f) \right) \phi(2 \cdot -k),$$

or, in terms of (semidiscrete) convolutions,

$$f = c(f) * \phi = (S_a c(f)) * \phi(2 \cdot), \qquad (S_a c)_j := \sum_{k \in \mathbb{Z}} a_{j-2k} \, c_k, \tag{11}$$

and the subdivision operator S_a. We will use this connection later.

In s variables, usually $s = 2, 3$, we use the respective tensor product scaling functions

$$\phi(x) = \prod_{r=1}^{s} \phi(x_r),$$

so that, for $\alpha \in \mathbb{Z}^s$,

$$\int_{\mathbb{R}^s} \phi(x)\,\phi(x-\alpha)\,dx = \prod_{r=1}^{s} \int_{\mathbb{R}} \phi(x_r)\,\phi(x_r-\alpha_r)\,dx_r = \prod_{r=1}^{s} \delta_{\alpha_r,0} = \delta_{\alpha,0}.$$

To build wavelets, we set $\psi_0 = \phi$, $\psi_1 = \psi$ and define the $2^s - 1$ wavelet functions

$$\psi_\eta(x) := \prod_{r=1}^{s} \psi_{\eta_r}(x_r), \qquad \eta \in H := \{0,1\}^s \setminus \{0\}$$

and the refinement equation is given by

$$\psi_\eta = \sum_{\alpha\in\mathbb{Z}^s} b_{\eta,\alpha}\,\phi(2\cdot-\alpha), \qquad \eta \in H,$$

with

$$b_{\eta,\alpha} := \prod_{r=1}^{s} \left((1-\eta_r)a_{\alpha_r} + \eta_r b_{\alpha_r}\right), \qquad \eta \in H, \quad \alpha \in \mathbb{Z}^s.$$

These function satisfy the orthonormality condition

$$\int_{\mathbb{R}^s} \psi_\eta(x)\,\psi_{\eta'}(x-\alpha)\,dx = \delta_{\eta,\eta'}\,\delta_{\alpha,0}, \qquad \eta,\eta' \in \{0,1\}, \quad \alpha,\beta \in \mathbb{Z}^s.$$

Thus, with $V_j = \operatorname{span}\left\{\phi(2^j\cdot-\alpha) : \alpha \in \mathbb{Z}^s\right\}$, we again have the multiresolution analysis

$$V_{j+1} = V_j \oplus W_j, \qquad W_j := \operatorname{span}\left\{2^{js/2}\,\psi_\eta(2^j\cdot-\alpha) : \eta \in H,\ \alpha \in \mathbb{Z}^s\right\},$$

and the wavelet decomposition

$$f = \sum_{\alpha\in\mathbb{Z}^s} c^n_\alpha(f)\,\phi(2^n\cdot-\alpha) = \sum_{\alpha\in\mathbb{Z}^s} c_\alpha(f)\,\phi(\cdot-\alpha) + \sum_{j=0}^{n-1}\sum_{\eta\in H}\sum_{\alpha\in\mathbb{Z}^s} d^j_{\eta,\alpha}(f)\,2^{js/2}\psi_\eta\left(2^j\cdot-\alpha\right). \quad (12)$$

Remark 1. There exist orthogonal wavelet decompositions for arbitrary scaling matrices and even a generic tensor-product like approach to construct them, see [3, 7], but the classical dyadic tensor product construction offers an extremely efficient way of localizing the supports of scaling functions and wavelets which is useful in the implementation of a fast decomposition and reconstruction as described in [12].

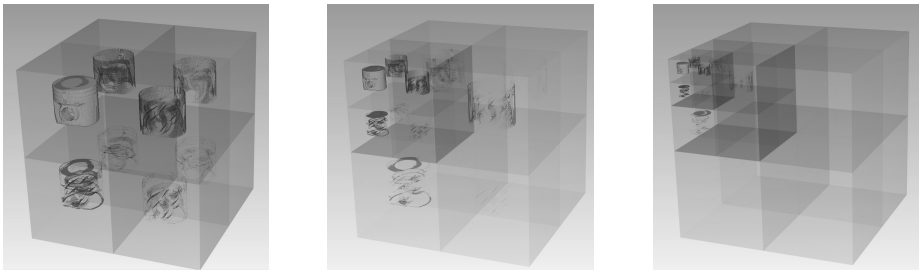


Figure 2: Different levels of wavelet decomposition of a typical industrial piece (engine piston). *(Left)* One level, the different types of edge and face dectections are clearly visible, *(middle)* even after two levels wavelet coefficients become almost invisible, *(right)* and after three iterations, they are mainly irrelevant.

§4. Wavelets and compression

The idea behind wavelet compression is rather simple: if a function $f : \mathbb{R}^s \to \mathbb{R}$ can be approximated well by V_k for some rather small k, then the wavelet coefficients $d^j_{\eta,\alpha}(f)$ for $j > k$ will be very small and can be discarded without significant loss of quality. This works for two reasons, cf. [5, 9]:

1. Any reasonable scaling function ϕ is compactly supported and (re)produces polynomials of a certain degree, but at least satisfies

$$\sum_{\alpha\in\mathbb{Z}^s} \phi(\cdot - \alpha) = 1.$$

   This means that the wavelet coefficients for *locally constant* functions vanish in the regions where the function is constant.

2. The spaces V_k usually have good approximation properties for smooth function which implies that in smooth regions the absolute values $\left|d^j_{\eta,\alpha}(f)\right|$ of the wavelet coefficients decay very fast with respect to j.

Starting from the high resolution $c^n(f)$ of

$$f \approx \sum_{\alpha\in\mathbb{Z}^s} c^n_\alpha(f)\, \phi(2^n \cdot -\alpha),$$

one computes the wavelet coefficients $d^{n-1}_\eta(f), d^{n-2}_\eta(f), \cdots, d^0_\eta(f)$ and the scaling coefficients $c^0(f)$ by means of a fast wavelet transform. Note that the index η also has an intuitive geometric meaning for the wavelet coefficient. Indeed, if $|\eta| = 1$, it detects faces parallel to the coordinate planes, if $|\eta| = 2$ the coefficients correspond to edges parallel to the axes and $\eta = (1, 1, 1)$ detects some "diagonal" feature, see Fig. 2.

While the originally sampled data $c^n(f)$ is usually dense, the wavelet transform is sparse if the underlying image is piecewise constant which is the case in most technical applications,

see again Fig. 2. In addition to deleting zero coefficients, lossy compression is obtained by thresholding the wavelet coefficients and replacing them by

$$\hat{d}^j_{\eta,\alpha} = t_\epsilon\left(d^j_{\eta,\alpha}\right), \qquad \eta \in H,\ \alpha \in \mathbb{Z}^s,\ j = 0,\dots,n-1,$$

by means of threshold function t_ε.

Definition 2. For a threshold level ε the *hard threshold* and the *soft threshold* use the functions

$$t^h_\varepsilon(x) = \begin{cases} 0, & |x| < \varepsilon, \\ x, & |x| \ge \varepsilon, \end{cases} \qquad t^s_\varepsilon(x) = \begin{cases} 0, & |x| < \varepsilon, \\ x - \varepsilon, & x \ge \varepsilon, \\ x + \varepsilon, & x \le -\varepsilon. \end{cases}$$

respectively.

While soft thresholding is known to perform a denoising operation, popular as *wavelet shrinkage*, hard threshold is more contrast and edge preserving. The choice of the threshold level can be made according to several strategies, for example

1. absolute choice of threshold level,
2. *best N–term approximation:* the threshold is chosen in such a way that only the N largest coefficients remain,
3. overall precision: ϵ is chosen such that

$$\sum_{j=0}^{n-1} \sum_{\eta\in H} \sum_{\alpha\in\mathbb{Z}^s} \left(d^j_{\eta,\alpha} - \hat{d}^j_{\eta,\alpha}\right)^2$$

   does not exceed a certain bound. Since these are the coefficients in an orthonormal expansion, this is also the norm of the L_2–error, hence a certain PSNR can be prescribed for the compression.

Recall also that after transformation and thresholding, the array of coefficients is encoded in an efficient way using a more or less standard entropy encoder, cf. [12] for details. We will not dwell on these issues here though they are of course important for the overall compression rates.

Definition 3. For a function f with a thresholded wavelet decomposition we define

$$N(f) := \#\{\alpha : \hat{c}_\alpha \ne 0\} + \sum_{j=0}^{n-1} \sum_{\eta\in H} \#\left\{\alpha : \hat{d}^j_{\eta,\alpha} \ne 0\right\}$$

as the number of nonzero coefficients in the representation.

§5. Wavelet correlation

The *correlation* between two functions f, g is defined as

$$f \star g(y) := \int_{\mathbb{R}^s} f(x)\, g(x+y)\, dx, \qquad y \in \mathbb{R}^s,$$

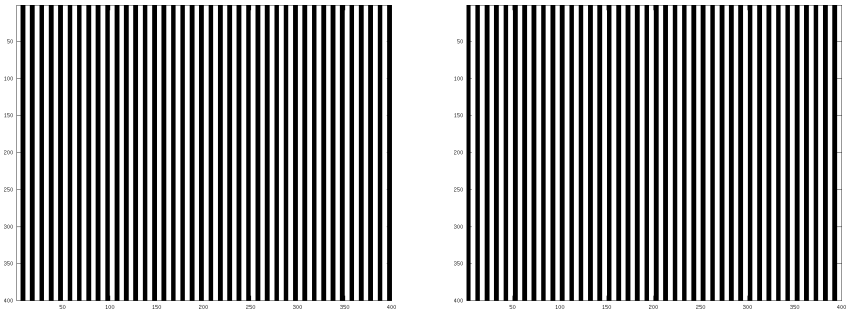


Figure 3: Two images with a PSNR value of 0, i.e., of maximal disparity, while $\sigma(f, g) = 1$.

and measures the best fit between f and a shifted version of g. It is useful for image comparison by using

$$\sigma(f, g) := \frac{1}{\|f\|_2 \, \|g\|_2} \max_{y \in \mathbb{R}^s} |f \star g(y)|$$

as a translation invariant similarity measure for images. An extremal example in this respect can be seen in Fig. 3. Moreover, local correlations are needed in order to stitch images together by finding a proper offset of one of the images such that the overlapping areas coincide as much as possible.

Taking into account that in many cases a complete reconstruction of a bigture is impossible due to memory limitations, we need an algorithm that computes the correlation entirely from the wavelet decomposition. We will develop such a method in this section. To that end, we set up some terminology first.

Definition 4. The *correlation* of two sequences $c, d \in \mathbb{Z}^s$ is defined as

$$(c \star d)_\alpha := \sum_{\beta \in \mathbb{Z}^s} c_{\alpha+\beta} \, d_\beta, \qquad \alpha \in \mathbb{Z}^s,$$

and the *translation operator* as

$$\left(\tau_\gamma c\right)_\alpha := c_{\alpha+\gamma}, \qquad \alpha \in \mathbb{Z}^s.$$

We start with $f, g \in V_n$, written as

$$f = \sum_{\alpha \in \mathbb{Z}^s} c_\alpha(f) \, \phi(\cdot - \alpha) + \sum_{j=0}^{n-1} 2^{js/2} \sum_{\eta \in H} \sum_{\alpha \in \mathbb{Z}^s} d^j_{\eta,\alpha}(f) \, \psi_\eta(2^j \cdot - \alpha),$$

$$g = \sum_{\alpha \in \mathbb{Z}^s} c_\alpha(g) \, \phi(\cdot - \alpha) + \sum_{j=0}^{n-1} 2^{js/2} \sum_{\eta \in H} \sum_{\alpha \in \mathbb{Z}^s} d^j_{\eta,\alpha}(g) \, \psi_\eta(2^j \cdot - \alpha),$$

and first observe that the integer correlations can be easily computed as shifted correlations of the discrete sequences.

Lemma 1. *For $\gamma \in \mathbb{Z}^s$ we have that*

$$f \star g(\gamma) = \tau_{-\gamma}\left(c(f) \star c(g)\right) + \sum_{j=0}^{n-1} \sum_{\eta \in H} \tau_{-2^j\gamma}\left(d_\eta^j(f) \star d_\eta^j(g)\right). \tag{13}$$

The number of arithmetic operations is bounded by $\min\{N(f), N(g)\}$ *and hence subadditive in the number of nonzero coefficients in the expansion of f or g, respectively.*

Proof. We compute

$$\begin{aligned}
&f \star g(\gamma)\\
&= \sum_{\alpha,\beta \in \mathbb{Z}^s} c_\alpha(f)\, c_\beta(g)\, \langle \phi(\cdot - \alpha), \phi(\cdot - \beta + \gamma)\rangle\\
&\quad + \sum_{j=0}^{n-1} 2^{-js/2} \sum_{\eta \in H} \sum_{\alpha,\beta \in \mathbb{Z}^s} \left(c_\alpha(f)\, d^j_{\eta,\beta}(g) + c_\alpha(g)\, d^j_{\eta,\beta}(f)\right) \left\langle \phi(\cdot - \alpha), \psi_\eta\left(2^j \cdot - \beta + 2^j\gamma\right)\right\rangle\\
&\quad + \sum_{j,k=0}^{n-1} 2^{-(j+k)s/2} \sum_{\eta,\eta' \in H} \sum_{\alpha,\beta \in \mathbb{Z}^s} d^j_{\eta,\alpha}(f)\, d^k_{\eta',\beta}(g) \left\langle \psi_\eta\left(2^j \cdot - \alpha\right), \psi_{\eta'}\left(2^k \cdot - \beta + 2^k\gamma\right)\right\rangle\\
&= \sum_{\alpha,\beta \in \mathbb{Z}^s} c_\alpha(f)\, c_\beta(g)\, \delta_{\alpha,\beta-\gamma} + \sum_{j,k=0}^{n-1} \sum_{\eta,\eta' \in H} \sum_{\alpha,\beta \in \mathbb{Z}^s} d^j_{\eta,\alpha}(f)\, d^k_{\eta,\beta}(g) \delta_{jk} \delta_{\alpha,\beta-2^k\gamma} \delta_{\eta,\eta'}\\
&= \sum_{\beta \in \mathbb{Z}^s} c_{\beta-\gamma}(f)\, c_\beta(g) + \sum_{j=0}^{n-1} \sum_{\eta \in H} \sum_{\beta \in \mathbb{Z}^s} d^j_{\eta,\beta-2^j\gamma}(f)\, d^j_{\eta,\beta}(g),
\end{aligned}$$

which gives (13). Since any contribution to the sum requests a nonzero coefficient of the expansion of f and g, the number of arithmetic operations is bounded by $\min\left(N(f), N(g)\right)$. □

For the general case, we define the bi–infinite matrix valued function

$$\Phi(y) := \left[\langle \phi(\cdot - \alpha), \phi(\cdot - \beta + y)\rangle : \alpha, \beta \in \mathbb{Z}^s\right], \qquad y \in \mathbb{R}^s, \tag{14}$$

which represents the correlation for $f, g \in V_0$ in the sense that $f \star g(y) = c(f)^T \Phi(y)\, c(g)$.

Lemma 2. *If $\phi \in L_2(\mathbb{R}^s)$ is a compactly supported orthonormal scaling function, then*

1. *the matrix $\Phi(y)$ is a banded* Toeplitz matrix*,*
2. *$\Phi(y)$ is almost* 1*-periodic: $\Phi(y + \gamma) = \Phi(y)\tau_\gamma = \tau_{-\gamma}\Phi(y)$, $\gamma \in \mathbb{Z}^s$,*
3. *$y \mapsto \Phi(y)$ is continuous with $\Phi(0) = I$, in the sense that the coefficients form a uniformly equicontinuous family.*

Proof. For 1) we note that

$$\Phi(y)_{\alpha,\beta} = \langle \phi(\cdot - \alpha), \phi(\cdot - \beta + y)\rangle = \langle \phi\left(\cdot - (\alpha - \beta)\right), \phi(\cdot + y)\rangle,$$

hence depends only on $\alpha - \beta$. Bandedness results from the compact support of ϕ: if $\|\alpha - \beta\|$ is large enough, then the supports of $\phi(\cdot - (\alpha - \beta))$ and $\phi(\cdot + y)$ are disjoint and the integral is zero. For 2) we consider

$$(\Phi(y+\gamma)c)_\alpha = \sum_{\beta\in\mathbb{Z}^s} \langle \phi(\cdot-\alpha), \phi(\cdot-(\beta-\gamma)+y)\rangle c_\beta = \sum_{\beta\in\mathbb{Z}^s} \langle \phi(\cdot-\alpha), \phi(\cdot-\beta+y)\rangle c_{\beta+\gamma}$$

and

$$\left(c^T\Phi(y+\gamma)\right)_\beta = \sum_{\alpha\in\mathbb{Z}^s} \langle \phi(\cdot-(\alpha+\gamma)), \phi(\cdot-\beta+y)\rangle c_\alpha = \left(\tau_{-\gamma}c^T\Phi(y)\right)_\beta.$$

To prove 3) we first note that due to 1) the matrix $\Phi(y)$ only contains finitely many nonzero entries of the form

$$\langle \phi(\cdot-\alpha), \phi(\cdot+y)\rangle, \qquad \alpha\in\mathbb{Z}^s.$$

For each $\alpha\in\mathbb{Z}^s$ and $\delta\in\mathbb{R}^s$ we then have that

$$|\langle \phi(\cdot-\alpha), \phi(\cdot+y+\delta)\rangle - \langle \phi(\cdot-\alpha), \phi(\cdot+y)\rangle| \le \|\phi\|_2 \ \|\phi(\cdot+\delta)-\phi\|_2$$

which tends to zero for $\|\delta\|\to 0$ uniformly in α and y, as $\phi\in L_2(\mathbb{R}^s)$. □

Example 1. In the simple case of Haar wavelets where $\phi = \chi = \chi_{[0,1]^s}$, the entries of $\Phi(y)$ can easily be computed explicitly, namely, for $y\in(0,1)^s$ as

$$\Phi(y)_{\alpha,\beta} = \int_{[0,1]^s} \chi(x+\alpha-\beta+y)\,dx = \prod_{r=1}^{s}\int_0^1 \chi(x+\alpha_r-\beta_r+y_r)\,dx.$$

Now, $[0,1]+\alpha_r-\beta_r+y_r\cap[0,1]\neq\emptyset$ only if $\beta_r=\alpha_r+1$ or $\beta_r=\alpha_r$, where

$$\int_0^1 \chi(x+\alpha_r-\beta_r+y_r)\,dx = \int_{-1}^0 \chi(x+y_r) = y_r$$

in the first case and

$$\int_0^1 \chi(x+\alpha_r-\beta_r+y_r)\,dx = \int_0^1 \chi(x+y_r) = 1-y_r$$

in the latter. Therefore, every row of $\Phi(y)$ contains exactly 2^s nonzero values, namely

$$\Phi(y)_{\alpha,\alpha+\eta} = \prod_{r=1}^{s}(\eta_r y_r + (1-\eta_r)(1-y_r)), \qquad \alpha\in\mathbb{Z}^s, \quad \eta\in\{0,1\}^s.$$

For arbitrary $y\in\mathbb{R}^s$, we apply Lemma 2. Note however that there is a shift in the matrices in this case:

$$\lim_{y\to(1,\dots,1)}\Phi(y) = \tau_{(1,\dots,1)} = \lim_{y\to 0}\Phi(y)\tau_{(1,\dots,1)}.$$

Since the finely sampled data $c^n(f)$ corresponds to evaluation of a function on the grid $2^{-n}\mathbb{Z}^s$, even if the sequence is indexed by integers, we have to be able to compute correlations at least for dyadic values $2^{-m}\gamma$, $\gamma\in\mathbb{Z}^s$, $0\le m\le n$. The next result shows that also in this case, correlations can be computed from wavelet coefficients by means of $\Phi(y)$.

Theorem 3. *For $\gamma \in \mathbb{Z}^s$ and $0 \le m \le n$ we have that*

$$\begin{aligned} f \star g\,(2^{-m}\gamma) = c(f)^T \Phi\,(2^{-m}\gamma)\, c(g) &+ 2^{-s} \sum_{j=0}^{m-1} \sum_{\eta,\eta' \in H} \left(S_{b_\eta} d_\eta^j(f)\right)^T \Phi\left(2^{j+1-m}\gamma\right) \left(S_{b_{\eta'}} d_{\eta'}^j(g)\right) \\ &+ \sum_{0 \le j < k \le m-1} 2^{-\left(1+\frac{k-j}{2}\right)s} \sum_{\eta,\eta' \in H} \left(S_a^{k-j} S_{b_\eta} d_\eta^j(f)\right)^T \Phi\,(2^{-m}\gamma) \left(S_{b_{\eta'}} d_{\eta'}^k(g)\right) \\ &+ \sum_{0 \le k < j \le m-1} 2^{-\left(1+\frac{j-k}{2}\right)s} \sum_{\eta,\eta' \in H} \left(S_{b_\eta} d_\eta^j(f)\right)^T \Phi\,(2^{-m}\gamma) \left(S_a^{j-k} S_{b_{\eta'}} d_{\eta'}^k(g)\right) \\ &+ \sum_{j=m}^{n-1} \sum_{\eta \in H} \tau_{-2^{j-m}\gamma} \left(d_\eta^j(f) \star d_\eta^j(g)\right). \end{aligned} \tag{15}$$

Proof. As all levels at least as fine as m are still orthogonal, we obtain

$$\begin{aligned} & f \star g(2^{-m}\gamma) \\ &= \sum_{\alpha,\beta \in \mathbb{Z}^s} c_\alpha(f)\, c_\beta(g) \left\langle \phi(\cdot - \alpha), \phi(\cdot - \beta + 2^{-m}\gamma) \right\rangle \\ &\quad + \sum_{j,k=0}^{m-1} 2^{(j+k)s/2} \sum_{\eta,\eta' \in H} \sum_{\alpha,\beta \in \mathbb{Z}^s} d_{\eta,\alpha}^j(f)\, d_{\eta',\beta}^k(g) \left\langle \psi_\eta\left(2^j \cdot - \alpha\right), \psi_{\eta'}\left(2^k \cdot - \beta + 2^{k-m}\gamma\right) \right\rangle \\ &\quad + \sum_{j=m}^{n-1} \sum_{\eta \in H} \sum_{\beta \in \mathbb{Z}^s} d_{\eta,\beta-2^{j-m}\gamma}^j(f)\, d_{\eta,\beta}^j(g) \\ &= c(f)^T \Phi\,(2^{-m}\gamma)\, c(g) + \sum_{j=m}^{n-1} \sum_{\eta \in H} \sum_{\beta \in \mathbb{Z}^s} d_{\eta,\beta-2^{j-m}\gamma}^j(f)\, d_{\eta,\beta}^j(g) \\ &\quad + \sum_{j,k=0}^{m-1} 2^{-(j+k)s/2} \sum_{\eta,\eta' \in H} \sum_{\alpha,\beta \in \mathbb{Z}^s} d_{\eta,\alpha}^j(f)\, d_{\eta',\beta}^k(g) \left\langle 2^{js} \psi_\eta\left(2^j \cdot - \alpha\right), 2^{ks} \psi_{\eta'}\left(2^k \cdot - \beta + 2^{k-m}\gamma\right) \right\rangle, \end{aligned} \tag{16}$$

which already gives the first and the last term in the right hand side of (15). We are left with calculating the correlations of all levels from zero to $m-1$.

Recalling that

$$\psi_\eta = \sum_{\alpha \in \mathbb{Z}^s} b_{\eta,\alpha}\, \phi(2 \cdot - \alpha), \qquad \eta \in H,$$

we obtain for $j \le n-1$ that

$$\begin{aligned} \sum_{\alpha \in \mathbb{Z}^s} d_{\eta,\alpha}^j(f)\, \psi_\eta\left(2^j \cdot - \alpha\right) &= \sum_{\alpha \in \mathbb{Z}^s} d_{\eta,\alpha}^j(f) \sum_{\beta \in \mathbb{Z}^s} b_{\eta,\beta} \phi\left(2^{j+1} \cdot - 2\alpha - \beta\right) \\ &= \sum_{\beta \in \mathbb{Z}^s} \left(\sum_{\alpha \in \mathbb{Z}^s} b_{\eta,\beta-2\alpha}\, d_\alpha^j(f) \right) \phi\left(2^{j+1} \cdot - \beta\right) = \sum_{\beta \in \mathbb{Z}^s} \left(S_{b_\eta} d_\eta^j(f)\right)_\beta \phi\left(2^{j+1} \cdot - \beta\right), \end{aligned}$$

and the refinement equation (6) for ϕ yields in the same way for $k > j$ that

$$\sum_{\alpha \in \mathbb{Z}^s} d_{\eta,\alpha}^j(f)\, \psi_\eta\left(2^j \cdot - \alpha\right) = \sum_{\beta \in \mathbb{Z}^s} \left(S_a^{k-j} S_{b_\eta} d_\eta^j(f)\right)_\beta \phi\left(2^{k+1} \cdot - \beta\right).$$

With these formulas at hand, we split the sum from (16) into its diagonal

$$\begin{aligned}
&\sum_{j=0}^{m-1} 2^{js} \sum_{\eta,\eta'\in H} \sum_{\alpha,\beta\in\mathbb{Z}^s} d^j_{\eta,\alpha}(f)\, d^j_{\eta',\beta}(g) \Big\langle \psi_\eta\left(2^j\cdot-\alpha\right), \psi_{\eta'}\left(2^j(\cdot+2^{-m}\gamma)-\beta\right)\Big\rangle \\
&= \sum_{j=0}^{m-1} 2^{js} \sum_{\eta,\eta'\in H} \sum_{\alpha,\beta\in\mathbb{Z}^s} \left(S_{b_\eta} d^j_{\eta,}(f)\right)_\alpha \left(S_{b_{\eta'}} d^j_{\eta'}(g)\right)_\beta \left\langle \phi\left(2^{j+1}\cdot-\alpha\right), \phi\left(2^{j+1}(\cdot+2^{-m}\gamma)-\beta\right)\right\rangle \\
&= 2^{-s} \sum_{j=0}^{m-1} \sum_{\eta,\eta'\in H} \left(S_{b_\eta} d^j_\eta(f)\right) \Phi\left(2^{j+1-m}\gamma\right) \left(S_{b_{\eta'}} d^j_{\eta'}(g)\right),
\end{aligned}$$

the lower triangle

$$\begin{aligned}
&\sum_{0\le j<k\le m-1} 2^{(j+k)s/2} \sum_{\eta,\eta'\in H} \sum_{\alpha,\beta\in\mathbb{Z}^s} d^j_{\eta,\alpha}(f)\, d^k_{\eta',\beta}(g) \Big\langle \psi_\eta\left(2^j\cdot-\alpha\right), \psi_{\eta'}\left(2^k(\cdot+2^{-m}\gamma)-\beta\right)\Big\rangle \\
&\qquad = \sum_{0\le j<k\le m-1} 2^{(j+k)s/2} \sum_{\eta,\eta'\in H} \sum_{\alpha,\beta\in\mathbb{Z}^s} \left(S_{b_\eta} d^j_{\eta,}(f)\right)_\alpha \left(S_{b_{\eta'}} d^j_{\eta'}(g)\right)_\beta \\
&\qquad\qquad \cdot \left\langle \phi\left(2^{j+1}\cdot-\alpha\right), \phi\left(2^{k+1}(\cdot+2^{-m}\gamma)-\beta\right)\right\rangle \\
&\qquad = \sum_{0\le j<k\le m-1} 2^{(j+k)s/2} \sum_{\eta,\eta'\in H} \sum_{\alpha,\beta\in\mathbb{Z}^s} \left(S_a^{k-j} S_{b_\eta} d^j_{\eta,}(f)\right)_\alpha \left(S_{b_{\eta'}} d^j_{\eta'}(g)\right)_\beta \\
&\qquad\qquad \cdot \left\langle \phi\left(2^{k+1}\cdot-\alpha\right), \phi\left(2^{k+1}(\cdot+2^{-m}\gamma)-\beta\right)\right\rangle \\
&\qquad = \sum_{0\le j<k\le m-1} 2^{\left(\frac{j-k}{2}-1\right)s} \sum_{\eta,\eta'\in H} \left(S_a^{k-j} S_{b_\eta} d^j_{\eta,}(f)\right)^T \Phi\left(2^{k+1-m}\gamma\right) \left(S_{b_{\eta'}} d^j_{\eta'}(g)\right)
\end{aligned}$$

and the upper triangle

$$\sum_{m-1\ge j>k\ge 0} \cdots = \sum_{m-1\ge j>k\ge 0} 2^{\left(\frac{k-j}{2}-1\right)s} \sum_{\eta,\eta'\in H} \left(S_{b_\eta} d^j_{\eta,}(f)\right)^T \Phi\left(2^{j+1-m}\gamma\right) \left(S_a^{j-k} S_{b_{\eta'}} d^j_{\eta'}(g)\right)$$

that is obtained in the same way. □

Correlation is a standard tool for matching objects whose mutual displacement is due to shifts. This goal can often be achieved by aligning calibrated reference objects which are first used to compensate rotational effects. After that, maximizing the correlation means finding the best alignment between the two objects f and g. For this purpose we propose the following hierarchical algorithm:

1. Determine the best integer shift γ^0 by computing the correlations with the formula (13) from Lemma 1.
2. For $m = 1, 2, \ldots, n$ determine the best dyadic shift $2^{-m}\gamma^m$ among the $3^s - 1$ neighbors $2^{1-m}\gamma^{m-1} + 2^{-m}\kappa$, $\kappa \in \{-1, 1, 1\}^s \setminus \{0\}$ of $2^{1-m}\gamma^{m-1}$.

Note that for $m = 1$ the iteration only requires to apply the subdivision operator S_b to $d^0_\eta(f)$ and $d^1_\eta(g)$ and the matrices $\Phi(2^{-1}\gamma)$ and $\Phi(\gamma) = I\,\tau_\gamma$ to compute

$$c(f)^T \Phi\left(2^{-1}\gamma\right) c(g) + 2^{-s} \sum_{j=0}^{m-1} \sum_{\eta,\eta'\in H} \tau_{-\gamma}\left(S_{b_\eta} d^j_\eta(f) \star S_{b_{\eta'}} d^j_{\eta'}(g)\right) + \sum_{j=1}^{n-1} \sum_{\eta\in H} \tau_{-2^{j-m}\gamma}\left(d^j_\eta(f) \star d^j_\eta(g)\right).$$

If we store the level correlations seperately for $j = 0, \dots, n$, we only need to recompute the first two sums of the above expression. This technique can in general be extended to later iterations, but the two middle sums in (15) also necessitate the application of the subdivision S_a to the low level wavelet coefficient vectors. Of course, the finer the resolution becomes, the effort increases, but keep in mind that also then, due to the hierarchical procedure, only $3^s - 1$ correlations have to be computed which even for $s = 3$ is still the relatively moderate value of 26.

§6. Conclusions

Tomography is much more than a medical diagnosis tool, and the technology available for industrial tomography enables us to generate spectacular measurements with very high resolution or of very large objects. The resulting amount of data, however, is no more tractable on even well-equipped computer systems without switching to sparse representations. This provides a lot of mathematical challenges starting from an efficient creation of such sparse representations, requiring the development of compression and storage strategies that allow fast access to full resolution data in certain regions of interest and leading to a redefinition of standard operations in terms of the sparse representation – the sparse perspective only makes sense if it is respected in all steps of computations and manipulations.

The example of correlation shows that this is doable, even efficiently, but requires some non-straightforward mathematical operations.

Acknowledgements

The work in this paper has been funded by the Bavarian Ministry of Economy by means of the research project “Big Picture”. We also thank Thomas Lang (FORWISS, University of Passau) for providing the examples on segmentation in Fig. 1.

References

[1] Cavaretta, A., Dahmen, W., and Micchelli, C. *Stationary Subdivision*. American Mathematical Society: Memoirs of the American Mathematical Society. American Mathematical Society, 1991.

[2] Censor, Y. Row-action methods for huge and sparse systems and their applications. *SIAM Review 23* (1981), 444–466.

[3] Cotronei, M., Rossini, M., Sauer, T., and Volontè, E. Filters for anisotropic wavelet decompositions. *Journal of Computational and Applied Mathematics 349* (2019), 316 – 330.

[4] Daubechies, I. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics 41*, 7 (1988), 909–996.

[5] Daubechies, I. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[6] Gordon, R., Bender, R., and Herman, G. T. Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *Journal of Theoretical Biology 29*, 3 (1970), 471 – 481.

[7] Han, B. Compactly supported tight wavelet frames and orthonormal wavelets of exponential decay with a general dilation matrix. *Journal of Computational and Applied Mathematics 155* (2003), 43 – 67.

[8] Mallat, S. Multiresolution approximations and wavelet orthonormal bases of l2(r). *Transactions of the American Mathematical Society 315*, 1 (1989), 69–87.

[9] Mallat, S. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., 2008.

[10] Natterer, F. *The Mathematics of Computerized Tomography*. John Wiley & Sons, 1986.

[11] Natterer, F., and Wübbeling, F. *Mathematical Methods in Image Reconstruction*. Society for Industrial and Applied Mathematics, 2001.

[12] Stock, A. M., Herl, G., Sauer, T., and Hiller, J. Edge preserving compression of CT scans using wavelet. *International Symposium on Structural Health Monitoring and Nondestructive Testing* (2018).

[13] Vetterli, M., and Kovačevic, J. *Wavelets and Subband Coding*. Prentice-Hall, Inc., 1995.

B. Diederichs
University of Passau & Fraunhofer IIS Research Group “Knowledge Based Image Processing”
Innstr. 43, D–94032 Passau, Germany
benedikt.diederichs@iis.fraunhofer.de

T. Sauer
Lehrstuhl für Digitale Bildverabeitung & FORWISS, University of Passau
Fraunhofer IIS Research Group “Knowledge Based Image Processing”
Innstr. 43, D–94032 Passau, Germany
tomas.sauer@uni-passau.de

A. M. Stock
FORWISS, University of Passau
Innstr. 43, D–94032 Passau, Germany
stock@forwiss.uni-passau.de

Monografías Matemáticas *García de Galdeano* **42**, 95–109 (2019)

CONVERGENCE AND ERROR ESTIMATES FOR THE COMPRESSIBLE NAVIER-STOKES EQUATIONS

Thierry Gallouët

Abstract. We are interested in the paper by the discretization of the (unsteady and stationary) compressible (isentropic) Navier-Stokes Equations with the Marker-And- Cell scheme. We present recent results for the convergence (as the discretization parameter goes to zero) of the approximate solutions to a weak solution of the continuous equations and error estimates when the solution of the continuous equations is regular enough.

Keywords: Keywords separated by commas.

AMS classification: AMS classification codes.

§1. Introduction

I present in this paper some results obtained with R. Eymard, R. Herbin, J. C. Latché, D. Maltese and A. Novotny.

Let Ω be a bounded open connected set of $\mathbb{R}^3$ with a Lipschitz continuous boundary, $T > 0$, $\gamma > 3/2$, $\mathbf{u_0} \in L^2(\Omega)$, $\rho_0 \in L^\gamma(\Omega)$ and $\mathbf{f} \in L^2(]0,T[, L^2(\Omega)^3)$. The compressible Navier-Stokes equations read

$$\partial_t \rho + \operatorname{div}(\rho \mathbf{u}) = 0 \text{ in } \Omega\times]0,T[, \qquad \text{(mass equation)} \quad (1)$$

$$\partial_t(\rho\mathbf{u}) + \operatorname{div}(\rho \mathbf{u}\otimes\mathbf{u}) - \Delta\mathbf{u} + \operatorname{grad} p = \mathbf{f} \text{ in } \Omega\times]0,T[, \qquad \text{(momentum equation)} \quad (2)$$

$$p = \rho^\gamma \text{ in } \Omega\times]0,T[. \qquad \text{(Equation Of State)} \quad (3)$$

To this system, we add a Dirichlet boundary condition,

$$\mathbf{u} = 0 \text{ on } \partial\Omega\times]0,T[, \quad (4)$$

and an initial condition

$$\mathbf{u}(\cdot,\mathbf{0}) = \mathbf{u_0},\ \rho(\cdot,0) = \rho_0 \text{ on } \partial\Omega. \quad (5)$$

The main unknowns of Problem (1)-(5) are $\mathbf{u}$ and ρ (then, p is given with (3)). Under the assumption $\rho_0 > 0$ a.e. on Ω and $\int_\Omega (\frac{1}{2}\rho_0|\mathbf{u_0}|^2 + \rho_0^\gamma/(\gamma-1))dx < +\infty$, existence of a weak solution $(\mathbf{u},\rho)$ to (1)-(5) is known (but no uniqueness in general) since the works of P.-L. Lions [18] and E. Feireisl and coauthors [5], [6]. This weak solution sastifies $\rho \in L^\infty(]0,T[, L^\gamma(\Omega))$, $\rho \geq 0$ a.e., $\mathbf{u} \in L^2([0,T[, H_0^1(\Omega)^3)$ and $\rho|\mathbf{u}|^2 \in L^\infty(]0,T[, L^1(\Omega))$. Futhermore, $\int_\Omega \rho(x,t)dx = \int_\Omega \rho_0(x)dx$ a.e.. In particular, such a weak solution has a finite energy. More precisely, for a.e. t in $]0,T[$, if $\mathbf{f} = \mathbf{0}$,

$$\int_\Omega (\frac{1}{2}\rho|\mathbf{u}|^2 + \frac{\rho^\gamma}{\gamma-1})(t)\,dx + \int_0^t \int_\Omega |\operatorname{grad}\mathbf{u}|^2\,dxd\tau \leq \int_\Omega (\frac{1}{2}\rho_0|\mathbf{u_0}|^2 + \frac{\rho_0^\gamma}{\gamma-1})dx. \quad (6)$$

It is said that this weak solution is a "suitable" solution.

We are also interested by the stationary compressible Navier Stokes equations. In this case, Ω is a bounded open set of $\mathbb{R}^3$, with a Lipschitz continuous boundary, $\gamma > 3/2$, $\mathbf{f} \in L^2(\Omega)^3$ and $M > 0$. The equations read

$$\mathrm{div}(\rho \mathbf{u}) = 0 \text{ in } \Omega, \tag{7}$$

$$\mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u}) - \Delta \mathbf{u} + \mathrm{grad}\, p = \mathbf{f} \text{ in } \Omega, \tag{8}$$

$$p = \rho^\gamma \text{ in } \Omega. \tag{9}$$

To this system, we add a Dirichlet boundary condition,

$$\mathbf{u} = 0 \text{ on } \partial\Omega \times]0, T[, \tag{10}$$

and

$$\rho \geq 0 \text{ a.e. } \int_\Omega \rho(x) dx = M. \tag{11}$$

Here also, the main unknowns of Problem (7)-(11) are $\mathbf{u}$ and ρ (and p is given with (9)). Existence of a weak solution $(\mathbf{u}, \rho)$ to (7)-(11) is known (but no uniqueness) with $\mathbf{u} \in H_0^1(\Omega)^3$ and $\rho \in L^\gamma(\Omega)$, at least for $\gamma > 5/3$, see for instance [19], [20]. Indeed, the "optimal" space for this weak solution depends on γ (except for $\mathbf{u}$ which always belongs to $H_0^1(\Omega)^3$). If $\gamma > 3$, $\rho \in L^{2\gamma}(\Omega)$ and then $p \in L^2(\Omega)$. If $\gamma < 3$, $\rho \in L^{\gamma\delta}(\Omega)$, with $\delta = 3(\gamma-1)/\gamma$, and then $p \in L^\delta(\Omega)$. In particular, the function ρ belongs to $L^2(\Omega)$ for $\gamma \geq 5/3$.

Remark 1. For $\gamma = 3/2$, one has $\bar{q} = 3(\gamma - 1)/\gamma = 1$, and $\gamma\delta = 3(\gamma - 1) = 3/2$, so that the natural spaces for $p, \rho, \mathbf{u}$ seem to be $p \in L^1(\Omega)$, $\rho \in L^{\frac{3}{2}}(\Omega)$, $\mathbf{u} \in H_0^1(\Omega)^3$. Using the Sobolev embedding $H_0^1(\Omega) \subset L^6(\Omega)$ these natural spaces gives $\rho \mathbf{u} \otimes \mathbf{u} \in L^1(\Omega)^3$. This is a reason for the limitation $\gamma > 3/2$. However, in the case of the stationary compressible Stokes equations (that is without this term $\rho \mathbf{u} \otimes \mathbf{u}$ in (8)), one has a weak solution with $p \in L^2(\Omega)$ (and $\rho \in L^{2\gamma}(\Omega)$) and there is no restriction on γ in the sense that we can take $\gamma \geq 1$ (see for instance [4, 3] for $\gamma > 1$ and [9] for $\gamma = 1$).

For this two problems (Compressible Navier-Stokes Equations and Stationary Compressible Navier-Stokes Equations, namely Problem (1)-(5) and Problem (7)-(11)) we are interested by the discretized models obtained with the Marker-And-Cell scheme (MAC in short) and, for the unsteady problem, with an implicit discretization in time. The reason of this choice is that the MAC scheme is widely used in computational fluid dynamics. It was introduced in [16] and considered (since the beginning) as a suitable space discretization for both incompressible and compressible flow problems (see [14, 15] for the seminal papers and [23] for a review). We refer to [3], [10], [12] for a description of the MAC scheme. Of course, we have to consider a domain Ω adapted to the discretization by the MAC scheme.

Admitting the existence of an approximate solution, that is a solution of the discretized problem (this existence can be proven), two questions ares interesting:

1. Is it possible to prove convergence (up to the subsequence) of the approximate solution to the weak solution of the continuous problem as the mesh size goes to 0 (and also the time step in the evolution case) ?
2. In case of uniqueness of the solution of the continuous problem, is it possible to obtain error estimates and what are they ?

The answer for this two questions are partially known, it remains some open questions (and the known results are completely different between the unsteady case and the steady case) :

1. For the stationary compressible Navier-Stokes problem (namely (7)-(11)), we prove, for $\gamma > 3$, convergence (up to the subsequence) of the approximate solution to the (weak) solution of (7)-(11) as the mesh size goes to 0, see [10]. But it is an open problem for $3/2 < \gamma \leq 3$. Note that for $\gamma > 3$ the proof of convergence given in [10] also gives existence of a weak solution to (7)-(11) since the existence of an approximate solution is also proven in [10].
2. For the compressible Navier-Stokes problem (namely (1)-(5)), the convergence of the approximate solution, up to a subsequence, to the solution of the continuous problem is probably true, but we do not have a complete proof.
3. For the compressible Navier-Stokes problem (namely (1)-(5)), if the solution of the continuous problem is regular enough (then we call it a "strong solution"), we obtain, for $\gamma > 3/2$ an error estimate, cf. [12] for the case $\mathbf{f} = \mathbf{0}$. The rate of convergence obtained in [12] depends on γ and is probably not optimal.
4. For the stationary compressible Navier-Stokes problem, even when the solution of the continuous problem is regular, we are not able to obtain error estimates.

Remark 2. It is possible to obtain some convergence results or some error estimates with other schemes than the MAC scheme. For instance, a convergence result is given for the unsteady compressible Navier-Stokes equations in [17] with a FV-FE scheme, albeit only in the case $\gamma > 3$ (the difficulty in the realistic case $\gamma \leq 3$ arise from the treatment of the non linear convection term). Some error estimates (when the solution of these unsteady compressible Navier-Stokes equations is regular enough) have been derived for this FV-FE scheme in [11] if $\gamma > 3/2$.

§2. Error estimates

2.1. For the compressible Navier-Stokes problem

For the compressible Navier-Stokes problem the proof of an error estimate, that is the comparison of a "strong" solution of Problem (1)-(5) and an approximate solution (that is a solution given by the MAC-scheme in space and an Euler-backward scheme in time) is very close to the so called "weak-strong uniqueness principle", which is the comparison of a "strong" solution and a weak solution of Problem (1)-(5). Indeed, the weak-strong uniqueness principle states that if Problem (1)-(5) has a regular enough solution (the main hypothesis on the solution is the fact that $\operatorname{div} \mathbf{u} \in L^1(]0,T[, L^\infty(\Omega))$ and $\operatorname{grad} p \in L^1(]0,T[, L^\infty(\Omega))$) then Problem (1)-(5) has a unique weak solution (and this solution is equal to the strong solution).

This idea of the weak-strong uniqueness principle comes back to G. Prodi [21] (1959) and J. Serrin [22] (1963) for the case of Incompressible Navier-Stokes Equations. For the compressible isentropic Navier-Stokes equations, the first result is probably in [13]. More general Equation Of State are considered in [6].

For the compressible Navier-Stokes equations, the proof of this weak-strong uniqueness principle uses the so-called "relative entropy" introduced by C. M. Dafermos for Euler Equations [2]. In other papers, the "relative entropy" is called "modulated energy". We will use below the term "relative energy" which seems to be more adapted to our system of equations.

We first describe in Sec. 2.2 this weak-strong uniqueness principle in a very simple case containing the main idea of the method.

2.2. Weak-strong uniqueness principle, simple case

We present in this section the weak-strong uniqueness principle in the case of the compressible Stokes equations with $\gamma = 2$ and $\mathbf{f} = \mathbf{0}$. The set Ω is still a bounded open connected set of $\mathbb{R}^3$, with a Lipschitz continuous boundary and $T > 0$. The problem read

$$\partial_t \rho + \operatorname{div}(\rho \mathbf{u}) = 0 \text{ in } \Omega\times]0,T[, \tag{12}$$

$$\partial_t \mathbf{u} - \Delta \mathbf{u} + \operatorname{grad} p = \mathbf{0} \text{ in } \Omega\times]0,T[, \tag{13}$$

$$p = \rho^2 \text{ in } \Omega\times]0,T[. \tag{14}$$

with a Dirichlet boundary condition,

$$\mathbf{u} = 0 \text{ on } \partial\Omega\times]0,T[, \tag{15}$$

and an initial condition

$$\mathbf{u}(\cdot,\mathbf{0}) = \mathbf{u_0},\ \rho(\cdot,0) = \rho_0 \text{ on } \partial\Omega. \tag{16}$$

Let $(\bar{\mathbf{u}}, \bar{\rho}, \bar{p})$ be a regular solution of (12)-(16) (we call it "strong solution") and let $(\mathbf{u}, \rho, p)$ be a suitable weak solution of (12)-(16).

The idea of the proof is to use a Gronwall inequality on the "relative energy" between $(\mathbf{u}, \rho)$ and $(\bar{\mathbf{u}}, \bar{\rho})$ which reads in this case (Stokes Equations, $\gamma = 2$), for $t \in [0,T]$,

$$E_t(\mathbf{u},\rho \,|\, \bar{\mathbf{u}},\bar{\rho}) = \int_\Omega (\frac{1}{2}|\mathbf{u}(t) - \bar{\mathbf{u}}(t)|^2 + |\rho(t) - \bar{\rho}(t)|^2)\, dx.$$

Note that this quantity is indeed well defined for any t, thanks to some continuity which can be proven for $\mathbf{u}$ and ρ. We now transform formally the quantity $E_t(\mathbf{u},\rho \,|\, \bar{\mathbf{u}},\bar{\rho})$ in three steps, using (12)-(16).

Step 1 Energy Inequalities for the suitable weak solution and for the strong solution

We formally take $\mathbf{u}$ as test function in the momentum equation for $\mathbf{u}$ (Equation (13)) to obtain, for $t \in [0,T]$,

$$\frac{1}{2}\int_\Omega |\mathbf{u}|^2(t)\, dx + \int_0^t \int_\Omega (|\operatorname{grad}\mathbf{u}|^2 - p \operatorname{div}\mathbf{u})\, dx d\tau = \frac{1}{2}\int_\Omega |\mathbf{u_0}|^2 dx. \tag{17}$$

We formally take ρ as test function in the mass equation (Equation (12)) to obtain

$$\frac{1}{2}\int_\Omega \rho^2(t)\, dx - \frac{1}{2}\int_\Omega \rho_0^2\, dx - \int_0^t \int_\Omega \rho \mathbf{u} \cdot \operatorname{grad}\rho\, dx d\tau = 0.$$

But, since $\rho^2 = p$,

$$\int_0^t \int_\Omega \rho \mathbf{u} \cdot \operatorname{grad} \rho \, dx d\tau = \frac{1}{2} \int_0^t \int_\Omega \mathbf{u} \cdot \operatorname{grad}(\rho^2) \, dx d\tau = -\frac{1}{2} \int_0^t \int_\Omega p \operatorname{div} \mathbf{u} \, dx d\tau$$

and then

$$\int_\Omega \rho^2(t) \, dx + \int_0^t \int_\Omega p \operatorname{div} \mathbf{u} \, dx d\tau = \int_\Omega \rho_0^2 \, dx. \tag{18}$$

Then, adding Equations (17) and (18) gives for all $t \in [0, T]$,

$$\frac{1}{2} \int_\Omega |\mathbf{u}|^2(t) \, dx + \int_\Omega \rho^2(t) \, dx + \int_0^t \int_\Omega (|\operatorname{grad} \mathbf{u}|^2 \, dx d\tau = \frac{1}{2} \int_\Omega |\mathbf{u_0}|^2 dx + \int_\Omega \rho_0^2 \, dx.$$

Indeed, this is Inequality (6) with an equality instead of an inequality, but the computation here is formal. For the suitable weak solution, one has Inequality (6) which is here

$$\frac{1}{2} \int_\Omega |\mathbf{u}|^2(t) \, dx + \int_\Omega \rho^2(t) \, dx + \int_0^t \int_\Omega (|\operatorname{grad} \mathbf{u}|^2 \, dx d\tau \leq \frac{1}{2} \int_\Omega |\mathbf{u_0}|^2 dx + \int_\Omega \rho_0^2 \, dx. \tag{19}$$

For the the strong solution (which is "more" than a suitable weak solution), one has also

$$\frac{1}{2} \int_\Omega |\bar{\mathbf{u}}|^2(t) \, dx + \int_\Omega \bar{\rho}^2(t) \, dx + \int_0^t \int_\Omega (|\operatorname{grad} \bar{\mathbf{u}}|^2 \, dx d\tau \leq \frac{1}{2} \int_\Omega |\mathbf{u_0}|^2 dx + \int_\Omega \rho_0^2 \, dx. \tag{20}$$

Using (19) and (20), for all t,

$$\begin{aligned} E_t(\mathbf{u}, \rho \,|\, \bar{\mathbf{u}}, \bar{\rho}) = \int_\Omega (\frac{1}{2}|\mathbf{u}(t) - \bar{\mathbf{u}}(t)|^2 + |\rho - \bar{\rho}|^2) \, dx \leq \\ - \int_\Omega \mathbf{u}(t) \cdot \bar{\mathbf{u}}(t) \, dx - 2 \int_\Omega \rho(t) \bar{\rho}(t) \, dx - \int_0^t \int_\Omega (|\operatorname{grad} \mathbf{u}|^2 + |\operatorname{grad} \bar{\mathbf{u}}|^2) \, dx d\tau \\ + \int_\Omega |\mathbf{u_0}|^2 \, dx + 2 \int_\Omega |\rho_0|^2 \, dx, \end{aligned} \tag{21}$$

We have now to transform the two first terms of the right hand side of (21).

Step 2 Transformation of $\int_\Omega \rho(t) \bar{\rho}(t) \, dx$.

Using the regularity of the strong solution, we can take $\bar{\rho}$ as test function in the mass equation for the weak solution (Equation (12)) and ρ as test function in the mass equation for the strong solution. This gives

$$\int_0^t \int_\Omega (\partial_t \rho) \bar{\rho} \, dx d\tau - \int_0^t \int_\Omega \rho u \cdot \operatorname{grad} \bar{\rho} \, dx d\tau = 0,$$

$$\int_0^t \int_\Omega (\partial_t \bar{\rho}) \rho \, dx d\tau + \int_0^t \int_\Omega \operatorname{div}(\bar{\rho} \bar{u}) \rho \, dx d\tau = 0.$$

The non-symmetry between these two equalities is due to fact that $(\mathbf{u}, \rho)$ is only a weak solution. Adding the two equations leads to

$$\int_\Omega \bar{\rho}(t) \rho(t) \, dx - \int_\Omega \rho_0^2 \, dx = \int_0^t \int_\Omega \rho \mathbf{u} \cdot \operatorname{grad} \bar{\rho} \, dx d\tau - \int_0^t \int_\Omega \operatorname{div}(\bar{\rho} \bar{\mathbf{u}}) \rho \, dx d\tau. \tag{22}$$

Step 3 Transformation of $\int_\Omega \mathbf{u}(t)\cdot\bar{\mathbf{u}}(t)\,dx$.

Using, here also, the regularity of the strong solution, we can take $\bar{\mathbf{u}}$ as test function in the momentum equation for the weak solution (Equation (13)) and $\mathbf{u}$ as test function in the momentum equation for the strong solution. This gives

$$\int_0^t\int_\Omega (\partial_t\mathbf{u})\bar{\mathbf{u}}\,dxd\tau + \int_0^t\int_\Omega (\operatorname{grad}\mathbf{u} : \operatorname{grad}\bar{\mathbf{u}} - p\operatorname{div}(\bar{\mathbf{u}}))\,dxd\tau = 0,$$

$$\int_0^t\int_\Omega (\partial_t\bar{\mathbf{u}})\mathbf{u}\,dxd\tau + \int_0^T\int_\Omega (\operatorname{grad}\mathbf{u} : \operatorname{grad}\bar{\mathbf{u}} + \mathbf{u}\cdot\operatorname{grad}\bar{p})\,dxd\tau = 0.$$

Adding the two equations leads to

$$\int_\Omega \bar{\mathbf{u}}(t)\cdot\mathbf{u}(t)\,dx - \int_\Omega |\mathbf{u_0}|^2\,dx = \int_0^t\int_\Omega (-2\operatorname{grad}\mathbf{u} : \operatorname{grad}\bar{\mathbf{u}} + p\operatorname{div}(\bar{\mathbf{u}}) - \mathbf{u}\cdot\operatorname{grad}\bar{p})\,dxd\tau. \quad (23)$$

Step 4 End of the proof of the weak strong uniqueness principle

We use (22) and (23) to transform (21). We obtain

$$E_t(\mathbf{u},\rho\,|\,\bar{\mathbf{u}},\bar{\rho}) \le -\int_0^t\int_\Omega |\operatorname{grad}\mathbf{u} - \operatorname{grad}\bar{\mathbf{u}}|^2\,dxd\tau - \int_0^t\int_\Omega (p\operatorname{div}(\bar{\mathbf{u}}) - \mathbf{u}\cdot\operatorname{grad}\bar{p})\,dxd\tau$$
$$- 2\int_0^t\int_\Omega \rho\mathbf{u}\cdot\operatorname{grad}\bar{\rho}\,dxd\tau + 2\int_0^t\int_\Omega \operatorname{div}(\bar{\rho}\bar{\mathbf{u}})\rho\,dxd\tau.$$

Using $p = \rho^2$, $\bar{p} = \bar{\rho}^2$, $\int_0^t\int_\Omega \operatorname{div}(\bar{\rho}\bar{\mathbf{u}})\rho\,dxd\tau = \int_0^t\int_\Omega(\bar{\rho}\rho\operatorname{div}(\bar{\mathbf{u}}) + \rho\bar{\mathbf{u}}\cdot\operatorname{grad}\bar{\rho})\,dxd\tau$ and $\int_0^t\int_\Omega(\operatorname{div}(\bar{\mathbf{u}})\bar{\rho}^2 + 2\bar{\mathbf{u}}\bar{\rho}\cdot\operatorname{grad}\bar{\rho})dxd\tau = 0$, this inequality can be rewritten as

$$E_t(\mathbf{u},\rho\,|\,\bar{\mathbf{u}},\bar{\rho}) \le \int_0^t\int_\Omega (-|\operatorname{grad}\mathbf{u} - \operatorname{grad}\bar{\mathbf{u}}|^2 - (\rho-\bar{\rho})^2\operatorname{div}(\bar{\mathbf{u}}) - 2(\bar{\rho}-\rho)(\bar{\mathbf{u}}-\mathbf{u})\cdot\operatorname{grad}\bar{\rho})\,dxd\tau.$$

and then

$$E_t(\mathbf{u},\rho\,|\,\bar{\mathbf{u}},\bar{\rho}) \le \int_0^t\int_\Omega (-(\rho-\bar{\rho})^2\operatorname{div}(\bar{\mathbf{u}}) - 2(\bar{\rho}-\rho)(\bar{\mathbf{u}}-\mathbf{u})\cdot\operatorname{grad}\bar{\rho})\,dxd\tau. \quad (24)$$

Setting $\varphi(t) = E_t(\rho,\mathbf{u}|\bar{\rho},\bar{\mathbf{u}}) = \frac{1}{2}\int_\Omega |\mathbf{u}(t)-\bar{\mathbf{u}}(t)|^2\,dx + \int_\Omega(\rho(t)-\bar{\rho}(t))^2\,dx$, using Cauchy-Schwarz Inequality for the last term, we obtain from (24), since $\operatorname{div}\bar{\mathbf{u}} \in L^1(]0,T[,L^\infty(\Omega))$ and $\operatorname{grad}\bar{\rho} \in L^1(]0,T[,L^\infty(\Omega))$,

$$\varphi(t) \le C\int_0^t a(\tau)\varphi(\tau)d\tau \text{ for all } t\in[0,T],$$

with some $a \in L^1(]0,T[)$. This gives, by Gronwall Inequality, $\varphi(t) \le \varphi(0)e^{\int_0^t a(\tau)d\tau}$ and then, since $\varphi(0) = 0$, $\varphi(t) = 0$ for all $t \in [0,T]$. The weak-strong uniqueness principle is then proven for this simple case (compressible Stokes equations with $\gamma = 2$ and $\mathbf{f} = \mathbf{0}$).

2.3. Error estimate for the compressible Navier Stokes equations

We consider here the compressible Navier Stokes system (1)-(5) with $\mathbf{f} = \mathbf{0}$, $\gamma > 3/2$ and a domain Ω adapted to the MAC scheme (for instance, $\Omega =]0,1[^3$). Mimicking the previous proof of uniqueness (given in Sec. 2.2) at the discrete level it is possible to obtain error estimates, that is an estimate between a strong solution (we assume existence of such a solution) and the approximate solution given by a numerical scheme (roughly speaking it is not so far of a weak solution with some errors due to the discretization). Instead of a suitable weak solution (ρ, u), we use now the solution of the scheme (that is the solution obtained with a space discretization using the MAC scheme and an Euler-backward discretiztation in time). This numerical solution is denoted by $(\mathbf{u}, \rho)$ and the strong solution is denoted by $(\mathbf{u}, \bar{\rho})$. The energy is now

$$E_t(\mathbf{u}, \rho \,|\, \bar{\mathbf{u}}, \bar{\rho}) = \int_\Omega (\frac{1}{2}\rho|u(t) - \bar{u}(t)|^2 + e(\rho(t)|\bar{\rho}(t)))\, dx,$$

with $e(\rho|\bar{\rho}) = \rho^\gamma - \bar{\rho}^{\gamma-1}\gamma(\rho - \bar{\rho}) - \bar{\rho}^\gamma$. Note that $e(\rho|\bar{\rho}) = 0$ if and only if $\rho = \bar{\rho}$.

If h is the mesh size and k the time step, the error estimate given in [12] is

$$E_t(\rho, u|\bar{\rho}, \bar{u}) \leq C(h^\alpha + k^{1/2}) \text{ for all } 0 \leq t \leq T,$$

where C depends only on the strong solution and on the regularity of the mesh and $\alpha = \min(\frac{2\gamma-3}{\gamma}, \frac{1}{2})$. For $\gamma = 2$, one has $\alpha = 1/2$ and E_t is the L^2-norm of $(\rho - \bar{\rho})$ plus the L^2-norm of $(u - \bar{u})$ weighted by ρ (and we have $\rho > 0$ a.e.).

2.4. For the stationary compressible Navier-Stokes problem

We are not able to give error estimates for the stationary compressible Navier-Stokes problem (that is problem (7)-(11)) as we did for the compressible Navier-Stokes problem in Sec. 2.3. The proof in Sec. 2.3 follows closely the proof of the weak-strong uniqueness principle. A crucial tool in the proof of weak-strong uniqueness principle is the use of the Gronwall inequality. Then a natural question is "What can play the role of Gronwall Inequality for stationary problems" ?

We present below a very simple example where uniqueness in the unsteady case follows easily from the Gronwall inequality and uniqueness is also true in the stationary case, with a trick which has some similarity with the Gronwall inequality. Unfortunately, we are not able to adapt the same trick in the case of the stationary compressible Navier-Stokes problem.

Let Ω be a bounded open set of $\mathbb{R}^3$, $T > 0$, $\mathbf{w} \in L^\infty(\Omega)^3$, $f \in L^2(]0,T[, L^2(\Omega))$, $u_0 \in L^2(\Omega)$ and φ be Lipschitz continuous function from $\mathbb{R}$ to $\mathbb{R}$. We consider the following problem,

$$\begin{aligned} \partial_t u + \operatorname{div}(\mathbf{w}\varphi(u)) - \Delta u &= f \text{ in } \Omega\times]0,T[, \\ u(\cdot,t) &= 0 \text{ on } \partial\Omega \text{ for all } t \in]0,T[, \\ u(\cdot,0) &= u_0 \text{ on } \partial\Omega. \end{aligned}$$

For this problem, one has existence of the solution in the space $L^2(]0,T[, H^1_0(\Omega))$ and the solution is continuous with value in $L^2(\Omega)$. Uniqueness easily follows from a Gronwall inequality.

We now consider the stationary case, that is $f \in L^2(\Omega)$ (and still $\mathbf{w} \in L^\infty(\Omega)^3$, φ Lipschitz continuous) and the stationary problem reads

$$\begin{aligned} \operatorname{div}(\mathbf{w}\varphi(u)) - \Delta u &= f \text{ in } \Omega, \\ u(\cdot, t) &= 0 \text{ on } \partial\Omega. \end{aligned}$$

Note that we do not have any hypothesis on $\operatorname{div}(\mathbf{w})$. Then, we may have a non-coercive differential operator.

For this problem, it is possible to prove existence in the space $H_0^1(\Omega)$ (for instance cf. [8], Exercice 3.5). But, for this problem, it is also possible to prove uniqueness. If u and $\bar{u}$ are two solutions, the idea is to take $T_\varepsilon(u - \bar{u})$ ($\varepsilon > 0$) as test function, where $T_\varepsilon(s) = \max(-\varepsilon, \min(s, \varepsilon))$ for $s \in \mathbb{R}$.

Using in particular Sobolev Injection of $W_0^{1,1}(\Omega)$ in $L^{1^\star}(\Omega)$ (with $1^\star = 3/2$ since $\Omega \subset \mathbb{R}^3$) and letting $\varepsilon \to 0$ allows us to conclude $u = \bar{u}$ a.e.. (for instance, cf. [1] or [8] Exercice 3.6.)

§3. Convergence results

3.1. For the stationary compressible Navier-Stokes problem

For the stationary compressible Navier-Stokes equations (7)-(11) discretized with a MAC scheme (of course, we assume that Ω is adapted to the MAC scheme), we prove (cf. [10]) convergence of the approximate solution (up to a subsequence) to a weak solution, in the case $\gamma > 3$ (and $f \in L^2(\Omega)^3$, $M > 0$) following the idea of P.L. Lions (cf. [18]) for proving existence of a solution.

Let $(\mathbf{u}_n, p_n, \rho_n)_{n\in\mathbb{N}}$ be a sequence a approximate solutions obtained with the MAC scheme (existence of such an approximate solution is proven, cf. [10]). We assume $\lim_{n\to+\infty} h_n = 0$, where h_n is the mesh size. The steps for proving the convergence result are

1. Estimates on the approximate solution $(\mathbf{u}_n, p_n, \rho_n)$;
2. Compactness result (convergence of the approximate solution, up to a subsequence);
3. Passage to the limit in the approximate equations.

The main difficulty is in the passage to the limit in the EOS ($p = \rho^\gamma$) since the EOS is a non linear function and Step 2 only leads to weak convergences of p_n and ρ_n.

The estimate on $\mathbf{u}_n$ is with a norm which mimic (at the discrete level) the $H_0^1(\Omega)^3$-norm. The estimate on p_n is in $L^2(\Omega)$-norm (thanks to $\gamma \geq 3$) and the estimate on ρ_n is in $L^{2\gamma}(\Omega)$-norm. Thanks to these estimates on $\mathbf{u}_n$, p_n, ρ_n, it is possible to assume (up to a subsequence) that, as $n \to +\infty$,

$$\begin{aligned} \mathbf{u}_n \to \mathbf{u} \text{ in } L^q(\Omega)^3 \text{ for } q < 6 \text{ and weakly in } L^6(\Omega)^3, \ \mathbf{u} \in H_0^1(\Omega)^3, \\ p_n \to p \text{ weakly in } L^2(\Omega), \quad \rho_n \to \rho \text{ weakly in } L^{2\gamma}(\Omega). \end{aligned}$$

We show now how to pass to the limit in the equations. For simplicity we will asume that $(\mathbf{u}_n, p_n, \rho_n)$ is a weak solution of (7)-(11) with $\mathbf{f}_n$ instead of $\mathbf{f}$, and $\mathbf{f}_n \to \mathbf{f}$ weakly in $L^2(\Omega)^3$ as $n \to +\infty$. The passage to the limit in the equation when $(\mathbf{u}_n, p_n, \rho_n)$ is an approximate

solution given by MAC scheme follows the same lines, with some modifications that we indicate when there are interesting.

For the mass equation, let $v \in C_c^\infty(\mathbb{R}^3)$, one has

$$\int_\Omega \rho_n \mathbf{u}_n \cdot \operatorname{grad} v = 0, \tag{25}$$

Since $\rho_n \to \rho$ weakly in $L^{2\gamma}(\Omega)$, with $2\gamma > 6/5$, and $\mathbf{u}_n \to \mathbf{u}$ in $L^q(\Omega)^3$ for all $q < 6$. Then $\rho_n \mathbf{u}_n \to \rho \mathbf{u}$ weakly in $L^1(\Omega)^3$. This gives $\int_\Omega \rho \mathbf{u} \cdot \operatorname{grad} v = 0$. Indeed, at the discrete level, in Equation (25), there is an additional term which allows us to prove $\int_\Omega \rho_n \, dx = M$. This term vanishes as $n \to +\infty$ since it is of ordre h_n^α, where $\alpha \in]0, 1[$ is a given parameter (cf. [10]).

The L^1-weak convergence of ρ_n gives non negativity of ρ and convergence of the total mass, that is $\rho \geq 0$ a.e. in Ω, $\int_\Omega \rho(x) dx = M$. For the momentum equation, let $\mathbf{v} \in C_c^\infty(\Omega)^3$,

$$\int_\Omega \operatorname{grad} \mathbf{u}_n : \operatorname{grad} \mathbf{v} \, dx - \int_\Omega \rho_n \mathbf{u}_n \otimes \mathbf{u}_n : \operatorname{grad} \mathbf{v} \, dx - \int_\Omega p_n \operatorname{div}(\mathbf{v}) dx = \int_\Omega \mathbf{f}_n \cdot \mathbf{u} \, dx \tag{26}$$

This is also true at the discrete level with an error term (vanishing as $n \to +\infty$) and a discret operator grad_n (acting on $\mathbf{u}_n$) mimicking grad. One has, as $n \to +\infty$, $\operatorname{grad} \mathbf{u}_n \to \operatorname{grad} \mathbf{u}$ weakly in $L^2(\Omega)^3$ (this is also true at the discrete level with grad_n isntead of grad). Furthemore, using $\rho_n \to \rho$ weakly in $L^{2\gamma}(\Omega)$, with $2\gamma > 3/2$, and $\mathbf{u}_n \to \mathbf{u}$ in $L^q(\Omega)^3$ for all $q < 6$ (and $\frac{2}{3} + \frac{1}{6} + \frac{1}{6} = 1$), $\rho_n \mathbf{u}_n \otimes \mathbf{u}_n \to \rho \mathbf{u} \otimes \mathbf{u}$ weakly in $L^1(\Omega)^{3\times 3}$. It remains to remark that $p_n \to p$ weakly in $L^2(\Omega)$ and $\mathbf{f}_n \to \mathbf{f}$ weakly in $L^2(\Omega)^3$. Then, we can pass to the limit in (26), it gives

$$\int_\Omega \operatorname{grad} \mathbf{u} : \operatorname{grad} \mathbf{v} \, dx - \int_\Omega \rho \mathbf{u} \otimes \mathbf{u} : \operatorname{grad} \mathbf{v} \, dx - \int_\Omega p \operatorname{div}(\mathbf{v}) \, dx = \int_\Omega \mathbf{f} \cdot \mathbf{v} \, dx.$$

For the moment, we proved that $(\mathbf{u}_n, p_n, \rho_n)$ is solution of the momentum equation and of the mass equation. We also proved non negativity of ρ and $\int_\Omega \rho dx = M$. It remains to prove $p = \rho^\gamma$. This is not easy since p_n and ρ_n converge only weakly... and $\gamma > 1$.

In order to prove $p = \rho^\gamma$ a.e. in Ω, the main step is to prove that

$$\liminf_{n \to +\infty} \int_\Omega p_n \rho_n \, dx \leq \int_\Omega p \rho \, dx. \tag{27}$$

(Then, we deduce the a.e. convergence of p_n and ρ_n and $p = \rho^\gamma$ using the fact that the function $y \mapsto y^\gamma$ is increasing and a variant of the Minty trick.) Note that for $\gamma < 3$ the natural spaces given in Sec. 1 are $L^{3(\gamma-1)}$ for p and $L^{3(\gamma-1)/\gamma}$ fro ρ. Then, we need here $\gamma \geq 2$, in order to have $p\rho \in L^1(\Omega)$.

In order to prove (27), we first remark that, for all $\bar{\mathbf{u}}, \bar{\mathbf{v}}$ in $H_0^1(\Omega)^3$,

$$\int_\Omega \operatorname{grad} \bar{\mathbf{u}} : \operatorname{grad} \bar{\mathbf{v}} \, dx = \int_\Omega \operatorname{div}(\bar{\mathbf{u}}) \operatorname{div}(\bar{\mathbf{v}}) \, dx + \int_\Omega \operatorname{curl}(\bar{\mathbf{u}}) \cdot \operatorname{curl}(\bar{\mathbf{v}}) \, dx. \tag{28}$$

A similar equality is true at the discrete level with the MAC scheme and the natural discrete operators grad_n and div_n (acting on discrete functions), cf. [3] (this is the first "miracle" with the Mac scheme). With other schemes, it seems that there is not a similar equality and this

introduces an additional difficulty, needing, for instance, a "regularization" term for proving the convergence of the scheme, cf. [4].

Using (28), the momentum equation is, for all $\bar{\mathbf{v}}$ in $H_0^1(\Omega)^3$,

$$\int_\Omega \mathrm{div}(\mathbf{u}_n)\,\mathrm{div}(\bar{\mathbf{v}})\,dx + \int_\Omega \mathrm{curl}(\mathbf{u}_n)\cdot\mathrm{curl}(\bar{\mathbf{v}})\,dx - \int_\Omega (\rho_n\mathbf{u}_n\otimes\mathbf{u}_n) : \mathrm{grad}\,\bar{\mathbf{v}}\,dx \\ - \int_\Omega p_n\,\mathrm{div}(\bar{\mathbf{v}})\,dx = \int_\Omega \mathbf{f}_n\cdot\bar{\mathbf{v}}\,dx \quad (29)$$

Our aim is now to choose $\bar{\mathbf{v}} = \bar{\mathbf{v}}_n$ with $\mathrm{curl}(\bar{\mathbf{v}}_n) = 0$, $\mathrm{div}(\bar{\mathbf{v}}_n) = \rho_n$ and $(\bar{\mathbf{v}}_n)_{n\in\mathbb{N}}$ bounded in $H_0^1(\Omega)^3$. Unfortunately, it is possible to choose such a $\bar{\mathbf{v}}_n$ in $H^1(\Omega)^3$ (as we will below) but not in $H_0^1(\Omega)^3$. Assuming anyway that we can have such a $\bar{\mathbf{v}}_n$ in $H_0^1(\Omega)^3$, then, up to a subsequence,

$$\bar{\mathbf{v}}_n \to v \text{ in } L^2(\Omega)^3 \text{ and weakly in } H_0^1(\Omega)^3,\ \mathrm{curl}(\mathbf{v}) = 0,\ \mathrm{div}(\mathbf{v}) = \rho,$$

and (29) becomes

$$\int_\Omega (\mathrm{div}(\mathbf{u}_n) - p_n)\rho_n\,dx = \int_\Omega \rho_n\mathbf{u}_n\otimes\mathbf{u}_n : \mathrm{grad}\,\bar{v}_n\,dx + \int_\Omega f_n\cdot\bar{\mathbf{v}}_n\,dx.$$

If we prove that $\int_\Omega \rho_n\mathbf{u}_n\otimes\mathbf{u}_n : \mathrm{grad}\,\bar{\mathbf{v}}_n\,dx \to \int_\Omega \rho\mathbf{u}\otimes\mathbf{u} : \mathrm{grad}\,\mathbf{v}\,dx$ then

$$\lim_{n\to+\infty}\int_\Omega (\mathrm{div}(\mathbf{u}_n) - p_n)\rho_n\,dx = \int_\Omega \rho\mathbf{u}\otimes\mathbf{u} : \mathrm{grad}\,\mathbf{v}\,dx + \int_\Omega \mathbf{f}\cdot\mathbf{v}\,dx.$$

But, since we already know that $-\Delta\mathbf{u} + \mathrm{div}(\rho\mathbf{u}\otimes\mathbf{u}) + \mathrm{grad}\,p = \mathbf{f}$,

$$\int_\Omega \mathrm{div}(\mathbf{u})\,\mathrm{div}(\mathbf{v})\,dx + \int_\Omega \mathrm{curl}(\mathbf{u})\cdot\mathrm{curl}(\mathbf{v})\,dx - \int_\Omega p\,\mathrm{div}(\mathbf{v}) = \int_\Omega \rho\mathbf{u}\otimes\mathbf{u} : \mathrm{grad}\,\mathbf{v}\,dx + \int_\Omega \mathbf{f}\cdot\mathbf{v}\,dx,$$

which gives (using $\mathrm{div}\,\mathbf{v} = \rho$ and $\mathrm{curl}\,\mathbf{v} = 0$)

$$\int_\Omega (\mathrm{div}(\mathbf{u}) - p)\rho\,dx = \int_\Omega \rho\mathbf{u}\otimes\mathbf{u} : \mathrm{grad}\,\mathbf{v}\,dx + \int_\Omega \mathbf{f}\cdot\mathbf{v}\,dx,$$

Then, $\lim_{n\to+\infty}\int_\Omega (p_n - \mathrm{div}(\mathbf{u}_n))\rho_n\,dx = \int_\Omega (p - \mathrm{div}(\mathbf{u}))\rho\,dx$.

Finally, thanks to the mass equations, we can prove $\int_\Omega \rho_n\,\mathrm{div}(\mathbf{u}_n)\,dx = 0$ and $\int_\Omega \rho\,\mathrm{div}(\mathbf{u})\,dx = 0$. Then, $\lim_{n\to+\infty}\int_\Omega p_n\rho_n\,dx = \int_\Omega p\rho\,dx$.

Indeed, at the discrete level, one has only $\int_\Omega \rho_n\,\mathrm{div}(\mathbf{u}_n)\,dx \le 0$ and (27) is proven (even with lim sup instead of lim inf). It remains to prove

$$\int_\Omega \rho_n\mathbf{u}_n\otimes\mathbf{u}_n : \mathrm{grad}\,\bar{\mathbf{v}}_n\,dx \to \int_\Omega \rho\mathbf{u}\otimes\mathbf{u} : \mathrm{grad}\,\mathbf{v}\,dx. \quad (30)$$

We remark that (since $\mathrm{div}(\rho_n\mathbf{u}_n) = 0$)

$$\int_\Omega \rho_n\mathbf{u}_n\otimes\mathbf{u}_n : \mathrm{grad}\,\bar{\mathbf{v}}_n\,dx = \int_\Omega (\rho_n\mathbf{u}_n\cdot\mathrm{grad})\mathbf{u}_n\cdot\bar{\mathbf{v}}_n\,dx,$$

and the sequence $((\rho_n \mathbf{u}_n \cdot \mathrm{grad})\mathbf{u}_n)_{n\in\mathbb{N}}$ is bounded in $L^r(\Omega)^3$ with $\frac{1}{r} = \frac{1}{2} + \frac{1}{6} + \frac{1}{2\gamma}$, and $r > \frac{6}{5}$ since $\gamma > 3$. Then, up to a subsequence, $(\rho_n \mathbf{u}_n \cdot \mathrm{grad})\mathbf{u}_n \to G$ weakly in $L^r(\Omega)^3$. and (since $\bar{v}_n \to \bar{v}$ in $L^r(\Omega)^3$ for all $r < 6$),

$$\int_\Omega (\rho_n \mathbf{u}_n \cdot \mathrm{grad})\mathbf{u}_n \cdot \bar{v}_n \, dx \to \int_\Omega G \cdot \bar{v} \, dx.$$

But, $G = (\rho \mathbf{u} \cdot \mathrm{grad})\mathbf{u}$, since for a fixed $\mathbf{w} \in H_0^1(\Omega)^3$,

$$\int_\Omega (\rho_n \mathbf{u}_n \cdot \mathrm{grad})\mathbf{u}_n \cdot \mathbf{w} \, dx = \int_\Omega \rho_n \mathbf{u}_n \otimes \mathbf{u}_n : \mathrm{grad}\, \mathbf{w} \, dx \to \int_\Omega \rho \mathbf{u} \otimes \mathbf{u} : \mathrm{grad}\, \mathbf{w} \, dx.$$

Then, (30) is proven and this gives (27) except that there is a mistake in the previous proof since it is not possible to have such a $\bar{\mathbf{v}}_n$ in $H_0^1(\Omega)^3$ such that $\mathrm{curl}\, \bar{\mathbf{v}}_n = 0$, $\mathrm{div}\, \bar{\mathbf{v}}_n = \rho_n$ and $(\bar{\mathbf{v}}_n)_{n\in\mathbb{N}}$ bounded in $H_0^1(\Omega)^3$. In order to correct to proof, we will use such a $\bar{\mathbf{v}}_n$ in $H^1(\Omega)^3$ but not in $H_0^1(\Omega)^3$.

Let $w_n \in H_0^1(\Omega)$, $-\Delta w_n = \rho_n$, It is well known that $w_n \in H^2_{loc}(\Omega)$ (equivalent to say here, since $w_n \in H^1(\Omega)$, $\Delta(w_n\varphi) \in L^2(\Omega)$ for all $\varphi \in C_c^\infty(\Omega)$). An easy way to prove this regularity result is to remark that, for $\varphi \in C_c^\infty(\Omega)$, with C_φ depending only on φ and of the bound of the L^2-norm of ρ_n,

$$\sum_{i,j=1}^3 \int_\Omega \partial_i\partial_j(w_n\varphi)\, \partial_i\partial_j(w_n\varphi)\, dx = \sum_{i,j=1}^3 \int_\Omega \partial_i\partial_i(w_n\varphi)\, \partial_j\partial_j(w_n\varphi)\, dx$$
$$= \int_\Omega (\Delta(w_n\varphi))^2 \, dx = C_\varphi < +\infty.$$

The main interest of this way to prove he H^2_{loc}-regularity of w_n is that it is possible to prove a discrete version of this result with the corresponding discrete problem obtained on the primal mesh of the MAC discretization. Namely, we obtain an H^2_{loc}-discrete estimate on w_n in term of the L^2-norm of ρ_n when w_n is the solution of the discrete problem (it is the second miracle for the MAC scheme).

To continue our proof of (27), we take $\mathbf{v}_n = \mathrm{grad}\, w_n$ so that $\mathrm{div}\, \mathbf{v}_n = \rho_n$ and $\mathrm{curl}\, \mathbf{v}_n = 0$ a.e. in Ω. Furthermore, thanks to the H^2_{loc}-discrete estimate, the sequence $(\mathbf{v}_n)_{n\in\mathbb{N}}$ is bounded in $(H^1_{loc}(\Omega))^3$. Then, up to a subsequence, as $n \to +\infty$, $\mathbf{v}_n \to \mathbf{v}$ in $L^2_{loc}(\Omega)$ and weakly in $H^1_{loc}(\Omega)$, $\mathrm{curl}(\mathbf{v}) = 0$, $\mathrm{div}(\mathbf{v}) = \rho$.

Let $\varphi \in C_c^\infty(\Omega)$ (so that $\mathbf{v}_n\varphi \in H_0^1(\Omega)^3)$). Taking $\bar{\mathbf{v}} = \mathbf{v}_n\varphi$ in (29) gives

$$\int_\Omega \mathrm{div}(\mathbf{u}_n)\, \mathrm{div}(\mathbf{v}_n\varphi)\, dx + \int_\Omega \mathrm{curl}(\mathbf{u}_n) \cdot \mathrm{curl}(\mathbf{v}_n\varphi)\, dx - \int_\Omega p_n\, \mathrm{div}(\mathbf{v}_n\varphi)\, dx$$
$$= \int_\Omega \rho_n \mathbf{u}_n \otimes \mathbf{u}_n : \mathrm{grad}(\mathbf{v}_n\varphi)\, dx + \int_\Omega \mathbf{f}_n \cdot (\mathbf{v}_n\varphi)\, dx.$$

Using a proof similar to that given if $\varphi = 1$ (with additional terms involving φ), we obtain, as $n \to +\infty$,

$$\lim_{n\to+\infty} \int_\Omega (p_n - \mathrm{div}(\mathbf{u}_n))\rho_n\varphi \, dx = \int_\Omega (p - \mathrm{div}(\mathbf{u}))\rho\varphi \, dx \text{ for all } \varphi \in C_c^\infty(\Omega),$$

that is $F_n = (p_n - \mathrm{div}(\mathbf{u}_n))\rho_n) \to F = (p - \mathrm{div}(\mathbf{u}))\rho$ in the distribution sense. But since $(F_n)_{n\in\mathbb{N}}$ bounded in L^q for some $q > 1$ (this is due to the fact that $p_n - \mathrm{div}(\mathbf{u}_n)$ is bounded in $L^2(\Omega)$ and ρ_n is bounded in $L^r(\Omega)$ with some $r > 2$, here we use $\gamma > 5/3$), one has also $F_n \to F$ weakly in $L^1(\Omega)$ and therefore

$$\int_\Omega (p_n - \mathrm{div}(\mathbf{u}_n))\rho_n \, dx \to \int_\Omega (p - \mathrm{div}(\mathbf{u}))\rho \, dx.$$

Finally, thanks to the mass equations, $\int_\Omega \mathrm{div}(\mathbf{u})\rho \, dx = 0$ and $\int_\Omega \mathrm{div}(\mathbf{u}_n)\rho_n \, dx = 0$ (or $\le$ in the case of the discrete setting) and one obtains (27), that is $\liminf_{n\to+\infty} \int_\Omega p_n\rho_n \, dx \le \int_\Omega p\rho \, dx$.

We prove now the a.e. convergence of ρ_n and p_n. Let $G_n = (\rho_n^\gamma - \rho^\gamma)(\rho_n - \rho)$ so that $G_n \in L^1(\Omega)$ and $G_n \ge 0$ a.e. in Ω. Futhermore $G_n = (p_n - \rho^\gamma)(\rho_n - \rho) = p_n\rho_n - p_n\rho - \rho^\gamma\rho_n + \rho^\gamma\rho$ and:

$$\int_\Omega G_n \, dx = \int_\Omega p_n\rho_n \, dx - \int_\Omega p_n\rho \, dx - \int_\Omega \rho^\gamma\rho_n \, dx + \int_\Omega \rho^\gamma\rho \, dx.$$

Using the weak convergence in $L^2(\Omega)$ of p_n and ρ_n and (27), $\liminf_{n\to+\infty} \int_\Omega G_n = 0$. Then (up to a subsequence), $G_n \to 0$ a.e. and then $\rho_n \to \rho$ a.e. (since $y \mapsto y^\gamma$ is an increasing function on $\mathbb{R}_+$). Finally, $\rho_n \to \rho$ in $L^q(\Omega)$ for all $1 \le q < 2\gamma$, $p_n = \rho_n^\gamma \to \rho^\gamma$ in $L^q(\Omega)$ for all $1 \le q < 2$ and $p = \rho^\gamma$ a.e. in Ω.

It is possible to adapt this proof of convergence when $(\mathbf{u}_n, \rho_n, p_n)$ is the approximate solution given by the MAC scheme as it is done in [10]. As we said before, two main tools are interesting with the MAC scheme:

1. There exists a discrete counterpart of

$$\int_\Omega \mathrm{grad}\,\mathbf{u} : \mathrm{grad}\,\mathbf{v} \, dx = \int_\Omega (\mathrm{div}\,\mathbf{u}\,\mathrm{div}\,\mathbf{v} + \mathrm{curl}\,\mathbf{u} \cdot \mathrm{curl}\,\mathbf{v}) \, dx.$$

2. If w_n, belonging to a discrete equivalent of the $H_0^1(\Omega)$-space, is the solution of $-\Delta_n w_n = \rho_n$ where $-\Delta_n$ is the natural discretization of $-\Delta$ on the primal mesh of the MAC-discretization, then one has an estimate on w_n in the "discrete local H^2-norm" of w_n in term of the L^2-norm of ρ_n.

   If $\gamma < 3$, a new difficulty appears since we have to work with the local L^p-norm of the second dicrete derivatives of w_n for some $p > 2$.

In order to conclude this section, we recall that the convergence of approximate solutions (given by the MAC scheme) if $3/2 < \gamma \le 3$ is, to our knowledge, still an open problem.

3.2. For the compressible Navier-Stokes problem

We consider in the section the compressible Navier-Stokes problem discretized with the MAC scheme and the Euler backward discretization in time, as in Sec. 2.3 (with $T > 0$, $\gamma > 3/2$ and $f \in L^2(]0, T[, L^2(\Omega))$). For $n \in \mathbb{N}$, the approximate solution $(\mathbf{u_n}, \rho_n, p_n)$ is solution of the discretization of Problem (1)-(5). We assume that $\lim_{n\to+\infty} h_n = \lim_{n\to+\infty} k_n = 0$, where h_n and k_n are the mesh size and the time step of the discretization. Our objective is to prove that the approximate solution converges, in an appropriate sense, up to a subsequence, to a weak solution of (1)-(5).

As usual, the first step, for proving such a convergence result, is to obtain estimates on the approximate solution. A quite easy estimate is in $L^\infty(]0,T[,L^\gamma(\Omega))$ for ρ_n and in $L^2(]0,T[,H_n)$ for $\mathbf{u}_n$ where the norm in H_n is a discrete counterpart of the $H_0^1(\Omega)$-norm (this gives also an $L^2(]0,T[,L^6(\Omega))$ estimate on $\mathbf{u}_n$).

Then, in order to pass to the limit in the equations (as $n \to +\infty$), a new difficulty appears (with respect to the stationary case) for passing to the limit on the non linear terms, namely $\rho_n\mathbf{u}_n$ and $\rho_n\mathbf{u}_n\otimes\mathbf{u}_n$. For instance, in the stationary case (Sec. 3.1), we pass to the limit on $\rho_n\mathbf{u}_n$ (up to a subsequence) using the (strong) convergence of $\mathbf{u}_n$ in a Lebesgue space $L^q(\Omega)^3$ for some $q<6$ and the weak convergence of ρ_n in the dual space $L^{q'}(\Omega)$, $q'=q/(q-1)>6/5$. It gives convergence of $\rho_n\mathbf{u}_n$ in $L^1(\Omega)$. This method does not work in the unsteady case since we do not have relative (strong) compactness of the sequence $(\mathbf{u}_n)_n$ in a Lebesgue space. However, we can also conclude in the stationary case by changing the roles of $\mathbf{u}_n$ and ρ_n. Assuming, for simplicity that $(\mathbf{u}_n)_{n\in\mathbb{N}}$ is bounded in $H_0^1(\Omega)^3$, one has, up to subsequence, $u_n \to u$ weakly in $H_0^1(\Omega)^3$, $\rho_n \to \rho$ in $H^{-1}(\Omega)$ (thanks to the compact embedding of $L^{q'}(\Omega)$ in $H^{-1}(\Omega)$) and then, for all $\psi \in C_c^\infty(\mathbb{R})^3$,

$$\int_\Omega \rho_n\mathbf{u}_n\cdot\psi\,dx = \langle\rho_n,\mathbf{u}_n\cdot\psi\rangle_{H^{-1},H_0^1} \to \langle\rho,\mathbf{u}\cdot\psi\rangle_{H^{-1},H_0^1} = \int_\Omega \rho\mathbf{u}\cdot\psi\,dx.$$

For the discrete setting, we also have to replace the $H_0^1(\Omega)$-norm by the so-called discrete-H_0^1-norm (which depends on n), cf. [7] for a complete proof.

The main interest of this new proof for passing to the limit on $\rho_n\mathbf{u}_n$ is that it works also for the unsteady case. Assuming also for simplicity that $(\mathbf{u}_n)_{n\in\mathbb{N}}$ is bounded in $L^2(]0,T[,H_0^1(\Omega)^3)$ (cf. [7] for the discrete case), one has (up to a subsequence) $\mathbf{u}_n \to \mathbf{u}$ weakly in $L^2(]0,T[,H_0^1(\Omega)^3)$. We also know that $(\rho_n)_n$ in $L^2(]0,T[,L^{q'}(\Omega))$ for some $q'>6/5$ and the mass equation (1) (together with the fact that $\mathbf{u}_n$ is bounded in $L^6(\Omega)$) gives that the sequence $(\partial_t\rho_n)_n$ is bounded in $L^2(]0,T[,W^{-1,1}(\Omega))$. Then $(\rho_n)_{n\in\mathbb{N}}$ is relatively compact in $L^2(]0,T[,H^{-1}(\Omega))$ (thanks to Aubin-Lions-Simon compactness results, since $L^{q'}(\Omega)$ is compactly embedded in $H^{-1}(\Omega)$). Then, up to a subsequence $\rho_n \to \rho$ in $L^2(]0,T[,H^{-1}(\Omega))$ and finally, for all $\psi \in C_c^\infty(\mathbb{R}\times\mathbb{R}^3)^3$,

$$\int_0^T\int_\Omega \rho_n\mathbf{u}_n\cdot\psi\,dxdt = \int_0^T \langle\rho_n,\mathbf{u}_n\cdot\psi\rangle_{H^{-1},H_0^1} \to = \int_0^T\int_\Omega \rho\mathbf{u}\cdot\psi\,dxdt.$$

The difficulty is similar for the term $\rho\mathbf{u}\otimes\mathbf{u}$. In Sec. 3.1 we pass to the limit on this term using $\mathbf{u}_n \to \mathbf{u}$ in $L^q(\Omega)^3$ for all $q<6$ and $\rho_n\mathbf{u}_n \to \rho\mathbf{u}$ weakly in $L^{q'}(\Omega)^3$, with some $q'>\frac{6}{5}$. It gives $\rho_n\mathbf{u}_n\otimes\mathbf{u}_n \to \rho\mathbf{u}\otimes\mathbf{u}$ weakly in $L^1(\Omega)^{3\times3}$. But an other method is possible. One can use $\mathbf{u}_n \to \mathbf{u}$ weakly in $H_0^1(\Omega)^3$ and $\rho_n\mathbf{u}_n \to \rho\mathbf{u}$ in $H^{-1}(\Omega)^3$ (thanks to the compact embedding of $L^{q'}(\Omega)$ in $H^{-1}(\Omega)$). It also gives convergence of $\rho_n\mathbf{u}_n\otimes\mathbf{u}_n$ to $\rho\mathbf{u}\otimes\mathbf{u}$, that is, for all $\psi \in C_c^\infty(\mathbb{R})^{3\times3}$, $\int_\Omega \rho_n\mathbf{u}_n\otimes\mathbf{u}_n : \psi\,dx \to \int_\Omega \rho\mathbf{u}\otimes\mathbf{u} : \psi\,dx$. Here also, the generalization of this second method is possible for the unsteady case cf. [7].

This does not conclude the convergence (as $n \to +\infty$, up to a subsequence) of the approximate solution to a weak solution of Problem (1)-(5). It remains to pass to the limit on p_n and on the EOS $p_n = \rho_n^\gamma$. It is an ongoing work.

References

[1] Boccardo, L., Gallouët, T., and Murat, F. Unicité de la solution de certaines équations elliptiques non linéaires. *C. R. Acad. Sci. Paris Sér. I Math. 315*, 11 (1992), 1159–1164.

[2] Dafermos, C. M. The second law of thermodynamics and stability. *Arch. Rational Mech. Anal. 70*, 2 (1979), 167–179. Available from: https://doi.org/10.1007/BF00250353, doi:10.1007/BF00250353.

[3] Eymard, R., Gallouët, T., Herbin, R., and Latché, J.-C. Convergence of the MAC scheme for the compressible Stokes equations. *SIAM J. Numer. Anal. 48*, 6 (2010), 2218–2246. Available from: http://dx.doi.org/10.1137/090779863, doi:10.1137/090779863.

[4] Eymard, R., Gallouët, T., Herbin, R., and Latché, J. C. A convergent finite element-finite volume scheme for the compressible Stokes problem. II. The isentropic case. *Math. Comp. 79*, 270 (2010), 649–675. Available from: http://dx.doi.org/10.1090/S0025-5718-09-02310-2, doi:10.1090/S0025-5718-09-02310-2.

[5] Feireisl, E., Novotný, A., and Petzeltová, H. On the existence of globally defined weak solutions to the Navier-Stokes equations. *J. Math. Fluid Mech. 3*, 4 (2001), 358–392. Available from: https://doi.org/10.1007/PL00000976, doi:10.1007/PL00000976.

[6] Feireisl, E., Novotný, A., and Sun, Y. Suitable weak solutions to the Navier-Stokes equations of compressible viscous fluids. *Indiana Univ. Math. J. 60*, 2 (2011), 611–631. Available from: https://doi.org/10.1512/iumj.2011.60.4406, doi:10.1512/iumj.2011.60.4406.

[7] Gallouët, T. Discrete functional analysis tools for some evolution equations. *Comput. Methods Appl. Math. 18*, 3 (2018), 477–493. Available from: https://doi.org/10.1515/cmam-2017-0059, doi:10.1515/cmam-2017-0059.

[8] Gallouët, T., and Herbin, R. Equations aux dérivées partielles. Lecture, Sept. 2015. Available from: https://hal.archives-ouvertes.fr/cel-01196782.

[9] Gallouët, T., Herbin, R., and Latché, J.-C. A convergent finite element-finite volume scheme for the compressible Stokes problem. I. The isothermal case. *Math. Comp. 78*, 267 (2009), 1333–1352. Available from: https://doi.org/10.1090/S0025-5718-09-02216-9, doi:10.1090/S0025-5718-09-02216-9.

[10] Gallouët, T., Herbin, R., Latché, J.-C., and Maltese, D. Convergence of the MAC scheme for the compressible stationary Navier-Stokes equations. *Math. Comp. 87*, 311 (2018), 1127–1163. Available from: https://doi.org/10.1090/mcom/3260, doi:10.1090/mcom/3260.

[11] Gallouët, T., Herbin, R., Maltese, D., and Novotny, A. Error estimates for a numerical approximation to the compressible barotropic Navier-Stokes equations. *IMA J. Numer. Anal. 36*, 2 (2016), 543–592. Available from: https://doi.org/10.1093/imanum/drv028, doi:10.1093/imanum/drv028.

[12] Gallouët, T., Maltese, D., and Novotny, A. Error estimates for the implicit mac scheme for the compressible navier–stokes equations. *Numerische Mathematik 141*, 2 (Feb 2019), 495–567. Available from: https://doi.org/10.1007/s00211-018-1007-x, doi:10.1007/s00211-018-1007-x.

[13] Germain, P. Weak-strong uniqueness for the isentropic compressible Navier-Stokes system. *J. Math. Fluid Mech. 13*, 1 (2011), 137–146. Available from: https://doi.org/10.1007/s00021-009-0006-1, doi:10.1007/s00021-009-0006-1.

[14] Harlow, F., and Amsden, A. Numerical calculation of almost incompressible flow. *Journal of Computational Physics 3* (1968), 80–93.

[15] Harlow, F., and Amsden, A. A numerical fluid dynamics calculation method for all flow speeds. *Journal of Computational Physics 8* (1971), 197–213.

[16] Harlow, F., and Welsh, J. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Physics of Fluids 8* (1965), 2182–2189.

[17] Karper, T. K. A convergent FEM-DG method for the compressible Navier-Stokes equations. *Numer. Math. 125*, 3 (2013), 441–510. Available from: https://doi.org/10.1007/s00211-013-0543-7, doi:10.1007/s00211-013-0543-7.

[18] Lions, P.-L. *Mathematical topics in fluid mechanics. Vol. 2*, vol. 10 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press, Oxford University Press, New York, 1998. Compressible models, Oxford Science Publications.

[19] Novo, S., and Novotný, A. On the existence of weak solutions to the steady compressible Navier-Stokes equations when the density is not square integrable. *J. Math. Kyoto Univ. 42*, 3 (2002), 531–550.

[20] Novotný, A., and Straškraba, I. *Introduction to the mathematical theory of compressible flow*, vol. 27 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2004.

[21] Prodi, G. Un teorema di unicità per le equazioni di Navier-Stokes. *Ann. Mat. Pura Appl. (4) 48* (1959), 173–182. Available from: https://doi.org/10.1007/BF02410664, doi:10.1007/BF02410664.

[22] Serrin, J. The initial value problem for the Navier-Stokes equations. In *Nonlinear Problems (Proc. Sympos., Madison, Wis., 1962)*. Univ. of Wisconsin Press, Madison, Wis., 1963, pp. 69–98.

[23] Wesseling, P. *Principles of computational fluid dynamics*, vol. 29 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2001. Available from: https://doi.org/10.1007/978-3-642-05146-3, doi:10.1007/978-3-642-05146-3.

T. Gallouët
I2M, CMI,
39 rue F. Joliot Curie
F-13453 Marseille cedex 13
thierry.gallouet@univ-amu.fr

Monografías Matemáticas *García de Galdeano* **42**, 111–125 (2019)

A COLLOCATION METHOD FOR A TWO-POINT BOUNDARY VALUE PROBLEM WITH A RIEMANN-LIOUVILLE-CAPUTO FRACTIONAL DERIVATIVE

José Luis Gracia, Eugene O'Riordan and Martin Stynes

Abstract. Numerical methods for a two-point boundary value problem, where the leading term in the differential operator is a Caputo fractional-order derivative of order $1 < \alpha < 2$, are examined. By reformulating the problem as a Volterra integral equation of the second kind, the problem can be discretized using a collocation method. The performance of this collocation method is compared to a finite difference method applied to the original two-point boundary value problem.

Keywords: Fractional differential equation, Riemann-Liouville-Caputo fractional derivative, two-point boundary value problem, collocation method, weak singularity.

AMS classification: AMS classification codes 34A08, 65L10, 65L60.

§1. Introduction

Judging by the ever-expanding literature on numerical methods for fractional differential equations, this topical area is of interest to many researchers. In numerous publications within this area, the classical numerical method of finite differences has been adapted to deal with the presence of a fractional derivative in the differential equation, which results in the associated system matrix being a relatively dense matrix. This matrix structure has practical implications in terms of the accumulation of rounding errors and in storage issues for problems in higher dimensions, Moreover, the associated numerical analysis of finite difference methods for fractional differential equations can be difficult.

In this paper, we examine an alternative approach to discretizing the following two-point boundary value problem

$$-D^{\alpha}_{RLC}u(x) + b(x)u'(x) + c(x)u(x) = f(x) \ \text{ for } x \in (0, L), \tag{1a}$$

$$D^{\alpha-1}_{C}u(0) = 0, \ u(L) + \beta_1 u'(L) = \gamma_1. \tag{1b}$$

The leading term in the differential operator is a fractional-order derivative of order α, $\alpha \in (1, 2)$, which is called a Riemann-Liouville-Caputo [7], Patie-Simon [1, 8, 12] or conservative Caputo derivative [15]. It is defined by

$$D^{\alpha}_{RLC}u(x) := \frac{d}{dx} D^{\alpha-1}_{C}u(x) \text{ for } x > 0,$$

where D_C^β denotes the Caputo fractional derivative of order β (see for example [4]) with $n-1 < \beta < n$ and n is a positive integer; that is,

$$D_C^\beta v(x) := \frac{1}{\Gamma(n-\beta)} \int_{t=0}^{x} (x-t)^{n-1-\beta}\, v^{(n)}(t)\, dt, \quad \text{where } v^{(n)}(t) := \frac{d^n v(t)}{dt^n}. \tag{2}$$

The constants $\beta_1 \geq 0$ and γ_1 and the functions b, c, f are given and it is assumed that

$$c(x) \geq 0 \text{ for } x \in [0, L].$$

The motivation for using the fractional derivative D_{RLC}^α in problem (1) instead of the more commonly used Riemann-Liouville or Caputo fractional derivatives comes from recent publications modelling physical processes [1, 3, 5, 12]. The use of the Caputo fractional Neumann boundary condition $D_C^{\alpha-1}u(0) = 0$ in combination with the fractional derivative D_{RLC}^α is suggested in [3]. In [6] it is proved that, in the case of problem (1), $D_C^{\alpha-1}u(0) = 0$ is equivalent to $u'(0) = 0$. We shall consider this commonly used boundary condition in the present paper.

The solution of problem (1) has a weak singularity at $x = 0$ so its numerical approximation is troublesome. To deal with it, the problem (1) is first reformulated as a Volterra integral equation of the second kind, which is then discretized using a collocation method on a graded mesh. Relative to the finite difference method, this approach is easier to implement and, moreover, the associated numerical analysis is more natural for such problems involving fractional derivatives. The numerical analysis follows the classical approach for collocation methods for Volterra integral equations [2].

This reformulation for fractional-derivative problems was first presented in [9] for certain types of boundary conditions. The convergence result established in [9] is a significant improvement on the corresponding convergence result in [14] where a finite difference scheme was considered. Reformulation was also applied successfully in [10] to a two-point boundary value problem where the highest-order derivative is of Riemann-Liouville type.

In this current paper, we demonstrate that both the method and analysis of [9] extend easily to problem (1), which was not covered in [9]. In [6], we examined the same problem using a finite difference method; there, to prove first-order convergence, we needed to impose the constraint $b \leq 0$ on the data. In the present paper, using the collocation approach, we derive a convergence result without imposing this constraint on the sign of b.

The paper is structured as follows: In Section 2 the two-point boundary value problem (1) is first shown to be equivalent to another boundary value problem whose highest-order derivative is of Caputo type and whose boundary condition at $x = 0$ is $u'(0) = 0$. This new problem is reformulated as a Volterra integral equation of the second kind. In Section 3 the collocation method for this integral equation is presented on a graded mesh condensing at the endpoint $x = 0$. Error estimates are obtained showing the convergence of the collocation method and the dependence of the order of convergence on the choice of the collocation points and on the grading exponent of the mesh. In Section 4 two examples are used in order to compare our collocation method and the finite difference scheme of [6]. They illustrate that the collocation method is more efficient with both a lower computational cost and a higher order of convergence.

Notation: In this paper C denotes a generic constant that can depend on the data of the boundary value problem (1) and possibly on the mesh grading but is independent of the mesh

diameter. Note that C can take different values in different places. For each $g \in C[0,1]$, set $\|g\|_\infty = \max_{0\le x\le L} |g(x)|$.

§2. Reformulations of the problem

In [6] it is proved that problem (1) is equivalent to the Caputo two-point boundary value problem

$$-D_C^\alpha u(x) + b(x)u'(x) + c(x)u(x) = f(x) \quad \text{for } x \in (0, L), \tag{3a}$$

$$u'(0) = 0, \; u(L) + \beta_1 u'(L) = \gamma_1. \tag{3b}$$

This follows because

(i) the condition $D_C^{\alpha-1}u(0) = 0$ implies $u'(0) = 0$ (which is proved in [6]);

(ii) using integration by parts, one can deriver the relationship

$$D_C^\alpha u(x) = D_{RLC}^\alpha u(x) - \frac{x^{1-\alpha}}{\Gamma(2-\alpha)} u'(0). \tag{4}$$

between the Caputo and Riemann-Liouville-Caputo fractional derivatives, provided that they exist;

(iii) in [4, Lemma 3.11] it is proved that $D_C^{\alpha-1}u(0) = 0$ if u' is absolutely continuous.

Hence, we shall approximate problem (3). Before considering any numerical method for its numerical approximation, some information about the behaviour of the solution is required. Assuming appropriate regularity conditions on the data problem, in [6] it is proved that the solution of (3) satisfies the bounds

$$|u^{(i)}(x)| \le \begin{cases} C & \text{if } i = 0, \\ Cx^{\alpha-i} & \text{if } i = 1, 2, 3, \ldots, \end{cases} \tag{5}$$

showing that a typical solution u of problem (1) has a weak singularity at $x = 0$.

In [6], the standard L2 discretization on a uniform mesh is used to approximate (3). In the present paper, similarly to [9], a collocation method is used instead. To this end we reformulate (3) as a Volterra integral equation of the second kind. Recall the definition of the Riemann-Liouville fractional integral operator of order r, which is

$$(J^r g)(x) := \frac{1}{\Gamma(r)} \int_{t=0}^{x} (x-t)^{r-1} g(t)\, dt. \tag{6}$$

Applying $J^{\alpha-1}$ to (3a) and using the fact that

$$J^{\alpha-1} D_C^\alpha g(x) = J^{\alpha-1} J^{2-\alpha} g''(x) = J g''(x) = g'(x) - g'(0),$$

and $u'(0) = 0$, one has

$$-u'(x) + J^{\alpha-1}(b(x)u'(x) + c(x)u(x)) = J^{\alpha-1} f(x).$$

Noting that $u(x) = \int_{s=0}^{x} u'(s)\,ds + u(0)$ and denoting $y(x) = u'(x)$ and $Y(x) = \int_{s=0}^{x} y(s)\,ds$, then (3a) is rewritten as

$$y(x) - J^{\alpha-1}(b(x)y(x) + c(x)Y(x)) = -J^{\alpha-1}f(x) + u(0)J^{\alpha-1}c(x). \tag{7}$$

Consider the decomposition

$$y(x) := v(x) + u(0)w(x),$$

where $v(x)$ is the solution of

$$v(x) - J^{\alpha-1}(b(x)v(x) + c(x)V(x)) = -J^{\alpha-1}f(x), \quad \text{for } x \in (0, L], \quad v(0) = 0, \tag{8}$$

with $V(x) = \int_{s=0}^{x} v(s)\,ds$, and $w(x)$ is the solution of

$$w(x) - J^{\alpha-1}(b(x)w(x) + c(x)W(x)) = J^{\alpha-1}c(x), \quad \text{for } x \in (0, L], \quad w(0) = 0, \tag{9}$$

with $W(x) = \int_{s=0}^{x} w(s)\,ds$. Hence, both v and w satisfy weakly singular Volterra integral equations of the second kind. From [9, Lemma 2.1], the problems (8) and (9) each have a unique solution.

Once v and w are obtained, then we calculate $u(0)$ from

$$u(x) = \int_{s=0}^{x} u'(s)\,ds + u(0) = \int_{s=0}^{x} v(s)\,ds + u(0)\left(1 + \int_{s=0}^{x} w(s)\,ds\right) \tag{10}$$

and imposing the boundary condition (1b) at $x = L$:

$$\begin{aligned}\int_{s=0}^{L} v(s)\,ds + u(0)\left(1 + \int_{s=0}^{L} w(s)\,ds\right) &= u(L)\\ &= \gamma_1 - \beta_1 u'(L)\\ &= \gamma_1 - \beta_1\left[v(L) + u(0)w(L)\right];\end{aligned}$$

Thus, one has

$$u(0) = \frac{\gamma_1 - \beta_1 v(L) - \int_{s=0}^{L} v(s)\,ds}{1 + \beta_1 w(L) + \int_{s=0}^{L} w(s)\,ds}. \tag{11}$$

The denominator in (11) must not be zero, i.e.,

$$1 + \beta_1 w(L) + \int_{s=0}^{L} w(s)\,ds \neq 0. \tag{12}$$

Since $c(x) \geq 0$, a proof by contradiction argument, as in the proof of [9, Lemma 4.1], can be used to prove that $w(x) \geq 0$ for $x \in [0, L]$. Hence (12) is satisfied. Moreover, from [9, Lemma 2.1], we have $\|v\|_\infty \leq C$ and $\|w\|_\infty \leq C$. Thus $|u(0)| \leq C$ from (11).

§3. The collocation method

Consider the problem

$$z(x) - J^{\alpha-1}(b(x)z(x) + c(x)Z(x)) = J^{\alpha-1}g(x), \quad \text{for } x \in (0, L], \quad z(0) = 0 \tag{13}$$

with $Z(x) := \int_{s=0}^{x} z(s)\,ds$. Observe that if $g(x) = -f(x)$ or $g(x) = c(x)$, then we have the problems (8) and (9) associated with the components v and w of u. From the definition (6), one can write (13) as: Find z such that

$$\begin{aligned} z(x) - \frac{1}{\Gamma(\alpha-1)} \int_{t=0}^{x} (x-t)^{\alpha-2} \left[b(t)z(t) + c(t) \int_{s=0}^{t} z(s)\,ds \right] dt \\ = \frac{1}{\Gamma(\alpha-1)} \int_{t=0}^{x} (x-t)^{\alpha-2} g(t)\,dt, \quad \text{for } x \in (0, L], \end{aligned} \tag{14}$$

which will be approximated using the collocation method.

Let N be a positive integer. Consider the graded mesh

$$x_i = L(i/N)^r \quad \text{for } i = 0, 1, \ldots, N, \quad h_i = x_{i+1} - x_i, \quad \text{for } i = 0, 1, \ldots, N-1, \tag{15}$$

where $r \geq 1$ is the grading exponent. If $r = 1$ the mesh is uniform, while the larger r is, the more the grid condenses near $x = 0$. We set $h = \max_{0 \leq i \leq N-1} h_i$ and $h_N = 0$.

The computed solution $z_h \in S_{m-1}^{-1}$, where

$$S_{m-1}^{-1} := \{v : \, v|_{(x_i, x_{i+1})} \in \pi_{m-1}, \; i = 0, 1, \ldots, N-1\}$$

and π_{m-1} denotes the space of polynomials of degree at most $m-1$. Thus, the elements of S_{m-1}^{-1} are piecewise polynomials of degree at most $m-1$ that may be discontinuous at the points x_i. The set of collocation points is

$$X_h = \{x_i + c_j h_i : \, 0 \leq c_1 < c_2 < \cdots < c_m \leq L, \; i = 0, 1, \ldots, N-1\},$$

where $\{c_i\}$ are chosen by the user. If $c_1 = 0$ and $c_m = 1$, then $z_h \in S_{m-1}^{-1} \cap C[0, L]$. The collocation solution $z_h \in S_{m-1}^{-1}$ is computed by imposing

$$\begin{aligned} z_h(x) - \frac{1}{\Gamma(\alpha-1)} \int_{t=0}^{x} (x-t)^{\alpha-2} \left[b(t)z_h(t) + c(t) \int_{s=0}^{t} z_h(s)\,ds \right] dt \\ = \frac{1}{\Gamma(\alpha-1)} \int_{t=0}^{x} (x-t)^{\alpha-2} g(t)\,dt \quad \text{for all } x \in X_h \cup \{L\}. \end{aligned} \tag{16}$$

Note that the collocation method solves mesh interval by mesh interval; thus on each interval one solves a system of m equations (or $m-1$ equations if $c_1 = 0$ and $c_m = 1$) where the unknowns are located at the collocation points. Therefore, collocation methods are more efficient than finite difference methods where one has to solve a single large linear system (see [6] for a comparison).

In practice, the integrals in (16) are evaluated using quadrature formulas with the collocation points as nodes and the functions b, c and g are replaced by polynomials of degree $m-1$ that interpolate to these functions at the collocation points. The computed solution is denoted by $\hat{z}_h$ and satisfies the following result.

Lemma 1. *Assume that $b, c, g \in C^m[0,1]$. Then the collocation solution $\hat{z}_h$ satisfies*

$$\max_{0\le i\le N} |(\hat{z}_h - z)(x_i)| \le Ch^{\min\{r(\alpha-1),m\}}.$$

If in addition the collocation points $\{c_j\}$ are such that

$$\int_{s=0}^{1} \prod_{j=1}^{m} (s - c_j)\, ds = 0, \tag{17}$$

and $b, c, g \in C^{m+1}[0,1]$, then if $r(\alpha-1) \ge m$,

$$\max_{0\le i\le N} \max_{1\le j\le m} |(\hat{z}_h - z)(x_i + c_j h_i)| \le Ch^{m+\alpha-1}.$$

Proof. See [2, Theorem 6.2.14] and [9, Corollary 3.1 and Corollary 3.2]. □

Using the quadrature formulas, one computes the approximations $\hat{v}_h$ and $\hat{w}_h$ to v_h and w_h, respectively. Assuming that h is sufficiently small so that

$$1 + \beta_1 \hat{w}_h(L) + \int_{s=0}^{L} \hat{w}_h(s)\, ds \ne 0,$$

by imitating (10) and (11) one constructs the following approximations of u and u':

$$\hat{u}_h(x) = \int_{s=0}^{x} \hat{v}_h(s)\, ds + \hat{u}_h(0)\left(1 + \int_{s=0}^{x} \hat{w}_h(s)\, ds\right), \tag{18}$$

$$\hat{u}_h'(x) = \hat{v}_h(x) + \hat{u}_h(0)\hat{w}_h(x), \tag{19}$$

with

$$\hat{u}_h(0) = \frac{\gamma_1 - \beta_1 \hat{v}_h(L) - \int_{s=0}^{L} \hat{v}_h(s)\, ds}{1 + \beta_1 \hat{w}_h(L) + \int_{s=0}^{L} \hat{w}_h(s)\, ds}.$$

Recalling (11), we see that

$$|(u - \hat{u}_h)(0)| \le C\left(|(v - \hat{v}_h)(L)| + \int_{s=0}^{L} |(v - \hat{v}_h)(s)|\, ds + |(w - \hat{w}_h)(L)| + \int_{s=0}^{L} |(w - \hat{w}_h)(s)|\, ds\right)$$

and from (10) and (18) we have

$$(u - \hat{u}_h)(x) = \int_{s=0}^{x} (v - \hat{v}_h)(s)\, ds + \hat{u}_h(0) \int_{s=0}^{x} (w - \hat{w}_h)(s)\, ds + (u - \hat{u}_h)(0)\left(1 + \int_{s=0}^{x} w(s)\, ds\right).$$

Using Lemma 1 to bound $\|v - \hat{v}_h\|_\infty$ and $\|w - \hat{w}_h\|_\infty$ (as in [9, Theorem 4.2]), one can establish the following error bound for the collocation method (16).

Theorem 2. *Assume that $b, c, f \in C^m[0,1]$. Let h be sufficiently small. Then the collocation solution u_h of* (1) *and its equivalent problem* (1)*, when product quadrature with collocation points as nodes is used, satisfies the error bound*

$$\max_{0\le i\le N} |(\hat{u}_h - u)(x_i)| + \max_{0\le i\le N} |(\hat{u}_h' - u')(x_i)| \le Ch^{\min\{r(\alpha-1),m\}}, \tag{20}$$

where r is the mesh grading exponent (15). *If in addition* (17) *is satisfied and $b, c, f \in C^{m+1}[0,1]$, then for $r(\alpha-1) \geq m$ one obtains*

$$\max_{0\leq i\leq N} |(\hat{u}_h - u)(x_i)| + \max_{0\leq i\leq N} \max_{1\leq j\leq m} |(\hat{u}_h' - u')(x_i + c_j h_i)| \leq Ch^{m+\alpha-1}. \tag{21}$$

Remark 1. The error estimate of Theorem 2 for our collocation method does not place any constraint on the sign of b. In contrast, the convergence analysis for the finite difference method in [6] is valid only when $b \leq 0$; its analysis without the restriction $b \leq 0$ is still open.

§4. Numerical experiments

Numerical results are given in this section for two examples with b and f constants and $c \equiv 0$. In the first example $b < 0$ while $b > 0$ in the second example. The exact solution of both examples can be obtained using Laplace transforms. The maximum error and maximum derivative errors in the computed solution $\{\hat{u}_h\}$ are denoted by

$$E_N := \max_{0\leq i\leq N} |(\hat{u}_h - u)(x_i)|, \quad D_N := \max_{0\leq i\leq N} \max_{1\leq j\leq m} |(\hat{u}_h' - u')(x_i + c_j h_j)|.$$

Note that the errors in the approximate solutions are computed at the mesh points and the errors in the approximate first derivative of the solution are computed at the collocation points. The orders of convergence are computed from these values in a standard way:

$$p_N := \log_2\left(\frac{E_N}{E_{2N}}\right), \quad q_N := \log_2\left(\frac{D_N}{D_{2N}}\right).$$

The solutions of both examples are approximated using our collocation method and the finite difference scheme considered in [6]. In the former, for the sake of brevity, we only consider a specific collocation method with $m = 1$ and $c_1 = 1/2$ using both a uniform and a graded mesh with $r = 1/(\alpha-1)$. The collocation point $c_1 = 1/2$ is special because (17) is satisfied, i.e.,

$$\int_{s=0}^{1} \left(s - \frac{1}{2}\right) ds = 0,$$

and, as a result, the collocation method with $c_1 = 1/2$ provides more accurate approximations than for $c_1 \neq 1/2$. Theorem 2 tells us that the optimal graded mesh is obtained when $r = m/(\alpha-1)$, as this then gives the highest possible rate of convergence $O(h^{m+\alpha-1})$ — any larger value of r would not improve the rate of convergence but would increase the mesh width near $x = L$ and consequently increase the constant multiplier C in the error bound. As we use $m = 1$ in our experiments, we choose $r = 1/(\alpha-1)$ to get the optimal mesh grading.

A numerical approximation $\hat{z}_h(x)$ (for all $x \in X_h \cup \{L\}$) to the solution of (16) is computed and then the maximum derivative errors D_N can be computed. To approximate the nodal values $u(x_i) = \int_{s=0}^{x_i} y(s)\, ds + u(0)$, all integrals of the form $\int_{s=0}^{x_i} \hat{z}_h(s)\, ds$ are approximated using the composite trapezoidal rule, i.e.,

$$\int_{s=0}^{x_i} \hat{z}_h(s)\, ds \approx \frac{x_1 - x_0}{2} \frac{\hat{z}_h(h_0/2)}{2} + \sum_{l=1}^{i-1} \frac{x_{l+1} - x_{l-1}}{2} \frac{\hat{z}_h(x_{l-1} + h_{l-1}/2) + \hat{z}_h(x_l + h_l/2)}{2}$$
$$+ \frac{x_i - x_{i-1}}{2} \frac{\hat{z}_h(x_{i-1} + h_{i-1}/2) + z_h(x_i + h_i/2)}{2}, \quad 0 < i \leq N,$$

with $\hat{z}_h(x_N + h_N/2) = \hat{z}_h(x_N)$. The maximum errors E_N can then be computed.

The finite difference scheme [6] is defined on a uniform mesh. It is given by

$$\left(-D^{\alpha}_{C,L2}u_h + bD^0u_h + cu_h\right)(jL/N) = f(jL/N) \text{ for } j = 1, 2, \ldots, N-1,$$
$$-D^+u_h(0) = 0, \quad u_h(L) + \beta_1 D^-u_h(L) = \gamma_1,$$

where $D^{\alpha}_{C,L2}$ is the well-known L2 approximation [11] of the Caputo fractional derivative D^{α}_C and D^0, D^- and D^+ are the standard central, backward and forward differences, respectively.

Example I

Consider the problem

$$-D^{\alpha}_{RLC}u - 0.5u' = 1 \text{ on } (0,1), \quad D^{\alpha-1}u(0) = 0,\ u(1) = 0. \tag{22}$$

Its exact solution can be obtained in closed form using Laplace transforms (see [6]):

$$u(x) = -x^{\alpha}E_{\alpha-1,\alpha+1}(-0.5x^{\alpha-1}) + E_{\alpha-1,\alpha+1}(-0.5) \text{ for } 0 \le x \le 1,$$

where $E_{\beta,\gamma}(\cdot)$ is the two-parameter Mittag-Leffler function defined by

$$E_{\beta,\gamma}(z) := \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(k\beta + \gamma)} \text{ for } \beta, \gamma > 0 \text{ and all real numbers } z.$$

In order to compute the errors, the Mittag-Leffler function $E_{\beta,\gamma}(z)$ is evaluated in our code using the function *mlf* provided at MatLab Central File exchange [13].

The maximum nodal errors and orders of convergence of the finite difference scheme proposed in [6] are given in Table 1. The computed orders of convergence indicate that this method is first-order convergent, in agreement with the convergence result proved in [6].

The solution of Example I is now approximated by the collocation method (16) for $m = 1$ and $c_1 = 1/2$. Numerical results using uniform and graded meshes are given in Tables 2 and 3. We see that the collocation method is more accurate on the graded mesh than on the uniform mesh and that both approaches are more accurate than the finite difference scheme [6] (see Table 1) for all the values of α. The order of convergence of the collocation method on the graded mesh is predicted by (21). Note that this theoretical error bound has not been established in the case of a uniform mesh (as (21) requires $r(\alpha - 1) \ge m$). Nevertheless, superconvergence is still observed in Table 2, when a uniform mesh is used. Unlike [6], the collocation theory also gives error estimates for numerical approximations $\hat{u}'_h$ to u'. In Tables 4 and 5 the maximum derivative errors for the collocation method are given using a uniform and a graded mesh. If N is sufficiently large, the approximation of u' is again more accurate when a graded mesh is used and the computed orders of convergence are in agreement with Theorem 2.

Example II

Consider the following problem

$$-D^{\alpha}_{RLC}u + 0.5u' = 1 \text{ on } (0,1), \quad D^{\alpha-1}u(0) = 0,\ u(1) = 0, \tag{23}$$

Table 1: Example I: Maximum nodal errors and orders of convergence using the finite difference scheme [6] on a uniform mesh

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	1.789E-02	8.946E-03	4.473E-03	2.237E-03	1.118E-03	5.592E-04	2.796E-04	1.398E-04
	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
$\alpha = 1.2$	1.840E-02	9.207E-03	4.605E-03	2.303E-03	1.152E-03	5.759E-04	2.880E-04	1.440E-04
	0.999	0.999	1.000	1.000	1.000	1.000	1.000	
$\alpha = 1.3$	1.888E-02	9.457E-03	4.734E-03	2.368E-03	1.185E-03	5.925E-04	2.963E-04	1.482E-04
	0.998	0.998	0.999	0.999	1.000	1.000	1.000	
$\alpha = 1.4$	1.929E-02	9.681E-03	4.852E-03	2.430E-03	1.216E-03	6.085E-04	3.044E-04	1.522E-04
	0.995	0.997	0.998	0.998	0.999	0.999	1.000	
$\alpha = 1.5$	1.956E-02	9.848E-03	4.948E-03	2.482E-03	1.244E-03	6.231E-04	3.119E-04	1.561E-04
	0.990	0.993	0.995	0.997	0.998	0.998	0.999	
$\alpha = 1.6$	1.958E-02	9.902E-03	4.994E-03	2.513E-03	1.263E-03	6.338E-04	3.178E-04	1.592E-04
	0.983	0.987	0.991	0.993	0.995	0.996	0.997	
$\alpha = 1.7$	1.915E-02	9.748E-03	4.943E-03	2.500E-03	1.261E-03	6.353E-04	3.195E-04	1.605E-04
	0.974	0.980	0.984	0.987	0.989	0.991	0.993	
$\alpha = 1.8$	1.804E-02	9.236E-03	4.711E-03	2.396E-03	1.216E-03	6.156E-04	3.111E-04	1.570E-04
	0.966	0.971	0.975	0.979	0.982	0.984	0.987	
$\alpha = 1.9$	1.590E-02	8.142E-03	4.160E-03	2.122E-03	1.080E-03	5.493E-04	2.789E-04	1.414E-04
	0.966	0.969	0.971	0.974	0.976	0.978	0.980	

whose exact solution is (see [6])

$$u(x) = -x^{\alpha} E_{\alpha-1,\alpha+1}(0.5x^{\alpha-1}) + E_{\alpha-1,\alpha+1}(0.5).$$

The error analysis in [6] does not apply to this example because $b > 0$, nevertheless the numerical results from Table 6 show that the finite difference method proposed in that paper on a uniform mesh also converges with first order to the solution u. On the other hand, the error estimates for our collocation method remain valid for this example; the numerical results given in Tables 7 and 8 show that the collocation method converges with order $O(h^{\alpha})$ using either a uniform or a graded mesh. Observe also that the maximum errors for the collocation method on both meshes are similar in this example but smaller than the finite difference errors in Table 6.

Finally, it is shown in Tables 9 and 10 that the numerical approximations $\hat{u}_h'$ generated by the collocation method on a uniform and a graded mesh also converge to u'. Conclusions similar to Example I are reached.

Acknowledgements

The research of José Luis Gracia was partly supported by the Institute of Mathematics and Applications (IUMA), the project MTM2016-75139-R and the Diputación General de Aragón (E24-17R). The research of Martin Stynes was supported in part by the National Natural Science Foundation of China under grant NSAF-U1530401.

References

[1] Baeumer, B., Kovács, M., Meerschaert, M. M., and Sankaranarayanan, H. Boundary conditions for fractional diffusion. *J. Comput. Appl. Math. 336* (2018), 408–424. Available from: https://doi.org/10.1016/j.cam.2017.12.053.

Table 2: Example I: Maximum errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a uniform mesh

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	3.741E-03	1.779E-03	8.451E-04	4.010E-04	1.901E-04	9.001E-05	4.259E-05	2.014E-05
	1.072	1.074	1.076	1.077	1.078	1.080	1.081	
$\alpha = 1.2$	2.670E-03	1.197E-03	5.345E-04	2.380E-04	1.056E-04	4.679E-05	2.068E-05	9.125E-06
	1.157	1.163	1.168	1.171	1.175	1.178	1.181	
$\alpha = 1.3$	1.877E-03	7.875E-04	3.283E-04	1.362E-04	5.628E-05	2.318E-05	9.525E-06	3.905E-06
	1.253	1.262	1.269	1.275	1.280	1.283	1.286	
$\alpha = 1.4$	1.296E-03	5.066E-04	1.965E-04	7.577E-05	2.910E-05	1.114E-05	4.251E-06	1.620E-06
	1.355	1.366	1.375	1.381	1.386	1.389	1.392	
$\alpha = 1.5$	8.770E-04	3.191E-04	1.151E-04	4.126E-05	1.473E-05	5.245E-06	1.864E-06	6.612E-07
	1.459	1.471	1.480	1.486	1.490	1.493	1.495	
$\alpha = 1.6$	6.289E-04	2.185E-04	7.460E-05	2.518E-05	8.434E-06	2.811E-06	9.337E-07	3.095E-07
	1.525	1.550	1.567	1.578	1.585	1.590	1.593	
$\alpha = 1.7$	5.005E-04	1.607E-04	5.082E-05	1.592E-05	4.957E-06	1.538E-06	4.759E-07	1.470E-07
	1.639	1.661	1.675	1.683	1.689	1.692	1.694	
$\alpha = 1.8$	3.795E-04	1.128E-04	3.311E-05	9.647E-06	2.798E-06	8.095E-07	2.338E-07	6.743E-08
	1.750	1.768	1.779	1.785	1.789	1.792	1.794	
$\alpha = 1.9$	2.749E-04	7.565E-05	2.062E-05	5.589E-06	1.510E-06	4.075E-07	1.098E-07	2.958E-08
	1.861	1.876	1.883	1.888	1.890	1.892	1.893	

Table 3: Example I: Maximum errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a graded mesh with $r = 1/(\alpha - 1)$

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	3.725E-04	1.249E-04	4.500E-05	1.735E-05	7.108E-06	3.050E-06	1.352E-06	6.122E-07
	1.576	1.473	1.375	1.288	1.220	1.173	1.143	
$\alpha = 1.2$	3.879E-04	1.320E-04	4.719E-05	1.777E-05	6.994E-06	2.850E-06	1.190E-06	5.047E-07
	1.556	1.483	1.409	1.345	1.295	1.260	1.237	
$\alpha = 1.3$	3.770E-04	1.219E-04	4.105E-05	1.444E-05	5.283E-06	1.996E-06	7.720E-07	3.037E-07
	1.629	1.570	1.507	1.451	1.405	1.370	1.346	
$\alpha = 1.4$	3.471E-04	1.055E-04	3.310E-05	1.076E-05	3.617E-06	1.254E-06	4.457E-07	1.615E-07
	1.717	1.673	1.622	1.572	1.528	1.492	1.465	
$\alpha = 1.5$	3.122E-04	8.958E-05	2.622E-05	7.867E-06	2.426E-06	7.678E-07	2.489E-07	8.233E-08
	1.801	1.773	1.736	1.697	1.660	1.625	1.596	
$\alpha = 1.6$	2.793E-04	7.639E-05	2.108E-05	5.902E-06	1.681E-06	4.873E-07	1.439E-07	4.327E-08
	1.870	1.857	1.837	1.812	1.786	1.759	1.734	
$\alpha = 1.7$	2.506E-04	6.622E-05	1.751E-05	4.654E-06	1.246E-06	3.367E-07	9.188E-08	2.533E-08
	1.920	1.919	1.912	1.901	1.888	1.874	1.859	
$\alpha = 1.8$	2.263E-04	5.847E-05	1.505E-05	3.874E-06	9.985E-07	2.580E-07	6.686E-08	1.739E-08
	1.952	1.958	1.958	1.956	1.952	1.948	1.943	
$\alpha = 1.9$	2.048E-04	5.242E-05	1.329E-05	3.364E-06	8.506E-07	2.150E-07	5.437E-08	1.375E-08
	1.966	1.979	1.982	1.984	1.984	1.984	1.983	

Table 4: Example I: Maximum derivative errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a uniform mesh

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	1.460E-03	1.337E-03	1.221E-03	1.112E-03	1.010E-03	9.156E-04	8.281E-04	7.474E-04
	0.128	0.131	0.135	0.138	0.142	0.145	0.148	
$\alpha = 1.2$	3.141E-03	2.551E-03	2.056E-03	1.645E-03	1.308E-03	1.035E-03	8.142E-04	6.378E-04
	0.300	0.311	0.321	0.330	0.339	0.346	0.352	
$\alpha = 1.3$	3.452E-03	2.437E-03	1.701E-03	1.176E-03	8.058E-04	5.486E-04	3.714E-04	2.502E-04
	0.502	0.519	0.533	0.545	0.555	0.563	0.570	
$\alpha = 1.4$	2.767E-03	1.683E-03	1.010E-03	6.000E-04	3.537E-04	2.072E-04	1.208E-04	7.017E-05
	0.717	0.736	0.751	0.763	0.772	0.778	0.783	
$\alpha = 1.5$	1.832E-03	9.572E-04	4.939E-04	2.526E-04	1.283E-04	6.490E-05	3.271E-05	1.645E-05
	0.936	0.955	0.968	0.977	0.984	0.988	0.992	
$\alpha = 1.6$	1.067E-03	4.795E-04	2.132E-04	9.411E-05	4.135E-05	1.811E-05	7.915E-06	3.454E-06
	1.154	1.169	1.180	1.187	1.191	1.194	1.196	
$\alpha = 1.7$	5.672E-04	2.198E-04	8.443E-05	3.227E-05	1.229E-05	4.673E-06	1.774E-06	6.731E-07
	1.368	1.380	1.388	1.392	1.395	1.397	1.398	
$\alpha = 1.8$	2.819E-04	9.439E-05	3.141E-05	1.041E-05	3.445E-06	1.138E-06	3.758E-07	1.241E-07
	1.578	1.587	1.593	1.596	1.598	1.599	1.599	
$\alpha = 1.9$	1.330E-04	3.858E-05	1.114E-05	3.208E-06	9.227E-07	2.652E-07	7.619E-08	2.188E-08
	1.786	1.792	1.796	1.798	1.799	1.799	1.800	

Table 5: Example I: Maximum derivative errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a graded mesh with $r = 1/(\alpha - 1)$

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	2.595E-03	1.264E-03	6.025E-04	2.842E-04	1.333E-04	6.240E-05	2.916E-05	1.361E-05
	1.038	1.069	1.084	1.092	1.096	1.098	1.099	
$\alpha = 1.2$	1.897E-03	8.453E-04	3.727E-04	1.634E-04	7.139E-05	3.114E-05	1.357E-05	5.912E-06
	1.166	1.182	1.190	1.194	1.197	1.198	1.199	
$\alpha = 1.3$	1.267E-03	5.246E-04	2.155E-04	8.809E-05	3.592E-05	1.462E-05	5.948E-06	2.418E-06
	1.272	1.284	1.290	1.294	1.296	1.298	1.299	
$\alpha = 1.4$	8.025E-04	3.100E-04	1.189E-04	4.543E-05	1.730E-05	6.579E-06	2.499E-06	9.482E-07
	1.372	1.382	1.388	1.393	1.395	1.397	1.398	
$\alpha = 1.5$	4.858E-04	1.756E-04	6.306E-05	2.254E-05	8.028E-06	2.853E-06	1.013E-06	3.590E-07
	1.468	1.478	1.484	1.489	1.492	1.495	1.496	
$\alpha = 1.6$	2.804E-04	9.509E-05	3.203E-05	1.073E-05	3.582E-06	1.192E-06	3.958E-07	1.312E-07
	1.560	1.570	1.577	1.583	1.587	1.590	1.593	
$\alpha = 1.7$	1.524E-04	4.857E-05	1.538E-05	4.842E-06	1.518E-06	4.742E-07	1.477E-07	4.592E-08
	1.650	1.659	1.667	1.673	1.678	1.683	1.686	
$\alpha = 1.8$	9.695E-05	2.619E-05	7.067E-06	2.002E-06	5.908E-07	1.737E-07	5.091E-08	1.488E-08
	1.888	1.890	1.819	1.761	1.766	1.771	1.774	
$\alpha = 1.9$	7.537E-05	1.966E-05	5.117E-06	1.330E-06	3.457E-07	8.979E-08	2.332E-08	6.056E-09
	1.939	1.942	1.943	1.944	1.945	1.945	1.945	

Table 6: Example II: Maximum errors and orders of convergence using a finite difference scheme on a uniform mesh

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	1.006E-01	5.061E-02	2.539E-02	1.271E-02	6.362E-03	3.182E-03	1.592E-03	7.959E-04
	0.991	0.995	0.998	0.999	0.999	1.000	1.000	
$\alpha = 1.2$	9.787E-02	4.953E-02	2.492E-02	1.250E-02	6.262E-03	3.134E-03	1.568E-03	7.841E-04
	0.982	0.991	0.995	0.998	0.999	0.999	1.000	
$\alpha = 1.3$	9.051E-02	4.605E-02	2.324E-02	1.168E-02	5.859E-03	2.934E-03	1.468E-03	7.346E-04
	0.975	0.986	0.992	0.996	0.998	0.999	0.999	
$\alpha = 1.4$	8.136E-02	4.162E-02	2.109E-02	1.063E-02	5.339E-03	2.677E-03	1.341E-03	6.713E-04
	0.967	0.981	0.989	0.993	0.996	0.997	0.998	
$\alpha = 1.5$	7.170E-02	3.691E-02	1.880E-02	9.508E-03	4.791E-03	2.407E-03	1.208E-03	6.053E-04
	0.958	0.974	0.983	0.989	0.993	0.995	0.997	
$\alpha = 1.6$	6.197E-02	3.213E-02	1.646E-02	8.373E-03	4.237E-03	2.137E-03	1.075E-03	5.403E-04
	0.948	0.965	0.975	0.983	0.987	0.991	0.993	
$\alpha = 1.7$	5.219E-02	2.725E-02	1.407E-02	7.204E-03	3.669E-03	1.861E-03	9.410E-04	4.747E-04
	0.938	0.954	0.965	0.973	0.979	0.984	0.987	
$\alpha = 1.8$	4.219E-02	2.213E-02	1.150E-02	5.931E-03	3.043E-03	1.555E-03	7.916E-04	4.020E-04
	0.931	0.945	0.955	0.963	0.969	0.974	0.978	
$\alpha = 1.9$	3.166E-02	1.653E-02	8.583E-03	4.436E-03	2.284E-03	1.173E-03	6.007E-04	3.070E-04
	0.937	0.946	0.952	0.957	0.962	0.965	0.968	

Table 7: Example II: Maximum errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a uniform mesh

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	5.498E-03	2.438E-03	1.085E-03	4.844E-04	2.170E-04	9.747E-05	4.391E-05	1.983E-05
	1.173	1.168	1.163	1.159	1.155	1.151	1.147	
$\alpha = 1.2$	1.525E-03	5.963E-04	2.684E-04	1.304E-04	6.160E-05	2.854E-05	1.304E-05	5.905E-06
	1.355	1.152	1.041	1.082	1.110	1.129	1.144	
$\alpha = 1.3$	9.601E-04	4.722E-04	2.177E-04	9.681E-05	4.207E-05	1.799E-05	7.607E-06	3.190E-06
	1.024	1.117	1.169	1.202	1.226	1.242	1.254	
$\alpha = 1.4$	7.204E-04	3.253E-04	1.388E-04	5.725E-05	2.312E-05	9.199E-06	3.622E-06	1.415E-06
	1.147	1.229	1.277	1.308	1.329	1.345	1.356	
$\alpha = 1.5$	4.571E-04	1.935E-04	7.718E-05	2.977E-05	1.124E-05	4.180E-06	1.538E-06	5.615E-07
	1.240	1.326	1.374	1.406	1.427	1.442	1.454	
$\alpha = 1.6$	2.723E-04	1.027E-04	3.867E-05	1.404E-05	4.983E-06	1.740E-06	6.004E-07	2.054E-07
	1.407	1.408	1.461	1.495	1.518	1.535	1.547	
$\alpha = 1.7$	2.189E-04	5.902E-05	1.732E-05	5.980E-06	2.009E-06	6.622E-07	2.154E-07	6.936E-08
	1.891	1.769	1.534	1.574	1.601	1.620	1.635	
$\alpha = 1.8$	1.954E-04	5.001E-05	1.266E-05	3.186E-06	7.994E-07	2.208E-07	6.824E-08	2.083E-08
	1.966	1.982	1.990	1.995	1.856	1.694	1.712	
$\alpha = 1.9$	1.957E-04	4.998E-05	1.263E-05	3.175E-06	7.960E-07	1.993E-07	4.986E-08	1.247E-08
	1.970	1.984	1.992	1.996	1.998	1.999	1.999	

Table 8: Example II: Maximum errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a graded mesh with $r = 1/(\alpha - 1)$

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	4.686E-03	1.347E-03	4.078E-04	2.305E-04	1.169E-04	5.667E-05	2.691E-05	1.266E-05
	1.799	1.724	0.823	0.979	1.045	1.074	1.088	
$\alpha = 1.2$	2.150E-03	1.057E-03	4.924E-04	2.204E-04	9.701E-05	4.238E-05	1.846E-05	8.034E-06
	1.024	1.102	1.159	1.184	1.195	1.199	1.200	
$\alpha = 1.3$	1.908E-03	8.072E-04	3.315E-04	1.345E-04	5.440E-05	2.199E-05	8.895E-06	3.602E-06
	1.241	1.284	1.301	1.306	1.307	1.306	1.304	
$\alpha = 1.4$	1.339E-03	5.069E-04	1.897E-04	7.084E-05	2.649E-05	9.934E-06	3.735E-06	1.407E-06
	1.401	1.418	1.421	1.419	1.415	1.411	1.408	
$\alpha = 1.5$	8.558E-04	2.953E-04	1.015E-04	3.498E-05	1.210E-05	4.208E-06	1.469E-06	5.143E-07
	1.535	1.540	1.537	1.531	1.524	1.519	1.514	
$\alpha = 1.6$	5.309E-04	1.679E-04	5.315E-05	1.690E-05	5.407E-06	1.740E-06	5.625E-07	1.826E-07
	1.661	1.659	1.653	1.644	1.636	1.629	1.623	
$\alpha = 1.7$	3.811E-04	9.938E-05	2.834E-05	8.311E-06	2.453E-06	7.287E-07	2.177E-07	6.535E-08
	1.939	1.810	1.770	1.760	1.751	1.743	1.736	
$\alpha = 1.8$	3.000E-04	7.754E-05	1.972E-05	4.972E-06	1.249E-06	3.289E-07	9.056E-08	2.506E-08
	1.952	1.976	1.987	1.994	1.925	1.861	1.854	
$\alpha = 1.9$	2.395E-04	6.144E-05	1.556E-05	3.916E-06	9.822E-07	2.460E-07	6.155E-08	1.539E-08
	1.963	1.981	1.991	1.995	1.998	1.999	1.999	

Table 9: Example II: Maximum derivative errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a uniform mesh

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	1.130E-02	8.939E-03	7.133E-03	5.737E-03	4.648E-03	3.789E-03	3.106E-03	2.560E-03
	0.339	0.326	0.314	0.304	0.295	0.287	0.279	
$\alpha = 1.2$	1.141E-02	7.808E-03	5.430E-03	3.826E-03	2.725E-03	1.958E-03	1.418E-03	1.034E-03
	0.547	0.524	0.505	0.490	0.477	0.466	0.456	
$\alpha = 1.3$	7.765E-03	4.703E-03	2.899E-03	1.812E-03	1.145E-03	7.297E-04	4.682E-04	3.020E-04
	0.723	0.698	0.678	0.662	0.650	0.640	0.632	
$\alpha = 1.4$	4.590E-03	2.469E-03	1.350E-03	7.476E-04	4.178E-04	2.351E-04	1.329E-04	7.545E-05
	0.895	0.871	0.853	0.840	0.830	0.822	0.817	
$\alpha = 1.5$	2.506E-03	1.195E-03	5.777E-04	2.822E-04	1.388E-04	6.859E-05	3.402E-05	1.691E-05
	1.069	1.048	1.034	1.024	1.017	1.012	1.008	
$\alpha = 1.6$	1.293E-03	5.444E-04	2.318E-04	9.946E-05	4.289E-05	1.855E-05	8.042E-06	3.491E-06
	1.248	1.232	1.221	1.214	1.209	1.206	1.204	
$\alpha = 1.7$	6.380E-04	2.363E-04	8.828E-05	3.316E-05	1.250E-05	4.721E-06	1.785E-06	6.757E-07
	1.433	1.420	1.412	1.408	1.405	1.403	1.402	
$\alpha = 1.8$	3.027E-04	9.834E-05	3.216E-05	1.055E-05	3.472E-06	1.143E-06	3.768E-07	1.242E-07
	1.622	1.613	1.607	1.604	1.602	1.601	1.601	
$\alpha = 1.9$	1.983E-04	5.246E-05	1.385E-05	3.656E-06	9.651E-07	2.657E-07	7.627E-08	2.190E-08
	1.918	1.921	1.922	1.922	1.861	1.801	1.800	

Table 10: Example II: Maximum derivative errors and orders of convergence using a collocation method for $m = 1$ and $c_1 = 1/2$ on a graded mesh with $r = 1/(\alpha - 1)$

	N=32	N=64	N=128	N=256	N=512	N=1024	N=2048	N=4096
$\alpha = 1.1$	2.130E-02	1.066E-02	5.165E-03	2.460E-03	1.160E-03	5.447E-04	2.550E-04	1.192E-04
	0.998	1.045	1.070	1.084	1.091	1.095	1.097	
$\alpha = 1.2$	1.338E-02	6.172E-03	2.778E-03	1.233E-03	5.427E-04	2.378E-04	1.039E-04	4.533E-05
	1.116	1.152	1.172	1.184	1.190	1.194	1.197	
$\alpha = 1.3$	7.353E-03	3.170E-03	1.335E-03	5.546E-04	2.283E-04	9.352E-05	3.818E-05	1.556E-05
	1.214	1.247	1.268	1.280	1.288	1.292	1.295	
$\alpha = 1.4$	3.743E-03	1.517E-03	6.000E-04	2.338E-04	9.019E-05	3.458E-05	1.321E-05	5.030E-06
	1.303	1.338	1.360	1.374	1.383	1.389	1.393	
$\alpha = 1.5$	1.783E-03	6.829E-04	2.547E-04	9.341E-05	3.387E-05	1.218E-05	4.361E-06	1.555E-06
	1.385	1.423	1.447	1.464	1.475	1.482	1.488	
$\alpha = 1.6$	1.062E-03	3.298E-04	1.030E-04	3.510E-05	1.202E-05	4.074E-06	1.371E-06	4.593E-07
	1.687	1.680	1.553	1.546	1.561	1.571	1.578	
$\alpha = 1.7$	6.272E-04	1.822E-04	5.304E-05	1.551E-05	4.562E-06	1.350E-06	4.017E-07	1.265E-07
	1.783	1.781	1.774	1.765	1.757	1.749	1.667	
$\alpha = 1.8$	3.720E-04	1.020E-04	2.793E-05	7.652E-06	2.101E-06	5.787E-07	1.599E-07	4.432E-08
	1.867	1.869	1.868	1.865	1.860	1.856	1.851	
$\alpha = 1.9$	2.228E-04	5.836E-05	1.523E-05	3.968E-06	1.033E-06	2.691E-07	7.011E-08	1.828E-08
	1.933	1.938	1.940	1.941	1.941	1.941	1.940	

[2] Brunner, H. *Collocation methods for Volterra integral and related functional differential equations*, vol. 15 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2004. Available from: https://doi.org/10.1017/CBO9780511543234, doi:10.1017/CBO9780511543234.

[3] del Castillo-Negrete, D. Fractional diffusion models of nonlocal transport. *Phys. Plasmas 13*, 082308 (2006).

[4] Diethelm, K. *The analysis of fractional differential equations*, vol. 2004 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2010. An application-oriented exposition using differential operators of Caputo type.

[5] Ervin, V. J., Heuer, N., and Roop, J. P. Regularity of the solution to 1-D fractional order diffusion equations. *Math. Comp. 87*, 313 (2018), 2273–2294. Available from: https://doi.org/10.1090/mcom/3295, doi:10.1090/mcom/3295.

[6] Gracia, J. L., O'Riordan, E., and Stynes, M. Convergence analysis of a finite difference scheme for a two-point boundary value problem with a Riemann-Liouville-Caputo fractional derivative. *Submitted for publication*.

[7] Jia, L., Huanzhen, C., and Ervin, V. J. Existence and regularity of solutions to 1-d fractional order diffusion equations. arXiv:1808.10555.

[8] Kelly, J. F., Sankaranarayanan, H., and Meerschaert, M. M. Boundary conditions for two-sided fractional diffusion. *J. Comput. Phys. 376* (2019), 1089–1107. Available from: http://www.sciencedirect.com/science/article/pii/S0021999118306673, doi:https://doi.org/10.1016/j.jcp.2018.10.010.

[9] Kopteva, N., and Stynes, M. An efficient collocation method for a Caputo two-point boundary value problem. *BIT 55*, 4 (2015), 1105–1123. Available from: https://doi.org/10.1007/s10543-014-0539-4, doi:10.1007/s10543-014-0539-4.

[10] Kopteva, N., and Stynes, M. Analysis and numerical solution of a Riemann-Liouville fractional derivative two-point boundary value problem. *Adv. Comput. Math. 43*, 1 (2017), 77–99. Available from: https://doi.org/10.1007/s10444-016-9476-x.

[11] Oldham, K. B., and Spanier, J. *The fractional calculus*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1974. Theory and applications of differentiation and integration to arbitrary order, With an annotated chronological bibliography by Bertram Ross, Mathematics in Science and Engineering, Vol. 111.

[12] Patie, P., and Simon, T. Intertwining certain fractional derivatives. *Potential Anal. 36*, 4 (2012), 569–587. Available from: https://doi.org/10.1007/s11118-011-9241-1.

[13] Podlubny, I. Mittag-Leffler function, 2012. Available from: https://es.mathworks.com/matlabcentral/fileexchange/8738-mittag-leffler-function.

[14] Stynes, M., and Gracia, J. L. A finite difference method for a two-point boundary value problem with a Caputo fractional derivative. *IMA J. Numer. Anal. 35*, 2 (2015), 698–721. Available from: https://doi.org/10.1093/imanum/dru011.

[15] Wang, H., and Yang, D. Wellposedness of Neumann boundary-value problems of space-fractional differential equations. *Fract. Calc. Appl. Anal. 20*, 6 (2017), 1356–1381. Available from: https://doi.org/10.1515/fca-2017-0072.

J. L. Gracia
Instituto Universitario de Aplicaciones y Matemáticas (IUMA)
Department of Applied Mathematics
University of Zaragoza, Spain
jlgracia@unizar.es

E. O'Riordan
School of Mathematical Sciences
Dublin City University, Ireland
eugene.oriordan@dcu.ie

M. Stynes
Applied and Computational Mathematics Division
Beijing Computational Science Research Center, China
m.stynes@csrc.ac.cn

Monografías Matemáticas *García de Galdeano* **42**, 127–141 (2019)

A DECOUPLED STAGGERED SCHEME FOR THE SHALLOW WATER EQUATIONS

Raphaèle Herbin, Jean-Claude Latché, Youssouf Nasseri and Nicolas Therme

Abstract. We present a first order scheme based on a staggered grid for the shallow water equations with topography in two space dimensions, which enjoys several properties: positivity of the water height, preservation of constant states, and weak consistency with the equations of the problem and with the associated entropy inequality.

Keywords: Shallow water, finite volumes, staggered grid.
AMS classification: 65M08,76B99.

§1. Introduction

The shallow water equations form a hyperbolic system of two conservation equations (mass and momentum) which are obtained when modelling a flow whose vertical height is considered small with respect to the plane scale. The solution of such a system may develop shocks, so that the finite volume method is usually preferred for numerical simulations. Two main approaches are found: one is the colocated approach which is usually based on some approximate Riemann solver, see e.g. [3] and references therein; the other one is based on a staggered arrangement of the unknowns on the grid. Indeed, staggered schemes have been used for some time in the hydraulic and ocean engineering community, see e.g. [1, 2, 12]. They have been recently analysed in the case of one space dimension [5, 8], following the works on the related barotropic Euler equations, see [11] and references therein. In the present work, we obtain a discrete local entropy inequality; furthermore, we extend the consistency analysis of the scheme to the case of two space dimensions, and we weaken the assumptions on the estimates, namely we no longer require a bound on the BV norm of the approximate solutions, at least for the weak formulation (the passage to the limit in the entropy still necessitates a time BV boundedness).

Let Ω be an open bounded domain of $\mathbb{R}^2$ and let $T > 0$. We consider the shallow water equations with topography over the space and time domain $\Omega \times (0, T)$:

$$\partial_t h + \operatorname{div}(h\boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{1a}$$

$$\partial_t (h\boldsymbol{u}) + \operatorname{div}(h\boldsymbol{u} \otimes \boldsymbol{u}) + \nabla p + gh\nabla z = 0 \qquad \text{in } \Omega \times (0,T), \tag{1b}$$

$$p = \frac{1}{2} g h^2 \qquad \text{in } \Omega \times (0,T), \tag{1c}$$

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0 \qquad \text{on } \partial\Omega \times (0,T), \tag{1d}$$

$$h(\boldsymbol{x}, 0) = h_0, \ \boldsymbol{u}(\boldsymbol{x}, 0) = \boldsymbol{u}_0 \qquad \text{in } \Omega. \tag{1e}$$

where t stands for the time, g is the standard gravity constant and z the (given) topography, which is supposed to be regular in this paper. These equations solve the water height h and the velocity $\boldsymbol{u}$.

Let us recall that if $(h, \boldsymbol{u})$ is a regular solution of (1), the following elastic potential energy balance and kinetic energy balance is obtained by manipulations on the mass and momentum equations:

$$\partial_t(\frac{1}{2}gh^2) + \mathrm{div}(\frac{1}{2}gh^2\boldsymbol{u}) + \frac{1}{2}gh^2\mathrm{div}\boldsymbol{u} = 0 \tag{2}$$

$$\partial_t(\frac{1}{2}h|\boldsymbol{u}|^2) + \mathrm{div}(\frac{1}{2}h|\boldsymbol{u}|^2\boldsymbol{u}) + \boldsymbol{u}\cdot\nabla p + gh\boldsymbol{u}\cdot\nabla z = 0. \tag{3}$$

Summing these equations, we obtain en entropy equality of the form $\partial_t\eta + \mathrm{div}\Phi = 0$, where the entropy-entropy flux pair (η, Φ) is given by:

$$\eta = \frac{1}{2}h|\boldsymbol{u}|^2 + \frac{1}{2}gh^2 + ghz \text{ and } \Phi = (\eta + \frac{1}{2}gh^2)\boldsymbol{u}. \tag{4}$$

For non regular functions the above manipulations are no longer valid, and the entropy inequality $\partial_t\eta + \mathrm{div}\Phi \leq 0$ is satisfied in a distributional sense.

In this paper, we build a decoupled scheme, involving only explicit steps; the resulting approximate solutions are shown to satisfy some discrete equivalent of (2) and (3); furthermore, under some convergence and boundedness assumptions, the approximate solutions are shown in Section 5 to converge to a weak solution of (1) and to satisfy a weak entropy inequality.

§2. Mesh and space discretizations

Let Ω be a connected subset of $\mathbb{R}^2$ consisting in a union of rectangles whose edges are assumed to be orthogonal to the canonical basis vectors, denoted by $(\boldsymbol{e}^{(1)}, \boldsymbol{e}^{(2)})$.

Definition 1 (MAC grid)**.** A discretization $(\mathcal{M}, \mathcal{E})$ of Ω with a staggered rectangular grid (or MAC grid), is defined by:

- A primal grid $\mathcal{M}$ which consists in a conforming structured partition of Ω in rectangles, possibly non uniform. A generic cell of this grid is denoted by K, and its mass center by $\boldsymbol{x}_K$. The scalar unknowns (water height and pressure) are associated to this mesh.
- The set of all edges of the mesh $\mathcal{E}$, with $\mathcal{E} = \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{ext}}$, where $\mathcal{E}_{\mathrm{int}}$ (resp. $\mathcal{E}_{\mathrm{ext}}$) are the edges of $\mathcal{E}$ that lie in the interior (resp. on the boundary) of the domain. The set of edges that are orthogonal to $\boldsymbol{e}^{(i)}$ is denoted by $\mathcal{E}^{(i)}$, for $i = 1, 2$. We then have $\mathcal{E}^{(i)} = \mathcal{E}^{(i)}_{\mathrm{int}} \cup \mathcal{E}^{(i)}_{\mathrm{ext}}$, where $\mathcal{E}^{(i)}_{\mathrm{int}}$ (resp. $\mathcal{E}^{(i)}_{\mathrm{ext}}$) are the edges of $\mathcal{E}^{(i)}$ that lie in the interior (resp. on the boundary) of the domain.

  For $\sigma \in \mathcal{E}_{\mathrm{int}}$, we write $\sigma = K|L$ if $\sigma = \partial K \cap \partial L$. A dual cell D_σ associated to an edge $\sigma \in \mathcal{E}$ is defined as follows:

  - if $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$ then $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$, where $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) is the half-part of K (resp. L) adjacent to σ (see Fig. 1);

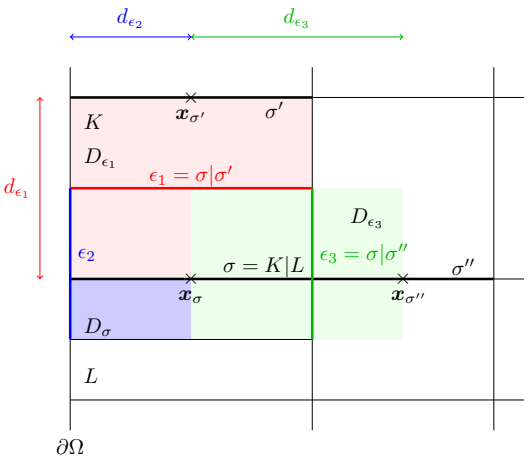


Figure 1: Notations for control volumes and dual cells (in two space dimensions, for the second component of the velocity).

- if $\sigma \in \mathcal{E}_{\text{ext}}$ is adjacent to the cell K, then $D_\sigma = D_{K,\sigma}$.

For each dimension $i = 1, 2$, the domain Ω is partitioned in dual cells: $\Omega = \cup_{\sigma\in\mathcal{E}^{(i)}} D_\sigma$, $i = 1, 2$; the i^{th} partition is refered to as the i^{th} dual mesh; it is associated to the i^{th} velocity component, in a sense which is clarified below. The set of the edges of the i^{th} dual mesh is denoted by $\widetilde{\mathcal{E}}^{(i)}$ (note that these edges may be orthogonal to any vector of the basis of $\mathbb{R}^2$ and not only $\boldsymbol{e}^{(i)}$) and is decomposed into the internal and boundary edges: $\widetilde{\mathcal{E}}^{(i)} = \widetilde{\mathcal{E}}^{(i)}_{\text{int}} \cup \widetilde{\mathcal{E}}^{(i)}_{\text{ext}}$. The dual edge separating two duals cells D_σ and $D_{\sigma'}$ is denoted by $\epsilon = \sigma|\sigma'$. We denote by D_ϵ the dual cell associated to a dual edge $\epsilon \in \widetilde{\mathcal{E}}$ defined as follows:

- if $\epsilon = \sigma|\sigma' \in \widetilde{\mathcal{E}}_{int}$ then $D_\epsilon = D_{\sigma,\epsilon} \cup D_{\sigma',\epsilon}$, where $D_{\sigma,\epsilon}$ (resp. $D_{\sigma',\epsilon}$) is the half-part of D_σ (resp. $D_{\sigma'}$) adjacent to ϵ (see Fig. 1);
- if $\epsilon \in \widetilde{\mathcal{E}}_{ext}$ is adjacent to the cell D_σ, then $D_\epsilon = D_{\sigma,\epsilon}$.

In order to define the scheme, we need some additional notations. The set of edges of a primal cell K and of a dual cell D_σ are denoted by $\mathcal{E}(K)$ and $\widetilde{\mathcal{E}}(D_\sigma)$ respectively. For $\sigma \in \mathcal{E}$, we denote by $\boldsymbol{x}_\sigma$ the mass center of σ. The vector $\boldsymbol{n}_{K,\sigma}$ stands for the unit normal vector to σ outward K. In some cases, we need to specify the orientation of various geometrical entities with respect to the axis:

- a primal cell K will be denoted $K = [\overrightarrow{\sigma\sigma'}]$ if $\sigma, \sigma' \in \mathcal{E}^{(i)}(K)$ for some $i = 1, 2$ are such that $(\boldsymbol{x}_{\sigma'} - \boldsymbol{x}_\sigma) \cdot \boldsymbol{e}^{(i)} > 0$;
- we write $\sigma = \overrightarrow{K|L}$ if $\sigma \in \mathcal{E}^{(i)}$, $\sigma = K|L$ and $\overrightarrow{\boldsymbol{x}_K\boldsymbol{x}_L} \cdot \boldsymbol{e}^{(i)} > 0$ for some $i = 1, 2$;
- the dual edge ϵ separating D_σ and $D_{\sigma'}$ is written $\epsilon = \overrightarrow{\sigma|\sigma'}$ if $\overrightarrow{\boldsymbol{x}_\sigma\boldsymbol{x}_{\sigma'}} \cdot \boldsymbol{e}^{(i)} > 0$ for some $i = 1, 2$.

The size $\delta_{\mathcal{M}}$ of the mesh and its regularity $\eta_{\mathcal{M}}$ are defined by:

$$\delta_{\mathcal{M}} = \max_{K\in\mathcal{M}} \operatorname{diam}(K), \text{ and } \eta_{\mathcal{M}} = \max\Big\{\frac{|\sigma|}{|\sigma'|},\ \sigma\in\mathcal{E}^{(i)},\ \sigma'\in\mathcal{E}^{(j)},\ i,j=1,2,\ i\neq j\Big\}, \tag{5}$$

where $|\cdot|$ stands for the one (or two) dimensional measure of a subset of $\mathbb{R}$ (or $\mathbb{R}^2$).

The discrete velocity unknowns are associated to the dual cells and are denoted by $(u_{i,\sigma})_{\sigma\in\mathcal{E}^{(i)}}$, $i=1,2$, while the scalar unknowns (discrete water height and pressure) are associated to the primal cells and are denoted respectively by $(h_K)_{K\in\mathcal{M}}$ and $(p_K)_{K\in\mathcal{M}}$. The scalar unknown space $L_{\mathcal{M}}$ is defined as the set of piecewise constant functions over each grid cell K of $\mathcal{M}$, and the discrete i^{th} velocity space $H_{\mathcal{E}^{(i)}}$ as the set of piecewise constant functions over each of the grid cells $D_\sigma,\ \sigma\in\mathcal{E}^{(i)}$. As in the continuous case, the Dirichlet boundary conditions are taken into account by defining the subspaces $H_{\mathcal{E}^{(i)},0}\subset H_{\mathcal{E}^{(i)}},\ i=1,2$ as follows

$$H_{\mathcal{E}^{(i)},0} = \Big\{u_i\in H_{\mathcal{E}^{(i)}},\ u_i(\boldsymbol{x})=0,\ \forall \boldsymbol{x}\in D_\sigma,\ \sigma\in\mathcal{E}^{(i)}_{\text{ext}}\Big\}.$$

We then set $\boldsymbol{H}_{\mathcal{E},0} = H_{\mathcal{E}^{(1)},0}\times H_{\mathcal{E}^{(2)},0}$. Defining the characteristic function $\mathbb{1}_A$ of any subset $A\subset\Omega$ by $\mathbb{1}_A(\boldsymbol{x})=1$ if $\boldsymbol{x}\in A$ and $\mathbb{1}_A(\boldsymbol{x})=0$ otherwise, the functions $\boldsymbol{u}=(u_1,u_2)\in\boldsymbol{H}_{\mathcal{E},0}$, may then be written:

$$u_i(\boldsymbol{x}) = \sum_{\sigma\in\mathcal{E}^{(i)}} u_{i,\sigma}\mathbb{1}_{D_\sigma}(\boldsymbol{x}),\ i=1,2. \tag{6}$$

For $\boldsymbol{u}\in\boldsymbol{H}_{\mathcal{E},0}$, let $[\![u_i]\!]_\epsilon = |u_{i,\sigma}-u_{i,\sigma'}|$, for $\epsilon=\sigma|\sigma'\in\widetilde{\mathcal{E}}^{(i)}_{int}, i=1,2$. In the same way the functions $h\in L_{\mathcal{M}}$ are defined by $h(\boldsymbol{x})=\sum_{K\in\mathcal{M}} h_K\mathbb{1}_K(\boldsymbol{x})$ and the notation $[\![\ ]\!]_\sigma$ refers to $[\![h]\!]_\sigma = |h_K-h_L|$, for $\sigma=K|L\in\mathcal{E}_{\text{int}}(K)$.

§3. A decoupled explicit scheme

Description of the scheme Let us consider a uniform discretisation $0=t_0<t_1<\cdots<t_N=T$ of the time interval $(0,T)$, and let $\delta t=t_{n+1}-t_n$ for $n=0,1,\cdots,N-1$ be the (constant) time step. The discrete velocity $\boldsymbol{u}$ and water height h unknowns are defined by:

$$\boldsymbol{u}(\boldsymbol{x},t) = \sum_{n=0}^{N-1}\boldsymbol{u}^{n+1}(\boldsymbol{x})\mathbb{1}_{[t_n,t_{n+1})}(t),\ \text{ with } \boldsymbol{u}^{n+1}\in\boldsymbol{H}_{\mathcal{E},0},$$
$$h(\boldsymbol{x},t) = \sum_{n=0}^{N-1}h^{n+1}(\boldsymbol{x})\mathbb{1}_{[t_n,t_{n+1})}(t),\ \text{ with } h^{n+1}\in L_{\mathcal{M}},$$

where $\mathbb{1}_{[t_n,t_{n+1})}$ is the characteristic function of the interval $[t_n,t_{n+1})$ and the space functions $\boldsymbol{u}^n$ and h^n take the form defined in the previous section. We propose the following decoupled discretisation of the system (1), written in compact form, with the various discrete operators

defined below.

$$\textbf{Initialisation}: \qquad \boldsymbol{u}^0 = \mathcal{P}_{\mathcal{E}}\boldsymbol{u}_0,\ h^0 = \mathcal{P}_{\mathcal{M}}h_0,\ p^0 = \frac{1}{2}g(h^0)^2. \tag{7a}$$

Iteration n, $0 \le n \le N-1$: solve for $\boldsymbol{u}^{n+1} \in \boldsymbol{H}_{\mathcal{E},0}, h^{n+1} \in L_{\mathcal{M}}$ and $p^{n+1} \in L_{\mathcal{M}}$:

$$\eth_t h^{n+1} + \mathrm{div}_{\mathcal{M}}(h^n\boldsymbol{u}^n) = 0, \tag{7b}$$

$$p^{n+1} = \frac{1}{2}g(h^{n+1})^2, \tag{7c}$$

$$\eth_t(h\boldsymbol{u})^{n+1} + \boldsymbol{C}_{\mathcal{E}}(h^n\boldsymbol{u}^n)\boldsymbol{u}^n + \nabla_{\mathcal{E}}p^{n+1} + g\,\boldsymbol{I}_{\mathcal{E}}h^{n+1}\,\nabla_{\mathcal{E}}z = 0, \tag{7d}$$

Projection operators - The operators $\mathcal{P}_{\mathcal{E}}$ and $\mathcal{P}_{\mathcal{M}}$ used in the initialisation step are defined by $\mathcal{P}_{\mathcal{E}} = (\mathcal{P}_{\mathcal{E}^{(i)}})_{i=1,\cdots,d}$ with

$$\mathcal{P}_{\mathcal{E}^{(i)}}: \left| \begin{array}{l} L^1(\Omega) \longrightarrow H_{\mathcal{E}^{(i)},0} \\ v \longmapsto \mathcal{P}_{\mathcal{E}^{(i)}}v = \displaystyle\sum_{\sigma\in\mathcal{E}^{(i)}_{\mathrm{int}}} v_\sigma\, \mathbb{1}_{D_\sigma} \text{ with } v_\sigma = \frac{1}{|D_\sigma|}\int_{D_\sigma} v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\ \text{for } \sigma \in \mathcal{E}^{(i)}_{\mathrm{int}}. \end{array} \right. \tag{8}$$

For $q \in L^2(\Omega)$, $\mathcal{P}_{\mathcal{M}}q \in L_{\mathcal{M}}$ is defined by:

$$\mathcal{P}_{\mathcal{M}}q = \sum_{K\in\mathcal{M}} q_K\,\mathbb{1}_K \text{ with } q_K = \frac{1}{|K|}\int_K q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \text{ for } K \in \mathcal{M}. \tag{9}$$

Discrete time derivative - The symbol $\eth_t$ denotes the discrete time derivative for both water height and momentum:

$$\eth_t h^{n+1} = \sum_{K\in\mathcal{M}} \frac{1}{\delta t}(h_K^{n+1} - h_K^n)\mathbb{1}_K, \qquad \eth_t(h\boldsymbol{u})^{n+1} = (\eth_t(hu_1)^{n+1}, \ldots, \eth_t(hu_d)^{n+1})$$
$$\text{with } \eth_t(hu_i)^{n+1} = \sum_{\sigma\in\mathcal{E}^{(i)}} \frac{1}{\delta t}(h_{D_\sigma}^{n+1}u_{i,\sigma}^{n+1} - h_{D_\sigma}^n u_{i,\sigma}^n)\mathbb{1}_{D_\sigma},\ i = 1,2,$$

where h_{D_σ} is the discrete water height in the dual cell, which is computed from the primal unknowns $(h_K^n)_{n\in\mathbb{N},K\in\mathcal{M}}$ and defined so as to satisfy a discrete mass balance, see below.

Discrete divergence and gradient operators - The discrete divergence operator $\mathrm{div}_{\mathcal{M}}$ is defined by:

$$\mathrm{div}_{\mathcal{M}}: \left| \begin{array}{l} \boldsymbol{H}_{\mathcal{E},0} \longrightarrow L_{\mathcal{M},0} \\ \boldsymbol{u} \longmapsto \mathrm{div}_{\mathcal{M}}(h\boldsymbol{u}) = \displaystyle\sum_{K\in\mathcal{M}} \mathrm{div}_K(h\boldsymbol{u})\mathbb{1}_K,\ \text{with } \mathrm{div}_K(h\boldsymbol{u}) = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}, \end{array} \right. \tag{10}$$

where $F_{K,\sigma}$ is the (conservative) numerical mass flux, defined by $F_{K,\sigma} = |\sigma|\ h_\sigma u_{K,\sigma}$ with $u_{K,\sigma} = u_{i,\sigma}\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)}$ for $\sigma \in \mathcal{E}^{(i)}_{\mathrm{int}}$, $i = 1,2$, while h_σ is approximated by the first order upwind scheme namely, for $\sigma = K|L \in \mathcal{E}_{int}$, $h_\sigma = h_K$ if $u_{K,\sigma} \ge 0$ and $h_\sigma = h_L$ otherwise.

The discrete gradient operator applies to the pressure and the topography and is defined by:

$$\nabla_{\mathcal{E}} : \quad \left| \begin{array}{l} L_{\mathcal{M}} \longrightarrow \boldsymbol{H}_{\mathcal{E},0} \\ p \longmapsto \nabla_{\mathcal{E}} p, \end{array} \right.$$

with for $i = 1, 2$:

$$(\nabla_{\mathcal{E}} p)_i = \sum_{\sigma \in \mathcal{E}_{int}^{(i)}} (\eth_i p)_\sigma \mathbb{1}_{D_\sigma} \text{ with for } \sigma = \overrightarrow{K|L},\ (\eth_i p)_\sigma = \frac{|\sigma|}{|D_\sigma|}\,(p_L - p_K). \tag{11}$$

The above defined discrete divergence and gradient operators satisfy the following div-grad duality relationship [7, Lemma 2.5]:

$$\text{for } p \in L_{\mathcal{M}},\ \boldsymbol{u} \in \boldsymbol{H}_{\mathcal{E},0}, \quad \int_\Omega p \,\mathrm{div}_{\mathcal{M}}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{x} + \int_\Omega \nabla_{\mathcal{E}} p \cdot (\boldsymbol{u})\,\mathrm{d}\boldsymbol{x} = 0.$$

Discrete convection operator – The discrete nonlinear convection operator $\boldsymbol{C}_{\mathcal{E}}(h\boldsymbol{u})$ is linked to the discrete divergence operator on the dual mesh by the relation $\boldsymbol{C}_{\mathcal{E}}(h\boldsymbol{u})\boldsymbol{u} = \mathrm{div}_{\mathcal{E}}(h\boldsymbol{u} \otimes \boldsymbol{u})$, where the full discrete convection operator $\boldsymbol{C}_{\mathcal{E}}(h\boldsymbol{u})$ is defined by:

$$\boldsymbol{C}_{\mathcal{E}}(h\boldsymbol{u})\,\boldsymbol{u} = (C_{\mathcal{E}^{(1)}}(h\boldsymbol{u})\,u_1, C_{\mathcal{E}^{(2)}}(h\boldsymbol{u})\,u_2),$$

and the i-th component $C_{\mathcal{E}^{(i)}}(h\boldsymbol{u})$ of the convection operator is defined by:

$$C_{\mathcal{E}^{(i)}}(h\boldsymbol{u}) : \quad \left| \begin{array}{l} H_{\mathcal{E}^{(i)},0} \longrightarrow H_{\mathcal{E}^{(i)},0} \\ u_i \longmapsto C_{\mathcal{E}^{(i)}}(h\boldsymbol{u})\,u_i = \displaystyle\sum_{\sigma \in \mathcal{E}_{\mathrm{int}}^{(i)}} \mathrm{div}_{\mathcal{E}^{(i)}}(hu_i\boldsymbol{u})\ \mathbb{1}_{D_\sigma}, \\ \text{with } \mathrm{div}_{\mathcal{E}^{(i)}}(hu_i\boldsymbol{u}) = \dfrac{1}{|D_\sigma|} \displaystyle\sum_{\epsilon \in \widetilde{\mathcal{E}}^{(i)}(D_\sigma)} F_{\sigma,\epsilon} u_{i,\epsilon}, \end{array} \right. \tag{12}$$

where $u_{i,\epsilon}$ is approximated by the upwind technique with respect to the sign of $F_{\sigma,\epsilon}$. The quantity $F_{\sigma,\epsilon}$ is the numerical mass flux through ϵ outward D_σ; it must be chosen carefully to ensure some stability properties of the scheme as in [7, 11]. Indeed we recall that in order to derive a discrete kinetic energy balance (Lemma 3 below), it is necessary that a discrete equation of the mass balance holds in the dual mesh, namely:

$$\frac{|D_\sigma|}{\delta t}(h_{D_\sigma}^{n+1} - h_{D_\sigma}^{n}) + \mathrm{div}_{\mathcal{E}}(h^n\boldsymbol{u}^n) = 0, \quad \text{with} \quad |D_\sigma|\,\mathrm{div}_{\mathcal{E}}(h^n\boldsymbol{u}^n) = \sum_{\epsilon \in \widetilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n. \tag{13}$$

The water height h_{D_σ} and the flux $F_{\sigma,\epsilon}$ are computed from the primal unknowns and fluxes so as to satisfy this latter relation thanks to the discrete mass balance on the primal mesh (7b). For $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$, the water height h_{D_σ} is defined as a weighted average between h_K and h_L:

$$|D_\sigma|\,h_{D_\sigma} = |D_{K,\sigma}|\,h_K + |D_{L,\sigma}|\,h_L, \tag{14}$$

where D_σ, $D_{K,\sigma}$ and $D_{L,\sigma}$ are defined in Definition 1. The numerical flux $F_{\sigma,\epsilon}$ on the internal dual edges, is defined according to the location of the edges as follows:

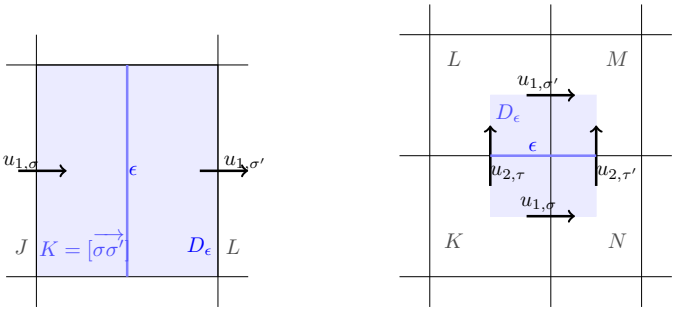


Figure 2: Notations for the definition of the momentum flux on the dual mesh for the first component of the velocity- left: first case - right: second case.

- First case – The vector $\boldsymbol{e}^{(i)}$ is normal to ϵ, and ϵ is included in a primal cell K, with $K = [\overrightarrow{\sigma\sigma'}]$ (see Definition 1 and Figure 2 on the left for $i = 1$). Then for a dual edge $\epsilon \in \widetilde{\mathcal{E}}^{(i)}$ such that $\epsilon = \overrightarrow{\sigma|\sigma'}$, the flux $F_{\sigma,\epsilon}$ through the edge ϵ is given by:

$$F_{\sigma,\epsilon} = \frac{1}{2}(F_{K,\sigma'} - F_{K,\sigma}) = \frac{1}{2}|\epsilon|\,(h_\sigma u_{i,\sigma} + h_{\sigma'} u_{i,\sigma'}), \tag{15}$$

  since $|\sigma| = |\sigma'| = |\epsilon|$.

- Second case – The vector $\boldsymbol{e}^{(i)}$ is tangent to ϵ, and ϵ is the union of the halves of two primal edges τ and τ' such that $\tau = \overrightarrow{K|L}$, $\tau \in \mathcal{E}(K)$ and $\tau' = \overrightarrow{N|M} \in \mathcal{E}(N)$ (see Definition 1 and Figure 2 on the right for $i = 2$). The flux numerical through ϵ is then given by:

$$F_{\sigma,\epsilon} = \frac{1}{2}\,(F_{K\tau} + F_{L\tau'}) = \frac{1}{2}\,(|\tau|\,h_\tau u_{,\tau} + |\tau'|\,h_{\tau'} u_{i,\tau'}). \tag{16}$$

Note that the numerical momentum flux on a dual edge is conservative. It is easy to check that the unknowns $h^n_{D_\sigma}$ and $F^n_{\sigma,\epsilon}$ thus defined satisfy the discrete dual mass balance (13).

Discrete water height on the dual mesh, for the topography term – In equation (7d) the interpolation operator $\boldsymbol{I}_{\mathcal{E}}$ is defined as the mean value of the water height:

$$\boldsymbol{I}_{\mathcal{E}} h = \sum_{\sigma\in\mathcal{E}_{int}} h_{\sigma,c}\,\mathbb{1}_{D_\sigma} \text{ with } h_{\sigma,c} = \begin{cases} \frac{1}{2}(h_K + h_L) \text{ for } \sigma = K|L \in \mathcal{E}_{int}, \\ h_K \text{ for } \sigma \in \mathcal{E}_{ext} \cap \mathcal{E}(K). \end{cases} \tag{17}$$

This choice is important to preserve steady states, see Lemma 2.

§4. Properties of the scheme

The scheme (7) enjoys some interesting properties, which we now state. First of all, thanks to the upwind choice for h^n in (1a), the positivity of the water height is preserved under a CFL like condition.

Lemma 1 (Positivity of the water height)**.** *Let $n \in [\![0, N-1]\!]$, let $(h_K^n,\ u_{i,\sigma}^n)_{K\in\mathcal{M},\ \sigma\in\mathcal{E}^{(i)}}$ be given and such that $h_K^n \geq 0$, for all $K \in \mathcal{M}$, and let h_K^{n+1} be computed by* (7b)*. Then $h_K^{n+1} \geq 0$, for all $K \in \mathcal{M}$ under the following CFL condition,*

$$\delta t \leq \frac{|K|}{\displaystyle\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,|\boldsymbol{u}_{K,\sigma}^n|}. \tag{18}$$

Second, thanks to the choice (17) for the reconstruction of the water height, the "lake at rest" steady state is preserved by the scheme.

Lemma 2 (Steady state "lake at rest")**.** *Let $n \in [\![0, N-1]\!]$, $C \in \mathbb{R}_+$; let $\boldsymbol{u}^{n+1} \in \boldsymbol{H}_{\mathcal{E},0}$ and $h^{n+1} \in L_{\mathcal{M}}$ be a solution to* (7b)-(7d) *with $\boldsymbol{u}^n = 0$ and $h^n + z = C$, where C is a given real number. Then $\boldsymbol{u}^{n+1} = 0$ and $h^{n+1} + z = C$.*

As a consequence of the careful discretisation of the convection term, the scheme satisfies a discrete kinetic energy balance, as stated in the following lemma. The proof of this result is an easy adaptation of [10, Lemma 3.2].

Lemma 3 (Discrete kinetic balance)**.** *A solution to the scheme* (7) *satisfies the following equality, for $i = 1, 2$, $\sigma \in \mathcal{E}^{(i)}$ and $0 \leq n \leq N-1$:*

$$\frac{1}{2\,\delta t}(h_{D_\sigma}^{n+1}(u_{i,\sigma}^{n+1})^2 - h_{D_\sigma}^n(u_{i,\sigma}^n)^2) + \frac{1}{2\,|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} F_{\sigma,\epsilon}^n(u_{i,\epsilon}^n)^2 + u_{i,\sigma}^{n+1}(\eth_i p^{n+1})_\sigma + g\,h_{\sigma,c}^{n+1}\,u_{i,\sigma}^{n+1}(\eth_i z)_\sigma = -R_{i,\sigma}^{n+1}, \tag{19}$$

with $R_{i,\sigma}^{n+1} \geq 0$ under the CFL like restriction:

$$\forall\sigma \in \mathcal{E}^{(i)}, \qquad \delta t \leq \frac{|D_\sigma|\,h_{D_\sigma}^{n+1}}{\displaystyle\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)} (F_{\sigma,\epsilon}^n)^-}. \tag{20}$$

The scheme also satisfies the following potential energy balance [10, Lemma 3.3].

Lemma 4 (Discrete elastic potential balance)**.** *Let, for $K \in \mathcal{M}$ and $0 \leq n \leq N$ the potential energy be defined by $(E_p)_K^n = \frac{1}{2}g\,(h_K^n)^2$. A solution to the scheme* (7) *satisfies the following equality, for $K \in \mathcal{M}$ and $0 \leq n \leq N-1$:*

$$\eth_t E_p^{n+1} + \mathrm{div}_K(E_p^n \boldsymbol{u}^n) + p_K^n \mathrm{div}_K(\boldsymbol{u}^n) = -R_K^{n+1}, \tag{21}$$

with

$$R_K^{n+1} \geq \frac{1}{|K|} g \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,u_{K,\sigma}^n h_\sigma^n (h_K^{n+1} - h_K^n). \tag{22}$$

Note that the right-hand side of Equation (22) may be negative, and thus the quantities R_K^{n+1} also. This is specific to explicit schemes (for implicit or pressure-correction schemes [9], this residual is non-negative) and prevents getting a stability estimate for the scheme.

However, combining the two previous lemmas allows to prove that convergent sequences of solutions to the scheme satisfy an entropy inequality, as depicted in the next section. To this purpose, we will pass to the limit in a discrete entropy balance which is built as follows. Let $K \in \mathcal{M}$ and let us denote by $(E_k)_K^n$ the following quantity, which may be seen as a kinetic energy associated to K:

$$(E_k)_K^n = \frac{1}{4\,|K|} \sum_{i=1}^{2} \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}^{(i)}} |D_\sigma|\, h_{D_\sigma}^n (u_{i,\sigma}^n)^2.$$

Then, for $\sigma_0 \in \mathcal{E}(K)$, we define a kinetic energy flux, which we denote by G_{K,σ_0}^n, as follows. Let us suppose, for instance, that $\sigma_0 \in \mathcal{E}^{(1)}$. We denote by ϵ the face of D_{σ_0} parallel to σ_0 and included in K and by ϵ' the opposite face of D_{σ_0}. In addition, σ_0 is the union of two half-faces of the dual mesh associated to the second component of the velocity, which we denote by τ and τ', and we denote by σ and σ' the two faces of K belonging to $\mathcal{E}^{(2)}$ such that $\tau \in \widetilde{\mathcal{E}}(D_\sigma)$ and $\tau' \in \widetilde{\mathcal{E}}(D_{\sigma'})$. We then have:

$$G_{K,\sigma_0}^n = \frac{1}{4}\Big[-F_{\sigma_0,\epsilon}^n\,(u_{1,\epsilon}^n)^2 + F_{\sigma_0,\epsilon'}^n\,(u_{1,\epsilon'}^n)^2 + F_{\sigma,\tau}^n\,(u_{2,\tau}^n)^2 + F_{\sigma,\tau'}^n\,(u_{2,\tau'}^n)^2\Big]$$

Multiplying the kinetic energy balance equation (19) associated to each face σ of K by $\frac{1}{2}\,|D_\sigma|$ and summing the four obtained relations with (21), we get

$$\frac{|K|}{\delta t}\left[(E_k)_K^{n+1} + (E_p)_K^{n+1} - (E_k)_K^n - (E_p)_K^n\right] + \sum_{\sigma \in \mathcal{E}(K)} \left[G_{K,\sigma}^n + F_{K,\sigma}^n\,(E_p)_\sigma^n\right]$$
$$+ \sum_{\sigma \in \mathcal{E}(K),\ \sigma = K|L} |\sigma|\,\frac{1}{2}\,(p_L^{n+1} - p_K^{n+1})\,u_{K,\sigma}^{n+1} + \sum_{\sigma \in \mathcal{E}(K),\ \sigma = K|L} |\sigma|\,\frac{1}{4} g\,(h_K^{n+1} + h_L^{n+1})\,(z_L - z_K)\,u_{K,\sigma}^{n+1} = -T_K^{n+1},$$

where T_K^{n+1} collects the residual terms in (19) and (21), and thus $T_K^{n+1} \geq R_K^{n+1}$. We now remark that, thanks to the discrete mass balance equation and the fact that the topography does not depend on time,

$$\frac{1}{2} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^n (z_L - z_K) = \frac{|K|}{\delta t}\,(h_K^{n+1} z_K - h_K^n z_K) + \frac{1}{2} \sum_{\sigma \in \mathcal{E}(K),\ \sigma = K|L} F_{K,\sigma}^n\,(z_K + z_L),$$

and we finally obtain the following discrete entropy balance:

$$\frac{|K|}{\delta t}\left[(E_k)_K^{n+1} + (E_p)_K^{n+1} + g\,h_K^{n+1} z_K - (E_k)_K^n - (E_p)_K^n - g\,h_K^n z_K\right]$$
$$+ \sum_{\sigma \in \mathcal{E}(K)} \left[G_{K,\sigma}^n + F_{K,\sigma}^n\,(E_p)_\sigma^n + \frac{1}{2}\,F_{K,\sigma}^n (z_K + z_L)\right]$$
$$+ \sum_{\sigma \in \mathcal{E}(K),\ \sigma = K|L} |\sigma|\,\frac{1}{2}\,(p_K^{n+1} + p_L^{n+1})\,u_{K,\sigma}^{n+1} = -(R_e)_K^{n+1}, \quad (23)$$

with

$$(R_e)_K^{n+1} \geq T_K^{n+1} + g \sum_{\sigma\in\mathcal{E}(K)} [\frac{1}{2}F_{K,\sigma}^n - \frac{1}{4}|\sigma|\,(h_K^{n+1} + h_L^{n+1})\,u_{K,\sigma}^{n+1}]\,(z_L - z_K) + \sum_{\sigma\in\mathcal{E}(K),\ \sigma=K|L} |\sigma|\,\frac{1}{2}\,(p_K^{n+1}u_{K,\sigma}^{n+1} - p_K^n u_{K,\sigma}^n). \quad (24)$$

§5. Consistency analysis

The objective of this section is to show that the schemes are consistent in the Lax-Wendroff sense, namely that if a sequence of solutions is controlled in suitable norms and converges to a limit, this latter necessarily satisfies a weak formulation of the continuous problem.

A weak solution to the continuous problem satisfies, for any $\varphi \in C_c^\infty(\Omega\times[0,T))$ ($\boldsymbol{\varphi} \in C_c^\infty(\Omega\times[0,T))^2$):

$$\int_0^T\int_\Omega \Big[h\,\partial_t\varphi + h\,u\,\cdot\nabla\varphi\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega h_0(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0, \quad (25a)$$

$$-\int_0^T\int_\Omega \Big[h\,\boldsymbol{u}\cdot\partial_t\boldsymbol{\varphi} + (h\boldsymbol{u}\otimes\boldsymbol{u}) : \boldsymbol{\varphi} + \frac{1}{2}\,g\,h^2\mathrm{div}(\boldsymbol{\varphi}) + g\,h\,\nabla(z)\boldsymbol{\varphi}\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \quad (25b)$$
$$-\int_\Omega h_0(\boldsymbol{x})\,\boldsymbol{u}_0(\boldsymbol{x})\cdot\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0.$$

This system is supplemented with a weak entropy inequality, for any nonnegative test functions $\varphi \in C_c^\infty(\Omega\times[0,T),\mathbb{R}_+)$:

$$-\int_0^T\int_\Omega \Big[\eta\,\partial_t\varphi + \boldsymbol{\Phi}\cdot\nabla\varphi\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\mathbf{t} - \int_\Omega \eta_{\mathbf{0}}(\boldsymbol{x})\,\varphi(\boldsymbol{x},\mathbf{0})\,\mathrm{d}\boldsymbol{x} \leq \mathbf{0}, \quad (26)$$

with η and $\boldsymbol{\Phi}$ defined by (4).

Before stating the global weak consistency of the scheme (7), some definitions and estimate assumptions are needed.

Let $(\mathcal{M}^{(m)},\mathcal{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes in the sense of Definition 1 and let $(h^{(m)}\ \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be the associated sequence of solutions of the scheme (7)).

Assumed estimates - We need also some a priori estimates on the sequence of discrete solutions $(h^{(m)},\ \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ in order to prove the consistency result we are seeking. First of all we assume that $h^{(m)} > 0$, $\forall m \in \mathbb{N}$ which can be obtained under the CFL condition (18). Furthermore:

– The water height $h^{(m)}$ and its inverse are uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))$, *i.e.*there exists some constants $C, C' \in \mathbb{R}_+^*$ such that for $m \in \mathbb{N}$ and $0 \leq n < N^{(m)}$:

$$1/C < (h^{(m)})_K^n \leq C, \quad 1/C' < 1/(h^{(m)})_K^n \leq C' \quad \forall K \in \mathcal{M}^{(m)} \quad (27)$$

– The velocity $\boldsymbol{u}^{(m)}$ is also uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))^2$:

$$|(\boldsymbol{u}^{(m)})^n_\sigma| \le C, \quad \forall \sigma \in \mathcal{E}^{(m)}. \tag{28}$$

Finally, the weak consistency to the entropy inequality is only proved under additional assumptions. First we need the following condition on the space and time steps, which is stronger than a CFL condition:

$$\frac{\delta t^{(m)}}{\delta_{\mathcal{M}^{(m)}}} \to 0 \text{ as } m \to +\infty \tag{29}$$

Second, the $L^1(\Omega, BV)$ norm of the height is required to be bounded, *i.e.* there exists one constant C such that, for $m \in \mathbb{N}$,

$$\sum_{n=0}^{N-1} \sum_{K\in\mathcal{M}} |K|\, |(h^{(m)})^{n+1}_K - (h^{(m)})^n_K| \le C. \tag{30}$$

We are now in position to state the following consistency result.

Theorem 5 (Weak consistency of the scheme)**.** *Let $(\mathcal{M}^{(m)}, \mathcal{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\delta t^{(m)}$ and $\delta_{\mathcal{M}^{(m)}} \to 0$ as $m \to +\infty$; assume that there exists $\eta > 0$ such that $\eta_{\mathcal{M}^{(m)}} \le \eta$ for any $m \in \mathbb{N}$ (with $\eta_{\mathcal{M}^{(m)}}$ defined by* (5)*); assume moreover that (27) and (28) hold. Let $(h^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be a sequence of solutions to the scheme* (7) *converging to $(\bar h, \bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$. Then $(\bar h, \bar{\boldsymbol{u}})$ satisfies the weak formulation* (25) *of the shallow water equations.*

If we furthermore assume the space and time steps satisfy (29) *and that the sequence of heights is uniformly bounded in $L^1(\Omega, BV)$, i.e. satisfy* (30)*, then $(\bar h, \bar{\boldsymbol{u}})$ satisfies the entropy inequality* (26)*.*

Proof. The proof is obtained by passing to the limit in the scheme and in the discrete entropy balance (23), using the tool of [6] (or, more precisely speaking, simplified versions of these tools adapted to Cartesian grids). The additional assumptions required for the entropy condition are used to prove that the residual term appearing in the discrete potential energy balance, given by (22), tends to zero. □

§6. Numerical results

We now assess the behaviour of the scheme on some numerical experiments. The computations presented here are performed with the CALIF^{3}S free software developed at IRSN [4].

6.1. Rotation in a paraboloid

This first test case consists in calculating the uniform rotation of a circular drop on a support of parabolic shape (see Figure 3). The computational domain is $(0, L)\times(0, L)$ and the elevation of the support is:

$$z = -h_0\,(1 - (x - \frac{L}{2})^2 - (y - \frac{L}{2})^2),$$

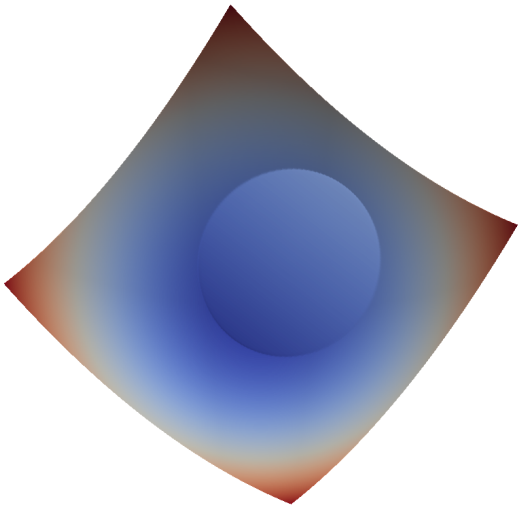


Figure 3: Sloshing of a drop on a parabolic support – State obtained after one revolution (very close to the initial state).

with $L = 4$ and $h_0 = 0.1$. The fluid height is given by

$$h = h_0 \ \max(0,\ (x - \frac{L}{2})\ \cos(\omega t) + (y - \frac{L}{2})\ \sin(\omega t) - z - 0.5),$$

and the velocity is

$$\boldsymbol{u} = \frac{1}{2}\,\omega \begin{bmatrix} -\sin(\omega t) \\ \cos(\omega t) \end{bmatrix}.$$

It is then easy to check that the mass and momentum balance equations are verified provided that $\omega^2 = 2\,g\,h_0$. The solution is thus regular, and this test features a regular topography and dry zones (*i.e.* zones where $h = 0$). We compare the numerical and theoretical height obtained after one rotation (*i.e.* $\omega t = 2\pi$), for different uniform grids and with a time step $\delta t = \delta x/8$. (the maximal speed of sound and the maximal velocity are both close to 1); results are gathered in the following table:

grid	error (discrete L^1 norm)
100×100	$3.02\ 10^{-3}$
200×200	$1.54\ 10^{-3}$
400×400	$0.896\ 10^{-3}$
800×800	$0.511\ 10^{-3}$

We observe an order of convergence between 0.8 and 1, which is consistent with a first-order approximation of the fluxes and the time derivative.

6.2. A dam-break problem

In this test, the computational domain is:

$$\Omega = (0, 200) \times (0, 200) \setminus \Omega_w \text{ with } \Omega_w = (95, 105) \times (0, 95) \cup (95, 170) \times (0, 200).$$

The fluid is supposed to be initially at rest, and the initial height is $h = 10$ for $x_1 \leq 100$ and $h = 5$ for $x_1 > 100$. A zero normal velocity is prescribed at all the boundaries of the computational domain. The computation is performed with a mesh obtained from a 1000×1000 regular grid, by removing the cells included in Ω_w. The time step is $\delta t = \delta x/25$ (the maximal speed of sound and the maximal velocity are both close to 10). The obtained fluid height is shown at different times on Figure 4; they confirm the efficiency of the scheme, and its capability to deal with reflexion phenomena very simply (*i.e.* just by setting the normal velocity at the boundary to zero, by contrast with schemes based on Riemann solvers which need to implement fictitious cells techniques).

Acknowledgements

The authors would like to thank Robert Eymard and Thierry Gallouët for several interesting discussions.

References

[1] Arakawa, A., and Lamb, V. A potential enstrophy and energy conserving scheme for the shallow water equations. *Monthly Weather Review 109* (1981), 18–36.

[2] Bonaventura, L., and Ringler, T. Analysis of discrete shallow-water models on geodesic delaunay grids with c-type staggering. *Monthly Weather Review 133*, 8 (2005), 2351–2373.

[3] Bouchut, F. *Nonlinear Stability of finite volume methods for hyperbolic conservation laws*. Birkhauser, 2004.

[4] CALIF^{3}S. A software components library for the computation of fluid flows. `https://gforge.irsn.fr/gf/project/califs`.

[5] Doyen, D., and Gunawan, H. An explicit staggered finite volume scheme for the shallow water equations. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, vol. 77 of *Springer Proc. Math. Stat.* Springer, Cham, 2014, pp. 227–235.

[6] Gallouët, T., Herbin, R., and Latché. On the weak consistency of finite volumes schemes for conservation laws on general meshes. *under revision* (2019). Available from: `https://hal.archives-ouvertes.fr/hal-02055794`.

[7] Gallouët, T., Herbin, R., Latché, J.-C., and Mallem, K. Convergence of the marker-and-cell scheme for the incompressible Navier-Stokes equations on non-uniform grids. *Foundations of Computational Mathematics 18* (2018), 249–289.

[8] Gunawan, H. *Numerical simulation of shallow water equations and related models*. PhD thesis, Université Paris-Est and Institut Teknologi Bandung, 2015.

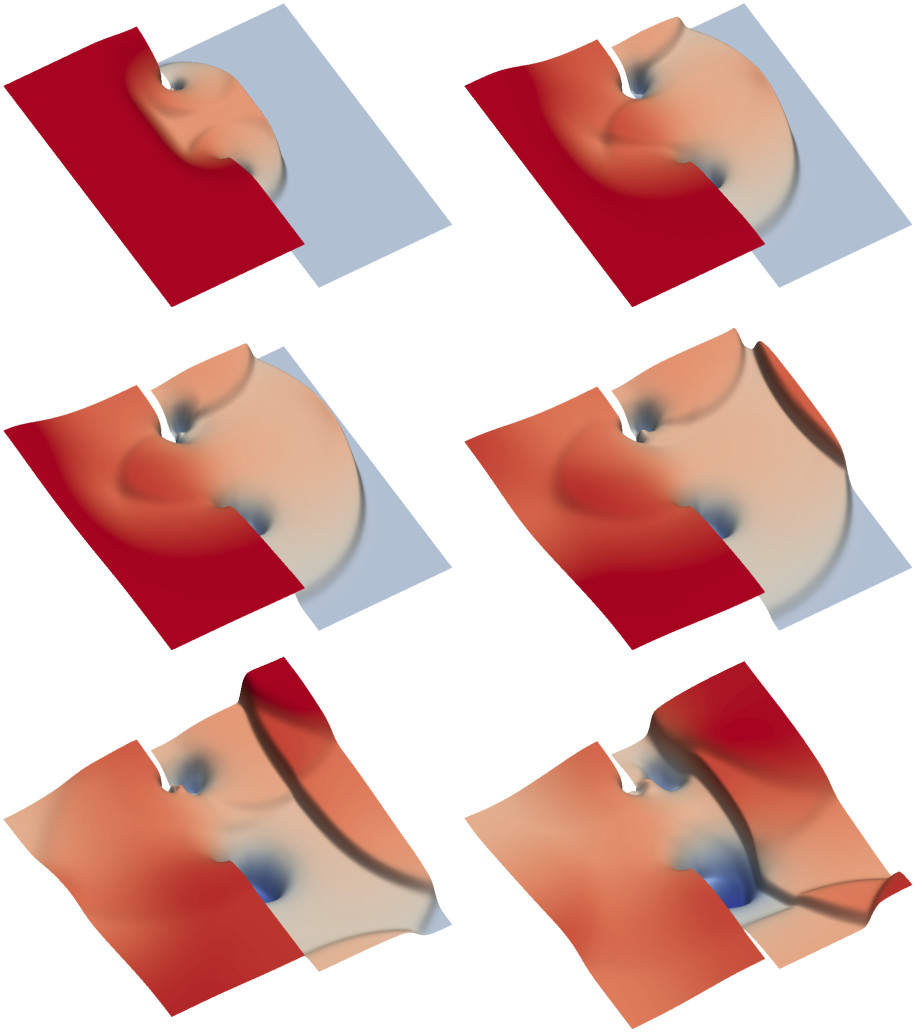


Figure 4: Partial dam break – Height obtained at $t = 4$, $t = 8$, $t = 10$, $t = 12$, $t = 16$ and $t = 20$ with a mesh obtained (by supression of the zones associated to the obstacles) from a 1000×1000 regular grid. In the last Figure ($t = 20$), the obtained minimal and maximal heights are $h = 2.149$ and $h = 9.306$ respectively.

[9] Herbin, R., Kheriji, W., and Latché, J.-C. On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis 48* (2014), 1807–1857.

[10] Herbin, R., Latché, J.-C., and Nguyen, T. Explicit staggered schemes for the compressible Euler equations. *ESAIM: Proceedings 40* (2013), 83–102.

[11] Herbin, R., Latché, J.-C., and Nguyen, T. Consistent segregated staggered schemes with explicit steps for the isentropic and full Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis 52* (2018), 893–944.

[12] Stelling, G., and Duinmeijer, S. A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. *International Journal for Numerical Methods in Fluids 43* (2003), 1329–1354.

R. Herbin and Y. Nasseri
Aix-Marseille Université, Institut de Mathématiques de Marseille,
39 rue Joliot Curie
13453 Marseille
raphaele.herbin@univ-amu.fr and youssouf.nasseri@univ-amu.fr

J.-C. Latché
Institut de Radioprotection et Sûreté Nucléaire,
13115, Saint-Paul-lez-Durance
jean-claude.latche@irsn.fr

N. Therme
CEA/CESTA
33116, Le Barp, France
nicolas.therme@cea.fr

Monografías Matemáticas *García de Galdeano* **42**, 143–152 (2019)

STABILIZED VIRTUAL ELEMENT METHOD FOR THE INCOMPRESSIBLE NAVIER-STOKES EQUATIONS

Diego Irisarri and Guillermo Hauke

Abstract. In this work, we present a discretization for the incompressible Navier-Stokes equations based on the stabilized virtual element method (VEM). Basically, VEM can be considered a generalization of FEM that enables a polynomial decomposition of the domain. In this work, the concepts of stabilized methods are introduced in the VEM formulation. Thus, stabilization terms are included in the variational form to circumvent the Babuška-Brezzi condition and to stabilize the solution for convection dominated flows. Numerical examples are presented to show the behavior of the method.

Keywords: Virtual element methods, Navier-Stokes problem, stabilized methods.
AMS classification: 76D05, 65M60.

§1. Introduction

The virtual element method (VEM) can be considered a generalization of the finite element method (FEM) that allows a greater versatility in the partition of the domain. The basis of VEM was established in [4, 5, 12]. Many works related to VEM have been published both in the field of elasticity [6, 13, 19] and fluid mechanics [23, 10, 8, 9].

In this work, we address the stabilized VEM formulation for incompressible Navier-Stokes equations. The VEM has already been applied to the Stokes problem [2, 8] and the Navier-Stokes equations [9]. It is well known that the space for the velocity and pressure cannot be selected arbitrarily since the Babuška-Brezzi condition or inf-sup condition must be satisfied. However, stabilization terms can be introduced in the discretization in order to circumvent the inf-sup condition. In this work, the concepts of stabilized methods [15, 21, 22, 18, 26, 20, 16] are introduced in the VEM formulation for Navier-Stokes equations. This formulation enables to select the velocity and pressure spaces with equal order interpolation functions. Thus, stabilization terms are included in the variational form to circumvent the Babuška-Brezzi condition and to stabilize the solution for convection dominated flows. We consider the transient incompressible Navier-Stokes using a semi-discrete scheme (see, for instance, [18, 17]).

§2. The incompressible Navier-Stokes equations

The problem is defined on a bounded domain $\Omega \subset \mathbb{R}^N$, $N = 2, 3$. The boundary is partitioned into two non-overlapping zones Γ_g and Γ_h such that $\Gamma_g \cup \Gamma_h = \Gamma$ and $\Gamma_g \cap \Gamma_h = \oslash$.

Let us set up the unsteady incompressible Navier-Stokes equations, given by

$$\begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} + (\nabla \boldsymbol{u})\boldsymbol{u} - 2\nu\nabla\cdot\boldsymbol{\varepsilon}(\boldsymbol{u}) + \dfrac{1}{\rho}\nabla p = \boldsymbol{f} & \text{in} \quad \Omega\times(0,T) \\ \nabla\cdot\boldsymbol{u} = 0 & \text{in} \quad \Omega\times(0,T) \\ \boldsymbol{u} = \mathbf{g} & \text{on} \quad \Gamma_g\times(0,T) \\ 2\nu\boldsymbol{\varepsilon}(\boldsymbol{u})\boldsymbol{n} = \boldsymbol{h} & \text{on} \quad \Gamma_h\times(0,T) \\ \boldsymbol{u} = \boldsymbol{u}_0 & \text{in} \quad \Omega \text{ at } t=0 \end{cases} \tag{1}$$

where $\boldsymbol{u}$ and p are the unknown velocity and pressure, respectively. ρ is the fluid density, ν represents the kinematic viscosity, $\boldsymbol{f}$ is the source term.

The tensor $\boldsymbol{\varepsilon}(\boldsymbol{u})$ is the symmetric part of the velocity gradient and is defined as

$$\varepsilon_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}) \text{ for } i,j = 1,...,n_{sd} \tag{2}$$

where n_{sd} is the number of spatial dimensions, i.e., $n_{sd} = 2$ for 2D and $n_{sd} = 3$ for 3D.

2.1. Variational formulation

Firstly, we define the spaces for the test and trial functions,

$$\begin{aligned} \mathcal{V} &= \{\boldsymbol{v}(\cdot,t) \in H^1(\Omega)^{n_{\text{sd}}}, t\in[0,T] \;\mid\; \boldsymbol{v}(\cdot,t) = \mathbf{0} \;\text{ on }\; \Gamma_g\} \\ \mathcal{S} &= \{\boldsymbol{u}(\cdot,t) \in H^1(\Omega)^{n_{\text{sd}}}, t\in[0,T] \;\mid\; \boldsymbol{u}(\cdot,t) = \mathbf{g} \;\text{ on }\; \Gamma_g\} \\ \mathcal{P} = \mathcal{Q} &= \{q(\cdot,t) \in L^2(\Omega)\cap H^1(\Omega), t\in[0,T] \;\text{ s.t. }\; \int_\Omega q(\cdot,t)\mathrm{d}\Omega = 0\} \end{aligned}$$

The variational formulation is defined as: Find $\boldsymbol{u}\in\mathcal{S}$ and $p\in\mathcal{P}$ such that

$$B(\boldsymbol{u},p;\boldsymbol{v},q) = F(\boldsymbol{v},q),\ (\boldsymbol{v},q)\in\mathcal{V}\times\mathcal{Q} \tag{4}$$

with

$$B(\boldsymbol{u},p;\boldsymbol{v},q) = d(\boldsymbol{u},\boldsymbol{v}) + a(\boldsymbol{u},\boldsymbol{v}) + c(\boldsymbol{u};\boldsymbol{u},\boldsymbol{v}) + b_m(p,\boldsymbol{v}) + b_c(\boldsymbol{u},q) \tag{5}$$

where $d(\cdot,\cdot)$, $a(\cdot,\cdot)$, $b_m(\cdot,\cdot)$, $b_c(\cdot,\cdot)$ are bilinear forms and $c(\cdot;\cdot,\cdot)$ is the trilinear form that represents the convective term,

$$\begin{aligned} &d(\boldsymbol{u},\boldsymbol{v}) = \left(\frac{\partial\boldsymbol{u}}{\partial t},\boldsymbol{v}\right), \quad a(\boldsymbol{u},\boldsymbol{v}) = (\nu\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v})), \quad c(\boldsymbol{w};\boldsymbol{u},\boldsymbol{v}) = ((\nabla\boldsymbol{u})\boldsymbol{w},\boldsymbol{v}) \\ &b_m(p,\boldsymbol{v}) = (\nabla p,\boldsymbol{v}), \quad b_c(\boldsymbol{u},q) = (\nabla\cdot\boldsymbol{u},q) \end{aligned} \tag{6}$$

and

$$F(\boldsymbol{v},q) = (f,\boldsymbol{v}) + (\boldsymbol{h},\boldsymbol{v})_{\Gamma_h} \tag{7}$$

§3. Virtual element method discretization

In this section, we show the VEM discretization of the variational form (4) using a first order approximation. The domain Ω is decomposed into a partition $\mathcal{T}_h$ composed of polygons K, and let $\mathcal{E}_h$ be the set of edges e of $\mathcal{T}_h$. Let $\widetilde{\Omega}$ denote the union of the polygons, $\widetilde{\Omega} = \bigcup_{e=1}^{n_{\text{el}}} K$ where n_{el} is the number of polygons. In this work, linear elements are employed. We define the following initial local space defined on each element:

$$\widetilde{V}_h(K) := \{v \in C^0(K) : \ v_{|e} \in \mathbb{P}_1(e) \ \forall \ e \subset \partial K, \Delta v \in \mathbb{P}_1(K)\},$$

where $\mathbb{P}_1(K)$ are the polynomials of degree 1 on the polygon K. In $\widetilde{V}_h(K)$, we can take the values of $v \in \widetilde{V}_h(K)$ at the vertices as degrees of freedom, *dof*. Then, the number of degrees of freedom in K is equal to the number of vertices N^V.

We define the following projectors in K:

- the H^1-seminorm projection $\Pi_1^{\nabla,K} : [\widetilde{V}_h(K)]^{n_{sd}} \ \rightarrow \ [\mathbb{P}_1(K)]^{n_{sd}}$,

$$\int_K \nabla(\Pi_1^{\nabla}\boldsymbol{v} - \boldsymbol{v}) : \nabla \boldsymbol{p}_1 \, dx = \mathbf{0} \ \text{ and } \ \int_{\partial K} (\Pi_1^{\nabla}\boldsymbol{v} - \boldsymbol{v}) \, \mathrm{d}s = \mathbf{0} \quad \forall p_1 \in \mathbb{P}_1, \tag{8}$$

- the L^2-projection for scalar functions $\Pi_k^{0,K}$: $\widetilde{V}_h \rightarrow \ \mathbb{P}_k(K)$ is defined locally as

$$\int_K (v - \Pi_k^0 v) \, p_k \, dx = 0 \quad \forall p_k \in \mathbb{P}_k \quad \text{for} \quad k = 0 \quad \text{and} \quad k = 1. \tag{9}$$

We can now introduce the local Virtual Element space:

$$V_h(K) := \{v \in \widetilde{V}_h(K) : \ \int_K v \, p_1 \, dx = \int_K \Pi_1^{\nabla} v \, p_1 \, dx \ \forall \ p_1 \in \mathbb{P}_1(K)\}. \tag{10}$$

The dimension of $V_h(K)$ is $N_{\text{dof}} = N^V$ as the same as the degrees of freedom which are unisolvent with respect to $V_h(K)$ [1].

The global virtual spaces defined for the unknown variables of the discrete problem are

$$V_h^u := \{\boldsymbol{v} \in [H^1(\Omega)]^{n_{sd}} : \ \boldsymbol{v}_{|K} \in [V_h(K)]^{n_{sd}} \ \forall K \in \mathcal{T}_h\} \tag{11}$$

$$Q_h := \{q \in H^1(\Omega) \ \text{ s. t. } \int_\Omega q \, \mathrm{d}\Omega = 0 : \ q_{|K} \in V_h(K) \ \forall K \in \mathcal{T}_h\}. \tag{12}$$

The basis functions on each element K, $\varphi_i \in \widetilde{V}_h(K)$, are defined, as happens in FEM, as the canonical basis functions, $\text{dof}_i(\varphi_j) = \delta_{ij}$ for $i, j = 1, ..., N_{\text{dof}}$. We recall that the basis functions for the velocity and pressure are the same. Thus, the unknown variables $(\boldsymbol{u}_h, p_h)$ are expressed as a linear combination of these basis functions,

$$\boldsymbol{u}_h = \sum_{i=1}^{N_{\text{dof}}} \text{dof}_i(\boldsymbol{u}_h)\varphi_i \qquad p_h = \sum_{i=1}^{N_{\text{dof}}} \text{dof}_i(p_h)\varphi_i. \tag{13}$$

The Galerkin formulation reads: Find $(\boldsymbol{u}_h, p_h) \in V_h^u \times Q_h$ such that

$$B(\boldsymbol{u}_h, p_h; \boldsymbol{v}_h, q_h) = F(\boldsymbol{v}_h, q_h), \text{ for all } (\boldsymbol{v}_h, q_h) \in \mathcal{V}_h \times \mathcal{P}_h \tag{14}$$

with

$$B(\boldsymbol{u}_h, p_h; \boldsymbol{v}_h, q_h) = \quad d(\boldsymbol{u}_h, \boldsymbol{v}_h) + a(\boldsymbol{u}_h, \boldsymbol{v}_h) + c(\boldsymbol{u}_h; \boldsymbol{u}_h, \boldsymbol{v}_h) + b_m(p_h, \boldsymbol{v}) + b_c(\boldsymbol{u}_h, q_h) \tag{15}$$

where the bilinear forms $d(\cdot,\cdot)$, $a(\cdot,\cdot)$, $b_m(\cdot,\cdot)$, $b_c(\cdot,\cdot)$ and the trilinear form $c(\cdot;\cdot,\cdot)$ are

$$\begin{aligned}
d(\boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \sum_K d^K(\boldsymbol{u}_h, \boldsymbol{v}_h) = \sum_K (\frac{\partial \boldsymbol{u}_h}{\partial t}, \boldsymbol{v}_h)_K \\
a(\boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \sum_K a^K(\boldsymbol{u}_h, \boldsymbol{v}_h) = \sum_K (\nu \boldsymbol{\varepsilon}(\boldsymbol{u}_h), \boldsymbol{\varepsilon}(\boldsymbol{v}_h))_K \\
c(\boldsymbol{u}_h; \boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \sum_K c^K(\boldsymbol{u}_h; \boldsymbol{u}_h, \boldsymbol{v}_h) = \sum_K ((\nabla \boldsymbol{u}_h)\boldsymbol{u}_h, \boldsymbol{v}_h)_K \\
b_m(p_h, \boldsymbol{v}_h) = &\quad \sum_K b_m^K(p_h, \boldsymbol{v}_h) = \sum_K (\nabla p_h, \boldsymbol{v}_h)_K \\
b_{c,h}(\boldsymbol{u}_h, q_h) = &\quad \sum_K b_c^K(\boldsymbol{u}_h, q_h) = \sum_K (\nabla \cdot \boldsymbol{u}, q_h)_K.
\end{aligned} \tag{16}$$

The discrete terms belonging to $B(\cdot,\cdot)$ are computable using the projector operators and the degrees of freedom. Thus, we define the approximate bilinear and trilinear forms:

$$\begin{aligned}
a_h^K(\boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \int_K \nu \Pi_0^0 \nabla \boldsymbol{u}_h : \Pi_0^0 \nabla \boldsymbol{v}_h \mathrm{d}\Omega + \mathcal{S}_\nu^K((I - \Pi_1^\nabla)\boldsymbol{u}_h, (I - \Pi_1^\nabla)\boldsymbol{v}_h) \\
d_h^K(\boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \int_K \frac{\partial}{\partial t} \Pi_1^0 \boldsymbol{u}_h \cdot \Pi_1^0 \boldsymbol{v}_h \mathrm{d}\Omega + \mathcal{S}_t^K((I - \Pi_1^0)\boldsymbol{u}_h, (I - \Pi_1^0)\boldsymbol{v}_h) \\
c_h^K(\boldsymbol{u}_h; \boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \int_K [(\Pi_0^0 \nabla \boldsymbol{u}_h)(\Pi_1^0 \boldsymbol{u}_h)] \cdot \Pi_1^0 \boldsymbol{v}_h \, \mathrm{d}\Omega \\
b_{m,h}^K(\boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \int_K \Pi_0^0 \nabla p_h \cdot \Pi_1^0 \boldsymbol{v}_h \, \mathrm{d}\Omega \\
b_{c,h}^K(\boldsymbol{u}_h, \boldsymbol{v}_h) = &\quad \int_K (\Pi_0^0 \nabla \cdot \boldsymbol{u}_h)(\Pi_1^0 q_h) \, \mathrm{d}\Omega
\end{aligned} \tag{17}$$

where the VEM-stabilization terms $\mathcal{S}_\alpha^K$ are necessary for stability [1] and will be explained later.

In this work, the stabilized VEM that is proposed uses a linear approximation ($k = 1$) both for the velocity and the pressure. Thus, the degrees of freedom of pressure and velocity are the values at the vertices. We have followed the work of Franca et al. [18] to stabilize the VEM formulation. As it is well known, in stabilized methods additional terms are included in the Galerkin formulation that consist in weighting the residual by a determined differential operator (related to the differential equation) applied to the test functions. Besides, a generalized trapezoidal method is employed for the temporal term in order to reach the steady-state solutions and deal with the nonlinearity of the equations.

The stabilized VEM formulation includes additional terms to circumvent the Babuška-Brezzi condition and to obtain a stable solution for convection dominated flows. This formulation can be written as:

Find $(\boldsymbol{u}_h, p_h) \in V_h^u \times Q_h$ such that

$$B(\boldsymbol{u}_h, p_h; \boldsymbol{v}_h, q_h) + B^\tau(\boldsymbol{u}_h, p_h; \boldsymbol{v}_h, q_h) = \quad F(\boldsymbol{v}_h, q_h) + F^\tau(\boldsymbol{v}_h, q_h) \quad \text{for all } (\boldsymbol{v}_h, q_h) \in \mathcal{V}_h \times \mathcal{P}_h \tag{18}$$

where

$$\begin{aligned} B^\tau(\boldsymbol{u}_h, p_h; \boldsymbol{v}, q) \quad = \sum_{K \in \widetilde{\Omega}} \Big(& (\frac{\partial \boldsymbol{u}_h}{\partial t} + (\nabla \boldsymbol{u}_h)\boldsymbol{u}_h + \nabla p_h - 2\nu \nabla \cdot \boldsymbol{\varepsilon}(\boldsymbol{u}_h), \\ & \tau((\nabla \boldsymbol{v}_h)\boldsymbol{u}_h + \nabla q_h \pm 2\nu \nabla \cdot \boldsymbol{\varepsilon}(\boldsymbol{u}_h))) + (\nabla \cdot \boldsymbol{u}_h, \delta \nabla \cdot \boldsymbol{v}_h) \Big)_K \end{aligned} \tag{19}$$

$$F^\tau(\boldsymbol{v}_h, q) = \sum_{K \in \widetilde{\Omega}} \Big(\boldsymbol{f}, \tau((\nabla \boldsymbol{v}_h)\boldsymbol{u}_h + \nabla q_h \pm 2\nu \nabla \cdot \boldsymbol{\varepsilon}(\boldsymbol{u}_h)) \Big)_K \tag{20}$$

where τ and δ are the stability parameters. They are taken from the work of Codina [16],

$$\tau = \left(\frac{c_1 \nu}{h^2} + \frac{c_2 \|\boldsymbol{u}_h\|_{L^\infty(K)}}{h} \right)^{-1} \qquad \delta = \frac{c_3 h^2}{\tau}. \tag{21}$$

The constants c_1, c_2 and c_3 are taken as $c_1 = 4$, $c_2 = 2$ and $c_3 = 1$. Other possibilities for non-regular elements can be found in [3]. The value of h (length of the element) is taken as $h = \sqrt{|K|}$, where $|K|$ is the area of the element.

We observe that the operators $B^\tau(\cdot,\cdot)$ and $F^\tau(\cdot)$ correspond to the stabilization terms. Since we only consider $k = 1$, the terms containing $\nabla \cdot \boldsymbol{\varepsilon}(\boldsymbol{u}_h)$ disappear because $\nabla \cdot \boldsymbol{\varepsilon}(\boldsymbol{u}_h) = 0$. The stabilized terms for the momentum and continuity equations are defined as follows.

· *Stabilized terms for the momentum equations*

$$\begin{aligned} \tau\left(\frac{\partial \boldsymbol{u}_h}{\partial t}, (\nabla \boldsymbol{v}_h)\boldsymbol{u}_h\right) \quad &= \tau \int_K [(\Pi_0^0 \nabla \boldsymbol{v}_h)(\Pi_1^0 \boldsymbol{u}_h)] \cdot \Pi_1^0 \frac{\partial \boldsymbol{u}_h}{\partial t} \, \mathrm{d}\Omega \\ \tau((\nabla \boldsymbol{u}_h)\boldsymbol{u}_h, (\nabla \boldsymbol{v}_h)\boldsymbol{u}_h) \quad &= \tau \int_K [(\Pi_0^0 \nabla \boldsymbol{v}_h)(\Pi_1^0 \boldsymbol{u}_h)] \cdot [(\Pi_0^0 \nabla \boldsymbol{v}_h)(\Pi_1^0 \boldsymbol{v}_h)] \, \mathrm{d}\Omega \\ \tau(\nabla p_h, (\nabla \boldsymbol{v}_h)\boldsymbol{u}_h) \quad &= \tau \int_K [(\Pi_0^0 \nabla \boldsymbol{v}_h)(\Pi_1^0 \boldsymbol{u}_h)] \cdot (\Pi_0^0 \nabla q_h) \, \mathrm{d}\Omega \\ \tau(\frac{\partial \boldsymbol{u}_h}{\partial t}, \nabla q_h) \quad &= \tau \int_K (\Pi_0^0 \nabla q_h) \cdot \Pi_1^0 \frac{\partial \boldsymbol{u}_h}{\partial t} \, \mathrm{d}\Omega \\ \tau((\nabla \boldsymbol{u}_h)\boldsymbol{u}_h, \nabla q_h) \quad &= \tau \int_K (\Pi_0^0 \nabla q_h) \cdot [(\Pi_0^0 \nabla \boldsymbol{v}_h)(\Pi_1^0 \boldsymbol{v}_h)] \, \mathrm{d}\Omega \\ \tau(\nabla p_h, \nabla q_h) \quad &= \tau \int_K (\Pi_0^0 \nabla q_h) \cdot (\Pi_0^0 \nabla p_h) + \mathcal{S}_p^K((I - \Pi_1^\nabla)p_h, (I - \Pi_1^\nabla)q_h) \, \mathrm{d}\Omega \end{aligned} \tag{22}$$

· *Stabilized terms for the continuity equation*

$$\delta(\nabla \cdot \boldsymbol{u}_h, \nabla \cdot \boldsymbol{v}_h) \quad = \delta \int_K (\Pi_0^0 \nabla \cdot \boldsymbol{v}_h)(\Pi_0^0 \nabla \cdot \boldsymbol{u}_h) \, \mathrm{d}\Omega \tag{23}$$

We observe that in the presented VEM formulation, there appear some terms called $\mathcal{S}_\alpha^K(\cdot,\cdot)$ which are the VEM-stabilization part, for $\alpha = \nu, t, p$. These terms are a peculiarity of VEM and they emerge from the projection of the basis functions. In order to compute the above mentioned matrices, we decompose the basis functions as $\varphi = \Pi\varphi + (I - \Pi)\varphi$. Therefore, we project the basis functions, $\Pi\varphi$, from the virtual space to a determined polynomial space. Thus, these terms that involve the projection of the variables, both Π_k^∇ and Π_k^0, can be computed exactly via numerical integration and they ensure consistency. However, a peculiarity of VEM is that the kernel of these projections, $(I - \Pi)\varphi$, must be considered for some terms to ensure the VEM stability [7, 4]. The terms $\mathcal{S}_\alpha^K(;\cdot)$ take into account the terms $(I - \Pi)\varphi$ which are not considered by the consistency part. The only condition is that $\mathcal{S}_\alpha^K(;\cdot)$ scales as the consistency part. In this case, it has been observed numerically that three stability terms must be considered.

The term $\mathcal{S}_\alpha^K(;\cdot)$ can be selected in different ways. A rigorous work on the stability term can be found in [7, 11]. In [27, 19] are proposed different definitions for the VEM stabilization term $\mathcal{S}^K$. The authors exploited the flexibility of selecting this term in order to improve the characteristics of the method. Here, we define them as follows:

- The diffusion term,

$$\begin{aligned}\mathcal{S}_\nu^K((I-\Pi_1^\nabla)\boldsymbol{u}_h,(I-\Pi_1^\nabla)\boldsymbol{v}_h) \quad &\approx \nu[(I-\Pi_1^\nabla)\overrightarrow{u}_h]^T[(I-\Pi_1^\nabla)\overrightarrow{v}_h^{M_x}] \\ &+\nu[(I-\Pi_1^\nabla)\overrightarrow{v}_h]^T[(I-\Pi_1^\nabla)\overrightarrow{v}_h^{M_y}]\end{aligned} \tag{24}$$

- The temporal term,

$$\begin{aligned}\mathcal{S}_t^K((I-\Pi_1^0)\boldsymbol{u}_h,(I-\Pi_1^0)\boldsymbol{v}_h) \quad &\approx h_K^2[(I-\Pi_1^0)\overrightarrow{u}_h]^T[(I-\Pi_1^0)\overrightarrow{v}_h^{M_x}] \\ &+h_K^2[(I-\Pi_1^0)\overrightarrow{v}_h]^T[(I-\Pi_1^0)\overrightarrow{v}_h^{M_y}]\end{aligned} \tag{25}$$

- The stability term, $\tau(\nabla q_h, \nabla p_h)$

$$\mathcal{S}_p^K((I-\Pi_1^\nabla)p_h,(I-\Pi_1^\nabla)q_h) \approx [(I-\Pi_1^\nabla)\overrightarrow{p}_h]^T[(I-\Pi_1^\nabla)\overrightarrow{q}_h] \tag{26}$$

with $\overrightarrow{u}_h$, $\overrightarrow{u}_h$ and $\overrightarrow{p}_h$ being the vector containing the degrees of freedom of u_h, v_h and p_h in the element K, respectively. That is to say,

$$u_h|_K = \sum_{i=1}^{N_{\text{dof,K}}}[\overrightarrow{u}_h]_i\varphi_i, \quad v_h|_K = \sum_{i=1}^{N_{\text{dof,K}}}[\overrightarrow{v}_h]_i\varphi_i \quad \text{and} \quad p_h|_K = \sum_{i=1}^{N_{\text{dof,K}}}[\overrightarrow{p}_h]_i\varphi_i, \tag{27}$$

where $N_{\text{dof,K}}$ are the degrees of freedom in K. Similarly, we have that $\overrightarrow{v}_h^{M_x}$, $\overrightarrow{v}_h^{M_y}$ and $\overrightarrow{q}_h$ are the degrees of freedom for the test function in the x-momentum equation, y-momentum equation and continuity equation.

Whereas the VEM-stabilization term in the diffusion $\mathcal{S}_\nu^K$ is the classical choice in VEM, see for instance [4], we have considered two more VEM stabilizing terms. The temporal term stabilization is only necessary to be considered when the temporal term is dominant. However it has been observed that it improves considerably the condition number of the matrix. On the other hand, the term $\mathcal{S}_p^K$ is very important in order to obtain a proper solution since it helps to penalize those non-physical oscillations of the pressure.

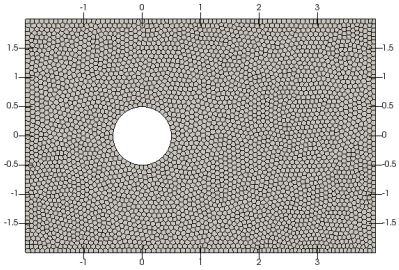


Figure 1: Domain dimensions and mesh.

As for the source term, it is approximated by

$$(\boldsymbol{f}_h, \boldsymbol{v}_h) = \sum_{K\in\mathcal{T}_h} (\boldsymbol{f}_h, \boldsymbol{v}_h)_K = \sum_{K\in\mathcal{T}_h} \int_K \Pi_1^0 \boldsymbol{f} \cdot \boldsymbol{v}_h \, \mathrm{d}\Omega = \sum_{K\in\mathcal{T}_h} \int_K \boldsymbol{f} \cdot \Pi_1^0 \boldsymbol{v}_h \, \mathrm{d}\Omega \tag{28}$$

where Eq. (28) expresses the RHS and it is computable using the degrees of freedom.

In the discrete problem, there are terms with derivatives of the velocities with respect to time that represent the evolution of the velocity field.

We consider the generalized trapezoidal rule given by the following predictor multi-corrector algorithm [18]. We name $\boldsymbol{a} = \frac{\partial \boldsymbol{u}}{\partial t}$ the acceleration. For the purpose of integrating in time, we write Eq. (18) separating the terms that include the acceleration $\boldsymbol{a}$ and the others,

$$\boldsymbol{M}(\boldsymbol{a}_h) + \boldsymbol{K}(\boldsymbol{u}_h, p_h) = \boldsymbol{F} + \boldsymbol{F}^{\tau} \tag{29}$$

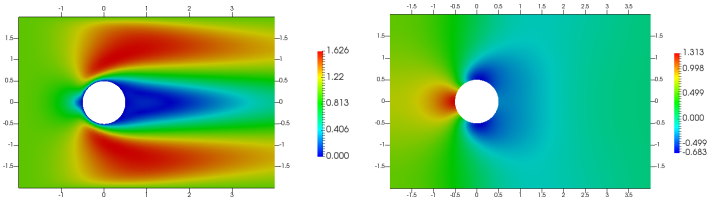
where, for the sake of simplicity, we use now $\boldsymbol{a}_h$, $\boldsymbol{u}_h$ and p_h to denote the vectors including the global degrees of freedom of the acceleration, velocity and pressure, respectively. In [24] the time integration algorithm is explained in more detail.

§4. Numerical examples: Flow around a circular cylinder

This problem has been studied extensively in the literature, see for instance [14] and its references. The flow around the circular cylinder depends on the Reynolds number which is defined as $Re = \frac{U \cdot D}{\nu}$, where U is the incoming flow velocity, D is the diameter, and ν is the kinematic viscosity. The domain is depicted in Fig. 1 and the mesh consists of hexagonal elements generated by *PolyMesher*, [25]. We have employed 8000 elements. We impose the velocity $(u, v) = (1, 0)$ on the outer boundary except on the right boundary where natural outflow boundary conditions are set. The no-slip boundary condition is applied on the cylinder surface. Fig. 2 represents the velocity and pressure magnitudes for $Re = 25$.

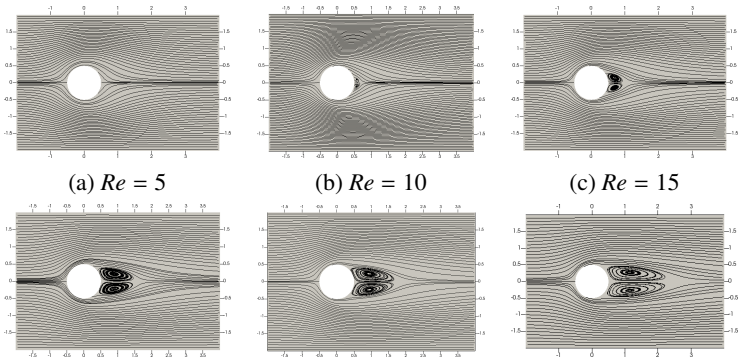
We have simulated this problem for Reynolds number up to 45 in which the steady flow becomes unstable. It is well-known that for Re that range from 6 to 45 approx. the flow is symmetric with two vortices behind the cylinder, see Fig. 3. In contrast, for higher Re, a Hopf bifurcation arises producing unstable flow.

As we can observed the numerical solution is stable and similar to the expected one for this problem. Also, in comparison with the use of FEM and stabilized methods, the solution



(a) Velocity field (b) Pressure

Figure 2: Velocity field and pressure. $Re = 25$.



(a) $Re = 5$ (b) $Re = 10$ (c) $Re = 15$

(d) $Re = 25$ (e) $Re = 30$ (f) $Re = 45$

Figure 3: Streamlines for several Reynolds numbers.

is close to the VEM solution we have presented [18, 24]. In [24], there are more numerical examples related to this work.

§5. Conclusions

In this work, the Navier-Stokes equations are discretized using VEM. The numerical method is based on the theory of stabilized methods. Thus, this method enables to select the velocity and pressure spaces with equal order interpolation functions circumventing the Babuška-Brezzi condition. Also, it can be applied to convection dominated flows since stabilization terms are considered. Numerical examples show the good performance of the method.

Acknowledgements

This work has been partially funded by Gobierno de Aragón and FEDER funding from the European Union (Grupo Consolidado de Mecánica de Fluidos Computacional T21) and by the Ministerio de Economía y Competitividad under contract MAT2016-76039-C4-4-R.

References

[1] Ahmad, B., Alsaedi, A., Brezzi, F., Marini, L. D., and Russo, A. Equivalent projectors for virtual element methods. *Computers & Mathematics with Applications 66*, 3 (2013), 376–391.

[2] Antonietti, P. F., Beirão da Veiga, L., Mora, D., and Verani, M. A stream virtual element formulation of the Stokes problem on polygonal meshes. *SIAM Journal on Numerical Analysis 52*, 1 (2014), 386–404.

[3] Bazilevs, Y., Calo, V., Cottrell, J., Hughes, T., Reali, A., and Scovazzi, G. Variational multiscale residual-based turbulence modeling for large eddy simulation of incompressible flows. *Computer Methods in Applied Mechanics and Engineering 197*, 1 (2007), 173 – 201.

[4] Beirão da Veiga, L., Brezzi, F., Cangiani, A., Manzini, G., Marini, L., and Russo, A. Basic principles of virtual element methods. *Math. Mod. Meth. Appl. S. 23*, 01 (2013), 199–214.

[5] Beirão da Veiga, L., Brezzi, F., Marini, L., and Russo, A. The hitchhiker's guide to the virtual element method. *Math. Mod. Meth. Appl. S. 24*, 08 (2014), 1541–1573.

[6] Beirão da Veiga, L., Brezzi, F., and Marini, L. D. Virtual elements for linear elasticity problems. *SIAM J. Numer. Anal. 51*, 2 (2013), 794–812.

[7] Beirão da Veiga, L., Lovadina, C., and Russo, A. Stability analysis for the virtual element method. *Mathematical Models and Methods in Applied Sciences 27*, 13 (2017), 2557–2594.

[8] Beirão da Veiga, L., Lovadina, C., and Vacca, G. Divergence free virtual elements for the Stokes problem on polygonal meshes. *ESAIM: Mathematical Modelling and Numerical Analysis 51*, 2 (2017), 509–535.

[9] Beirão da Veiga, L., Lovadina, C., and Vacca, G. Virtual elements for the Navier–Stokes problem on polygonal meshes. *SIAM Journal on Numerical Analysis 56*, 3 (2018), 1210–1242.

[10] Benedetto, M., Berrone, S., Borio, A., Pieraccini, S., and Scialo, S. Order preserving SUPG stabilization for the virtual element formulation of advection–diffusion problems. *Comput. Methods in Appl. Mech. Eng. 311* (2016), 18–40.

[11] Brenner, S. C., and Sung, L.-Y. Virtual element methods on meshes with small edges or faces. *Mathematical Models and Methods in Applied Sciences* (2018), 1–46.

[12] Brezzi, F., Falk, R. S., and Marini, L. D. Basic principles of mixed virtual element methods. *ESAIM: Mathematical Modelling and Numerical Analysis 48*, 4 (2014), 1227–1240.

[13] Brezzi, F., and Marini, L. D. Virtual element methods for plate bending problems. *Comput. Methods in Appl. Mech. Eng. 253* (2013), 455–462.

[14] Brøns, M., Jakobsen, B., Niss, K., Bisgaard, A. V., and Voigt, L. K. Streamline topology in the near wake of a circular cylinder at moderate reynolds numbers. *Journal of Fluid Mechanics 584* (2007), 23–43.

[15] Brooks, A., and Hughes, T. Streamline upwind/petrov-galerkin formulations for convection dominated flows with particular emphasis on the incompressible navier-stokes equations. *Comput. Meth. Appl. Mech. Engrng. 32* (1982), 199–259.

[16] Codina, R. Stabilization of incompressibility and convection through orthogonal subscales in finite element methods. *Computer Methods in Applied Mechanics and Engineering 190*, 13 (2000), 1579 – 1599.

[17] Dettmer, W., and Peric, D. An analysis of the time integration algorithms for the finite element solutions of incompressible Navier-Stokes equations based on a stabilised formulation. *Computer Methods in Applied Mechanics and Engineering 192*, 9 (2003), 1177 – 1226.

[18] Franca, L., and Frey, S. Stabilized finite element methods: II. The incompressible Navier-Stokes equations. *Comput. Meth. Appl. Mech. Engrng. 99* (1992), 209–233.

[19] Gain, A. L., Talischi, C., and Paulino, G. H. On the virtual element method for three-dimensional linear elasticity problems on arbitrary polyhedral meshes. *Computer Methods in Applied Mechanics and Engineering 282* (2014), 132 – 160.

[20] Hauke, G., and Hughes, T. A unified approach to compressible and incompressible flows. *Comput. Meth. Appl. Mech. Engrg. 113* (1994), 389–395.

[21] Hughes, T., Franca, L., and Mallet, M. A new finite element formulation for computational fluid dynamics: VI. convergence analysis of the generalized supg formulation for linear time-dependent multidimensional advective-diffusive systems. *Comput. Meth. Appl. Mech. Engrg. 63* (1987), 97–112.

[22] Hughes, T. J., and Franca, L. P. A new finite element formulation for computational fluid dynamics: VII. the Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. *Comput. Meth. Appl. Mech. Engrg. 65*, 1 (1987), 85–96.

[23] Irisarri, D. Virtual element method stabilization for convection-diffusion-reaction problems using the link-cutting condition. *Calcolo 54*, 1 (2017), 141–154.

[24] Irisarri, D., and Hauke, G. Stabilized virtual element methods for the unsteady incompressible navier-stokes equations. *Submitted*.

[25] Talischi, C., Paulino, G. H., Pereira, A., and Menezes, I. F. Polymesher: a general-purpose mesh generator for polygonal elements written in matlab. *Structural and Multidisciplinary Optimization 45*, 3 (2012), 309–328.

[26] Tezduyar, T., Mittal, S., Ray, S., and Shih, R. Incompressible flow computations with stabilized bilinear and linear equal-order-interpolation velocity-pressure elements. *Computer Methods in Applied Mechanics and Engineering 95*, 2 (1992), 221 – 242.

[27] Wriggers, P., Rust, W., and Reddy, B. A virtual element method for contact. *Computational Mechanics 58*, 6 (2016), 1039–1050.

D. Irisarri and G. Hauke
LIFTEC (CSIC) – University of Zaragoza,
C/María de Luna 3, 50018 Zaragoza, Spain
dirisarri@unizar.es and ghauke@unizar.es

Monografías Matemáticas *García de Galdeano* **42**, 153–164 (2019)

ANALYSIS OF THE EQUILIBRIA AND LIMIT CYCLE OSCILLATIONS OF FLIGHT DYNAMICS AND AIRFOIL AEROELASTICITY

Sébastien Kolb

Abstract. In aeronautics some phenomena require a nonlinear approach because the linear analysis is not sufficient to catch the underlying physics. Some issues met in the fields of flight dynamics and aeroelasticity are concerned with this feature. This study aims at showing so-called bifurcations implying unpredictable behaviours in the linear frame such as jumps or appearances of limit cycles and thus for which a nonlinear analysis is mandatory in order to catch the real behaviour. The methodology is based on the continuation algorithm amongst others. Practical aspects necessary to perform such an analysis of airplane design are here exposed.

Keywords: bifurcation theory, flight dynamics, aeroelasticity.
AMS classification: 34A34, 34K18, 37G10, 37G15.

Introduction

Some phenomena of aircraft flight dynamics and airfoil aeroelasticity must be examined thanks to a nonlinear approach. In this context, the bifurcation theory allows to set a mathematical frame, to perform an analysis and to understand the underlying dynamics.

As far as the longitudinal flight dynamics of the studied aircraft is concerned, a Hopf bifurcation is diagnosed and gives rise to periodic orbits. Moreover there is a range of elevator deflections δ_e for which there are multiple equilibria. A pitchfork bifurcation seems responsible for this feature leading to a possible stabilization at a nonzero bank angle ϕ. Both situations may surprise the pilot and can be hazardous to manage (especially during a critical phase such as a landing).

The other topic deals with the aeroelasticity of an airfoil whose nonlinear physics come from the pitch stiffness (torsion) or the plunge stiffness (bending) amongst others. Computing the equilibria and the envelope of the periodic orbits (with the continuation algorithm of the *matcont* toolbox of *matlab*) may help investigating some types of nonlinear behaviour.

For example, the plunge stiffness can be hardened $k_h : h \mapsto K_h\left(1 + \xi_h h^2\right)$. The observation of the bifurcation diagrams shows that the Hopf bifurcation associated to a high ξ_h is supercritical whereas the one associated to a low ξ_h is subcritical. This last case may be a dangerous situation since limit cycles may appear before the critical flutter speed determined in the classical linear frame.

In this article, after presenting the employed mathematical framework of dynamical systems, the flight dynamics of the F-18 fighter aircraft is first studied. The modeling of the

flight dynamics and the variables employed are explicited. The phases of longitudinal flight and turn are analysed mathematically then the results are interpreted from the point of view of flight dynamics. Afterwards the nonlinear aeroelasticity of a 2D airfoil section is examined. Two cases of stiffness hardening are examined. The type of the associated Hopf bifurcations is determined and the dangerousness of each situation is assessed.

§1. Mathematical framework and modeling

The models are described under the form of an ordinary differential equation (ODE) whose function $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ (vector field corresponding to the dynamics) is supposed sufficiently regular [5]:

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

where X is the so-called state vector of dimension n and U is the control vector of dimension p.

Bifurcation theory studies how the structure of the trajectories solution of a dynamical system evolves qualitatively when the control parameters are varying. When limiting the approach to the local bifurcations, then the focus is set on the equilibria, their changes of stability and the apparition of multiple equilibria and limit cycles for certain values of control parameters.

Definition 1. Equilibria are linked to zero dynamics and are the solutions (X, U) of the equation

$$F(X, U) = 0 \tag{2}$$

For most of the equilibria (X_e, U_e) i.e. the non critical ones, the methodology of analysis is based on the theorem of Hartman-Grobman [5] which states that

Theorem 1 (Hartman-Grobman)**.** *If $D_X F(X_e, U_e)$ has no zero or purely imaginary eigenvalues then there is a homeomorphism locally taking orbits of the nonlinear flow to those of the linear flow.*

There are several types of bifurcations which are met in this study i.e. the Hopf bifurcation which is the main one for this issue, the pitchfork bifurcation, the saddle-node bifurcation [5].

Theorem 2 (Hopf)**.** *If $F(X, U) = 0$ has an equilibrium (X_e, U_e) for which $D_X F(X_e, U_e)$ has:*

1. *a pair of purely imaginary eigenvalues $\lambda, \bar{\lambda}$ and no other eigenvalues with zero real parts,*
2. *$\frac{\partial Re\lambda(u)}{\partial u}|_{u_e} \neq 0$ (derivative of the real part of one eigenvalue λ with respect to one control state u of the control vector U),*

then there is a surface of periodic solutions in the center manifold which has a quadratic tangency with the eigenspace of $\lambda(u_e), \bar{\lambda}(u_e)$.

As far as the practical aspects are concerned, the computations will be made with the numerical bifurcation analysis toolbox *matcont* [2]. First the analysis will be focused on the flight dynamics of a fighter aircraft and next on the aeroelasticity of a two-dimensional airfoil.

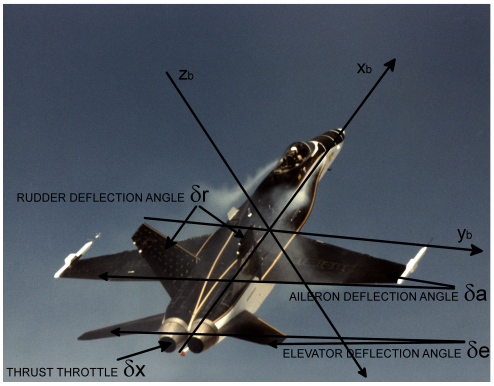


Figure 1: F/A-18 High Alpha Research Vehicle (HARV) flown by NASA's Dryden Flight Research Center, Edwards, CA (`https://www.dfrc.nasa.gov/Gallery/Photo/F-18HARV/Large/EC89-0096-149.jpg`) with added annotations for controls and body-fixed frame

§2. Flight dynamics

The flight dynamics of a F-18 fighter aircraft is studied here. After presenting the model used, two flight phases are studied, that is to say longitudinal flight and turn. Each time, the link is made between the mathematical results and a practical interpretation of the aircraft behaviour.

2.1. Description of the flight dynamics model

The flight dynamics model [4] is taken "as is" that is to say phenomena are observed and analysed but the inner content of the model is not deeply studied.

Concerning the mathematical model, as for the (smooth) function F associated to the dynamical system (1), its expression is polynomial or piecewise polynomial due to the identification of the aerodynamic forces and moments. When studying the whole flight dynamics, the control vector is $U = \{\delta_a, \delta_e, \delta_r, \delta_x\}$ (figure 1 explicits the parts of aircraft especially the tails involved for each control) and the state vector $X = \{\mathcal{M}, \alpha, \beta, p, q, r, \phi, \theta, \psi, x, y, h\}$ contains the variables of airspeed, angles, rotation rates and position (illustrated in figure 2). For a fighter aircraft, the Mach $\mathcal{M}$ which corresponds to the dimensionless ratio of the airspeed to the local speed of sound v_s is often preferred to the classical airspeed V ($\mathcal{M} = V/v_s$).

As far as the pure longitudinal model is concerned, it takes only into account the movement in the vertical plane (no transverse motion), that's why there remain only the state vector $X = \{\mathcal{M}, \alpha, q, \theta, h\}$ and the control vector $U = \{\delta_e, \delta_x\}$ (states presented in the middle of figure 2 and two controls including elevator deflection δ_e of the horizontal tail and thrust throttle δ_x) and besides the other variables are fixed to zero.

The equations of flight dynamics follow the formalism of [4]. The six first ones describe the physics and come from the Newton law (forces and moments), the six last ones are linked to the dynamics of the Euler angles (ϕ, θ, ψ) and of the position components (x, y, h) and

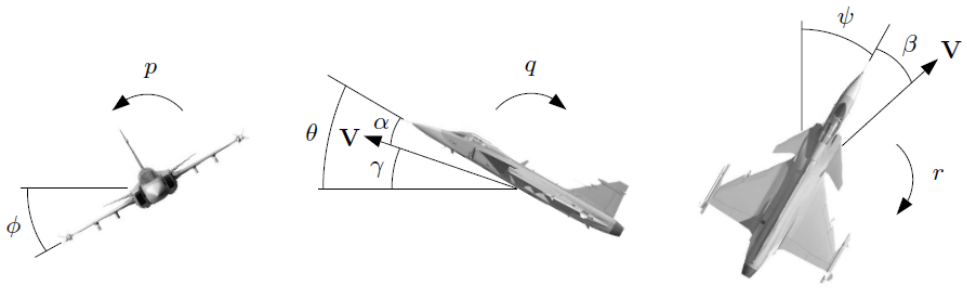


Figure 2: Flight dynamics variables: Euler orientation angles (ϕ, θ, ψ), aerodynamic angles (α, β), angular velocities (p, q, r), flight-path angle γ and airspeed V [6]

traduce some kinematic relations between the variables:

$$
\begin{aligned}
\dot{\mathcal{M}} &= \frac{1}{m v_s}\left[T_m \delta_x \cos\alpha\cos\beta - C_D \frac{1}{2}\rho(v_s\mathcal{M})^2 S - mg\sin\gamma\right] \\
\dot{\alpha} &= q - \frac{1}{\cos\beta}\Bigg[(p\cos\alpha + r\sin\alpha)\sin\beta \\
&\qquad + \frac{1}{m v_s \mathcal{M}}\left(T_m\delta_x\sin\alpha + C_L\frac{1}{2}\rho(v_s\mathcal{M})^2 S - mg\cos\mu\cos\gamma\right)\Bigg] \\
\dot{\beta} &= \frac{1}{m v_s \mathcal{M}}\left[-T_m\delta_x\cos\alpha\sin\beta + C_Y\frac{1}{2}\rho(v_s\mathcal{M})^2 S + mg\sin\mu\cos\gamma\right] \\
\dot{p} &= \frac{I_y - I_z}{I_x} qr + \frac{1}{2I_x}\rho(v_s\mathcal{M})^2 S\, b C_l \\
\dot{q} &= \frac{I_z - I_x}{I_y} pr + \frac{1}{2I_y}\rho(v_s\mathcal{M})^2 S\, c C_m \\
\dot{r} &= \frac{I_x - I_y}{I_z} pq + \frac{1}{2I_z}\rho(v_s\mathcal{M})^2 S\, b C_n \\
\dot{\phi} &= p + q\sin\phi\tan\theta + r\cos\phi\tan\theta \\
\dot{\theta} &= q\cos\phi - r\sin\phi \\
\dot{\psi} &= (q\sin\phi + r\cos\phi)\sec\theta \\
\dot{x} &= v_s\mathcal{M}\cos\gamma\cos\chi \\
\dot{y} &= v_s\mathcal{M}\cos\gamma\sin\chi \\
\dot{h} &= -v_s\mathcal{M}\sin\gamma
\end{aligned}
\tag{3}
$$

The aerodynamic coefficients of drag C_D, side force C_Y, lift C_L, roll moment C_l, pitch moment C_m and yaw moment C_n are piecewise polynomial functions of the angles of attack α, sideslip β and deflection angles of elevator δ_e, aileron δ_a, rudder δ_r and rates of roll p, pitch q and yaw r. More precisely the aerodynamic coefficients are functions of the following variables:

$$C_D(\alpha), C_Y(\beta, \alpha, \delta_r, \delta_a), C_L(\alpha, \delta_e), C_l(\alpha, \beta, p, r, \delta_a, \delta_r), C_m(\alpha, q, \delta_e), C_n(\alpha, \beta, r, \delta_a, \delta_r) \quad (4)$$

In the aforementioned equations (3), χ, γ, μ are the wind axes orientation angles (between the aerodynamic and body frames) whereas ϕ, θ, ψ are the Euler orientation angles (between the body and Earth frames). Moreover the speed of sound v_s and the air density ρ depend on the altitude h. Some data correspond to characteristic dimensions of the aircraft such as the wing span b, the mean aerodynamic chord c, the mass m, the reference area S (wing surface) and the principal moments of inertia I_x, I_y, I_z. Besides the thrust throttle δ_x is here the percentage of maximum available thrust ($T = T_m \delta_x$).

The steps of the analysis methodology are the following ones. The locus of equilibrium points (bifurcation diagram) is first determined. Then the values of critical control parameters (bifurcation values) are calculated and afterwards potentially the locus of bifurcation points. Finally the link is made between the mathematical results (bifurcation theory) and the physical interpretation from the flight dynamics viewpoint. Beneath time simulations are performed so as to illustrate concretely the results of the nonlinear analysis.

2.2. Longitudinal flight

In this section, the longitudinal flight is studied that is to say only the flight in the vertical plane is considered and there are no sideslip and no lateral rotations (for the equilibria of the nominal flight). A classical result of flight dynamics is that for one elevator deflection angle and one thrust throttle position, there is only one (longitudinal) equilibrium. Especially since for a longitudinal equilibrium the sum of the pitching moments must be zero, one elevator deflection δ_e correspond to one angle-of-attack α [8]. But for this F-18 aircraft, several critical behaviours are observed.

2.2.1. Mathematical analysis and numerical results

In order to conduct the analysis, the bifurcation diagram is plotted in figure 3. It presents the angles-of-attack α (equilibria and limit cycles) versus elevator deflection angle δ_e for a thrust throttle fixed at $\delta_x \approx 40\%$. A Hopf bifurcation [5] is diagnosed at $\delta_e \approx -14.9$ deg and creates limit cycles.

Another classical diagram is the locus of bifurcation points presented in figure 4 which shows the critical controls of elevator deflection δ_e and thrust throttle δ_x for which a bifurcation occurs.

There are Hopf bifurcations and branch points. The generalized Hopf bifurcation (where the first Lyapounov coefficient vanishes [2]) at the critical control parameters $\delta_x \approx 53\%, \delta_e \approx -14.3$ deg changes the way periodic orbits are created i.e. for lower or higher elevator deflections than the bifurcation value. The branch points indicates a new phenomenon. Indeed there is a range of elevator deflections ($\delta_e \in [-12\,\text{deg}, -9\,\text{deg}]$, $\delta_x \geq 50\%$) with multiple equilibria (two stable and one unstable). Besides the two distinct curves intersect at a zero Hopf bifurcation (corresponding to a pair of purely imaginary eigenvalues and a zero eigenvalue [2]) at $\delta_x \approx 82\%, \delta_e \approx -11.3$ deg.

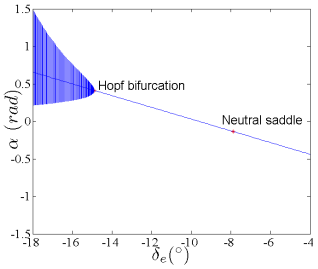


Figure 3: Bifurcation diagram for the longitudinal flight of a F-18 aircraft showing angle-of-attack α in function of elevator deflection angle δ_e

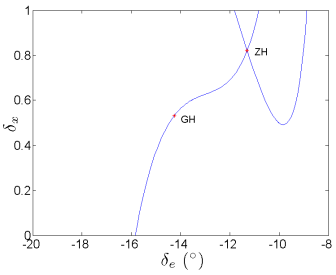


Figure 4: Locus of the bifurcation points for the (δ_e, δ_x) controls

We will next consider the case of a thrust throttle fixed at $\delta_x \approx 70\%$ and study the different bifurcations appearing and especially which physical variables and aircraft mode are involved.

On the one hand, the Hopf bifurcation at the elevator deflection $\delta_e \approx -11.9\,\text{deg}$ involves the variables $(\mathcal{M}, \alpha, q, \theta)$ and is associated to the (pair of complex conjugate) eigenvalues $\lambda_H = \pm 0.298i$. The aircraft begins suddenly to oscillate at a flight path angle of $\gamma = 3.7\,\text{deg}$ after this destabilization. This is a similar phenomenon as the one illustrated figure 3.

On the other hand, the branch point at the elevator deflection $\delta_e \approx -10.9\,\text{deg}$ involves the lateral variables (β, p, r, ϕ) and is associated to the real eigenvalue $\lambda_{BP} = 0$ (the whole model of flight dynamics is exploited for this calculation). That's why from this equilibrium point, it is possible to have stable equilibria with nonzero bank angle ϕ in an asymmetric configuration.

The following time simulations (figures 5 and 6) illustrates both behaviours. Figure 5 shows the behaviour for elevator deflections δ_e higher and lower than the critical Hopf bifurcation value. A stable limit cycle exist for $\delta_e = -13\,\text{deg}$ and a stable equilibrium for $\delta_e = -11.5\,\text{deg}$.

Besides between the two branch point values at elevator deflection angles of $\delta_e = -9.1\,\text{deg}$ and $\delta_e = -10.9\,\text{deg}$, the classical longitudinal equilibrium becomes unstable and the aircraft stabilizes itself at a nonzero bank angle. Figure 6 shows time simulations for an initial bank angle $\phi = -0.3\ rad$ and different elevator deflection angles of $\delta_e = -12\,\text{deg}$ and

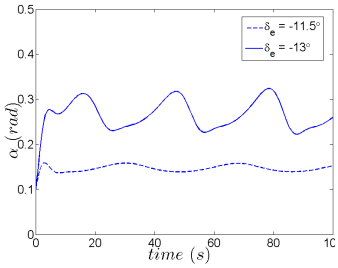


Figure 5: Time simulations for elevator deflection angles δ_e higher and lower than the critical Hopf bifurcation value

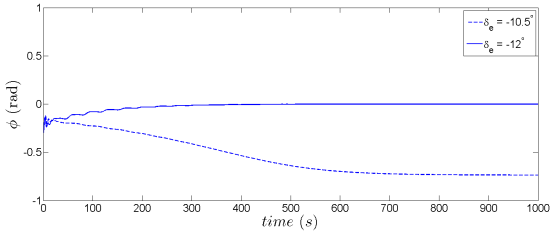


Figure 6: Time simulations for different elevator deflection angles δ_e (between and outside the critical branch point values)

$\delta_e = -10.5$ deg.

Near the bifurcation point, a small change of the elevator deflection angle δ_e can render the classical equilibrium unstable (at a zero bank angle) and leads to a stabilization at a nonzero bank angle ϕ.

Figure 7 is the bifurcation diagram associated to the longitudinal flight. The bank angle ϕ at equilibrium is given in function of the elevator deflection angle δ_e for a throttle $\delta_x \approx 70\%$. In particular, there is a range of elevator deflection angles δ_e with two stable equilibria (nonzero bank angles) and one unstable equilibrium.

Pitchfork bifurcations occur for $\alpha \approx 0.1 rad \approx 5.8$ deg and $\alpha \approx -0.037 rad \approx -2.1$ deg and give rise to several branches of equilibria.

2.2.2. Physical interpretation

In the longitudinal flight dynamics of the F-18 fighter, Hopf bifurcations and pitchfork bifurcations are met. A practical consequence of the existence of a Hopf bifurcation is that the pilot can be astonished by the sudden apparition of peridic orbits during a seemingly normal flight at equilibrium. For example, during a phase with a nonzero flight-path angle γ such as a landing, this sudden change of behaviour can be very hazardous. The loss of stability of the phugoid mode (exchange of airspeed and altitude [8]) seems to be responsible for that.

Moreover there exist also pitchfork bifurcations implying the existence of multiple equilibria for a range of elevator deflection angles δ_e: two stable equilibria with nonzero bank

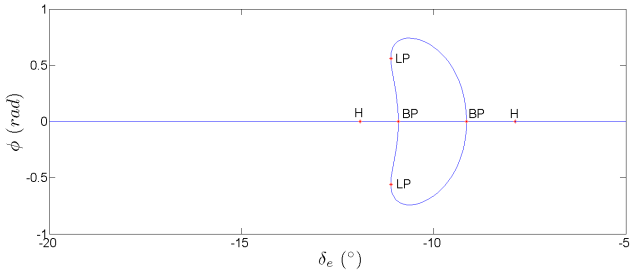


Figure 7: Bifurcation diagram for the longitudinal flight of a F-18 aircraft showing bank angle ϕ in function of the elevator deflection angle δ_e

angle ϕ besides the classical symmetric longitudinal equilibrium (unfortunately unstable). Thus the aircraft can stabilize itself in an unusual asymmetric configuration. Since this motion is very slow and is recoverable by a reverse control action, it seems manageable by a pilot. Nevertheless this propensity of the aircraft to engage itself in a turn due to the loss of stability of the so-called spiral mode [8] may be unpleasant for a pilot.

After analysing the longitudinal flight dynamics and showing some interesting bifurcations and unusual behaviours, the next examined flight phase will be the turn.

2.3. Turn

The effect of aileron deflections on the turn properties (and especially on the roll rate p) are studied here. We will see that it may give rise to nonlinear phenomena and to unexpected behaviour.

The bifurcation diagram is first plotted in figure 8 with the thrust throttle fixed at $\delta_x = 0.5$ and the elevator deflection angle at $\delta_e = -15°$. Limit points (also called fold or saddle-node bifurcation [5]) appear. Theses last ones lead to a jump of roll rate p near the aileron deflections $\delta_a = \pm 32$ deg and is illustrated in the time simulations of figure 9. Nevertheless at an aileron deflections $\delta_a = \pm 18$ deg, the high roll rate may suddenly disappear.

From the physical point of view, at the mechanical limits of the authorized aileron deflection range, the pilot must be careful since the flight dynamics meets some jumps and hysteresis phenomena. The irreversible, quick and unexpected nature of such phenomenona can lead to a hazardous situation with a high roll rate. Thus an advice for the the pilot is to avoid using the ailerons too closely of their mechanical limits.

After examining the flight dynamics of a F-18 fighter aircraft during the phases of longitunal flight and turn thus revealing the existence of diverse types of bifurcations and of unintuitive behaviours, we will next treat the case of nonlinear aeroelasticity of a two-dimensional airfoil.

§3. Airfoil aeroelasticity

After describing the classical model for the aeroelasticity of a 2D airfoil section and its nonlinear terms, the influence of these last ones on the global system behaviour is assessed.

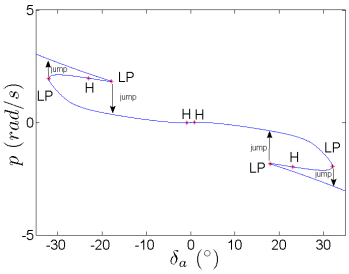


Figure 8: Bifurcation diagram associated to a F-18 turn whose roll rate p is piloted with the control δ_a of aileron deflection

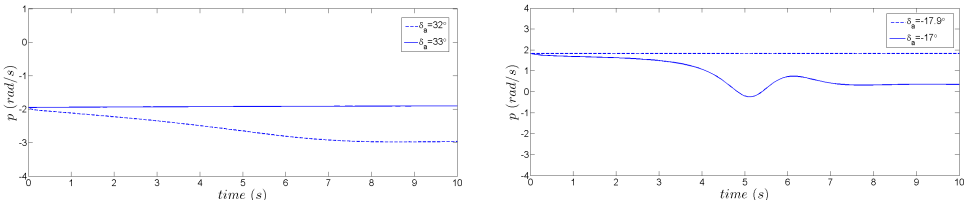


Figure 9: Time simulations for different aileron deflections ($\delta_a = 32\,\text{deg}, \delta_a = 33\,\text{deg}$ and $\delta_a = -17.9\,\text{deg}, \delta_a = -17\,\text{deg}$)

3.1. Airfoil aeroelasticity model

The classical mathematical model is based on a force equation (including lift force, plunge stiffness) and a moment equation (including pitching moment, pitch stiffness) [3]:

$$\begin{pmatrix} m_T & m_W x_\alpha b \\ m_W x_\alpha b & I_\alpha \end{pmatrix} \begin{pmatrix} \ddot{h} \\ \ddot{\alpha} \end{pmatrix} + \begin{pmatrix} c_h & 0 \\ 0 & c_\alpha \end{pmatrix} \begin{pmatrix} \dot{h} \\ \dot{\alpha} \end{pmatrix} + \begin{pmatrix} k_h(h) & 0 \\ 0 & k_\alpha(\alpha) \end{pmatrix} \begin{pmatrix} h \\ \alpha \end{pmatrix} = \begin{pmatrix} -L \\ M \end{pmatrix} \tag{5}$$

Writing the second order ordinary differential equation under the canonical form, the state vector is $X = \left\{h, \alpha, \dot{h}, \dot{\alpha}\right\}$ and the control vector is $U = \{V, \beta\}$ (variables of airspeed and flap deflection angle intervening in the calculation of lift and pitching moment).

The aeroelasticity of an airfoil may present nonlinear features. The ones which are considered here come from the pitch (torsional spring k_α) or plunge stiffness (translational spring k_h). As far as the overall behaviour is considered, apart from the classical change of equilibrium stability at the critical flutter speed, they impact the way limit cycle oscillations are created near the corresponding Hopf bifurcation point.

In order to perform the concrete analysis, the different diagrams of bifurcation theory are plotted and allow to determine the underlying dynamics and the structural changes. Generally in order to determine the Hopf bifurcation type, the algebraic expressions (normal forms) are used and allows to calculate the Lyapounov coefficient [5]. Here numerical simulations are performed so as to see the behaviours linked to the different situations e.g. periodic orbits, equilibria which are stable or unstable.

Nevertheless for the aeroelasticity problem, since the main equilibrium state value is

clearly known to be zero, the most important points consist in determining the critical flutter speed, the Hopf bifurcation type that is to say whether it is supercritical with stable limit cycles or subcritical with unstable limit cycles and potentially the enveloppe of periodic orbits. It implies respectively slowly growing oscillations or oscillations of large amplitude even before reaching the bifurcation critical speed. From the practical point of view, the last situation is quite dangerous and must be avoided [3].

3.2. Sensitivity to physical parameters

Several conclusions can be drawn concerning the sensitivity to physical parameters. The plunge stiffness seems to be favourable that is to say to imply a supercritical Hopf bifurcation. On the contrary, the pitch stiffness seems to be unfavourable in the sense that they induce a subcritical Hopf bifurcation. These statements will be illustrated in the following section.

In the model furnished in [9] and [7], the stiffness is hardened towards either plunge or pitch. The plunge stiffness comes from the spring constant for plunge degree of freedom. A nonlinear law is taken into account $k_h : h \mapsto K_h\left(1+\xi_h h^2\right)h$ with $\xi_h = 0.09$ and $\xi_h = 50$ instead of the standard linear $k_h : h \mapsto K_h h$. The Hopf bifurcation associated to $\xi_h = 0.09$ is subcritical and the one associated to $\xi_h = 50$ is supercritical as can be seen in the bifurcation diagrams (figure 10). As a consequence, the plunge stiffness hardening seems to have a nefast effect on the overall behaviour.

For the the second case study, the benchmark described in [1] is exploited for the linear part and the pitch stiffness follows the chosen nonlinear law

$$k_\alpha : \alpha \mapsto K_\alpha\left(1+10\alpha^2\right)\alpha \tag{6}$$

In figure 11, the bifurcation diagram is plotted and contains two bifurcations. There are a classical Hopf bifurcation which is supercritical and also a branch point (pitchfork bifurcation), this last one is linked to a real negative eigenvalue which becomes positive (the system remains globally unstable). The linear frame would only determine the respective flutter speed and divergence speed of the zero equilibrium. But in the nonlinear frame, the presence of stable limit cycles created at the Hopf bifurcation reduces the negative impact of the unstable equilibrium since the amplitude of the oscillations are limited. If the destabilization due to flutter is managed thanks to a feedback loop, then the presence of stable equilibria after the pitchfork bifurcation limits also the amplitudes of plunge and pitch. Thus the pitch stiffness hardening ($k_\alpha : \alpha \mapsto K_\alpha\left(1+\xi_\alpha\alpha^2\right)\alpha$ instead of $k_\alpha : \alpha \mapsto K_\alpha\alpha$) seems to have a beneficial effect.

Conclusion

The bifurcation theory allows to show and to explain some phenomena of aircraft flight dynamics and airfoil aeroelasticity. The sudden apparitions of periodic orbits and of multiple equilibria are diagnosed. The characterisation of the associated bifurcations in terms of type and control parameter values permits to assess their level of hazardousness from the practical point of view.

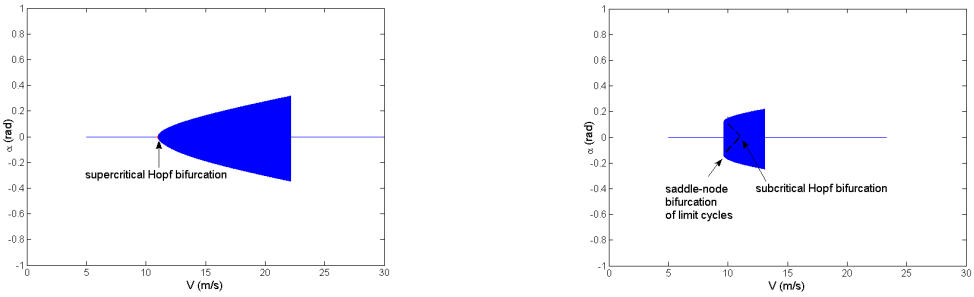


Figure 10: Bifurcation diagrams with airspeed V as control parameter presenting limit cycles and equilibria for nonlinear plunge stiffness with $\xi_h = 50$ (left) and $\xi_h = 0.09$ (right)

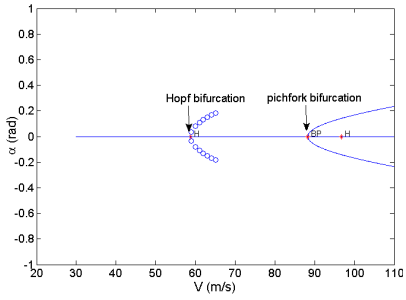


Figure 11: Bifurcation diagrams with airspeed V as control parameter presenting limit cycles and equilibria for nonlinear pitch stiffness with $\xi_\alpha = 10$

Concerning the longitudinal flight dynamics, the existence of periodic orbits and of equilibria at a nonzero bank angle present a risk for the flight safety in a situation which seems apparently completely normal. For the aircraft turns, the fold bifurcations and associated jumps reveal the effective range of the lateral control which can be used without any problem.

As far as the airfoil aeroelasticity is concerned, the hardening of the stiffness in pitch or in plunge can have a positive or a negative effect. Determining the type of the associated Hopf bifurcation that is to say supercritical or subcritical is the main feature so as to evaluate the dangerousness of the configuration and especially the sufficiency of the determination of the classical critical flutter speed.

Nomenclature

Common

U	control vector	X	state vector
L	lift (N)	M	pitching moment ($N.m$)
V	airspeed (m/s)		

	Flight dynamics
α	angle-of-attack (rad)
β	sideslip angle (rad)
δ_a	aileron deflection angle (deg)
δ_e	elevator deflection angle (deg)
δ_r	rudder deflection angle (deg)
δ_x	thrust throttle (%)
γ	flight-path angle (rad)
ϕ	bank angle (rad)
θ	pitch angle (rad)
ψ	heading angle (rad)
x, y	aircraft position coordinates (m)
h	altitude (m)
p	roll rate (rad/s)
q	pitch rate (rad/s)
r	yaw rate (rad/s)

	Aeroelasticity
m_T	total mass of the wing (kg)
m_W	wing mass alone (kg)
I_α	mass moment of inertia about the elastic axis
b	half chord length (m)
x_α	nondimensionalized distance between the center of mass and the elastic axis
h	plunge (m)
α	angle-of-attack/pitch angle (rad)
β	flap deflection angle (rad)
ρ	air density (kg/m^3)
c_h	plunge structural damping coefficient
c_α	pitch structural damping coefficient
k_h	plunge stiffness
k_α	pitch stiffness

References

[1] Axisa, F. *Vibrations sous écoulements*. Hermes, 2001.

[2] Dhooge, A., Govaerts, W., and Kuznetsov, Y. A. Matcont: A matlab package for numerical bifurcation analysis of odes. *ACM Trans. Math. Softw. 29*, 2 (2003), 141–164.

[3] Dimitriadis, G. *Introduction to nonlinear Aeroelasticity*. John Wiley & Sons Ltd, 2017.

[4] Fan, Y., Lutze, F. H., and Cliff, E. M. Time-optimal lateral maneuvers of an aircraft. *Journal of Guidance, Control, and Dynamics 18*, 5 (1995), 1106–1111.

[5] Guckenheimer, J., and Holmes, P. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, vol. 42 of *Applied Mathematical Sciences*. Springer, 2002.

[6] Härkegård, O. Backstepping and control allocation with applications to flight control. Master's thesis, Linköping University, 2003.

[7] O'Neil, T., and Strganac, T. W. Aeroelastic response of rigid wing supported by nonlinear springs. *Journal of Aircraft 35*, 4 (1998).

[8] Sinha, N. K., and Ananthkrishnan, N. *Elementary Flight Dynamics with an Introduction to Bifurcation and Continuation Methods*. Taylor & Francis Group, 2017.

[9] Strganac, T. W., Ko, J., Thompson, D. E., and Kurdila, A. J. Identification and control of limit cycle oscillations in aeroelastic systems. *Journal of Guidance, Control, and Dynamics 23*, 6 (2000).

S. Kolb
CReA, French Air Force Research Centre
BA 701, 13661 Salon Air, France
sebastien.kolb@ecole-air.fr

Monografías Matemáticas *García de Galdeano* **42**, 165–172 (2019)

PERIODIC SOLUTIONS IN THE HÉNON-HEILES ROTATING SYSTEM

Víctor Lanchares, Manuel Iñarrea, Jesús Palacián, Ana Isabel Pascual, José Pablo Salas and Patricia Yanguas

Abstract. We consider a generalized Hénon-Heiles system in a rotating frame. Our aim is to prove the existence of periodic orbits in a neighborhood of the origin for appropriate values of the rotating frequency. To this end, we use classical averaging theory to demonstrate that the number of periodic orbits is in correspondence with the equilibrium solutions of the original system, with the same type of stability.

Keywords: Generalized Hénon-Heiles system, periodic orbits, averaging.

AMS classification: 70H08, 70H09, 70H12, 70H15, 34C25, 37C27.

§1. Introduction

Equilibrium points and periodic orbits of dynamical systems are of special interest to understand its dynamics. They organize the phase structure and, some times, the appearance of heteroclitic connections allows migration of orbits giving rise to a kind of transport phenomena. For instance, this is what happens in Celestial Mechanics in the framework of the three body problem [6, 10], but also in the context of galactic dynamics, where the existence of heteroclitic connections are proposed as a way to explain the formation of spiral arms [12]. The model considered in [12] is based on a logarithmic potential. However, many galactic models consider cubic or quartic polynomial potentials [3]. This is the case of the well known Hénon-Heiles system, used to describe stellar orbits under the action of the galaxy's core [7]. Although this model has been considered as a paradigmatic system to study chaos and other properties of planar dynamical systems in many different fields, it does not take into account the effect of a rotating framework. In this way, de Zeeuw & Merritt [5] consider the cubic potential of the Hénon-Heiles system for a rotating galaxy and other authors consider a similar model in the context of atomic physics [2, 9]. The presence of the rotating frequency makes the system more interesting, from a dynamics point of view, with the appearance of Lagrangian type equilibrium points. In [8], a detailed analysis of the stability of these points is performed. One of the remarkable facts of this system is the existence of a critical value of the rotating frequency in such a way that the nature of the critical points, as critical points of the effective potential, reverses. This is an interesting situation that deserves more insight. In particular, the existence of periodic orbits is the next step in understanding the dynamic of the system. To prove the existence of periodic orbits, we will use the classical averaging theory [13] used successfully to find periodic orbits in many different dynamical systems [1, 4, 11].

§2. The system

Let us consider the Hamiltonian system defined by

$$\mathcal{H} = \frac{1}{2}(X^2 + Y^2) - \omega(xY - yX) + \frac{1}{2}(x^2 + y^2) + ayx^2 + by^3, \tag{1}$$

which can be viewed as a generalized Hénon-Heiles system in a rotating reference frame with angular velocity ω, where we assume, without loss of generality, $a > 0$ and $\omega > 0$. The equations of the motion are given by

$$\begin{aligned} \dot{x} &= \frac{\partial \mathcal{H}}{\partial X} = X + \omega y, \qquad & \dot{X} &= -\frac{\partial \mathcal{H}}{\partial x} = -x + \omega Y - 2axy, \\ \dot{y} &= \frac{\partial \mathcal{H}}{\partial Y} = Y - \omega x, \qquad & \dot{Y} &= -\frac{\partial \mathcal{H}}{\partial y} = -y - \omega X - ax^2 - 3by^2. \end{aligned} \tag{2}$$

It is clear that the origin is always an equilibrium point. Moreover, three more equilibrium points can appear, depending on the values of the parameters a and b. An interesting fact is that if

$$E_0 \equiv (x_0, y_0, X_0, Y_0)$$

is an equilibrium point for $\omega = \omega_0$, then

$$\hat{E}_0 \equiv (-x_0/\omega_0^2, -y_0/\omega_0^2, -X_0/\omega_0^4, -Y_0/\omega_0^4)$$

is also a critical point for $\omega = 1/\omega_0$. In this way, there is a correspondence between the cases $0 < \omega < 1$ and $\omega > 1$. However, there is a slight difference. Indeed, equilibrium points are related to the critical points of the effective potential

$$\Phi_{\text{eff}} = \mathcal{H} - \frac{1}{2}(\dot{x}^2 + \dot{y}^2) = \frac{1}{2}(x^2(2ay - \omega^2 + 1) + y^2(2by - \omega^2 + 1)), \tag{3}$$

in such a way that if E_0 is an equilibrium point of the system (1), then (x_0, y_0) is a critical point of the effective potential Φ_{eff}. In this way, if E_0 is a minimum (maximum) of the effective potential, then $\hat{E}_0$ is a maximum (minimum) of Φ_{eff}. In the case E_0 is a saddle point, the same happens for $\hat{E}_0$. As a consequence, linear stability properties cannot be extended directly from the case $0 < \omega < 1$ to the case $\omega > 1$ if the corresponding critical point is a minimum (maximum). While a minimum of Φ_{eff} is always a linear stable equilibrium, the same cannot be said for a maximum. Nevertheless, if the critical point is the origin, then it is always a linear stable equilibrium, it does not matter a minimum or a maximum. Indeed, the associated eigenvalues are

$$\lambda_{1,2} = \pm i(\omega - 1), \quad \lambda_{3,4} = \pm i(\omega + 1). \tag{4}$$

For a detailed study of equilibrium points and their stability properties the reader is referred to [8].

It is worth noting that in the transition case, $\omega = 1$, the origin loses its elliptic character, as two zero eigenvalues appear, precisely those coming from $\pm i(\omega - 1)$. Moreover, the origin is the unique equilibrium point of the system and a bifurcation occurs when all the

equilibria come into coincidence. Thus, what happens in the vicinity of the origin as $\omega \to 1$ deserves some analysis. In particular, we focus on the existence of periodic orbits and their bifurcations, assuming that $ab \neq 0$, in order to avoid degenerate situations, when non isolated equilibria appear. To begin with, we observe that, being the origin an elliptic point with associated eigenvalues given by (4), the Hamiltonian function can be transformed into an equivalent one made of two coupled harmonic oscillators with frequencies $1-\omega$ and $1+\omega$. To this end, we transform the system by means of the canonical change of variables

$$\begin{aligned} x &= -\frac{x_1}{\sqrt{2}} + \frac{x_2}{\sqrt{2}}, \qquad & X &= -\frac{X_1}{\sqrt{2}} + \frac{X_2}{\sqrt{2}}, \\ y &= \frac{X_1}{\sqrt{2}} + \frac{X_2}{\sqrt{2}}, \qquad & Y &= -\frac{x_1}{\sqrt{2}} - \frac{x_2}{\sqrt{2}}. \end{aligned} \tag{5}$$

The new Hamiltonian is given by

$$\mathcal{H}_2 = \frac{1}{2}(1-\omega)(x_1^2+X_1^2) + \frac{1}{2}(1+\omega)(x_2^2+X_2^2) + \frac{X_1+X_2}{2\sqrt{2}}(a(x_1-x_2)^2 + b(X_1+X_2)^2). \tag{6}$$

§3. Averaging and periodic orbits

Taking into account that $\omega \approx 1$, one of them oscillates with high frequency with respect to the other one and the theory of averaging is suitable to study the system. In particular, the following Theorem [13] can be applied

Theorem 1. *Let us consider the differential system*

$$\dot{x} = \varepsilon f(t,x) + \varepsilon^2 g(t,x,\varepsilon), \tag{7}$$

with $x \in D \subseteq \mathbb{R}^n$, $t \geq 0$. Moreover f, g, $\partial f/\partial x$, $\partial^2 f/\partial x^2$, $\partial g/\partial x$ are defined, continuous and bounded by a constant M independent of ε in $[0,\infty)\times D$, $0 \leq \varepsilon \leq \varepsilon_0$. In addition f and g are T-periodic in t (T independent of ε). Then, if p is a non degenerate critical point of the system

$$\dot{y} = \varepsilon f^0(y),$$

where

$$f^0(y) = \frac{1}{T}\int_0^T f(t,y)\,dt,$$

there exists a T-periodic solution $\phi(t,\varepsilon)$ of (7) which is close to p such that

$$\lim_{\varepsilon\to 0}\phi(t,\varepsilon) = p.$$

The key point is to transform the Hamiltonian differential system defined by (6) into a system in the form (7). This can be done in several steps. First of all, taking into account that we are considering $\omega \approx 1$, we scale the variables and the frequency according to

$$x_j, X_j \to \varepsilon x_j, \varepsilon X_j, \quad j=1,2, \qquad 1-\omega \to \varepsilon\nu.$$

Substituting into the Hamiltonian function, and taking out the common factor ε^2, we arrive to

$$\mathcal{H}_2 = x_2^2 + X_2^2 + \frac{\varepsilon}{2\sqrt{2}}\left[\sqrt{2}\nu(x_1^2 + X_1^2 - x_2^2 - X_2^2) + b(X_1 + X_2)^3 + a(X_1 + X_2)(x_1 - x_2)^2\right].$$

The equations of the motion are given by

$$\dot{x}_i = \frac{\partial \mathcal{H}_2}{\partial X_j}, \quad \dot{X}_j = -\frac{\partial \mathcal{H}_2}{\partial x_j}, \quad j = 1, 2.$$

Now, we introduce polar coordinates for the pair of variables (x_2, X_2) in the form

$$x_2 = r\cos\theta, \quad X_2 = r\sin\theta.$$

Thus, the differential equations for r and θ turn to be

$$\dot{r} = \left(\frac{\partial \mathcal{H}_2}{\partial X}\cos\theta - \frac{\partial \mathcal{H}_2}{\partial x}\sin\theta\right), \quad \dot{\theta} = -\frac{1}{r}\left(\frac{\partial \mathcal{H}_2}{\partial x}\cos\theta + \frac{\partial \mathcal{H}_2}{\partial X}\sin\theta\right).$$

Explicitly, these equations read as

$$\begin{cases} \dot{r} = \dfrac{\varepsilon}{4\sqrt{2}}\,(a(x_1 - r\cos\theta)(r + 2x_1\cos\theta - 3r\cos 2\theta - 4X_1\sin\theta) + \\ \qquad\qquad 6b\cos\theta(X_1 + r^{1/2}\sin\theta)^2\Big), \\ \dot{\theta} = -2 + \dfrac{\varepsilon}{2\sqrt{2}r}\Big(2\sqrt{2}\nu r + a(x_1 - r\cos\theta)(2X_1\cos\theta - (x_1 - 3r\cos\theta)\sin\theta) - \\ \qquad\qquad 3b(X_1 + r\sin\theta)^2\sin\theta\Big), \end{cases} \tag{8}$$

whereas for the variables x_1, X_1 we obtain

$$\begin{cases} \dot{x}_1 = \dfrac{\varepsilon}{4}\left(4\nu X_1 + 3\sqrt{2}b(X_1 + r^{1/2}\sin\theta)^2 + \sqrt{2}a(y - r^{1/2}\cos\theta)^2\right), \\ \dot{X}_1 = \dfrac{\varepsilon}{2\sqrt{2}}\left(\sqrt{2}\nu x_1 - a(x_1 - r^{1/2}\cos\theta)(X_1 + r^{1/2}\sin\theta)\right). \end{cases} \tag{9}$$

On the other hand, the Hamiltonian function is expressed as

$$\mathcal{H}_2 = r^2 + \frac{\varepsilon}{2\sqrt{2}}\left(\sqrt{2}\nu(x_1^2 + X_1^2 - r^2) + a(X_1 + r\sin\theta)(x_1 - r\cos\theta)^2 + b(X_1 + r\sin\theta)^2\right). \tag{10}$$

From this expression, the radial variable can be obtained as a power series in ε. Up to the first order we get

$$\begin{aligned} r \approx\ & \sqrt{h} + \frac{\varepsilon}{4\sqrt{2h}}\Big(\sqrt{2}\nu(h - x_1^2 - X_1^2) \\ & - a(X_1 + h^{1/2}\sin\theta)(x_1 - h^{1/2}\cos\theta)^2 - b(X_1 + h^{1/2}\sin\theta)^3\Big). \end{aligned} \tag{11}$$

Now, we introduce the variable θ as a new time in the differential equations for x_1 and X_1. After replacing r by (11), and expanding in power series of ε up to the second order, we get

$$\frac{dx_1}{d\theta} = -\frac{\varepsilon}{4\sqrt{2}}\left(2\sqrt{2}\nu X_1 + a(x_1 - h^{1/2}\cos\theta)^2 + 3b(X_1 + h^{1/2}\sin\theta)^2\right) + \varepsilon^2 F(x_1, X_1, \theta; h, \varepsilon),$$
$$\frac{dX_1}{d\theta} = \frac{\varepsilon}{2\sqrt{2}}\left(\sqrt{2}\nu x_1 + a(x_1 - h^{1/2}\cos\theta)(X_1 + h^{1/2}\sin\theta)\right) + \varepsilon^2 G(x_1, X_1, \theta; h, \varepsilon),$$

where F and G are 2π-periodic in θ and satisfy the conditions of Theorem 1 for $h > 0$. Thus, according to this Theorem, the non degenerate equilibrium points of the averaged system give rise to periodic orbits. The equations of the averaged system are given by

$$\frac{x_1}{d\theta} = -\frac{\varepsilon}{8\sqrt{2}}\left(4\sqrt{2}\nu X_1 + a(h + 2x_1^2) + 3b(h + 2X_1^2)\right),$$
$$\frac{dX_1}{d\theta} = \frac{\varepsilon}{2\sqrt{2}}\left(\sqrt{2}\nu x_1 + a x_1 X_1\right).$$

By equating to zero these equations, we obtain the equilibrium points

$$E_{1,2} \equiv \left(\pm\sqrt{\frac{4\nu^2(2a - 3b) - a^2h(1 + 3b)}{2a^3}}, -\frac{\sqrt{2}\nu}{a}\right),$$
$$E_{3,4} \equiv \left(0, \frac{-2\nu \pm \sqrt{4\nu^2 - 3bh(a + 3b)}}{3\sqrt{2}b}\right).$$

Consequently, based on Theorem 1, we can establish the following result

Theorem 2. *For $\varepsilon \neq 0$ sufficiently small and at energy level $h > 0$ of the Hamiltonian $\mathcal{H}$ given in (1) and ω close to one, we find for its associated Hamiltonian system (2) periodic solutions bifurcating from the origin. The number of these periodic solutions depends on the parameters a, b, h and ν. Assuming $a > 0$*

1. *If $4\nu^2(2a - 3b) - a^2h(1 + 3b) > 0$ and $4\nu^2 - 3bh(a + 3b) > 0$, there are four periodic solutions.*
2. *If $(4\nu^2(2a - 3b) - a^2h(1 + 3b))(4\nu^2 - 3bh(a + 3b)) < 0$, there are two periodic solutions.*
3. *If $4\nu^2(2a - 3b) - a^2h(1 + 3b) < 0$ and $4\nu^2 - 3bh(a + 3b) < 0$, there are not periodic solutions.*

Even more, the linear stability of these orbits follows from the stability character of the equilibrium points, which is summarized in Figure 1. A remarkable fact is that, for h small enough, the number of periodic orbits, their bifurcations and stability match with the number of critical points, bifurcations and character of the critical points of the effective potential associated to the original Hamiltonian system given by (1).

The periodic orbits can be computed by inverting the process of averaging. Thus, starting with a value of h and ν and the coordinates of an equilibrium points, once fixed a and b, we recover r from (11) to obtain the original set of coordinates (x, y, X, Y), after using (5). As an example, we depict the four periodic orbits when $\omega = 0.9$, $h = 0.08$ and a and b are the

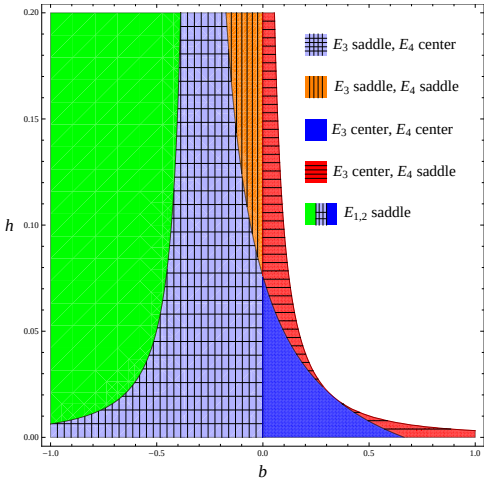


Figure 1: Stability character of the equilibrium points of the averaged system, when $a = 1$, in terms of b and h.

classical Hénon-Heiles parameters ($a = 1$, $b = -1/3$), that can be viewed in the left panel of Figure 2. There are four periodic orbits, a stable one centered at the origin and three unstable orbits that are at the same energy level. As a consequence, there is a heteroclitic connection between the three unstable orbits, allowing a mechanism of transport between different zones of the phase space (see the right panel of Figure 2).

Acknowledgements

This work has been partly supported from the Spanish Ministry of Science and Innovation through the projects MTM2014-59433-CO (subprojects MTM2014-59433-C2-1-P and MTM2014-59433-C2-2-P), MTM2017-88137-CO (subprojects MTM2017-88137-C2-1-P and MTM2017-88137-C2-2-P), and by University of La Rioja through project REGI 2018751.

References

[1] Alfaro, F., Llibre, J., and Pérez-Chavela, E. Periodic orbits for a class of galactic potentials. *Astrophys. Space Sci. 344* (2013), 39–44.

[2] Barrabés, E., Ollè, M., Borondo, F., Farrelly, D., and Mondelo, J. M. Phase space structure of a hydrogen atom in a circularly polarized microwave field. *Rhys. D 241* (2012), 333–349.

[3] Contopoulos, G. *Order and Chaos in Dynamical Astronomy*. Springer-Verlag, New York, 2002.

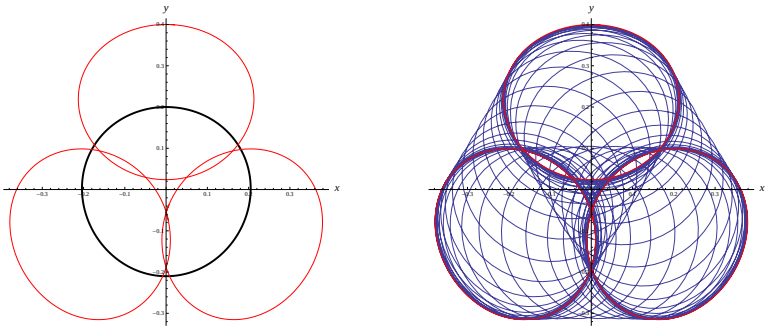


Figure 2: In the left panel, the four periodic orbits for the case $a = 1$, $b = -1/3$, $\omega = 0.9$ and $h = 0.08$. The red color indicates unstable orbits, whereas the black one stands for the stable orbit, centered at the origin. In the right panel, the heteroclitic connection between the three unstable orbits, at the same energy level, is depicted.

[4] Corbera, M., Llibre, J., and Valls, C. Periodic orbits of perturbed non-axially symmetric potentials in 1:1:1 and 1:1:2 resonances. *Discrete Cont. Dyn-B 23* (2018), 2299–2337.

[5] de Zeeuw, T., and Merritt, D. Stellar orbits in a triaxial galaxy. i. orbits in the plane of rotation. *Astrophys. J. 267* (1983), 571–595.

[6] Gómez, G., Koon, W. S., Lo, M. W., Marsden, J. E., Masdemont, J. J., and Ross, S. D. Connecting orbits and invariant manifolds in the spatial restricted three-body problem. *Nonlinearity 17* (2014), 1571–1606.

[7] Hénon, M., and Heiles, C. The applicability of the third integral of motion: Some numerical experiments. *Astron. J. 69* (1964), 73–79.

[8] Iñarrea, M., Lanchares, V., Palacián, J., Pascual, A. I., Salas, J. P., and Yanguas, P. Lyapunov stability for a generalized hénon-heiles system in a rotating reference frame. *Appl. Math. Comput. 253* (2015), 159–171.

[9] Kawai, S., Bandrauk, A. D., Jaffé, C., Bartsch, T., Palacián, J., and Uzer, T. Transition state theory for laser-driven reactions. *J. Chem. Phys. 126* (2007), 164306.

[10] Koon, W. S., Lo, M. W., Marsden, J. E., and Ross, S. D. Heteroclitic connections between periodic orbits and resonance transitions in celestial mechanics. *Chaos 10* (2000), 427–469.

[11] Llibre, J., Pasca, D., and Valls, C. Periodic solutions of a galactic potential. *Chaos Soliton Fract. 61* (2014), 38–43.

[12] Romero-Gómez, M., Masdemont, J. J., García-Gómes, C., and Athanassoula, E. The role of the unstable equilibrium points in the transfer of matter in galactic potentials. *Commun. Nonlinear Sci. 14* (2009), 4123–4138.

[13] Verhulst, F. *Nonlinear Differential Equations and Dynamical Systems*. Springer-Verlag, New York, 1990.

V. Lanchares, M. Iñarrea, A. I. Pascual and J. P. Salas
Universidad de La Rioja.
C/ Madre de Dios, 53. Edificio CCT.
26006, Logroño, La Rioja, Spain.
vlancha@unirioja.es, manuel-inarrea@unirioja.es, aipasc@unirioja.es,
josepablo.salas@unirioja.es

J. Palacián and P. Yanguas
Universidad Púbica de Navarra.
Campus Arrosadía. Edificio Las Encinas.
31006, Pamplona, Navarra, Spain.
palacian@unavarra.es, yanguas@unavarra.es

Monografías Matemáticas *García de Galdeano* **42**, 173–181 (2019)

LIPSCHITZ SPACES ASSOCIATED TO THE HARMONIC OSCILLATOR

Marta de León-Contreras and José L. Torrea

Abstract.
We define Lipschitz classes adapted to the Harmonic Oscillator

$$\mathcal{H} = -\Delta + |x|^2, \quad x \in \mathbb{R}^n.$$

These classes will be defined either through a pointwise condition or through some integral conditions, in this case by using a semigroup approach. We will prove that the different definitions are equivalent. The semigroup approach will allow us to prove regularity properties of some Bessel operators associated to $\mathcal{H}$.

Keywords: Harmonic Oscillator, Lipschitz Hölder spaces, Semigroups.
AMS classification: 35R11, 35R09, 34A08, 26A33.

§1. Introduction

Along this note, we shall denote by $\mathcal{H}$ the Harmonic Oscillator

$$\mathcal{H} = -\Delta + |x|^2, \quad x \in \mathbb{R}^n.$$

Our purpose is to define Lipschitz classes adapted to $\mathcal{H}$. These classes will be defined either through a pointwise condition or through some integral conditions, in this case by using a semigroup approach. We will prove that the different definitions are equivalent. The semigroup approach will allow us to prove regularity properties of some Bessel operators associated to $\mathcal{H}$. Several of the results contained in this note can be found in [1] and [2].

Lipschitz (also called Hölder) spaces are classes of smooth functions which are basic in functional analysis, Fourier analysis and partial differential equations. Roughly speaking, for certain $k \in \mathbb{N} \cup \{0\}$ and $k < \alpha < k + 1$, the space Lipschitz-α is the class of functions that are more regular than C^k (the space of functions whose k-order derivatives are continuous) and less regular than C^{k+1}. Lipschitz spaces are usually defined through pointwise estimates but this approach is not convenient when we want to prove regularity results of some differential operators, because in most of cases it leads to quite involved computations. However, the semigroup description of Lipschitz spaces is really useful for this purpose. This approach was introduced by Taibleson and Stein in the 60's, see [8, 13, 14, 15]. They characterized classical Lipschitz spaces through the heat semigroup, $e^{y\Delta}$ and the Poisson semigroup $e^{-y\sqrt{-\Delta}}$. These characterizations raise the question of analyzing some Hölder spaces associated to different Laplacians and to find the pointwise and semigroup estimate characterizations. In the case of the Ornstein-Ulhenbeck operator $O = -\frac{1}{2}\Delta + x \cdot \nabla$, in [3] some Lipschitz classes were defined by means of its Poisson semigroup, $e^{-y\sqrt{O}}$, and in [4] a pointwise characterization

was obtained for $0 < \alpha < 1$. In the case of the classical parabolic operator $\partial_t - \Delta$, in [12] Lipschitz classes adapted to this operator were characterized through the Poisson semigroup. In the case of the Hermite operator on $\mathbb{R}^n$, $n \geq 1$, $\mathcal{H} = -\Delta + |x|^2$, adapted Hölder classes were defined pointwise in [11]. These last spaces were characterized in [1] by means of the Poisson semigroup, $e^{-y\sqrt{\mathcal{H}}}$, also in the parabolic case. Laplacians More recently, see [2], some spaces have been defined in the case of Schrödinger operators $\mathcal{L} = -\Delta + V$ in $\mathbb{R}^n$, $n \geq 3$, where V is a nonnegative potential satisfing

$$\left(\frac{1}{|B|}\int_B V(y)^q dy\right)^{1/q} \leq \frac{C}{|B|}\int_B V(y)dy, \ \ q > n/2, \text{ for every ball } B. \tag{1}$$

It could be said that the breakdown of the analysis of Schrödinger operators was the paper by Shen, [7]. It relays on estimates of the heat kernel of $e^{-t\mathcal{L}}$. However this method only covers the range $n \geq 3$.

The Harmonic Oscillator is probably the most important example among the family of Schrödinger operators. It has the advantage that the kernel of the heat semigroup, $e^{-t\mathcal{H}}f(x)$ is known explicitly. Along this note we shall show how this fact allows us to built a satisfactory theory of Lipschitz spaces for all $n \geq 1$. In this way we shall complement some results of [1] and [2]. We do not want to be exhaustive in this presentation. However we shall remark those results that are new. Sometimes the proofs will be only suggested.

Definition 1 (Hermite Hölder spaces)**.** Let $0 < \alpha < 2$. We consider the space of functions

$$C^\alpha_{\mathcal{H}}(\mathbb{R}^n) = \left\{f : (1+|\cdot|)^\alpha f(\cdot) \in L^\infty(\mathbb{R}^n), \text{ and } \sup_{|z|>0}\frac{\|f(\cdot+z)+f(\cdot-z)-2f(\cdot)\|_\infty}{|z|^\alpha} < \infty.\right\}$$

with associated norm

$$\|f\|_{C^\alpha_{\mathcal{H}}} = [f]_{M^\alpha} + [f]_{C^\alpha_{\mathcal{H}}}.$$

Where $[f]_{M^\alpha} = \|(1+|\cdot|)^\alpha f(\cdot)\|_\infty$ and $[f]_{C^\alpha_{\mathcal{H}}} = \sup_{|z|>0}\frac{\|f(\cdot+z)+f(\cdot-z)-2f(\cdot)\|_\infty}{|z|^\alpha}$.

Remark 1. This definition was already considered for Schrödinger operators $\mathcal{L}$, where the function $1+|x|$ was substituted by the inverse of the so called critical radius $\rho(x)$, see (4). It can be seen that, if $0 < \alpha < 1$, the last space coincides with the space such that $[f]_{M^\alpha} < \infty$ and $\sup_{|z|>0}\frac{\|f(\cdot+z)-f(\cdot)\|_\infty}{|z|^\alpha}$. This space was defined in [10], [11].

Definition 2. Let $e^{-y\mathcal{H}} = W_y$ and $e^{-y\sqrt{\mathcal{H}}} = P_y$ be the heat and Poisson semigroups associated to $\mathcal{H}$. For $\alpha > 0$ we define the spaces $\Lambda^W_{\alpha/2}$ and Λ^P_α as

(A) $\Lambda^W_{\alpha/2} = \left\{f : [f]_{M^\alpha} < \infty \text{ and } \left\|\partial_y^k W_y f\right\|_{L^\infty(\mathbb{R}^n)} \leq C_k y^{-k+\alpha/2}, \ k = [\alpha/2]+1\right\}$. We shall denote by $S^W_\alpha[f]$ the infimum of the constants C_k above.

(B) $\Lambda^P_\alpha = \left\{f : M^P[f] = \int_{\mathbb{R}^n}\frac{|f(x)|}{(1+|x|)^{n+1}}dx < \infty, \text{ and } \left\|\partial_y^k P_y f\right\|_{L^\infty(\mathbb{R}^n)} \leq B_k y^{-k+\alpha}, \ k = [\alpha]+1\right\}$. We shall denote by $S^P_\alpha[f]$ the infimum of the constants B_k above.

We observe that condition $[f]_{M^\alpha} < \infty$ implies that in particular the function f must be bounded. Moreover, if $f \in \Lambda^P_\alpha$ then $\rho(\cdot)^{-\alpha}f \in L^\infty(\mathbb{R}^n)$, see [2] Theorem 1.9. For $\mathcal{H}$ it is known that $\rho(x) = \frac{1}{1+|x|}$. Therefore we get that $f \in L^\infty(\mathbb{R}^n)$, so Λ^P_α coincides with the space

defined in [1]. We will also see that this condition is natural as soon as either $S^P_\alpha[f] < \infty$ or $S^W_\alpha[f] < \infty$.

The main Theorem of this note is the following

Theorem 1. *Let $0 < \alpha < 2$, $n \geq 1$. The following statements are equivalent:*

$$(1)\ f \in C^\alpha_{\mathcal{H}}(\mathbb{R}^n), \qquad (2)\ f \in \Lambda^W_{\alpha/2}, \qquad (3)\ f \in \Lambda^P_{\alpha/2}.$$

Moreover, the norms of the function f in these spaces are equivalent.

The Theorem was proved to be true in the case $\mathcal{L}$ for $n \geq 3$ and $0 < \alpha \leq 2 - \frac{n}{q}$, where q is the exponent in (1). In [1] it was proved that (1) is equivalent to (3). On the other hand, the proof of (2) implies (3) was given in [2] and it remains valid in this case for any $\alpha > 0$. We include it here as Theorem 6, we believe that is of independent interest. Finally, we sketch the proof of (1) implies (2) at the end of Section 2.

Our second aim is to study the regularity of operators in the Lipschitz spaces. As example of the technique we shall present here the *Bessel potential of order $\beta > 0$*,

$$(Id + \mathcal{H})^{-\beta/2} f(x) = \frac{1}{\Gamma(\beta/2)} \int_0^\infty e^{-t} e^{-t\mathcal{H}} f(x) t^{\beta/2} \frac{dt}{t}.$$

Theorem 2. *Let $\alpha, \beta > 0$. Then, the Bessel potential satisfies*

(i) $\|(Id + \mathcal{H})^{-\beta/2} f\|_{\Lambda^W_{\frac{\alpha+\beta}{2}}} \leq C \|f\|_{\Lambda^W_{\alpha/2}}$.

(ii) $\|(Id + \mathcal{H})^{-\beta/2} f\|_{\Lambda^W_{\beta/2}} \leq C \|f\|_\infty$.

All the results in this note have been proved for the elliptic operator $\mathcal{H}$ but they can be done in the parabolic case parallely, as we did in [1] for the Poisson case.

§2. Proof of Theorem 1.

It is well-known that, for $f \in L^p(\mathbb{R}^n)$, $1 \leq p \leq \infty$, see [9], the heat semigroup associated to $\mathcal{H}$ is given by the Mehler's formula

$$e^{-y\mathcal{H}} f(x) = W_y f(x) = \int_{\mathbb{R}^n} W_y(x,z) f(x-z) dz = \int_{\mathbb{R}^n} \frac{e^{-\frac{|z|^2}{4}\coth y} e^{-\frac{|2x-z|^2}{4}\tanh y}}{(2\pi \sinh(2y))^{n/2}} f(x-z) dz, \tag{2}$$

In addition, by Bochner subordination, for $f \in L^p(\mathbb{R}^n)$, $1 \leq p \leq \infty$, the Poisson semigroup is given by

$$e^{-y\sqrt{\mathcal{H}}} f(x) = P_y f(x) = \frac{y}{2\sqrt{\pi}} \int_0^\infty \int_{\mathbb{R}^n} e^{-y^2/4\tau} \frac{e^{-\frac{|x-z|^2}{4}\coth\tau} e^{-\frac{|x+z|^2}{4}\tanh\tau}}{(2\pi \sinh 2\tau)^{n/2}} f(z)\, dz \frac{d\tau}{\tau^{3/2}}. \tag{3}$$

See [1] for more details.

Remark 2. The following results will be use along the paper. Let $\tau > 0$.

(1) If $\tau < 1$, then $\sinh\tau \sim \tau$, $\cosh\tau \sim C$, $\coth\tau \sim \frac{1}{\tau}$ and $\tanh\tau \sim \tau$.

(2) If $\tau > 1$, then $\sinh\tau \sim e^\tau$, $\cosh\tau \sim e^\tau$, $\coth\tau \sim C$ and $\tanh\tau \sim C$.

(3) Let $z \geq 0$ and $\alpha \geq 0$ there exist a constant $C_\alpha > 0$ such that $z^\alpha e^{-z} \leq C_\alpha e^{-z/2}$.

As usual, by $A \sim B$ we mean there exist constants C_1, C_2 such that $C_1 A \leq B \leq C_2 A$.

We shall also need some expressions of hat kernel acting over constants functions. The reader can be found the proofs of the following formulas in [1].

Lemma 3. *For each $x \in \mathbb{R}^n$ and $\tau > 0$, we have:*

(1) $e^{-y\mathcal{H}}1(x) = \dfrac{e^{-\frac{\tanh(2y)}{2}|x|^2}}{(\cosh(2y))^{n/2}}$.

(2) $|\partial_y e^{-y\mathcal{H}}1(x)| \leq C(\min\{y,1\} + |x|^2)\dfrac{e^{-\frac{\tanh(2y)}{2}|x|^2}}{(\cosh(2y))^{n/2}}$.

Lemma 4. *Let $k \in \mathbb{N}$. Then for every $x \in \mathbb{R}^n$ and $y > 0$,*

$$\Big|\int_{\mathbb{R}^n} \partial_y^k W_y(x,z)dz\Big| \leq \frac{C_k}{y^k}.$$

Proof. For $k = 1$, the proof follows easily by using Remark 2 and the estimate

$$|\partial_y W_y(x,z)| \leq C\frac{e^{-\frac{|2x-z|^2}{c}\tanh y}e^{-\frac{|z|^2}{c}\coth y}}{(\sinh(2y))^{n/2}y}.$$

For $k \geq 1$ the proof is parallel. □

Remark 3. Observe that for bounded functions f, Lemma 4 assures that $\big\|\partial_y W_y f\big\|_{L^\infty(\mathbb{R}^n)} \leq C\|f\|_\infty y^{-1}$. Therefore we can assume in the definition of $\Lambda^W_{\alpha/2}$ that $y < 1$. By subordination the same fact occurs for Λ^P_α.

Proposition 5. *Let $\alpha > 0$,*

- *If $k = [\alpha/2] + 1$ and f is a function satisfying $M^\alpha[f] < \infty$, then $\|\partial_y^k W_y f\|_{L^\infty(\mathbb{R}^n)} \leq C_\alpha y^{-k+\alpha/2}$ if, and only if, for $m \geq k$, $\|\partial_y^m W_y f\|_{L^\infty(\mathbb{R}^n)} \leq C_m y^{-m+\alpha/2}$. Moreover, for each m, C_m and C_α are comparable.*
- *If $k = [\alpha] + 1$ and f is a function satisfying $M^P[f] = \int_{\mathbb{R}^n} \frac{|f(x)|}{(1+|x|)^{n+1}}dx < \infty$, then, $\|\partial_y^k P_y f\|_{L^\infty(\mathbb{R}^n)} \leq C_k y^{-k+\alpha}$ if, and only if, for $m \geq k$, $\|\partial_y^m P_y f\|_{L^\infty(\mathbb{R}^n)} \leq C_m y^{-m+\alpha}$.*

Proof. Let $m \geq [\alpha/2] + 1 = k$. By the semigroup property and Lemma 4 we have

$$\Big|\partial_y^m W_y f(x)\Big| = C\Big|\partial_y^{m-k} W_{y/2}(\partial_u^k W_u f(x)\big|_{u=y/2})\Big| \leq C'_\alpha \frac{1}{y^{m-k}} y^{-k+\alpha/2} = C_m y^{-m+\alpha/2}.$$

For the converse, the fact $|\partial_y^\ell W_y f(x)| \to 0$ as $y \to \infty$, allows us to integrate on y as many times as we need to get $\|\partial_y^k W_y f\|_{L^\infty(\mathbb{R}^n)} \leq C_\alpha\, y^{-k+\alpha/2}$.

For the Poisson semigroup the proof is parallel. □

The following result appears for the first time in [2] in the case of Schrödinger operators.

Theorem 6. *Let $\alpha > 0$. If $f \in \Lambda^W_{\alpha/2}$, then $f \in \Lambda^P_\alpha$. Moreover, $S^P_\alpha[f] \leq C S^W_\alpha[f]$.*

Proof. Let $k = [\alpha/2]+1$ and $f \in \Lambda^W_{\alpha/2}$, then $[\alpha]+1 = [\alpha/2+\alpha/2]+1 \le [\alpha/2]+[\alpha/2]+2 = 2k$. By Proposition 5 it is enough to prove that $\|\partial_y^{2k} P_y f\|_\infty \le C y^{-(2k)+\alpha}$.

Since $\partial_y^2\Big(\frac{ye^{-\frac{y^2}{4\tau}}}{\tau^{3/2}}\Big) = \partial_\tau\Big(\frac{ye^{-\frac{y^2}{4\tau}}}{\tau^{3/2}}\Big)$, k-times integration by parts give

$$
\begin{aligned}
|\partial_y^{2k} P_y f(x)| &= \left|\frac{1}{2\sqrt{\pi}}\int_0^\infty \partial_y^{2k}\left(\frac{ye^{-\frac{y^2}{4\tau}}}{\tau^{3/2}}\right)e^{-\tau\mathcal{H}}f(x)d\tau\right| = \left|\frac{1}{2\sqrt{\pi}}\int_0^\infty \partial_\tau^{k}\left(\frac{ye^{-\frac{y^2}{4\tau}}}{\tau^{3/2}}\right)e^{-\tau\mathcal{H}}f(x)d\tau\right| \\
&= \frac{1}{2\sqrt{\pi}}\left|\int_0^\infty (-1)^k\left(\frac{ye^{-\frac{y^2}{4\tau}}}{\tau^{3/2}}\right)\partial_\tau^k e^{-\tau\mathcal{L}}f(x)d\tau\right| \le C\, S^W_\alpha[f]\int_0^\infty \frac{ye^{-\frac{y^2}{4\tau}}}{\tau^{3/2}}\tau^{-k+\alpha/2}d\tau \\
&\le C\, S^W_\alpha[f] y^{-2k+\alpha}.
\end{aligned}
$$

□

Remark 4. It is clear that if f is a function such that $(1+|\cdot|)^\alpha f \in L^\infty(\mathbb{R}^n)$, then $f \in L^\infty(\mathbb{R}^n)$. Therefore, the following Remark 5 establishes that in the definition of $\Lambda^W_{\alpha/2}$, we can consider indistinctly $f \in L^\infty(\mathbb{R}^n)$ or $(1+|x|)^\alpha f \in L^\infty(\mathbb{R}^n)$. The proposition was proved in the case $\mathcal{L}$ in [2].

Remark 5. Let $\alpha > 0$. If f is a bounded function such that $\|\partial_y^m W_y f\|_{L^\infty(\mathbb{R}^n)} \le C_m y^{-m+\alpha/2}$, $m = [\alpha/2]+1$, then $|x|^\alpha f \in L^\infty(\mathbb{R}^n)$.

Proof. We shall do the proof only in the case $0 < \alpha < 1$, for the other cases see [2]. Now for $|x| > 1$ and $0 < \alpha < 1$ we have

$$
\begin{aligned}
|x|^\alpha|f(x)| &\le |x|^\alpha \sup_{0<y<\frac{1}{|x|}}|W_y f(x)| \le |x|^\alpha \sup_{0<y<\frac{1}{|x|}}\Big(|W_y f(x) - W_{\frac{1}{|x|}}f(x)| + |W_{\frac{1}{|x|}}f(x)|\Big) \\
&\le |x|^\alpha \sup_{0<y<\frac{1}{|x|}}\left|\int_y^{\frac{1}{|x|}}\partial_{z_1}W_{z_1}f(x)dz_1\right| + C\|f\|_{\Lambda^{W_y}_{\alpha/2}} \\
&\le C|x|^\alpha \sup_{0<y<\frac{1}{|x|}}\left|\int_y^{\frac{1}{|x|}} z_1^{-1+\alpha}dz_1\right| + C\|f\|_{\Lambda^{W_y}_{\alpha/2}} \le C.
\end{aligned}
$$

□

Now we shall prove that (1) implies (2) in Theorem 1.

Suppose that f is a function that satisfies the conditions in (1). Let $y < 1$. By using that $\int_{\mathbb{R}^n}\partial_y W_y(x,z)f(x+z)dz = \int_{\mathbb{R}^n}\partial_y W_y(x,-z)f(x-z)dz$, we can write

$$
\begin{aligned}
\partial_y W_y f(x) &= \frac{1}{2}\int_{\mathbb{R}^n}\partial_y W_y(x,z)(f(x-z)+f(x+z)-2f(x))dz \\
&\quad + \frac{1}{2}\int_{\mathbb{R}^n}\Big(\partial_y W_y(x,z) - \partial_y W_y(x,-z)\Big)f(x-z)dz + f(x)\partial_y e^{-y\mathcal{H}}1(x) \\
&= I + II + III.
\end{aligned}
$$

On the one hand, by using Remark 2 and Lemma 4 we have that

$$|I| \leq C \int_{\mathbb{R}^n} \frac{e^{-\frac{|z|^2 \coth y}{c}} e^{-\frac{|2x-z|^2 \tanh y}{c}} |z|^\alpha}{(\sinh(2y))^{n/2} y} dz \leq C y^{-1+\alpha/2}.$$

Regarding II, observe that

$$\begin{aligned}
\left| \partial_y W_y(x,z) - \partial_y W_y(x,-z) \right| &= \left| \partial_y \left(\frac{e^{-\frac{|z|^2 \coth y}{4}}}{(2\pi \sinh(2y))^{n/2}} \left[e^{-\frac{|2x-z|^2 \tanh y}{4}} - e^{-\frac{|2x+z|^2 \tanh y}{4}} \right] \right) \right| \\
&= \left| \partial_y \left(\frac{e^{-\frac{|z|^2 \coth y}{4}}}{(2\pi \sinh(2y))^{n/2}} \right) \left[e^{-\frac{|2x-z|^2 \tanh y}{4}} - e^{-\frac{|2x+z|^2 \tanh y}{4}} \right] \right. \\
&\quad \left. + \frac{e^{-\frac{|z|^2 \coth y}{4}}}{(2\pi \sinh(2y))^{n/2}} \partial_y \left[e^{-\frac{|2x-z|^2 \tanh y}{4}} - e^{-\frac{|2x+z|^2 \tanh y}{4}} \right] \right| \\
&\leq C e^{-\frac{|z|^2 \coth y}{4}} \left(\frac{|z|^2}{(\sinh(y))^2 (\sinh(2y))^{n/2}} + \frac{\coth(2y)}{(\sinh(2y))^{n/2}} \right) \left| e^{-\frac{|2x-z|^2 \tanh y}{4}} - e^{-\frac{|2x+z|^2 \tanh y}{4}} \right| \\
&\quad + \frac{e^{-\frac{|z|^2 \coth y}{4}}}{(2\pi \sinh(2y))^{n/2}} \left| \int_{-1}^{1} \partial_\theta \partial_y \left(e^{-\frac{|2x-\theta z|^2 \tanh y}{4}} \right) d\theta \right| \\
&= II_a + II_b.
\end{aligned}$$

Observe that

$$\begin{aligned}
\left| e^{-\frac{|2x-z|^2 \tanh y}{4}} - e^{-\frac{|2x+z|^2 \tanh y}{4}} \right| &= \left| \int_{-1}^{1} \partial_\theta e^{-\frac{|2x-\theta z|^2 \tanh y}{4}} d\theta \right| = \left| \int_{-1}^{1} \nabla_z (e^{-\frac{|2x-\theta z|^2 \tanh y}{4}}) \cdot z \, d\theta \right| \\
&= \left| \int_{-1}^{1} e^{-\frac{|2x-\theta z|^2 \tanh y}{4}} \left(\frac{\theta \tanh y}{2} (2x - \theta z) \cdot z \right) d\theta \right| \\
&\leq C |z| (\tanh y)^{1/2}.
\end{aligned}$$

Therefore, by using Remark 2 we have that

$$\begin{aligned}
|II_a| &\leq C e^{-\frac{|z|^2 \coth y}{4}} \left(\frac{|z|^3 (\tanh y)^{1/2}}{(\sinh(y))^2 (\sinh(2y))^{n/2}} + \frac{\coth(2y) |z| (\tanh y)^{1/2}}{(\sinh(2y))^{n/2}} \right) \\
&\leq C e^{-\frac{|z|^2}{cy}} \left(\frac{|z|^3}{y^{3/2+n/2}} + \frac{|z|}{y^{1/2+n/2}} \right) \leq C \frac{e^{-\frac{|z|^2}{cy}}}{y^{n/2}}.
\end{aligned}$$

On the other hand, since

$$\begin{aligned}\Big|\int_{-1}^{1}\partial_\theta\partial_y\left(e^{-\frac{|2x-\theta z|^2\tanh y}{4}}\right)d\theta\Big| &= \left|\int_{-1}^{1}\nabla_z\partial_y\left(e^{-\frac{|2x-\theta z|^2\tanh y}{4}}\right)\cdot z\,d\theta\right| \\ &= \left|\int_{-1}^{1}\partial_y\left(e^{-\frac{|2x-\theta z|^2\tanh y}{4}}\frac{\theta\tanh y}{2}(2x-\theta z)\cdot z\right)d\theta\right| \\ &= \left|\int_{-1}^{1}e^{-\frac{|2x-\theta z|^2\tanh y}{4}}\left(-\frac{\theta\tanh y}{2}\frac{|2x-\theta z|^2}{4\cosh^2(y)}(2x-\theta z)\cdot z+\frac{\theta(2x-\theta z)\cdot z}{2\cosh^2 y}\right)d\theta\right| \\ &\le C\frac{|z|}{(\tanh y)^{1/2}\cosh^2 y},\end{aligned}$$

we have that $|II_b|\le C\frac{e^{-\frac{|z|^2\coth y}{4}}}{(2\pi\sinh(2y))^{n/2}}\frac{|z|}{(\tanh y)^{1/2}\cosh^2 y}\le C\frac{e^{-\frac{|z|^2}{cy}}}{y^{n/2}}$. Estimates II_a and II_b and the fact that $y<1$ allow us to get $|II|\le C\|f\|_\infty y^{-1+\alpha/2}$.

Finally, by using Remark 2 and Lemma 3 (2) we get

$$\begin{aligned}|III| &\le C|f(x)|(1+|x|^2)\frac{e^{-\frac{\tanh(2y)|x|^2}{2}}}{(\cosh(2y))^{n/2}}\le C|f(x)|(1+|x|^2)e^{-cy|x|^2} \\ &\le C([f]_{M^\alpha}+\|f\|_\infty)y^{-1+\alpha/2}.\end{aligned}$$

This is the end of the proof of Theorem 1.

§3. Proof of Theorem 2.

Since $\|W_y f\|_\infty\le C\|f\|_\infty$ and $\|\partial_y^\ell W_y f\|_\infty\le C\frac{\|f\|_\infty}{y^\ell}$ for $\ell\in\mathbb{N}$, we can apply Fubini's Theorem and the derivatives and the integral commute.

Let $f\in\Lambda^W_{\alpha/2}$ and $\ell=[\alpha/2+\beta/2]+1$. Then

$$\begin{aligned}|\partial_y^\ell W_y((Id+\mathcal{H})^{-\beta/2}f(x))| &= \left|\frac{1}{\Gamma(\beta/2)}\int_0^\infty e^{-t}\partial_y^\ell W_y(W_t f)(x)t^{\beta/2}\frac{dt}{t}\right| \\ &\le C\int_0^\infty e^{-t}(\partial_w^\ell W_w f(x)\Big|_{w=y+t})t^{\beta/2}\frac{dt}{t} \\ &\le C\int_0^\infty e^{-t}(y+t)^{-\ell+\alpha/2}t^{\beta/2}\frac{dt}{t} \\ &\overset{\frac{t}{y}=u}{\le} Cy^{\alpha/2+\beta/2-\ell}\int_0^\infty\frac{u^{\beta/2}e^{-yu}}{(1+u)^{\ell-\alpha/2}}\frac{du}{u} \\ &\le Cy^{\alpha/2+\beta/2-\ell}.\end{aligned}$$

When $f\in L^\infty(\mathbb{R}^n)$ we proceed analogously by using that, for $\ell=[\beta/2]+1$, $\|\partial_y^\ell W_y W_\nu f\|_\infty\le C\frac{\|f\|_\infty}{y^\ell}$.

§4. Remarks about Schrödinger operators

It can be checked that the Harmonic Oscillator, $V(x) = |x|^2$, satisfies condition (1) for all $q < \infty$. On the other hand, one of the fundamental tools in the theory of the operator $\mathcal{L}$ is the so called "critical radius" $\rho(x)$, $x \in \mathbb{R}^n$, defined as

$$\rho(x) := \sup\Big\{r > 0 : \frac{1}{r^{n-2}} \int_{B(x,r)} V(y)dy \leq 1\Big\}. \tag{4}$$

In the case $\mathcal{H}$, $\rho(x) = \frac{1}{1+|x|}$. The function $\frac{1}{1+|x|}$, appeared before the paper by Shen, [7], in the historical work of Muckenhoupt, [6]. It was related with the Ornstein-Uhlenbeck operator on the line. For the interested reader we refer to the paper [5]. In that paper it is shown that $\frac{1}{1+|x|}$, $n \geq 1$, shares all the properties of $\rho(x)$. Of particular interest is the existence of a covering of the space with balls of type $B(x, \frac{1}{1+|x|})$. All these remarks together say that Theorem 1 could also be proved by changing in an appropriated way the proof given for the operator $\mathcal{L}$ in [2].

Acknowledgements

The authors have been partially supported by MTM2015-66157-C2-1-P, MINECO-FEDER.

References

[1] De León-Contreras, M., and Torrea, J. L. Fractional powers of the parabolic Hermite operator. Regularity properties. *arXiv:1708.02788*.

[2] De León-Contreras, M., and Torrea, J. L. Lipschitz spaces adapted to Schrödinger operators and regularity properties. *arXiv:1901.06898*.

[3] Gatto, A. E., and Urbina R, W. O. On Gaussian Lipschitz spaces and the boundedness of fractional integrals and fractional derivatives on them. *Quaestiones Mathematicae 38*, 1 (2015), 1–25.

[4] Liu, L., and Sjögren, P. A characterization of the Gaussian Lipschitz space and sharp estimates for the Ornstein-Uhlenbeck Poisson kernel. *Revista matemática iberoamericana 32*, 4 (2016), 1189–1210.

[5] Martínez, T. Extremal spaces related to Schrödinger operators with potentials satisfying a reverse Hölder inequality. *Rev. Un. Mat. Argentina 45*, 1 (2004), 43–61 (2005).

[6] Muckenhoupt, B. Poisson integrals for Hermite and Laguerre expansions. *Trans. Amer. Math. Soc. 139* (1969), 231–242.

[7] Shen, Z. W. L^p estimates for Schrödinger operators with certain potentials. *Ann. Inst. Fourier (Grenoble) 45*, 2 (1995), 513–546.

[8] Stein, E. M. *Singular integrals and differentiability properties of functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970.

[9] Stempak, K., and Torrea, J. L. Poisson integrals and Riesz transforms for Hermite function expansions with weights. *J. Funct. Anal. 202*, 2 (2003), 443–472.

[10] Stinga, P. R., and Torrea, J. L. Extension problem and Harnack's inequality for some fractional operators. *Comm. Partial Differential Equations 35*, 11 (2010), 2092–2122.

[11] Stinga, P. R., and Torrea, J. L. Regularity theory for the fractional harmonic oscillator. *J. Funct. Anal. 260*, 10 (2011), 3097–3131.

[12] Stinga, P. R., and Torrea, J. L. Regularity theory and extension problem for fractional nonlocal parabolic equations and the master equation. *SIAM J. Math. Anal. 49*, 5 (2017), 3893–3924.

[13] Taibleson, M. H. On the theory of Lipschitz spaces of distributions on Euclidean n-space. I. Principal properties. *J. Math. Mech. 13* (1964), 407–479.

[14] Taibleson, M. H. On the theory of Lipschitz spaces of distributions on Euclidean n-space. II. Translation invariant operators, duality, and interpolation. *J. Math. Mech. 14* (1965), 821–839.

[15] Taibleson, M. H. On the theory of Lipschitz spaces of distributions on Euclidean n-space. III. Smoothness and integrability of Fourier tansforms, smoothness of convolution kernels. *J. Math. Mech. 15* (1966), 973–981.

M. de León-Contreras
Departamento de Matemáticas,
Facultad de Ciencias, Universidad Autónoma de Madrid,
28049 Madrid, Spain.
marta.leon@uam.es

J. L. Torrea
Departamento de Matemáticas,
Facultad de Ciencias, Universidad Autónoma de
Madrid,
28049 Madrid, Spain.
joseluis.torrea@uam.es

Monografías Matemáticas *García de Galdeano* **42**, 183–192 (2019)

ACCURATE LEAST SQUARES FITTING WITH A GENERAL CLASS OF SHAPE PRESERVING BASES

Esmeralda Mainar, Juan Manuel Peña and Beatriz Rubio

Abstract. In this paper we consider the problem of least squares fitting with a very general class of bases with interest in Computer Aided Geometric Design and Approximation Theory. We compute a factorization of the collocation matrix A of these bases that allows us to obtain a QR decomposition of A. Then the triangular system corresponding to the matrix factor R is solved using a bidiagonal factorization of this matrix. Numerical experiments show the accuracy of this procedure.

Keywords: B-basis, Bidiagonal decompositions, Least Squares, Accurate computations.
AMS classification: 65D17, 65F05, 65D05, 41A05, 42A10.

§1. Introduction

The accurate computation with structured classes of matrices is an important issue in Numerical Linear Algebra and it is receiving increasing attention in the recent years (cf. [10, 23, 7]). For this purpose, a parametrization adapted to the structure of the considered matrices is needed. Let us recall that an algorithm can be performed with high relative accuracy (HRA) when it only uses products, quotients, additions of numbers with the same sign or subtractions of initial data (cf. [11]). Performing an algorithm with HRA is a very desirable goal because it implies that the relative errors of the computations are of the order of the machine precision, independently of the size of the condition number of the considered problem. Bidiagonal factorizations provide a parametrization that has played a crucial role to derive algorithms with HRA for some classes of totally positive (TP) matrices. In this case, the mentioned bidiagonal factorizations can be explicitly computed by means of an elimination process called Neville elimination (cf. [12]). When the bidiagonal factorization of the considered matrix is obtained with HRA, the computation of the inverse matrix, its eigenvalues and singular values, the solutions of some linear systems or the computation of its QR factorization can be also performed with HRA using the algorithms presented by Koev in [17] and [16]. Up to now, this has been achieved with some relevant subclasses of TP matrices with applications to Computer Aided Geometric Design (cf. [22, 6, 7, 23, 19]), to Finance (cf. [5]) or to Combinatorics (cf. [8]).

In Computer Aided Geometric Design shape preserving representations are associated with normalized totally positive (NTP) bases because parametric curves inherit the geometric properties of their control polygons with respect to these bases. Among all NTP bases of a given space of functions, there exists a unique normalized B-basis, which is the basis with optimal shape preserving properties (cf. [24], [4]). The Bernstein bases and the B-spline bases are the normalized B-bases of their corresponding spaces. The matrices considered in [6, 9]

are collocation matrices of polynomial rational functions. However the rational model has several drawbacks (see [20]). Rational curves require additional parameters (weights), which do not have an evident geometric meaning and whose selection is often unclear. In addition, the behavior of rational bases with respect to differentiation and integration operations, is particularly unpleasant and the exact integration of rational curves is hard and requires (whenever possible) involved non rational forms. On the other hand, the rational model cannot encompass transcendental curves such as the helix or the cycloid, which are of interest in many applications. Furthermore the parametrization of conic sections does not correspond to the natural arc-length parametrization, so given uniform partitions in the parameter space we can get unevenly spaced points. Therefore, non-polynomial basis functions (such as trigonometric functions, hyperbolic functions or their mixtures with polynomials) are often used to represent some typical curves or surfaces without rational forms. In [19] algorithms for the computation of the bidiagonal decomposition of square collocation matrices of a very general class of non-polynomial bases with interest in Computer Aided Geometric Design and Approximation Theory are provided. The obtained algorithms are used in [19] to perform accurate algebraic computations, such as the calculation of their inverses, their eigenvalues or their singular values. In this paper, following the approach of [21] for a polynomial case, we generalize the mentioned bidiagonal factorizations to the case of rectangular collocation matrices. Using their QR decompositions, we focus on the problem of least squares fitting in the spaces generated by the general class of bases defined in [19]. By computing the bidiagonal decomposition of the coefficient matrix of the least squares problem, an algorithm for the computation of its QR decomposition is then applied. Finally, using the bidiagonal decomposition of the matrix factor R, a triangular system is solved.

The layout of the paper is as follows. Section 2 includes matrix notations basic concepts and tools. We also recall the Neville elimination procedure, which allows us to introduce the bidiagonal factorization of a square strictly totally positive matrix. Section 3 introduces the class of fg-Bernstein bases and recalls the bidiagonal factorization of the collocation matrices associated to these bases derived in [19]. In Section 4, we generalize these decompositions to the case of rectangular matrices. Then a procedure for computing the solution of the least squares problems in the space generated by fg-Bernstein bases is obtained. Finally, Section 5 shows numerical examples with accurate results obtained when we apply the explained procedure.

§2. Basic notations and auxiliary results

A matrix is *totally positive* (TP) if all its minors are nonnegative and *strictly totally positive* (STP) if they are positive (see [1]). A system of functions $(u_0, \dots, u_n)$ defined on $I \subseteq \mathbb{R}$ is TP if all its collocation matrices

$$\left(u_{j-1}(t_i)\right)_{1\le i\le l+1;1\le j\le n+1}, \quad t_1 < \cdots < t_{l+1} \text{ in } I$$

are TP. A TP system of functions on I is *normalized* (NTP) if $\sum_{i=0}^{n} u_i(t) = 1$, for all $t \in I$. NTP bases are commonly used in Computer Aided Geometric Design due to their shape preserving properties (see [3], [24]).

Among all NTP bases of a space, we can find a unique normalized B-basis, which is the optimal shape preserving basis (cf. [4]). For instance, the Bernstein bases and the B-spline bases are the normalized B-bases of their corresponding spaces. The following characterization of a B-basis is a consequence of Corollary 3.10 of [4] and Proposition 3.11 of [4].

Theorem 1. *Let $(u_0, \dots, u_n)$ be a TP basis of a space $\mathcal{U}$. Then $(u_0, \dots, u_n)$ is a B-basis if for any other TP basis $(v_0, \dots, v_n)$ of $\mathcal{U}$ the matrix K of change of basis such that $(v_0, \dots, v_n) = (u_0, \dots, u_n)K$ is TP.*

Let us now recall some basic matrix notations and results on Neville elimination. Our notation follows the notation used in [12, 15]. Given $n \in \mathbb{N}$ and $k \in \{1, \dots, n\}$, let $Q_{k,n}$ be the set of increasing sequences of k positive integers less than or equal to n. If $\alpha, \beta \in Q_{k,n}$, we denote by $A[\alpha|\beta]$ the $k \times k$ submatrix of A containing rows of places α and columns of places β.

Neville elimination is a procedure to make zeros in a column of a matrix by adding to a given row an appropriate multiple of the previous one (see [12, 15]). For a given nonsingular matrix $A = (a_{i,j})_{1 \le i,j \le n}$, let us present this elimination procedure for the case that no row exchanges are necessary. Neville elimination consists of at most $n-1$ successive major steps, resulting in the sequence of matrices:

$$A^{(1)} := A \to A^{(2)} \to \cdots \to A^{(n)} = U.$$

For $1 \le k \le n-1$, $A^{(k+1)} = (a_{i,j}^{(k+1)})_{1 \le i,j \le n}$ is obtained from $A^{(k)} = (a_{i,j}^{(k)})_{1 \le i,j \le n}$ by defining

$$a_{i,j}^{(k+1)} := a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{i-1,k}^{(k)}} a_{i-1,j}^{(k)} \quad \text{if } a_{i-1,k}^{(k)} \neq 0, \quad k+1 \le i, j \le n,$$

so that $A^{(k+1)}$ has zeros below its main diagonal in the k first columns. Finally, U is an upper triangular matrix. The element $p_{i,j} := a_{i,j}^{(j)}$, $1 \le j \le i \le n$, is called the (i, j) pivot of the Neville elimination of A. The pivots $p_{i,i}$ are called diagonal pivots. The Neville elimination can be performed without row exchanges if all the pivots are nonzero and, in this case, Lemma 2.6 of [12] implies that $p_{i,1} = a_{i,1}$, $1 \le i \le n$, and

$$p_{i,j} = \frac{\det A[i-j+1, \dots, i | 1, \dots, j]}{\det A[i-j+1, \dots, i-1 | 1, \dots, j-1]}, \quad 1 < j \le i \le n. \tag{1}$$

Furthermore, the (i, j) multiplier of the Neville elimination of A is

$$m_{i,j} := \frac{a_{i,j}^{(j)}}{a_{i-1,j}^{(j)}} = \frac{p_{i,j}}{p_{i-1,j}}, \quad 1 \le j < i \le n. \tag{2}$$

Neville elimination has been used to characterize TP and STP matrices (see [12, 15]). From Theorem 4.1 of [12] and p. 116 of [15], a given matrix A is STP if and only if the Neville elimination of A and A^T can be performed without row exchanges, all the multipliers of the Neville elimination of A and A^T are positive and all the diagonal pivots of the Neville elimination of A are positive.

Bidiagonal factorizations have played a crucial role to derive, for TP matrices, algorithms with HRA (cf. [16]). According to the arguments of p.116 of [15], an STP matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ can be factorized, in a unique way under certain conditions, in the form

$$A = F_n F_{n-1} \cdots F_1 D G_1 \cdots G_{n-1} G_n, \tag{3}$$

where F_i and G_i are the lower and upper triangular bidiagonal matrices

$$F_i = \begin{pmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 0 & 1 & & & & \\ & & & m_{i+1,1} & 1 & & & \\ & & & & \ddots & \ddots & & \\ & & & & & m_{n+1,n+1-i} & 1 \end{pmatrix},$$

$$G_i^T = \begin{pmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 0 & 1 & & & & \\ & & & \hat{m}_{i+1,1} & 1 & & & \\ & & & & \ddots & \ddots & & \\ & & & & & \hat{m}_{n+1,n+1-i} & 1 \end{pmatrix}, \tag{4}$$

and $D = \operatorname{diag}(p_{1,1}, \ldots, p_{n+1,n+1})$. The entries $m_{i,j}$ and $\hat{m}_{i,j}$ are the multipliers of the Neville elimination of A and A^T, respectively, and the diagonal entries $p_{i,i}$ are the diagonal pivots of the Neville elimination of A. In fact a unique bidiagonal factorization can be obtained for nonsingular TP matrices (see [14, 15]).

§3. The class of fg-Bernstein bases

Let us suppose that $I = [a, b]$ and $f, g : I \to \mathbb{R}$ are nonnegative continuous functions. For a given $n \in \mathbb{N}$, the corresponding fg-Bernstein basis of order n was defined in [19] as

$$(u_0^n, \ldots, u_n^n), \quad u_k^n(t) := \binom{n}{k} f^k(t) g^{n-k}(t), \quad t \in [a, b], \quad k = 0, \ldots, n. \tag{5}$$

The following result corresponds to Proposition 19 of [18] and characterizes when the fg-Bernstein basis defined in (5) is a B-basis.

Proposition 2. *The system given in* (5) *is a B-basis if and only if the function f/g defined on $I_0 := \{t \in I \mid g(t) \neq 0\}$ is increasing and satisfies*

$$\inf\left\{\frac{f(t)}{g(t)} \mid t \in I_0\right\} = 0, \quad \sup\left\{\frac{f(t)}{g(t)} \mid t \in I_0\right\} = +\infty. \tag{6}$$

Theorem 2 of [19] proves that, given nonnegative $f, g : I \to \mathbb{R}$ such that $f(t) \neq 0, g(t) \neq 0$, $\forall t \in (a, b)$ and f/g is a strictly increasing function, then

$$A := \left(\binom{n}{j-1} f^{j-1}(t_i) g^{n-j+1}(t_i)\right)_{1\leq i,j\leq n+1}, \quad a < t_1 < \cdots < t_{n+1} < b, \tag{7}$$

is STP. Moreover, in Theorem 3 of [19], the following bidiagonal decomposition (3) of the collocation matrices (7) was deduced

$$A = F_n F_{n-1} \cdots F_1 D G_1 \cdots G_{n-1} G_n, \tag{8}$$

where F_i and G_i, $1 \leq i \leq n$, are the lower and upper triangular bidiagonal matrices of the form (4) and $D = \operatorname{diag}(p_{1,1}, \ldots, p_{n+1,n+1})$. The entries $m_{i,j}, \hat{m}_{i,j}$ and $p_{i,i}$ are given by

$$\begin{aligned}
m_{i,j} &= \frac{g^{n-j+1}(t_i) g(t_{i-j})}{g^{n-j+2}(t_{i-1})} \frac{\prod_{k=1}^{j-1} \left(f(t_i)g(t_{i-k}) - f(t_{i-k})g(t_i)\right)}{\prod_{k=2}^{j} \left(f(t_{i-1})g(t_{i-k}) - f(t_{i-k})g(t_{i-1})\right)}, \\
\hat{m}_{i,j} &= \frac{n-i+2}{i-1} \frac{f(t_j)}{g(t_j)}, \quad 1 \leq j < i \leq n+1, \\
p_{i,i} &= \binom{n}{i-1} \frac{g^{n-i+1}(t_i)}{\prod_{k=1}^{i-1} g(t_k)} \prod_{k=1}^{i-1} \left(f(t_i)g(t_k) - f(t_k)g(t_i)\right), \quad 1 \leq i \leq n+1.
\end{aligned} \tag{9}$$

Let us observe that a sufficient condition to obtain the bidiagonal decomposition of A with HRA is that the expressions $f(t_i)$, $g(t_i)$ and $f(t_i)g(t_k) - f(t_k)g(t_i)$, for all $k < i$, can be computed with HRA.

There are many interesting choices of functions f and g satisfying conditions (6) and allowing us the definition of B-bases whose STP collocation matrices can be factorized as in (8). For example, if

$$f(t) := \frac{t-a}{b-a}, \quad g(t) := \frac{b-t}{b-a}, \quad t \in [a, b],$$

the basis (5) is the Bernstein basis of the space of polynomials of degree not greater than n on the compact interval $[a, b]$. Let us observe that, in this case, the computation of $f(t_i)$, $g(t_i)$ and $f(t_i)g(t_k) - f(t_k)g(t_i) = (t_i - t_k)/(b-a)$, $k < i$, can be performed with HRA because it only requires quotients and subtractions of the initial data. Therefore we can also guarantee that the bidiagonal decomposition (8) of the corresponding collocation matrices (7) can be obtained with HRA. We can also consider

$$f(t) := t^2, \quad g(t) := 1 - t^2, \quad t \in [0, 1].$$

Taking into account Proposition 2, we deduce that the system (5) is the normalized B-basis of the space $\langle 1, t^2, \ldots, t^{2n} \rangle$ of even polynomials of degree less than or equal to $2n$ on $[0, 1]$. Let us also observe that the computation of $f(t_i)$, $g(t_i)$ and $f(t_i)g(t_k) - f(t_k)g(t_i) = t_i^2 - t_k^2 = (t_i + t_k)(t_i - t_k)$, $k < i$, requires additions, products and subtractions of the initial data, therefore it can be done with HRA. Again, we can guarantee that the bidiagonal decomposition (8) of the corresponding collocation matrices (7) can be obtained with HRA.

Another particular case can be given by considering the functions

$$f(t) := \sin^2(t/2) = (1 - \cos t)/2, \quad g(t) := \cos^2(t/2) = (1 + \cos t)/2, \quad t \in I = [0, \pi]. \tag{10}$$

In [24] it was proved that the system (5) is the normalized B-basis of the space of even trigonometric polynomials $\langle 1, \cos t, \cos 2t, \ldots, \cos nt\rangle$ on I. On the other hand, if we consider $0 < \Delta < \pi/2$ and

$$f(t) := \sin((\Delta + t)/2), \quad g(t) := \sin((\Delta - t)/2), \quad t \in I = [-\Delta, \Delta], \tag{11}$$

for a given $n = 2m$, the system (5) is a basis that coincides, up to a positive scaling, with the normalized B-basis of the space $\langle 1, \cos t, \sin t, \ldots, \cos mt, \sin mt\rangle$ of trigonometric polynomials of degree less than or equal to m on I (see Section 3 of [25]). Finally, for any $\Delta > 0$, we can also consider

$$f(t) := \sinh((\Delta + t)/2)), \quad g(t) := \sinh((\Delta - t)/2), \quad t \in I = [-\Delta, \Delta]. \tag{12}$$

For $n = 2m$, the system (5) is a B-basis of the space $\langle 1, e^t, e^{-t}, \ldots, e^{mt}, e^{-mt}\rangle$ of hyperbolic polynomials of degree less than or equal to m on I.

In the last three cases, taking into account that $f(t_i)g(t_k) - f(t_k)g(t_i)$ is equal to $(\cos(t_k) - \cos(t_i))/2$, for the functions f and g defined in (10), $\sin(\Delta)\sin((t_i - t_k)/2)$, for the functions f and g defined in (11) and $\sinh(\Delta)\sinh((t_i - t_k)/2)$, for the functions f and g defined in (12), the computation with HRA of the corresponding bidiagonal decomposition (8) should require the evaluation with HRA of the involved trigonometric or hyperbolic functions. Although this cannot be guaranteed, Section 5 and the numerical experiments in [19] show that accurate algebraic computations with the collocation matrices associated to these non-polynomial bases functions can be performed.

§4. Accurate least squares fitting with fg-Bernstein bases

Let us suppose that f and g are functions defined on $[a, b]$ such that $f(t) \neq 0$, $g(t) \neq 0$, $\forall t \in (a, b)$, and f/g is a strictly increasing function. Given a set of parameters $a < t_1 < \cdots < t_{l+1} < b$ and real values $p_1 < \cdots < p_{l+1}$, for some $n \leq l$, we want to compute a function

$$p(t) := \sum_{j=1}^{n+1} c_j \binom{n}{j-1} f^{j-1}(t) g^{n-j+1}(t), \quad t \in [a, b],$$

minimizing the sum of the squares of the deviations from the data $\sum_{i=1}^{l+1} |p_i - p(t_i)|^2$. In order to compute the coefficients of $p(t)$ with respect to the considered fg-Bernstein basis we have to solve, in the least square sense, the overdeterminated linear system $Ac = p$, where

$$A := \left(\binom{n}{j-1} f^{j-1}(t_i) g^{n-j+1}(t_i)\right)_{1 \leq i \leq l+1; 1 \leq j \leq n+1}$$

is the rectangular collocation matrix of the fg-Bernstein basis corresponding to the nodes $t_1 < \cdots < t_{l+1}$, $p = (p_1, \ldots, p_{l+1})^T$ is the data vector and $c = (c_1, \ldots, c_{n+1})^T$ is the vector with the coefficients we want to compute. Using Theorem 2 of [19], we can easily deduce that A is STP and so has maximal rank $n + 1$. Therefore this problem has a unique solution, which is given by the solution of the linear system

$$A^T A c = A^T p.$$

Solving the previous normal equations is a worse conditioned problem than computing the solution through the QR decomposition of the coefficient matrix A, which is the usual approach. In [16] an efficient algorithm for computing the QR decomposition of an STP matrix A is presented. In [17] the Matlab or Octave library `TNQR`, containing an implementation of the mentioned last algorithm, is available. Assuming that the bidiagonal factorization of A is known, `TNQR` computes the matrix Q and the bidiagonal factorization of the matrix R with HRA. Now, following the approach of [21], we shall describe how to solve our least squares problem by means of a bidiagonal decomposition for rectangular matrices that generalizes the bidiagonal factorization described, for the square case, in the previous section and the QR decomposition provided by `TNQR`.

In order to compute the solution of the least squares problem, we define the $(l+1)\times(n+1)$ matrix M such that

$$\begin{aligned} &M_{i,i} := p_{i,i}, \quad && i = 1,\dots,n+1,\\ &M_{i,j} := m_{i,j}, \quad && j = 1,\dots,n+1;\ \ i = j+1,\dots,l+1,\\ &M_{i,j} := \hat{m}_{i,j}, \quad && i = 1,\dots,n;\ \ j = i+1,\dots,n+1, \end{aligned}$$

where the $m_{i,j}$, $\hat{m}_{i,j}$ and $p_{i,i}$ are obtained as in (9). Then, using `TNQR`, we can obtain the QR decomposition of A such that

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix},$$

where $Q \in \mathbb{R}^{(l+1)\times(l+1)}$ is an orthogonal matrix and $R \in \mathbb{R}^{(n+1)\times(n+1)}$ is an upper triangular matrix with positive diagonal entries. Following Section 1.3.1 in [2], the solution of the least squares problem is obtained from

$$\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = Q^T p, \quad Rc = d_1, \quad r = Q\begin{pmatrix} 0 \\ d_2 \end{pmatrix}, \tag{13}$$

where $d_1 \in \mathbb{R}^{n+1}$, $d_2 \in \mathbb{R}^{l-n}$ and $r = f - Ac$. The matrices Q and R have an special structure described in [13]. In particular, R is nonsingular and TP. In order to obtain the solution of the upper triangular system $Rc = d_1$, we have used the routine `TNSolve` of [16], which uses the bidiagonal decomposition of the upper triangular TP matrix R.

§5. Numerical experiments

Now let us illustrate the accuracy of the method explained in the previous section for the computation of the solution of the least squares minimization problem with fg-Bernstein bases. For different choices of f and g, we have considered fg-Bernstein bases of order n defined on $[a,b]$ and computed with Matlab two approximations of the vector $c = (c_1,\dots,c_{n+1})$ such that the function

$$p(t) = \sum_{j=1}^{n+1} c_j \binom{n}{j-1} f^{j-1}(t) g^{n-j+1}(t), \quad t \in [a,b],$$

minimizes $\sum_{k=1}^{100}(p_k - p(t_k))^2$, where $p_1,\dots,p_{100}$ are given integer values and $t_1,\dots,t_{100}$, $l > n$, are equidistant parameters in (a,b). One approximation has been obtained using the procedure explained in the previous section and the other approximation has been obtained using

n+1	TNQR	$A \setminus p$	TNQR	$A \setminus \tilde{p}$
15	4.89151×10^{-15}	8.09049×10^{-13}	8.18912×10^{-15}	4.57057×10^{-13}
20	2.97354×10^{-15}	1.95465×10^{-12}	2.68153×10^{-15}	4.60592×10^{-12}
25	4.20615×10^{-15}	9.55201×10^{-10}	3.864845×10^{-15}	9.49291×10^{-10}
30	8.16195×10^{-16}	2.56043×10^{-8}	9.15474×10^{-16}	4.55599×10^{-8}

Table 1: Relative errors with $f(t) = (1 + t)/2$, $g(t) = (1 - t)/2$, $t \in [-1, 1]$.

n+1	TNQR	$A \setminus p$	TNQR	$A \setminus \tilde{p}$
15	8.4759×10^{-16}	1.31186×10^{-12}	1.80073×10^{-15}	4.36905×10^{-14}
20	1.74157×10^{-15}	3.8791×10^{-13}	1.77785×10^{-15}	1.39799×10^{-13}
25	7.41971×10^{-15}	4.14554×10^{-10}	1.92262×10^{-14}	2.00229×10^{-10}
30	2.36573×10^{-15}	1.67828×10^{-9}	1.10435×10^{-14}	1.18769×10^{-8}

Table 2: Relative errors with $f(t) = t^2$, $g(t) = 1 - t^2$, $t \in [0, 1]$.

the Matlab command \. We have also computed the solution of these least squares problems using the Mathematica command LeastSquares with a precision of 100 digits and considered this solution c as the exact solution of the problem. Let us recall that in general we cannot guarantee HRA. However the numerical experiments show great accuracy in all the considered cases.

We have computed the relative error of every approximation $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c}_{n+1})$ of the solution c of the least squares problems by means of the formula

$$e = \frac{\|c - \tilde{c}\|_2}{\|c\|_2}.$$

We have considered $p_k := k \times (-1)^k$, $k = 1, \ldots, 100$ and also $\tilde{p}_k := k$, $k = 1, \ldots, 50$, $\tilde{p}_k := -k$, $k = 51, \ldots, 100$. The obtained errors are included in Table 1 (for the choice $f(t) = (1+t)/2$, $g(t) = (1-t)/2$, $t \in [-1, 1]$), in Table 2 (for the choice $f(t) = t^2$, $g(t) = 1-t^2$, $t \in [0, 1]$), in Table 3 (for the choice $f(t) = (1 - \cos t)/2$, $g(t) = (1 + \cos t)/2$, $t \in [0, \pi]$), in Table 4 (for the choice $f(t) = \sin((1 + t)/2)$, $g(t) = \sin((1 - t)/2)$, $t \in [-1, 1]$) and, finally, in Table 5 (for the choice $f(t) = \sinh((1+t)/2)$, $g(t) = \sinh((1-t)/2)$, $t \in [-1, 1]$). The computed results confirm the accuracy of the proposed method that, clearly, keeps the accuracy when the dimension of the problem increases.

In conclusion, we have presented a method for solving least squares problems with collocation matrices of fg-Bernstein bases that can be performed, in some cases, with HRA. We think that the proposed method exploits the structural properties of totally positive matrices and this could explain the great accuracy, even though HRA cannot be guaranteed, providing results much more accurate than those obtained by Matlab using the standard method for the resolution of least squares problems.

Acknowledgements

Partially supported by PGC2018-096321-B-I00 Spanish Research Grant and by Gobierno de Aragón E41_17R and Feder 2014-2020 “Construyendo Europa desde Aragón”.

n+1	TNQR	$A \setminus p$	TNQR	$A \setminus \tilde{p}$
15	2.95136×10^{-14}	1.99703×10^{-12}	1.2823×10^{-14}	4.93654×10^{-13}
20	1.81561×10^{-14}	6.64461×10^{-11}	1.7008×10^{-15}	1.64829×10^{-12}
25	4.61364×10^{-14}	1.72592×10^{-9}	1.15137×10^{-14}	7.41444×10^{-10}
30	7.88557×10^{-14}	4.34384×10^{-8}	4.88251×10^{-15}	6.41024×10^{-9}

Table 3: Relative errors with $f(t) = (1 - \cos t)/2$, $g(t) = (1 + \cos t)/2$, $t \in [0, \pi]$.

n+1	TNQR	$A \setminus p$	TNQR	$A \setminus \tilde{p}$
15	1.23029×10^{-15}	2.63437×10^{-12}	1.05317×10^{-14}	3.25159×10^{-12}
20	4.79405×10^{-15}	4.12448×10^{-11}	1.34693×10^{-15}	3.39137×10^{-11}
25	6.14711×10^{-16}	1.05147×10^{-9}	1.57822×10^{-14}	3.70748×10^{-10}
30	6.47177×10^{-15}	6.98518×10^{-8}	9.14979×10^{-15}	1.6299×10^{-7}

Table 4: Relative errors with $f(t) = \sin((1 + t)/2)$, $g(t) = \sin((1 - t)/2)$, $t \in [-1, 1]$.

References

[1] Ando, T. Totally positive matrices. *Linear algebra and its applications 90* (1987), 165–219.

[2] Björck, A. Numerical methods for least squares problems.

[3] Carnicer, J. M., and Peña, J. M. Shape preserving representations and optimality of the bernstein basis. *Advances in Computational Mathematics 1*, 2 (1993), 173–196.

[4] Carnicer, J. M., and Peña, J. M. Totally positive bases for shape preserving curve design and optimality of b-splines. *Computer Aided Geometric Design 11*, 6 (1994), 633–654.

[5] Delgado, J., Peña, G., and Peña, J. M. Accurate and fast computations with positive extended schoenmakers-coffey matrices. *Numerical Linear Algebra with Applications 23*, 6 (2016), 1023–1031.

[6] Delgado, J., and Peña, J. M. Accurate computations with collocation matrices of rational bases. *Applied Mathematics and Computation 219*, 9 (2013), 4354–4364.

[7] Delgado, J., and Peña, J. M. Fast and accurate algorithms for jacobi-stirling matrices. *Applied Mathematics and Computation 236* (2014), 253–259.

[8] Delgado, J., and Peña, J. M. Accurate computations with collocation matrices of q-bernstein polynomials. *SIAM Journal on Matrix Analysis and Applications 36*, 2 (2015), 880–893.

[9] Delgado, J., and Peña, J. M. Accurate computations with lupaş matrices. *Numerical Linear Algebra with Applications 303* (2017), 171–177.

[10] Demmel, J., Dumitriu, I., Holtz, O., and Koev, P. Accurate and efficient expression evaluation and linear algebra. *Acta Numerica 17* (2008), 87–145.

[11] Demmel, J., and Koev, P. The accurate and efficient solution of a totally positive generalized vandermonde linear system. *SIAM Journal on Matrix Analysis and Applications 27*, 1 (2005), 142–152.

n+1	TNQR	$A \setminus p$	TNQR	$A \setminus \tilde{p}$
15	4.66821×10^{-15}	5.93865×10^{-13}	3.33749×10^{-15}	2.32175×10^{-14}
20	1.8994×10^{-15}	1.84095×10^{-11}	2.55009×10^{-15}	7.65487×10^{-12}
25	2.29248×10^{-15}	3.38356×10^{-10}	4.82278×10^{-14}	6.31661×10^{-10}
30	5.00391×10^{-15}	4.98638×10^{-9}	1.89331×10^{-14}	3.22028×10^{-9}

Table 5: Relative errors with $f(t) = \sinh((1+t)/2)$, $g(t) = \sinh((1-t)/2)$, $t \in [-1, 1]$.

[12] Gasca, M., and Peña, J. M. Total positivity and neville elimination. *Linear algebra and its applications 165* (1992), 25–44.

[13] Gasca, M., and Peña, J. M. Total positivity, *qr* factorization, and neville elimination. *SIAM Journal on Matrix Analysis and Applications 14*, 4 (1993), 1132–1140.

[14] Gasca, M., and Peña, J. M. A matricial description of neville elimination with applications to total positivity. *Linear algebra and its applications 202* (1994), 33–53.

[15] Gasca, M., and Peña, J. M. On factorizations of totally positive matrices. In *M. Gasca C.A. Micchelli (Eds.), Total positivity and Its applications*. Kluver Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 109–130.

[16] Koev, P. Accurate computations with totally nonnegative matrices. *SIAM Journal on Matrix Analysis and Applications 29*, 3 (2007), 731–751.

[17] Koev, P. http://www.math.sjsu.edu/ koev/software/tntool.html.

[18] Mainar, E., and Peña, J. M. Corner cutting algorithms associated with optimal shape preserving representations. *Computer Aided Geometric Design 16*, 9 (1999), 883–906.

[19] Mainar, E., and Peña, J. M. Accurate computations with collocation matrices of a general class of bases. *Numerical Linear Algebra with Applications 25* (2018).

[20] Mainar, E., Peña, J. M., and Sánchez-Reyes, J. Shape preserving alternatives to the rational bézier model. *Computer aided geometric design 18*, 1 (2001), 37–60.

[21] Marco, A., and Martínez, J. J. Polynomial least squares fitting in the bernstein basis. *Linear Algebra and its Applications 433*, 7 (2010), 1254–1264.

[22] Marco, A., and Martínez, J. J. Accurate computations with totally positive bernstein-vandermonde matrices. *Electronic Journal of Linear Algebra 26*, 1 (2013), 24.

[23] Marco, A., and Martínez, J. J. Bidiagonal decomposition of rectangular totally positive said-ball-vandermonde matrices: Error analysis, perturbation theory and applications. *Linear Algebra and its Applications 495* (2016), 90–107.

[24] Peña, J. M. Shape preserving representations for trigonometric polynomial curves. *Computer Aided Geometric Design 14*, 1 (1997), 5–11.

[25] Sánchez-Reyes, J. Harmonic rational bézier curves, p-bézier curves and trigonometric polynomials. *Computer Aided Geometric Design 15*, 9 (1998), 909–923.

E. Mainar, J. M. Peña and B. Rubio
Departamento de Matemática Aplicada/IUMA
Universidad de Zaragoza, Spain
esmemain@unizar.es and jmpena@unizar.es and brubio@unizar.es

Monografías Matemáticas *García de Galdeano* **42**, 193–202 (2019)

On the Laplacian flow and coflow of G_2-structures

Víctor Manero, Antonio Otal, and Raquel Villacampa

Abstract. We review some recent results on the study of the Laplacian flow and coflow of G_2-structures.

Keywords: G_2-structures, Laplacian flow, Laplacian coflow, Lie groups, warped products.
AMS classification: 53C23, 22E25, 53C10.

§1. Introduction

In the 50's Berger [3] obtained the list of possible holonomy groups of simply connected, irreducible and non-symmetric Riemannian manifolds. In that list for the particular case of 7-dimensional manifolds appeared the exceptional holonomy Lie group G_2. A first tool in order to describe manifolds with holonomy G_2 is the concept of G_2-structure introduced by Bonan in [4]. A G_2-structure on a 7-dimensional manifold M can be characterized by the existence of a certain globally defined 3-form σ which is called the fundamental 3-form. The presence of such a structure on a manifold defines a metric g_σ on it, a volume form, and hence a Hodge star operator, namely $*$. Fernández and Gray in [12] gave a characterization for a manifold endowed with a G_2-structure to have holonomy restricted to the group G_2.

Theorem 1. *[12]. Let M be a manifold endowed with the G_2-structure σ. Denote by ∇^σ the Levi-Civita connection of the metric induced by the G_2-structure. Then, the following conditions are equivalent:*

- $Hol(\nabla^\sigma) \subseteq G_2$.
- $\nabla^\sigma \sigma = 0$.
- $d\sigma = d * \sigma = 0$.

The problem of obtaining manifolds with holonomy group G_2 was not a straightforward task and until the 80's the first examples were not described. In particular the first local example is due to Bryant [5], and later in a joint work with Salamon [6] obtained the first complete examples. These examples are obtained by considering 7-dimensional manifolds endowed with SO(3) or SO(4)-structures and a splitting of type 3+4. On those manifolds can be described a G_2-structure σ such that $d\sigma = 0$ and $d * \sigma = 0$. Concerning compact examples with holonomy G_2 the first ones were described by Joyce in [20] using the Kümmer construction for K3 surfaces. Later, Kovalev [22] and more recently Corti, Haskins, Nordstrom and Pacini have obtained new compact examples of manifolds with holonomy G_2 with the twisted connected sum construction and an extension of that technique respectively.

The torsion of a G_2-structure can be identified with the covariant derivative of the fundamental form σ and, as it is described in [12], it can be decomposed into four G_2 irreducible components, namely X_1, X_2, X_3 and X_4. Thus, a G_2-structure is said to be of type

$\mathcal{P}, \mathcal{X}_i, \mathcal{X}_i \oplus \mathcal{X}_j, \mathcal{X}_i \oplus \mathcal{X}_j \oplus \mathcal{X}_k$ or $\mathcal{X}$ if the covariant derivative $\nabla^\sigma \sigma$ lies in $\{0\}, X_i, X_i \oplus X_j, X_i \oplus X_j \oplus X_k$ or $X = X_1 \oplus X_2 \oplus X_3 \oplus X_4$, respectively. Hence, there exist 16 different classes of G_2-structures.

Another technique that allows to obtain examples of manifolds with holonomy in the group G_2 is via the study of flows of G_2-structures. These flows consist on one-parameter families of G_2-structures with certain initial conditions and such that satisfy an appropriated evolution equation. If this evolution equation is chosen appropriately, a solution for that flow is such that the initial value for the G_2-structures, which can have torsion, evolves to a G_2-structure without torsion. In this note we summarize some known results concerning the study of flows of G_2-structures, concretely we focus our attention on the Laplacian flow and the Laplacian coflow of a G_2-structure.

§2. Preliminars

We start explainning the basics about SU(3) and G_2-structures which are helpful for a brief introduction to the topic.

2.1. G_2-structures

A G_2-structure on a 7-dimensional manifold M consists of a reduction of the structure group of its frame bundle to the Lie group G_2. The existence of such structure on a manifold M can also be characterized by the presence of a global non-degenerate 3-form σ which can be locally written as

$$\sigma = e^{127} + e^{347} + e^{567} + e^{135} - e^{146} - e^{236} - e^{245}, \tag{1}$$

where $\{e^1, \dots, e^7\}$ is a local basis of 1-forms on M which we call the adapted basis. As usual in the related literature the notation $e^{i_1 \dots i_k}$ stands for the wedge product $e^{i_1} \wedge \cdots \wedge e^{i_k}$.

A manifold M endowed with a G_2-structure σ is called a G_2 manifold and the corresponding structure defines also a volume form vol_7 and a Riemannian metric g_σ satisfying

$$g_\sigma(X, Y) vol_7 = \frac{1}{6} \iota_X \sigma \wedge \iota_Y \sigma \wedge \sigma,$$

for every X, Y vector fields on M.

In order to describe the different classes of G_2-structures we consider first the G_2 type decomposition of the space of forms (see [5] for details). Let (M, σ) be a G_2 manifold, consider the action of the group G_2 on the space of differential p-forms on the manifold M, namely $\Omega^p(M)$. This action is irreducible on $\Omega^1(M)$ and $\Omega^6(M)$, but it is reducible for $\Omega^p(M)$ with $2 \leq p \leq 5$. The G_2 irreducible decompositions for $p = 2$ and 3 are

$$\Omega^2(M) = \Omega^2_7(M) \oplus \Omega^2_{14}(M),$$

where those irreducible spaces can characterized by

$$\Omega^2_7(M) = \{*_7(\alpha \wedge *_7 \sigma) \mid \alpha \in \Omega^1(M)\},$$
$$\Omega^2_{14}(M) = \{\beta \in \Omega^2(M) \mid \beta \wedge \sigma = - *_7 \beta\} = \{\beta \in \Omega^2(M) \mid \beta \wedge *_7 \sigma = 0\},$$

Class	Torsion forms	Condition	Structure
$\mathcal{P}$	$\tau_0 = \tau_1 = \tau_2 = \tau_3 = 0$	$d\sigma = d *_7 \sigma = 0$	Parallel
$\mathcal{X}_2$	$\tau_0 = \tau_1 = \tau_3 = 0$	$d\sigma = 0$	Closed
$\mathcal{X}_4$	$\tau_0 = \tau_2 = \tau_3 = 0$	$d\sigma = 3\tau_1 \wedge \sigma$, $d *_7 \sigma = 4\tau_1 \wedge *_7\sigma$	Locally Conformal Parallel
$\mathcal{X}_1 \oplus \mathcal{X}_3$	$\tau_1 = \tau_2 = 0$	$d *_7 \sigma = 0$	Coclosed
$\mathcal{X}_2 \oplus \mathcal{X}_4$	$\tau_0 = \tau_3 = 0$	$d\sigma = 3\tau_1 \wedge \sigma$	Locally Conformal Closed

Table 1: Principal classes of G_2-structures

and

$$\Omega^3(M) = \Omega^3_1(M) \oplus \Omega^3_7(M) \oplus \Omega^3_{27}(M),$$

with

$$\begin{aligned}
\Omega^3_1(M) &= \{f\sigma \mid f \in C^\infty(M)\},\\
\Omega^3_7(M) &= \{*_7(\alpha \wedge \sigma) \mid \alpha \in \Omega^1(M)\},\\
\Omega^3_{27}(M) &= \{\gamma \in \Omega^3(M) \mid \gamma \wedge \sigma = 0,\ \gamma \wedge *_7\sigma = 0\},
\end{aligned}$$

where $\Omega^p_k(M)$ denotes a G_2 irreducible space of p-forms of dimension k at every point. Note that the description on the other degrees are obtained via the isomorphism described by the Hodge star operator, i.e. $*_7\, \Omega^p_k(M) \cong \Omega^{7-p}_k(M)$.

The G_2 type decomposition of forms on M allows to express the exterior derivative of σ and $*_7\sigma$ as follows

$$\begin{aligned}
d\sigma &= \tau_0 *_7\sigma + 3\,\tau_1 \wedge \sigma + *_7\,\tau_3,\\
d *_7\sigma &= 4\,\tau_1 \wedge *_7\sigma + \tau_2 \wedge \sigma,
\end{aligned} \tag{2}$$

where $\tau_0 \in C^\infty(M)$, $\tau_1 \in \Omega^1(M)$, $\tau_2 \in \Omega^2_{14}(M)$ and $\tau_3 \in \Omega^3_{27}(M)$ are called the torsion forms of the G_2-structure.

Notice that all the information of the torsion of a G_2-structure is encoded on the covariant derivative of the fundamental form σ but also on the exterior derivatives of σ and $*\sigma$. Thus the different classes of G_2-structures can be described in terms of their behavior or equivalently, in view of (2), by the torsion forms $\tau_0, \tau_1, \tau_2, \tau_3$. In Table 1 some Fernández-Gray classes of G_2-structures are given.

The presence of certain G_2-structures on a manifold give information concerning its geometrical properties. Manifolds endowed with a parallel G_2-structure have holonomy contained in G_2, manifolds with a closed G_2-structure have non-positive scalar curvature. However, the scalar curvature of a manifold endowed with a coclosed G_2-structure has no sign restrictions. Locally Conformal Parallel and Locally Conformal Closed G_2-structures are (locally) Parallel and Closed G_2-structures which can be described by a conformal change of the original G_2-structure.

2.2. SU(3)-structures

An SU(3)-structure on a 6-dimensional manifold N consists of a triple (g, J, Ψ) such that g is a Riemannian metric, J is an almost complex structure compatible with the metric, and Ψ is

Class	Condition	Structure
$\{0\}$	$d\omega = d\psi_+ = d\psi_- = 0$	Calabi-Yau
$\mathcal{W}_1^-$	$d\omega = 3\psi_+, d\psi_- = -2\omega^2$	Nearly Kähler
$\mathcal{W}_2^-$	$d\omega = d\psi_+ = 0$	Symplectic half-flat
$\mathcal{W}_1^- \oplus \mathcal{W}_2^- \oplus \mathcal{W}_3$	$d\omega^2 = d\psi_+ = 0$	Half-flat

Table 2: Principal classes of SU(3)-structures

a complex volume form satisfying

$$\frac{3}{4} i \Psi \wedge \overline{\Psi} = \omega^3,$$

where ω is the fundamental form associated to the almost Hermitian structure (g, J). Note that an SU(3)-structure on a 6-dimensional manifold N can be described by the pair (ω, ψ_+), where ψ_+ is the real part of the complex volume form Ψ. Indeed, for the imaginary part ψ_- of the form Ψ one has that $\psi_- = J\psi_+$, so ψ_- is determined by ψ_+ and the almost complex structure J (see [18]). We will denote by g_{ω,ψ_+} the Riemannian metric induced by the SU(3)-structure.

Note that SU(3) and G_2-structures are closely related, in particular the presence of an SU(3)-structure (ω, ψ_+), on a 6-dimensional manifold N induces a G_2-structure on the 7-dimensional manifold $N \times L$ with $L = \mathbb{R}$ or S^1 which can be defined by

$$\sigma = \omega \wedge ds + \psi_+,$$

being s the coordinate on L.

As it is described in [9] the torsion of an SU(3)-structure, namely T, is identified with the covariant derivatives of ω and J and lies in a space of the form

$$T \in \mathcal{W}_1^{\pm} \oplus \mathcal{W}_2^{\pm} \oplus \mathcal{W}_3 \oplus \mathcal{W}_4 \oplus \mathcal{W}_5,$$

where $\mathcal{W}_i$ are the irreducible components under the action of the group SU(3). Analogously than for the G_2 case, this torsion can also be given in terms of the derivatives of the forms ω, ψ_+ and ψ_-. Equivalently the torsion forms of an SU(3)-structure can be defined (see [2] for details), but we will not care about this description on this note.

There exist many different classes of SU(3)-structures but the most relevant in the construction of G_2-structures are given in Table 2.

Calabi-Yau manifolds have holonomy in the group SU(3). Concerning nearly Kähler SU(3)-structures, not many examples of manifolds endowed with such structure are known, see [8] for homogeneous examples or in [16] can be found complete inhomogeneous examples on S^6 and $S^3 \times S^3$. Other well-known SU(3)-structures are the half-flat ones. These structures were first considered in [19] (see also [9]) and can be evolved to a parallel G_2-structure. Symplectic half-flat structures have been considered for several authors (see, for example, [10] and [13]) in order to obtain closed G_2-structures.

§3. Laplacian flow and coflow

The first author considering flows of G_2-structures was Bryant in [5]. The objective of considering flows of G_2-structures was to obtain examples of G_2-structures without torsion as the result of certain evolution of other G_2-structures with torsion. Thus, Bryant considered the so-called Laplacian flow of a G_2-structure σ_0 which is given by

$$\begin{cases} \frac{d}{dt}\sigma(t) = \Delta_t\sigma(t), \\ \sigma(0) = \sigma_0, \\ d\sigma(t) = 0, \end{cases} \tag{3}$$

where Δ_t denotes the corresponding Hodge Laplacian operator. On compact manifolds short time existence and uniqueness of solution for the Laplacian flow of a closed G_2-structure has been proved by Bryant and Xu in [7]. Xu and Ye in [29] proved long time existence and convergence of solution of the Laplacian flow starting near a torsion-free G_2-structure. In the last years Lotay and Wei in the series of papers [25, 26, 27] have obtained important results concerning long time existence and convergence of solution of the Laplacian flow.

On the other hand, in [21] Karigiannis, McKay and Tsui introduced the Laplacian coflow. This latter flow can be considered as the analogue to the Laplacian flow in which the fundamental 3-form is claimed to be coclosed instead of closed. Thus, this flow is given by the equations

$$\begin{cases} \frac{d}{dt}\psi(t) = -\Delta_t\psi(t), \\ \psi(0) = \psi_0, \\ d\psi(t) = 0, \end{cases}$$

with $\psi(t) = *_t\sigma(t)$ and $*_t$ denoting the Hodge star operator. As far as the authors know, short time existence and uniqueness of solution for this latter flow is not known. In [17] Grigorian introduced a modified version of this flow called modified Laplacian coflow for which he proved short time existence and uniqueness of solution.

3.1. Solutions of the Laplacian flow and coflow on Lie groups

The first examples of long time existence of solution for the Laplacian flow of closed G_2-structures were described in [11]. Concretely those examples are nilpotent Lie groups endowed with a one parameter family of left-invariant closed G_2-structures.

Theorem 2. *[11]. Consider the simply connected Lie group with Lie algebra given by the structure equations*

$$de^5 = e^1 \wedge e^2, \quad de^6 = e^1 \wedge e^3, \text{ and } de^i = 0 \text{ for all } i = 1, 2, 3, 4, 7.$$

The family of closed G_2 *forms* $\sigma(t)$ *on* N *given by*

$$\sigma(t) = e^{147} + e^{267} + e^{357} + f(t)^3 e^{123} + e^{156} + e^{245} - e^{346}, \qquad t \in \left(-\frac{3}{10}, +\infty\right),$$

where $f(t)$ is the function

$$f(t) = \Big(\frac{10}{3}t + 1\Big)^{\frac{1}{5}}.$$

is the solution of the Laplacian flow (3) *with initial value*

$$\sigma_0 = e^{147} + e^{267} + e^{357} + e^{123} + e^{156} + e^{245} - e^{346}.$$

Moreover, the underlying metrics $g(t)$ of this solution converge smoothly, up to pull-back by time-dependent diffeomorphisms, to a flat metric, uniformly on compact sets, as t goes to infinity.

More examples of long time solutions can also be found in [11] or in [23, 24]. Analogously in [1] have been given explicit long time solutions for the Laplacian coflow and the modified Laplacian coflow. These examples consist of one-parameter families of left-invariant coclosed G_2-structures on the 7-dimensional Heisenberg Lie group H_7 which is given by the matrices of the form

$$a = \begin{pmatrix} 1 & x_1 & x_3 & x_5 & x_7 \\ & 1 & & & x_2 \\ & & 1 & & x_4 \\ & & & 1 & x_6 \\ & & & & 1 \end{pmatrix}$$

with $x_i \in \mathbb{R}$ for all $i = 1, \dots 7$. Then a global system of coordinates x_i for H_7 is defined by $x_i(a) = x_i$. A standard calculation shows that a basis for the left invariant 1-forms on H_7 can be described by

$$e^1 = dx_1, \quad e^2 = dx_2, \quad e^3 = dx_3, \quad e^4 = dx_4,$$
$$e^5 = dx_5, \quad e^6 = dx_6, \quad \text{and} \quad e^7 = dx_7 - x_1 dx_2 - x_3 dx_4 - x_5 dx_6.$$

Thus, the corresponding Lie algebra, namely $\mathfrak{h}_7$ is given by the structure equations

$$de^7 = -e^1 \wedge e^2 - e^3 \wedge e^4 - e^5 \wedge e^6, \text{ and } de^i = 0 \text{ for all } i = 1, \dots, 6.$$

Theorem 3. *[1]. Consider H_7 the 7-dimensional Heisenberg Lie group. Then, the solution of the Laplacian coflow on H_7 with the initial coclosed G_2 form,*

$$\sigma_0 = e^{127} + e^{347} + e^{567} + e^{135} - e^{146} - e^{236} - e^{245},$$

is given by

$$\sigma(t) = \frac{1}{f(t)}(e^{127} + e^{347} + e^{567}) + f(t)^3(e^{135} - e^{146} - e^{236} - e^{245}), \quad t \in \Big(-\infty, \frac{3}{5}\Big)$$

where $f(t)$ is the positive function

$$f(t) = \Big(1 - \frac{5}{3}t\Big)^{\frac{1}{10}}.$$

Recently the study of the Laplacian flow and coflow of G_2-structures on Lie groups has been extended to different classes of G_2-structures like Locally Conformal Parallel G_2-structures (LCP for short) or Locally Conformal Closed ones (LCC for short). In particular, in [28] the authors consider the Laplacian flow, resp. coflow, of a LCP G_2-structure which can be defined as:

$$\begin{cases} \dfrac{d}{dt}\sigma(t) = \Delta_t\sigma(t), \\ \sigma(0) = \sigma_0, \\ d\sigma(t) = 3\,\tau(t) \wedge \sigma(t), \\ d *_t \sigma(t) = 4\,\tau(t) \wedge *_t\sigma(t). \end{cases} \qquad \begin{cases} \dfrac{d}{dt}\psi(t) = -\Delta_t\psi(t), \\ \psi(0) = \psi_0, \\ d\psi(t) = 4\,\tau(t) \wedge \psi(t), \\ d *_t \psi(t) = 3\,\tau(t) \wedge *_t\psi(t), \end{cases}$$

obtaining the following results:

Theorem 4. *[28]. Every 7-dimensional rank-one solvable extension of a nilpotent Lie group with a Locally Conformal Parallel* G_2 *form,* σ_0*, admits a long time solution* $\sigma(t)$ *to the Laplacian flow, preserving the LCP condition along the flow, such that* $\sigma(0) = \sigma_0$*.*

Theorem 5. *[28]. Every 7-dimensional rank-one solvable extension of a nilpotent Lie group with a Locally Conformal Parallel* G_2 *form admits a long time LCP solution to the Laplacian coflow.*

On the other hand the Laplacian flow of LCC G_2-structures can be described by

$$\begin{cases} \frac{d}{dt}\sigma(t) = \Delta_t\,\sigma(t), \\ d\,\sigma(t) = 3\tau(t) \wedge \sigma(t), \\ \sigma(0) = \sigma_0. \end{cases} \tag{4}$$

For this latter flow explicit examples of long time solutions are given in [14].

Theorem 6. *[14]. Consider the simply connected, solvable Lie group whose Lie algebra has structure equations*

$$de^1 = \frac{1}{2}e^1 \wedge e^7, \qquad de^2 = \frac{1}{2}e^2 \wedge e^7, \qquad de^3 = \frac{1}{2}e^3 \wedge e^7, \qquad de^4 = \frac{1}{2}e^4 \wedge e^7,$$
$$de^5 = e^1 \wedge e^4 + e^2 \wedge e^3 + e^5 \wedge e^7, \quad de^6 = e^1 \wedge e^3 - e^2 \wedge e^4 + e^6 \wedge e^7, \text{ and } de^7 = 0.$$

The family of locally conformal closed G_2*-structures* $\sigma(t)$ *given by*

$$\sigma(t) = (1-4t)^{3/4}\,e^{127} + (1-4t)^{3/4}\,e^{347} + e^{567} + e^{135} - e^{146} - e^{236} - e^{245}, \text{ where } t \in \left(-\infty, \tfrac{1}{4}\right)$$

is the solution for the Laplacian flow (4) *of the* G_2 *form*

$$\sigma_0 = e^{127} + e^{347} + e^{567} + e^{135} - e^{146} - e^{236} - e^{245}.$$

The Lee 1-form $\theta(t)$ *of* $\sigma(t)$ *is* $\theta(t) = -e^7$*. Moreover, the underlying metrics* $g(t)$ *of this solution converge smoothly, up to pull-back by time-dependent diffeomorphisms, to a flat metric, uniformly on compact sets, as* t *goes to* $-\infty$*, and they blow-up as* t *goes to* $\frac{1}{4}$*.*

3.2. Solutions of the Laplacian flow and coflow on warped products

Solutions of the Laplacian flow and coflow have also been obtained using warped products. The warped product of two Riemannian manifolds (F, g_F) and (B, g_B) is denoted by $B \times_f F$ and consists on the product manifold $B \times F$ endowed with the metric $g = \pi_1^*(g_B) + f^2\pi_2^*(g_F)$ with f a non-vanishing real differentiable function on B and π_1, π_2 the projections of $B \times F$ onto B and F, respectively.

As it is described in [15] if we consider $(\omega, \psi_\pm)$ an SU(3)-structure over a 6-dimensional manifold M^6 the 3-form

$$\sigma = f\omega \wedge ds + \psi_+$$

defines a G_2-structure on $M^7 = M^6 \times L$ with $L = \mathbb{R}$ or S^1 where f is a non-vanishing function on L and s the coordinate in L. This G_2-structure is called warped G_2-structure since the induced metric, namely g_σ, is exactly $g_{\omega,\psi_+} + f^2 ds^2$. Considering warped G_2-structures Fino and Raffero in [15] obtained sufficient conditions on the SU(3)-structure and the warping function f that guarantee the existence of solution for the Laplacian flow of a closed G_2-structure.

Concerning the Laplacian coflow of a coclosed G_2-structure Karigiannis, MacKay and Tsui in [21] showed that using warped products solutions for this flow could be obtained from 6-dimensional manifolds endowed with Nearly Kähler or Calabi Yau structures.

Let us finish by noticing that the Nearly Kähler or Calabi Yau conditions are very restrictive and thus not many examples of these classes are known. On the contrary with the approach of Fino and Raffero in [15] solutions for the Laplacian flow of a closed G_2-structure can be obtained from less restrictive conditions on the SU(3)-structure (concretely symplectic half-flat condition). Thus the following question naturally arises:

Question: Is it possible to obtain solutions for the Laplacian coflow as warped products of 6-dimensional manifolds endowed with less restrictive SU(3)-structures, like half-flat ones?

Acknowledgements

The three authors have been partially supported by the project MTM2017-85649-P (AEI/Feder, UE) and Gobierno Aragón/Fondo Social Europeo–Grupo Consolidado E22-17R Algebra y Geometría. The third author would also like to thank to the Fields Institute for its support during her stay in Toronto.

References

[1] Bagaglini, L. Fernández, M., and Fino, A. Laplacian coflow on the 7-dimensional heisenberg group. *arXiv:1704.00295v1 [math.DG]*.

[2] Bedulli, L., and Vezzoni, L. The Ricci tensor of $SU(3)$-manifolds. *J. Geom. Phys. 57* (2007), 1125–1146.

[3] Berger, M. Sur les groupes d'holonomie homogène des variétés à connexion affine et des variétés riemanniennes. *Bull. Soc. Math. France 83* (1955), 279–330.

[4] Bonan, E. Sur les variétés riemanniennes à groupe d'holonomie G_2 ou Spin(7). *C. R. Acad. Sci. Paris 262* (1966), 127–129.

[5] Bryant, R. Some remarks on G_2 structrures. *Proceedings of Gökova Geometry-Topology Conference 2005* (2006), 75–109.

[6] Bryant, R., and Salamon, S. On the construction of some complete metrics with exceptional holonomy. *Duke Math. J. 58* (2007), 829–850.

[7] Bryant, R., and Xu, F. Laplacian flow for closed G_2-structures: short time behavior. *arXiv:1101.2004 [math.DG]*.

[8] Butruille, J. B. Classification des variétés approximativement kähleriennes homogènes. *Ann. Glob. Anal. Geom. 27* (2005), 201–225.

[9] Conti, D., and Salamon, S. Generalized Killing spinors in dimension 5. *Trans. Amer. Math. Soc. 359* (2007), 5319–5343.

[10] de Andrés, L., Fernández, M., Fino, A., and Ugarte, L. Contact 5-manifolds with su(2) structure. *Q. J. Math. 60* (2009), 429–459.

[11] Fernández, M. Fino, A., and Manero, V. Laplacian flow of closed G_2-structures inducing nilsolitons. *J. Geom. Anal. 26* (2016), 1808–1837.

[12] Fernández, M., and Gray, A. Riemannian manifolds with structure group G_2. *Ann. Mat. Pura Appl. 132* (1982), 19–45.

[13] Fernández, M., Manero, V., Otal, A., and Ugarte, L. Symplectic half-flat solvmanifolds. *Ann. Global Anal. Geom. 43* (2013), 367–383.

[14] Fernández, M. Manero, V., and Sánchez, J. The Laplacian flow of locally conformal calibrated G_2-structures. *Axioms 8* (2019), 1–15.

[15] Fino, A., and Raffero, A. Closed warped G_2-structures evolving under the laplacian flow. *arXiv:1708.00222v1 [math.DG]. To appear in Ann. Sc. Norm. Super. Pisa*.

[16] Foscolo, L. Haskins, M. New G_2-holonomy cones and exotic nearly kähler structures on S^6 and $S^3 \times S^3$. *Ann. of Math. 185* (2017), 59–130.

[17] Grigorian, S. Short-time behavior of a modified laplacian coflow of G_2-structures. *Adv. Math. 248* (2013), 378–415.

[18] Hitchin, N. J. The geometry of three-forms in six dimensions. *J. Differ. Geom. 55* (2000), 547–576.

[19] Hitchin, N. J. Stable forms and special metrics. *Global Differential Geometry: The Mathematical Legacy of Alfred Gray (Bilbao, 2000), Contemp. Math. Amer. Math. Soc., Providence, RI, 288* (2001), 70–89.

[20] Joyce, D. Compact Riemannian 7-manifolds with holonomy G_2. i, ii,. *J. Differential Geom. 43* (1996), 291–328, 329–375.

[21] Karigiannis, S. McKay, B., and Tsui, M. P. Soliton solutions for the laplacian coflow of some G_2-structures with symmetry. *Diff. Geom. Appl. 30* (2012), 318–333.

[22] Kovalev, A. Twisted connected sums and special Riemannian holonomy. *J. Reine Angew. Math. 565* (2003), 125–160.

[23] Lauret, J. Laplacian flow of homogeneous G_2-structures and its solitons. *Proc. London Math. Soc. 114* (2017), 1–34.

[24] Lauret, J. Laplacian solitons: questions and homogeneous examples. *Diff. Geom. Appl. 54* (2017), 345–360.

[25] Lotay, J. D., and Wei, Y. Laplacian flow for closed G_2-structures: real analyticity. *arXiv:1601.04258. To appear in Comm. Anal. Geom.*.

[26] Lotay, J. D., and Wei, Y. Stability of torsion free G_2 structures along the laplacian flow. *arXiv:1504.07771. To appear in J. Diff. Geom.*.

[27] Lotay, J. D., and Wei, Y. Laplacian flow for closed G_2-structures: Shi-type estimates, uniqueness and compactness. *Geom. Funct. Anal 27* (2017), 165–233.

[28] Manero, V. Otal, A., and Villacampa, R. Solutions of the Laplacian flow and coflow of a locally conformal parallel G_2-structure. *arXiv:1711.08644v1 [math.DG]*.

[29] Xu, F., and Ye, R. Existence, convergence and limit map of the laplacian flow. *arXiv:0912.0074 [math.DG]*.

V. Manero
Departamento de Matemáticas - I.U.M.A.,
Universidad de Zaragoza, Facultad de Ciencias Humanas y de la Educación,
22003 Huesca, Spain
vmanero@unizar.es

A. Otal and R. Villacampa
Centro Universitario de la Defensa - I.U.M.A.,
Academia General Militar, Crta. de Huesca s/n,
50090 Zaragoza, Spain
aotal@unizar.es and raquelvg@unizar.es

Monografías Matemáticas *García de Galdeano* **42**, 203–217 (2019)

On some generalizations of B-splines

Peter Massopust

Abstract. In this article, we consider some generalizations of polynomial and exponential B-splines. Firstly, the extension from integral to complex orders is reviewed and presented. The second generalization involves the construction of uncountable families of self-referential or fractal functions from polynomial and exponential B-splines of integral and complex orders. As the support of the latter B-splines is the set $[0, \infty)$, the known fractal interpolation techniques are extended in order to include this setting.

Keywords: B-splines, cardinal splines, exponential splines, self-referential function, fractal interpolation, fractal function.

AMS classification: 26A33, 28A80, 41A05, 46F05, 65D07.

§1. Introduction

Schoenberg's polynomial B-splines [25] are a powerful tool in approximation theory because of their favorable analytic and computational properties. Unfortunately, polynomial B-splines also have some disadvantages. Amongst them, we list:

- Polynomial B-splines have only integer smoothness which is linked to the integer order n. However, for approximation-theoretic purposes, it is useful to fill in the gaps in the smoothness spectrum C^n, $n \in \mathbb{N}$. There are many functions that are elements of, for instance, Hölder spaces $C^{n,\alpha}$, $0 \leq \alpha < 1$.
- Polynomial B-splines do not contain phase information. The importance of approximation functions to be able to provide phase information is exemplified by the so-called Oppenheim-Lim Experiment [23]. In their paper, Oppenheim & Lim showed that the Fourier reconstruction of an image using only the modulus of the complex-valued Fourier coefficients results in less informative content than a reconstruction from the phase of the Fourier coefficients (and setting the modulus equal to 1). The reconstruction from phase showed singularities such as corners and edges quite clearly but they were hard to see in the reconstruction from the modulus.

  In addition, there are sometimes requirements for a single-band frequency analysis. For some applications, e.g., for phase retrieval tasks, complex-valued analysis bases are needed since real-valued bases can only provide a symmetric spectrum.
- Polynomial splines are ill-suited for describing functions or data that exhibit sudden growth or decay because of their oscillatory behavior near the points where such an increase or decrease occurs [28].

The first two items in the above list can be resolved by extending the order of B-splines from integral n to complex z with $\operatorname{Re} z > 1$. The thus obtained so-called complex B-splines [7] generate a two-parameter family of functions with a continuous smoothness spectrum and built-in phase information.

The third issue can be rectified by introducing exponential splines and B-splines. These splines are employed to model phenomena that follow differential systems of the form $\dot{x} = Ax$, where A is a constant matrix. For such equations the solutions are linear combinations of functions of the type e^{ax} and $x^n e^{ax}$, $a \in \mathbb{R}$. Like polynomial B-splines, exponential B-splines can be defined as finite convolution products of exponential functions. See [1, 5, 19, 24, 27, 30, 31] for an incomplete list of references for exponential splines. The extension of exponential B-splines to complex order [14] adds the option of applying them for the retrieval of phase information.

Neither the original nor extended polynomial and exponential B-splines are appropriate approximants when functions exhibit complex intrinsic characteristics such as self-referential or fractal behavior. In these cases, one needs to resort to fractal interpolation and approximation techniques to describe them. The extension of polynomial B-splines to an uncountable family of self-referential or fractal functions indexed by a finite tuple of real numbers $\alpha_i \in (-1, 1)$ was presented in, i.e., [12, 13, 22]. Here we consider the case of exponential B-splines of integral order and also the fractal generalization of polynomial and exponential B-splines of complex orders. The latter requires extending fractal interpolation techniques to unbounded domains.

The structure of this article is as follows. For the sake of presentation and completeness, we briefly introduce polynomial and exponential B-splines and their complex order extensions in Sections 2, respectively, 3. A brief introduction to self-referential functions is provided in Section 4.1 and in the final Section 4 uncountably many families of self-referential polynomial and exponential B-splines of complex orders are constructed.

§2. Polynomial B-Splines

In this section, we briefly review polynomial splines and their basis functions, polynomial B-splines. The interested reader may consult the large literature on splines for more details and further results.

To this end, let $\mathcal{X} = \{a = x_0 < x_1 < \cdots < x_k < x_{k+1} = b\}$ be a set of points, called knots, supported on the real line $\mathbb{R}$.

Definition 1. A *spline of order* n on $[a, b]$ with knot set $\mathcal{X}$ is a function $s : [a, b] \to \mathbb{R}$ such that

(i) On each subinterval $[x_{i-1}, x_i)$, s is a polynomial of order at most n (degree at most $n - 1$);

(ii) $s \in C^{n-2}[a, b]$.

s is called a *cardinal spline* if the knot set is a contiguous subset of $\mathbb{Z}$.

The set $\mathcal{S}_{\mathcal{X},n}$ of all spline functions s of order n over a knot set $\mathcal{X}$ forms an $\mathbb{R}$-vector space of dimension $n + k$. A convenient and powerful basis of $\mathcal{S}_{\mathcal{X},n}$ is given

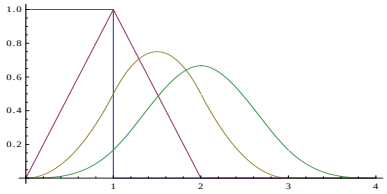


Figure 1: Some graphs of polynomial B-splines: $n = 1, 2, 3, 4$.

by Schoenberg's cardinal polynomial B-Splines [25]. They are recursively defined as follows. Denote by χ the characteristic function on $[0, 1]$ and set

$$B_1(x) := \chi(x),$$
$$B_n(x) := (B_{n-1} * B_1)(x) = \int_0^1 B_{n-1}(x-t)dt, \qquad 2 \le n \in \mathbb{N}, \tag{2.1}$$

where $*$ denotes the convolution between functions. An immediate consequence of this definition is that $\operatorname{supp} B_n = [0, n]$ and that $B_n \in C^{n-2}$, $n \in \mathbb{N}$, with C^{-1} denoting the family of piecewise continuous functions. Some graphs of these cardinal polynomial B-splines are shown in Figure 1.

Taking the Fourier transform of (2.1) yields the Fourier representation of B_n, which is sometimes used to define the B-splines.

$$\widehat{B}_n(\omega) := \mathcal{F}(B_n)(\omega) := \int_{\mathbb{R}} B_n(x)e^{-i\omega x}dx = \left(\frac{1-e^{-i\omega}}{i\omega}\right)^n. \tag{2.2}$$

It can be shown, either using (2.1) or (2.2) that the n-order B-spline has an explicit representation in the form

$$B_n(x) = \frac{1}{\Gamma(n)} \sum_{k=0}^{\infty} (-1)^k \binom{n}{k} (x-k)_+^{n-1}, \tag{2.3}$$

where $x_+ := \max\{0, x\}$.

The collection $\{B_n : n \in \mathbb{N}\}$ is thus a discrete family of functions with increasing smoothness and support. Both the support and the smoothness are tied to the integral order n.

The next result justifies the term B-spline with B standing for basis. For a proof, see for instance [6].

Proposition 1. *Every cardinal spline function $s : [a, b] \to$ of order n has a unique representation in terms of a finite shifted sequence of cardinal B-splines of order n:*

$$s(x) = \sum_{j=-n+1}^{k} c_j B_n(x-j),$$

where $c_j \in \mathbb{R}$.

Hence, investigating properties of splines reduces to those of B-splines.

Terminology. As we are dealing exclusively with cardinal splines and B-splines in the remainder of this paper, we will drop the adjective "cardinal."

2.1. Some properties of polynomial B-splines

The polynomial B-splines enjoy among others the following properties.

(i) *Recursion Relation*:

$$\forall n \in \mathbb{N}\ \forall x \in \mathbb{R} : B_n(x) = \frac{x}{n-1} B_{n-1}(x) + \frac{n-x}{n-1} B_{n-1}(x-1)$$

(ii) *Convolution Relation*:

$$\forall m, n \in \mathbb{N} : \quad B_m * B_n = B_{m+n}$$

(iii) *Convergence to Gaussians*: As $n \to \infty$, B_n converges in L^p-norm, $2 \leq p \leq \infty$, to a modulated Gaussian.

(iv) The *Error of Approximation* for an $f \in C^n[a,b]$ on a uniform grid of mesh size h by polynomial B-splines of order n is $O(h^n)$.

The interested reader may consult the extensive literature on B-splines to learn about additional properties of this important family of functions in approximation theory.

2.2. Polynomial B-splines of complex orders

Both the first and second obstacle of polynomial B-splines mentioned in the introduction can be overcome by extending them to include complex orders (or complex degrees). This can be done in the Fourier domain as follows. (Cf. [7].)

Definition 2. Suppose $z \in \mathbb{C}$ with $\operatorname{Re} z > 1$. The B-spline of complex order z, for short complex B-spline, is given by $\widehat{B} : \mathbb{R} \to \mathbb{C}$,

$$\widehat{B}_z(\omega) := \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^z, \tag{2.4}$$

or more precisely,

$$\widehat{B}_z(\omega) = \underbrace{\widehat{B}_{\operatorname{Re} z}(\omega)}_{\text{continuous smoothness}} \underbrace{e^{i \operatorname{Im} z \ln \Omega(\omega)}}_{\text{phase}} \underbrace{e^{-\operatorname{Im} z \arg \Omega(\omega)}}_{\text{modulation}}, \tag{2.5}$$

where $\Omega(\omega) := \frac{1-e^{-i\omega}}{i\omega}$.

We remark that $\widehat{B}_z$ is well-defined as $\operatorname{graph} \Omega$ does not intersect the real axis.

The first factor in the product appearing in (2.5) is the Fourier transform of a so-called *fractional B-spline* [29]. Some graphs of such B-splines of real order are depicted in Figure 2. The second and third factors in (2.5) are a modulating and a damping factor. The presence of the imaginary part $\operatorname{Im} z$ causes the frequency components on

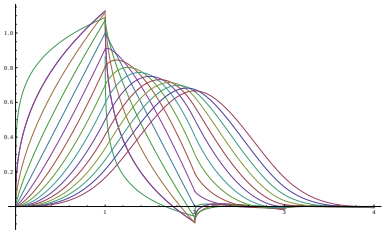


Figure 2: A family of B-splines of real order for $\alpha = \operatorname{Re} z = 0.6 + m \cdot 0.2$, $m = 1, \ldots, 17$.

the negative and positive real axis to be enhanced with different signs. This has the effect of shifting the frequency spectrum towards the negative or positive frequency side, depending on the sign of $\operatorname{Im} z$. The corresponding bases can be interpreted as approximate single-band filters [7].

The time domain representation of a complex B-spline was derived in [7] and is given in the next theorem.

Theorem 2 (Time domain representation)**.**

$$B_z(x) = \frac{1}{\Gamma(z)} \sum_{k=0}^{\infty} (-1)^k \binom{z}{k} (x-k)_+^{z-1}, \quad \operatorname{Re} z > 1. \tag{2.6}$$

Equality holds point-wise for all $x \in \mathbb{R}$ and in the $L^2(\mathbb{R})$–norm.

Complex B-splines enjoy among others the following properties.

1. $B_z \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, $\operatorname{Re} z > 1$.
2. $\int_{\mathbb{R}} B_z(x)dx = \widehat{B}_z(0) = 1$.
3. $B_z \in W^{p,2}(\mathbb{R})$ for $p < \operatorname{Re} z - \frac{1}{2}$.
4. $B_z(x) = O(|x|^{-m})$, for $m < \operatorname{Re} z + 1$ and $|x| \to \infty$.
5. B_z converges in L^p-norm, $2 \le p \le \infty$, to a modulated and shifted Gaussian as $\operatorname{Re} z \to \infty$.
6. $B_{\operatorname{Re} z}$ reproduces polynomials up to order $\lceil \operatorname{Re} z \rceil$.
7. For $\operatorname{Re} z > 1$, $B_{\operatorname{Re} z}$ is $(\operatorname{Re} z - 1)$-Hölder continuous.
8. $\{B_z(\cdot - k)\}_{k \in \mathbb{Z}}$ is a Riesz sequence in $L^2(\mathbb{R})$. This allows the construction of spline scaling functions and spline wavelets of complex order.

Some graphical examples of complex polynomial B-splines are shown in Figure 3.

In summary, complex B-splines are a continuous two-parameter family of functions which enjoy the properties:

(a) $\operatorname{Re} z > 1$ gives a continuous family of functions of increasing smoothness $\operatorname{Re} z$;

(b) $\operatorname{Im} z$ contains phase information and can be used to describe and resolve singularities in signals and images.

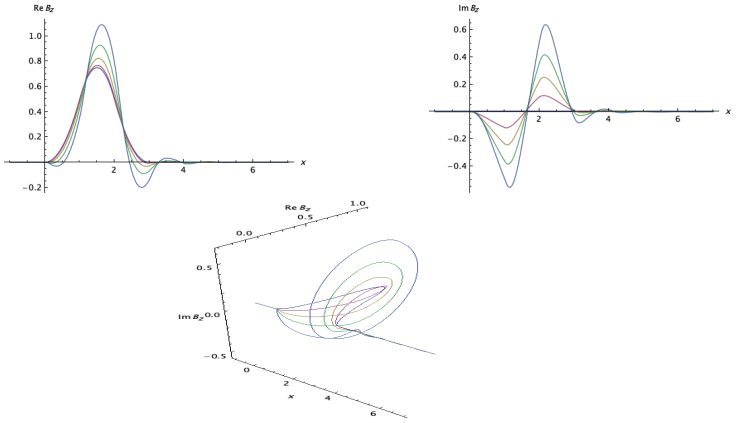


Figure 3: B_z for $z = (3 + \frac{k}{4}) + i$, $k = 0, 1, \ldots, 4$.

§3. Exponential B-Splines

Exponential B-splines can be used to interpolate or approximate data that exhibit sudden growth or decay and for which polynomial B-splines are not well-suited because of their oscillatory behavior near the points where the sudden growth or decay occurs [28]. The interested reader is referred to the following albeit incomplete list of references on exponential B-splines [5, 19, 27, 30, 31].

To define the class of exponential B-splines, let $N \in \mathbb{N}$ and let $\boldsymbol{a} := (a_1, \ldots, a_N)$, where $a_1, \ldots, a_N \in \mathbb{R}$ with $a_i \neq 0$ for at least one $i \in \mathbb{N}_N$.

Definition 3. An exponential B-spline $E_{N,\boldsymbol{a}} : \mathbb{R} \to \mathbb{R}$ of order N for the N-tuple $\boldsymbol{a}$ is a function of the form

$$E_N := E_{N,\boldsymbol{a}} := \mathop{*}_{k=1}^{N} e^{a_k(\cdot)}\chi.$$

To simplify notation, we set $\varepsilon^{a(\cdot)} := e^{a(\cdot)}\chi$. A closed formula for E_n was derived in [4]. Note that $\operatorname{supp} E_N = [0, N]$, $N \in \mathbb{N}$.

For any $a \in \mathbb{R}$, the Fourier transform of $\varepsilon^{-a(\cdot)}$ is given by

$$\mathcal{F}(\varepsilon^{-a(\cdot)})(\omega) = \frac{1 - e^{-a}e^{-i\omega}}{i\omega + a}.$$

and, therefore,

$$\mathcal{F}(E_n)(\omega) = \prod_{k=1}^{n} \frac{1 - e^{-a_k}e^{-i\omega}}{i\omega + a_k} \overset{a_k=a}{=} \left(\frac{1 - e^{-a}e^{-i\omega}}{i\omega + a}\right)^n. \tag{3.1}$$

3.1. Exponential B-splines of complex order

Let $z \in \mathbb{C}_{>1} := \{\zeta \in \mathbb{C} : \operatorname{Re}\zeta > 1\}$ and $a > 0$. Taking the left-hand-side of (3.1) as a starting point, we define an exponential B-spline of complex order z, for short complex

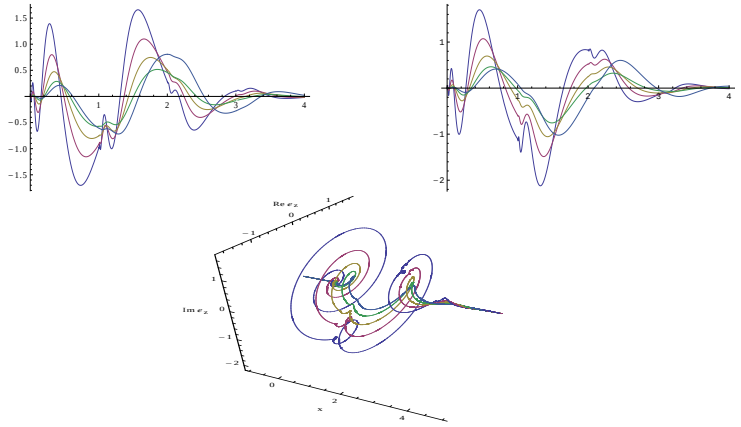


Figure 4: $E_{z,a}$ for $z = (3 + \frac{k}{4}) + 3i$, $k = 0, 1, \ldots, 4$ and $a = 1.7$. Top right: $\operatorname{Re} E_{z,a}$, top left: $\operatorname{Im} E_{z,a}$, bottom: Three-dimensional rendering of $E_{z,a}$.

exponential B-spline, by (see [14])

$$\begin{aligned}\widehat{E}_{z,a}(\omega) &:= \left(\frac{1 - e^{-(a+i\omega)}}{a + i\omega}\right)^z \\ &= \widehat{E}_{\operatorname{Re} z,a}(\omega)\, e^{i\,\Omega_a(\omega)\,\operatorname{Im} z}\, e^{-\arg \Omega_a(\omega)\,\operatorname{Im} z},\end{aligned} \tag{3.2}$$

where $\Omega_a(\omega) := \frac{1-e^{-(a+i\omega)}}{a+i\omega}$. An investigation of the function $\Omega_a : \mathbb{R} \to \mathbb{C}$ shows that $\widehat{E}_{z,a}$ is well-defined only if $a > 0$. (See [14].) The second and third terms in the product of (3.2) play the same role as they did in the case of complex polynomial B-splines.

Using properties of the exponential difference operator and the definition of $E_{z,a}$, the following time domain representation of $E_{z,a}$ was proved in [14].

Theorem 3. *Suppose $z \in \mathbb{C}_{>1}$ and $a > 0$. Then,*

$$E_{z,a}(x) = \frac{1}{\Gamma(z)} \sum_{k=0}^{\infty} (-1)^k \binom{z}{k} e^{-ka} e_+^{-a(x-k)}\, (x-k)_+^{z-1},$$

where $e_+^{(\cdot)} := \chi_{[0,\infty)}\, e^{(\cdot)}$. The sum converges both point-wise in $\mathbb{R}$ and in the L^2–sense.

Figure 4 depicts some graphs of exponential B-splines of complex order.

Remark 1. Complex polynomial and exponential B-splines of order $z \in \mathbb{C}_{>1}$ are two-parameter families of functions assigning to each point $x \in [0, \infty)$ both a real value and a single direction given by $\operatorname{Im} z$. For several applications however, such as geophysical data processing or multichannel data, more than one independent direction is required. For this purpose, the complex order is replaced by a quaternionic or more generally a hypercomplex order. We refer the interested reader to [8, 9, 17] for these extensions in the case of polynomial B-splines.

§4. Self-Referential Polynomial and Exponential B-Splines

In this section, we consider some fractal extensions of the classical as well as the complex polynomial and exponential B-splines.

4.1. Self-Referential Functions

First, we briefly review the concept of self-referential function. For more details and proofs we refer the interested reader to, for instance, [2, 3, 10, 13, 18, 20] or any other of the numerous publications in fractal interpolation theory.

In the following, the symbol $\mathbb{N}_N := \{1, \dots, N\}$ denotes the initial segment of length N of $\mathbb{N}$. Further, we assume that $N \geq 2$.

Let I be a nonempty interval in $\mathbb{R}$ and suppose that $\{L_n : I \to I : n \in \mathbb{N}_N\}$ is a family of bijections with the property that $\{L_n(I) : n \in \mathbb{N}_N\}$ forms a partition of I, i.e.,

$$I = \bigcup_{n=1}^{N} L_n(I), \quad \text{and} \quad L_n(I) \cap L_m(I) = \emptyset, \quad \forall n \neq m \in \mathbb{N}_N. \tag{4.1}$$

Remark 2. Condition (4.1) cannot be relaxed without adding compatibility conditions to guarantee the form (4.2) of the RB operator T. For more details, we refer the interested reader to [26].

Denote by $B(I) := B(I, \mathbb{R})$ the set

$$B(I) := \{f : I \to \mathbb{R} : f \text{ bounded}\}.$$

$(B(I), d)$ becomes a complete metric space when endowed with the metric

$$d(f, g) := \sup_{x \in I} |f(x) - g(x)|,$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}$.

Let $f, b \in B(I)$ be arbitrary. Consider the Read-Bajraktarevíc (RB) operator $T : B(I) \to B(I)$ defined on each subinterval $L_n(I)$ by

$$Tg = f + \alpha_n \cdot (g - b) \circ L_n^{-1}, \quad n \in \mathbb{N}_N, \tag{4.2}$$

with $\alpha_n \in \mathbb{R}$. Under the assumption that $\alpha := \max\{|\alpha_n| : n \in \mathbb{N}_N\} < 1$, it follows from the Banach fixed point theorem that T has a unique fixed point $f^* \in B(I)$. This fixed point satisfies the *self-referential equation*

$$f^* = f + \alpha_n \cdot (f^* - b) \circ L_n^{-1}, \quad \text{on} \quad L_n(I),\ n \in \mathbb{N}_N. \tag{4.3}$$

Any function in $B(I)$ which satisfies an equation of the form (4.3) is termed a *self-referential function of type* $B(I)$. The functions f and b are called *seed function*, respectively, *base function*.

Note that f^* can be iteratively obtained as the limit of the sequence $\{g_k\}$ defined by

$$g_k := Tg_{k-1} = f + \alpha_n \cdot (g_{k-1} - b) \circ L_n^{-1}, \quad \text{on } L_n(I), \quad k \in \mathbb{N}, \tag{4.4}$$

for an arbitrary $g_0 \in B(I)$.

Remark 3. The fixed point f^* of an RB operator has the property that $\operatorname{graph} f^*$ is made up of a finite number of copies of itself and is therefore, in general, a fractal set. For this reason, f^* is also called a *fractal function* [2, 10, 18].

Remark 4. Self-referential functions defined on function spaces other than $B(I)$ can be constructed as well. Examples include, among others, the smoothness spaces $C^r(I)$, the Lebesgue spaces $L^p(I)$, and the Besov spaces $B^s_{p,q}(I)$. (Cf., for instance, [3, 11, 15, 16, 18].) To ensure that the RB operator T maps a function space into itself, additional conditions at the points $\{L_n(\partial I)\}$, $n \in \mathbb{N}_N$, may have to be imposed.

Remark 5. For a given finite set of bijections $\{L_n\}$ or, equivalently, a given partition of I yielding a finite set of bijections, the fixed point f^* depends on the functions f and b as well as the vertical scaling factors $\{\alpha_n\}$. The interested reader may want to consult [21] in the former case.

Remark 6. For a varying N-tuple $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_N)$, the fixed point f^* actually defines an uncountable family $f^{\boldsymbol{\alpha}}$ of self-referential functions indexed by $\boldsymbol{\alpha} \in (-1, 1)^N$. Such sets of self-referential functions were termed $\boldsymbol{\alpha}$-fractal functions and considered as the image of an operator $\mathcal{F}^{\boldsymbol{\alpha}}$, $f \mapsto f^{\boldsymbol{\alpha}}$. (Cf., i.e., [20].)

4.2. Polynomial and exponential splines of integral order

For this purpose, let B_N be the cardinal polynomial B-spline of order $N \geq 2$ as in (2.3). Let $I := \operatorname{supp} B_N = [0, N]$ and define bijections $L_n : I \to I$, $n \in \mathbb{N}_N$, by

$$L_n(I) := \begin{cases} [n-1, n), & n \in \mathbb{N}_{N-1}; \\ [N-1, N], & n = N. \end{cases}$$

Now choose $f := B_N$ and $b \equiv 0$. The case where b is null is considered in [21]. Suppose $\alpha := \max\{|\alpha_n| : n \in \mathbb{N}_N\} < 1$. Then the RB operator T reads

$$Tg = B_N + \alpha_n \cdot g \circ L_n^{-1}, \quad \text{on } L_n(I), \quad n \in \mathbb{N}_N,$$

for any $g \in B(I)$. As $B_N \in C^{N-2}$, we additionally require $g \in C^{N-2}(I)$ and impose the join-up conditions

$$\forall\, m \in \mathbb{N}_{N-1} \,:\, (Tg)^{(\nu)}(m-) = (Tg)^{(\nu)}(m+), \quad \nu = 0, 1, \ldots, N-2. \tag{4.5}$$

Conditions (4.5) guarantee that $Tg \in C^{N-2}(I)$ and as $C^{N-2}(I)$ becomes a Banach space under the norm $\sum_{\nu=0}^{N-2} \|(\cdot)^{(\nu)}\|_\infty$, the unique fixed point $\mathfrak{B}_N$ of T is an element of $C^{N-2}(I)$ and a self-referential function:

$$\mathfrak{B}_N = B_N + \alpha_n \cdot \mathfrak{B}_N \circ L_n^{-1}, \quad \text{on } L_n(I), \quad n \in \mathbb{N}_N. \tag{4.6}$$

As the fixed point $\mathfrak{B}_N$ depends continuously on the set of parameters $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_N) \in (-1, 1)^N$, we also write $\mathfrak{B}_N(\boldsymbol{\alpha})$ should the need arise. Hence, (4.6) defines an uncountable family of functions parametrized by $\boldsymbol{\alpha}$. Clearly, $\boldsymbol{\alpha} = 0$ reproduces the seed function B_N. (See also [22].)

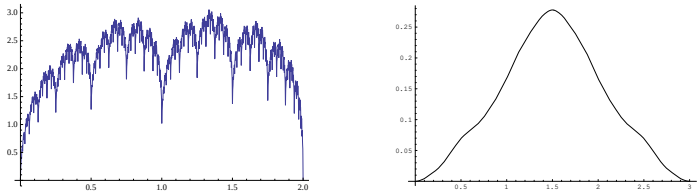


Figure 5: A linear (left) and a quadratic (right) fractal polynomial B-spline.

Figure 5 depicts two such fractal polynomial B-splines: the linear $\mathfrak{B}_2((\frac{3}{4},\frac{3}{4}))$ and the quadratic $\mathfrak{B}_3((\frac{1}{4},\frac{1}{4},\frac{1}{4}))$. Note that $\mathfrak{B}_3((\frac{1}{4},\frac{1}{4},\frac{1}{4}))$ is differentiable on $[0,3]$ and its graph is made up of three copies of itself.

In a similar fashion, we can take $I := [0,N]$, $f := E_{N,\boldsymbol{a}}$, and set again $b \equiv 0$ to generate an uncountable family of fractal analogues of the classical exponential B-splines $E_{N,\boldsymbol{a}}$. The RB operator then reads

$$Tg = E_{N,\boldsymbol{a}} + \alpha_n \cdot g \circ L_n^{-1}, \quad \text{on } L_n(I), \quad n \in \mathbb{N}_N,$$

for any $g \in C(I)$ (as the functions $E_{N,\boldsymbol{a}}$ are continuous on I). As above, we choose $\boldsymbol{\alpha} \in (-1,1)^N$ and impose the continuity conditions

$$Tg(m-) = Tg(m+), \quad m \in \mathbb{N}_{N-1}.$$

Under these conditions, T is well-defined and contractive from $C(I)$ into itself. Its unique fixed point $\mathfrak{E}_{N,\boldsymbol{a}} := \mathfrak{E}_{N,\boldsymbol{a}}(\boldsymbol{\alpha})$ satisfies the self-referential equation

$$\mathfrak{E}_{N,\boldsymbol{a}} = E_{N,\boldsymbol{a}} + \alpha_n \cdot \mathfrak{E}_{N,\boldsymbol{a}} \circ L_n^{-1}, \quad \text{on } L_n(I), \quad n \in \mathbb{N}_N. \tag{4.7}$$

In Figures 6 and 7, two fractal exponential B-splines are depicted.

4.3. Polynomial and exponential B-splines of complex order

In order to derive the fractal extensions of polynomial and exponential B-splines of complex order, we need to take into account the fact that the support of B_z and $E_{z,a}$ is the unbounded interval $I := [0,\infty)$ and extend the above construction to this setting.

To be specific, suppose that the bijections L_n, $n \in \mathbb{N}_N$, are such that $L_n(I)$, $n \in \mathbb{N}_{N-1}$, is bounded on $\mathbb{R}$ and $L_N(I)$ unbounded. As before, we require that Eqn. (4.1) holds. We note that this set-up is an important special case of a general approach investigated in [16].

To this end, we introduce the Banach space $(C_{0,0}(\mathbb{R}_0^+), \|\cdot\|_\infty)$ given by

$$C_{0,0}(\mathbb{R}_0^+) := C_{0,0}(\mathbb{R}_0^+,\mathbb{R}) := \left\{ f \in C(\mathbb{R}_0^+,\mathbb{R}) : f(0) = 0 \wedge \lim_{x\to\infty} f(x) = 0 \right\}.$$

As B_z and $E_{z,a}$ are continuous functions of the time variable x, vanish at $x = 0$, and satisfy $\lim_{x\to\infty} B_z(x) = 0 = \lim_{x\to\infty} E_{z,a}(x)$, we need to impose conditions on the RB operator T in Eqn. (4.2) to map $C_{0,0}(\mathbb{R}_0^+)$ into itself.

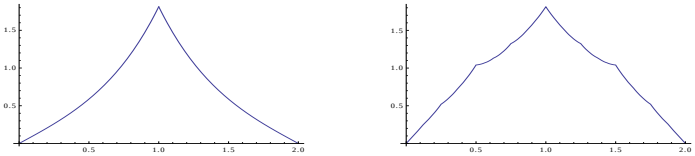


Figure 6: The graphs of $E_{2,(2,-2)}$ and $\mathfrak{E}_{2,(2,-2)}(\frac{1}{4},\frac{1}{4})$

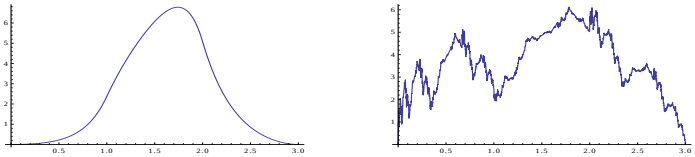


Figure 7: The graphs of $E_{3,(4,-3,1)}$ and $\mathfrak{E}_{3,(4,-3,1)}(\frac{3}{4},-\frac{1}{4},\frac{1}{2})$

These conditions read as follows. For $n \in \mathbb{N}_{N-1}$, denote

$$\begin{aligned} L_n(0) &=: x_{n-1}, \\ L_n(\infty) &=: x_n, \end{aligned} \tag{4.8}$$

and for $n := N$:

$$\begin{aligned} L_N(0) &=: x_{N-1}, \\ L_N(\infty) &= \infty. \end{aligned} \tag{4.9}$$

Here, we used the shorthand notation $f(\infty) := \lim_{x\to\infty} f(x)$ for a function f.

As a base function, we choose again $b \equiv 0$ on $[0,\infty)$ and require that, for $n \in \mathbb{N}_{N-1}$,

$$Tg(x_n-) = Tg(x_n+) \tag{4.10}$$

or, equivalently,

$$Tg(L_n(\infty)) = Tg(L_{n+1}(0)), \tag{4.11}$$

with the obvious modification for $n = N$.

Theorem 4. *Suppose bijections $L_n : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ are chosen such that $\{L_n(\mathbb{R}_0^+)\}_{n\in\mathbb{N}_N}$ forms a partition of $[0,\infty)$ subject to (4.8) and (4.9). Further suppose that $T : C_{0,0}(\mathbb{R}_0^+) \to C_{0,0}(\mathbb{R}_0^+)$ is given by*

$$Tg = f + \alpha_n \cdot g \circ L_n^{-1}, \tag{4.12}$$

and satisfies (4.10), where $f \in C_{0,0}(\mathbb{R}_0^+)$ is arbitrary and $\alpha := \max\{|\alpha_n| : n \in \mathbb{N}_N\} < 1$. Then T is well-defined and contractive on $(C_{0,0}(\mathbb{R}_0^+), \|\cdot\|_\infty)$ with Lipschitz constant α.

Proof. The conditions on the bijections $\{L_n\}$ and the join-up conditions (4.10) guarantee that T is well-defined and maps $C_{0,0}(\mathbb{R}_0^+)$ into itself. To establish that T is contractive on $(C_{0,0}(\mathbb{R}_0^+), \|\cdot\|_\infty)$ with Lipschitz constant α is straightforward. □

The unique fixed point $f^* \in C_{0,0}(\mathbb{R}_0^+)$ of T as defined in Eqn. (4.12) is called a *self-referential function of class $C_{0,0}(\mathbb{R}_0^+)$*.

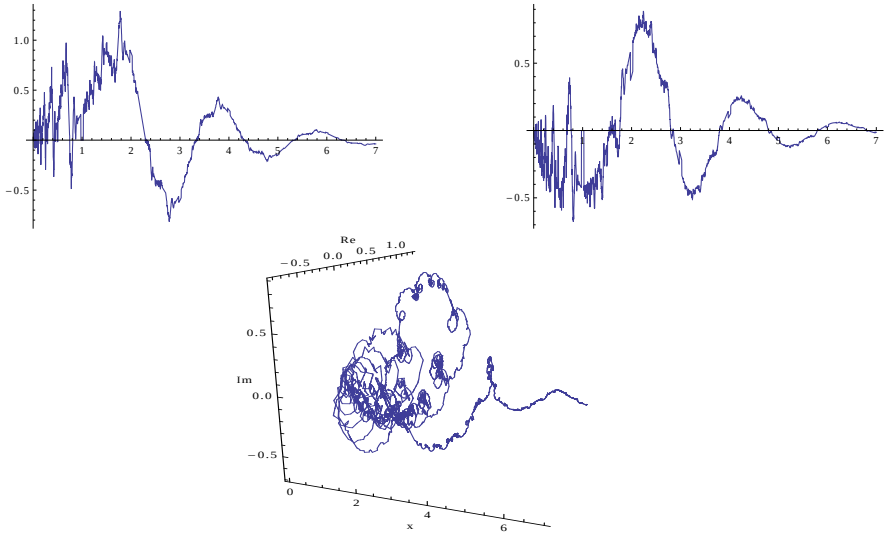


Figure 8: Top right: $\operatorname{Re}\mathfrak{B}_{\pi+i}(\frac{3}{4},-\frac{1}{2})$, top left: $\operatorname{Im}\mathfrak{B}_{\pi+i}(\frac{3}{4},-\frac{1}{2})$, bottom: Three-dimensional rendering of $\mathfrak{B}_{\pi+i}(\frac{3}{4},-\frac{1}{2})$.

Remark 7. Note that Theorem 4 also holds for the Banach spaces $(C_{0,0}(\mathbb{R}_0^+,\mathbb{C}),\|\cdot\|_\infty)$.

As noted above, for varying $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_N)$ subject to $\alpha:=\max\{|\alpha_n|:n\in\mathbb{N}_N\}<1$, f^* actually defines an uncountably infinite family $f^{\boldsymbol{\alpha}}$ of self-referential functions containing the seed function f.

As two prominent examples of how to obtain the fractal extension of functions in $C_{0,0}(\mathbb{R}_0^+)$, we consider $f=B_z$ and $f=E_{z,a}$. For this purpose and the sake of presentation, we choose $N:=2$ and define bijections $L_n:\mathbb{R}_0^+\to\mathbb{R}_0^+$ by

$$L_1(x):=2\,\pi^{-1}\arctan x \quad\text{and}\quad L_2(x):=x+1.$$

Then $[0,\infty)=L_1([0,\infty))\cup L_2([0,\infty))=[0,1)\cup[1,\infty)$.

Now select $f:=B_z$, respectively, $f=E_{z,a}$, choose $\alpha_n\in(-1,1)$, $n=1,2$, and define RB operators

$$T_1g:=B_z+\alpha_1\,g\circ\tan(\tfrac{\pi x}{2})\big|_{[0,1)}+\alpha_2\,g(x-1)\big|_{[1,\infty)}.$$

and

$$T_2g:=E_{z,a}+\alpha_1\,g\circ\tan(\tfrac{\pi x}{2})\big|_{[0,1)}+\alpha_2\,g(x-1)\big|_{[1,\infty)}.$$

By Theorem 4 and Remark 7, we obtain the fractal extensions of B_z and $E_{z,a}$ as the fixed points $\mathfrak{B}_z(\boldsymbol{\alpha})$, respectively, $\mathfrak{E}_{z,a}(\boldsymbol{\alpha})$ of the RB operators T_1 and T_2:

$$\mathfrak{B}_z=B_z+\alpha_1\,\mathfrak{B}_z\circ\tan(\tfrac{\pi x}{2})\big|_{[0,1)}+\alpha_2\,\mathfrak{B}_z(x-1)\big|_{[1,\infty)}.$$

and

$$\mathfrak{E}_{z,a}=E_z^a+\alpha_1\,\mathfrak{E}_{z,a}\circ\tan(\tfrac{\pi x}{2})\big|_{[0,1)}+\alpha_2\,\mathfrak{E}_{z,a}(x-1)\big|_{[1,\infty)}.$$

In Figures 8 and 9, the graphs of $\mathfrak{B}_{\pi+i}(\frac{3}{4},-\frac{1}{2})$ and $\mathfrak{E}_{\sqrt{2}+i,1}(\frac{3}{4},-\frac{1}{2})$ are displayed.

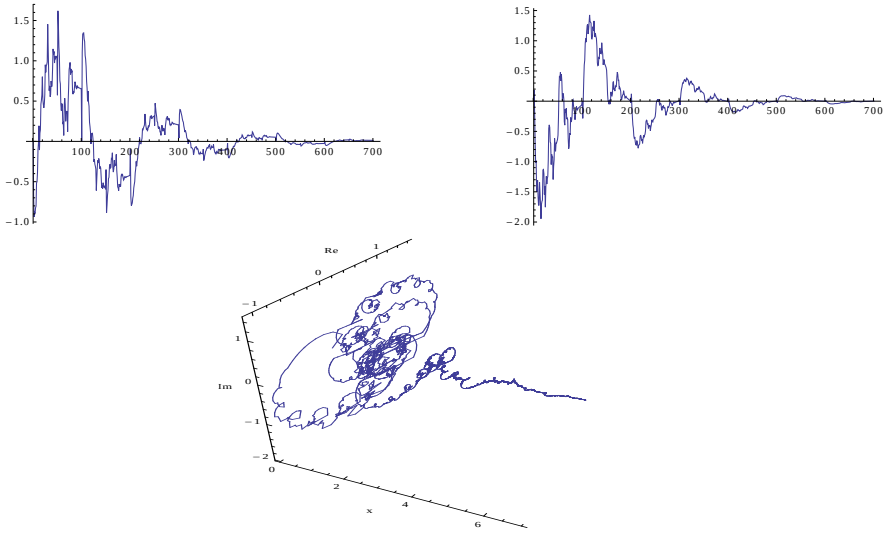


Figure 9: Top right: $\mathbf{Re}\,\mathfrak{E}_{\sqrt{2}+i,1}(\frac{3}{4},-\frac{1}{2})$, top left: $\mathbf{Im}\,\mathfrak{E}_{\sqrt{2}+i,1}(\frac{3}{4},-\frac{1}{2})$, bottom: Three-dimensional rendering of $\mathfrak{E}_{\sqrt{2}+i,1}(\frac{3}{4},-\frac{1}{2})$. The length unit on the x-axis for the graphs on top is $\frac{1}{100}$.

Remark 8. The families of self-referential functions supported on the interval $I = [0,\infty)$ not only depend on $\boldsymbol{\alpha}$ but also on the partition induced by the bijections L_n on I. Denoting the collection of all such partitions by $\Pi = \Pi_N$, the set of fixed points $f^{\boldsymbol{\alpha}}$ should more precisely be written as $f^{\boldsymbol{\alpha}}_{\Pi}$ and regarded as a function $(-1,1)^N \times \Pi \to f^*$.

References

[1] Ammar, G., Dayawansa, W., and Martin, C. Exponential interpolation theory: Theory and numerical algorithms. *Appl. Math. Comput. 41* (1991), 189–232.

[2] Barnsley, M. F. Fractal functions and interpolation. *Const. Approx. 2* (1986), 303–329.

[3] Barnsley, M. F., Hegland, M., and Massopust, P. Numerics and fractals. *Bull. Inst. Math. Acad. Sinica (N.S.) 9*, 3 (2014), 389–430.

[4] Christensen, O., and Massopust, P. Exponential B-splines and the partition of unity property. *Adv. Comput. Math. 37* (2012), 301–318.

[5] Dahmen, W., and Micchelli, C. A. On the theory and applications of exponential splines. In *Topics in Multivariate Approximation*, C. K. Chui, L. L. Schumaker, and F. I. Utreras, Eds. Academic Press, Boston, 1987.

[6] de Boor, C. *A Practical Guide to Splines.* No. 27 in Applied Mathematical Sciences. Springer Verlag, 2001.

[7] Forster, B., Unser, M., and Blu, T. Complex B-splines. *Appl. Comput. Harm. Anal. 20* (2006), 261–282.

[8] Hogan, J., and Massopust, P. Quaternionic B-Splines. *J. Approx. Th. 224* (2017), 43–65.

[9] Hogan, J., and Massopust, P. Quaternionic fundamental cardinal splines: Interpolation and sampling. *arxiv.org/abs/1804.06638* (2018), 1–24.

[10] Hutchinson, J. E. Fractals and self-similarity. *Indiana J. Math. 30*, 5 (1981), 713–747.

[11] Massopust, P. Fractal functions and their applications. *Chaos, Solitons, & Fractals 8*, 2 (1997), 171–190.

[12] Massopust, P. Fractal functions, splines, and Besov and Triebel-Lizorkin spaces. In *Fractals in Engineering: New trends and applications*, J. Lévy-Véhel and E. Lutton, Eds. Springer Verlag, London, 2005, pp. 21–32.

[13] Massopust, P. *Interpolation and Approximation with Splines and Fractals*. Oxford University Press, 2010.

[14] Massopust, P. Exponential splines of complex order. *Contemp. Math. 626* (2014), 87–105.

[15] Massopust, P. Local fractal functions and function spaces. In *Fractals, Wavelets, and their Applications*, C. B. et al., Ed., Springer Proceedings in Mathematics & Statistics. Springer Verlag, 2014, pp. 245–270.

[16] Massopust, P. Local fractal interpolation on unbounded domains. *Proc. Edinburgh Math. Soc. 61* (2018), 151–167.

[17] Massopust, P. Splines and fractional differential operators. *arxiv.org/abs/1901.11304* (2019), 1–17.

[18] Massopust, P. R. *Fractal Functions, Fractal Surfaces, and Wavelets*, 2nd ed. Academic Press, New York, 2016.

[19] McCartin, B. J. Theory of exponential splines. *J. Approx. Th. 66* (1991), 1–23.

[20] Navascués, M. A. Fractal polynomial interpolation. *Z. Anal. Anwendungen 24*, 2 (2005), 401–418.

[21] Navascués, M. A., and Massopust, P. Fractal convolution - a new operation between functions. *arxiv:1805.11316v1* (2018), 1–21.

[22] Navascués, M. A., and Sebastián, M. V. Fractal splines. *Monografías del Seminario Matemático García de Galdeano 33*, 161–168 (2006).

[23] Oppenheim, A., and Lim, J. The importance of phase in signals. *IEEE 69*, 5 (1981), 529–541.

[24] Sakai, M., and Usmani, R. A. On exponential B-splines. *J. Approx. Th. 47* (1986), 122–131.

[25] Schoenberg, I. J. Contributions to the problem of approximation of equidistant data by analytic functions. *Quart. Appl. Math. 4* (1946), 45–99, 112–141.

[26] Serpa, C., and Buescu, J. Constructive solutions for systems of iterative functional equations. *Const. Approx. 45*, 2 (2017), 273–299.

[27] Spaeth, H. Exponential spline interpolation. *Computing 4* (1969), 225–233.

[28] Stoer, J., and Bulirsch, R. *Introduction to Numerical Analysis*, 2nd ed. Springer Verlag, New York, 1993.

[29] Unser, M., and Blu, T. Fractional Splines and Wavelets. *SIAM Review 42*, 1 (2000), 43–67.

[30] Unser, M., and Blu, T. Cardinal exponential splines: Part I – theory and filtering algorithms. *IEEE Trans. Signal Processing 53*, 4 (2005), 1425–1438.

[31] Zoppou, C., Roberts, S., and Renka, R. J. Exponential spline interpolation in characteristic based scheme for solving the advective–diffusion equation. *Int. J. Numer. Meth. Fluids 33* (2000), 429–452.

P. Massopust
Center of Mathematics
Technical University of Munich
Garching b. München, Germany
massopust@ma.tum.de

Monografías Matemáticas *García de Galdeano* **42**, 219–228 (2019)

FRACTAL JACKSON APPROXIMATION ON THE TORUS

María Antonia Navascués, Sangita Jha, María Victoria Sebastián, Arya Kumar Bedabrata Chand

Abstract. In this article we generalize an approximation formula for three dimensional periodic data on a grid using fractal techniques which helps us to construct both smooth and non-smooth approximants depending on the choice of scale factors. We obtain bounds of the approximation error and showed the convergence with very weak conditions, when the sampling frequency is indefinitely increased. The density of the mappings involved in the space of two-dimensional periodic and continuous functions is proved using certain ranges of the scaling factors. A numerical example is presented to illustrate the proposed approximation methods.

Keywords: Fractals, Fractal Interpolation Functions, Fractal Surfaces, Two Dimensional Approximation.

AMS classification: 28A80, 42A10, 42A15.

§1. Introduction

Current major investigations in the theory of approximation concern smooth approximation. However it would be good to have mathematical structures to describe real life models which are non-smooth in nature. Such a structure is provided, for instance, by the theory of fractal functions (see for instance cf. [1], [2], [3], [8], [9]). Barnsley (cf. [1], [2]) first introduced the concept of fractal interpolation functions (FIFs) using the theory of iterated function system (IFS) (cf. [5]). FIFs form the basis of iterative constructive approximation theory. Barnsley and Harrington (cf. [3]) derived the calculus of FIF and showed that depending on the parameters of the IFS, one can construct smooth or non-smooth FIFs. Adapting the notion of FIF, Navascués (cf. [10]) constructed an entire family of fractal functions f^α, parameterized by an appropriate vector α, beginning from a given continuous function f on a compact interval I. This type of maps tend to bridge the gap between the smoothness of the classical mathematical objects and the pseudo-randomness of experimental data.

In the theory of classical trigonometric approximation, D. Jackson (cf. [6], [7]) described the degree of approximation of a continuous function by means of algebraic trigonometric polynomials. For the one dimensional case, he introduced an approximation formula (cf. [6]) for 2π periodic continuous functions as

$$\Sigma_m f(x) = H_m \sum_{i=1}^{2m} f(x_i) \left(\frac{\sin\left(\frac{m(x_i-x)}{2}\right)}{m \sin\left(\frac{x_i-x}{2}\right)} \right)^4 , \tag{1.1}$$

where

$$x_{i+1} - x_i = \frac{\pi}{m}, i = 1, 2, \ldots, 2m-1 \text{ and } H_m^{-1} = \sum_{i=1}^{2m} \left(\frac{\sin\left(\frac{m(x_i - x)}{2}\right)}{m \sin\left(\frac{x_i - x}{2}\right)} \right)^4 .$$

We generalize the previous formula (cf. [11]) using a positive exponent γ, and derive the convergence properties with very weak conditions on the original function. Recently, (cf. [14]), Navascués and Sebastián extended the approximation formula (1.1) for the two dimensional case. The formula proposed in [14] has an explicit representation in terms of the sample data on a two dimensional grid.

The approximation problem considered here is the representation of a prescribed periodic continuous and real-valued function of two variables using fractal techniques. In addition, we prove the density of the mappings involved in the space of two-dimensional periodic and continuous functions using certain ranges of the scaling factors. Numerical examples are given in the last section to illustrate the proposed process.

§2. Preliminaries

First we shall review the materials from the references (cf. [1], [7], [10], [13]) which will be used in the sequel.

2.1. Construction of fractal functions

Let us recall the construction of fractal interpolation functions in this section. Consider an interpolation data set $\{(x_i, y_i), i \in \mathbb{N}_N \cup \{0\}\}$, where $\mathbb{N}_N = \{1, 2, \ldots, N\}$. Let $\Delta := x_0 < x_1 < \cdots < x_N$ be a partition of the interval $I = [x_0, x_N]$. Let $L_i : I \to I_i = [x_{i-1}, x_i], i \in \mathbb{N}_N$ be contractive homeomorphisms such that

$$L_i(x_0) = x_{i-1},\ L_i(x_N) = x_i. \tag{2.1}$$

Let $K = I \times \mathbb{R}$ and N continuous mappings, $F_i : K \to \mathbb{R}$ be satisfying

$$F_i(x_0, y_0) = y_{i-1}, F_i(x_N, y_N) = y_i,\ |F_i(x, y) - F_i(x, y')| \leq |c_i||y - y'|, \tag{2.2}$$

where $(x, y), (x, y') \in K,\ \ c_i \in (-1, 1),\ \ i \in \mathbb{N}_N$. Now define functions $w_i : \mathbb{R}^2 \to \mathbb{R}^2$ as $w_i(x, y) = (L_i(x), F_i(x, y))\ \forall\ i \in \mathbb{N}_N$.

Theorem 1. *The Iterated Function System (IFS) $\mathcal{I} = \{K; w_i, i = 1, 2, \ldots, N\}$ admits a unique attractor G, which is the graph of a continuous function $f : I \to \mathbb{R}$ which obeys $f(x_i) = y_i$ for $i = 0, 1, 2, \ldots, N$.*

The previous function is called a Fractal Interpolation Function (FIF) corresponding to the IFS $\mathcal{I} = \{L_i(x), F_i(x, y)\}_{i=1}^N$, and it satisfies the following functional equation:

$$f(x) = F_i(L_i^{-1}(x), f \circ L_i^{-1}(x)), x \in I_i, i \in \mathbb{N}_N. \tag{2.3}$$

In this paper we choose $L_i(x) = a_i x + b_i$ satisfying (2.1) and $F_i(x, y) = \alpha_i y + q_i(x)$, where $q_i : I \to \mathbb{R}$ are continuous functions verifying (2.2). The vector $\alpha = (\alpha_1, \ldots, \alpha_N)$ is called a vertical scaling factor and it must satisfy the inequality $|\alpha|_\infty = \max\{|\alpha_i|; i = 1, 2, \ldots, N\} < 1$.

2.2. α-fractal function

Let $f : I \to \mathbb{R}$ be a continuous function. Consider $q_i(x) = f \circ L_i(x) - \alpha_i b(x)$, where b is defined from f through a linear map L ($Lf = b$) satisfying $b(x_0) = f(x_0), b(x_N) = f(x_N)$. The fixed point function associated with the above IFS is known as the α-fractal function f^α, and it enjoys the following equation:

$$f^\alpha(x) = f(x) + \alpha_i(f^\alpha - b)(L_i^{-1}(x)), x \in I_i, i \in \mathbb{N}_N. \tag{2.4}$$

The previous equation provides the inequality

$$\|f^\alpha - f\|_\infty \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty}\|f - b\|_\infty = \frac{|\alpha|_\infty}{1 - |\alpha|_\infty}\|f - Lf\|_\infty, \tag{2.5}$$

which bounds the uniform distance between f^α and f. Navascués (cf. [10]) proposed the linear and continuous operator $\mathcal{F}^\alpha$ defined by $\mathcal{F}^\alpha(f) = f^\alpha$.

§3. One dimensional fractal Jackson approximant

Let $C(T^1)$ denote the set of all continuous periodic function on $[-\pi, \pi]$. Let $\Delta_m : -\pi = x_0 < \cdots < x_{2m-1} < x_{2p} = \pi$ be such that $x_{i+1} = x_i + \frac{\pi}{m}$ for all $i = 0, 1, 2, \ldots, 2m-1$. Let us consider the continuous and periodic basis $\left\{P_{mi\gamma}(x) = \left|\frac{\sin\left(\frac{m(x_i - x)}{2}\right)}{m\sin\left(\frac{x_i - x}{2}\right)}\right|^\gamma ; i = 0, 1, 2, \ldots, 2m\right\}$. Let us define the set $\tau_m = \text{span}\{P_{mi\gamma}\}_{i=0}^{2m}$. Let us consider a Jackson type operator $\mathcal{T}_{m\gamma} : C(T^1) \mapsto \tau_m$ assigning a periodic approximant belonging to τ_m for every $g \in C(T^1)$ (with respect to the data $\{(x_i, g(x_i))\}_{i=0}^{2m}$), defined as

$$\mathcal{T}_{m\gamma}(g)(x) = H_{m\gamma}(x)\sum_{i=0}^{2m} g(x_i)P_{mi\gamma}(x),$$

where $(H_{m\gamma}(x))^{-1} = \sum_{i=0}^{2m}\left|\frac{\sin\left(\frac{m(x_i - x)}{2}\right)}{m\sin\left(\frac{x_i - x}{2}\right)}\right|^\gamma$. It is easy to see that

$$\|\mathcal{T}_{m\gamma}g\|_{C(T^1)} \leq \|g\|_{C(T^1)}.$$

In fact, the equality holds if we choose $g(x) = 1$. In the one dimensional case, the error of discrete Jackson approximation was studied in cf. [12]. According to this reference, for $g \in C(T^1)$, and $\gamma > 2$, the error of the approximation can be bounded as

$$\|\mathcal{T}_{m\gamma}(g) - g\|_{C(T^1)} \leq \left(\frac{\pi}{2}\right)^\gamma \omega_g\left(\frac{\pi}{4m}\right)(1 + 2^\gamma\zeta(\gamma - 1)), \tag{3.1}$$

where ζ is the Riemann zeta function. We define the α-fractal Jackson approximant of $g \in C(T^1)$ as

$$\mathcal{T}_{m\gamma}^\alpha(g)(x) = H_{m\gamma}(x)\mathcal{F}^\alpha\left(\sum_{i=0}^{2m} g(x_i)P_{mi\gamma}(x)\right) = H_{m\gamma}(x)\left(\sum_{i=0}^{2m} g(x_i)P_{mi\gamma}^\alpha(x)\right),$$

where $P^{\alpha}_{mi\gamma}(x)$ is the α-fractal function of $P_{mi\gamma}$ with respect to the partition Δ of $I = [-\pi, \pi]$ and a linear bounded operator L. Let us denote $\mathcal{P}_{m\gamma}(g)(x) = \sum_{i=0}^{2m} g(x_i)P_{mi\gamma}(x)$. Then

$$\|\mathcal{P}_{m\gamma}(g)\|_\infty \leq \|g\|_\infty \| \sum_{i=0}^{2m} P_{mi\gamma}(x)\|_\infty. \tag{3.2}$$

Thus $\|\mathcal{P}_{m\gamma}(g)\|_\infty \leq \|g\|_\infty \|H^{-1}_{m\gamma}\|_\infty$ which provides the inequality $\|\mathcal{P}_{m\gamma}\| \leq \|H^{-1}_{m\gamma}\|_\infty$, where $\|\mathcal{P}_{m\gamma}\|$ represents the norm of the operator with respect to the supremum norm $\|.\|_\infty$ in $C(T^1)$. Here $H^{-1}_{m\gamma}$ represents the inverse with respect to the product. For the operator $\mathcal{T}^{\alpha}_{m\gamma}$,

$$\begin{aligned}\|\mathcal{T}^{\alpha}_{m\gamma}(g)\|_\infty \leq & \|H_{m\gamma}\|_\infty \|\mathcal{F}^{\alpha}(\mathcal{P}_{m\gamma})\|_\infty \\ & \leq \|\mathcal{F}^{\alpha}\| \|H_{m\gamma}\|_\infty \|H^{-1}_{m\gamma}\|_\infty \|g\|_\infty \\ & = R_{m\gamma\alpha}\|g\|_\infty, \end{aligned} \tag{3.3}$$

where $R_{m\gamma\alpha} = \|\mathcal{F}^{\alpha}\| \|H_{m\gamma}\|_\infty \|H^{-1}_{m\gamma}\|_\infty$. Then $\|\mathcal{T}^{\alpha}_{m\gamma}\| \leq R_{m\gamma\alpha}$. Let us consider the error term $\mathcal{T}^{\alpha}_{m\gamma}(g) - g$:

$$\begin{aligned}\mathcal{T}^{\alpha}_{m\gamma}(g)(x) - g(x) = & H_{m\gamma} \sum_{i=0}^{2m} g(x_i)P^{\alpha}_{mi\gamma}(x) - H_{m\gamma} \sum_{i=0}^{2m} g(x_i)P_{mi\gamma}(x) + H_{m\gamma} \sum_{i=0}^{2m} g(x_i)P_{mi\gamma}(x) - g(x) \\ & = H_{m\gamma}\mathcal{P}^{\alpha}_{m\gamma}(g)(x) - H_{m\gamma}\mathcal{P}_{m\gamma}(g)(x) + H_{m\gamma}\mathcal{P}_{m\gamma}(g)(x) - g(x),\end{aligned}$$

where

$$\mathcal{P}^{\alpha}_{m\gamma}(g)(x) = \sum_{i=0}^{2m} g(x_i)P^{\alpha}_{mi\gamma}(x) = \mathcal{F}^{\alpha}(\mathcal{P}_{m,\gamma}(g))(x).$$

Thus, using the above computations we obtain

$$\|\mathcal{T}^{\alpha}_{m\gamma}(g) - g\|_\infty \leq \|H_{m\gamma}\|_\infty \|\mathcal{P}^{\alpha}_{m\gamma}(g) - \mathcal{P}_{m\gamma}(g)\|_\infty + \|\mathcal{T}_{m\gamma}(g) - g\|_\infty. \tag{3.4}$$

Using (2.5), the first term of the above inequality can be bounded as

$$\begin{aligned}\|\mathcal{P}^{\alpha}_{m\gamma}(g) - \mathcal{P}_{m\gamma}(g)\|_\infty \leq & \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|\mathcal{P}_{m\gamma}(g) - L\mathcal{P}_{m\gamma}(g)\|_\infty \\ & \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|I - L\| \|\mathcal{P}_{m\gamma}(g)\|_\infty \\ & \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|I - L\| \|H^{-1}_{m\gamma}\|_\infty \|g\|_\infty.\end{aligned} \tag{3.5}$$

Finally, using (3.1) and (3.5) in (3.4) we get

$$\|\mathcal{T}^{\alpha}_{m\gamma}(g) - g\|_{C(T^1)} \leq \|H_{m\gamma}\|_\infty \frac{|\alpha|_\infty \|I - L\|}{1 - |\alpha|_\infty} \|H^{-1}_{m\gamma}\|_\infty \|g\|_\infty + \left(\frac{\pi}{2}\right)^{\gamma} \omega_g\left(\frac{\pi}{4m}\right)(1 + 2^{\gamma}\zeta(\gamma - 1)).$$

(3.6)

§4. Fractal Jackson approximation on T^2

In this section, the approximation process described above is extended to data on a two dimensional torus. Let be given two partitions $\Delta_m^1 : -\pi = x_0 < x_1 < \cdots < x_{2m-1} < x_{2m} = \pi$ and $\Delta_n^2 : -\pi = y_0 < y_1 < \cdots < y_{2n-1} < y_{2n} = \pi$ of the circle. Let us consider the grid $\Delta = \Delta_m^1 \times \Delta_n^2$ of $T^2 = T^1 \times T^1$ and data $\{(x_i, y_j, z_{ij}) : i = 0, 1, 2, \ldots, 2m; j = 0, 1, 2, \ldots, 2n\}$ with 2π-periodicity condition in both variables. Let $\alpha \in (-1, 1)^{2m}$ and $\beta \in (-1, 1)^{2n}$ be scale vectors for Δ_m^1 and Δ_n^2 respectively. Let us define the operator using different exponents γ_1, γ_2 for both single functions as

$$\mathcal{J}_{mn\gamma_1\gamma_2}(f)(x, y) = K_{mn\gamma_1\gamma_2}(x, y) \sum_{i=0}^{2m} \sum_{j=0}^{2n} f(x_i, y_j) P_{mi\gamma_1}(x) Q_{nj\gamma_2}(y),$$

where $x_{i+1} - x_i = \frac{\pi}{m}; i = 0, 1, 2, \ldots, 2m-1$, $y_{j+1} - y_j = \frac{\pi}{n}; j = 0, 1, 2, \ldots, 2n-1$,

$$P_{mi\gamma_1}(x) = \left| \frac{\sin\left(\frac{m(x_i - x)}{2}\right)}{m \sin\left(\frac{x_i - x}{2}\right)} \right|^{\gamma_1},$$

$$Q_{nj\gamma_2}(y) = \left| \frac{\sin\left(\frac{n(y_j - y)}{2}\right)}{n \sin\left(\frac{y_j - y}{2}\right)} \right|^{\gamma_2},$$

and

$$K^{-1}_{mn\gamma_1\gamma_2}(x, y) = \sum_{i=0}^{2m} \sum_{j=0}^{2n} \left| \frac{\sin\left(\frac{m(x_i - x)}{2}\right)}{m \sin\left(\frac{x_i - x}{2}\right)} \right|^{\gamma_1} \left| \frac{\sin\left(\frac{n(y_j - y)}{2}\right)}{n \sin\left(\frac{y_j - y}{2}\right)} \right|^{\gamma_2}.$$

Lemma 2. *(cf. [11]) For all $k = 1, 2, \ldots, \gamma$ and $z \in \mathbb{R}$:*

$$\left| \frac{\sin(kz)}{k \sin(z)} \right|^{\gamma} \leq 1.$$

Definition 1. (cf. [4]) Let f be a continuous function defined on T^2. The modulus of continuity of f is defined as

$$\omega_f(\delta) := \sup_{\|x_1 - x_2\| \leq \delta} \{|f(x_1) - f(x_2)| : x_1, x_2 \in T^2\}.$$

We will use the following properties of the modulus of continuity:

1. $\omega_f(\delta_1 + \delta_2) \leq \omega_f(\delta_1) + \omega_f(\delta_2)$.
2. $\omega_f(\lambda\delta) \leq \lambda\omega_f(\delta)$ for $\lambda \in \mathbb{N}$.

Lemma 3. *For any $\gamma_1, \gamma_2 > 0$, the norm of $K_{mn\gamma_1\gamma_2}$ can be bounded as*

$$\|K_{mn\gamma_1\gamma_2}\|_\infty \leq \frac{1}{4} \left(\frac{\pi}{2}\right)^{2\gamma_{max}},$$

where $\gamma_{max} = \max\{\gamma_1, \gamma_2\}$.

Proof. From the definition of $K_{mn\gamma_1\gamma_2}$ we have

$$\begin{aligned} K^{-1}_{mn\gamma_1\gamma_2}(x,y) &= \sum_{i=0}^{2m}\sum_{j=0}^{2n}\left|\frac{\sin(\frac{m(x_i-x)}{2})}{m\sin(\frac{x_i-x}{2})}\right|^{\gamma_1}\left|\frac{\sin(\frac{n(y_j-y)}{2})}{n\sin(\frac{y_j-y}{2})}\right|^{\gamma_2} \\ &= H^{-1}_{m\gamma_1}(x)H^{-1}_{n\gamma_2}(y) \\ &\leq \frac{1}{2}\left(\frac{\pi}{2}\right)^{\gamma_1}\frac{1}{2}\left(\frac{\pi}{2}\right)^{\gamma_2} \\ &\leq \frac{1}{4}\left(\frac{\pi}{2}\right)^{2\gamma_{max}}, \end{aligned}$$

where $H^{-1}_{m\gamma}(x)$ is defined in Section 3 and considering that $H^{-1}_{m\gamma}(x) \geq 2\left(\frac{2}{\pi}\right)^{\gamma}$ for any $\gamma > 0$ (cf. [12]). □

Theorem 4. *Let $f \in C(T^2)$. Then for any $\gamma_1, \gamma_2 > 2$, the approximant $\mathcal{J}_{mn\gamma_1\gamma_2}(f)$ converges uniformly to f whenever m, n tend to infinity.*

Proof. Consider the approximation error as $E_{mn\gamma_1\gamma_2}(f)(x,y) = \mathcal{J}_{mn\gamma_1\gamma_2}(f)(x,y) - f(x,y)$. Applying the definition of $K_{mn\gamma_1\gamma_2}$, modulus of continuity of f, and the changes $x_i - x = 2u_i, y_j - y = 2v_j$ we obtain

$$|E_{mn\gamma_1\gamma_2}(f)(x,y)| \leq 2K_{mn\gamma_1\gamma_2}(x,y)\sum_{i=0}^{2m}\sum_{j=0}^{2n}(\omega_f(\bar{u}_i)+\omega_f(\bar{v}_j))\left|\frac{\sin m\bar{u}_i}{m\sin\bar{u}_i}\right|^{\gamma_1}\left|\frac{\sin n\bar{v}_j}{n\sin\bar{v}_j}\right|^{\gamma_2},$$

where $\bar{u}_i, \bar{v}_j$ are constructed as increasing order in $|u_i|, |v_j|$ respectively. From the inequalities (15) and (16) of the reference [14], for all $i, j \geq 2$,

$$\left|\frac{\sin(\frac{m(x_i-x)}{2})}{m\sin(\frac{x_i-x}{2})}\right|^{\gamma_1} \leq \left(\frac{2}{i}\right)^{\gamma_1} \text{ and } \left|\frac{\sin(\frac{n(y_j-y)}{2})}{n\sin(\frac{y_j-y}{2})}\right|^{\gamma_2} \leq \left(\frac{2}{j}\right)^{\gamma_2}. \tag{4.1}$$

Using (4.1) and similar lines as given in [14], we obtain an error bound as

$$|E_{mn\gamma_1\gamma_2}(f)(x,y)| \leq \omega_f\left(\frac{\pi}{4m}+\frac{\pi}{4n}\right)F(\gamma_1,\gamma_2),$$

where $F(\gamma_1, \gamma_2)$ is independent of m, n. Thus the error term tends to zero when the partition is indefinitely refined. □

Definition 2. The fractal operator of Jackson approximation of a continuous f on the torus is defined as

$$\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f)(x,y) = K_{mn\gamma}(x,y)\sum_{i=0}^{2m}\sum_{j=0}^{2n}f(x_i,y_j)P^{\alpha}_{mi\gamma_1}(x)Q^{\beta}_{nj\gamma_2}(y).$$

Theorem 5. *Let $f \in C(T^1 \times T^1)$ and $\gamma_1, \gamma_2 > 2$, then*

$$\|\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f)-f\|_\infty \leq mn\left(\frac{\pi}{2}\right)^{2\gamma_{\max}}\left(\frac{|\alpha|_\infty}{1-|\alpha|_\infty}\|I-L\|\|\mathcal{F}^{\beta}\|_\infty+\frac{|\beta|_\infty}{1-|\beta|_\infty}\|I-L^*\|\right)+\omega_f\left(\frac{1}{m}+\frac{1}{n}\right)F(\gamma_1,\gamma_2),$$

where α, β are suitable scaling vectors used to construct the fractal perturbation of the basis functions $P_{mi\gamma_1}$ and $Q_{nj\gamma_2}$.

Proof. To attain the prescribed upper bound we will use

$$\|\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f) - f\|_\infty \le \|\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f) - \mathcal{J}_{mn\gamma_1\gamma_2}(f)\|_\infty + \|\mathcal{J}_{mn\gamma_1\gamma_2}(f) - f\|_\infty.$$

According to the definition of $\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f)$ and $K_{mn\gamma_1\gamma_2}$,

$$\|\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f) - \mathcal{J}_{mn\gamma_1\gamma_2}(f)\|_\infty \le \|K_{mn\gamma_1\gamma_2}\|_\infty \left\| \sum_{i=0}^{2m}\sum_{j=0}^{2n} f(x_i, y_j)\left(P^\alpha_{mi\gamma_1}(x)Q^\beta_{nj\gamma_2}(y) - P_{mi\gamma_1}(x)Q_{nj\gamma_2}\right)\right\|_\infty .$$

The norm of the sum in the previous expression can be bounded as

$$\begin{aligned}
&\left\| \sum_{i=0}^{2m}\sum_{j=0}^{2n} f(x_i, y_j)\left(P^\alpha_{mi\gamma_1}Q^\beta_{nj\gamma_2} - P_{mi\gamma_1}Q_{nj\gamma_1}\right)\right\|_\infty \\
&\le \|f\|_\infty \sum_{i=0}^{2m}\sum_{j=0}^{2n} \|P^\alpha_{mi\gamma_1}Q^\beta_{nj\gamma_2} - P_{mi\gamma_1}Q_{nj\gamma_2}\|_\infty \qquad (4.2)\\
&\le \|f\|_\infty \sum_{i=0}^{2m}\sum_{j=0}^{2n} \left(\|P^\alpha_{mi\gamma_1}Q^\beta_{nj\gamma_2} - P_{mi\gamma_1}Q^\beta_{nj\gamma_2}\|_\infty + \|P_{mi\gamma_1}Q^\beta_{nj\gamma_2} - P_{mi\gamma_1}Q_{nj\gamma_2}\|_\infty\right).
\end{aligned}$$

Now the first norm of the expression (4.2) in the parenthesis can be bounded as

$$\begin{aligned}
\|P^\alpha_{mi\gamma_1}Q^\beta_{nj\gamma_2} - P_{mi\gamma_1}Q^\beta_{nj\gamma_2}\|_\infty &\le \|P^\alpha_{mi\gamma_1} - P_{mi\gamma_1}\|_\infty \|Q^\beta_{nj\gamma_2}\|_\infty \\
&\le \frac{|\alpha|_\infty}{1-|\alpha|_\infty}\|P_{mi\gamma_1} - LP_{mi\gamma_1}\|_\infty \|\mathcal{F}^\beta\|\,\|Q_{nj\gamma_2}\|_\infty \qquad (4.3)\\
&\le \frac{|\alpha|_\infty}{1-|\alpha|_\infty}\|I - L\|\|P_{mi\gamma_1}\|_\infty \|\mathcal{F}^\beta\|\,\|Q_{nj\gamma_2}\|_\infty,
\end{aligned}$$

where we have assumed $b_{mi\gamma_1} = LP_{mi\gamma_1}$ for a bounded linear operator L. But $\|P_{mi\gamma_1}\|_\infty \le 1$, $\|Q_{nj\gamma_2}\|_\infty \le 1$ due to Lemma 2. Similarly, the second norm of (4.2) in the parenthesis can be bounded as

$$\begin{aligned}
\|P_{mi\gamma_1}Q^\beta_{nj\gamma_2} - P_{mi\gamma_1}Q_{n,j,\gamma_2}\|_\infty &\le \|P_{mi\gamma_1}\|_\infty \|Q^\beta_{nj\gamma_2} - Q_{n,j,\gamma_2}\|_\infty \\
&\le \frac{|\beta|_\infty}{1-|\beta|_\infty}\|I - L^*\|, \qquad (4.4)
\end{aligned}$$

where $b^*_{nj\gamma_2} = L^*Q_{nj\gamma_2}$ for a bounded linear operator L^*. Finally, using (4.3), (4.4) in (4.2) we obtain

$$\left\| \sum_{i=0}^{2m}\sum_{j=0}^{2n} f(x_i, y_j)\left(P^\alpha_{m,i,\gamma_1}Q^\beta_{n,j,\gamma_2} - P_{m,i,\gamma_1}Q_{n,j,\gamma_2}\right)\right\|_\infty \le \frac{|\alpha|_\infty}{1-|\alpha|_\infty}\|I - L\|\,\|\mathcal{F}^\beta\| + \frac{|\beta|_\infty}{1-|\beta|_\infty}\|I - L^*\|.$$

Using Lemma 3, Theorem 4 and the above expression, the final bound for the error is

$$\|\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f) - f\|_\infty \le mn\left(\frac{\pi}{2}\right)^{2\gamma_{\max}}\left(\frac{|\alpha|_\infty}{1-|\alpha|_\infty}\|I - L\|\,\|\mathcal{F}^\beta\| + \frac{|\beta|_\infty}{1-|\beta|_\infty}\|I - L^*\|\right) +$$

$$\omega_f\left(\frac{1}{m}+\frac{1}{n}\right)F(\gamma_1,\gamma_2).$$

□

Corollary 6. *If $f \in C([-\pi,\pi]\times[-\pi,\pi]), \gamma_1,\gamma_2 > 2$ and if we choose scaling vectors α,β such that $mn|\alpha|_\infty$, $mn|\beta|_\infty$ have the same rate of convergence as that of ω_f, then the discrete fractal approximant $\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f)$ converges uniformly to f as m,n tend to infinity. The order of convergence does not depend on γ_1,γ_2.*

Remark 1. The present approach may be extended to high-dimensional settings, for functions defined on hypertori. The convergence results would remain qualitatively equal to those exposed in this paper.

§5. Example

In this section we give the numerical explanation of the proposed approximants for different exponents and scale vectors. Figure 1(a) represents the graph of the smooth function $f(x,y) = 2\sin^2(x) + 3\cos^2(y)$ over the interval $[-\pi,\pi]\times[-\pi,\pi]$. Figure 1(b) represents the surface corresponding to the discrete approximant $\mathcal{J}_{mn\gamma_1\gamma_2}(f)$ for the values of $m = n = 10$ and $\gamma_1 = \gamma_2 = 4$. In order to get the fractal surface $\mathcal{J}^{\alpha\beta}_{mn\gamma_1\gamma_2}(f)$ corresponding to the discrete surface data, we consider a uniform partition of $[-\pi,\pi]$ in both directions with $M = N = 10$. Figure 1(c) depicts the fractal surface corresponding to $\alpha_i = \beta_i = 0.12$ for $i = 1,2,\ldots,N$ and $\gamma_1 = \gamma_2 = 4$. Figure 1(d) represents another periodic fractal surface for $\alpha_i = 0.08, \beta_i = 0.1$ for $i = 1,2,\ldots,N$ and $\gamma_1 = 3, \gamma_2 = 4$. Usually, we tend to think that the sample points come from a smooth function, but in practice this is not always the case. Thus for non-smooth periodic surface data, these procedures may help to provide a better approximation.

Acknowledgements

This work has been partially supported by the Projects CUD-ID: 2015-05 and CUD-ID: 2017-03 of the Centro Universitario de la Defensa de Zaragoza.

References

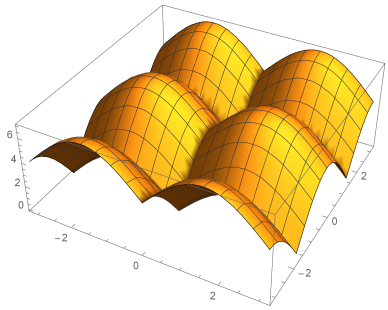
[1] Barnsley, M. F. Fractal functions and interpolation. *Constr. Approx. 2*, 1 (1986), 303–329. Available from: https://doi.org/10.1007/BF01893434.

[2] Barnsley, M. F. *Fractals Everywhere*. Academic Press, Inc., Boston, MA, 1988.

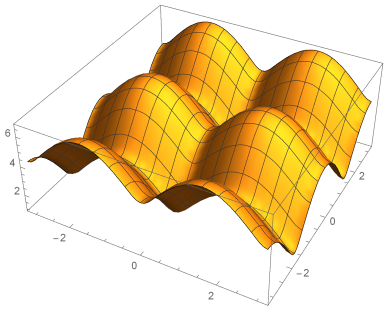
[3] Barnsley, M. F., and Harrington, A. N. The calculus of fractal interpolation functions. *J. Approx. Theory 57*, 1 (1989), 14–34. Available from: https://doi.org/10.1016/0021-9045(89)90080-4.

[4] Cheney, E. W. *Introduction to Approximation Theory*. AMS Chelsea Publishing, Providence, RI, 1998.

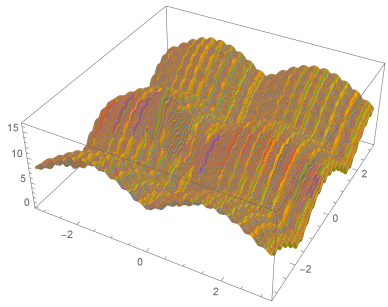
[5] Hutchinson, J. E. Fractals and self-similarity. *Indiana Univ. Math. J. 30*, 5 (1981), 713–747. Available from: https://doi.org/10.1512/iumj.1981.30.30055.



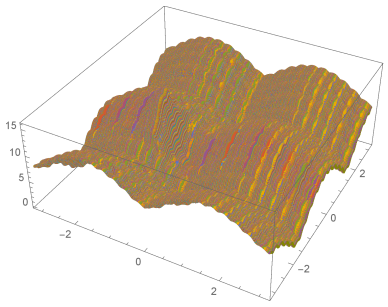
(a) Graph of f.



(b) Graph of classical discrete approximant.



(c) Graph of the fractal discrete approximant for $\gamma_1 = \gamma_2$.



(d) Graph of the fractal discrete approximant for different γ_1, γ_2.

Figure 1: Graph of f, its discrete classical and fractal approximants for different values of γ_1, γ_2, α and β.

[6] Jackson, D. On approximation by trigonometric sums and polynomials. *Trans. Amer. Math. Soc. 13*, 4 (1912), 491–515. Available from: https://doi.org/10.2307/1988583.

[7] Jackson, D. On the accuracy of trigonometric interpolation. *Trans. Amer. Math. Soc. 14*, 4 (1913), 453–461. Available from: https://doi.org/10.2307/1988698.

[8] Massopust, P. R. *Interpolation and Approximation with Splines and Fractals*. Oxford University Press, Oxford, 2010.

[9] Massopust, P. R. *Fractal Functions, Fractal Surfaces, and Wavelets*, second ed. Elsevier/Academic Press, London, 2016.

[10] Navascués, M. A. Fractal polynomial interpolation. *Z. Anal. Anwendungen 24*, 2 (2005), 401–418. Available from: https://doi.org/10.4171/ZAA/1248.

[11] Navascués, M. A., Jha, S., Chand, A. K. B., and Sebastián, M. V. Fractal approximants on the circle. *Chaotic Modeling and Simulation 3* (2018), 343–353.

[12] Navascués, M. A., Jha, S., Chand, A. K. B., and Sebastián, M. V. Fractal approximation of Jackson type for periodic phenomena. *Fractals 26*, 5 (2018), 1850079. Available from: https://doi.org/10.1142/S0218348X18500792.

[13] Navascués, M. A., Jha, S., Chand, A. K. B., and Sebastián, M. V. Generalized trigonometric interpolation. *J. Comput. Appl. Math. 354* (2019), 152–162. Available from: https://doi.org/10.1016/j.cam.2018.08.003.

[14] Navascués, M. A., and Sebastián, M. V. Fitting functions of Jackson type for three-dimensional data. *Int. J. Comput. Math.* (2018), 1–18. Available from: https://doi.org/10.1080/00207160.2018.1458099.

M. A. Navascués
Departamento de Matemática Applicada
Universidad de Zaragoza
500018 Zaragoza, Spain
manavas@unizar.es

S. Jha and A. K. B. Chand
Department of Mathematics
Indian Institute of Technology Madras
600036 Chennai, India
sangitajha285@gmail.com and chand@iitm.ac.in

M. V. Sebastián
Centro Universitario de la Defensa de Zaragoza
Academia General Militar
50090 Zaragoza, Spain
msebasti@unizar.es

Monografías Matemáticas *García de Galdeano* **42**, 229–241 (2019)

NON-ASSOCIATIVE ALGEBRAIC HYPERSTRUCTURES AND ITS APPLICATIONS TO BIOLOGICAL INHERITANCE

Oyeyemi O. Oyebola and Temitope G. Jaiyeola

Abstract. In this paper, we investigate non-associative properties in algebraic hyperstructures as it plays out in the biological inheritance which is expressed in the genotypic and phenotypic information that are passed to the progenies from the parental traits. This is with the intention to valuate with precision the non-associativity of weak associative properties in algebraic structures derived from some biological inheritance crossing. Examples of biological inheritance crossing which obey the WASS condition $x \cdot (y \cdot z) \cap (x \cdot y) \cdot z \neq \emptyset$ for the $1, 2, 3$-variable forms were found (though the corresponding identities were not obeyed). The structures $(H, \otimes)$ were found to be hypergroupoids or hyperquasigroups which obey 1-variable identity (3-power associativity) or 2-variable identities (LAP, RAP or flexibility) or 3-variable identities (extra-1 or extra-2 or extra-3). Such hyperstructures can be termed to be 3-power associative, flexible, left (right) alternative or extra; in their precise measure of weakness in associativity.

Keywords: hypergroup, hypersemigroup, H_v-group, H_v-semigroup, H_v-structures, Filial generations.

AMS classification: 20N20.

§1. Introduction

The study of algebraic hyperstructures was born in 1934 by F. Marty [6] when he gave the definitions of hypergroups and illustrated with some applications. It had since been a motivating platform for further studies in hyperstructures and its applications to other issues of life. Hyperstructures are algebraic structures equipped with at least one multi-valued operation, called a hyperoperation. The largest classes of hyperstructures are the ones called H_v-structures. Algebraic hyperstructures are suitable generalizations of classical algebraic structures. In a classical algebraic structure, the composition of two elements is an element, while in an algebraic hyperstructure, the composition of two elements is a set. Algebraic hyperstructure theory has a multiplicity of applications to other disciplines such as geometry, graphs and hypergraphs, binary relations, lattices, groups, fuzzy sets and rough sets, automata, cryptography, codes, median algebras, relation algebras, C-algebras, artificial intelligence and probability theory.

Etherington presented Genetic algebras in 1939, Non-associative algebra and the symbolism of genetics in 1941. Schafer published Structure of genetic algebras in 1949. Mendel authored Experiments in Plant-Hybridization in 1866.

In this work, our main objective is to showcase that non-associativity in hyperstructures is associated with biological inheritance. We explore some properties in algebraic hyperstructures that naturally occur as genetic information gets passed down through generations. Mathematically, the algebraic hyperstructures that arise in genetics are very interesting ones. They are generally commutative but not associative. It is noteworthy that the order in which genes interact in a given filial generation matters, thus, this necessitated the idea of non-associativity. Hence, the need to valuate with precision the relationship that exist between the progenies of each crosses. Thus, the import of the idea of weak associativity property which the study of hyperstructures availed us. Interestingly, many of the algebraic properties of these hyperstructures have genetic import. This work is furtherance to ideas presented by Davvaz et al. [4], contributions made by Al-Tahan and Davvaz [2, 1], Anvariyeh and Momeni [3] and recent compilations of reports in Davvaz and Vougiouklis [5]

§2. Preliminaries and Basic Definitions

In this section, some basic definitions related to hyperstructures and biological inheritance are presented. It is known that an operation $(\circ)$ on a set H is any map from $H \times H$ to H. In other words, to any two elements $x, y \in H$ there correspond an element of H which we denote $x \circ y$. This map is written as follows

$$\circ : H \times H \to H : (x, y) \mapsto x \circ y \in H$$

Usual operations are the addition $(+)$ and the multiplication $(\cdot)$. Hyperoperation or multivalued operation in a set is any operation which maps to two elements x, y of H into a subset $x * y$ of H. Thus, we write

$$* : H \times H \to P(H)\backslash : (x, y) \mapsto x * y \subset H$$

where P is the power set of H.

A pair $(H, *)$, consisting of a set equipped with a hyperoperation, is called an hypergroupoid. This is the hyperstructure or multivalued structure. Hyperstructure is every algebraic structure in which at least one hyperoperation is defined.

Definition 1. A hypergroup is a pair $(H, \circ)$, where $\circ : H \times H \longrightarrow P^*(H)$, such that the following conditions hold for all x, y, z of H:

1. $(x \circ y) \circ z = x \circ (y \circ z)$ for all $x, y, z \in H$ which means that

$$\bigcup_{u \in x \circ y} u \circ z = \bigcup_{v \in y \circ z} x \circ v$$

2. $H \circ x = x \circ H = H$, where

$$H \circ x = \bigcup_{h \in H} h \circ x \text{ and } x \circ H = \bigcup_{h \in H} x \circ h$$

This condition is called *the reproduction axiom*.

A commutative hypergroup $(H, \circ)$ is a join space if for all x, y, z of H, the following implication holds:

$$x/y \cap z/w \neq \emptyset \Longrightarrow x \circ w \cap y \circ z \neq \emptyset \qquad (\textit{transposition axiom}).$$

where $x/y = \{u \in H | x \in u \circ y\}$.

In 1934, Marty introduced the concept of a hypergroup. The motivation example was the following: Let G be a group and H be any subgroup of G. Then $G/H = \{xH \mid x \in G\}$ becomes a hypergroup where the hyperoperation is defined in a usual manner:

$$xH \circ yH = \{zH \mid z \in xH \cdot yH\},$$

for all $x, y \in G$.

Definition 2. Let $(H, \circ)$ be a hypergroupoid.

(i) An element $e \in H$ is called an identity if, for all $x \in H$, $x \in x \circ e \cap e \circ x$.
An identity e is called scalar identity if, for all $x \in H$, $x \circ e = e \circ x = x$.
An identity e is called partial identity if, for any $x \in H$, $x \in x \circ e$ or $x \in e \circ x$.

(ii) An element $x' \in H$ is called an inverse of $x \in H$ if there is an identity $e \in H$, such that $e \in x \circ x' \cap x' \circ x$.

Definition 3. Let H be a non-empty set and $\cdot : H \times H \longrightarrow P^*(H)$ be a hyperoperation.

(i) Then, the hypergroupoid $(H, \cdot)$ is said to be weak associative if

$$x \cdot (y \cdot z) \cap (x \cdot y) \cdot z \neq \emptyset$$

WASS: the weak associativity

(ii) Then, the hypergroupoid $(H, \cdot)$ is said to be weakly commutative if

$$x \cdot y \cap y \cdot x \neq \emptyset$$

COW: the weak commutativity

(iii) Then, the hypergroupoid $(H, \cdot)$ is said to be strongly commutative if

$$x \cdot y = y \cdot x$$

Remark 1. If $(H, \cdot)$ is an hypergroupoid with WASS, then, it is called an H_v-semigroup. In addition, if $(H, \cdot)$ has the reproduction axiom, then it is called an H_v-group.

Definition 4. Let $(H, \cdot)$ be an hypergroupoid and let $x, y, z \in H$.

(i) $(H, \cdot)$ is said to have the 3-power associativity property (3-PA) if it obeys the identity $(x \cdot x) \cdot x = x(x \cdot x)$.

(ii) $(H, \cdot)$ is said to have the left alternative property (LAP) if it obeys the identity $x \cdot (x \cdot y) = (x \cdot x) \cdot y$.

$\otimes$	RY	Ry	rY	ry
RY	$RRYY$	$RRYy$	$RrYY$	$RrYy$
Ry	$RRYy$	$RRyy$	$RrYy$	$Rryy$
rY	$RrYY$	$RrYy$	$rrYY$	$rrYy$
ry	$RrYy$	$Rryy$	$rrYy$	$rryy$

Table 1: Dihybrid crosses with Pea plants

(iii) $(H, \cdot)$ is said to have the right alternative property (RAP) if it obeys the identity $(y \cdot x) \cdot x = y(x \cdot x)$.

(iv) $(H, \cdot)$ is said to have the flexibility or elasticity if it obeys the identity $(x \cdot y) \cdot x = x(y \cdot x)$.

(v) $(H, \cdot)$ is said to have the extra-1 law if obeys the identity $((x \cdot y) \cdot z) \cdot x = x \cdot (y \cdot (z \cdot x))$.

(vi) $(H, \cdot)$ is said to have the extra-1 law if it obeys the identity $((x \cdot y) \cdot z) \cdot x = x \cdot (y \cdot (z \cdot x))$.

(vii) $(H, \cdot)$ is said to have the extra-2 law if it obeys the identity $(y \cdot x) \cdot (z \cdot x) = (y \cdot (x \cdot z)) \cdot x$.

(viii) $(H, \cdot)$ is said to have the extra-3 law if it obeys the identity $(y \cdot x) \cdot (z \cdot x) = x \cdot ((y \cdot x) \cdot z)$.

(ix) $(H, \cdot)$ is called an hyperquasigroup if it has the reproduction axiom.

Remark 2. For any other weak law (aside WASS and COW), an hypergroupoid $(H, \cdot)$ with such weak law will be called an H_v-structure.

§3. Examples of Different Genetic Inheritance

In his dihybrid crosses with pea plants, Gregor Mendel simultaneously examined two different genes that controlled two different traits. For instance, in one series of experiments, Mendel began by crossing a plant that was homozygous for both round seed shape and yellow seed color (RRYY) with another plant that was homozygous for both wrinkled seed shape and green seed color (rryy). Then, when Mendel crossed two of the F_1 (First Filial generation) progeny plants with each other ($RrYy \times RrYy$), he obtained an F_2(Second Filial generation).

$$P : (\textit{Round and yellow})\ RRYY \ \otimes\ (\textit{wrinkled and green})\ rryy$$

$$F_1 : RrYy$$

$$F_2 : F_1 \ \otimes\ F_1$$

$$F_2 : RrYy \ \otimes\ RrYy$$

Theorem 1. *Let $H = \{RY, Ry, rY, ry\}$ with $\otimes$ defined on H as given in Table 1. Then,*

(i) *$(H, \otimes)$ is a non-associative hyperquasigroup and H_v-group.*

(ii) *$(H, \otimes)$ is an H_v-structure which does not satisfy the left alternative property.*

(iii) *$(H, \otimes)$ is an H_v-structure which does not satisfy the right alternative property.*

(iv) *$(H, \otimes)$ is an H_v-structure which does not satisfy the flexibility property.*

(v) $(H, \otimes)$ *is an* H_v*-structure which is a* 3*-power associative hyperquasigroup*

Proof. $(H, \otimes)$ is an hyperquasigroup based on the multiplication Table 1.

(i) Let us check if $(H, \otimes)$ is associative or not:

$$(Ry \otimes rY) \otimes ry \neq Ry \otimes (rY \otimes ry)$$

$$RrYy \otimes ry \neq Ry \otimes rrYy$$

$\{RrYy, Rryy, rrYy, rryy\} \neq \{RrYy, Rryy\}$ but, $(Ry \otimes rY) \otimes ry \cap Ry \otimes (rY \otimes ry) \neq \emptyset$.

Hence, $(H, \otimes)$ is a non-associative hyperquasigroup and H_v-group.

(ii) Let us check if the left alternative property is satisfied:

$$x \cdot xy = xx \cdot y$$

Then,

$$Ry \otimes (Ry \otimes ry) \neq (Ry \otimes Ry) \otimes ry$$

$$Ry \otimes Rryy \neq RRyy \otimes ry$$

$\{RRyy, Rryy\} \neq \{Rryy, Rryy\}$ but, $Ry \otimes (Ry \otimes ry) \cap (Ry \otimes Ry) \otimes ry \neq \emptyset$.

Hence, $(H, \otimes)$ is an H_v-structure which does not satisfy the left alternative property.

(iii) Let us check if the right alternative property is satisfied:

$$x \cdot yy = xy \cdot y$$

Then,

$$rY \otimes (ry \otimes ry) \neq (rY \otimes ry) \otimes ry$$

$$rY \otimes rryy \neq rrYy \otimes ry$$

$rrYy \neq \{rrYy, rryy\}$ but, $rY \otimes (ry \otimes ry) \cap (rY \otimes ry) \otimes ry \neq \emptyset$.

Hence, $(H, \otimes)$ is an H_v-structure which does not satisfy the right alternative property.

(iv) Let us check if the flexibility property is satisfied:

$$x \cdot yx = xy \cdot x$$

$$Ry \otimes (rY \otimes Ry) \neq (Ry \otimes rY) \otimes Ry$$

$$Ry \otimes RrYy \neq RrYy \otimes Ry$$

$\{RRYy, RRyy, Rryy\} \neq \{RRYy, RRyy, RrYy, Rryy\}$ but, $Ry \otimes (rY \otimes Ry) \cap (Ry \otimes rY) \otimes Ry \neq \emptyset$.

Hence, $(H, \otimes)$ is an H_v-structure which does not satisfy the flexibility property.

(v) It can be deduced from the multiplication Table 1 that the 3-power associativity property holds:

$$x \cdot xx = xx \cdot x \; \forall \; x \in H.$$

For instance, $Ry \otimes (Ry \otimes Ry) = (Ry \otimes Ry) \otimes Ry$

$$Ry \otimes RRyy = RRyy \otimes Ry$$

$$RRyy = RRyy. \quad \square$$

Remark 3. Therefore, $(H, \otimes)$ is an H_v-structure, which is a 3-power associative hyperquasigroup whose weakness in associativity is 1-variable measurable and not 2-variable measurable because it failed LAP, RAP and flexibility property.

$\otimes$	AA	Aa	aa
AA	AA	$\{AA, Aa\}$	Aa
Aa	$\{AA, Aa\}$	$\{AA, Aa, aa\}$	$\{Aa, aa\}$
aa	Aa	$\{Aa, aa\}$	aa

Table 2: Hereditary information inherited from crosses

3.1. Simple Mendelian Inheritance

The zygotes AA and aa are called homozygous, since they carry two copies of the same allele. In this case, simple Mendelian inheritance means that there is no chance involved as to what genetic information will be inherited in the next generation; i.e., AA will pass on the allele A and aa will pass on a. However, the zygotes Aa and aA (which are equivalent) each carry two different alleles. These zygotes are called heterozygous. The rules of simple Mendelian inheritance indicate that the next filial generation will inherit either A or a with equal measure. So, when two gametes reproduce, a multiplication is induced which indicates how the hereditary information will be passed down to the next filial generation. This multiplication is given by the following rules:

1. $A \times A = A$
2. $A \times a = \{A, a\}$
3. $a \times A = \{a, A\}$
4. $a \times a = a$

In 1. and 4. above, both gametes carry the same allele, while there is equal presence of the two alleles in 2. and 3.

Theorem 2. *Let $H = \{AA, Aa, aa\}$ with $\otimes$ defined on H as given in Table 2. Then,*

(i) $(H, \otimes)$ *is a non-associative hypergroupoid, not a hyperquasigroup and a H_v-semigroup.*

(ii) $(H, \otimes)$ *is an H_v-structure which does not satisfy the left alternative property.*

(iii) $(H, \otimes)$ *is an H_v-structure which does not satisfy the right alternative property.*

(iv) $(H, \otimes)$ *is an H_v-structure which satisfies the flexibility property.*

(v) $(H, \otimes)$ *is an H_v-structure which is a 3-power associative hypergroupoid.*

(vi) $(H, \otimes)$ *is an H_v-structure which satisfies the extra-1 identity.*

(vii) $(H, \otimes)$ *is an H_v-structure which does not satisfy the extra-2 identity.*

(viii) $(H, \otimes)$ *is an H_v-structure which does not satisfy the extra-3 identity.*

Proof. $(H, \otimes)$ is an hypergroupoid and not a hyperquasigroup based on the multiplication Table 2.

(i) We shall show that the hypergroupoid $(H, \otimes)$ is non-associative:

$$(AA \otimes Aa) \otimes aa \neq AA \otimes (Aa \otimes aa)$$

$$\{AA, Aa\} \otimes aa \neq AA \otimes \{Aa, aa\}$$

$$\{Aa, aa\} \neq \{AA, Aa\} \text{ but, } (AA \otimes Aa) \otimes aa \cap AA \otimes (Aa \otimes aa) \neq \emptyset$$

Hence, $(H, \otimes)$ is non-associative and an H_v-semigroup.

(ii) Let us check if the left alternative property (LAP) is satisfied:

$$xx \cdot y = x \cdot xy$$

$$(Aa \otimes Aa) \otimes aa \neq Aa \otimes (Aa \otimes aa)$$

$$\{AA, Aa, aa\} \otimes aa \neq Aa \otimes \{Aa, aa\}$$

$$\{Aa, aa\} \neq \{AA, Aa, aa\} \text{ but, } (Aa \otimes Aa) \otimes aa \cap Aa \otimes (Aa \otimes aa) \neq \emptyset$$

Hence, LAP is not satisfied by $(H, \otimes)$. So, $(H, \otimes)$ is an H_v-structure.

(iii) Let us check if the right alternative property (RAP) is also satisfied:

$$xy \cdot y = x \cdot yy$$

$$(AA \otimes Aa) \otimes Aa \neq AA \otimes (Aa \otimes Aa)$$

$$\{AA, Aaa\} \otimes Aa \neq AA \otimes \{AA, Aa, aa\}$$

$$\{AA, Aa, aa\} \neq \{AA, Aa\} \text{ but, } (AA \otimes Aa) \otimes Aa \cap AA \otimes (Aa \otimes Aa) \neq \emptyset$$

Hence, RAP is not satisfied by $(H, \otimes)$. So, $(H, \otimes)$ is an H_v-structure.

(iv) We shall show that flexibility property holds in $(H, \otimes)$ by considering the following and others:

$$x \cdot yx = xy \cdot x \ \forall x, y \in H.$$

(a) $$AA \otimes (aa \otimes AA) = (AA \otimes aa) \otimes AA$$

$$AA \otimes Aa = Aa \otimes AA$$

$$\{AA, Aa\} = \{AA, Aa\}.$$

(b) $$AA \otimes (Aa \otimes AA) = (AA \otimes Aa) \otimes AA$$

$$AA \otimes \{AA, Aa\} = \{AA, Aa\} \otimes AA$$

$$\{AA, Aa\} = \{AA, Aa\}.$$

(c) $$Aa \otimes (aa \otimes Aa) = (Aa \otimes aa) \otimes Aa$$

$$Aa \otimes \{Aa, aa\} = \{Aa, aa\} \otimes Aa$$

$$\{AA, Aa, aa\} = \{Aa, Aa, aa\}.$$

Hence, we see that $(H, \otimes)$ satisfies the flexibility property.

(v) We shall now show that the 3-power associative property is true:

$$x \cdot xx = xx \cdot x$$

Then,

$$\text{(d)} \qquad AA \otimes (AA \otimes AA) = (AA \otimes AA) \otimes AA$$

$$AA = AA$$

$$\text{(e)} \qquad Aa \otimes (Aa \otimes Aa) = (Aa \otimes Aa) \otimes Aa$$

$$Aa \otimes \{Aa, Aa, aa\} = \{AA, Aa, aa\} \otimes Aa$$

$$\{AA, Aa, aa\} = \{AA, Aa, aa\}$$

$$\text{(e)} \qquad aa \otimes (aa \otimes aa) = (aa \otimes aa) \otimes aa$$

$$aa = aa$$

Hence, by (d), (e) and (f), we see that $(H, \otimes)$ satisfies the 3-power associative property.

(vi) We shall show that extra-1 identity holds in $(H, \otimes)$ by considering the following and others:

$$(xy \cdot z)x = x(y \cdot zx) \ \forall \ x, y, z \in H$$

Then,

$$((AA \otimes Aa) \otimes aa) \otimes AA = AA \otimes (Aa \otimes (aa \otimes AA))$$

$$(\{AA, Aa\} \otimes aa) \otimes AA = AA \otimes (Aa \otimes Aa)$$

$$\{Aa, aa\} \otimes AA = AA \otimes \{AA, Aa, aa\}$$

$$\{AA, Aa\} = \{AA, Aa\}$$

Hence, $(H, \otimes)$ satisfies extra-1 identity.

(vii) Let us check if $(H, \otimes)$ satisfies extra-2 identity:

$$yx \cdot zx = (y \cdot xz)x$$

Then,

$$(Aa \otimes AA) \otimes (aa \otimes AA) \neq (Aa \otimes (AA \otimes aa)) \otimes AA$$

$$\{AA, Aa\} \otimes Aa \neq (Aa \otimes Aa) \otimes AA$$

$\{AA, Aa, aa\} \neq \{AA, Aa\}$ but, $(Aa \otimes AA) \otimes (aa \otimes AA) \cap (Aa \otimes (AA \otimes aa)) \otimes AA \neq \emptyset$

Hence, $(H, \otimes)$ does not satisfy extra-2 identity.

(viii) Let us check if $(H, \otimes)$ satisfies extra-3 identity:

$$xy \cdot xz = x(yx \cdot z)$$

Then,

$$(AA \otimes Aa) \otimes (AA \otimes aa) \neq AA \otimes ((Aa \otimes AA) \otimes aa)$$

$$\{AA, Aa\} \otimes Aa \otimes AA \otimes (\{\{AA, Aa\}) \otimes aa)$$

$$\{AA, Aa, aa\} \neq AA \otimes \{AA, Aa, aa\}$$

$\{AA, Aa, aa\} \neq \{AA, Aa\}$ but, $(AA \otimes Aa) \otimes (AA \otimes aa) \cap AA \otimes ((Aa \otimes AA) \otimes aa) \neq \emptyset$

Hence, $(H, \otimes)$ does not satisfy extra-3 identity. □

Remark 4. Therefore, $(H, \otimes)$ is an H_v-structure, which is a 3-power associative, flexible and extra-1 hyperqroupoid. Its weakness in associativity is $1, 2, 3$-variable measurable even though it failed LAP and RAP. Depending on the algebraic properties that are satisfied, these can be used to categorise each cross mating that takes place. It can also be used as counsel to guide in experimentation procedures in cross breeding, in order to cut cost, manage time, energy and materials. It gives an added advantage over being probabilistic in experimentation.

3.2. Combs in Chicken

The research conducted by the British geneticists, William Bateson and R. C. Punnett (4) showed that the shape of the comb in chickens was caused by the interaction between two different genes. Bateson and Punnett were aware of the fact that different varieties of chickens possess distinctive combs. For instance, Wyandottes have a “rose" comb, Brahmas have a “pea" comb, and Leghorns have a “single" comb. When Bateson and Punnett crossed a Wynadotte chicken with a Brahma chicken, all of the F_1 progeny had a new type of comb, which the duo termed a “walnut" comb. In this case, neither the rose comb of the Wyandotte nor the pea comb of the Brahma appeared to be dominant, because the F_1 offspring had their own unique phenotype.

$$P : RRpp \otimes rrPP$$

$$F_1 : RrPp$$

$$F_2 : RrPp \otimes RrPp$$

Moreover, when two of these F_1 progeny were crossed with each other, some of the members of the resulting F_2 generation had walnut combs, some had rose combs, some had pea combs, and some had a single comb. Because the four comb shapes appeared in a 9:3:3:1 ratio (i.e., nine walnut chickens per every three rose chickens per every three pea chickens per every one single-comb chicken), it seemed that two different genes must play a role in comb shape. Through continued research, Bateson and Punnett deduced that Wyandotte (rose-combed) chickens must have the genotype $RRpp$, while Brahma chickens must have the genotype $rrPP$. A cross between a Wyandotte and a Brahma would yield offspring that all had the $RrPp$ genotype, which manifested as the walnut-comb phenotype. Indeed, any chicken with at least one rose-comb allele (R) and one pea-comb allele (P) would have a walnut comb. Thus, when two F_1 walnut chickens were crossed, the resulting F_2 generation would yield rose-comb chickens ($RRpp$), pea-comb chickens ($rrPP$), and walnut-comb chickens ($RrPp$), as well as chickens with a new, fourth phenotype—the single-comb phenotype. Based on the process of elimination, it could be assumed that these single-comb chickens had the $rrpp$ genotype (Bateson & Punnett, 1905; 1906; 1908).

Lemma 3. *Let $H = \{RP, Rp, rP, rp\}$ with $\otimes$ defined on H as given in Table 3. Then, $(H, \otimes)$ is a non-associative hyperquasigroup and an H_v-group.*

Proof. $(H, \otimes)$ is an hyperquasigroup based on the multiplication Table 3.

$$(RP \otimes rP) \otimes rp \neq RP \otimes (rP \otimes rp)$$

$$RrPP \otimes rp \neq RP \otimes rrPp$$

$\otimes$	RP	Rp	rP	rp
RP	$RRPP$	$RRPp$	$RrPP$	$RrPp$
Rp	$RRPp$	$RRpp$	$RrPp$	$Rrpp$
rP	$RrPP$	$RrPp$	$rrPP$	$rrPp$
rp	$RrPp$	$Rrpp$	$rrPp$	$rrpp$

Table 3: Crosses of Combs in Chicken

$$\{RrPp, rrPp\} \neq \{RrPP, RrPp\} \text{ but, } (RP \otimes rP) \otimes rp \cap RP \otimes (rP \otimes rp) \neq \emptyset$$

Hence, $(H, \otimes)$ is a non-associative hyperquasigroup and an H_v-group. □

§4. Non-associativity of Genetic Inheritance

Algebraic hyperstructure with genetic realization are not necessarily associative but may be weakly associative. It seems logical that the order in which populations mate is significant. i.e., if parents A and B mate and then the resulting progenies mates with C, the resulting progeny is not the same as the offsprings resulting from A mating with the progenies obtained from mating parents B and C originally. Symbolically, $(A \times B) \times C$ is not equal to $(A \times (B \times C)$. Epistasis: One set of alleles (a gene) may mask or inhibit the expression of another gene's alleles.

4.1. Epistasis of Dominant Traits in Eye Color

The two allelomorphs responsible for eye color, christened $OCA2$ and $HERC2$ may be represented by Oo and Hh. O and H are dominant over o and h. The alleles interact as shown below:
$Omhh$ and $oomn$ have phenotype blue and $OmHn$ has phenotype brown.
In this case, $m = O$ or o and $n = H$ or h. Hence, we have the result as stated below:

$$P : OOHH \otimes oohh$$

$$F_1 : OoHh$$

and

$$F_1 \otimes F_1 : OoHh \otimes OoHh$$

$$F_2 : Brown, Blue, Blue$$

Brown is represented by D_1, Blue by D_2 and Blue by D_3.

Remark 5. Note that, phenotypically there is no distinction between D_2 and D_3 but there is a clear distinction between their genotypic composition. Hence, the genotypic representation of the resulting offsprings in F_2 is given as:

$$F_2 : \hat{D_1}(of genotype OOHH), \hat{D_2}(of genotype OOhh), \hat{D_3}(of genotype oohh)$$

OOHH (Brown)
OOHh (Brown)
OOhh (Blue)
OoHH (Brown)
OoHh (Brown)
Oohh (Blue)
ooHH (Blue)
ooHh (Blue)
oohh (Blue)

Table 4: Different genetic combinations of eye colors

$\otimes$	OH	Oh	oH	oh
OH	$OOHH$ (Brown)	$OOHh$ (Brown)	$OoHH$ (Brown)	$OoHh$ (Brown)
Oh	$OOHh$ (Brown)	$OOhh$ (Blue)	$OoHh$ (Brown)	$Oohh$ (Blue)
oH	$OoHH$ (Brown)	$OoHh$ (Brown)	$ooHH$ (Blue)	$ooHh$ (Blue)
oh	$OoHh$ (Brown)	$Oohh$ (Blue)	$ooHh$ (Blue)	$oohh$ (Blue)

Table 5: Genes that are far apart or on different chromosomes

Genes come in different versions (or alleles). OCA2 comes in brown (O) and blue (o) versions. HERC2 also comes in two different versions, brown (H) and blue (h). Since people have two copies of each gene, there are nine different possible genetic combinations. This is expressed in Table 4.

Thus, from the result of above experiment, we have that:

$$(\hat{D}_1 \otimes \hat{D}_2) \otimes \hat{D}_3 \neq \hat{D}_1 \otimes (\hat{D}_2 \otimes \hat{D}_3)$$

and

$$(D_1 \otimes D_2) \otimes D_3 \neq D_1 \otimes (D_2 \otimes D_3)$$

Since genes come in different versions, resulting in epistatic representation of the phenotypes, we have that:

$$(\hat{D}_1 \otimes \hat{D}_2) \otimes \hat{D}_3 \ \cap \ \hat{D}_1 \otimes (\hat{D}_2 \otimes \hat{D}_3) \neq \emptyset$$

and

$$(D_1 \otimes D_2) \otimes D_3 \ \cap \ D_1 \otimes (D_2 \otimes D_3) \neq \emptyset$$

Based on Table 4, we have the multiplication table given in Table 5.

Lemma 4. *Let $H = \{OH, Oh, oH, oh\}$ with $\otimes$ defined on H as given in Table 5. Then, $(H, \otimes)$ is a non-associative hyperquasigroup and an H_v-group.*

Proof. $(H, \otimes)$ is an hyperquasigroup based on the multiplication Table 3. Now,

$$(OH \otimes oH) \otimes oh \neq OH \otimes (oH \otimes oh)$$

$$OoHH \otimes oh \neq OH \otimes ooHh$$

$$\{OoHh, ooHh\} \neq \{OoHH, OoHh\}$$

Hence, $(H, \otimes)$ is a non-associative hyperquasigroup. In fact,

$$(OH \otimes oH) \otimes oh \bigcap OH \otimes (oH \otimes oh) \neq \emptyset$$

$$\text{because } \{OoHh, ooHh\} \bigcap \{OoHH, OoHh\} = \{OoHh\}.$$

Thus, considering other triplets as well, $(H, \otimes)$ is a H_v-group. □

§5. Summary and Conclusion

After the introduction of the notion of hyperstructures about 80 years ago, a number of researches, including its applications have been carried out. Vougiouklis (1990) introduced and studied weak hyper-algebraic structures (H_v-group) for a pair $(H, \cdot)$ where H is a set and "$\cdot$ " is an hyperoperation, with the axiom

$$x \cdot (y \cdot z) \cap (x \cdot y) \cdot z \neq \emptyset \text{ for all } x, y, z \in H \tag{5.1}$$

some other authors have found the genotypes of F_2-offspring to be a cyclic H_v-semigroup and relationship between algebraic hyperstructures and biological inheritance have been established (Al-Tahan et al. 2017).

The main objective of this paper was to valuate with precision the non-associativity of weak associative properties in algebraic structures derived from some biological inheritance crossing. In this work, examples of biological inheritance crossing which obey axiom (5.1) in the 2, 3-variable forms were found. Though the corresponding identities were not obeyed. The structure $(H, \otimes)$ were found to be hypergroupoids or hyperquasigroups which obey 1-variable identity (3-power associativity) or 2-variable identities (LAP. RAP or flexibility) or 3-variable identities (extra-1 or extra-2 or extra-3). Such hyperstructures can be termed to be 3-power associative, flexible, left (right) alternative or extra; in their precise measure of weakness in associativity.

References

[1] Al-Tahan, M., and Davvaz, B. Algebraic hyperstructures associated to biological inheritance. *Mathematical Bioscience 285* (2017), 112–118.

[2] Al-Tahan, M., and Davvaz, B. N-ary hyperstructures associated to the genotypes of f_2-offspring. *International Journal of Biomathematics 10* (2017).

[3] Anvariyeh, S. M., and Momeni, S. Nn-ary hypergroups and associated with n-ary relations. *Bull. Korean Math. Soc. 50* (2013), 507–524.

[4] Davvaz, B., Dehghan-Nezad, A., and Heidari, M. M. Inheritance of algebraic hyperstructures. *Information Sciences 224* (2013), 180–187.

[5] Davvaz, B., and Vougiouklis, T. *A Walk Through Weak Hyperstructures H_v-Structures*. World Scientific Publishing Co. Pte. Ltd, Singapore, 2019.

[6] Marty, F. Sur une generalization de la notion de groupe. *8th Congress Math. Scandenaves* (1934), 45–49.

O. O. Oyebola
Department of Mathematics
Federal University of Agriculture Abeokuta
Ogun State, Nigeria.
oooyeyemi@gmail.com, oyebolaoo@funaab.edu.ng

T. G. Jaiyeola
Department of Mathematics
Obafemi Awolowo University
Osun State, Nigeria
tjayeola@oauife.edu.ng, jaiyeolatemitope@yahoo.com

Monografías Matemáticas *García de Galdeano* **42**, 243–252 (2019)

BEST REGULARITY FOR A SCHRÖDINGER TYPE EQUATION WITH NON SMOOTH DATA AND INTERPOLATION SPACES

Jean Michel Rakotoson

Keywords: Grand and Small Lebesgue spaces, classical Lorentz-spaces, Interpolation, very weak solution.

AMS classification: Primary 46E30,46B70, 35J65.

Abstract. Given a vector field $U(x)$ and a nonnegative potential $V(x)$ on a smooth open bounded set Ω of $\mathbb{R}^n$, we shall discuss some regularity results for the following equation

$$-\Delta\omega + U\cdot\nabla\omega + V\omega = f \qquad \text{in } \Omega \tag{0.1}$$

whenever δf is a bounded Radon measure with $\delta(x)$ is the distance between x and the boundary $\partial\Omega$.

§1. Introduction

To explain the origin of our study, let us recall some recent results concerning the very weak solution in the sense of Brezis concerning the Laplacian operator, (say $U = V = 0$ in the above equation)
and when f belongs to $L^1_+(\Omega,\delta)\backslash L^1(\Omega;\delta(1+|\ln\delta|)))$ with $\delta(x) = \text{dist}\,(x,\partial\Omega)$, then (see [10])

$$\omega \notin W^1_0 L(\text{Log}\, L) = \left\{v \in W^{1,1}_0(\Omega) : \nabla v \in L(\text{Log}\, L)^n\right\},$$

and

$$\int_\Omega |\nabla\omega|\,|\text{Log}\,\delta| dx = +\infty.$$

More, we have (see [11]) the

Theorem 1. *Let*

$$W_+ = \left\{\psi \in W^{2,n}(\Omega) \cap H^1_0(\Omega) : -\Delta\psi \geqslant 0\right\}$$

and

$$L_+ = \left\{f \in L^1_+(\Omega;\delta) : \exists \psi \in W_+ \ s.t \int_\Omega f(x)\psi(x)dx = +\infty\right\}.$$

Then the unique solution $u \in L^{n',\infty}(\Omega)$ of

$$\int_\Omega u\Delta\varphi = \int_\Omega f\varphi,\ \forall\,\varphi \in C^2_0(\overline{\Omega}) = \left\{\varphi \in C^2(\overline{\Omega}) : \varphi = 0 \text{ on } \partial\Omega\right\}$$

verifies

$$\int_\Omega |\nabla u| dx = +\infty : u \notin W^{1,1}(\Omega).$$

But we know (see [1]), that

$$W^1(L(\mathrm{Log}\, L)) \subset L^{n'}(\mathrm{Log}\, L)^{\beta(n'-1)} \quad \forall \beta > 1,\ n' = \frac{n}{n-1}.$$

and this last set is included in the so called small Lebesgue spaces

$$L^{(n',1} \subset L^{(n',\alpha}, \quad 0 < \alpha < 1.$$

Nevertheless, we have shown in [6] that if f is in $L^1(\Omega; \delta(1 + |\mathrm{Log}\, \delta|)^\alpha)$, $\frac{1}{n'} < \alpha \leqslant 1$ then the unique solution u of the equation (0.1) belongs to $L^{(n',\theta}(\Omega)$ for some θ.
More precisely, we have shown in [4, 6] the following

Theorem 2. *Let Ω be a bounded open set of class C^2 of $\mathbb{R}^n$, $|\Omega| = 1$, $\alpha \geqslant \frac{1}{n'}$ where $n' = \frac{n}{n-1}$, $f \in L^1(\Omega; \delta)$. Consider $u \in L^{n',\infty}(\Omega)$, the* v.w.s. *of*

$$-\int_\Omega u \Delta \varphi dx = \int_\Omega f\varphi dx \qquad \forall\, \varphi \in C^2(\overline{\Omega}),\ \varphi = 0 \text{ on } \partial\Omega. \tag{1.1}$$

Then,

1. *if $f \in L^1\Big(\Omega; \delta(1 + |\mathrm{Log}\, \delta|)^\alpha\Big)$, and $\alpha > \frac{1}{n'}$*

$$u \in L^{(n',n\alpha-n+1}(\Omega) = G\Gamma(n', 1; w_\alpha),\ w_\alpha(t) = t^{-1}(1 - \mathrm{Log}\, t)^{\alpha-1-\frac{1}{n'}}$$

and

$$\|u\|_{G\Gamma(n',1;w_\alpha)} \leqslant K_0 |f|_{L^1\left(\Omega;\delta(1+|\mathrm{Log}\,\delta|)^\alpha\right)} \tag{1.2}$$

2. *if $\alpha = \frac{1}{n'}$ then*

$$u \in L^{n'}(\Omega) \text{ and similar estimate as (1.2) holds.}$$

In a recent paper [5], we improve the inequality (1.2) namely for the dimension 2 by getting similar information for $\alpha \leqslant \frac{1}{2}$. Here, we want to extend those results replacing the Laplacian operator by a more general one as it is given in (0.1). Namely, we shall prove the following:

Theorem 3. *Let U be in $L^p(\Omega)^n$, $p > n$, $\mathrm{div}\,(U) = 0$ in $\mathcal{D}'(\Omega)$, $U \cdot \nu = 0$ on $\partial\Omega$, $V \in L^p(\Omega)$, $V \geqslant 0, \beta > \frac{n-1}{n}$, $f \in L^1(\Omega; \delta(1 + |\mathrm{Log}\, \delta|)^\beta)$, $\beta = \frac{n-1+\theta}{n}$, $\theta = n\beta - n + 1$.*
Then the unique solution $u \in L^{n',\infty}(\Omega)$ of

$$\int_\Omega u\Big[-\Delta\varphi - U \cdot \nabla\varphi + V\varphi\Big] dx = \int_\Omega f\varphi dx \quad \forall\, \varphi \in C_0^2(\overline{\Omega}) \tag{1.3}$$

belongs to $L^{(n',\theta}(\Omega)$ *and there exists a constant c (depending only of the data U V and* Ω*) such that*

$$\|u\|_{L^{(n',\theta}} \leqslant c \int_\Omega |f|\delta(1+|\mathrm{Log}\,|)^\beta dx.$$

When f is in $L(\mathrm{Log}\,L)^\beta$, we may obtain a similar result concerning the gradient of u but under weaker assumptions on the operator, we will show for $\beta > \dfrac{1}{n'}$

$$\|\nabla u\|_{L^{(n',n\beta-n+1}} \leqslant c\|f\|_{L(\mathrm{Log}\,L)^\beta}. \tag{1.4}$$

§2. Notation Primary results

For a measurable function $f : \Omega \to \mathbb{R}$, we set for $t \geqslant 0$

$$D_f(t) = \text{measure}\left\{x \in \Omega : |f(x)| > t\right\}$$

and f_* the decreasing rearrangement of $|f|$, for $s \in (0, |\Omega|)$

$$f_*(s) = \inf\left\{t : D_f(t) \leqslant s\right\},\ |\Omega| \text{ is the measure of } \Omega,$$

that we shall assume to be equal to 1 for simplicity.

If A_1 and A_2 are two quantities depending on some parameters, we shall write

$$A_1 \lesssim A_2 \text{ if there exists } c > 0 \text{ (independent of the parameters) such that } A_1 \leqslant cA_2$$

$$A_1 \simeq A_2 \text{ if and only if } A_1 \lesssim A_2 \text{ and } A_2 \lesssim A_1$$

We recall also the following definition of interpolation spaces. Let $(X_0, \|\cdot\|_0)$, $(X_1, \|\cdot\|_1)$ two Banach spaces contained continuously in a Hausdorff topological vector space (that is (X_0, X_1) is a compatible couple). For $g \in X_0 + X_1$, $t > 0$ one defines the so called K functional $K(g,t;X_0,X_1) \doteq K(g,t)$ by setting

$$K(g,t) = \inf_{g=g_0+g_1}(\|g_0\|_0 + t\|g_1\|_1). \tag{2.1}$$

For $0 \leqslant \theta \leqslant 1$, $1 \leqslant p \leqslant +\infty$, $\alpha \in \mathbb{R}$ we shall consider

$$(X_0,X_1)_{\theta,p;\alpha} = \left\{g \in X_0 + X_1,\ \|g\|_{\theta,p;\alpha} = \|t^{-\theta-\frac{1}{p}}(1-\mathrm{Log}\,t)^\alpha K(g,t)\|_{L^p(0,1)} \text{ is finite}\right\}.$$

Here $\|\cdot\|_V$ denotes the norm in a Banach space V. The weighted Lebesgue space $L^p(0,1;\omega)$, $0 < p \leqslant +\infty$ is endowed with the usual norm or quasi norm, where ω is a weight function on $(0,1)$, $L^p_+(0,1,\omega) = \left\{f \in L^p(0,1;\omega),\ f \geqslant 0\right\}$. Our definition of the interpolation space is different from the usual one (see [2, 13]) since we restrict the norms on the interval $(0,1)$.

If we consider ordered couple, i.e. $X_1 \hookrightarrow X_0$ and $\alpha = 0$,

$$(X_0,X_1)_{\theta,p;0} = (X_0,X_1)_{\theta,p}$$

is the interpolation space as it is defined by J. Peetre (see [2, 13, 3]).

$$C_0^2(\overline{\Omega}) = \left\{\varphi : \overline{\Omega} \to \mathbb{R},\ \text{twicely differentiable and vanishing at the boundary}\right\}$$

$$W^1V = \left\{\varphi \in L^1_{loc}(\Omega) : \nabla\varphi \in V^n\right\}.$$

2.1. A few description of $G\Gamma(p,m;w_1,w_2)$

Definition 1 (of a Generalized Gamma space with double weights)**.** Let w_1, w_2 be two weights on $(0,1)$, $m \in [1,+\infty]$, $1 \leqslant p < +\infty$. We assume the following conditions:

c1) There exists $K_{12} > 0$ such that $w_2(2t) \leqslant K_{12} w_2(t)\ \forall\, t \in (0,1/2)$. The space $L^p(0,1;w_2)$ is continuously embedded in $L^1(0,1)$.

c2) The function $\displaystyle\int_0^t w_2(\sigma)d\sigma$ belongs to $L^{\frac{m}{p}}(0,1;w_1)$.

A generalized Gamma space with double weights is the set

$$G\Gamma(p,m;w_1,w_2) = \left\{ v : \Omega \to \mathbb{R} \text{ measurable } \int_0^t v_*^p(\sigma) w_2(\sigma) d\sigma \text{ is in } L^{\frac{m}{p}}(0,1;w_1) \right\}.$$

A similar definition has been considered in [8]. They were interested in the embeddings between $G\Gamma$-spaces.

Properties. Let $G\Gamma(p,m;w_1,w_2)$ be a Generalized Gamma space with double weights and let us define for $v \in G\Gamma(p,m;w_1,w_2)$

$$\rho(v) = \left[\int_0^1 w_1(t) \left(\int_0^t v_*^p(\sigma) w_2(\sigma) d\sigma \right)^{\frac{m}{p}} dt \right]^{\frac{1}{m}}$$

with the obvious change for $m = +\infty$.
Then,

1. ρ is a quasinorm.
2. $G\Gamma(p,m;w_1,w_2)$ endowed with ρ is a quasi-Banach function space.
3. If $w_2 = 1$
$$G\Gamma(p,m;w_1,1) = G\Gamma(p,m;w_1).$$

Example 1 (of weights)**.** Let $w_1(t) = (1 - \operatorname{Log} t)^{\gamma}$, $\quad w_2(t) = (1 - \operatorname{Log} t)^{\beta}$ wit $(\gamma,\beta) \in \mathbb{R}^2$. Then

$$w_2 \text{ satisfies condition c1) and } w_1 \text{ and } w_2 \text{ are in } L_{exp}^{\max(\gamma;\beta)}(]0,1[)\,.$$

Definition 2 (of the small Lebesgue space)**.** The small Lebesgue space associated to the parameter $p \in]1,+\infty[$ and $\theta > 0$ is the set

$$L^{(p,\theta}(\Omega) = \left\{ f : \Omega \to \mathbb{R} \text{ measurable such that } \right.$$
$$\left. \|f\|_{(p,\theta} = \int_0^1 (1 - \operatorname{Log} t)^{-\frac{\theta}{p}+\theta-1} \left(\int_0^t f_*^p(\sigma) d\sigma \right)^{1/p} \frac{dt}{t} < +\infty \right\}.$$

Let us notice that the small Lebesgue space is a G-gamma space.

Definition 3 (of the Grand Lebesgue space)**.** The associate space of the small Lebesgue space is denoted by $L^{p),\theta}(\Omega)$ for $1 < p < +\infty$, $\theta > 0$ and is defined as

$$L^{p),\theta}(\Omega) = \left\{ f : \Omega \to \mathbb{R} \text{ measurable such that } \|f\|_{p),\theta} = \sup_{0<\varepsilon<p-1} \left(\varepsilon^{\theta} \int_\Omega |f|^{p-\varepsilon} dx \right)^{\frac{1}{p-\varepsilon}} \text{ is finite} \right\}.$$

Properties of small and Grand Lebesgue spaces.

1. They are rearrangement invariant Banach function spaces. One has the following equivalent norm :

$$\|u\|_{L^{(p,\theta}(\Omega)} = \inf_{u=\sum_k u_k} \left\{ \sum_k \inf_{0<\varepsilon<p'-1} \varepsilon^{-\frac{\theta}{(p'-\varepsilon)}} \left(\int_\Omega |u_k|^{(p'-\varepsilon)'} dx \right)^{\frac{1}{(p'-\varepsilon)'}} \right\}$$

$$\|u\|_{p),\theta} \approx \sup_{0<t<|\Omega|} (1 - \operatorname{Log} t)^{-\frac{\theta}{p}} \left(\int_t^{|\Omega|} u_*(s)^p ds \right)^{\frac{1}{p}}.$$

2. $\bigcup_{\varepsilon>0} L^{p+\varepsilon}(\Omega) \overset{\subset}{\neq} \bigcup_{\beta>1} L^p(\operatorname{Log} L)^{\frac{\beta\theta}{p'-1}}(\Omega) \overset{\subset}{\neq} L^{(p,\theta}(\Omega) \subset L^p(\operatorname{Log} L)^{\frac{\theta}{p'-1}}$.

3. $L^p(\Omega) \overset{\subset}{\neq} \dfrac{L^p}{\operatorname{Log}^\theta L}(\Omega) \overset{\subset}{\neq} L^{p),\theta}(\Omega) \overset{\subset}{\neq} \bigcap_{\alpha>1} \dfrac{L^p}{\operatorname{Log}^{\alpha\theta} L}(\Omega) \overset{\subset}{\neq} \bigcap_{0<\varepsilon<p-1} L^{p-\varepsilon}$

4. $\int_\Omega u \cdot v dx \leqslant \|u\|_{L^{(p',\theta}} \|v\|_{L^{p),\theta}}$, $\dfrac{1}{p} + \dfrac{1}{p'} = 1.$

$$VMO(\Omega) = \left\{ f \in L^1(\Omega) : \lim_{R\to 0} \sup_{r<R, x_0\in\Omega} \frac{1}{r^n} \int_{B(x_0,r)\cap\Omega} |f - f_r| dx = 0 \right\}$$

here $f_r = \dfrac{1}{|B(x_0; r) \cap \Omega|} \displaystyle\int_{B(x_0;r)\cap\Omega} f(x)dx.$

§3. Proof of Theorem 3

The proof of Theorem 3 follows the same scheme as in [6] by considering the following dual problem

Lemma 4. *For any $g \in L_+^{n),\theta}(\Omega)$, $V \in L^{n),\theta}(\Omega)$ and $\theta > 0$ the unique solution $\varphi \in H_0^1(\Omega) \cap L^\infty(\Omega)$ of*

$$-\Delta\varphi + U \cdot \nabla\varphi + V\varphi = g \text{ in } H^{-1}(\Omega) \tag{3.1}$$

satisfies $\varphi \in W^2 L^{n),\theta}(\Omega)$ and there exists a constant $c_n > 0$ independent of θ such that

$$\|\varphi\|_{W^2 L^{n),\theta}(\Omega)} \leqslant c_n \|g\|_{L^{n),\theta}(\Omega)}.$$

Here, we assume the same integrability for U as in Theorem 3.

Proof. The existence, uniqueness of φ is given in [4]. Indeed, we have for $n \geqslant 2$,

$$L^{n),\theta}(\Omega) \subset L^{n-\varepsilon}(\Omega), \ \forall\, 0 < \varepsilon < \frac{1}{2}.$$

Thus $g \in L^{\frac{n}{2},1}(\Omega) \subset H^{-1}(\Omega)$.
To obtain the $W^2 L^{n),\theta}$ regularity, we may assume first V and g bounded. Then following Proposition 11 of [4], we have $\varphi \in W^2 L^p(\Omega)$, $p > n$.

Let us show that, we have $\varepsilon_0 > 0$ and a constant $c_0 > 0$ depending only on the data U, V, Ω such that $\forall \varepsilon \in [0, \varepsilon_0]$

$$\|\varphi\|_{W^2 L^{n-\varepsilon}} \leqslant c_0 \|g\|_{L^{n-\varepsilon}}. \tag{3.2}$$

Let $0 < \varepsilon < \dfrac{1}{2}$. Then from the equation satisfied by φ, one has :

$$\|\Delta\varphi\|_{L^{n-\varepsilon}} \leqslant \|U \cdot \nabla \varphi\|_{L^{n-\varepsilon}} + \|V\varphi\|_{L^{n-\varepsilon}} + \|g\|_{L^{n-\varepsilon}}. \tag{3.3}$$

Since $\varphi \in L^\infty(\Omega)$ and

$$\|\varphi\|_{L^\infty} \leqslant c_n \|g\|_{L^{\frac{n}{2},1}} \leqslant c_n \|g\|_{L^{n),\theta}}. \tag{3.4}$$

So that

$$\|V\varphi\|_{L^{n-\varepsilon}} \leqslant c\|V\|_{L^{n-\varepsilon}} \|\varphi\|_{L^\infty} \leqslant c\|V\|_{L^{n-\varepsilon}} \|g\|_{L^{n),\theta}}. \tag{3.5}$$

By Hölder inequality, for $p > n$,

$$\|U \cdot \nabla \varphi\|_{L^{n-\varepsilon}} \leqslant \|U\|_{L^p} \|\nabla \varphi\|_{L^{\frac{p(n-\varepsilon)}{p-n+\varepsilon}}} \leqslant c\|U\|_{L^p} \|\nabla \varphi\|_{L^{p(n)}} \text{ where } p(n) = \frac{pn}{p-n}. \tag{3.6}$$

We shall choose $\varepsilon_0 > 0 : (n - \varepsilon_0)^* > p(n)$ i.e $0 < \varepsilon < \min\Big(\dfrac{1}{2}; \dfrac{n(p-n)}{2p-n}\Big)$. In that case, we have the compact embedding $W^2 L^{n-\varepsilon_0}(\Omega) \subsetneq W^1 L^{p(n)}(\Omega)$. Therefore $\forall\, \eta > 0$, there exists $c_\eta > 0$ such that

$$\|\nabla \varphi\|_{L^{p(n)}} \leqslant \eta \|\varphi\|_{W^2 L^{n-\varepsilon_0}} + c_\eta \|\varphi\|_{L^2}. \tag{3.7}$$

From Agmon-Douglis-Niremberg's theorem and Marcienkiewicz interpolation's theorem, one has a constant $c_n > 0$ such that

$$\|\varphi\|_{W^2 L^{n-\varepsilon}} \leqslant c_n \|\Delta\varphi\|_{L^{n-\varepsilon}} \qquad \forall\, \varphi \in W^2 L^{n-\varepsilon}(\Omega) \cap H_0^1(\Omega) \text{ and } \forall\, \varepsilon \in [0, \varepsilon_0]. \tag{3.8}$$

Combining relations (3.3) to (3.8), we deduce for all $\eta > 0$, one has a constant $c_\eta > 0$, for all $\varepsilon \in [0, \varepsilon_0]$

$$\|\varphi\|_{W^2 L^{n-\varepsilon}} \leqslant \eta \|U\|_{L^p} \|\varphi\|_{W^2 L^{n-\varepsilon}} + c_\eta \|U\|_{L^p} \|\varphi\|_{L^\infty} + c' \|V\|_{L^{n-\varepsilon}} \|g\|_{L^{n),\theta}} + \|g\|_{L^{n-\varepsilon}}. \tag{3.9}$$

Since we have

$$\|g\|_{L^{n),\theta}} \simeq \sup_{0<\varepsilon<\frac{n-1}{2}} \left(\varepsilon^\theta \int_\Omega |g|^{n-\varepsilon}(x)dx\right)^{\frac{1}{n-\varepsilon}};$$

we deduce from relation (3.9) :

$$\|\varphi\|_{W^2 L^{n),\theta}} (1 - \eta \|U\|_{L^p}) \leqslant c_\eta \|U\|_{L^p} \|g\|_{L^{n),\theta}} + c(1 + \|V\|_{L^{n),\theta}}) \|g\|_{L^{n),\theta}}.$$

Choosing $\eta \|U\|_{L^p} \leqslant \dfrac{1}{2}$, we then have a constant c depending only on U and Ω.

$$\|\varphi\|_{W^2 L^{n),\theta}} \leqslant c(1 + \|V\|_{L^{n),\theta}}) \|g\|_{L^{n),\theta}}. \tag{3.10}$$

We conclude by usual density argument, say

$$\text{replacing } g \text{ by } g_k(x) = \min(k; |g(x)|)\text{sign}\,(g(x)),\ V_k = \min(V; k).$$

the solution of φ_k of $\begin{cases} -\Delta\varphi_k + U \cdot \nabla \varphi_k + V_k \varphi = g_k \\ \varphi_k \in H_0^1(\Omega) \cap L^\infty(\Omega) \end{cases}$ satisfies (3.10).

Let $k \to \infty$, the uniqueness of solution (1.3) gives the result. □

§4. Regularity for data in $L(\mathrm{Log}\, L)^\alpha$ for a full linear operator

In [6], we have shown the following

Theorem 5. *Let Ω be a bounded open set of $\mathbb{R}^n$, $n \geqslant 3$ of class $C^{1,1}$, $A(x) = (a_{ij}(x))_{i,j}$, $x \in \Omega$ a bounded coercitive matrix. Assume that $a_{ij} \in VMO(\Omega)$ and let f be in $L(\mathrm{Log}\, L)^\alpha$, $\alpha > \dfrac{n-1}{n}$. Then, the weak solution of*

$$\begin{cases} \mathrm{div}\,(A(x)\nabla u) = f \text{ in } \Omega \\ u \in W_0^1 L^{n'}(\Omega) \end{cases}$$

satisfies

$$\|\nabla u\|_{L^{(n',n\alpha-n+1}} \leqslant c(n;\alpha)\|f\|_{L(\mathrm{Log}\, L)^\alpha}. \tag{4.1}$$

We want to extend the above result replacing the main operator by

$$\mathcal{L}u \doteq -\,\mathrm{div}\,(A(x)\nabla u) + B(x)\cdot\nabla u - \mathrm{div}\,(C(x)u) + V(x)u.$$

For this, we will assume that

H1. $C(x) = (c_i(x))_{i\in\{1,\dots,n\}}$, $B(x) = (b_i(x))$ are such that c_i, b_i are in $L^n(\Omega)$ for all i and $V \in L^{\frac{n}{2}}(\Omega)$, A is symmetric.

H2. There exists a constant $c_0 > 0$: $V - \mathrm{div}\,(C) \geqslant c_0 > 0$ in $\mathcal{D}'(\Omega)$

We recall the following results (see [9]).

Lemma 6. *Under the above assumptions on A, B, C and V, $F \in L^p(\Omega)^n$, $1 < p < n$. There exists an unique solution $u \in W_0^{1,p}(\Omega)$ of*

$$\mathcal{L}u = -\mathrm{div}\,(F) \text{ in } \mathcal{D}'(\Omega).$$

Moreover, there exists a constant $k(p) > 0$ (independent of f and u) such that

$$\|\nabla u\|_{L^{\frac{np}{n-p}}(\Omega)} \leqslant k(p)\|F\|_{L^p(\Omega)^n}. \tag{4.2}$$

Lemma 7 (see [7])**.** *Let $1 < p < n$, $f \in L^p(\Omega)$ and v the unique solution of $-\Delta v = f$ in $\mathcal{D}'(\Omega)$, $v \in W_0^{1,p}(\Omega)$. Then there exist a constant c_n independent of p, f and v such that*

$$\|\nabla v\|_{L^{\frac{np}{n-p}}(\Omega)} \leqslant \frac{c_n}{(p-1)^{\frac{n-1}{n}}}\|f\|_{L^p(\Omega)}. \tag{4.3}$$

Lemma 8. *Let $f \in L^p(\Omega)$, $1 < p < \dfrac{n(n-1)}{n^2-n-1}$, $p^* = \dfrac{pn}{n-p}$.*

Then, there exist a constant c_n' independent of p, f such that the unique solution $u \in W_0^{1,p}(\Omega)$ of $\mathcal{L}u = f$ in $\mathcal{D}'(\Omega)$ satisfies

$$\|\nabla u\|_{L^{p^*}} \leqslant \frac{c_n'}{(p-1)^{\frac{n-1}{n}}}\|f\|_{L^p(\Omega)}. \tag{4.4}$$

Proof. Let $r \in \left[\frac{n}{n-1} = n', \frac{n-1}{n-2} = (n-1)'\right]$ and $v \in W_0^{1,r}(\Omega) : -\Delta v \in L^m(\Omega)$ with $\frac{1}{m} = \frac{1}{r} + \frac{1}{n}$. From Lemma 6 for any solution $u \in W_0^{1,n'}(\Omega)$ of $\mathcal{L}u = -\Delta v$, one has

$$\|\nabla u\|_{L^{n'}(\Omega)} \leqslant k(n')\|\nabla v\|_{L^{n'}(\Omega)}. \tag{4.5}$$

and

$$\|\nabla u\|_{L^{(n-1)'}(\Omega)} \leqslant k((n-1)')\|\nabla v\|_{L^{(n-1)'}(\Omega)}. \tag{4.6}$$

Applying the Marcinkiewicz real interpolation method, we deduce that we have

$$\|\nabla u\|_{L^r(\Omega)} \leqslant \text{Max}\,(k(n'); k(n-1)')\|\nabla v\|_{L^r(\Omega)}. \tag{4.7}$$

Taking $1 < p < \frac{n(n-1)}{n^2-n-1}$, we have $n' < p^* < (n-1)'$ and choosing v such that $-\Delta v = f \in L^p(\Omega)$, $v \in W^{1,p}(\Omega)$ then applying Lemma 7, relation (4.7) leads to :

$$\|\nabla u\|_{L^{p^*}(\Omega)} \leqslant \frac{c_n}{(p-1)^{\frac{n-1}{n}}}\|f\|_{L^p(\Omega)}. \quad \square$$

Theorem 9. *Let $f \in L(\text{Log}\,L)^\alpha$, $\alpha > \frac{n-1}{n}$, u satisfying $\mathcal{L}u = f$ in $\mathcal{D}'(\Omega)$, $u \in W_0^1 L^{n',\infty}(\Omega)$. Then*

1. $|\nabla u| \in L^{(n',n\alpha-n+1}(\Omega)$.
2. $\|\nabla u\|_{L^{(n',n\alpha-n+1}(\Omega)} \leqslant c(n;\alpha)\|f\|_{L(\text{Log}\,L)^\alpha}$.

Proof. Its follows the same arguments as in [6] using relation (4.4) and a suitable decomposition of f, whenever $f \geqslant 0$. We drop the details. $\square$

We may weaken hypothesis H2. on V and $C(x)$ by assuming

H3. $V - \text{div}\,(C) \geqslant 0$ in $\mathcal{D}'(\Omega)$.

But we shall add an assumption as

H4. $V - \frac{1}{2}\text{div}\,(C+B) \geqslant 0$ in $\mathcal{D}'(\Omega)$.

Hypothesis H4. ensures that for all $T \in H^{-1}(\Omega)$ the problem $\mathcal{L}u = T$ in $\mathcal{D}'(\Omega)$ (resp. $\mathcal{L}^*u = T$) possesses an unique solution $u \in H_0^1(\Omega)$, $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$. As a by product of such result and Lemma 6 one has :

Lemma 10. *Let $r \in \left[\frac{2n}{n+2}, \frac{2n}{n-2}\right]$, $n \geqslant 3$, $F \in L^r(\Omega)^n$. Then, there exists an unique $u \in W_0^{1,r}(\Omega)$ of $\mathcal{L}u = -\text{div}\,(F)$ in $\mathcal{D}'(\Omega)$. Moreover,*

$$\exists c(r) > 0 : \|\nabla u\|_{L^r} \leqslant c(r)\|F\|_{L^r}. \tag{4.8}$$

Proof. Let $F \in L^r(\Omega)^n$, $r \in \left[2, \frac{2n}{n-2}\right]$. Since $F \in L^2(\Omega)^n$, we may use hypothesis H4. to deduce that the problem $\mathcal{L}u = -\text{div}\,(F)$ has an unique solution $u \in H_0^1(\Omega)$. Let $F_0 \in L^{2^*}(\Omega)^n$ such $-\text{div}\,(F_0) = u$ and

$$\|F_0\|_{L^{2^*}} \leqslant c\|u\|_{L^2} \leqslant c\|F\|_{L^2} \leqslant c\|F\|_{L^r}. \tag{4.9}$$

We may write the equation $\mathcal{L}u = -\mathrm{div}\,(F)$ as

$$-\mathrm{div}\,(A(x)\nabla u) + B(x)\nabla u - \mathrm{div}\,(C(x)u) + (V+1)u = -\mathrm{div}\,(F_0 + F), \quad F_0 + F \in L^r(\Omega)^n$$

One has $V + 1 - \mathrm{div}\,(C) \geqslant 1 > 0$.
Applying Lemma 6, we deduce that $u \in W_0^{1,r}(\Omega)$ and

$$\|\nabla u\|_{L^r} \leqslant c(r)\|F_0 + F\|_{L^r} \leqslant c(r)\|F\|_{L^r}. \tag{4.10}$$

For $r \in \left[\dfrac{2n}{n+2}, 2\right]$, we argue by duality to conclude that one has an unique function $u \in W_0^{1,r}(\Omega)$ such that $\mathcal{L}u = -\mathrm{div}\,(F)$ in $\mathcal{D}'(\Omega)$

$$\|\nabla u\|_{L^r} \leqslant c(r)\|F\|_{L^r}. \quad \square \tag{4.11}$$

Thank to the above Lemma, we have:

Lemma 11. *Let $r \in [n', (n-1)']$ then there exists a constant $k(n) > 0$*

$$\|\nabla u\|_{L^r} \leqslant k(n)\|F\|_{L^r}$$

whenever u satisfies: $\mathcal{L}u = -\mathrm{div}\,(F)$ in $\mathcal{D}'(\Omega)$.

We conclude as before to derive the following:

Lemma 12. *Let $f \in L^p(\Omega)$, $1 < p < \dfrac{n(n-1)}{n^2-n-1}$, $p^* = \dfrac{pn}{n-p} = -p(n)$. Then the unique solution u of $\mathcal{L}u = f$, $u \in W_0^{1,p}(\Omega)$ satisfies*

$$\|\nabla u\|_{L^{p^*}} \leqslant \frac{c_n}{(p-1)^{\frac{n-1}{n}}}\|f\|_{L^p(\Omega)}.$$

Theorem 13. *Assume H1. H3. and H4. Then for $f \in L(\mathrm{Log}\,L)^\alpha$, $\alpha > \dfrac{n-1}{n}$, $n \geqslant 3$. There exists an unique solution $u \in L^{(n',n\alpha-n+1}(\Omega)$ satisfying $\mathcal{L}u = f$ in $\mathcal{D}'(\Omega)$. Moreover, there exists a constant $c(n;\alpha) > 0$ such that:*

$$\|\nabla u\|_{L^{(n',n\alpha-n+1}(\Omega)} \leqslant c(n;\alpha)\|f\|_{L(\mathrm{Log}\,L)^\alpha}.$$

Proof. The proof follows the same argument as in [6]. $\square$

Recent developments concerning equation (2.1) but with singular potential as Colomb's potential is given in [12].

References

[1] Adams, R. *Sobolev spaces*. Academic Press, 1975.

[2] Bennett, C., and Sharpley, R. *Interpolation of Operators*. Academic Press, 1988.

[3] Bergh, J., and Lofstrom, J. *Interpolation spaces. An introduction*. Springer-Verlag, Berlin-New-York, 1976.

[4] Díaz, J. I., Gómez-Castro, D., Rakotoson, J. M., and Temam, R. Linear diffusion with singular absorption potential and/or unbounded convective flow: the weighted space approach. *D.C.D.S. 38*, 2 (2018), 509–546.

[5] Fiorenza, A., Formica, M. R., Gogatishvili, A., and Rakotoson, J. M. Some new results related to $G\Gamma$-spaces and interpolation. *submitted for a publication*.

[6] Fiorenza, A., Formica, M. R., and Rakotoson, J. M. Pointwise estimates for $G\Gamma$-functions and applications. *Differential and Integral equations 30*, 11-12 (2017), 809–824.

[7] Fiorenza, A., and Krbec., M. On an optimal decomposition in Zygmund spaces. *Georg. Math. J. 9*, 2 (2002), 271–286.

[8] Gogatishvili, A., Krepela, M., Pick, L., and Soudsky, F. Embeddings of Lorentz-type spaces involving weighted integral means. *J. Funct. Anal 273*, 9 (2017), 2939–2980.

[9] Ragusa, A. Elliptic boundary value problem in vanishing mean oscillations hypothesis. *Comment. Math. Univ Carolina 40*, 4 (1999), 651–663.

[10] Rakotoson, J. M. New hardy inequalities and behaviour of linear elliptic equations. *Journal of Functional Analysis 263* (2012), 2893–2920.

[11] Rakotoson, J. M. A sufficient condition for a bow-up in the space of absolutely continuous functions for very weak solution. *Applied Math. Optim. 73*, 1 (2016), 153–163.

[12] Rakotoson, J. M. Potential-capacity and some applications. *Asymptotic Analysis 1-28. DOI:10.3233/ASY-191523, see also arxiv 182.04061V1.* (2019).

[13] Tartar, L. *An introduction to Sobolev spaces and Interpolation spaces*. Springer-Verlag Berlin, 2007.

J. M. Rakotoson
Laboratoire de Mathématiques et Applications - Université de Poitiers,
Avenue Marie et Pierre Curie,Téléport 2,
BP 30179, 86692 Futuroscope Chasseneuil Cedex, France
rako@math.univ-poitiers.fr, jean.michel.rakotoson@univ-poitiers.fr

Monografías Matemáticas *García de Galdeano* **42**, 253–263 (2019)

RENORMALIZED SOLUTIONS FOR A STOCHASTIC p-LAPLACE EQUATION WITH L^1 INITIAL DATA

Niklas Sapountzoglou and Aleksandra Zimmermann

Abstract. For $1 < p < \infty$, we consider a stochastic p-Laplace equation on a bounded domain with homogeneous Dirichlet boundary conditions. The technical difficulties arise from the L^1 random initial data under consideration. We introduce the notion of renormalized solutions.

Keywords: Renormalized solutions, stochastic forcing, L^1 random initial data.
AMS classification: 35K92, 35K55, 60H15.

§1. Introduction

Let $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t\in[0,T]}, (\beta_t)_{t\in[0,T]})$ be a stochastic basis with a complete, countably generated probability space $(\Omega, \mathcal{F}, P)$, a filtration $(\mathcal{F}_t)_{t\in[0,T]} \subset \mathcal{F}$ satisfying the usual assumptions and a real valued, $\mathcal{F}_t$-Brownian motion $(\beta_t)_{t\in[0,T]}$. Let $D \subset \mathbb{R}^d$ a bounded Lipschitz domain, $T > 0$, $Q_T = (0,T) \times D$ and $1 < p < \infty$. Furthermore, let $u_0 : \Omega \to L^1(D)$ be $\mathcal{F}_0$-measurable and $\Phi \in L^2(\Omega \times Q_T)$ be progressively measurable. In this contribution, we study the nonlinear evolution problem:

$$\begin{aligned} du - \operatorname{div}(|\nabla u|^{p-2}\nabla u)\, dt &= \Phi\, d\beta && \text{in } \Omega \times Q_T, \\ u &= 0 && \text{on } \Omega \times (0,T) \times \partial D, \\ u(0,\cdot) &= u_0 && \in L^1(\Omega \times D). \end{aligned} \tag{1.1}$$

The diffusion operator in our equation is the p-Laplace operator for $1 < p < \infty$, i.e.,

$$\Delta_p(u) := \operatorname{div}(|\nabla u|^{p-2}\nabla u).$$

Obviously, $\Delta_2 = \Delta$, while Δ_p is a nonlinear monotone operator for $p \neq 2$. In the last decades, there has been an extensive study on (1) (see, e.g., [16], [15], [17], [14] and [4]). In our case, the main technical difficulty arises from the random initial data in $L^1(\Omega \times D)$. In this setting, variational solutions are out of range and therefore we consider the more general notion of renormalized solutions which has been introduced by [11] for the study of global existence and weak stability of the Boltzmann equation. Renormalized solutions of (1) with a deterministic right hand side have been studied by many authors, (see, e.g., [7], [5], [8]). Later, this solution concept has been extended to more general problems of parabolic, elliptic-parabolic and hyperbolic type (see, e.g., [9],[10], [6], [1]). For stochastic conservation laws the notion of entropy solutions has been considered in [3]. For a quasilinear, degenerate hyperbolic-parabolic SPDE with L^1 random initial data, the well-posedness and regularity of kinetic

solutions has been studied in [13], but, to the best of our knowledge, these results do not apply in the situation of (1). Our aim is to extend the notion of renormalized solutions for the stochastic setting. The well-posedness of (1.1) in the framework of renormalized solutions is the subject of a forthcoming research article.

The well-posedness for $\mathcal{F}_0$-measurable initial data $u_0 \in L^2(\Omega \times D)$ is an easy consequence of classical well-posedness results:

Theorem 1. *Let the conditions in the introduction be satisfied. Furthermore, let $u_0 \in L^2(\Omega \times D)$. Then there exists a unique strong solution to (1.1), i.e., there is an $\mathcal{F}_t$-adapted stochastic process $u : \Omega \times [0,T] \to W_0^{1,p}(D)$ such that $u \in L^p(\Omega; L^p(0,T; W_0^{1,p}(D))) \cap L^2(\Omega; C([0,T]; L^2(D)))$, $u(0,\cdot) = u_0$ in $L^2(\Omega \times D)$ and*

$$u(t) - u_0 - \int_0^t \operatorname{div}(|\nabla u|^{p-2}\nabla u)\, ds = \int_0^t \Phi\, d\beta$$

in $W^{-1,p'}(D) + L^2(D)$ for all $t \in [0,T]$ and a.s. in Ω.

Remark 1. Since we know from all terms except the term $\int_0^t \operatorname{div}(|\nabla u|^{p-2}\nabla u)\, ds$ that these terms are elements of $L^2(D)$ for all $t \in [0,T]$ and a.s. in Ω it follows that $\int_0^t \operatorname{div}(|\nabla u|^{p-2}\nabla u)\, ds \in L^2(D)$ for all $t \in [0,T]$ and a.s. in Ω. Therefore this equation is an equation in $L^2(D)$.

Proof. This result is a consequence of [14], Chapter II, Theorem 2.1 and Corollary 2.1. We only have to check the assumptions of this theorem. Following the notations therein, we set $V = W_0^{1,p}(D) \cap L^2(D)$ in the case $1 < p < 2$ and $V = W_0^{1,p}(D)$ in the case $p \geq 2$, $H = L^2(D)$, $E = \mathbb{R}$, $A : V \to V^*$, $A(u) = -\operatorname{div}(|\nabla u|^{p-2}\nabla u)$, $B = \Phi$, $f(t,\omega) = 2 + \|B(t,\omega)\|_2^2$ for almost each $(t,\omega) \in (0,T) \times \Omega$ and $z = 0$. Then we have $\mathcal{L}_Q(E;H) = \mathcal{L}_2(\mathbb{R}, L^2(D)) = L^2(D)$.
We remark that A does not depend on $(t,\omega) \in [0,T] \times \Omega$ and that B does not depend on $u \in V$. Obviously, conditions (A1), (A2) and (A5) in [14] are satisfied. Moreover, in the case $p \geq 2$ the validity of conditions (A3) and (A4) is well known in the theory of monotone operators. Therefore we only consider the case $1 < p < 2$.
In this case we check condition (A3). Using the norms

$$\|v\|_V := \left(\|v\|^p_{W_0^{1,p}(D)} + \|v\|_2^p\right)^{\frac{1}{p}},\ \|v\|_{W_0^{1,p}(D)} := \|\nabla v\|_{L^p(D)^d}$$

we have

$$\begin{aligned} |B|_Q^2 + 2\|v\|_V^p &= \|B\|_2^2 + 2\|v\|_V^p = f - 2 + 2\|v\|_V^p \\ &= f - 2 + 2\|v\|^p_{W_0^{1,p}(D)} + 2\|v\|_2^p = f - 2 + 2\|v\|_2^p + 2\langle Av, v\rangle_{V^*,V} \\ &\leq f + \|v\|_2^2 + 2\langle Av, v\rangle_{V^*,V} \end{aligned}$$

for all $v \in V$ since $x^p \leq 1 + x^2$ for all $x \geq 0$. This proves condition (A3) for $\alpha = K = 2$.
Now we check condition (A4). We estimate

$$\|A(u)\|_{V^*} \leq \|A(u)\|_{W^{-1,p'}(D)} \leq \|\nabla u\|^{p-1}_{L^p(D)^d} \leq \|u\|_V^{p-1}.$$

Therefore [14], Chapter II, Theorem 2.1, Corollary 2.1 and Theorem 2.2 provide the existence of a strong solution to (1.1). □

§2. Itô formula and renormalization

For two Banach spaces X, Y, let $L(X;Y)$ denote the Banach space of bounded, linear operators from X to Y.
In order to find an appropriate notion of renormalized solutions to (1.1), we use the methods of [12] to prove a particular version of the Itô formula. For the sake of completeness, we recall the following regularization procedure:

Lemma 2. *Let $D \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary, $1 \le p < \infty$ and $r = \min\{p, 2\}$. There exists a sequence of operators*

$$\Pi_n : W^{-1,p'}(D) + L^r(D) \to W_0^{1,p}(D) \cap L^2(D),\ n \in \mathbb{N}$$

such that

i.) $\Pi_n(v) \in W_0^{1,p}(D) \cap L^2(D) \cap C^\infty(\overline{D})$ *for all* $v \in W^{-1,p'}(D) + L^r(D)$ *and all* $n \in \mathbb{N}$

ii.) For any $n \in \mathbb{N}$ *and any Banach space*

$$F \in \{W_0^{1,p}(D),\ L^2(D),\ W^{-1,p'}(D),\ W_0^{1,p}(D) \cap L^2(D),\ W^{-1,p'}(D) + L^2(D)\}$$

$\Pi_n : F \to F$ *is a bounded linear operator such that* $\lim_{n\to\infty} \Pi_{n|F} = I_F$ *in* $L(F;F)$, *where* I_F *is the identity on* F.

Proof. We follow the ideas of [12], p. 200, Exemple 2.1 and let $\Pi_n(v) := (\phi_n \cdot v) * \rho_n$ be the convolution of the multiplication of $v \in W^{-1,p'}(D) + L^r(D)$ with an appropriate cutoff function ϕ_n and a standard mollifier ρ_n with support in $B_{1/n}(0)$ for $n \in \mathbb{N}$. Then, the assertion follows using Hardy and Young inequality. □

Proposition 3. *Let $G \in L^{p'}(\Omega \times Q_T)^d$, $\Phi \in L^2(\Omega \times Q_T)$ be progressively measurable, $u_0 \in L^2(\Omega \times D)$ be $\mathcal{F}_0$-measurable and $u \in L^2(\Omega; C([0,T]; L^2(D))) \cap L^p(\Omega; L^p(0,T; W_0^{1,p}(D)))$ satisfying the equality*

$$u(t) - u_0 - \int_0^t \operatorname{div} G\, ds = \int_0^t \Phi\, d\beta \tag{2.1}$$

in $L^2(D)$ for all $t \in [0,T]$ and a.s. in Ω.
Then, for all $\psi \in C^\infty([0,T]\times\overline{D})$ and all $S \in C^2(\mathbb{R})$ with $\operatorname{supp}(S'')$ compact such that $S'(0) = 0$ or $\psi(t,x) = 0$ for all $(t,x) \in [0,T] \times \partial D$ we have

$$\begin{aligned}&\int_D S(u(t))\psi(t) - S(u_0)\psi(0)\,dx + \int_0^t \int_D S''(u)\nabla u G\psi\,dx\,ds + \int_0^t \int_D S'(u)G\nabla\psi\,dx\,ds\\ &= \int_0^t \int_D S'(u)\psi\Phi\,dx\,d\beta + \int_0^t \int_D S(u)\psi_t\,dx\,ds + \frac{1}{2}\int_0^t \int_D S''(u)\psi\Phi^2\,dx\,ds\end{aligned} \tag{2.2}$$

for all $t \in [0,T]$ and a.s. in Ω.
Especially for $\psi \in C^\infty(\overline{D})$ not depending on t we get

$$\begin{aligned}&\int_D (S(u(t)) - S(u_0))\psi\,dx + \int_0^t \int_D S''(u)\nabla u G\psi\,dx\,ds + \int_0^t \int_D S'(u)G\nabla\psi\,dx\,ds\\ &= \int_0^t \int_D S'(u)\psi\Phi\,dx\,d\beta + \frac{1}{2}\int_0^t \int_D S''(u)\psi\Phi^2\,dx\,ds\end{aligned}$$

for all $t \in [0,T]$ and a.s. in Ω.

Proof. We choose the regularizing sequence (Π_n) according to Lemma 2 and set $u_n := \Pi_n(u)$, $u_0^n := \Pi_n(u_0)$, $(\operatorname{div} G)_n := \Pi_n(\operatorname{div} G)$ and $\Phi_n := \Pi_n(\Phi)$. We apply the operator Π_n to both sides of this equality. Since $\Pi_n \in L(W^{-1,p'}(D) + L^2(D); W_0^{1,p}(D) \cap L^2(D))$, we may conclude

$$u_n(t) - u_0^n - \int_0^t (\operatorname{div} G)_n \, ds = \int_0^t \Phi_n \, d\beta$$

in D, for all $t \in [0,T]$ and a.s. in Ω. Now we apply pointwise in $x \in D$ the classic Itô formula for $h(t,u) := S(u)\psi(t,x)$ with respect to the time variable t. Integration over D afterwards yields

$$\begin{aligned}&\int_D S(u_n(t))\psi(t) - S(u_0^n)\psi(0)\,dx - \int_0^t \langle (\operatorname{div} G)_n, S'(u_n)\psi\rangle_{W^{-1,p'}(D),W_0^{1,p}(D)}\,ds\\ &= \int_0^t \int_D S'(u_n)\psi\Phi_n\,dx\,d\beta + \int_0^t\int_D S(u_n)\psi_t\,dx\,ds + \frac{1}{2}\int_0^t\int_D S''(u_n)\psi\Phi_n^2\,dx\,ds\end{aligned}$$

for all $t \in [0,T]$ and a.s. in Ω. Again by [12] we may pass to the limit with $n \to \infty$. Thus, we get

$$\begin{aligned}&\int_D S(u(t))\psi(t) - S(u_0)\psi(0)\,dx - \int_0^t \langle \operatorname{div} G, S'(u)\psi\rangle_{W^{-1,p'}(D),W_0^{1,p}(D)}\,ds\\ &= \int_0^t \int_D S'(u)\psi\Phi\,dx\,d\beta + \int_0^t\int_D S(u)\psi_t\,dx\,ds + \frac{1}{2}\int_0^t\int_D S''(u)\psi\Phi^2\,dx\,ds\end{aligned}$$

for all $t \in [0,T]$ and a.s. in Ω. This concludes the equality

$$\begin{aligned}&\int_D S(u(t))\psi(t) - S(u_0)\psi(0)\,dx + \int_0^t\int_D S''(u)\nabla u G\psi\,dx\,ds + \int_0^t\int_D S'(u)G\nabla\psi\,dx\,ds\\ &= \int_0^t \int_D S'(u)\psi\Phi\,dx\,d\beta + \int_0^t\int_D S(u)\psi_t\,dx\,ds + \frac{1}{2}\int_0^t\int_D S''(u)\psi\Phi^2\,dx\,ds\end{aligned}$$

for all $t \in [0,T]$ and a.s. in Ω. □

§3. Renormalized solution

Let us assume that there exists a strong solution u to (1.1) in the sense of Theorem 1. We observe that for initial data u_0 merely in L^1, the Itô formula for the square of the norm (see, e.g., [15]) can not be applied and consequently the natural a priori estimate for ∇u in $L^p(\Omega \times Q_T)^d$ is not available. Choosing $\psi \equiv 1$ and

$$S(u) = \int_0^u T_k(r)\,dr$$

in (2.2), where $T_k : \mathbb{R} \to \mathbb{R}$ is the truncation function at level $k > 0$ defined by

$$T_k(r) = \begin{cases} r & , \ |r| \le k,\\ k\operatorname{sign}(r) & , \ |r| > k,\end{cases}$$

we find that there exists a constant $C(k) \geq 0$ depending on the truncation level $k > 0$, such that

$$\mathbb{E} \int_0^T \int_D |\nabla T_k(u)|^p \, dx \, ds \leq C(k).$$

As in the deterministic case, the notion of renormalized solutions takes this information into account :

Definition 1. Let the assumptions in the introduction be fulfilled with $u_0 \in L^1(\Omega \times D)$. Then $u \in L^1(\Omega; C([0,T]; L^1(D)))$ is called a renormalized solution to (1.1) with initial value u_0, if and only if

(i) $T_k(u) \in L^p(\Omega; L^p(0,T; W_0^{1,p}(D)))$ for all $k > 0$.

(ii) For all $\psi \in C^\infty([0,T] \times \bar{D})$ and all $S \in C^2(\mathbb{R})$ such that S' has compact support with $S'(0) = 0$ or $\psi(t,x) = 0$ for all $(t,x) \in [0,T] \times \partial D$ the equality

$$\begin{aligned} &\int_D S(u(t))\psi(t) - S(u_0)\psi(0) \, dx + \int_0^t \int_D S''(u)|\nabla u|^p \psi \, dx \, ds \\ &+ \int_0^t \int_D S'(u)|\nabla u|^{p-2}\nabla u \cdot \nabla \psi \, dx \, ds \\ &= \int_0^t \int_D S'(u)\psi\Phi \, dx \, d\beta + \int_0^t \int_D S(u)\psi_t \, dx \, ds + \frac{1}{2}\int_0^t \int_D S''(u)\psi\Phi^2 \, dx \, ds \end{aligned} \tag{3.1}$$

holds true for all $t \in [0,T]$ and a.s. in Ω.

(iii) The following energy dissipation condition holds true:

$$\lim_{k\to\infty} \mathbb{E} \int_{\{k<|u|<k+1\}} |\nabla u|^p \, dx \, dt = 0.$$

Several remarks about Definition 1 are in order: Let u be a renormalized solution in the sense of Definition 1. Since $\operatorname{supp}(S') \subset [-M, M]$, it follows that S is constant outside $[-M, M]$ and for all $k \geq M$, $S(u(t)) = S(T_k(u(t)))$ a.s. in $\Omega \times D$ for all $t \in [0,T]$. In particular, we have

$$S(u) \in L^p(\Omega; L^p(0,T; W^{1,p}(D))) \cap L^\infty(\Omega \times Q_T).$$

From the chain rule for Sobolev functions it follows that

$$S'(u)(|\nabla u|^{p-2}\nabla u) = S'(u)\chi_{\{|u|<M\}}(|\nabla u|^{p-2}\nabla u) = S'(T_M(u))(|\nabla T_M(u)|^{p-2}\nabla T_M(u)) \tag{3.2}$$

a.s. in $\Omega \times Q_T$ and therefore from (*i*) it follows that all the terms in (3.1) are well-defined. In general, for the renormalized solution u, ∇u may not be in $L^p(\Omega \times Q_T)^d$ and therefore (*iii*) is an additional condition which can not be derived from (*ii*). However, for $u \in L^1(\Omega \times Q_T)$ satisfying (*i*), we can define a generalized gradient (still denoted by ∇u) by setting

$$\nabla u(\omega, t, x) := \nabla T_k(u)\chi_{\{|u|<k\}}$$

a.s. in $\Omega \times Q_T$. From (*ii*) it follows that u satisfies the equation

$$\begin{aligned} &S(u(t)) - S(u(0)) - \int_0^t \operatorname{div}(S'(u)|\nabla u|^{p-2}\nabla u) \, ds \\ &= -\int_0^t S''(u)|\nabla u|^p \, ds + \int_0^t \Phi S'(u) \, d\beta + \frac{1}{2}\int_0^t S''(u)\Phi^2 \, ds, \end{aligned} \tag{3.3}$$

or equivalently the SPDE

$$\begin{aligned} &dS(u) - \operatorname{div}(S'(u)|\nabla u|^{p-2}\nabla u)\,dt + S''(u)|\nabla u|^p\,dt \\ &= \Phi S'(u)\,d\beta + \frac{1}{2}S''(u)\Phi^2\,dt \end{aligned} \tag{3.4}$$

in $L^1(D)$ for all $t \in [0,T]$, a.s. in Ω and for any $S \in C^2(\mathbb{R})$ such that $S'(0) = 0$ with $\operatorname{supp}(S')$ compact.

Remark 2. Let u be a renormalized solution to (1.1) with $\nabla u \in L^p(\Omega \times Q_T)^d$. For fixed $l > 0$, let $h_l : \mathbb{R} \to \mathbb{R}$ be defined by

$$h_l(r) = \begin{cases} 0 & , \ |r| \geq l+1 \\ l+1-|r| & , \ l < |r| < l+1 \\ 1 & , \ |r| \leq l. \end{cases}$$

Taking $S(u) = \int_0^u h_l(r)\,dr$ as a test function in (3.5), we may pass to the limit with $l \to \infty$ and we find that u is a strong solution to (1.1).

3.1. The Itô product rule

In the well-posedness theory of renormalized solutions in the deterministic setting (see, e.g., [7]), the product rule is a crucial part. In the following Lemma, we propose an Itô product rule for strong solutions to (1.1). In the following, we will call a function $f : \mathbb{R} \to \mathbb{R}$ *piecewise continuous*, iff it is continuous except for finitely many points.

Proposition 4. *For $1 < p < \infty$, $u_0, v_0 \in L^2(\Omega \times D)$ $\mathcal{F}_0$-measurable let u be a strong solution to* (1.1) *with initial datum u_0 and v be a strong solution to* (1.1) *with initial datum v_0 respectively. Then, for any $H \in C_b^2(\mathbb{R})$ and any $Z \in W^{2,\infty}(\mathbb{R})$ with Z'' piecewise continuous such that $Z(0) = Z'(0) = 0$*

$$\begin{aligned} &(Z((u-v)(t)), H(u(t)))_2 = (Z(u_0 - v_0), H(u_0))_2 \\ &+ \int_0^t \langle \Delta_p(u) - \Delta_p(v), H(u)Z'(u-v)\rangle_{W^{-1,p'}(D), W_0^{1,p}(D)}\,ds \\ &+ \int_0^t \langle \Delta_p(u), H'(u)Z(u-v)\rangle_{W^{-1,p'}(D), W_0^{1,p}(D)}\,ds + \int_0^t (\Phi H'(u), Z(u-v))_2\,d\beta \\ &+ \frac{1}{2}\int_0^t \int_D \Phi^2 H''(u)Z(u-v)\,dx\,ds \end{aligned} \tag{3.5}$$

for all $t \in [0,T]$ a.s. in Ω.

Proof. We fix $t \in [0,T]$. Since u, v are strong solutions to (1.1), it follows that

$$u(t) = u_0 + \int_0^t \Delta_p(u)\,ds + \int_0^t \Phi\,d\beta, \tag{3.6}$$

$$v(t) = v_0 + \int_0^t \Delta_p(v)\,ds + \int_0^t \Phi\,d\beta$$

and consequently

$$(u - v)(t) = u_0 - v_0 + \int_0^t \Delta_p(u) - \Delta_p(v)\,ds \tag{3.7}$$

holds in $L^2(D)$, a.s. in Ω. For $n \in \mathbb{N}$ we define Π_n according to Lemma 2 and set $\Phi_n := \Pi_n(\Phi)$, $u_0^n := \Pi_n(u_0)$, $v_0^n := \Pi_n(v_0)$, $u_n := \Pi_n(u)$, $v_n := \Pi_n(v)$, $g_n := \Pi_n(\Delta_p(u))$, $h_n := \Pi_n(\Delta_p(v))$. Applying Π_n on both sides of (3.7) yields

$$(u_n - v_n)(t) = u_0^n - v_0^n + \int_0^t g_n - h_n\,ds \tag{3.8}$$

and applying Π_n on both sides of (3.6) yields

$$u_n(t) = u_0^n + \int_0^t g_n\,ds + \int_0^t \Phi_n\,d\beta \tag{3.9}$$

in $W_0^{1,p}(D) \cap L^2(D) \cap C^\infty(\overline{D})$ a.s. in Ω. The pointwise Itô formula in (3.8) and (3.9) leads to

$$Z(u_n - v_n)(t) = Z(u_0^n - v_0^n) + \int_0^t (g_n - h_n)Z'(u_n - v_n)\,ds \tag{3.10}$$

and

$$H(u_n)(t) = H(u_0^n) + \int_0^t g_n H'(u_n)\,ds + \int_0^t \Phi_n H'(u_n)\,d\beta + \frac{1}{2}\int_0^t \Phi_n^2 H''(u_n)\,ds \tag{3.11}$$

in D, a.s. in Ω. From (3.10), (3.11) and the product rule for Itô processes, which is just and easy application of the two-dimensional classical Itô formula (see, e.g., [2], Proposition 8.1, p. 218), applied pointwise in t for fixed $x \in D$ it follows that

$$\begin{aligned} Z(u_n - v_n)(t)H(u_n)(t) &= Z(u_0^n - v_0^n)H(u_0^n) \\ &+ \int_0^t (g_n - h_n)Z'(u_n - v_n)H(u_n)\,ds + \int_0^t g_n H'(u_n)Z(u_n - v_n)\,ds \\ &+ \int_0^t \Phi_n H'(u_n)Z(u_n - v_n)\,d\beta + \frac{1}{2}\int_0^t \Phi_n^2 H''(u_n)Z(u_n - v_n)\,ds \end{aligned} \tag{3.12}$$

in D, a.s. in Ω. Integration over D in (3.12) yields

$$I_1 = I_2 + I_3 + I_4 + I_5 + I_6 \tag{3.13}$$

where

$$
\begin{aligned}
I_1 &= (Z((u_n - v_n)(t)), H((u_n)(t))_2 \\
I_2 &= (Z(u_0^n - v_0^n), H(u_0^n))_2 \\
I_3 &= \int_0^t \int_D (g_n - h_n) Z'(u_n - v_n) H(u_n)\, dx\, ds \\
I_4 &= \int_0^t \int_D g_n H'(u_n) Z(u_n - v_n)\, dx\, ds \\
I_5 &= \int_0^t (\Phi_n H'(u_n), Z(u_n - v_n))_2\, d\beta \\
I_6 &= \frac{1}{2} \int_0^t \int_D \Phi_n^2 H''(u_n) Z(u_n - v_n)\, dx\, ds
\end{aligned}
$$

a.s. in Ω. For any fixed $s \in [0,t]$ and almost every $\omega \in \Omega$, $u_n(\omega, s) \to u(\omega, s)$ and $v_n(\omega, s) \to v(\omega, s)$ for $n \to \infty$ in $L^2(D)$. Since Z, H, H' are continuous and bounded functions, it follows that

$$
\lim_{n\to\infty} I_1 = (Z((u - v)(t)), H'(u(t))_2, \tag{3.14}
$$

$$
\lim_{n\to\infty} I_2 = (Z(u_0 - v_0), H'(u_0))_2 \tag{3.15}
$$

in $L^2(\Omega)$ and a.s. in Ω. Note that

$$
I_3 = \int_0^t \langle (g_n - h_n), Z'(u_n - v_n) H(u_n) \rangle_{W^{-1,p'}(D), W_0^{1,p}(D)}\, ds
$$

a.s. in Ω and from the properties of Π_n it follows that

$$
\lim_{n\to\infty} g_n(\omega, s) - h_n(\omega, s) = \Delta_p(u(\omega, s)) - \Delta_p(v(\omega, s))
$$

in $W^{-1,p'}(D)$ for all $s \in [0,t]$ and a.e. $\omega \in \Omega$. Recalling the convergence result for (Π_n) from Lemma 2, there exists a constant $C_1 \geq 0$ not depending on s, ω and $n \in \mathbb{N}$ such that

$$
\begin{aligned}
\|g_n(\omega, s) - h_n(\omega, s)\|_{W^{-1,p'}(D)} &= \|\Pi_n(\Delta_p(u(\omega, s)) - \Delta_p(v(\omega, s)))\|_{W^{-1,p'}(D)} \\
&\leq C_1 \|\Delta_p(u(\omega, s)) - \Delta_p(v(\omega, s))\|_{W^{-1,p'}(D)}.
\end{aligned}
$$

Since the right-hand side of the above equation is in $L^{p'}(\Omega \times (0,t))$, from Lebesgue's dominated convergence theorem it follows that

$$
\lim_{n\to\infty} g_n - h_n = \Delta_p(u) - \Delta_p(v)
$$

in $L^{p'}(\Omega \times (0,t); W^{-1,p'}(D))$ and, with a similar reasoning, also in $L^{p'}(0,t; W^{-1,p'}(D))$ a.s. in Ω. From the chain rule for Sobolev functions it follows that

$$
\nabla(Z'(u_n - v_n) H(u_n)) = Z''(u_n - v_n) \nabla(u_n - v_n) H(u_n) + Z'(u_n - v_n) H'(u_n) \nabla u_n \tag{3.16}
$$

a.s. in $(0,t)\times\Omega$. Moreover, there exists a constant $C_2 = C_2(\|Z'\|_\infty, \|Z''\|_\infty, \|H\|_\infty, \|H'\|_\infty) \geq 0$ such that

$$\int_0^t \|\nabla(Z'(u_n - v_n)H(u_n))\|_p^p \, ds \leq C_2 \int_0^t (\|\nabla u\|_p^p + \|\nabla v\|_p^p) \, ds \tag{3.17}$$

a.s. in Ω. Consequently, for almost every $\omega \in \Omega$ there exists $\chi(\omega) \in L^p(0,t; W_0^{1,p}(D))$ such that, passing to a not relabeled subsequence that may depend on $\omega \in \Omega$,

$$Z'(u_n - v_n)H(u_n) \rightharpoonup \chi(\omega) \tag{3.18}$$

weakly in $L^p(0,t; W_0^{1,p}(D))$. Since in addition,

$$\lim_{n\to\infty} Z'(u_n - v_n)H(u_n) \to Z'(u-v)H(u)$$

in $L^p((0,t)\times D)$ a.s. in Ω, we get

$$\chi(\omega) = Z'(u-v)H(u) \tag{3.19}$$

in $L^p(0,t; W_0^{1,p}(D))$ a.s. in Ω and the weak convergence in (3.18) holds for the whole sequence. Therefore,

$$Z'(u_n - v_n)H(u_n) \rightharpoonup Z'(u-v)H(u)$$

for $n \to \infty$ weakly in $L^p(0,t; W_0^{1,p}(D))$ for almost every $\omega \in \Omega$. Resuming the above results it follows that

$$\lim_{n\to\infty} I_3 = \int_0^t \langle \Delta_p(u) - \Delta_p(v), Z'(u-v)H(u)\rangle_{W^{-1,p'}(D), W_0^{1,p}(D)} \, ds \tag{3.20}$$

a.s. in Ω. With analogous arguments we get

$$\lim_{n\to\infty} I_4 = \int_0^t \langle \Delta_p(u), H'(u)Z(u-v)\rangle_{W^{-1,p'}(D), W_0^{1,p}(D)} \, ds \tag{3.21}$$

a.s. in Ω. By Itô isometry,

$$\begin{aligned}
&\mathbb{E}\left| \int_0^t \int_D \Phi_n H'(u_n)Z(u_n - v_n) - \Phi H'(u)Z(u-v) \, dx \, d\beta \right|^2 \\
&= \mathbb{E} \int_0^t \int_D |\Phi_n H'(u_n)Z(u_n - v_n) - \Phi H'(u)Z(u-v)|^2 \, dx \, ds.
\end{aligned}$$

From the convergence

$$\Phi_n H'(u_n)Z(u_n - v_n) \to \Phi H'(u)Z(u-v)$$

in $L^2(D)$ for $n \to \infty$ a.s. in $\Omega \times (0,t)$ and since, for almost any (ω, s), there exists a constant $C_3 \geq 0$ not depending on the parameters n, s, ω such that

$$\|\Phi_n(\omega,s)H'(u_n(\omega,s))Z(u_n(\omega,s) - v_n(\omega,s))\|_2 \leq C_3 \|\Phi(\omega,s)\|_2$$

for all $n \in \mathbb{N}$, a.s. in $\Omega \times (0,t)$, it follows that

$$\lim_{n\to\infty} \Phi_n H'(u_n) Z(u_n - v_n) = \Phi H'(u) Z(u - v)$$

in $L^2(\Omega \times (0,t) \times D)$ and consequently

$$\lim_{n\to\infty} I_5 = \int_0^t \int_D \Phi H'(u) Z(u - v)\, dx\, d\beta \tag{3.22}$$

in $L^2(\Omega)$ and, passing to a subsequence if necessary, also a.s. in Ω. According to the properties of (Π_n), $\Phi_n^2 \to \Phi^2$ in $L^1((0,t) \times D)$ for $n \to \infty$ a.s. in Ω. From the boundedness and the continuity of H'' and Z we get

$$\lim_{n\to\infty} H''(u_n) Z(u_n - v_n) = H''(u) Z(u - v)$$

in $L^q((0,t) \times D)$ for all $1 \leq q < \infty$ and weak-$*$ in $L^\infty((0,t) \times D)$ a.s. in Ω, thus it follows that

$$\lim_{n\to\infty} I_6 = \frac{1}{2} \int_0^t \int_D \Phi^2 H''(u) Z(u - v)\, dx\, ds \tag{3.23}$$

a.s. in Ω. Passing to a subsequence if necessary, taking the limit in (3.12) for $n \to \infty$ a.s. in Ω the assertion follows from (3.14)-(3.23). □

Corollary 5. *Proposition 4 still holds true for $H \in W^{2,\infty}(\mathbb{R})$ such that H'' is piecewise continuous.*

Proof. There exists an approximating sequence $(H_\delta)_{\delta>0} \subset C_b^2(\mathbb{R})$ such that $\|H_\delta\|_\infty \leq \|H\|_\infty$, $\|H_\delta'\|_\infty \leq \|H'\|_\infty$, $\|H_\delta''\|_\infty \leq \|H''\|_\infty$ for all $\delta > 0$ and $H_\delta \to H$, $H_\delta' \to H'$ uniformly on compact subsets, $H_\delta'' \to H''$ pointwise in $\mathbb{R}$ for $\delta \to 0$. With this convergence we are able to pass to the limit with $\delta \to 0$ in (3.5). □

References

[1] Ammar, K., and Wittbold, P. Existence of renormalized solutions of degenerate elliptic-parabolic problems. *Proceedings of the Royal Society of Edinburgh Section A 133*, 3 (2003), 477–496.

[2] Baldi, P. *Stochastic Calculus. An Introduction Through Theory and Exercises*. Universitext. 2017. Springer.

[3] Bauzet, C., Vallet, G., and Wittbold, P. The Cauchy problem for conservation laws with a multiplicative stochastic perturbation. *Journal of Hyperbolic Differential Equations 9*, 4 (2012), 661–709.

[4] Bauzet, C., Vallet, G., Wittbold, P., and Zimmermann, A. On a p(t,x)-Laplace evolution equation with stochastic force. *Stochastic Partial Differential Equations. Analysis and Computations 1* (2013), 552–570.

[5] Bénilan, P., Boccardo, L., Gallouët, T., Gariepy, R., Pierre, M., and Vázquez, J. An L^1-theory of existence and uniqueness of solutions of nonlinear elliptic equations. *Annali della Scuola Normale Superiore di Pisa. Classe di scienze 22*, 2 (1995), 241–273.

[6] Bénilan, P., Carillo, J., and Wittbold, P. Renormalized entropy solutions of scalar conservation laws. *Annali della Scuola Normale Superiore di Pisa. Classe di scienze 29*, 2 (2000), 313–327.

[7] Blanchard, D. Truncations and monotonicity methods for parabolic equations. *Nonlinear Analysis: Theory, Methods & Applications 21*, 10 (1993), 725–743.

[8] Blanchard, D., and Murat, F. Renormalised solutions of nonlinear parabolic problems with L^1 data: existence and uniqueness. *Proceedings of the Royal Society of Edinburgh Section A 127*, 6 (1997), 1137–1152.

[9] Blanchard, D., Murat, F., and Redwane, H. Existence and Uniqueness of a Renormalised Solution for a Fairly General Class of Nonlinear Parabolic Problems. *Journal of Differential Equations 177* (2001), 331–374.

[10] Carillo, J., and Wittbold, P. Uniqueness of renormalized solutions of degenerate elliptic-parabolic problems. *Journal of Differential Equations 156*, 1 (1999), 93–121.

[11] DiPerna, R., and Lions, P. On the Cauchy problem for Boltzmann equations: global existence and weak stability. *Annals of Mathematics 130*, 2 (1989), 321–366.

[12] Fellah, D., and Pardoux, E. *Une formule d'Itô dans des espaces de Banach, et application*, vol. 31 of *Stochastic Analysis and Related Topics. Progress in Probability*. Körezlioglu, H and Üstünel, A.S., Boston, 1992. Birkhäuser.

[13] Gess, B., and Hofmanová, M. Well-posedness and regularity for quasilinear degenerate parabolic-hyperbolic SPDE. *The Annals of Probability 46*, 5 (2018), 2495–2544.

[14] Krylov, N., and Rozovskii, B. Stochastic evolution equations. *Journal of Soviet mathematics 16:4* (1981), 1233–1277.

[15] Pardoux, E. *Equations aux dérivées partielles stochastiques non linéaires monotones*. University of Paris, 1975. PhD-thesis.

[16] Vallet, G., Wittbold, P., and Zimmermann, A. On a stochastic evolution equation with random growth conditions. *Stochastic Partial Differential Equations. Analysis and Computations 4* (2016), 246–273.

[17] Vallet, G., and Zimmermann, A. Well-posedness for a pseudomonotone evolution problem with multiplicative noise. *Journal of Evolution Equations 19*, 1 (2019), 153–202.

N. Sapountzoglou and A. Zimmermann
Faculty of Mathematics
University of Duisburg-Essen
Thea-Leymann-Str. 9
45127 Essen
Germany
niklas.sapountzoglou@stud.uni-due.de and aleksandra.zimmermann@uni-due.de

Monografías Matemáticas *García de Galdeano* **42**, 265–274 (2019)

THE MATROID STRUCTURE OF VECTORS OF THE MORDELL-WEIL LATTICE AND THE TOPOLOGY OF PLANE QUARTICS AND BITANGENT LINES

Ryutaro Sato and Shinzo Bannai

Abstract. In this paper, we introduce the terminology of matroids into the study of Zariski-pairs related to rational elliptic surfaces, aiming to simplify the presentation and arguments involved. As an application, we provide new examples of Zariski N-ples of relatively low degree. Namely we show that a Zariski 102-ple of degree 18 exists.

Keywords: Elliptic Surfaces, Mordell-Weil lattice, Matroids, Zariski-pairs.

AMS classification: 14J27, 14E20, 05B35.

§1. Introduction

In this paper, we study the embedded topology of plane curves. We are interested in the following situation. Let $C_1, C_2 \subset \mathbb{P}^2$ be plane curves. Then $(\mathbb{P}^2, C_1)$ and $(\mathbb{P}^2, C_2)$ form a Zariski-pair if the following conditions are satisfied

1. There exist tubular neighborhoods $T(C_i)$ of C_i $(i = 1, 2)$ such that the pairs $(T(C_1), C_1)$ and $(T(C_2), C_2)$ are homeomorphic as pairs.
2. The pairs $(\mathbb{P}^2, C_1)$ and $(\mathbb{P}^2, C_2)$ are not homeomorphic as pairs.

The notion of a Zariski-pair was first defined in [1] by E. Artal–Bartolo and has been an object of interest to many mathematicians. The key in studying Zariski pairs is finding a suitable method to distinguish the curves. Many invariants have been used, such as the fundamental groups of the complements $\pi_1(\mathbb{P}^2 \setminus C_i)$, the Alexander polynomials $\Delta_{C_i}(t)$ and the existence/non-existence of certain Galois covers branched along C_i (see [2] for a survey on these topics). More recently, newer types of invariants such as "linking invariants" and "splitting invariants" have been developed in studying reducible plane curves ([3, 7, 12]). However, as the number of irreducible components of C_i increases, these invariants become more increasingly complex, and it becomes hard to grasp the situation clearly. Hence, we are especially interested in formulating a method in order to present the differences in the curves and the classification comprehensively.

An attempt at this was done in [5],[4] where the second author together with colleagues considered invariants of subsets of the set of irreducible components. This approach proved to be effective and was able to produce new examples of Zariski pairs. However the examples produced were relatively simple, maybe too simple, to appreciate the usefulness of the approach fully. In this paper, we introduce the terminology of *matroids* into our setting in order

to make the results more accessible to a wider audience and also to present more complex examples to demonstrate the usefulness of considering subarrangements more fully.

We introduce some notation to explain the kind of arrangements that we will study. Let Q be a smooth quartic curve and $z_o \in Q$ be a general point of Q. It is known that a rational elliptic surface S_{Q,z_0} can be associated to Q and z_o as follows (see [14, 5] for details): Let $\tilde{f}_Q : \tilde{S}_Q \to \mathbb{P}^2$ be the double cover of $\mathbb{P}^2$ branched along Q, and let $\mu : S_Q \to \tilde{S}_Q$ be the canonical resolution of singularities. Also, let Λ_{z_o} be the pencil of lines through z_o. Then the inverse image $\overline{\Lambda}_{z_o}$ of Λ_{z_o} in $\overline{S}_Q$ gives rise to a pencil of curves with genus 1. Next, the base points of $\overline{\Lambda}_{z_o}$ can be resolved by two consecutive blow-ups, whose composition is denoted by $\nu_{z_o} : S_{Q,z_o} \to \overline{S}_Q$. The morphism $\phi_{z_o} : S_{Q,z_o} \to \mathbb{P}^1$ induced by $\overline{\Lambda}_{z_o}$ gives a genus 1 fibration, and the exceptional divisor of the second blow-up in μ_{z_o} gives a section denote by O. Hence, we have an elliptic surface $\phi_{z_o} : S_{Q,z_o} \to \mathbb{P}^1$ associated to Q and z_o. Note that the covering transformation of $\widehat{S}_Q$ induces an involution on S_{Q,z_o} which we will denote by σ.

$$\begin{array}{ccccc} \widehat{S}_Q & \xleftarrow{\mu} & \overline{S}_Q & \xleftarrow{\nu_{z_o}} & S_{Q,z_o} \\ \downarrow{\scriptstyle \widehat{f}_Q} & & \downarrow{\scriptstyle \overline{f}_Q} & & \downarrow{\scriptstyle \phi_{z_o}} \\ \mathbb{P}^2 & \xleftarrow[q]{} & \overline{\mathbb{P}^2} & & \mathbb{P}^1 \end{array}$$

We denote the set of sections of ϕ_{z_o} by $\mathrm{MW}(S_{Q,z_o})$. The sections will be identified with their images and considered as curves on S_{Q,z_o}. It is known that $\mathrm{MW}(S_{Q,z_o})$ can be endowed with an abelian group structure with a pairing $\langle , \rangle : \mathrm{MW}(S_{Q,z_o}) \to \mathbb{Q}$ called the *height pairing* (see [10]). When considering the height pairing, $\mathrm{MW}(S_{Q,z_o})$ is called the Mordell-Weil lattice of S_{Q,z_o}.

Let $f = \widehat{f_Q} \circ \mu \circ \nu_{z_o}$. For a section $s \in \mathrm{MW}(S_{Q,z_o})$, let $C_s = f(s)$, the image of s under f. The curve C_s is a rational curve in $\mathbb{P}^2$ whose local intersection numbers with Q become even. Such curves are called contact curves of Q. Note that $f(s) = f(-s)$ where $-s$ is the negative of s with respect to the group structure of $\mathrm{MW}(S_{Q,z_o})$. The curves C that we will study are reducible curves of the form

$$C = Q + C_{s_1} + \cdots + C_{s_r}$$

for some choice of $s_1, \ldots, s_r \in \mathrm{MW}(S_{Q,z_o})$. The additional data related to $\mathrm{MW}(S_{Q,z_o})$ allows us to distinguish the curves.

Assume for simplicity that $\mathrm{MW}(S_{Q,z_o})$ is torsion free. Let $E_i = \{s_1^i, \ldots, s_r^i\} \subset \mathrm{MW}(S_{Q,z_o})$ $(i = 1, 2)$ be subsets of $\mathrm{MW}(S_{Q,z_o})$ such that $C_{s_j^i} \neq C_{s_k^i}$ for $j \neq k$. We will consider the matroid structure on E_1, E_2 induced by the linear dependence relations in $\mathrm{MW}(S_{Q,z_o}) \otimes \mathbb{Q}$. Let $C_i = Q + C_{s_1^i} + \cdots + C_{s_r^i}$ $(i = 1, 2)$.

Theorem 1. *Under the above settings, if* $\mathrm{MW}(S_{Q,z_o})$ *is torsion free and* E_1, E_2 *have distinct matroid structures, then there exist no homeomorphisms* $h : \mathbb{P}^2 \to \mathbb{P}^2$ *with* $h(C_1) = C_2$ *and* $h(Q) = Q$.

Moreover, if $h(C_1) = C_2$ *implies* $h(Q) = Q$ *necessarily and the combinatorics of* C_1, C_2 *are the same, then* $(\mathbb{P}^2, C_1)$ *and* $(\mathbb{P}^2, C_2)$ *form a Zariski-pair.*

Theorem 1 allows us to distinguish Zariski pairs and Zariski N-ples by simply calculating the matroid structures of the subsets of $\mathrm{MW}(S_{Q,z_o})$. However, to actually construct Zariski

pairs, we need to choose the subsets $\{s_1^i, \ldots, s_r^i\}$ so that they have the same combinatorics, which is a somewhat delicate matter. Fortunately, we were able to use classical results on smooth quartics and bitangent lines, which can be found in [6], to overcome this difficulty.

In the case where Q is a smooth quartic, it is known that $\mathrm{MW}(S_{Q,z_o}) \cong E_7^*$. The E_7^* lattice has 28 pairs of minimal vectors $\pm l_1, \ldots, \pm l_{28}$ of height $\frac{3}{2}$. Furthermore, $L_i = C_{l_i} = C_{-l_i}$ become bitangent lines of Q, and there is a bijection between the set of pairs $\pm l_i$ and the set of bitangent lines L_i. The combinatorics of these bitangent lines are known, as in the following proposition which will be proved later in Section 4.2.

Proposition 2. *For a general smooth quartic Q, its bitangent lines $L_1, \ldots, L_{28}$ and a fixed value $r = 1, \ldots, 28$, the combinatorics of curves of the form*

$$Q + L_{i_1} + \cdots + L_{i_r}$$

are the same for any $\{i_1, \ldots, i_r\} \subset \{1, \ldots, 28\}$. Namely, all L_{i_k} are true bitangents, i.e. they are tangent to Q at two distinct points, and any three of $L_{i_1}, \ldots, L_{i_r}$ are non-concurrent.

For curves C_1, C_2 of the form above, it is immediate that $h(C_1) = C_2$ implies $h(Q) = Q$ necessarily. Now, Proposition 2 together with Theorem 1 gives us the following theorem.

Theorem 3. *Let N_r be the number of distinct matroid structures on subsets of the form $\{l_{i_1}, \ldots, l_{i_r}\}$, where l_{i_k} is a representative of the pair $\pm l_{i_k}$. Then there exists a Zariski N_r-ple of curves having the combinatorics as in Proposition 2.*

At present, we have not been able to calculate the exact value of N_r due to a lack of computer skills of the authors. However, we have a lower bound as follows:

Proposition 4. *For $r = 1, \ldots, 28$, the value of N_r is greater than or equal to n_r given in the following table.*

r	1	2	3	4	5	6	7	8	9	10	11	12	13	14
n_r	1	1	1	2	2	4	6	11	19	37	52	80	95	102
r	15	16	17	18	19	20	21	22	23	24	25	26	27	28
n_r	100	90	70	54	37	23	16	10	5	3	2	1	1	1

We remark that Zariski-pairs involving smooth quartics and its bitangent lines have already been studied by E. Artal-Bartolo and J. Vallès. They gave an example of a pair consisting of a smooth quartic and three bitangent lines. The results were privately communicated to the authors. Also, the second author together with H. Tokunaga and M. Yamamoto have studied the case of four bitangent lines where a Zariski triple exists. Our approach using matroids fails to detect these examples but we think that our work is still worthwhile as it is easy to increase the number of bitangent lines involved and can be applied to non-smooth quartic curves. It also introduces a new point of view that is possibly relatively easier for a wider audience to access and hopefully will connect to other research areas.

The organization of this paper is as follows. In Section 2, we review the basic terminology of matroids and results concerning elliptic surfaces and dihedral covers, which will give the connection between the matroid structure of sections and the topology of the curves. In Section 3, we will prove Theorem 1. In Section 4, we will discuss the case where Q is a smooth quartic and prove Theorem 3 and also give the proof of Proposition 4.

§2. Preliminaries

2.1. Matroids

As will be seen later, the (in)dependence of elements of $\mathrm{MW}(S_{Q,z_o})$ is deeply related to the (non)existence of certain Galois covers of $\mathbb{P}^2$, hence it is important to understand the structure of (in)dependence. Here, Matroid Theory provides a nice framework as it was precisely designed to study generalizations of the notion of linear independence in vector spaces. In this section we briefly review the basic terminology of matroids. We refer to [9] for more details.

There are many different cryptomorphic definitions of Matroids. In our paper, we are interested in the dependence of elements of $\mathrm{MW}(S_{Q,z_o})$, hence we adopt the definition based on *independent sets*. Let E be a finite set and 2^E be the set of subsets of E.

Definition 1. A matroid structure (or simply a matroid) on E is a pair $(E, \mathcal{I})$, where $\mathcal{I} \subset 2^E$ satisfies

1. $\mathcal{I} \neq \emptyset$. (nontriviality)
2. For any $I_1, I_2 \subset E$, if $I_1 \subset I_2$ and $I_2 \in \mathcal{I}$, then $I_1 \in \mathcal{I}$. (descending)
3. For every $I_1, I_2 \in \mathcal{I}$, if $|I_1| < |I_2|$, then there exists $x \in I_2 - I_1$ such that $I_1 \cup \{x\} \in \mathcal{I}$. (augmentation)

Elements of $\mathcal{I}$ will be called *independent sets* and the other subsets will be said to be *dependent*.

Example 1. Let V be a vector space, and $E = \{v_1, \ldots, v_r\} \subset V$. Let $\mathcal{I} = \{I \subset E \mid I \text{ is linearly independent}\}$. Then $\mathcal{I}$ clearly satisfies the conditions (1), (2), (3) in Definition 1. Hence $(E, \mathcal{I})$ is a matroid structure on E.

Definition 2. Let $(E, \mathcal{I})$ be a matroid. A subset $C \subset E$ is called a *circuit* if $C \notin \mathcal{I}$ and all proper subsets of C are independent sets. Moreover, C is a minimal dependent set.

Example 2. Let $V = \mathbb{R}^3$ and $v_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ and $v_4 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$. Let $E = \{v_1, v_2, v_3, v_3\}$ and consider the matroid structure induced by linear independence. Then E itself forms a circuit.

§3. Proof of Theorem 1

In this section, we use the criterion for the existence of dihedral covers given in Section 3.2 to connect the data of matroids of subsets of $\mathrm{MW}(S_{Q,z_o})$ to the data of the embedded topology of the curves in $\mathbb{P}^2$, and prove Theorem 1.

Let $E_i = \{s_1^i, \ldots, s_r^i\} \subset \mathrm{MW}(S_{Q,z_o})$ $(i = 1, 2)$ be subsets of $\mathrm{MW}(S_{Q,z_o})$ such that $C_{s_j^i} \neq C_{s_k^i}$ for $j \neq k$. Consider the matroid structure $(E_i, \mathcal{I}_i)$ on E_i $(i = 1, 2)$ induced by the linear dependence relation in $\mathrm{MW}(S_{Q,z_o}) \otimes \mathbb{Q}$. Let $C_i = Q + C_{s_1^i} + \cdots + C_{s_r^i}$ $(i = 1, 2)$.

Proposition 5. *If there exists a homeomorphism $h : \mathbb{P}^2 \to \mathbb{P}^2$ such that $h(C_1) = C_2$ and $h(Q) = Q$, then $(E_1, \mathcal{I}_1)$ and $(E_2, \mathcal{I}_2)$ are equivalent as matroids.*

Proof. By the assumption that h is a homeomorphism such that $h(C_1) = C_2$ and $h(Q) = Q$, h induces a bijection $\{C_{s_1^1}, \dots, C_{s_r^1}\} \to \{C_{s_1^2}, \dots, C_{s_r^2}\}$ which in turn induces a bijection $h_* : E_1 \to E_2$. Let $I_1 \in \mathcal{I}_1$ be an independent set. Then by Lemma 10, there exists only a finite number of primes such that a D_{2p} cover branched at $2Q + p(\sum_{s\in J_1} C_s)$ for some subset $J_1 \subset I_1$ exists. Since h is a homeomorphism, the same is true for $h_*(I_1)$ which implies that $h_*(I_1) \in \mathcal{I}_2$, by Lemma 9. The converse is also true so we have $I_1 \in \mathcal{I}_1$ if and only if $h_*(I_1) \in \mathcal{I}_2$. Therefore $(E_1, \mathcal{I}_1)$ and $(E_2, \mathcal{I}_2)$ are equivalent as matroids. □

The contrapositive of Proposition 5 gives Theorem 1.

Remark 1. The statement of Proposition 5 concerns the matroid structure over $\mathbb{Q}$. However, from the proof, it is evident that if we consider the matroid structures of the sections in $\mathrm{MW}(S) \otimes \mathbb{Z}/p\mathbb{Z}$ for all p we would be able to distinguish the arrangements in more detail.

Definition 3. Let $(E_1, \mathcal{I}_1)$, $(E_2, \mathcal{I}_2)$ be matroids. The matroids $(E_1, \mathcal{I}_1)$, $(E_2, \mathcal{I}_2)$ are said to be equivalent as matroids if there exists a bijection $\varphi : E_1 \to E_2$ such that $I_1 \in \mathcal{I}_1$ if and only if $\varphi(I_1) \in \mathcal{I}_2$.

3.1. Elliptic surfaces and the Mordell-Weil lattice

In this subsection, we list the basic facts about quartics, rational elliptic surfaces and the Mordell-Weil lattice. We refer the reader to [10], [8] for more details.

In this paper, an *elliptic surface* is a smooth projective surface S, with a relatively minimal genus 1 fibration $\phi : S \to C$ over a smooth projective curve C having a section $O : C \to S$. We identify O with its image in S. We also assume that S has at least one singular fiber. Let $\mathrm{Sing}(\phi) = \{v \in C \mid \phi^{-1}(v) \text{ is singular }\}$. For $v \in \mathrm{Sing}(\phi)$, we put $F_v = \phi^{-1}(v)$ and denote its irreducible decomposition by $F_v = \Theta_{v,0} + \sum_{i=1}^{m_i-1} a_{v,i}\theta_{v,i}$, where $m_{v,i}$ is the number of irreducible components and $\Theta_{v,0}$ is the unique irreducible component with $\Theta_{v,0}.O = 1$. The subset of $\mathrm{Sing}(\phi)$ that corresponds to reducible singular fibers will be denoted by R. Let $\mathrm{MW}(S)$ be the set of sections of $\phi : S \to C$.

The set $\mathrm{MW}(S)$ can be endowed with a group structure as follows. Let E_S be the generic fiber of ϕ and $\mathbb{C}(C)$ be the function field of C. It is known that there is a bijection between $\mathbb{C}(C)$ rational points $E_S(\mathbb{C}(C))$ of E_S and $\mathrm{MW}(S)$. Furthermore, since we have $O \in \mathrm{MW}(S)$, $(E(S), O)$ can be considered as an elliptic curve over $\mathbb{C}(C)$ and has a group structure where O acts as the identity element.

Furthermore, under these circumstances, $\mathrm{MW}(S)$ becomes a finitely generated abelian group with a pairing $\langle , \rangle : \mathrm{MW}(S) \to \mathbb{Q}$ called the height pairing ([10]). The explicit formula to calculate the pairing for $s_1, s_2 \in \mathrm{MW}(S)$ is given by

$$\langle s_1, s_2\rangle = \chi(S) + s_1.O + s_2.O - s_1.s_2 - \sum_{v\in R} \mathrm{contr}_v(s_1, s_2).$$

The formulas for calculating $\mathrm{contr}_v(s_1, s_2)$ can be found in [10]. In the following we will only need the values of $\mathrm{contr}_v(s_1, s_2)$ for singular fibers of type I_2 of the form $F_v = \Theta_{v,0} + \Theta_{v,1}$. In this case we have

$$\mathrm{contr}_v(s_1, s_2) = \begin{cases} 1/2 & (s_1.\Theta_{v,1} = s_2.\Theta_{v,1} = 1) \\ 0 & \text{otherwise} \end{cases}.$$

3.2. Criterion for existence of dihedral covers

Let D_{2n} be the dihedral group of order $2n$. We present a criterion for the existence of certain dihedral covers of $\mathbb{P}^2$ in terms of $\mathrm{MW}(S)$. The existence/non-existence of the dihedral covers will enable us to distinguish the topology of the curves.

Let Q be a quartic plane curve, $z_o \in Q$ be a general point of Q, $s_1, \ldots, s_r \in \mathrm{MW}(S_{Q,z_o})$ be sections such that $C_{s_i} \neq C_{s_j}$, where $C_{s_i} = f(s_i)$ as in the Introduction.

Theorem 6 ([5, Corollary 4])**.** *Let p be an odd prime. Under the above setting, there exists a D_{2p}-cover of $\mathbb{P}^2$ branched at $2Q + p(C_{s_1} + \cdots + C_{s_r})$ if and only if there exists integers $a_i \in \{1, \ldots, p-1\}$ for $i = 1, \ldots r$ such that $\sum_{i=1}^{r} a_i s_i \in p\,\mathrm{MW}(S)$.*

Corollary 7. *If there exists a D_{2p} cover branched at $2Q + p(C_{s_1} + \cdots + C_{s_r})$, then the images of $s_1, \ldots, s_r$ in $\mathrm{MW}(S) \otimes \mathbb{Z}/p\mathbb{Z}$ become linearly dependent.*

Note that the converse of Corollary 7 is not true, as it is necessary for the images of $s_1, \ldots, s_r$ to have a linear dependence relation where all coefficients are non-zero for there to exist a dihedral cover. If there does not exist such linear dependence relation, the branch locus will not be the whole of $2Q + p(C_{s_1} + \cdots + C_{s_r})$. To exclude such cases, the notion of circuits is useful.

Corollary 8. *If the images of $s_1, \ldots, s_r$ in $\mathrm{MW}(S) \otimes \mathbb{Z}/p\mathbb{Z}$ forms a circuit, then there exists a D_{2p}-cover branched at $2Q + p(C_{s_1} + \cdots + C_{s_r})$.*

If $s_1, \ldots, s_r$ form a circuit over $\mathbb{Q}$, then their images in $\mathrm{MW}(S) \otimes \mathbb{Z}/p\mathbb{Z}$ form a circuit for infinitely many prime numbers p. Hence we have:

Lemma 9. *If $s_1, \ldots, s_r$ are linearly dependent, then there are infinitely many prime numbers p such that there exists a D_{2p}-cover branched at $2Q + p(C_{s_{i_1}} + \cdots + C_{s_{i_t}})$ for some nonempty subset $\{i_1, \ldots, i_t\} \subset \{1, \ldots, r\}$.*

On the other hand, if $s_1, \ldots, s_r$ are independent over $\mathbb{Q}$, then they are independent over $\mathbb{Z}/p\mathbb{Z}$ except for a finite number of primes. This implies the following.

Lemma 10. *If $s_1, \ldots, s_r$ are independent over $\mathbb{Q}$, then there are only a finite number of prime numbers p such that there exists a D_{2p}-cover branched at $2Q + p(C_{s_{i_1}} + \cdots + C_{s_{i_t}})$ for some nonempty subset $\{i_1, \ldots, i_t\} \subset \{1, \ldots, r\}$.*

§4. The smooth case

In this section, we will consider the case where Q is a smooth quartic.

4.1. The bitangents of Q and sections of S_{Q,z_o}

We will use the notation given in the Introduction. Let Q be a smooth plane quartic and $z_o \in Q$ be a general point of Q. Since Q is smooth, $\widehat{S}_Q = \overline{S}_Q$. In this case S_{Q,z_o} has only one reducible singular fiber $F_0 = \Theta_{0,0} + \Theta_{0,1}$ of type I_2. The component $\Theta_{0,0}$ is the exceptional divisor of the first blow up of μ_{z_o} in the introduction, and $\Theta_{0,1}$ is the strict transform of the preimage of the tangent line of Q at z_o. All other singular fibers are irreducible. By [8], we

have $\mathrm{MW}(S_{Q,z_o}) \cong E_7^*$ where E_7^* is the dual lattice of the root lattice E_7. It is known that the E_7^* lattice has 56 minimal vectors $\pm l_1, \ldots, \pm l_{28}$ of height $\frac{3}{2}$. It is also well known that Q has 28 bitangent lines $L_1, \ldots, L_{28}$. The correspondence between the 28 pairs of minimal vectors and the 28 bitangent lines is given in [11], but we describe the relation below for the readers convenience.

Lemma 11. *Let $l \in \mathrm{MW}(S_{Q,z_o})$ be a minimal vector of height $\frac{3}{2}$. Then $L = f(l)$ is a bitangent line of Q, where f is the morphism $f : S_{Q,z_o} \to \mathbb{P}^2$ given in the Introduction.*

Proof. By the explicit formula for the height pairing, and since $\chi(S_{Q,z_o}) = 1$ and $l.l = -1$, we have

$$\langle l, l\rangle = 2 + 2l.O - \mathrm{contr}(l, l) = \frac{3}{2}.$$

Where $\mathrm{contr}(l, l)$ is the contribution from the unique reducible singular fiber F_0. Since the possible values of $\mathrm{contr}(l, l) = 0, \frac{1}{2}$, we have $l.O = 0$ and $\mathrm{contr}(l, l) = \frac{1}{2}$ which implies that $l.\Theta_{0,1} = 1$. This implies that l is disjoint with the exceptional set of ν_{z_o}. Also, if we consider the section $-l = \sigma^*(l)$, the preimage of l under the involution σ, we have

$$\langle l, -l\rangle = 1 + l.O + (-l).O - l.(-l) - \mathrm{contr}(l, -l) = -\frac{3}{2}$$

Hence we obtain $l.(-l) = 2$. Let $\widehat{l} = \nu_{z_o}(l)$ and $\widehat{-l} = \nu_{z_o}(-l)$. The above implies that $\widehat{l}.\widehat{-l} = \widehat{l}.\widehat{Q} = 2$, where $\widehat{Q}$ is the ramification locus of $\widehat{f_Q}$. Now since $(\widehat{f_Q})^*(L) = \widehat{l} + \widehat{-l}$ we have $2L.L = (\widehat{l} + \widehat{-l}).(\widehat{l} + \widehat{-l})$. Hence we obtain $L.L = 1$ which implies that L is a line in $\mathbb{P}^2$. Also, the local intersection numbers of L and Q must be even by construction, hence L is a bitangent line. □

Remark 2. Note that the two points of tangency may coincide to give a line L intersecting Q at a single point with multiplicity 4, which we will still consider to be a bitangent line.

Lemma 12. *Let L be a bitangent line of Q and let $f^*(L) = l + l'$. Then l, l' become minimal sections with height $\frac{3}{2}$ and $l' = \sigma^* l = -l$.*

Proof. By following through the proof of Lemma 11 backwards, we have the desired result. □

The above two lemmas give us the following propositon.

Proposition 13. *There is a bijection between the set of 28 bitangent lines of Q and the set of 28 pairs of minimal vectors of the E_7^* lattice.*

4.2. The configuration of bitangents

In this subsection we explain the proof of Proposition 2. In [6], an explicit set of equations for computing the equations of the 28 bitangents, called Riemann's Equations, is given. Using these equations, it is possible to calculate the equations of all 28 bitangents provided that one has the data of seven bitangent lines $L_1, \ldots, L_7$ of Q, which form an Aronhold set (i.e. a septuple of bitangents such that, for any subset of three bitangents the six points of tangency

do not lie on a conic.). We can assume that $L_1, \ldots, L_7$ are given by the following equations for a suitable choice of coordinates where $[t_0, t_1, t_2]$ are homogeneous coordinates of $\mathbb{P}^2$:

$$L_1 = V(t_0), L_2 = V(t_1), L_3 = V(t_2), L_4 = V(t_0 + t_1 + t_2)$$
$$L_{4+i} = V(a_{0i}t_0 + a_{1i}t_1 + a_{2i}t_2), (i = 1, 2, 3)$$

Reimann's Equations gives us the explicit equations of the bitangent lines in terms of the coefficients a_{ij}. It also allows us to recover the equation of Q similarly. Once we have explicit equations it is possible to calculate the combinatorics of the quartic and bitangent lines. We used the open-source mathematics software system SageMath [13] for the actual calculations.

Lemma 14. *Let $L_1, \ldots, L_7$ be lines defined as above. Then, for a general choice of a_{0i}, a_{1i}, a_{2i} $(i = 1, 2, 3)$ the following hold:*

1. *There exists a smooth quartic Q having $L_1, \ldots, L_7$ as an Aronhold set of bitangents.*
2. *Any three bitangent lines of Q are non-concurrent.*
3. *Every bitangent line of Q is a true bitangent, i.e. it is tangent to Q at two distinct points.*

Proof. The equations of Q, and its bitangents $L_1, \ldots, L_{28}$ are given in terms of a_{ij} by Riemann's equation as in [6]. Since all three condition in the statements are closed conditions on a_{0i}, a_{1i}, a_{2i} $(i = 1, 2, 3)$, it is enough to find one example where the statements hold. Almost any choice will serve our purpose. We omit the details of the equations and calculations do to lack of space. □

Lemma 14 immediately implies Proposition 2.

4.3. The proof of Proposition 4

In this subsection, we describe the method we used to distinguish the matroid structures of minimal vectors of the E_7^* lattice in order to calculate n_r. We used SageMath [13] for the actual calculations.

The object that we want to classify are the matroid structures on the sets of the form $\{l_{i_1}, \ldots, l_{i_r}\}$ where l_{i_r} are representatives of pairs $\pm l_{i_r}$ of minimal vectors of height $\frac{3}{2}$. It is known that the E_7^* lattice can be representation in $\mathbb{Q}^8$ in a way so that the minimal vectors are of the form

$$\pm\frac{1}{4}(1, 1, 1, 1, 1, 1, -3, -3)$$

and its permutations. We use this representation in our calculations.

We used an inductive argument on the number of vectors r. For each subset $E \subset \{l_1, \ldots, l_{28}\}$ having r-elements, we assign an $(n_{r-1} + 1)$-ple of integers inductively as follows. The values of n_r will also be determined inductively along the way.

- **Step** (1)

  For every subset with a single element, we assign the pair $\alpha_{1,1} = (1; 1)$.

- **Step** $(k+1)$

  Suppose that every subset having k elements has been assigned an $(n_{k-1}+1)$-ple of integers. We set n_k to be the number of distinct $(n_{k-1}+1)$-ples that have been assigned and label them by $\alpha_{k,1}, \ldots, \alpha_{k,n_k}$. Next, to each subset $E \subset \{l_1, \ldots, l_{28}\}$ having $k+1$ elements, we assign an (n_k+1)-ple as follows:

  (i) Consider the linear dependence/independence of E. Put $i = 0$ if it is dependent and $i = 1$ if it is independent.

  (ii) Let m_j^k be the number of subsets of E of k elements that have the $(n_{k-1}+1)$-ple $\alpha_{k,j}$ assigned.

  (iii) Assign the (n_k+1)-ple $(i; m_1^k, \ldots, m_{n_k}^k)$ to E.

Lemma 15. *Let E_1, E_2 be subsets of $\{l_1, \ldots, l_{28}\}$ and $|E_1| = |E_2| = r$. If E_1 and E_2 have the same matroid structure, then the $(n_{r-1}+1)$-ples of integers assigned above are equivalent.*

Proof. We use induction on r to prove this lemma. The case for $r = 1$ is trivial as every subset having a single element has the same pair assigned and has the same matroid structure.

Assume the statement holds for $r = k$. If $|E_1| = |E_2| = k+1$ and E_1, E_2 have equivalent matroid structure, there exists a bijection $\varphi : E_1 \to E_2$ that preserves independent sets. Hence E_1 is independent if and only if E_2 is independent and the value of i must be equal. Also, φ induces a bijection from $\{E \subset E_1 \mid |E| = k\}$ to $\{E \subset E_2 \mid |E| = k\}$ and an equivalence of matroid structures among the corresponding subsets. Hence the values of m_j^k must be equal do to the hypothesis of induction, and the assigned (n_k+1)-ple are equivalent. □

Lemma 15 and calculations done by computer using SageMath gives Proposition 4.

Acknowledgements

The second author is partially supported by Grant-in-Aid for Scientific Research C (18K03263).

References

[1] Artal Bartolo, E. Sur les couples de Zariski. *J. Algebraic Geom. 3*, 2 (1994), 223–247.

[2] Artal Bartolo, E., Cogolludo, J. I., and Tokunaga, H.-o. A survey on Zariski pairs. In *Algebraic geometry in East Asia—Hanoi 2005*, vol. 50 of *Adv. Stud. Pure Math.* Math. Soc. Japan, Tokyo, 2008, pp. 1–100.

[3] Bannai, S. A note on splitting curves of plane quartics and multi-sections of rational elliptic surfaces. *Topology Appl. 202* (2016), 428–439.

[4] Bannai, S., Guerville-Ballé, B., Shirane, T., and Tokunaga, H.-o. On the topology of arrangements of a cubic and its inflectional tangents. *Proc. Japan Acad. Ser. A Math. Sci. 93*, 6 (2017), 50–53.

[5] Bannai, S., and Tokunaga, H.-o. Geometry of bisections of elliptic surfaces and Zariski N-plets for conic arrangements. *Geom. Dedicata 178* (2015), 219–237.

[6] Dolgachev, I. V. *Classical algebraic geometry*. Cambridge University Press, Cambridge, 2012. A modern view.

[7] Guerville-Ballé, B., and Meilhan, J.-B. A linking invariant for algebraic curves. arXiv:1602.04916.

[8] Oguiso, K., and Shioda, T. The Mordell-Weil lattice of a rational elliptic surface. *Comment. Math. Univ. St. Paul. 40*, 1 (1991), 83–99.

[9] Oxley, J. *Matroid theory*, second ed., vol. 21 of *Oxford Graduate Texts in Mathematics*. Oxford University Press, Oxford, 2011.

[10] Shioda, T. On the Mordell-Weil lattices. *Comment. Math. Univ. St. Paul. 39*, 2 (1990), 211–240.

[11] Shioda, T. Plane quartics and Mordell-Weil lattices of type E_7. *Comment. Math. Univ. St. Paul. 42*, 1 (1993), 61–79.

[12] Shirane, T. A note on splitting numbers for Galois covers and π_1-equivalent Zariski k-plets. *Proc. Amer. Math. Soc. 145*, 3 (2017), 1009–1017.

[13] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 8.1)*. `https://www.sagemath.org`.

[14] Tokunaga, H.-o. Sections of elliptic surfaces and Zariski pairs for conic-line arrangements via dihedral covers. *J. Math. Soc. Japan 66*, 2 (2014), 613–640.

Sato Ryutaro and Shinzo Bannai
National Institute of Technology, Ibaraki College
866 Nakane, Hitachinaka-shi, Ibaraki-Ken 312-8508 JAPAN
sbannai@ge.ibaraki-ct.ac.jp

Monografías Matemáticas *García de Galdeano* **42**, 275–284 (2019)

SPARSE POLYNOMIAL SURROGATES FOR UNCERTAINTY QUANTIFICATION IN COMPUTATIONAL FLUID DYNAMICS

Éric Savin

Abstract. This paper is concerned with the construction of polynomial surrogates of complex models typically arising in computational fluid dynamics for the purpose of propagating uncertainties pertaining to geometrical and/or operational parameters. Polynomial chaos expansions are considered and different techniques for the intrusive and non intrusive reconstruction of the polynomial expansion coefficients are outlined. A sparsity-based reconstruction approach is more particularly emphasized since it benefits from the "sparsity-of-effects" trend commonly observed on global quantities of interest such as the aerodynamic coefficients of a profile.

Keywords: Computational fluid dynamics, polynomial chaos, stochastic Galerkin method, stochastic collocation method, compressed sensing.

AMS classification: 76H05, 74F10, 65C20, 65K05.

§1. Introduction

The polynomial chaos or Wiener-Hermite expansion consists in the decomposition of a second-order random variable in a series of multivariate Hermite polynomials in a countable sequence of independent Gaussian random variables [3, 22]. Specifically, truncations of such an expansion are considered as approximations of random vectors, tensors or fields for the purpose of quantifying uncertainties in complex models. They have been intensively used in recent years in computational methods for solving ordinary or partial differential equations with poorly known, indeed random data or parameterized by Gaussian random parameters. As the solutions of these equations are stochastic processes indexed by spatial and/or time coordinates, which are typically second-order random fields from physical considerations (they have finite energy), polynomial chaos expansions are used to approximate them along the stochastic dimension. Mean-square convergence is guaranteed by the Cameron-Martin theorem [3] and is optimal (*i.e.* exponential) for Gaussian probability measures. For random fields parameterized by non Gaussian, arbitrary random variables the numerical study in [24] has shown that the convergence rates of Hermite polynomial chaos are not optimal. This observation has prompted the development of generalized polynomial chaos expansions involving other families of polynomials which are orthogonal with respect to the probability measures of the random parameters [9, 24].Optimal convergence rates can be achieved once this substitution has been done.

The earlier approach of using polynomial chaos to numerically solve differential equations proposed truncated expansions as trial functions in a Galerkin formulation, resulting

in the spectral stochastic finite element method, or stochastic Galerkin method [20] subsequently developed in [10, 13]. More precisely, the chaos expansions of the sought solutions are substituted in the model equations, which in turn yield evolution equations for their expansion coefficients from Galerkin projections using the orthogonal polynomials as test functions. The stochastic Galerkin method is intrusive in that it may require significant alterations of the existing deterministic codes utilized for solving the differential equations of interest. Generally it also yields coupled equations for the expansion coefficients. This situation has prompted the development of non intrusive approaches, which require only repetitive executions of existing deterministic codes. Stochastic collocation and regression methods [2, 14, 23] have become widely popular in computational fluid dynamics (CFD), for the applicability and precision of these uncertainty quantification techniques is not affected by the complexity and high nonlinearity of the existing flow solvers so long as they achieve a reasonable accuracy. Complex aerodynamic analysis and design of aircraft make use of such high-fidelity CFD tools for shape optimization for example, whereby some robustness is achieved by considering uncertain operational, environmental, or manufacturing parameters represented by random variables.

Both the intrusive and non intrusive approaches yield polynomial representations of the solution processes, known as surrogate models or response surfaces in the space of random parameters. They approximate the original stochastic processes solving the differential equations of interest accurately (in the mean-square sense), while being many orders faster to evaluate. One can thus consider these surrogates to compute the probability measures, moments, and/or sensitivities of the solutions or output quantities of interest related to them such as integrals, supremum norms, *etc*. The robust and most popular way to quantify uncertainties is Monte-Carlo estimation, but it may become intractable for complex models in which a simulation for one single value of the parameters may take several hours or days and a large number of model outputs have to be evaluated. Polynomial chaos is essentially a spectral representation in the random space, which typically exhibits fast convergence when the expended processes depend smoothly on the random parameters. Exponentially fast convergence can even be achieved under certain circumstances [24]. This rest of the paper is organized as follows. In section 2 we formulate our problem and introduce the probabilistic framework, focusing on the polynomial chaos expansion methodology. The actual methods for computing the polynomial chaos expansion coefficients are outlined in section 3. Here we also briefly review how these polynomial surrogates are used for uncertainty quantification. In the last section 4 we discuss the different approaches for their implementation in computational fluid dynamics with applications in aerodynamics and aeroelasticity, where complex configurations have been considered in recent works. We more particularly stress the "sparsity-of-effects" trend observed there that favors regression techniques benefitting from the sparsity of the output quantities of interest, such as compressed sensing.

§2. Problem setup

2.1. Model equations

Let $\mathcal{D} \subset \mathbb{R}^3$ be a fixed domain with a boundary $\partial\mathcal{D}$ and $\mathbf{x} \in \mathcal{D}$ the physical coordinates. Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a probability space where Ω is the abstract set of elementary events, $\mathcal{A}$ is a σ-

algebra of subsets of Ω, and $\mathcal{P}$ is a probability measure on $\mathcal{A}$. Our aim is to approximate the random field $u(\mathbf{x};\boldsymbol{\xi}) : \mathcal{D}\times\Gamma \to \mathbb{R}^m$ satisfying the parameterized partial differential equations:

$$\begin{aligned} \mathcal{L}(\mathbf{x},\boldsymbol{\xi};u) &= 0 \quad \text{in } \mathcal{D}\,, \\ \mathcal{B}(\mathbf{x},\boldsymbol{\xi};u) &= 0 \quad \text{on } \partial\mathcal{D}\,, \end{aligned} \tag{2.1}$$

where $\mathcal{L}$ is a linear or non linear differential operator and $\mathcal{B}$ is a boundary operator. Here $\boldsymbol{\xi}(\omega) = (\xi_1(\omega), \xi_2(\omega), \ldots, \xi_d(\omega)) : \Omega \to \Gamma \subseteq \mathbb{R}^d$ is a vector of d random parameters defined on $(\Omega, \mathcal{A}, \mathcal{P})$ with probability distribution $\mathcal{P}_{\boldsymbol{\Xi}}$, of which components $\xi_1, \xi_2, \ldots, \xi_d$ are mutually independent random variables with values in subsets of $\mathbb{R}$, $\Gamma_1, \Gamma_2, \ldots \Gamma_d$ respectively. We consider without loss of generality that the random field u has scalar values, *i.e.* $m = 1$. In practice one may also be interested in quantities:

$$y = F(u(\cdot;\boldsymbol{\xi}))\,, \tag{2.2}$$

that are functions of the solution u of the boundary value problem (2.1), in addition to the solution itself. In CFD for instance, u may be the pressure field satisfying the compressible Navier-Stokes equations about a fixed profile, and y may be the aerodynamic forces (*e.g.* drag, lift) exerted by the flow on that profile. In this latter case, the differential operator $\mathcal{L}$ may also depend on time t, and the boundary value problem (2.1) needs be supplemented with initial conditions. We do not consider that more general situation in the following discussion, for its main features basically extend to time-dependent problems.

2.2. Probabilistic framework

The vector of random parameters $\boldsymbol{\xi}$ is representative of variable geometrical characteristics, boundary conditions, loads, physical or mechanical properties, or combinations of them. It can be discrete, continuous, or a combination of both. In the continuous case, it is understood that its probability distribution $\mathcal{P}_{\boldsymbol{\Xi}}$ admits a probability density function $\boldsymbol{\xi} \mapsto W_{\boldsymbol{\Xi}}(\boldsymbol{\xi})$ with values in $\mathbb{R}_+ = [0, +\infty[$ such that $\mathcal{P}_{\boldsymbol{\Xi}}(B) = \int_B W_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) d\boldsymbol{\xi}$ for any subset B of $\mathbb{R}^d$. In addition, $\mathcal{P}_{\boldsymbol{\Xi}}(d\boldsymbol{\xi}) = \mathcal{P}_1(d\xi_1)\times\mathcal{P}_2(d\xi_2)\times\cdots\times\mathcal{P}_d(d\xi_d)$ owing to the assumption of mutual independence. In the present setting, it is further assumed that the random parameters are exponentially integrable, that is there exists $\beta > 0$ such that:

$$\mathbb{E}\{e^{\beta\|\boldsymbol{\xi}\|}\} = \int_{\mathbb{R}^d} e^{\beta\|\boldsymbol{\xi}\|} \mathcal{P}_{\boldsymbol{\Xi}}(d\boldsymbol{\xi}) < +\infty\,, \tag{2.3}$$

where $\|\boldsymbol{\xi}\| = (\sum_{n=1}^d \xi_n^2)^{\frac{1}{2}}$ is the usual Euclidean norm in $\mathbb{R}^d$, and $\mathbb{E}\{\cdot\}$ is mathematical expectation. Together with mutual independence, it ensures that each random variable ξ_n possesses finite moments of all orders, that is $\mathbb{E}\{|\xi_n|^k\} = \int_{\mathbb{R}} |\xi|^k \mathcal{P}_n(d\xi) < +\infty$ for all $k \in \mathbb{N}$. This uniquely defines a sequence of univariate orthonormal polynomials $\{\psi_j^{(n)}\}_{j\in\mathbb{N}}$ associated with the probability measure $\mathcal{P}_n$ for all $1 \le n \le d$, and a sequence of multivariate orthonormal polynomials $\{\psi_{\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}^d}$ associated with the probability measure $\mathcal{P}_{\boldsymbol{\Xi}}$ given by:

$$\psi_{\mathbf{j}}(\boldsymbol{\xi}) = \prod_{n=1}^d \psi_{j_n}^{(n)}(\xi_n)\,, \quad \mathbf{j} = (j_1, j_2, \ldots, j_d) \in \mathbb{N}^d\,, \tag{2.4}$$

such that $\{\psi_{\mathbf{j}}(\boldsymbol{\xi})\}_{\mathbf{j}\in\mathbb{N}^d}$ constitutes an orthonormal sequence of random variables in the space $L^2(\Omega,\sigma(\boldsymbol{\xi}),\mathcal{P})$ of second-order random variables defined on the probability space endowed with the σ-algebra $\sigma(\boldsymbol{\xi})$ generated by the random parameters $\boldsymbol{\xi}$; see [9, Theorem 3.6]. Alternatively, the polynomial set $\{\psi_{\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}^d}$ constitutes an orthonormal basis of the functional space $L^2(\mathbb{R}^d,\sigma_B(\mathbb{R}^d),\mathcal{P}_{\boldsymbol{\Xi}}(d\boldsymbol{\xi}))$ of square-integrable functions with respect to $\mathcal{P}_{\boldsymbol{\Xi}}$, where $\sigma_B(\mathbb{R}^d)$ is the Borel σ-algebra on $\mathbb{R}^d$.

Consequently, any random variable u in $L^2(\Omega,\sigma(\boldsymbol{\xi}),\mathcal{P})$ can be expanded in a series of multivariate orthonormal polynomials in the random parameters $\boldsymbol{\xi}$ as:

$$u=\sum_{\mathbf{j}\in\mathbb{N}^d} u_{\mathbf{j}}\psi_{\mathbf{j}}(\boldsymbol{\xi})\,,\quad u_{\mathbf{j}}=\mathbb{E}\{u\psi_{\mathbf{j}}(\boldsymbol{\xi})\}=\int_{\mathbb{R}^d} u\psi_{\mathbf{j}}(\boldsymbol{\xi})\mathcal{P}_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})\,. \tag{2.5}$$

This is the so-called generalized polynomial chaos expansion. Likewise, the random field $\mathbf{x}\mapsto u(\mathbf{x};\boldsymbol{\xi})$ satisfying (2.1) has finite energy from physical considerations, so it belongs to $L^2(\mathbb{R}^d,\sigma_B(\mathbb{R}^d),\mathcal{P}_{\boldsymbol{\Xi}}(d\boldsymbol{\xi}))$ and can be expanded as:

$$u(\mathbf{x};\boldsymbol{\xi})=\sum_{\mathbf{j}\in\mathbb{N}^d} u_{\mathbf{j}}(\mathbf{x})\psi_{\mathbf{j}}(\boldsymbol{\xi})\,,\quad u_{\mathbf{j}}(\mathbf{x})=\mathbb{E}\{u(\mathbf{x};\boldsymbol{\xi})\psi_{\mathbf{j}}(\boldsymbol{\xi})\}=\int_{\mathbb{R}^d} u(\mathbf{x};\boldsymbol{\xi})\psi_{\mathbf{j}}(\boldsymbol{\xi})\mathcal{P}_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})\,. \tag{2.6}$$

In practical numerical applications the foregoing expansions are truncated up to a total order p such that $|\mathbf{j}|=j_1+j_2+\cdots+j_d\le p$. Denoting by $\mathbb{P}^p[\cdot]$ the orthogonal projection onto the space of d-variate polynomials of total degree p in $\xi_1,\xi_2,\ldots,\xi_d$, say V_d^p, we seek for an approximate solution $\mathbb{P}^p[u]$ of (2.1) in V_d^p as:

$$u(\mathbf{x};\boldsymbol{\xi})\simeq\mathbb{P}^p[u](\mathbf{x};\boldsymbol{\xi})=\sum_{|\mathbf{j}|\le p} u_{\mathbf{j}}(\mathbf{x})\psi_{\mathbf{j}}(\boldsymbol{\xi})=\sum_{j=0}^{P-1} u_j(\mathbf{x})\psi_j(\boldsymbol{\xi})\,,\quad P=\binom{p+d}{d}, \tag{2.7}$$

by reordering the P multi-indices $\mathbf{j}$ such that $|\mathbf{j}|\le p$. From [9, Theorem 2.2], the sequence $\mathbb{P}^p[u]$ converges to u in the mean-square sense in $L^2(\Omega,\sigma(\boldsymbol{\xi}),\mathcal{P})$ as $p\to+\infty$ provided that the condition (2.3) is fulfilled.

Now the deterministic functional coefficients $\mathbf{x}\mapsto u_j(\mathbf{x})$ in the truncated series remain unknown since the random field u is unknown. Collocational or weighted versions of (2.1) together with the above approximation are considered in order to determine them.

§3. Construction of the polynomial chaos expansion

The different methods considered for computing the polynomial expansion coefficients are quoted as intrusive or non intrusive in the mechanical engineering literature. The stochastic Galerkin method is intrusive in that it may require significant alterations of the existing deterministic codes utilized for solving numerically the boundary value problem (2.1). Generally it also yields coupled equations for the expansion coefficients of its solution. Hence new codes need be developed to handle the larger and coupled systems of equations arising from the Galerkin formulation. The stochastic collocation and regression methods are non intrusive in that they require only repetitive executions of the existing deterministic codes for carefully selected parameter sets. They are the preferred methodologies in CFD, for their applicability is not affected by the complexity and high nonlinearity of the existing flow solvers.

3.1. Stochastic Galerkin method

Similarly to the weak formulation of deterministic problems, on can form the weak form of (2.1) and seek an approximate solution $u^p \in V_d^p$ such that:

$$\begin{aligned} &\mathbb{E}\{\mathcal{L}(\mathbf{x},\boldsymbol{\xi};u^p)v(\boldsymbol{\xi})\} = 0 \quad \forall v(\boldsymbol{\xi}) \in V_d^p\,, \mathbf{x} \in \mathcal{D}\,, \\ &\mathbb{E}\{\mathcal{B}(\mathbf{x},\boldsymbol{\xi};u^p)v(\boldsymbol{\xi})\} = 0 \quad \forall v(\boldsymbol{\xi}) \in V_d^p\,, \mathbf{x} \in \partial\mathcal{D}\,. \end{aligned} \tag{3.1}$$

The resulting system becomes a deterministic one in the physical domain $\mathcal{D}$ for the functional coefficients $u_j(\mathbf{x})$, and may be solved by standard discretization techniques *e.g.* finite elements, finite volumes, finite differences, boundary elements, *etc.*; see [13] and references therein for an extensive presentation of this method.

3.2. Stochastic collocation method

Alternatively, one may seek an approximate solution formed by interpolation between solutions of (2.1) for Q particular choices of the random parameters $\boldsymbol{\xi}$, namely the sampling set $\{\boldsymbol{\xi}^l\}_{1\le l\le Q}$ of so-called nodes, such that:

$$\begin{aligned} &\mathcal{L}(\mathbf{x},\boldsymbol{\xi}^l;u(\mathbf{x};\boldsymbol{\xi}^l)) = 0 \quad \forall l = 1,2,\ldots,Q\,, \mathbf{x} \in \mathcal{D}\,, \\ &\mathcal{B}(\mathbf{x},\boldsymbol{\xi}^l;u(\mathbf{x};\boldsymbol{\xi}^l)) = 0 \quad \forall l = 1,2,\ldots,Q\,, \mathbf{x} \in \partial\mathcal{D}\,. \end{aligned} \tag{3.2}$$

Then the approximate solution $\mathbb{I}^Q[u]$ to (2.1) reads as the Lagrange interpolation [14, 23]:

$$u(\mathbf{x};\boldsymbol{\xi}) \simeq \mathbb{I}^Q[u](\mathbf{x};\boldsymbol{\xi}) = \sum_{l=1}^{Q} u(\mathbf{x};\boldsymbol{\xi}^l) L_l(\boldsymbol{\xi})\,, \tag{3.3}$$

where $\{L_l\}_{1\le l\le Q}$ is the set of d-variate Lagrange polynomials based on the nodes $\{\boldsymbol{\xi}^l\}_{1\le l\le Q}$ chosen so that uniqueness of the interpolation is ensured.

3.2.1. Link with polynomial chaos

Choosing the nodes within a quadrature rule $\boldsymbol{\Theta}(d,Q) = \{\boldsymbol{\xi}^l, w^l\}_{1\le l\le Q}$ tailored such that $\sum_{l=1}^{Q} w^l f(\boldsymbol{\xi}^l)$ is a good approximation of the d-dimensional integral $\int_{\mathbb{R}^d} f(\boldsymbol{\xi})\mathcal{P}_{\boldsymbol{\Xi}}(d\boldsymbol{\xi}) = \mathbb{E}\{f(\boldsymbol{\xi})\}$ for sufficiently smooth functions f, the collocation approach may be considered to compute an approximate solution $\mathbb{P}_Q^p[u]$ defined by:

$$\begin{aligned} \mathbb{P}_Q^p[u](\mathbf{x};\boldsymbol{\xi}) &= \sum_{j=0}^{P-1}\left(\sum_{l=1}^{Q} w^l u(\mathbf{x};\boldsymbol{\xi}^l)\psi_j(\boldsymbol{\xi}^l)\right)\psi_j(\boldsymbol{\xi}) = \sum_{l=1}^{Q} u(\mathbf{x};\boldsymbol{\xi}^l)\left(w^l\sum_{j=0}^{P}\psi_j(\boldsymbol{\xi}^l)\psi_j(\boldsymbol{\xi})\right) \\ &= \sum_{l=1}^{Q} u(\mathbf{x};\boldsymbol{\xi}^l)\tilde{L}_l(\boldsymbol{\xi}) \end{aligned} \tag{3.4}$$

in view of (2.7); that is, the quadrature set $\boldsymbol{\Theta}(d,Q)$ is used to evaluate the coefficients $u_j(\mathbf{x})$ in (2.7). Provided that the quadrature rule $\boldsymbol{\Theta}(d,Q)$ integrates exactly all d-variate polynomials of total order $2p$ and $L_l \in V_d^p$, one has $\tilde{L}_l \equiv L_l$ owing to the orthonormalization of the polynomials $\{\psi_j\}_{0\le j\le P-1}$ which are such that $\mathbb{E}\{\psi_j(\boldsymbol{\xi})\psi_k(\boldsymbol{\xi})\} = \delta_{jk}$, the Kronecker symbol.

3.2.2. Choices of nodal set

The key issue of stochastic collocation is the selection of appropriate sampling sets. A straightforward choice is quadrature nodes and weights as in (3.4). Multi-dimensional quadrature sets $\boldsymbol{\Theta}(d, Q) = \{\boldsymbol{\xi}^l, w^l\}_{1\leq l\leq Q}$, where $\boldsymbol{\xi}^l$ is the l-th node in $\Gamma = \prod_{n=1}^d \Gamma_n$ and w^l is the corresponding weight, may be constructed from one-dimensional (univariate) quadrature sets by full tensorization or sparse tensorization, following Smolyak's algorithm [18].

Univariate Gauss quadratures $\Theta(1, q_1)$ based on q_1 integration points are tailored to integrate a smooth function $\xi \mapsto f(\xi)$ on $\Gamma_1 \equiv [a, b]$ by:

$$\int_{\Gamma_1} f(\xi) W_\Xi(\xi) d\xi \simeq \sum_{l=1}^{q_1-r} w^l f(\xi^l) + \sum_{m=1}^{r} w^{q_1-r+m} f(\xi^{q_1-r+m}), \tag{3.5}$$

such that this rule turns to be exact for univariate polynomials up to the order $2q_1 - 1 - r$. Here r is the number of fixed nodes of the rule, typically the bounds a, b. Depending on the choice of r, different terminologies are used:

- $r = 0$ is the classical Gauss rule;
- $r = 1$ is the Gauss-Radau rule, choosing $\xi^{q_1} = a$ or $\xi^{q_1} = b$ for instance;
- $r = 2$ is the Gauss-Lobatto (GL) rule, choosing $\xi^{q_1-1} = a$ and $\xi^{q_1} = b$ for instance.

Multivariate quadratures may subsequently be obtained by full or sparse tensorization of these one-dimensional rules. Firstly, a fully tensorized grid is obtained by the product rule:

$$\boldsymbol{\Theta}(d, Q) = \bigotimes_{n=1}^{d} \Theta(1, q_n), \tag{3.6}$$

which contains $Q = \prod_{n=1}^d q_n$ grid points in Γ. Secondly, a sparse quadrature rule can be derived thank to the Smolyak algorithm [18]. The so-called k-th level, d-dimensional Smolyak sparse rule $\widehat{\boldsymbol{\Theta}}(d, k)$ is obtained by the following linear combination of product formulas:

$$\widehat{\boldsymbol{\Theta}}(d, k) = \sum_{l=k-d}^{k-1} \sum_{q_1+\cdots+q_d=d+l} \Theta(1, q_1) \otimes \cdots \otimes \Theta(1, q_d). \tag{3.7}$$

Clearly, the above sparse grid is a subset of the full tensor product grids. It typically contains $Q \sim (2d)^k/k!$ nodes in Γ whenever $d \gg 1$ and k is fixed. By a direct extension of the arguments divised in [15], it can be shown that provided the univariate quadrature rules $\Theta(1, q)$ are exact for all univariate polynomials of order up to $2q - 1$ (Gauss rules) or $2q - 3$ (GL rules), the foregoing rule is exact for all d-variate polynomials of total order up to $2k - 1$ or $2k-3$, respectively. In [17] it has been observed that sparse quadratures outperform tensorized quadratures with non-nested underlying one-dimensional rules whenever $d \geq 4$, though. If $\Theta(1, q_i)$ is now Clenshaw-Curtis univariate quadrature of i-th level for $i > 1$, such that:

$$\xi^l = -\cos\frac{(l-1)\pi}{q_i - 1}, \quad 1 \leq l \leq q_i = 2^{i-1} + 1,$$

then $\Theta(1, q_i) \subset \Theta(1, q_{i+1})$, that is the univariate Clenshaw-Curtis rules $\Theta(1, q_i)$ are nested. Consequently the multivariate rules are nested as well, $\widehat{\boldsymbol{\Theta}}(d, k) \subset \widehat{\boldsymbol{\Theta}}(d, k + 1)$. The total number of nodes is significantly reduced compared to non nested rules. Nested Clenshaw-Curtis rules are however exact at least for all multivariate polynomials of total order k [1].

3.3. Regression methods

In regression approaches, the P expansion coefficients in (2.7) are determined on the basis of a set of observations $\{u(\cdot;\boldsymbol{\xi}^l\}_{1\leq l\leq Q}$, obtained by computations or measurements, of the random variable or field u for some particular choices of the random parameters $\boldsymbol{\xi}$, again the sampling set $\{\boldsymbol{\xi}^l\}_{1\leq l\leq Q}$. They consist in solving a weighted least-squares minimization problem:

$$\mathbf{U} \simeq \mathbf{U}^\star = \arg\min_{\mathbf{V}\in\mathbb{R}^P} \frac{1}{2}\,(\mathbf{y}-\boldsymbol{\Phi}\mathbf{V})^\mathsf{T}\,\mathbf{W}\,(\mathbf{y}-\boldsymbol{\Phi}\mathbf{V})\ , \tag{3.8}$$

where $\mathbf{y} = (u(\cdot;\boldsymbol{\xi}^1), u(\cdot;\boldsymbol{\xi}^2), \ldots, u(\cdot;\boldsymbol{\xi}^Q))^\mathsf{T}$ is the vector of observations, $[\boldsymbol{\Phi}]_{lj} = \psi_j(\boldsymbol{\xi}^l)$ is the so-called $Q\times P$ measurement matrix, $\mathbf{W}$ is a $Q\times Q$ weighting matrix, and $\mathbf{U} = (u_0, u_1, \ldots, u_{P-1})^\mathsf{T}$ is the sought vector of coefficients. This is the approach retained in *e.g.* [2], for which numerous methods are available to solve this optimization problem whenever $Q \geq P$. Alternatively, one may consider the situation whereby $Q < P$ and more particularly $Q \ll P$, that is, underdetermined systems. This can be achieved thanks to some recent results pertaining to the resolution of under-sampled linear systems promoting sparsity of the sought solution, known as compressed sensing or compressive sampling [4, 8]. A review of the application of this approach to generalized polynomial chaos expansions is proposed in [11]; see also [12, 17] for applications in aerodynamics and aeroelasticity. The compressed sensing approach consists in reformulating the least-squares minimization problem (3.8) as a convex minimization problem with some sparsity constraint, namely:

$$\mathbf{U} \simeq \mathbf{U}^\star = \arg\min_{\mathbf{V}\in\mathbb{R}^P} \{\|\mathbf{V}\|_1\ ;\ \|\mathbf{y}-\boldsymbol{\Phi}\mathbf{V}\|_2 \leq \varepsilon\}\ , \tag{3.9}$$

for some tolerance $0 \leq \varepsilon \ll 1$ on the polynomial chaos truncation (2.7). Here the ℓ_m-norm is $\|\mathbf{a}\|_m = (\sum_{j=0}^{P-1} |a_j|^m)^{\frac{1}{m}}$ for $m > 0$, and $\|\mathbf{a}\|_0 = \#\{j;\, a_j \neq 0\}$ otherwise. Sparsity means that only a small fraction of the sought coefficients $\mathbf{U}$ are non negligible. The latter problem is known as basis pursuit denoising [6]. It is uniquely solvable thanks to some *ad hoc* mixing properties of the measurement matrix $\boldsymbol{\Phi}$.

One of them is the restricted isometry property (RIP) or uniform uncertainty principle. For each integer $S \in \mathbb{N}^*$, the isometry constant δ_S of $\boldsymbol{\Phi}$ is defined as the smallest number such that:

$$(1-\delta_S)\,\|\mathbf{U}_S\|_2^2 \leq \|\boldsymbol{\Phi}\mathbf{U}_S\|_2^2 \leq (1+\delta_S)\,\|\mathbf{U}_S\|_2^2$$

for all S-sparse vectors $\mathbf{U}_S \in \{\mathbf{V}\in\mathbb{R}^P;\ \|\mathbf{V}\|_0 \leq S\}$. Then $\boldsymbol{\Phi}$ is said to satisfy the RIP of order S if, say, δ_S is not too close to 1. This property amounts to saying that all S-column submatrices of $\boldsymbol{\Phi}$ are numerically well-conditioned, or S (or less) columns selected arbitrarily in $\boldsymbol{\Phi}$ are nearly orthogonal. Consequently, they form a near isometry so that $\boldsymbol{\Phi}$ approximately preserves the Euclidean norm of S-sparse vectors. The following theorem by Candès *et al.* [4, 5] then states that (3.9) can be solved efficiently:

Theorem 1. *Assume $\delta_{2S} < \sqrt{2}-1$. Then the solution $\mathbf{U}^\star$ to (3.9) satisfies:*

$$\left\|\mathbf{U}^\star - \mathbf{U}\right\|_2 \leq C_0\,\frac{\|\mathbf{U}_S - \mathbf{U}\|_1}{\sqrt{S}} + C_1\varepsilon$$

for some $C_0, C_1 > 0$ depending only on δ_{2S}. Here $\mathbf{U}_S$ is $\mathbf{U}$ with all but the S largest entries set to zero.

This result calls for several comments. First, the coefficients $\mathbf{U}$ are actually nearly sparse, rather than strictly sparse, in the sense that only a small fraction of them contribute significantly to the output statistics while the others are not strictly null. Opportunely, the foregoing theorem deals with all signals and not only the S-sparse ones. In addition, it also allows noiseless recovery if $\varepsilon = 0$. Second, it is deterministic and does not involve any probability for a successful recovery. Lastly, the ℓ_1-minimization strategy is non adapted because it identifies the sparsity pattern, that is the order (location) of the negligible coefficients in the polynomial chaos basis, and the leading coefficients at the same time. The algorithm can therefore efficiently capture the relevant information of a sparse vector without trying to comprehend that vector [5]. This is clearly a much desirable feature for practical industrial applications. Additionally, the RIP prompts the use of unstructured observation sets $\{\boldsymbol{\xi}^l\}_{1\leq l\leq Q}$, typically selected randomly, for an efficient recovery by basis pursuit. Structured sets may also be considered, though, as proposed in [21].

3.4. Application to uncertainty quantification

Once the polynomial expansion (2.7) has been derived, the first moments and/or cumulants of the random field u can be computed with this expansion. Owing to the orthonormality of the polynomials, the mean and variance for example are:

$$\mu(\mathbf{x}) = \mathbb{E}\{u(\mathbf{x};\boldsymbol{\xi})\} = u_0(\mathbf{x})\,, \quad \sigma^2(\mathbf{x}) = \mathbb{E}\{(u(\mathbf{x};\boldsymbol{\xi}) - \mu(\mathbf{x}))^2\} = \sum_{j=1}^{P-1} u_j^2(\mathbf{x})\,.$$

Sensitivity indices may be computed alike [19]. They quantify the fraction of variance of the solution u which can be related to the variation of each random parameter. Denoting by $\mathcal{J}_n$ the set of indices corresponding to the polynomials depending only on the n-th variable parameter ξ_n, the main-effect Sobol' indices are given by:

$$\mathcal{S}_n(\mathbf{x}) = \frac{\operatorname{Var}\mathbb{E}\{u(\mathbf{x};\boldsymbol{\xi})|\xi_n\}}{\operatorname{Var}u(\mathbf{x};\boldsymbol{\xi})} = \frac{1}{\sigma^2(\mathbf{x})}\sum_{j\in\mathcal{J}_n} u_j^2(\mathbf{x})\,. \tag{3.10}$$

More generally, if $\mathcal{J}_{n_1 n_2 \ldots n_s}$ is the set of indices corresponding to the polynomials depending only on the parameters $\xi_{n_1}, \xi_{n_2}, \ldots \xi_{n_s}$, the s-fold joint sensitivity indices are:

$$\mathcal{S}_{n_1 n_2 \ldots n_s}(\mathbf{x}) = \frac{\operatorname{Var}\mathbb{E}\{u(\mathbf{x};\boldsymbol{\xi})|\xi_{n_1}, \xi_{n_2}, \ldots \xi_{n_s}\}}{\operatorname{Var}u(\mathbf{x};\boldsymbol{\xi})} = \frac{1}{\sigma^2(\mathbf{x})}\sum_{j\in\mathcal{J}_{n_1 n_2 \ldots n_s}} u_j^2(\mathbf{x})\,.$$

§4. Discussion

Because of the high complexity of fluid flow solvers, non intrusive uncertainty quantification techniques have been primarily developed in aerodynamic and aeroelastic simulations. They are used to compute the sensitivities of output quantities of interest that are required to evaluate the objective function of an optimization process, for example. Polynomial surrogate models have commonly been considered in this respect. In most applications the polynomial expansion coefficients are evaluated by Gauss quadratures (see section 3.2.2). However

this approach becomes computationally very demanding for parametric spaces of high dimensions, even if sparse rules are utilized: this is the so-called curse of dimensionality. Observing that the output quantities of interest of complex systems depend only weekly on the multiple cross-interactions between the variable inputs, one may argue that only low-order polynomials significantly contribute to their surrogates. This feature favors reconstruction techniques benefiting from such a sparse structure, as compressed sensing (see section 3.3). It should be noted that the "sparsity-of-effects" principle invoked here has already been outlined in [16]. It may be established rigorously for parameterized, possibly non linear elliptic-parabolic equations in the framework analyzed in [7]. The results obtained with aerodynamic and aeroelastic simulations involving complex fluid flows solved by Reynolds-averaged Navier-Stokes equations (RANS) with turbulence transport closure models corroborate to a large extend this expected trend. Such examples are described in [17] for the case of a two-dimensional rigid profile with random Mach number, angle-of-attack, and thickness-to-chord ratio; and in [12] for the case of a three-dimensional flexible wing-fuselage configuration with random Mach number, lift force, and wing structural stiffness. Efficient non-adapted polynomial reconstructions with sampling sets orders of magnitude smaller than the ones required by the usual techniques are achieved. The (global) quantities of interest considered in these applications are typically the drag force and pitching moment of the profiles, which integrate the (local) pressure fields along them.

References

[1] Barthelmann, V., Novak, E., and Ritter, K. High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics 12*, 4 (2000), 273–288.

[2] Berveiller, M., Sudret, B., and Lemaire, M. Stochastic finite element: a non-intrusive approach by regression. *Revue Européenne de Mécanique Numérique 15*, 1-3 (2006), 81–92.

[3] Cameron, R., and Martin, W. The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals. *Annals of Mathematics 48*, 2 (1947), 385–392.

[4] Candès, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics 59*, 8 (2006), 1207–1223.

[5] Candès, E. J., and Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Processing Magazine 25*, 2 (2008), 21–30.

[6] Chen, S. C., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing 20*, 1 (1998), 33–61.

[7] Chkifa, A., Cohen, A., and Schwab, C. Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *Journal de Mathématiques Pures et Appliquées 103*, 2 (2014), 400–428.

[8] Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory 52*, 4 (2006), 1289–1306.

[9] Ernst, O. G., Mugler, A., Starkloff, H.-J., and Ullmann, E. On the convergence of generalized polynomial chaos expansions. *ESAIM: Mathematical Modelling and Numerical Analysis 46*, 2 (2012), 317–339.

[10] Ghanem, R., and Spanos, P. D. *Stochastic finite elements: A Spectral Approach*. Springer, New York, 1991.

[11] Hampton, J., and Doostan, A. Compressive sampling methods for sparse polynomial chaos expansions. In *Handbook of Uncertainty Quantification*, R. G. Ghanem, D. Higdon, and H. Owhadi, Eds. Springer, Cham, 2016. 29 pages.

[12] Hantrais-Gervois, J.-L., and Savin, É. Application of efficient non-intrusive UQ prediction methods to aeroelastic test cases. UMRIDA Technical Report D3.3-36(9), 2016.

[13] Le Maître, O., and Knio, O. *Spectral Methods for Uncertainty Quantification. With Applications to Computational Fluid Dynamics*. Springer, Dordrecht, 2010.

[14] Mathelin, L., Hussaini, M., and Zang, T. Stochastic approaches to uncertainty quantification in CFD simulations. *Numerical Algorithms 38*, 1 (2005), 209–236.

[15] Novak, E., and Ritter, K. Simple cubature formulas with high polynomial exactness. *Constructive Approximation 15*, 4 (1999), 499–522.

[16] Rabitz, H., Aliş, Ö. F., Shorter, J., and Shim, K. Efficient input–output model representations. *Computer Physics Communications 117*, 1-2 (1999), 11–20.

[17] Savin, É., Resmini, A., and Peter, J. Sparse polynomial surrogates for aerodynamic computations with random inputs. AIAA paper 2016-0433, 2016.

[18] Smolyak, S. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Mathematics Doklady 4* (1963), 240–243.

[19] Sobol', I. M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation 55*, 1-3 (2001), 271–280.

[20] Sun, T.-C. A finite element method for random differential equations with random coefficients. *SIAM Journal on Numerical Analysis 16*, 6 (1979), 1019–1035.

[21] Tang, G., and Iaccarino, G. Subsampled Gauss quadrature nodes for estimating polynomial chaos expansions. *SIAM/ASA Journal on Uncertainty Quantification 2*, 1 (2014), 423–443.

[22] Wiener, N. The homogeneous chaos. *American Journal of Mathematics 60*, 4 (1938), 897–936.

[23] Xiu, D., and Hesthaven, J. S. High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing 27*, 3 (2005), 1118–1139.

[24] Xiu, D., and Karniadakis, G. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing 24*, 2 (2002), 619–644.

E. Savin
Onera – Computational Fluid Dynamics Dept.
29, avenue de la Division Leclerc
F-92322 Châtillon cedex, France
eric.savin@onera.fr

Monografías Matemáticas *García de Galdeano* **42**, 285–295 (2019)

TOOLS TO PROVE A PARABOLIC LEWY-STAMPACCHIA'S INEQUALITY

Yassine Tahraoui

Abstract. We studied a quasilinear parabolic variational inequality of Lewy-Stampacchia type governed by a pseudomonotone operator of Leray-Lions type in a joint work with O. Guibé, A. Mokrane and G. Vallet [11]. We propose here some tools and techniques used to deal with the difficulties, which appear in the study of the problem.

Keywords: Variational inequalities, penalization, pseudomonotone operator, Lewy-Stampacchia's inequality.

AMS classification: 35K86, 35R35.

§1. Introduction

We are interessted in a nonlinear parabolic problem with contraint and homogeneous Dirichlet boundary conditions. More precisely, we prove the existence of a solution satisfying the following Lewy-Stampacchia's inequality

$$0 \leq \partial_t u - \operatorname{div}[a(\cdot,\cdot,u,\nabla u)] - f \leq g^- = (f - \partial_t \psi + \operatorname{div}[a(\cdot,\cdot,\psi,\nabla\psi)])^-,$$

associated with the following problem

$$\int_0^T \langle \partial_t u, v-u\rangle dt + \int_Q a(t,x,u,\nabla u)\nabla(v-u)dxdt \geq \int_0^T \langle f, v-u\rangle dt, \quad u_0(0) = u_0$$

where $u \mapsto -\operatorname{div}[a(t,x,u,\nabla u)]$ is a pseudomomotone operator under the constraint $u \geq \psi$. We propose to present tools to show the existence of a solution for the above mentioned problem.

After the first results of H. Lewy and G. Stampacchia [7] concerning inequalities in the context of superharmonic problems, many authors have been interested in the so-called Lewy-Stampacchia's inequality associated with obstacle problems. Without exhaustiveness, let us cite the papers of A. Mokrane and F. Murat [9] for pseudo-monotone elliptic problems, A. Mokrane and G.Vallet [10] in the context of Sobolev spaces with variable exponents. The literature on Lewy-Stampacchia's inequality is mainly aimed at elliptic problems, or close to elliptic problems and fewer papers are concerned with other type of problems. Let us cite J. F. Rodrigues [12] for hyperbolic problems, F. Donati [4] for parabolic problems with a monotone operator or L. Mastroeni and M. Matzeu [8] in the case of a double obstacle.

The aim of O. Guibé, A. Mokrane, Y. Tahraoui and G. Vallet [11] was to extend F. Donati's work [4] to pseudo-monotone parabolic problems with a Leray-Lions operator. The authors proposed a result with very general assumptions on the Carathéodory function a, by using a method of penalization of the constraint associated with a suitable perturbation of the operator. As proposed e.g. by [6, p.102], this perturbation is one of the main new point

of the proof. Indeed, without it, one is usually only concerned by Lewy-Stampacchia's inequality in the elliptic case, and one needs to assume, as in [9], some additional, now useless, holder-continuity assumptions with respect to u and ∇u. Thus, this perturbation allows us on the one hand to prove Lewy-Stampacchia's inequality in the pseudomonotone parabolic case, and on the other hand to reduce significantly the list of assumptions. Let us mention also that, with this method, one is able to revisit Lewy-Stampacchia's inequality proposed in [9, 10] by assuming only basic assumptions. The second essential result is an extension of the formula of time-integration by parts of Mignot-Bamberger[2] & Alt-Luckhaus[1] to non-classical situations. Some information are also given too about the time-continuity of an element u when u and $\partial_t u$ are not in spaces in duality relation. We propose in this paper to present tools and techniques used by the authors to deal with the difficulties in the study of some terms in [11].

First of all, we need to precise the functional setting and the assumptions on the data. Denote by $D \subset \mathbb{R}^d, d \geq 1$ a Lipshitz bounded domain, $T > 0$, $Q = D\times]0,T[$and $p, p' \in]1,+\infty[$ such that$\frac{1}{p} + \frac{1}{p'} = 1$. $V = W_0^{1,p}(D)$ if $p \geq 2$ and $V = W_0^{1,p}(D) \cap L^2(D)$ with the graph-norm else. Then, $V' = W^{-1,p'}(D)$ if $p \geq 2$ and $V' = W^{-1,p'}(D) + L^2(D)$ else and the Lions-Guelfand triple $V \underset{d}{\hookrightarrow} H \underset{d}{\hookrightarrow} V'$ holds.

$W(0,T) = \{u \in L^p(0,T,V),\ \partial_t u \in L^{p'}(0,T,V')\}$ and $\mathcal{K}(\psi) := \{u \in W(0,T),\ u \geq \psi\}$.

Assume in the sequel the following:

H_1 :

$$A : W^{1,p}(D) \to W^{-1,p'}(D) \quad v \mapsto A(v) = -\operatorname{div}\Big[a(t,x,v,\nabla v)\Big],$$

where

$\mathrm{H}_{1,1}$ $a : (t,x,u,\xi) \in Q \times \mathbb{R} \times \mathbb{R}^d \mapsto a(t,x,u,\xi) \in \mathbb{R}^d$ is a Carathéodory function on $Q \times \mathbb{R}^{d+1}$,

$\mathrm{H}_{1,2}$ $\forall (t,x) \in Q$ a.e., $u \in \mathbb{R}$, $\forall \xi, \eta \in \mathbb{R}^d$,

$$\xi \neq \eta \Rightarrow [a(t,x,u,\xi) - a(t,x,u,\eta)].(\xi - \eta) > 0.$$

$\mathrm{H}_{1,3}$ $\exists \bar{\alpha} > 0, \bar{\beta} > 0$ and $\bar{\gamma} \geq 0$, functions $\bar{h} \in L^1(Q), \bar{k} \in L^p(Q)$ and two exponents $q, r < p$ such that, for a.e. $(t,x) \in Q$, $\forall u \in \mathbb{R}, \forall \xi \in \mathbb{R}^d$,

$$a(t,x,u,\xi).\xi \geq \bar{\alpha}|\xi|^p - \Big[\bar{\gamma}|u|^q + |\bar{h}(t,x)|\Big],$$
$$|a(t,x,u,\xi)| \leq \bar{\beta}\Big[|\bar{k}(t,x)| + |u|^{r/p} + |\xi|\Big]^{p-1}.$$

H_2 : $\psi \in L^p(0,T,W^{1,p}(D)) \cap L^p(0,T,L^2(D))$; that $\partial_t \psi$ belongs to $L^{p'}(0,T,V')$ and $\psi \leq 0$ on ∂D.

H_3 : the right hand side f, which is assumed to be such that $g = f - \partial_t \psi - A(\psi) = g^+ - g^-$ belongs to the order dual $L^p(0,T,V)^* = (L^{p'}(0,T,V'))^+ - (L^{p'}(0,T,V'))^+$, i.e. $g^+, g^- \in (L^{p'}(0,T,V'))^+$ the non-negative elements of $L^{p'}(0,T,V')$.

H_4 : $u_0 \in L^2(D)$ satisfies the constraint, *i.e.* $u_0 \geq \psi(0)$.

Let us now recall the main result in [11].

Theorem 1. *Under the above assumptions (H_1)-(H_4), there exists at least $u \in \mathcal{K}(\psi)$ with $u(t=0) = u_0$ and such that, for any $v \in L^p(0,T,V)$, $v \geq \psi$*

$$\int_0^T \langle \partial_t u, v-u\rangle dt + \int_Q a(t,x,u,\nabla u)\nabla(v-u)dxdt \geq \int_0^T \langle f, v-u\rangle dt.$$

Moreover, the following Lewy-Stampacchia's inequality holds

$$0 \leq \partial_t u - div[a(\cdot,\cdot,u,\nabla u)] - f \leq g^- = (f - \partial_t\psi + div[a(\cdot,\cdot,\psi,\nabla\psi)])^-.$$

§2. Strong continuity in $L^2(D)$

Let us denote by $V(D)$ ($V_0(D)$ resp.) the following space $W^{1,p}(D) \cap L^2(D)$ ($W_0^{1,p}(D) \cap L^2(D)$ resp.) and $V'(D) = W^{-1,p'}(D) + L^2(D)$. We have the following result.

Lemma 2. *If $u \in L^p(0,T;V(D))$ and $\partial_t u \in L^{p'}(0,T;V'(D))$ then $u \in C([0,T],L^2(D))$.*

Remark 1. This result is not the usual one since u and $\partial_t u$ are not in spaces being in duality relation and few words are needed concerning the time-derivative. Note that both $V(D)$ and $V_0(D)$ are dense subspaces of the chosen pivot space $L^2(D)$ so that it can be identify to a subspace of $V'(D)$ or $(V(D))'$. Therefore, u, as an element of $L^p(0,T;V(D)) \hookrightarrow L^P(0,T;L^2(D))$, has a time derivative in the sense of $\mathcal{D}'(0,T;L^2(D)) \hookrightarrow \mathcal{D}'(0,T;V'(D))$ and it is assumed to belong to $L^{p'}(0,T;V'(D))$.

Remark 2. Note that Lemma 2 ensures that the obstacle $\psi \in C([0,T],L^2(D))$ and therefore $u_0 \geq \psi(0)$ has a sense as elements of $L^2(D)$.

Sketch of the proof. This result is based on a classical method: first in $\mathbb{R}^d$, then in the half-space $\mathbb{R}^d_+$ and finally in D thanks to an atlas of charts.

For $D = \mathbb{R}^N$, we have $W_0^{1,p}(\mathbb{R}^N) = W^{1,p}(\mathbb{R}^N)$, therefore we can identify $V'(\mathbb{R}^N)$ with the dual of $V(\mathbb{R}^N)$. By considering the triple $V(\mathbb{R}^N) \underset{d}{\hookrightarrow} L^2(\mathbb{R}^N) \underset{d}{\hookrightarrow} V'(\mathbb{R}^N)$, thanks to [14] (Prop. 1.2 p. 106), one has $u \in C([0,T],L^2(\mathbb{R}^N))$.

If $D = \mathbb{R}^N_{+/resp.-} = \{(x',x_d) \in \mathbb{R}^d;\ x_d > 0$ (resp. $x_d < 0\}$, the method is based on a suitable extension of u to $\mathbb{R}^d$. Following a recommendation of F. Murat, we consider the following extension, used *e.g* in [5]

$$\tilde{u}(t,x',x_d) = \begin{cases} u(t,x',x_d); & x_d > 0 \\ -3u(t,x',-x_d) + 4u(t,x',-2x_d); & x_d < 0. \end{cases}$$

Note that $\tilde{u} \in L^p(0,T;V(\mathbb{R}^d))$ and, thanks to a change of variables, that for any $\varphi \in C_c^\infty(]0,T[\times\mathbb{R}^d)$ one gets

$$\int_0^T \int_{\mathbb{R}^d} \tilde{u}(t,x)\partial_t\varphi(t,x)\,dx\,dt = \int_0^T \int_{\mathbb{R}^d_-} (-3u(t,x',-x_d) + 4u(t,x',-2x_d))\partial_t\varphi(t,x',x_d)\,dx\,dt + \int_0^T \int_{\mathbb{R}^d_+} u(t,x)\partial_t\varphi(t,x)\,dx\,dt.$$

Then

$$\int_0^T \int_{\mathbb{R}^d} \tilde{u}(t,x)\partial_t\varphi(t,x)\,dx\,dt$$
$$= \int_0^T \int_{\mathbb{R}^d_+} (\partial_t(\varphi(t,x',x_d) - 3\varphi(t,x',-x_d) + 2\varphi(t,x',-\frac{x_d}{2}))u(t,x,x_d)\,dx\,dt.$$

Remark that $\psi(t,x) = \varphi(t,x',x_d) - 3\varphi(t,x',-x_d) + 2\varphi(t,x',-\frac{x_d}{2}) = 0 \quad if \quad x_d = 0$ and $\partial_t\psi(t,x) = 0$ if $x_d = 0$, which implies $\psi \in W^{1,\infty}(0,T;V_0(\mathbb{R}^d_+))$.
Note that $\|\psi\|_{L^p(0,T;V_0(\mathbb{R}^d_+))} \leq 8\|\varphi\|_{L^p(0,T;V(\mathbb{R}^d))}$. Therefore,

$$|\int_0^T \langle\partial_t\tilde{u},\varphi\rangle dt| = |\int_0^T \int_{\mathbb{R}^d_+} u\partial_t\psi\,dx\,dt| \leq \|\partial_t u\|_{L^{p'}(0,T;V'(\mathbb{R}^d_+))}\|\psi\|_{L^p(0,T;V_0(\mathbb{R}^d_+))} \leq C\|\varphi\|_{L^p(0,T;V(\mathbb{R}^d))}$$

Thus $\partial_t\tilde{u} \in L^{p'}(0,T;V'(\mathbb{R}^d))$.Then, one concludes that $\tilde{u} \in C([0,T],L^2(\mathbb{R}^d))$ i.e $u \in C([0,T],L^2(\mathbb{R}^d_+))$. Finally, the result holds in the general case by considering an atlas of charts as proposed *e.g* in [5]. □

§3. Penalization and perturbation of the operator

Denote by $\tilde{q} = \min(p,2)$ and let us define the function Θ

$$\Theta : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -[x^-]^{\tilde{q}-1},$$

and the perturbed operator

$$\tilde{a}(t,x,u,\xi) : Q\times\mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d \quad (x,t,u,\xi) \mapsto \tilde{a}(t,x,u,\xi) = a(t,x,\max(u,\psi(t,x)),\xi). \quad (3.1)$$

Remark 3. We wish to draw the reader's attention to the fact that with the proposed perturbation: $\tilde{a}(t,x,u,\xi) = a(t,x,\max(u,\psi),\xi)$, the idea is to make formally the operator monotone and not pseudomonotone any more on the free-set where the constraint is violated.

We define $\mathcal{A} : L^p(0,T;V) \to L^{p'}(0,T;V')$ such that $[\mathcal{A}(u)](t) := \tilde{A}(u(t)) = -\operatorname{div}[\tilde{a}(t,x,u,\nabla u)]$ and note that, the above assumption H_1 still holds.
For any positive ε, a cosmetic modification of [13, Section 8.4] yields the following result.

Theorem 3. *There exists $u_\epsilon \in W(0,T)$ such that $u_\epsilon(t=0) = u_0$ and*

$$\partial_t u_\epsilon - \operatorname{div}\left[\tilde{a}(t,x,u_\epsilon,\nabla u_\epsilon)\right] + \frac{1}{\varepsilon}\Theta(u_\epsilon - \psi) = f. \quad (3.2)$$

§4. From regular to general case

To prove the main result. On the one hand, we need some estimate for the penalization term. For that we impose an additional regularity on some data to get the desired estimate which permits to prove that the solution satisfies the constraint. On the other hand, we need some additional regularity to use an integration by part formula given in Section 5 to prove Lewy-Stampacchia's inequality. Then, we obtain the general case thanks the following density lemma.

Lemma 4. *The positive cone of $L^p(0,T;V)\cap L^2(Q)$ is dense in the positive cone of $\mathcal{V}'$, the dual set of $\mathcal{V}=L^p(0,T,V)$.*

Note that by truncation argument, the same result holds for the positive cone of $L^p(0,T;V)\cap L^{p'}(Q)$ when $p<2$. This result is given in [4, Lemma p.593]. We propose in [11] a sketch of a proof following the idea of [9].

§5. Mignot-Bamberger / Alt -Luckhaus integration by part formula

Note that $\mu_\varepsilon := \partial_t u_\varepsilon - \mathrm{div}[\tilde{a}(\cdot,\cdot,u_\varepsilon,\nabla u_\varepsilon)] - f = \frac{1}{\varepsilon}[(u_\varepsilon-\psi)^-]^{\tilde{q}-1} \geq 0$, so that the limit $\mu := \partial_t u - \mathrm{div}[\tilde{a}(\cdot,\cdot,u,\nabla u)] - f$ is a non-negative Radon measure which is also an element of $L^{p'}(0,T;V')$.

Using an idea from A. Mokrane and F. Murat [9], denote by $z_\varepsilon := g^- - \frac{1}{\varepsilon}[(u_\varepsilon-\psi)^-]^{\tilde{q}-1}$, we have

$$\partial_t u_\varepsilon + A(u_\varepsilon) + z_\varepsilon = g^+ + \partial_t\psi + A(\psi) \quad i.e. \quad \partial_t(u_\varepsilon-\psi) + A(u_\varepsilon) - A(\psi) + z_\varepsilon = g^+.$$

Observing that

$$\partial_t u_\varepsilon + A(u_\varepsilon) - f = -z_\varepsilon + g^-.$$

as in [9] in the elliptic case and under more restrictive assumptions on the operator a, proving that z_ε^- converges to 0 in an appropriate space leads to the Lewy-Stampacchia's inequality. Due to the time variable and the weak assumption on a we have to face to additional difficulties. For technical reasons, we will assume only that, on top of $g^- \in L^{p'}(Q)\cap L^p(0,T;V)$, $g^-\geq 0$, that $\partial_t g^- \in L^{\tilde{q}'}(Q)$. Roughly speaking it allows one to use a test function depending on g^- and together with Lemma 5 to perform an integration by part formula and then the convergence analysis of z_ε^-.

Lemma 5. *Consider $u\in L^p(0,T,W^{1,p}(D))\cap L^p(0,T,L^2(D))$ such that $\partial_t u\in L^{p'}(0,T,V')$. Let $\Psi : Q\times\mathbb{R}\to\mathbb{R}$ be a function such that $(t,x)\mapsto\Psi(t,x,\lambda)$ is measurable, $\lambda\mapsto\Psi(t,x,\lambda)$ is non-decreasing (càdlàg[1], or càglàd[2]) and denote by $\Lambda : Q\times\mathbb{R}\to\mathbb{R}, (t,x,\lambda)\mapsto\int_a^\lambda\Psi(t,x,\tau)d\tau$ where a is any arbitrary real number. Assume moreover that $|\Psi(t=0)|\leq h+|\lambda|^\alpha$ and that $\partial_t\Psi$ exists with $|\Psi(\lambda=0)|+|\partial_t\Psi|\leq h$ where $h\in L^2(Q)$ and $\alpha\in[0,1]$. If $\Psi(t,x,u)\in L^p(0,T,V)$, then, for any $\beta\in W^{1,\infty}(0,T)$ and any $0\leq s<t\leq T$,*

$$\int_s^t <\partial_t u,\Psi(\sigma,x,u)>\beta d\sigma = \int_D\Lambda(t,x,u(t))\beta(t)dx - \int_D\Lambda(s,x,u(s))\beta(s)dx - \int_s^t\int_D\Lambda(\sigma,x,u)\beta' dxd\sigma - \int_s^t\int_D\partial_t\Lambda(\sigma,x,u)\beta dxd\sigma.$$

Proof. We propose here to present the proof introduced in [11].Thanks to the assumptions, Ψ is a measurable function on $Q\times\mathbb{R}$ and Λ is a Carathéodory function on $Q\times\mathbb{R}$. Moreover,

$$|\Psi(t,x,\lambda)|\leq|\Psi(t=0)|+\int_0^t|\partial_t\Psi(s,x,\lambda)|ds\leq(T+1).h(t,x)+|\lambda|^\alpha,$$

$$|\Lambda(t,x,\lambda)|\leq|\lambda-a|\Big[(T+1).h(t,x)+|\lambda|^\alpha\Big]\leq C(T,a)\Big[|\lambda|^2+h^2(t,x)+h(t,x)+1\Big]$$

[1] right continuous with left limit

[2] left continuous with right limit

so that $\Lambda, \Psi \in L^2_{loc}(\mathbb{R}, L^2(Q))$ and the Nemitskii operator associated with Λ is continuous from $L^2(Q)$ to $L^1(Q)$. Concerning the time-derivation of Λ, for any $\varphi \in D(Q \times \mathbb{R})$, Fubini's theorem yields

$$-\int_{Q\times\mathbb{R}} \Lambda(t,x,\lambda)\partial_t\varphi(t,x,\lambda)dtdxd\lambda = -\int_{Q\times\mathbb{R}}\int_a^\lambda \Psi(t,x,\tau)d\tau\partial_t\varphi(t,x,\lambda)dtdxd\lambda$$
$$= \int_{Q\times\mathbb{R}}\int_a^\lambda \partial_t\Psi(t,x,\tau)d\tau\varphi(t,x,\lambda)dtdxd\lambda.$$

As a consequence,

$$\partial_t\Lambda(t,x,\lambda) = \int_a^\lambda \partial_t\Psi(t,x,\tau)d\tau, \quad \left|\partial_t\Lambda(t,x,\lambda)\right| \le |\lambda - a|h(t,x) \le |\lambda|^2 + h^2(t,x)/4 + |a|h(t,x)$$

so that the Nemitskii operator associated with $\partial_t\Lambda$ is continuous from $L^2(Q)$ to $L^1(Q)$.

Thanks to the assumptions, $u \in C([0,T], L^2(D))$ and one extends u to $\bar{u}$ in $\mathbb{R}$ by $\bar{u}(t) = u_0$ if $t < 0$ and $\bar{u}(t) = u(T)$ si $t > T$. Therefore, if $I_1 := (-1, T+1)$, $\bar{u} \in L^p(I_1, W^{1,p}(D)) \cap L^\infty(I_1, L^2(D)) \cap C(\bar{I}_1, L^2(D))$ such that $\partial_t\bar{u} \in L^{p'}(I_1, V')$ with $\partial_t\bar{u} = 0$ when $t < 0$ or $t > T$. Similarly to u, denote by $\bar{\Psi}$ the extension to I_1 of Ψ in the same way and by $\bar{\Lambda}$ the corresponding integral as introduced in the Lemma.

For any fixed $0 < h << 1$, let us denote by

$$v_h : t \mapsto \frac{\bar{u}(t+h) - \bar{u}(t)}{h}, \quad w_h : t \mapsto \frac{\bar{u}(t) - \bar{u}(t-h)}{h}.$$

Consider $\beta \in \mathcal{D}(I_1)$ and h, small enought so that $\text{supp}\beta + [-h, h] \subset I_1$. Then,

$$\int_{I_1} v_h(t)\beta(t)dt = \frac{1}{h}\int_{I_1}[\bar{u}(t+h) - \bar{u}(t)]\beta(t)dt$$
$$= \frac{1}{h}\int_{I_1}\bar{u}(t)\beta(t-h)dt - \frac{1}{h}\int_{I_1}\bar{u}(t)\beta(t)dt = \frac{1}{h}\int_{I_1}\bar{u}(t)[\beta(t-h) - \beta(t)]dt$$
$$\longrightarrow -\int_{-1}^{T+1}\bar{u}(t)\beta'(t)dt = -\int_0^T u(t)\beta'(t)dt + u(T)\beta(T) - u_0\beta(0) \quad \text{in } L^2(D);$$

similarly,

$$\int_{I_1} w_h(t)\beta(t)dt = \frac{1}{h}\int_{I_1}[\bar{u}(t) - \bar{u}(t-h)]\beta(t)dt$$
$$= \frac{1}{h}\int_{I_1}\bar{u}(t)\beta(t)dt - \frac{1}{h}\int_{I_1}\bar{u}(t)\beta(t+h)dt = \frac{1}{h}\int_{I_1}\bar{u}(t)[\beta(t) - \beta(t+h)]dt$$
$$\longrightarrow -\int_{-1}^{T+1}\bar{u}(t)\beta'(t)dt = -\int_0^T u(t)\beta'(t)dt + u(T)\beta(T) - u_0\beta(0) \quad \text{in } L^2(D),$$

so that v_h and w_h converge to $\partial_t\bar{u}$ in $\mathcal{D}'[I_1, L^2(D)]$, thus in $\mathcal{D}'[I_1, V']$; and to $\partial_t u$ in $\mathcal{D}'[0, T, L^2(D)]$ and $\mathcal{D}'[0, T, V']$. Moreover, by [3, Corollary A.2 p.145], the properties of Bochner integral

and since $\partial_t \bar{u} = 0$ outside $(0, T)$,

$$\int_{I_1} \|v_h(t)\|_{V'}^{p'} dt = \int_{I_1} \frac{1}{h^{p'}} \| \int_t^{t+h} \partial_t \bar{u}(s) ds \|_{V'}^{p'} dt \leq \int_{I_1} \frac{1}{h} \int_t^{t+h} \|\partial_t \bar{u}(s)\|_{V'}^{p'} ds dt$$
$$\leq \frac{1}{h} \int_{I_1} \int_{-1}^{t+h} \|\partial_t \bar{u}(s)\|_{V'}^{p'} ds dt - \frac{1}{h} \int_{I_1} \int_{-1}^{t} \|\partial_t \bar{u}(s)\|_{V'}^{p'} ds dt = \int_0^T \|\partial_t u(s)\|_{V'}^{p'} ds.$$

Since v_h already converges in the sense of Distributions, as a consequence of the above estimate, one may conclude that v_h converges weakly to $\partial_t \bar{u}$ in $L^{p'}[I_1, V']$ and to $\partial_t u$ in $L^{p'}[0, T, V']$. Similarly, w_h converges weakly to $\partial_t \bar{u}$ in $L^{p'}[I_1, V']$ and to $\partial_t u$ in $L^{p'}[0, T, V']$.

For any $\beta \in D(I_1)$, one has that $\Psi(\cdot, \bar{u})\beta \in L^p(I_1, V)$, since $L^2(D)$ is identified with its dual, one gets that

$$\int_{I_1 \times D} v_h \bar{\Psi}(\cdot, u(t))\beta \, dx \, dt = \int_{I_1} < v_h, \bar{\Psi}(\cdot, \bar{u}(t)) > \beta dt \to \int_{I_1} < \partial_t \bar{u}, \bar{\Psi}(\cdot, \bar{u}) > \beta dt,$$
$$\int_{I_1 \times D} w_h \bar{\Psi}(\cdot, \bar{u}(t))\beta \, dx \, dt = \int_{I_1} < w_h, \bar{\Psi}(\cdot, \bar{u}(t)) > \beta dt \to \int_{I_1} < \partial_t \bar{u}, \bar{\Psi}(\cdot, \bar{u}) > \beta dt.$$

Let us recall that a is a given real and $\bar{\Lambda}(t, x, \lambda) = \int_a^\lambda \bar{\Psi}(t, x, \tau) d\tau$. Since $\bar{\Psi}$ is a non-decreasing function of its third variable, for any real numbers u and v, one has

$$(v - u)\bar{\Psi}(t, x, u) \leq \bar{\Lambda}(t, x, v) - \bar{\Lambda}(t, x, u) = \int_u^v \bar{\Psi}(t, x, \tau) d\tau \leq (v - u)\bar{\Psi}(t, x, v).$$

Thus, assuming moreover that β is non-negative,

$$[\bar{u}(t + h, x) - \bar{u}(t, x)]\bar{\Psi}(t, x, \bar{u}(t))\beta \leq [\bar{\Lambda}(t, x, \bar{u}(t + h)) - \bar{\Lambda}(t, x, \bar{u}(t))]\beta$$
$$\leq [\bar{u}(t + h, x) - \bar{u}(t, x)]\bar{\Psi}(t, x, \bar{u}(t + h))\beta,$$
$$[\bar{u}(t, x) - \bar{u}(t - h, x)]\bar{\Psi}(t, x, \bar{u}(t - h))\beta \leq [\bar{\Lambda}(t, x, \bar{u}(t)) - \bar{\Lambda}(t, x, \bar{u}(t - h))]\beta$$
$$\leq [\bar{u}(t, x) - \bar{u}(t - h, x)]\bar{\Psi}(t, x, \bar{u}(t))\beta.$$

and, for h small enough to have $\operatorname{supp}\beta + [-h, h] \subset I_1$,

$$\int_{I_1 \times D} v_h \beta \bar{\Psi}(\cdot, u(t)) \, dx \, dt \leq \int_{I_1 \times D} \frac{\bar{\Lambda}(\cdot, \bar{u}(t + h)) - \bar{\Lambda}(\cdot, \bar{u}(t))}{h} \beta \, dx \, dt$$
$$\leq \int_{I_1 \times D} v_h \beta \bar{\Psi}(\cdot, \bar{u}(t + h)) \, dx \, dt,$$
$$\int_{I_1 \times D} w_h \beta \bar{\Psi}(\cdot, \bar{u}(t - h)) \, dx \, dt \leq \int_{I_1 \times D} \frac{\bar{\Lambda}(\cdot, \bar{u}(t)) - \bar{\Lambda}(\cdot, \bar{u}(t - h))}{h} \beta \, dx \, dt$$
$$\leq \int_{I_1 \times D} w_h \beta \bar{\Psi}(\cdot, \bar{u}(t)) \, dx \, dt,$$

so that

$$\begin{aligned}\liminf \int_{I_1\times D} \frac{\bar\Lambda(\cdot,\bar u(t+h))-\bar\Lambda(\cdot,\bar u(t))}{h}\beta\,dx\,dt &\geq \int_{I_1} <\partial_t\bar u,\bar\Psi(\cdot,\bar u)>\beta dt\\ &= \int_0^T <\partial_t u,\Psi(\cdot,u)>\beta dt,\\ \limsup \int_{I_1\times D} \frac{\bar\Lambda(\cdot,\bar u(t))-\bar\Lambda(\cdot,\bar u(t-h))}{h}\beta\,dx\,dt &\leq \int_{I_1} <\partial_t\bar u,\bar\Psi(\cdot,\bar u)>\beta dt\\ &= \int_0^T <\partial_t u,\Psi(\cdot,u)>\beta dt.\end{aligned}$$

Moreover,

$$\begin{aligned}&\int_{I_1\times D} \frac{\bar\Lambda(t,x,\bar u(t+h))-\bar\Lambda(t,x,\bar u(t))}{h}\beta(t)\,dx\,dt\\ &= \frac{1}{h}\int_{I_1\times D}\bar\Lambda(t-h,x,\bar u(t))\beta(t-h)\,dx\,dt-\frac{1}{h}\int_{I_1\times D}\bar\Lambda(t,x,\bar u(t))\beta(t)\,dx\,dt\\ &= \int_{I_1\times D}\frac{\bar\Lambda(t-h,x,\bar u(t))-\bar\Lambda(t,x,\bar u(t))}{h}\beta(t-h)\,dx\,dt+\int_{I_1\times D}\frac{\beta(t-h)-\beta(t)}{h}\bar\Lambda(t,x,\bar u(t))\,dx\,dt\end{aligned}$$

and

$$\begin{aligned}&\int_{I_1\times D} \frac{\bar\Lambda(t,x,\bar u(t))-\bar\Lambda(t,x,\bar u(t-h))}{h}\beta(t)\,dx\,dt\\ &= \int_{I_1\times D}\frac{\bar\Lambda(t,x,\bar u(t))-\bar\Lambda(t+h,x,\bar u(t))}{h}\beta(t+h)\,dx\,dt+\int_{I_1\times D}\frac{\beta(t)-\beta(t+h)}{h}\bar\Lambda(t,x,\bar u(t))\,dx\,dt\end{aligned}$$

one gets, by passing to the limit, and thanks to the time-extension procedure,

$$\begin{aligned}&\liminf \int_{I_1\times D}\frac{\bar\Lambda(t-h,x,\bar u(t))-\bar\Lambda(t,x,\bar u(t))}{h}\beta(t-h)\,dx\,dt\\ &\geq \int_0^T <\partial_t u,\Psi(\cdot,u)>\beta dt+\int_{I_1\times D}\bar\Lambda(\cdot,\bar u)\beta'\,dt\\ &\geq \limsup \int_{I_1\times D}\frac{\bar\Lambda(t,x,\bar u(t))-\bar\Lambda(t+h,x,\bar u(t))}{h}\beta(t+h)\,dx\,dt\end{aligned}$$

Note that

$$\begin{aligned}&\int_{I_1\times D}\frac{\bar\Lambda(t-h,x,\bar u(t))-\bar\Lambda(t,x,\bar u(t))}{h}\beta(t-h)\,dx\,dt\\ &\qquad = -\int_{I_1\times D}\frac{1}{h}\int_{t-h}^t\partial_t\bar\Lambda(s,x,\bar u(t))\beta(t-h)ds\,dx\,dt.\end{aligned}$$

Since, $|\partial_t \bar{\Lambda}(s,x,\bar{u}(t))\beta(t-h)| \leq \|\beta\|_\infty |\bar{u}(t,x)-a|h(s,x)$ is an integrable function, the properties of the point of Lebesgue (steklov average) yields

$$\int_{I_1\times D} \frac{\bar{\Lambda}(t-h,x,\bar{u}(t))-\bar{\Lambda}(t,x,\bar{u}(t))}{h}\beta(t-h)\,dx\,dt \to -\int_{I_1\times D}\partial_t\bar{\Lambda}(t,x,\bar{u}(t))\beta(t)\,dx\,dt$$
$$= -\int_Q \partial_t\Lambda(t,x,u(t))\beta(t)\,dx\,dt.$$

Since the same holds for $\limsup \int_{I_1\times D} \frac{\bar{\Lambda}(t,x,\bar{u}(t))-\bar{\Lambda}(t+h,x,\bar{u}(t))}{h}\beta(t+h)\,dx\,dt$, and if β is regular and non negative, one gets that, for all $\beta \in D^+([0,T])$,

$$\int_0^T <\partial_t u, \Psi(\cdot,u)>\beta dt = \int_D \Lambda(T,x,u(T))\beta(T)dx - \int_D \Lambda(0,x,u_0)\beta(0)dx$$
$$-\int_Q \Lambda(\cdot,u)\beta' dt - \int_Q \partial_t\Lambda(t,x,u(t))\beta(t)\,dx\,dt.$$

Since β is involved in linear integral terms, a classical argument of regularisation yields the result for any non-negative elements of $W^{1,\infty}(0,T)$, then for any elements of $W^{1,\infty}(0,T)$.

Since T is arbitrary, the result holds for any t and $s=0$, then for any t and s by subtracting the integral from 0 to s to the one from 0 to t. □

A priori, following Lemma's 5 notations, one should denote by $\Psi(t,x,\lambda) = -(g^- - \frac{1}{\varepsilon}[\lambda^-]^{\tilde{q}-1})^-$ and $\Lambda(t,x,\lambda) = \int_0^\lambda \Psi(t,x,\sigma)d\sigma$. For that, we need $\Psi(t,x,u)$ to be a test-function. Since $x \mapsto [x^-]^{\tilde{q}-1}$ is not *a priori* a Lipschitz-continuous function (*e.g.* if $p < 2$[3]), therefore, for any positive k, we will denote by

$\eta_k(x) = (\tilde{q}-1)\int_0^{x^+} \min(k, s^{\tilde{q}-2})ds$, $\Psi_k(t,x,\lambda) = -(g^- - \frac{1}{\varepsilon}\eta_k(\lambda^-))^-$ and $\Lambda_k(t,x,\lambda) = \int_0^\lambda \Psi_k(t,x,\sigma)d\sigma$. Note that $\Psi_k(t,x,0)=0$ and $\partial_t\Psi_k(t,x,\lambda) = \partial_t g^- 1_{\{g^- - \frac{1}{\varepsilon}\eta_k(\lambda^-)<0\}}$ so that, since $\Psi_k(t,x,u)$ is a test-function, by Lemma 5, for any t,

$$-\int_0^t\int_D \partial_t\Lambda_k(s,x,u_\varepsilon-\psi)dxds + \int_D \Lambda_k(t,x,u_\varepsilon(t)-\psi(t))dx - \int_D \Lambda_k(0,x,u_\varepsilon(0)-\psi(0))dx$$
$$-\int_0^t \langle A(u_\varepsilon)-A(\psi), (g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon-\psi)^-])^-\rangle ds - \int_Q z_\varepsilon(g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon-\psi)^-])^- dxds$$
$$= -\int_0^t \langle g^+, (g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon-\psi)^-])^-\rangle ds \leq 0.$$

Remark 4. Note that the perturbation of the operator will play a main role in the study of the

[3] $\tilde{q} = \min(2,p)$

principal term. Indeed, denote by E the set $\{g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon - \psi)^-] < 0\}$

$$
\begin{aligned}
&- \int_0^T \langle A(u_\varepsilon) - A(\psi), (g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon - \psi)^-])^- \rangle dt \\
&\quad = \int_Q 1_E \Big[\tilde{a}(t,x,u_\varepsilon,\nabla u_\varepsilon) - \tilde{a}(t,x,\psi,\nabla\psi)\Big]\nabla[g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon - \psi)^-]]\,dx\,dt \\
&\quad = \int_Q 1_E \Big[\tilde{a}(t,x,\psi,\nabla u_\varepsilon) - \tilde{a}(t,x,\psi,\nabla\psi)\Big]\nabla[g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon - \psi)^-]]\,dx\,dt,
\end{aligned}
$$

therefore,

$$
\begin{aligned}
&- \int_0^T \langle A(u_\varepsilon) - A(\psi), (g^- - \frac{1}{\varepsilon}\eta_k[(u_\varepsilon - \psi)^-])^- \rangle dt \\
&\qquad\qquad \geq - \int_Q \Big|\tilde{a}(t,x,\psi,\nabla u_\varepsilon) - \tilde{a}(t,x,\psi,\nabla\psi)\Big| |\nabla g^-| 1_{\{u_\varepsilon < \psi\}}\,dx\,dt.
\end{aligned}
$$

We prove that the last term goes to zero and by analysing the other terms, we obtain Lewy-Stampacchia inequality with regular data.

Finally, we present remark concerning the uniqueness of the solution.

Remark 5. Note that the pseudomonotone assumption of the operator doesn't ensure the uniqueness of the solution. Observe that under additional assumptions on the operator a, namely a local Lipschitz continuity with respect to the third variable, standard arguments allow one to prove the uniqueness of the solution obtained in Theorem 1.

References

[1] Alt, H. W., and Luckhaus, S. Quasilinear elliptic-parabolic differential equations. *Math. Z. 183*, 3 (1983), 311–341.

[2] Bamberger, A. étude d'une équation doublement non linéaire. *J. Functional Analysis 24*, 2 (1977), 148–155.

[3] Brézis, H. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Publishing Co., New York, 1973.

[4] Donati, F. A penalty method approach to strong solutions of some nonlinear parabolic unilateral problems. *Nonlinear Analysis, Th. Meth & App. 6*, 6 (1982), 585–597.

[5] Droniou, J. Inégalité de necas et quelques applications [online]. Available from: `http://users.monash.edu.au/~jdroniou/polys/polydroniou_ineg-necas.pdf`.

[6] Hess, P. On a second-order nonlinear elliptic boundary value problem. *In Nonlinear analysis (collection of papers in honor of Erich H. Rothe) Academic Press, New York* (1978), 99–107.

[7] Lewy, H., and Stampacchia, G. On the smoothness of superharmonics which solve a minimum problem. *J. Analyse Math. 23* (1970), 227–236.

[8] Mastroeni, L., and Matzeu, M. Strong solutions for two-sided parabolic variational inequalities related to an elliptic part of p-Laplacian type. *Z. Anal. Anwend. 31*, 4 (2012), 379–391.

[9] Mokrane, A., and Murat, F. A proof of the lewy-stampacchia's inequality by a penalization method. *Potential Analysis 9* (1998), 105–142.

[10] Mokrane, A., and Vallet, G. A Lewy-Stampacchia inequality in variable sobolev spaces for pseudomonotone operators. *Differential Equations and Applications 6*, 2 (2014), 233–254.

[11] O. Guibé, A. Mokrane, Y. T., and Vallet, G. Lewy-stampacchia's inequality for a pseudomonotone parabolic problem. *Advances in Nonlinear Analysis 9* (2020), 591–612.

[12] Rodrigues, J. F. On the hyperbolic obstacle problem of first order. *Chinese Ann. Math. Ser. B 23*, 2 (2002), 253–266.

[13] Roubíček., T. *Nonlinear partial differential equations with applications*, vol. 153 of *International Series of Numerical Mathematics*. Birkhäuser Verlag, Basel, 2005.

[14] Showalter, R. E. *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*. American Mathematical Society, 1997.

Y. Tahraoui
Laboratoire de Mathématiques et leurs Applications
UMR CNRS 5142, BP 1155
64013 Pau cedex, France
tahraouiyacine@yahoo.fr

Monografías Matemáticas *García de Galdeano* **42**, 297–306 (2019)

PERIODIC SOLUTIONS FOR IMPULSIVE DIFFERENTIAL EQUATIONS

José Manuel Uzal

Abstract. In this note we present some results on the existence of periodic solutions for some impulsive differential equations. Two different problems will be considered. First, a first order differential equations with the possible presence of singularities and impulses is studied. The impulses are assumed to happen on the position and at instants of time fixed beforehand. Second, a second order differential equation is considered with state-dependent impulses at both the position and its derivative. This means that the instants of impulsive effects depend on the solutions and they are not fixed beforehand, making the study of this problem more difficult.

Keywords: impulsive differential equations, periodic solutions.
AMS classification: 34B37, 34A37.

§1. Introduction

Some evolutions processes are subject to sudden changes. The mathematical description of these processes leads to impulsive differential equations. The changes are assumed to be instantaneous, since their length is negligible in comparison with the duration of the process. Thus, solutions of impulsive differential equations are, in general, piecewise continuous functions. Furthermore, the existence of impulsive effects could cause complicated phenomena. This type of differential equations can describe population dynamics, biological phenomena or several physical situations [12]. Moreover, impulses can be introduced on a system to generate a particular dynamic (for example periodic motions) or to control a process.

There are two large classes of impulsive differential equations, with impulses at fixed times or with state-dependent impulses. On the first class, the moments of impulsive effect are known beforehand. Techniques and tools used on the classical theory of differential equations can sometimes be generalized and applied to this case rather easily. On the second case, the times of impulsive effect change depending on the solution, making its study much more difficult, because the space of solutions does not have such good properties and some solutions could have unexpected behaviors. We refer the reader to [1, 7, 11, 12] for some results and applications of the impulsive differential equations.

In this note we will study two different problems. First, we consider a first order impulsive problem with impulses at fixed times and singularities. For example, the following problem could be considered

$$\begin{aligned} &x'(t) = -\frac{1}{(x(t))^{\alpha}} + e(t), \quad t \neq t_k; \\ &\Delta x(t_k) = I_k(x(t_k)); \\ &x(0) = x(T). \end{aligned} \tag{1.1}$$

Here $\Delta x(t_k) = x(t_k^+) - x(t_k^-)$ and $x(t_k^-), x(t_k^+)$ denote the limits of x as t approaches t_k from the left and right, respectively. The main difficulty of this problem is the presence of the term $1/x^\alpha$, because it makes it difficult to find a region where the possible solutions could be located. Differential equations with singularities have been studied in recent years because they appear in a lot of physical models [13], and the introduction of impulses makes the number of applications even larger, although its study could become much more difficult.

We present a result on the existence of periodic solutions for a problem much more general that (1.1) and including a large class of nonlinearities.

The second problem considered is a more classical second order differential equation. In this case, the presence of state-dependent impulses is studied. This makes its study much more difficult. Our aim is to guarantee the existence of periodic solutions of

$$\begin{aligned} &x''(t) + g(x(t)) = p(t, x(t), x'(t)), && t \neq \gamma_i(x(t), x'(t)); \\ &x(t^+) = x(t) + I_i(x(t), x'(t)), && t = \gamma_i(x(t), x'(t)); \\ &x'(t^+) = x(t) + J_i(x(t), x'(t)), && t = \gamma_i(x(t), x'(t)); \end{aligned} \tag{1.2}$$

This type of problems is harder because the moments of impulse depend on the solution of the differential equation. For example, the equation $t = \gamma_i(x(t), x'(t))$ could have no solutions, one solution or infinitely many; and the solutions of this equation need not to depend continuously on an initial data.

The rest of this note is organized as follows: in Section 2 we state some general facts about impulsive differential equations. In Section 3 we state our existence result for problem (1.1) and in Section 4 for problem (1.2).

§2. General facts about impulsive differential equations

Let A be a subset of $\mathbb{R}^n$, $f : \mathbb{R} \times A \longrightarrow \mathbb{R}^n$, $\gamma_i : A \longrightarrow \mathbb{R}$ and $\phi_i : A \longrightarrow A$. An impulsive differential equation is an expression of the form

$$\begin{aligned} &x'(t) = f(t, x(t)), && t \neq \gamma_i(x(t)); \\ &x(t^+) = \phi_i(x(t)), && t = \gamma_i(x(t)). \end{aligned} \tag{2.1}$$

There are mainly two large classes of impulsive differential equations:

- Equations with fixed moments of impulsive effect: in this case, γ_i is a constant function, i.e., $\gamma_i(x) = t_i$. The moments of impulsive effect are fixed and they are the same for every solution. Then, (2.1) can be written as

  $$\begin{aligned} &x'(t) = f(t, x(t)), && t \neq t_i; \\ &x(t_i^+) = \phi_i(x(t_i)). \end{aligned}$$

  A problem of this type will be considered in Section 3. The solutions of this problem are piecewise continuous functions with (possible) discontinuities at t_i.

- Equations with unfixed moments of the impulsive effect: these equations have the form (2.1) with γ_i non-constant functions. The moments of the impulsive effect occur when the point (t, x) meets a "hypersurface" given by $t = \gamma_i(x)$. The points of

discontinuity depend on the solution, and sometimes solutions can not be extended over a large interval, especially if the solution intersects a hypersurface $t = \gamma_i(x)$ more than once. Therefore, it is interesting to impose some hypotheses in order to ensure that solutions intersect each hypersurface only once. For simplicity, in Section 4 we consider the case with just one hypersurface.

We state some general results for equations with unfixed moments of the impulsive effect. Consider

$$\begin{aligned} x'(t) &= f(t, x(t)), \quad t \neq \gamma(x(t)); \\ x(t^+) &= \phi(x(t)), \quad t = \gamma(x(t)). \end{aligned} \tag{2.2}$$

The following hypotheses will be needed in Section 4:

1. $f : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a continuous function and locally Lipschitz in the second variable and all the solutions of $u' = f(t, u)$ exist for all $t \in \mathbb{R}$;
2. $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a continuous function;
3. $\gamma : \mathbb{R}^n \longrightarrow \mathbb{R}$, $\gamma \in C^1(\mathbb{R}^n, \mathbb{R})$, there exist $\gamma_-, \gamma_+ \in (0, T)$ such that $\gamma_- < \gamma_+$ and $0 < \gamma_- \leq \gamma(x) \leq \gamma_+ < T \quad \forall\, x \in \mathbb{R}^n$;
4. $f(t, x) = f(t + T, x) \quad \forall\, (t, x) \in \mathbb{R} \times \mathbb{R}^n$;
5. $\gamma(x) > \gamma(\phi(x)) \quad \forall\, x \in \mathbb{R}^n$;
6. $D\gamma(x) \cdot f(t, x) < 1 \quad \forall\, (t, x) \in \mathbb{R} \times \mathbb{R}^n$.

Under these hypotheses the following lemmas hold:

Lemma 1. *For any $x_0 \in \mathbb{R}^n$, there is a unique solution $x(\cdot; 0, x_0)$ of (2.2) satisfying $x(0) = x_0$.*

Lemma 2. *For any $x_0 \in \mathbb{R}^n$, there is a unique $t_{x_0} \in (0, T)$ such that $t_{x_0} = \gamma(x(t_{x_0}))$.*

Lemma 3. *The map $\Gamma : x_0 \in \mathbb{R}^n \longrightarrow t_{x_0} \in (0, T)$ is continuous.*

Lemma 4. *The map $P : x_0 \in \mathbb{R}^n \longrightarrow x(T; x_0) \in \mathbb{R}^n$ is continuous.*

Proof. Let

$$\begin{aligned} &f_1 : \zeta \in \mathbb{R}^n \longrightarrow (\zeta, \zeta) \in \mathbb{R}^n \times \mathbb{R}^n; \\ &f_2 : (\zeta, \sigma) \in \mathbb{R}^n \times \mathbb{R}^n \longrightarrow (t_\zeta, \sigma) \in [0, T] \times \mathbb{R}^n; \\ &f_3 : (t, \zeta) \in [0, T] \times \mathbb{R}^n \longrightarrow (t, x(t; 0, \zeta)) \in [0, T] \times \mathbb{R}^n; \\ &f_4 : (t, \zeta) \in [0, T] \times \mathbb{R}^n \longrightarrow (t, \varphi(\zeta)) \in [0, T] \times \mathbb{R}^n; \\ &f_5 : (t, \zeta) \in [0, T] \times \mathbb{R}^n \longrightarrow x(T; t, \zeta) \in \mathbb{R}^n. \end{aligned}$$

Each of these functions is continuous and $P(x_0) = (f_5 \circ f_4 \circ f_3 \circ f_2 \circ f_1)(x_0)$. □

§3. A first order problem

In this section we study the existence of a T-periodic solution of

$$x'(t) = -\frac{1}{(x(t))^\alpha} + e(t) \tag{3.1}$$

under impulsive effects

$$\Delta x(t_k) = x(t_k^+) - x(t_k^-) = I_k(x(t_k)), \tag{3.2}$$

with $\alpha > 0$, $T > 0$, $e : \mathbb{R} \longrightarrow \mathbb{R}$ continuous and T-periodic, $I_k : \mathbb{R} \longrightarrow \mathbb{R}$ continuous and $0 = t_0 < t_1 < \cdots < t_k < t_{k+1} = T$. We reduce the previous problem to a boundary value problem, so by a T-periodic solution of (3.1)–(3.2) we understand a piecewise continuous function $u : [0, T] \longrightarrow (0, \infty)$, with discontinuities at the points t_k, $u(0) = u(T)$ and satisfying (3.1) and (3.2).

Instead of considering (3.1)–(3.2), we are going to study

$$\begin{aligned} &x'(t) = f(x(t)) + e(t), \quad t \neq t_k; \\ &\Delta x(t_k) = I_k(x(t_k)); \\ &x(0) = x(T). \end{aligned} \tag{3.3}$$

In this case, $f : (0, \infty) \longrightarrow (a, b)$ is a continuous function with $a \in [-\infty, \infty)$ and $b \in (-\infty, \infty]$.

In order to prove the existence of periodic solutions [8] we use a classical result due to Mawhin [5]. We briefly present some definitions and results.

Definition 1. Let X and Y be two normed vector spaces and consider the linear mapping $L : D(L) \subset X \longrightarrow Y$. L is called a Fredholm mapping of index 0 if $\mathrm{Im}(L)$ is a closed subset of Y and $\dim(\ker(L)) = \mathrm{codim}(\mathrm{Im}(L)) < \infty$.

If L is a Fredholm mapping of index 0, there exist two projectors $P : X \longrightarrow X$ and $Q : Y \longrightarrow Y$ such that $\mathrm{Im}(P) = \ker(L)$ and $\mathrm{Im}(L) = \ker(Q) = \mathrm{Im}(I - Q)$. This implies that $L_{|D(L)\cap\ker(P)} : (I - P)X \longrightarrow \mathrm{Im}(L)$ is an invertible map, and its inverse will be denoted by K_P.

Definition 2. Let $N : X \longrightarrow Y$ be a continuous map between two normed spaces and Ω an open bounded subset of X. We say that N is L-compact on $\overline{\Omega}$ if $QN(\overline{\Omega})$ is bounded and $K_P(I - Q)N : \overline{\Omega} \longrightarrow X$ is a compact map.

Theorem 5. *Let X and Y be two Banach spaces, $L : D(L) \subset X \longrightarrow Y$ a Fredholm mapping of index 0, Ω an open bounded subset of X and $N : \overline{\Omega} \subset X \longrightarrow Y$ L-compact on $\overline{\Omega}$. Suppose that*

1. $Lx \neq \lambda Nx \quad \forall\, x \in \partial\Omega \cap D(L), \quad \forall\, \lambda \in (0, 1)$;
2. $QNx \neq 0 \quad \forall\, x \in \partial\Omega \cap \ker(L)$;
3. $\deg(JQN, \Omega \cap \ker(L), 0) \neq 0$, *where* $J : \mathrm{Im}(Q) \longrightarrow \ker(L)$ *is an isomorphism and* deg *represents the Brouwer's degree.*

Then the equation $Lx = Nx$ has at least one solution in $D(L) \cap \overline{\Omega}$.

We introduce the following hypotheses:

(H1) $\lim\limits_{s\to 0^+} f(s) = a^+$, $\lim\limits_{s\to\infty} f(s) = b^-$.

(H2) There exist $m_k, M_k \in \mathbb{R}$ such that $m_k \leq I_k(s) \leq M_k \quad \forall\, s > 0$.

(H3) If $c_1 = \dfrac{-m_1 - \cdots - m_q}{T} - \dfrac{1}{T}\displaystyle\int_0^T e(t)\,dt$ and $c_2 = \dfrac{-M_1 - \cdots - M_q}{T} - \dfrac{1}{T}\displaystyle\int_0^T e(t)\,dt$, then $c_1, c_2 \in (a, b)$.

(H4) For $\widetilde{M}_k = \max\{|M_k|, |m_k|\}$ and $r_2 = \inf\{s > 0 : f(s) \geq c_2\}$, it holds that

$$r_2 - 2\left(\widetilde{M}_1 + \cdots + \widetilde{M}_q\right) - \int_0^T e(t) + |e(t)|\,dt > 0.$$

We define

$$\begin{aligned}
X &= \{x : [0,T] \longrightarrow \mathbb{R} \mid x(0) = x(T), x \text{ continuous except at } t_k,\\
&\qquad \text{there exist } x(t_k^-), x(t_k^+) \text{ and } x(t_k) = x(t_k^-)\};\\
\|x\| &= \sup\{|x(t)| : t \in [0,T]\}\\
Y &= X \times \mathbb{R}^q;\\
Lx &= (g_1, \Delta x(t_1), \cdots, \Delta x(t_q)), \text{ with } g_1(t) = x'(t);\\
Nx &= (g_2, I_1(x(t_1)), \cdots, I_q(x(t_q))), \text{ with } g_2(t) = f(x(t)) + e(t).
\end{aligned}$$

Lemma 6. *Suppose that hypotheses (H1)–(H4) are satisfied. Then there exist two positive constants A_1 and A_2 such that $A_2 \leq x(t) \leq A_1$ for all $t \in [0,T]$, and for x any solution of the equation $Lx = \lambda Nx$, $\lambda \in (0,1]$. The constants A_1 and A_2 are independent of λ.*

Proof. Let $x \in X$ with $\min\{x(t) : t \in [0,T]\} > 0$ such that there exists $\lambda \in (0,1)$ with

$$\begin{cases} x'(t) = \lambda f(x(t)) + \lambda e(t), \ t \in [0,T], \ t \neq t_k; \\ \Delta x(t_k) = \lambda I_k(x(t_k)), \ k \in \{1, \ldots, q\}. \end{cases}$$

Integrating over $[0,T]$ we obtain

$$\int_0^T x'(t)\,dt = \lambda \int_0^T f(x(t))\,dt + \lambda \int_0^T e(t)\,dt.$$

The first integral is equal to

$$\int_0^T x'(t)\,dt = \sum_{k=1}^{q+1} \int_{t_{k-1}^+}^{t_k^-} x'(t)\,dt = -x(0) - \sum_{k=1}^{q} (x(t_k^+) - x(t_k^-)) + x(T) = -\lambda \sum_{k=1}^{q} I_k(x(t_k)).$$

We can deduce

$$(m_1 + \cdots + m_q) + \int_0^T e(t)\,dt \leq \int_0^T -f(x(t))\,dt \leq (M_1 + \ldots + M_q) + \int_0^T e(t)\,dt.$$

by using hypothesis (H2). We obtain that there exist $\xi, \eta \in [0,T] \setminus \{t_1, \ldots, t_q\}$ such that

$$-Tf(x(\xi)) \leq (M_1 + \cdots + M_q) + \int_0^T e(t)\,dt, \qquad -Tf(x(\eta)) \geq (m_1 + \cdots + m_q) + \int_0^T e(t)\,dt$$

by using the mean value theorem for definite integrals. Then $f(x(\xi)) \geq c_2$ and $f(x(\eta)) \leq c_1$, so there exist $r_1, r_2 > 0$ such that $x(\xi) \geq r_2$ and $x(\eta) \leq r_1$ (r_2 as defined by (H4) and r_1 analogously). It can be checked that

$$x(t) \leq r_1 + \left(\widetilde{M}_1 + \cdots + \widetilde{M}_q\right) + \int_0^T |x'(u)|\,du, \quad x(t) \geq r_2 - \left(\widetilde{M}_1 + \cdots + \widetilde{M}_q\right) - \int_0^T |x'(u)|\,du$$

Furthermore,

$$\int_0^T |x'(t)|\,dt \le \widetilde{M}_1 + \cdots + \widetilde{M}_q + \int_0^T e(t)\,dt + \int_0^T |e(t)|\,dt$$

This implies that

$$A_2 := r_2 - 2\sum_{k=1}^{q} \widetilde{M}_k - \int_0^T e(t) + |e(t)|\,dt \le x(t) \le A_1 := r_1 + 2\sum_{k=1}^{q} \widetilde{M}_k + \int_0^T e(t) + |e(t)|\,dt$$

and $A_2 > 0$ by hypothesis (H4). Therefore the lemma is proved. □

Theorem 7. *Suppose that hypotheses (H1)–(H4) are satisfied. Then problem* (3.3) *has at least one solution.*

Proof. We define

$$\Omega = \{x \in X : \min\{x(t) : t \in [0,T]\} > A_2 - \sigma_2,\ A_2 - \sigma_2 < \|x\| < A_1 + \sigma_1\},$$

with $0 < \sigma_2 < A_2$ and $\sigma_1 > 0$ two constants. The set Ω is bounded and open, $QN(\overline{\Omega})$ is bounded and $(K_P(I-Q)N)(\overline{\Omega})$ is relatively compact. Furthermore, for each $\lambda \in (0,1)$

$$Lx = \lambda Nx \Longrightarrow A_2 \le x(t) \le A_1 \quad \forall\, t \in [0,T] \Longrightarrow x \notin \partial\Omega.$$

Define $J : (b,0,\ldots,0) \in \mathrm{Im}(Q) \longrightarrow b \in \ker(L)$ an isomorphism. We must prove that $QNx \neq 0$ for every $x \in \partial\Omega \cap \ker(L)$ and $\deg(JQN, \Omega \cap \ker(L), 0)$ is not equal to 0.

Let $x \in \ker(L)$ with $QNx = 0$. We must check that $x \notin \partial\Omega$.

$$QNx = 0 \Longrightarrow \frac{1}{T}\int_0^T \left[f(x(t)) + e(t)\right] dt + \frac{1}{T}\sum_{k=1}^{q} I_k(x(t_k)) = 0,$$

$$x \in \ker(L) \Longrightarrow x \text{ constant} \Longrightarrow x(t) = x(0) \quad \forall\, t \in [0,T].$$

We obtain from the previous equations that

$$-f(x(0)) = \frac{1}{T}\int_0^T e(t)\,dt + \frac{1}{T}\sum_{k=1}^{q} I_k(x(0)).$$

This implies that $c_2 \le f(x(0)) \le c_1$ by hypothesis (H3). Then we can conclude the following: $A_2 - \sigma_2 < A_2 \le r_2 \le x(t) \le r_1 \le A_1 < A_1 + \sigma_1$, which implies that $x \notin \partial\Omega$.

We identify $\ker(L) \cap \Omega$ with the interval $(A_2 - \sigma_2, A_1 + \sigma_1)$ of $\mathbb{R}$. Then the degree of JQN in $\Omega \cap \ker(L)$ with respect to 0 is $\deg(\varphi, (p,q), 0)$, where $(p,q) = (A_2 - \sigma_2, A_1 + \sigma_1)$ and the function $\varphi : [p,q] \longrightarrow \mathbb{R}$ is given by

$$\varphi(x) = f(x) + \frac{1}{T}\int_0^T e(t)\,dt + \frac{1}{T}\sum_{k=1}^{q} I_k(x).$$

It can be proved that $\varphi(p) < 0 < \varphi(q)$. Then $\deg(\varphi, (p,q), 0) \neq 0$ using properties of the Brouwer's degree. Therefore we can use Theorem 5, so there exists $x \in D(L) \cap \overline{\Omega}$ such that $Lx = Nx$, which implies that the impulsive boundary value problem (3.3) has one positive T-periodic solution. □

§4. A second order problem

In this section we study the following second order differential equation

$$x''(t) + g(x(t)) = p(t, x(t), x'(t)) \tag{4.1}$$

with $g : \mathbb{R} \longrightarrow \mathbb{R}$ continuous and $p : \mathbb{R}^3 \longrightarrow \mathbb{R}$ bounded, continuous and T-periodic on the first variable. We consider state-dependent impulses

$$\begin{aligned} x(t^+) &= x(t) + I_i(x(t), x'(t)) \\ x'(t^+) &= x(t) + J_i(x(t), x'(t)) \end{aligned} \tag{4.2}$$

when $t = \gamma_i(x(t), x'(t))$, $i \in \{1, \dots, q\}$. Here $I_i, J_i, \gamma_i \in C(\mathbb{R}^2, \mathbb{R})$. For simplicity we consider $q = 1$.

There are few existence results for periodic problems with state-dependent impulses, some examples include [2, 4, 10, 11].

In order to prove the existence of periodic solutions, the idea is to reduce (4.1)–(4.2) to a first order planar system and to consider the map P defined in Lemma 4. Then a T-periodic solution would be a fixed point of P.

The first idea was to use Poincaré-Birkhoff fixed point theorem, which states that every area-preserving, orientation-preserving homeomorphism of an annulus that rotates the boundaries in opposite directions has at least two fixed points. There are some extensions of this result. It has been applied to second-order problems and to second order problems with impulses at fixed times (see [3, 6, 9]). We were not able to apply it to our problem other than in some trivial cases.

Consider the following simplification of a partial extension stated in [9]:

Theorem 8. *Let Γ_- and Γ_+ be two closed and convex curves surrounding the origin,* $\operatorname{int}(\Gamma_+)$ *the interior of Γ_+ in the sense of Jordan curve theorem, $\mathcal{A}$ the annulus bounded by Γ_- and Γ_+ and $F : \overline{\operatorname{int}(\Gamma_+)} \longrightarrow \mathbb{R}^2$ a continuous map. We denote*

$$E = \{z \in \mathcal{A} : |F(z)| \leq |z|\},$$

with $U(O)$ a neighborhood of the origin and L a real orthogonal matrix with $\det(L) = 1$.

If $\gamma : [a, b] \longrightarrow \mathbb{R}^2$ is a curve connecting Γ_- and Γ_+ and $\gamma([a, b]) \cap (J \cup E)$ is nonempty, then F has at least one fixed point.

The proof of this result is based on Brouwer's degree and some of its properties. We apply this result to our problem. First, we define a very important family of maps from $\mathbb{R}^2$ to $\mathbb{R}^2$, which will be fundamental in the proof of our result.

Definition 3. Let $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be a continuous map. We say that the map F has the property of partial boundedness if there is a bounded set $D \subset \mathbb{R}^2$, a convex cone and a curve $\Gamma : \lambda \in [0, \infty) \longrightarrow (x(\lambda), y(\lambda)) \in \mathbb{R}^2$, contained in the cone, such that

$$\lim_{\lambda \to \infty} (|x(\lambda)| + |y(\lambda)|) = +\infty \qquad \text{and} \qquad (F \circ \Gamma)([0, \infty)) \subset D.$$

The associated first-order planar system of (4.1) is

$$\begin{aligned} x' &= y \\ y' &= -g(x) + p(t, x, y). \end{aligned} \tag{4.3}$$

This system has been widely studied. Suppose that g and p are locally Lipschitz and

$$\lim_{x\to+\infty} g(x) = +\infty; \quad \lim_{x\to-\infty} g(x) = -\infty. \tag{g_0}$$

Consider the autonomous differential equation $x'' + g(x) = 0$ and define

$$G(x) = \int_0^x g(s)\, ds.$$

We have that G is bounded from below, has an absolute minimum and $G(x)$ goes to ∞ as $|x| \longrightarrow \infty$.

It can be checked that if $\zeta > 0$ is large enough and $x(0) = \zeta$, $y(0) = 0$, then the solutions are periodic and the least period of this solution, $\tau(\zeta)$, is given by the expression

$$\tau(\zeta) = \sqrt{2} \int_{h(\zeta)}^{\zeta} \frac{1}{\sqrt{c - G(y)}}\, dy,$$

where $h(\zeta)$ is a negative number such that $G(h(\zeta)) = G(\zeta)$ and $c = G(\zeta)$. Assume the following hypothesis:

$$\lim_{\zeta\to+\infty} \tau(\zeta) = 0. \tag{τ_0}$$

We can use polar coordinates on (4.3) for sufficiently large r, so

$$\begin{cases} \theta' = -\sin^2\theta - \dfrac{(g(r\cos\theta) - p(t, r\cos\theta, r\sin\theta))\cos\theta}{r}, \\ r' = r\cos\theta\sin\theta - (g(r\cos\theta) - p(t, r\cos\theta, r\sin\theta))\sin\theta. \end{cases} \tag{4.4}$$

Given an initial condition (r_0, θ_0), with r_0 sufficiently large, let $(r(t; r_0, \theta_0), \theta(t; r_0, \theta_0))$ be the solution of (4.4) verifying the initial condition (r_0, θ_0) at time $t = 0$.

The proof of the following lemma is a consequence of results that can be found on [3].

Lemma 9. *Suppose* (g_0) *is satisfied. Then there exists* $d > 0$ *sufficiently large such that*

$$r_0 > d \Longrightarrow \frac{d}{dt}\theta(t; r_0, \theta_0) < 0 \quad \forall\, \theta_0 \in \mathbb{R}.$$

Furthermore, there exists a continuous and non-decreasing function β *from* $[d, \infty)$ *to* $(0, \infty)$ *such that*

$$r_0 > d \Longrightarrow \frac{d}{dt}\theta(t; r_0, \theta_0) \geq -\beta(r_0) \quad \forall\, \theta \in \mathbb{R}.$$

Define $n_*(r, t)$ and $n^*(r, t)$ as the two non-negative integers such that for any solution of (4.3) with initial values $\sqrt{x(0)^2 + y(0)^2} = r$, the solution makes at least $n_*(r, t)$ and at most $n^*(r, t)$ turns around the origin on the interval $[0, t]$.

Lemma 10. *Let $t \in (0, T]$. If (g_0) and (τ_0) are satisfied, then*

$$\lim_{r\to\infty} n_*(r, t) = +\infty.$$

Lemma 11. *Suppose (g_0) and (τ_0) are satisfied and $t \in (0, T]$. Then*

$$\forall\, N \in \mathbb{N},\ \exists\, \rho > 0 : r_0 > \rho \Longrightarrow \theta(t; r_0, \theta_0) - \theta_0 < -2N\pi \quad \forall\, \theta_0 \in \mathbb{R}.$$

Next, we state and prove the existence of T-periodic solutions for problem (4.1)–(4.2).

Theorem 12. *Suppose (g_0), (τ_0) and the six hypotheses in Section 2 are satisfied and let*

$$\phi : (x, y) \in \mathbb{R}^2 \longrightarrow (x + I_1(x, y), y + J_1(x, y)) \in \mathbb{R}^2.$$

If ϕ has the property of partial boundedness, then P has at least one fixed point, that is, there exists at least one T-periodic solution of

$$\begin{aligned} &x''(t) + g(x(t)) = p(t, x(t), x'(t)), && t \neq \gamma(x(t), x'(t));\\ &x(t^+) = x(t) + I_1(x(t), x'(t)), && t = \gamma(x(t), x'(t));\\ &x'(t^+) = x(t) + J_1(x(t), x'(t)), && t = \gamma(x(t), x'(t)). \end{aligned} \tag{4.5}$$

Proof. There exist D a compact subset $\mathbb{R}^2$, a convex cone and a curve Γ starting at the origin, $\Gamma : \lambda \in [0, \infty) \longrightarrow (x(\lambda), y(\lambda)) \in \mathbb{R}^2$ contained in the cone such that

$$\lim_{\lambda\to\infty} (|x(\lambda)| + |y(\lambda)|) = +\infty \qquad \text{and} \qquad (\phi \circ \Gamma)([0, \infty)) \subset D.$$

The function f_5 as defined in the proof of Lemma 4 is also continuous. There exists $M_D > 0$ such that $|f_5(t, x)| \leq M_D$ for all $(t, x) \in [\gamma_-, \gamma_+] \times D$.

Take R_1 and R_2 sufficiently large with $R_2 > R_1 > M_D$ and satisfying

$$\theta(\gamma_+; \theta_0, R_1) - \theta(0; \theta_0, R_1) > -a, \qquad \theta(\gamma_-; \theta_0, R_2) - \theta(0; \theta_0, R_2) < -a - 4\pi.$$

for some $a > 0$. We can restrict ourselves to $\theta \in [0, 2\pi]$. Then we have that

$$\theta(t_{(\theta_0,R_2)}; \theta_0, R_2) - \theta(t_{(\theta_1,R_1)}; \theta_1, R_1) < -2\pi \tag{4.6}$$

for $\theta_0, \theta_1 \in [0, 2\pi)$, with $t_{(\theta_i,R_j)}$ the unique impulsive point given by Lemma 2. Take the curves $C_i = \{z \in \mathbb{R}^2 : |z| = R_i\}$, $i \in \{1, 2\}$. In order to use Theorem 8, let $\beta : I \subset \mathbb{R} \longrightarrow \mathbb{R}^2$ a curve connecting C_1 and C_2, with z_1 and z_2 its initial and final points. The associated curve $\tilde{\beta}(t) = (P_2 \circ f_3 \circ f_2 \circ f_1 \circ \beta)(t)$ makes at least one turn around the origin because of (4.6), where $P_2 : [0, T] \times \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ denotes the projection. So the curves $\tilde{\beta}$ and Γ intersect at least in one point. Let z be that point. Then $\phi(z) \in D$ and furthermore $|P(z)| \leq M_D < |z|$. This implies that the map P satisfies the hypotheses of Theorem 8, so $P : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ has at least one fixed point, which implies that there exists a T-periodic solution of (4.5). □

References

[1] Bainov, D., and Simeonov, P. *Impulsive Differential Equations: Periodic Solutions and Applications*. CRC Press, 1993.

[2] Bajo, I., and Liz, E. Periodic boundary value problem for first order differential equations with impulses at variable times. *J. Math. Anal. Appl. 204*, 1 (1996), 65–73. doi:10.1006/jmaa.1996.0424.

[3] Ding, T. R., and Zanolin, F. Periodic solutions of Duffing's equations with superquadratic potential. *J. Differ. Equ. 97*, 2 (1992), 328–378. doi:10.1016/0022-0396(92)90076-Y.

[4] Frigon, M., and O'Regan, D. First order impulsive initial and periodic problems with variable moments. *J. Math. Anal. Appl. 233*, 2 (1999), 730–739. doi:10.1006/jmaa.1999.6336.

[5] Gaines, R. E., and Mawhin, J. L. *Coincidence Degree and Nonlinear Differential Equations*. Springer, 1977. doi:10.1007/BFb0089537.

[6] Jacobowitz, H. Periodic solutions of $x'' + f(x,t) = 0$ via the Poincaré-Birkhoff theorem. *Journal of Differential Equations 20*, 1 (1976), 37–52. doi:10.1016/0022-0396(76)90094-2.

[7] Lakshmikantham, V., Bainov, D. D., and Simeonov, P. S. *Theory of Impulsive Differential Equations*, vol. 6. World scientific, 1989. doi:10.1142/0906.

[8] Nieto, J. J., and Uzal, J. M. Pulse positive periodic solutions for some classes of singular nonlinearities. *Appl. Math. Lett. 86* (2018), 134–140. doi:10.1016/j.aml.2018.06.025.

[9] Qian, D., Chen, L., and Sun, X. Periodic solutions of superlinear impulsive differential equations: a geometric approach. *J. Differ. Equ. 258*, 9 (2015), 3088–3106. doi:10.1016/j.jde.2015.01.003.

[10] Rachůnková, I., and Tomeček, J. *State-Dependent Impulses: Boundary Value Problems on Compact Interval*, vol. 6. Springer, 2015.

[11] Samoilenko, A. M., and Perestyuk, N. A. *Impulsive Differential Equations*, vol. 14 of *World Scientific Series on Nonlinear Science.* World Scientific Publishing, 1995. doi:10.1142/9789812798664.

[12] Stamova, I., and Stamov, G. *Applied Impulsive Mathematical Models*. CMS Books in Mathematics. Springer International Publishing, 2016. doi:10.1007/978-3-319-28061-5.

[13] Torres, P. J. *Mathematical models with singularities*, vol. 1 of *Atlantis Briefs in Differential Equations*. Atlantis Press, Paris, 2015.

J. M. Uzal
Departamento de Estatística, Análise Matemática e Optimización
Universidade de Santiago de Compostela
josemanuel.uzal@rai.usc.es

MONOGRAFÍAS DEL SEMINARIO MATEMÁTICO GARCÍA DE GALDEANO

Desde 2001, el Seminario ha retomado la publicación de la serie *Monografías* en un formato nuevo y con un espíritu más ambicioso. El propósito es que en ella se publiquen tesis doctorales dirigidas o elaboradas por miembros del Seminario, actas de congresos en cuya organización participe o colabore el Seminario, y monografías en general. En todos los casos, se someten al sistema habitual de arbitraje anónimo.

Los manuscritos o propuestas de publicaciones en esta serie deben remitirse a alguno de los miembros del Comité editorial. Los trabajos pueden estar redactados en español, francés o inglés.

Las monografías son recensionadas en *Mathematical Reviews* y en *Zentralblatt MATH*.

Últimos volúmenes de la serie:

21. A. Elipe y L. Floría (eds.): *III Jornadas de Mecánica Celeste*, 2001, ii + 202 pp., ISBN: 84-95480-21-2.

22. S. Serrano Pastor: *Modelos analíticos para órbitas de satélites artificiales de tipo quasi-spot*, 2001, vi + 76 pp., ISBN: 84-95480-35-2.

23. M. V. Sebastián Guerrero: *Dinámica no lineal de registros electrofisiológicos*, 2001, viii + 251 pp., ISBN: 84-95480-43-3.

24. Pedro J. Miana: *Cálculo funcional fraccionario asociado al problema de Cauchy*, 2002, 171 pp., ISBN: 84-95480-57-3.

25. Miguel Romance del Río: *Problemas sobre Análisis geométrico convexo*, 2002, xvii + 214 pp., ISBN: 84-95480-76-X.

26. Renato Álvarez-Nodarse: *Polinomios hipergeométricos y q-polinomios*, 2003, vi + 341 pp., ISBN: 84-7733-637-7.

27. M. Madaune-Tort, D. Trujillo, M. C. López de Silanes, M. Palacios y G. Sanz (eds.): *VII Jornadas Zaragoza-Pau de Matemática Aplicada y Estadística*, 2003, xxvi + 523 pp., ISBN: 84-96214-04-4.

28. Sergio Serrano Pastor: *Teorías analíticas del movimiento de un satélite artificial alrededor de un planeta. Ordenación asintótica del potencial en el espacio fásico*, 2003, 164 pp., ISBN: 84-7733-667-9.

29. Pilar Bolea Catalán: *El proceso de algebrización de organizaciones matemáticas escolares*, 2003, 260 pp., ISBN: 84-7733-674-1.

30. Natalia Boal Sánchez: *Algoritmos de reducción de potencial para el modelo posinomial de programación geométrica*, 2003, 232 pp., ISBN: 84-7733-667-9.

31. M. C. López de Silanes, M. Palacios, G. Sanz, J. J. Torrens, M. Madaune-Tort y D. Trujillo (eds.): *VIII Journées Zaragoza-Pau de Mathématiques Appliquées et de Statistiques*, 2004, xxvi + 578 pp., ISBN: 84-7733-720-9.

32. Carmen Godés Blanco: *Configuraciones de nodos en interpolación polinómica bivariada*, 2006, xii + 163 pp., ISBN: 84-7733-841-9.

33. M. Madaune-Tort, D. Trujillo, M. C. López de Silanes, M. Palacios, G. Sanz y J. J. Torrens (eds.): *Ninth International Conference Zaragoza-Pau on Applied Mathematics and Statistics*, 2006, xxxii + 440 pp., ISBN: 84-7733-871-X.

34. B. Lacruz, F. J. López, P. Mateo, C. Paroissin, A. Pérez-Palomares y G. Sanz (eds.): *Pyrenees International Workshop on Statistics, Probability and Operations Research, SPO 2007*, 2008, 205 pp., ISBN: 978-84-92521-18-0.

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.

37. J. Giacomoni, M. Madaune-Tort, C. Paroissin, G. Vallet, M. C. López de Silanes, M. Palacios, G. Sanz, J. J. Torrens, (eds.): *Eleventh International Conference Zaragoza-Pau on Applied Mathematics and Statistics*, 2012, xxvi+208 pp., ISBN: 978-84-15538-15-8.

38. L. M. Esteban, B. Lacruz, F. J. López, P. M. Mateo, A. Pérez-Palomares, G. Sanz, C. Paroissin, (eds.): *The Pyrenees International Workshop and Summer School on Statistics, Probability and Operations Research SPO 2011*, 2013, 132 pp., ISBN: 978-84-15770-81-7.

39. M. C. López de Silanes, M. Palacios, G. Sanz, É. Ahusborde y C. Amrouche, (eds.): *Twelfth International Conference Zaragoza-Pau on Mathematics*, 2014, xxx + 224 pp., ISBN: 978-84-16028-35-1.

40. É. Ahusborde, C. Amrouche, G. Warnault, M. C. López de Silanes, M. Palacios y G. Sanz, (eds.): *Thirteenth International Conference Zaragoza-Pau on Mathematics and its Applications*, 2016, xxvi + 134 pp., ISBN: 978-84-16515-68-4.

41. M. C. López de Silanes, M. Palacios, É. Ahusborde, C. Amrouche y G. Carbou, (eds.): *Fourteenth International Conference Zaragoza-Pau on Mathematics and its Applications*, 2017, xxiii + 222 pp., ISBN: 978-84-17358-00-6.