Mathematics in Industry 22
The European Consortium for Mathematics in Industry

Giovanni Russo
Vincenzo Capasso
Giuseppe Nicosia
Vittorio Romano *Editors*

Progress in Industrial Mathematics at ECMI 2014

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Springer

MATHEMATICS IN INDUSTRY **22**

THE EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

More information about this series at http://www.springer.com/series/4650

Giovanni Russo • Vincenzo Capasso •
Giuseppe Nicosia • Vittorio Romano
Editors

Progress in Industrial Mathematics at ECMI 2014

Springer

ECMI
EUROPEAN CONSORTIUM FOR
MATHEMATICS IN INDUSTRY

Editors
Giovanni Russo
University of Catania
Department of Mathematics
and Computer Science
Catania, Italy

Vincenzo Capasso
University of Milan
Department of Mathematics
Milan, Italy

Giuseppe Nicosia
University of Catania
Department of Mathematics
and Computer Science
Catania, Italy

Vittorio Romano
University of Catania
Department of Mathematics
and Computer Science
Catania, Italy

Photo of Angelo Marcello Anile by courtesy of Giovanni Russo

ISSN 1612-3956 ISSN 2198-3283 (electronic)
Mathematics in Industry
The European Consortium for Mathematics in Industry
ISBN 978-3-319-23412-0 ISBN 978-3-319-23413-7 (eBook)
DOI 10.1007/978-3-319-23413-7

Library of Congress Control Number: 2017950222

Mathematics Subject Classification (2010): 34-xx, 35-xx, 41-xx, 49-xx, 60-xx, 62-xx, 65-xx, 68-xx, 70-xx, 74-xx, 86-xx, 90-xx, 91-xx, 92-xx, 94-xx, 97-xx

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedication

This book is dedicated to the memory of Professor Angelo Marcello Anile (1948–2007). Worldwide known mathematician, genial scientist, professor dedicated to the creation and spreading of knowledge, gifted with exceptional human qualities, and highest moral. He produced original results in several fields, among which relativistic astrophysics and cosmology, relativistic thermodynamics, mathematical modeling of semiconductors, wave propagation, fuzzy logic, optimization, industrial mathematics. Leader in scientific research, he founded a school of Mathematical Physics with several students. Contributed to support research also with national and international projects, and various forms of contracts. Helped find job opportunities for many young people. Active member of ECMI, created and strengthened relationship between University and Industry in Italy and abroad.

Preface

This volume presents a snapshot of current activities in industrial mathematics in Europe and will be highly relevant for anyone interested in the latest applications of mathematics to industrial problems. It features papers based on contributions to the 18th conference of ECMI, the European Consortium for Mathematics in Industry. The biannual conference has established itself as a jour fixe for researchers interested in the applications of mathematical and computational methods to relevant problems in many different areas of social and economic importance. It brings together applied mathematicians and experts from industry, offering them a unique opportunity to exchange ideas, problems, and methodologies and bridging the gap between mathematics and industry to further the advancement of science and technology.

The conference focused on various aspects of industrial and applied mathematics, such as aerospace, information and communication, energy, imaging, medicine and biotechnology, finance, and education.

The event continued for five full days. In addition to the nine plenary speakers, the conference hosted 56 mini-symposia, each of which consisted of one or more 2-h sessions. Six parallel sessions were held at the same time, allowing roughly three hundred and fifty speakers to present the results of their research. A 2-h poster session was also held, offering participants an alternate form of presentation. The participation of young researchers was encouraged by a reduced fee with respect to previous years.

Several topics were discussed at the conference. Specifically, five main topics were addressed by the nine plenary speakers and discussed in several mini-symposia: the life sciences, material science and semiconductors, the environment, design automation, and finance.

Several other interesting mini-symposia on different fields of industrial mathematics enriched the program. A complete list can be found on the conference website.[1]

[1] http://www.taosciences.it/ecmi2014/.

The conference also included an extended mini-symposium organized by EU-Maths-In (European Network of Mathematics for Industry and Innovation),[2] a recent joint initiative of ECMI and EMS (the European Mathematical Society). In keeping with tradition, this year several prizes and awards were announced during the conference.

Alan Tayler Lecture—This lecture was set up by the ECMI Council to honor Alan Tayler, who passed away on January 28, 1995. Alan was one of the founding members of ECMI and served as its third president in 1989. The Alan Tayler Lecture has been a key feature of the biannual ECMI conferences since 1996 and was delivered this year by Sylvie Méléard.[3]

Anile-ECMI Prize for Mathematics in Industry[4]—This prize is cosponsored by the Associazione Angelo Marcello Anile and ECMI. The prize was established in memory of Angelo Marcello Anile, a brilliant astrophysicist and applied mathematician and a highly active member of ECMI; it is awarded to a young researcher for an excellent PhD thesis in industrial mathematics. This year it has been awarded to Dr. Paolo Pintus, from Scuola Superiore Sant'Anna of Pisa, in recognition of the excellent results achieved in his PhD thesis "Design of silicon based integrated optical devices using the finite element method."

Hansjörg Wacker Memorial Prize—This prize is cosponsored by ECMI and RICAM (the Johann Radon Institute for Computational and Applied Mathematics[5]) at the Austrian Academy of Sciences (ÖAW). This year it was awarded to Kishan Patel, who completed his MSc in scientific computing and mathematical modeling at the University of Oxford in 2012.

ECMI Honorary Membership—It is a tradition that ECMI, during its biannual conference, offers an honorary lifetime membership to a scientist of the hosting country, in recognition of his or her important contributions to Mathematics in Industry. This year the honorary membership was awarded to Professor Alfio Quarteroni,[6] in recognition of his important contributions to Mathematics in Industry.

With over four hundred participants, this year was one of the most successful ECMI conferences to date. The number of authors involved in the organization of the mini-symposia and preparation of abstracts and short papers totaled more than 800 people from 41 different countries. The authors came mainly, but not exclusively, from Europe, with the largest group from Germany (167), followed by Italy (152), Spain (79), the UK (69), and France (65).

Given the large variety of subjects, it has been difficult to classify the contributions into broader topics. Therefore, we decided to merely draw a distinction between short papers related to talks at the mini-symposia (chapter "MS 1 Minisymposium: Advanced Imaging for Industrial Application"), and short papers from

[2] http://www.eu-maths-in.eu.

[3] http://www.cmap.polytechnique.fr/spip.php?rubrique61.

[4] http://asso-ama.dmi.unict.it/en.

[5] http://www.ricam.oeaw.ac.at/.

[6] http://http://cmcs.epfl.ch/people/quarteroni.

the contributed sessions (chapter "A Customized System for Vehicle Tracking and Classification").

The mini-symposia are ordered alphabetically by title, as are the corresponding papers.

The last chapter contains the paper related to the Anile prize lecture.

The 18th ECMI conference, managed by the organizing committee, with the collaboration of ECMI and Tao Sciences Research Center, took place in Taormina, Sicily, on the beautiful premises of the Hotel Villa Diodoro, conveniently located near the Villa Comunale of Taormina and the Teatro Greco, offering a gorgeous view of Taormina Bay.

Acknowledgments

Many people contributed to the success of the conference, and it is impossible to list them all in a limited space. First of all, we would like to thank Mrs. Marisa Lappano Anile, who helped us a lot for all organizational, administrative, and practical problems, with professionality and enthusiasm. Even more, with a strong motivation behind, which bears the name of our friend and mentor Marcello Anile, she was one of the inspirers of the event. We thank all the members of the Scientific Committee, who determined the selection of the plenary speakers. In particular, we thank the past presidents of ECMI, such as L. Bonilla, and M. Guenther, and W. Schilders, for valuable practical advices about the organization of the conference. A thank goes to the Board of ECMI, for valuable advices and, even more, for trusting our actions and leaving us a large autonomy both at a scientific and organizational level. A special thank goes to the rest of the organizing committee and in particular the young volunteers: Alice, Armando, Gianluca, Gianni, Ivano, Leonardo, Mario, Riccardo, Sergio, and Silvia: they worked hard on virtually all organizational aspects of the conference. We warmly thank all the plenary speakers for accepting the invitation and delivering such interesting talks. Many thanks to the sponsors of the event, namely, the Associazione Angelo Marcello Anile, the Office of Naval Research Global, and European Patent's Office, for their generous contributions. We would like to thank all the mini-symposia organizers, all the contributors, and all the participants, who determined the success of the event and contributed to the pleasant general atmosphere of the conference. We all thank our wives for the patience and support during the several months the organization of the conference kept us busier than usual. We thank Prof. Nicole Marheineke for help in the management of the final version of the proceedings. Last, but not least, we would like to express our

gratitude to Ruth Allewelt, from Springer, for her help, patience, and competence granted us during the preparation of the present book of proceedings.

Catania, Italy — Giovanni Russo
Catania, Italy — Giuseppe Nicosia
Catania, Italy — Vittorio Romano
Milan, Italy — Vincenzo Capasso

Contents

Contents for Volume 2

Part I
Minisymposia

MS 1
MINISYMPOSIUM: ADVANCED IMAGING FOR INDUSTRIAL APPLICATION

Organizers

Sebastiano Battiato[1], Giovanni Gallo[2] and Filippo Stanco[3]

Speakers

Marco Leo[4] and Cosimo Distante[5]
Iris Segmentation: A New Strategy for Real Biometric Applications

Giovanni Maria Farinella[6], Sebastiano Battiato[1], Giovanni Giuffrida[7] and Davide Moltisanti[8]
Web Scraping of Online Newspapers Via Image Matching

Sebastiano Battiato[1], Giovanni Maria Farinella[6], Antonino Furnari[9], Giovanni Puglisi[10]
A Customized System for Vehicle Tracking and Classification

[1]Sebastiano Battiato, University of Catania, Italy.
[2]Giovanni Gallo, University of Catania, Italy.
[3]Filippo Stanco, University of Catania, Italy.
[4]Marco Leo, ISASI-CNR, Italy.
[5]Cosimo Distante, ISASI-CNR, Italy.
[6]Giovanni Maria Farinella, University of Catania, Italy.
[7]Giovanni Giuffrida, Neodata Group, Italy.
[8]Davide Moltisanti, Bristol, UK.
[9]Antonino Furnari, University of Catania, Italy.
[10]Giovanni Puglisi, University of Cagliari, Italy.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_1

Keywords

Computer vision
Image matching
Segmentation
Tracking

Short Description

The aim of this mini-symposium is to provide an overview of state of the art methods for imaging applications in different industrial contexts (consumer devices, e-health, digital signage, forensics, Cultural Heritage, etc.) to stimulate the creation of appropriate benchmark dataset to be used as reference for the development of novel algorithms.

A Customized System for Vehicle Tracking and Classification

Sebastiano Battiato, Giovanni Maria Farinella, Antonino Furnari, and Giovanni Puglisi

Abstract We present a customized system for vehicle tracking and classification. The main purpose of the system is tracking the vehicles in order to understand lane changes, gates transits and other behaviors useful for traffic analysis. The classification of the vehicles into two classes (short vehicles vs. tall vehicles) is also performed for electronic truck-tolling as well as to optimize the performances of the tracker module. The whole system has been developed through a data driven approach based on video sequences acquired by QFree. (Q-Free (www.q-free.com) is a global supplier of solutions and products for Road User Charging and Advanced Transportation Management having applications mainly within electronic toll collection for road financing, congestion charging, truck-tolling, law enforcement and parking/access control.) The sequences are acquired by wide angle cameras from the top of the road and are preprocessed in order to obtain a normalized, low-resolution representation of the scene where the distance between neighboring pixels is constant in the real world. The sequences exhibit high variability in terms of lighting changes, contrast changes and distortion. We assume that the vehicle detection is performed by an external module for plate recognition.

Keywords Image matching • Tracking

The tracking algorithm is based on Template Matching [1, 2] and the Normalized Cross Correlation is used as similarity measure. The vehicle template is updated at each frame to cope with the vehicles' changes of appearance. In order to deal with the main variabilities, four modules are designed: a multicorrelation module to deal with the appearance of artifacts on the vehicles; a refinement module to deal with the change of the vehicle horizontal scale due to distortion (see Fig. 1a–c); a background subtraction module to deal with perspective issues on tall vehicles; a

S. Battiato • G.M. Farinella • A. Furnari (✉)
Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: battiato@dmi.unict.it; gfarinella@dmi.unict.it; furnari@dmi.unict.it

G. Puglisi
Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy
e-mail: puglisi@unica.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_2

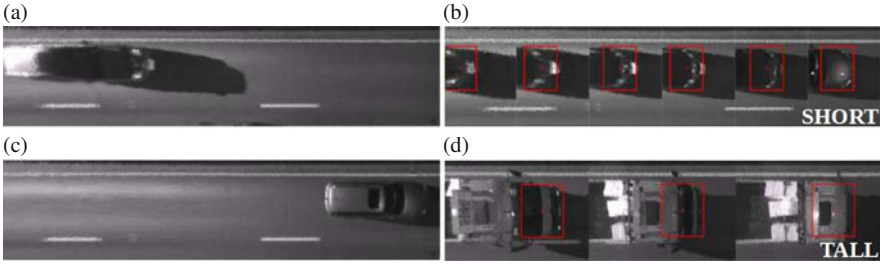


Fig. 1 Some sample images from the developed system. (**a**), (**b**) Vehicle deformation. (**c**) Tracking results. (**d**) Classification results

selective update module to avoid the propagation of a wrong template in the slow scenes. A controller has been realized to switch on or off the modules depending on the vehicle estimated speed and the vehicle estimated class. Two measuring methods are defined to assess the performances of the proposed tracker with respect to other standard tracking pipelines [3–6] in a supervised way. The performance analysis points out that the vehicles are correctly tracked for nearly the 99 % of the scene. Fig. 1b shows some tracking results.

The classification is performed when the vehicle approaches the central part of the scene, where the variabilities are less significant. The image patches are extracted from the frame taking into account the estimated vehicle bounding box. The training set is built considering the image patches extracted from the input sequences. To make the learning procedure more robust, the training set is augmented generating artificial patches tailored to introduce alignment, perspective, rotation and photometric variabilities. The patches are normalized to the training set mean patch size and the HOG (Histogram of Oriented Gradients) features are extracted [7]. The feature vectors dimensionality is reduced through the Principal Component Analysis (PCA) [8]. The labeled samples are then projected through the Linear Discriminant Analysis (LDA) [9] to the most discriminant dimension. This unidimensional feature is aggregated to the patch height in pixels, obtaining a two-dimensional vector. A new LDA projection is hence performed on the two-dimensional samples and the two projected populations are modeled as unidimensional Gaussian distributions. In the classification step, the sample is projected using the previously learned PCA and LDA bases. The Mahalanobis distances [10] are computed between the projected sample and the Gaussian distributions related to the two classes (short vs. tall vehicles). The sample is assigned to the class giving the smallest distance. The classification performances are evaluated with the Leave One Out strategy and the overall classification accuracy is over the 98 %. Fig. 1d shows some classification results.

This work has been performed in the project PANORAMA, co-funded by grants from Belgium, Italy, France, the Netherlands, and the United Kingdom, and the ENIAC Joint Undertaking.

References

1. Maggio, E., Cavallaro, A.: Video Tracking: Theory and Practice. Wiley, New York (2011)
2. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. **38**(4), 13-es (2006)
3. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. Imaging **130**, 674–679 (1981)
4. Baker, S., Gross, R., Ishikawa, T., Matthews, I.: Lucas-kanade 20 years on: a unifying framework: Part 2. Int. J. Comput. Vis. **56**(3), 221–255 (2004)
5. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**, 564–577 (2003)
6. Bradski, G.R.: Computer vision face tracking as a component of a perceptual user interface. In: IEEE Workshop on Applications in Computer Vision (1998)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. **1**, 886–893 (2005)
8. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**(6), 417 (1933)
9. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
10. Mahalanobis, P.C.: On the generalized distance in statistics. Proc. Natl. Inst. Sci. **2**, 49–55 (1936)

Iris Segmentation: A New Strategy for Real Biometric Applications

Marco Leo, Tommaso De Marco, and Cosimo Distante

Abstract Iris segmentation is driven by three different quality factors: accuracy, usability and speed. Unfortunately the deeply analysis of the literature shows that the greatest efforts of the researchers mainly focus on accuracy and speed. Proposed solutions, in fact, do not meet the usability requirement since they are based on specific optimizations related to the operating context and they impose binding conditions on the sensors to be used for the acquisition of periocular images. This paper tries to fill this gap by introducing an innovative iris segmentation technique that can be used in unconstrained environments, under non-ideal imaging conditions and, above all, that does not require any interaction for adaptation to different operating conditions. Experimental results, carried out on challenging databases, demonstrate that the high usability of the proposed solution does not penalize segmentation accuracy which, in some respects, outperforms that of the leading approaches in the literature.

Keywords Image matching • Segmentation

1 Introduction

Human identification leads to mutual trust that is essential for the proper functioning of society. With an increasing attention to security, biometric authentication has grown in popularity as an alternative way to provide personal identification that can overcome the limits of traditional authentication systems based on credentials (documents and PIN) which may be lost, stolen, or easily forgotten [10, 11]. The design and suitability of biometric technology for person identification depends on the application requirements. These requirements are typically specified in terms of identification accuracy, throughput, user acceptance, system security, robustness, and return on investment. The next generation biometric technology must overcome many hurdles and challenges to improve the recognition accuracy.

M. Leo (✉) • T. De Marco • C. Distante
National Research Council of Italy, Institute of Optics, via della libertà, Arnesano (Lecce), Italy
e-mail: marco.leo@cnr.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_3

These include ability to handle poor quality and incomplete data, achieve scalability to accommodate hundreds of millions of users, ensure interoperability, and protect user privacy while reducing system cost and enhancing system integrity. One of the most attractive and promising biometric modalities is based on the recognition of the iris texture that is stable and distinctive, even among identical twins (similar to fingerprints), and extremely difficult to surgically spoof. There are now various national identity schemes in progress that make use of iris recognition technology and there is also a large and vibrant research community studying ways to make it even more accurate in even larger-scale applications [1]. The fundamental components of an iris recognition system are: image acquisition, iris segmentation, iris feature extraction, iris template generation, iris template matching, and iris identification. Iris segmentation includes pupillary boundary and limbic boundary detection, and eyelids and eyelash exclusion [7] and its performances strongly affect the accuracy of the person's identification accuracy.

The most relevant and widely used algorithms require NIF camera to segment the iris images [3]. The well-known integro differential operator [4] is then used to search a circle to separate iris clearly from other parts of the imagery and remains in use widely today in commercial applications. Another classical circle-based model is the edge detection-based techniques [17], where the circular Hough transform is followed by edge detection to localise iris boundary. The above algorithms significantly decrease their accuracy if noisy iris images taken in visible wavelength and under non-ideal imaging conditions are used as input. In these cases, some of the factors that make the segmentation very challenging are: occlusions caused by the anatomical features of the eye (eyelids, eyelashes,...), illumination (poor illumination, specular reflections), user cooperation (off-angled iris, motion blur, eye glasses or contact lenses,...) [6, 13]. To address such a challenge, some efforts have been recently made [2, 14–16]. Unfortunately, most of the above segmentation schemes are very complex (many different sequential algorithmic steps involved), are not parameter free and, above all, they make use of some a priori knowledge. In other words, the experimented accuracy results on challenging datasets, are obtained through the use of specific optimizations which reduce the usability in real identity check applications. Starting from these considerations, this paper aims at introducing an iris segmentation algorithm that overcomes some of the aforementioned limitations of the state of the art methods: the core is a randomized circle detection algorithm which uses an alternative multiple-evidence strategy to define a valid circles set. The algorithm is iteratively applied to find limbic, pupil and eyelid boundaries. The proposed segmentation scheme can be used in unconstrained environments, under non-ideal imaging conditions, and above all it does not require any interaction for adaptation to different operating conditions. This has been widely proved by testing it on a challenging dataset containing noisy image and by comparing its outcomes with those of the leading approaches in literature.

2 Overview of the Proposed Approach

The proposed approach works on input iris image taken in visible wavelength. In Fig. 1 a schematic representation of how the system works is reported.

The system takes as input a periocular image and then a randomized circle detection algorithm (described in [5]) is firstly applied in order to find the outer boundaries of the iris region. No constraints about the searched radius are imposed: the algorithm searches for all possible radius and for all the possible locations of its center. The robustness of the circle detection algorithm allows the system to highlight the limbic boundaries among the number of possible false circular shapes that can emerge around the eye. The radius R of the detected circular region is then used as a reference for the following steps. Then, inside the detected iris region, the same circle detection algorithm is applied to search the region corresponding to the pupil boundaries. In this case the searched region is expected to have a radius in the range $[\frac{R}{3}; \frac{2R}{3}]$ depending on the pupil dilation [8] and a center quite coincident with the center of the iris region previously detected. Detected pupil area is then removed and the next step is performed with the aim to locate the upper and lower eyelids: this is done by dividing the image in two small rectangles from the outer two sides of the iris. On the rectangle corresponding to the upper part of the image the circle detection algorithm is performed to search a circular shape having radius in the range $[\frac{2R}{3}; 3R]$ whereas on the rectangle corresponding to the lower part of the image, the circle detection algorithm, is performed to search a circular shape having radius in the range $[4R; 6R]$. External regions (with respect to the center of the iris) are then removed and the remaining pixels approximatively correspond to the only iris region that can be supplied as input to the biometric identification algorithm.

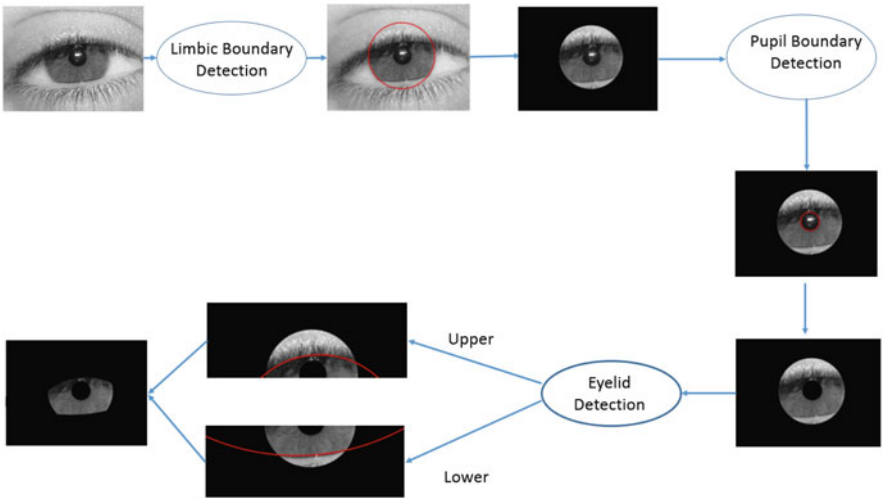


Fig. 1 A schematic representation of how the proposed system works

The strength of the proposed approach is undoubtedly the circle solver that uses an alternative multiple-evidence strategy to define a valid circles set: maintaining the four edge pixels approach, the circle solver adds further constraints based on the curvature of the isophotes. Isophotes are curves connecting pixels in the image with equal intensity, whose properties make them particularly suitable for objects detection [9]. Preliminarily it discards edge pixels with too high or low local isophote curvature, then the remaining pixels are classified into subsets of adjacent pixels with the same isophotes curvature (and consequently belonging to possible candidate circles with the same radius). This leads to three improvements with respect classical randomized approaches: firstly, the sampling process can be limited on each subset, so increasing probability to sample edge pixels belonging to the same circle, secondly candidate circles with radius not compliant with the subset under exam (false positives) can be discarded before the voting process and finally, dependency of the results from the used edge map is reduced. For each candidate circle, a kernel density based estimation voting process is performed: this provides better results than simple counting of edge pixels, because inliers are automatically defined according to the distribution of the distances between each edge pixel and the circle center. Then, detected circles parameters are refined with an error linear compensation algorithm, in order to provide a better fitting with the recognized circle and the inliers. The next subsections explain in more detail the algorithms involved in the proposed system.

3 Experimental Results

To evaluate the accuracy, usability and speed of proposed iris segmentation algorithm, it has been implemented using MATLAB®—R2012a software on a ASUS N56V with Intel® Core™—i7 3630 QM Processor (2.54 GHz, RAM 16 GB). Experiments were carried out on the UBIRISv1 database [12]. This database is composed of 1877 images collected from 241 persons in two distinct sessions: its most relevant characteristic is to incorporate images with several noise factors, simulating less constrained image acquisition environments. This enables the strict evaluation of the robustness of the proposed iris segmentation method. The database consists of two sessions of image capture: in the first session the noise factors are minimized, specially those relative to reflections, luminosity and contrast, since the image capture framework was installed inside a dark room. In the second session the capture place was changed in order to introduce natural luminosity factors. This propitiates the appearance of heterogeneous images with respect to reflections, contrast, luminosity and focus problems. Images collected at this stage simulate the ones captured by a vision system without or with minimal active participation from the subjects, adding several noise problems.

In Fig. 2 on the left an image of the UBIRIS dataset is shown, whereas, on the right, the four circles, detected by sequentially applying the algorithm in [5], are superimposed. The circle in red corresponds to the limbic boundaries (the iris), the

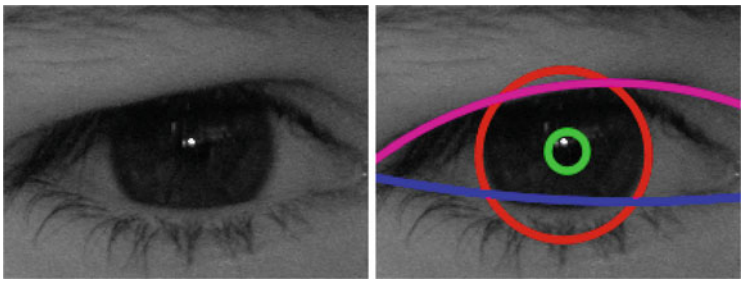


Fig. 2 The four circles detected on a periocular image

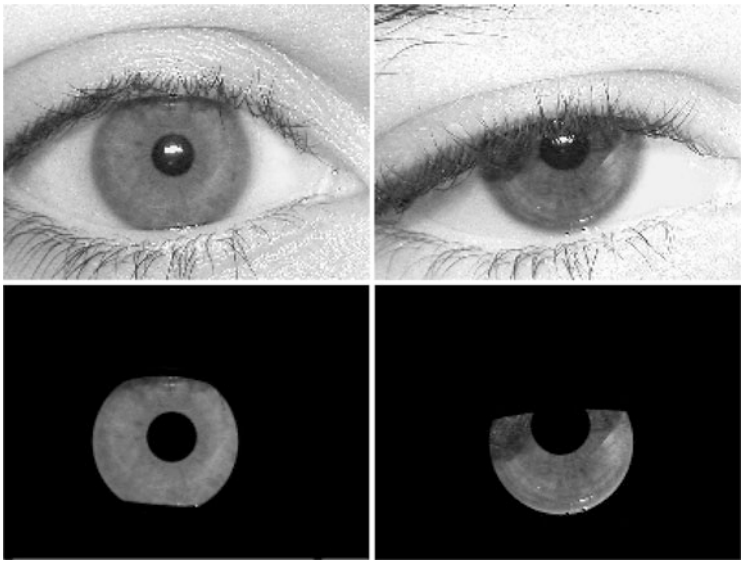


Fig. 3 Segmentation results on the UBIRISv1 dataset (Session 2)

green one corresponds to the pupil boundaries, the blue one to the lower eyelid and finally the purple one to the upper eyelid boundaries.

Figure 3 shows some periocular images of the UBIRIS dataset (session 1) and the corresponding final segmented images.

Figure 4 shows, instead, the segmented images after applying the proposed iris segmentation algorithm on session 2 of the UBIRISv1 database. These images are very challenging due to the noise and the iris occlusions: in particular image in the first row contains an eye acquired with a strong frontal light source and then some regions are overexposed (with pixel saturation effect) whereas the image in the second row is out of focus. Finally the image in the third row reports a semi-closed eye. Anyway, as proved by the corresponding segmented image on the right, these challenging conditions (that simulate acquisition in unconstrained environments) did not affect the segmentation process.

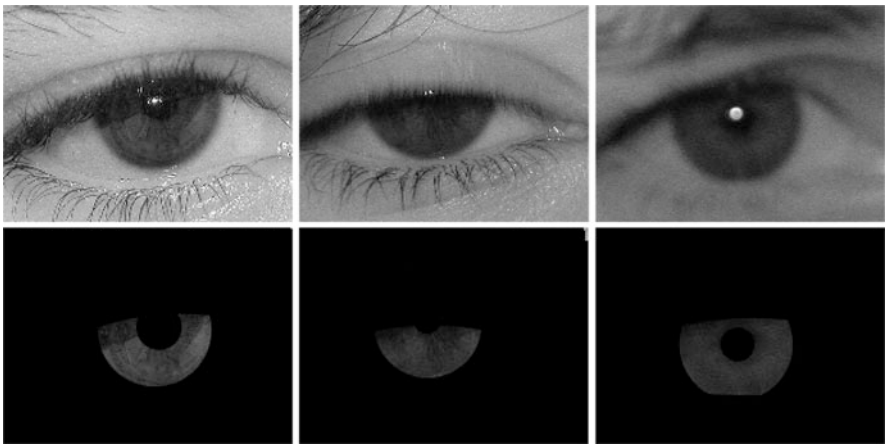


Fig. 4 Segmentation results on the UBIRISv1 dataset (Session 1)

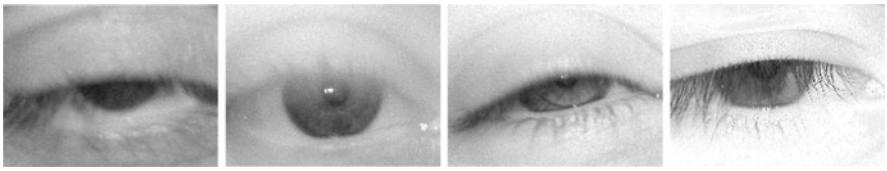


Fig. 5 Some images on which the proposed approach failed

Table 1 Segmentation accuracy

Step	Time (s)
Detection of outer borders of iris	1.1
Detection of inner borders of iris	0.4
Detection of lower eyelid	0.3
Detection of upper eyelid	0.3
Overall	2.1

Anyway, the proposed algorithm fails in segmenting noisy irises when eyelids and eyelashes obstruct big portions of the iris (more than 50 %) or when the upper or lower eyelids cover the pupil of the iris (note that most segmentation methods fail in these cases). Some of the images on which the proposed approach failed are reported in Fig. 5.

Concerning the computational remarks, the whole segmentation process took an average time about 2 s for each image of the dataset UBIRISv1 (image size 200×150 pixels). In particular the average computational time for each step involved in the segmentation process is reported in Table 1.

As expected, the first step (search for outer boundaries of the iris) is more computationally expensive since it can not exploit any a priori knowledge about the position and the size of the searched circle. The following steps instead, using the

position and radius information of the iris extracted in the step 1, are faster although they are based on the same iterative algorithm. In practice they exploit the available information to optimize the selection of initial points used to fit the searched circle and thus allowing a fast convergence to the optimal solution.

4 Conclusion

This paper introduced an innovative iris segmentation technique that can be used in unconstrained environments, under non-ideal imaging conditions, and above all that does not require any interaction for adaptation to different operating conditions. Experimental results, carried out on a challenging database, demonstrated that the high usability of the proposed solution does not penalize segmentation accuracy which, in terms of capability to extract the inner (pupillarity) and outer borders (limbic) of the iris, outperforms that of the leading approach in the literature. Future work will address the test of the system on different datasets and the implementation in an intermediate level language in order to speed up the calculation.

References

1. Burge, M.J., Bowyer, K.W.: Handbook of Iris Recognition. Springer Publishing Company, Incorporated, Berlin (2013)
2. Chen, Y., Adjouadi, M., Han, C., Wang, J., Barreto, A., Rishe, N., Andrian, J.: A highly accurate and computationally efficient approach for unconstrained iris segmentation. Image Vis. Comput. **28**(2), 261–269 (2010)
3. Daugman, J.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1148–1161 (1993)
4. Daugman, J.: How iris recognition works. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 21–30 (2004)
5. De Marco, T., Leo, M., Distante, C.: Soccer ball detection with isophotes curvature analysis. In: Image Analysis and Processing' ICIAP 2013. Lecture Notes in Computer Science, vol. 8156, pp. 793–802 (2013)
6. D'Orazio, T., Leo, M., Distante, A.: Eye detection in face images for a driver vigilance system. In: 2004 IEEE Intelligent Vehicles Symposium, pp. 95–98 (2004). doi: 10.1109/IVS.2004.1336362
7. Du, Y., Arslanturk, E., Zhou, Z., Belcher, C.: Video-based noncooperative iris image segmentation. IEEE Trans. Syst. Man Cybern. B Cybern. **41**(1), 64–74 (2011). doi: 10.1109/TSMCB.2010.2045371
8. Hollingsworth, K., Baker, S., Ring, S., Bowyer, K.W., Flynn, P.J.: Recent research results in iris biometrics. In: Proceedings of SPIE, vol. 7306, pp. 73061Y-1 (2009)
9. Lichtenauer, J., Hendriks, E., Reinders, M.: Isophote properties as features for object detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 2, pp. 649–654 (2005). doi: 10.1109/CVPR.2005.196
10. Martiriggiano, T., Leo, M., Spagnolo, P., D'Orazio, T.: Facial feature extraction by kernel independent component analysis. In: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005. AVSS 2005, pp. 270–275 (2005). doi: 10.1109/AVSS.2005.1577279

11. Martiriggiano, T., Leo, M., Spagnolo, P., D'Orazio, T.: Facial feature extraction by kernel independent component analysis. In: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005. AVSS 2005, pp. 270–275 (2005). doi: 10.1109/AVSS.2005.1577279
12. Proenca, H., Alexandre, L.: UBIRIS: a noisy iris image database. In: 13th International Conference on Image Analysis and Processing - ICIAP 2005. Lecture Notes in Computer Science, vol. 3617, pp. 970–977. Springer, Berlin (2005)
13. Proenca, H., Alexandre, L.A.: Introduction to the special issue on the recognition of visible wavelength iris images captured at-a-distance and on-the-move. Pattern Recogn. Lett. **33**(8), 963–964 (2012)
14. Sahmoud, S.A., Abuhaiba, I.S.: Efficient iris segmentation method in unconstrained environments. Pattern Recogn. **46**(12), 3174–3185 (2013)
15. Tan, T., He, Z., Sun, Z.: Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition. Image Vis. Comput. **28**(2), 223–230 (2010)
16. Wan, H.L., Li, Z.C., Qiao, J.P., Li, B.S.: Non ideal iris segmentation using anisotropic diffusion. Image Process. IET **7**(2), 111–120 (2013)
17. Wildes, R.: Iris recognition: an emerging biometric technology. Proc. IEEE **85**(9), 1348–1363 (1997)

Web Scraping of Online Newspapers via Image Matching

D. Moltisanti, G.M. Farinella, S. Battiato, and G. Giuffrida

Abstract Reading is an activity which takes place widely on the web: almost all newspapers have his own digital version on the internet and there are even a lot of magazines only on the web. In such a scenario, Computer Vision can offer a useful set of tools that can help web editors to improve the quality of the provided service. One of these tools is here presented: given a webpage of a newspaper or journal, the proposed framework localizes news items remotely clicked by users, giving the bounding box of the content of an article in its relative homepage. The tool is hence able to track an article in the page in which is contained at any time during the day: such an information is very useful for web editors to understand the trend of the published items and to rearrange the contents of the homepage accordingly.

Keywords Computer vision • Image matching

1 Introduction

The system has been developed with an hybrid approach: first we manipulate the HTML source of the homepage in order to generate a visual template (a HTML page) for the news item we want to localize. Thereafter we take the screenshot of such template in order to obtain an image to be used for the localization of the article, which is done via Keypoint and Template Matching.

The results depend on the layout of the websites and on the manner they arrange the contents in the homepage. Our tests show that good performances can be reached, giving also an interesting analysis and comparison among different keypoint descriptors.

D. Moltisanti (✉) • G.M. Farinella • S. Battiato
Image Processing Laboratory, Universita' degli Studi di Catania, Viale Andrea Doria 6, Catania, Italy
e-mail: davidemoltisanti@gmail.com; gfarinella@dmi.unict.it; battiato@dmi.unict.it

G. Giuffrida
Neodata Group SRL, Viale Vittorio Veneto 42, Catania, Italy
e-mail: giovanni.giuffrida@neodatagroup.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_4

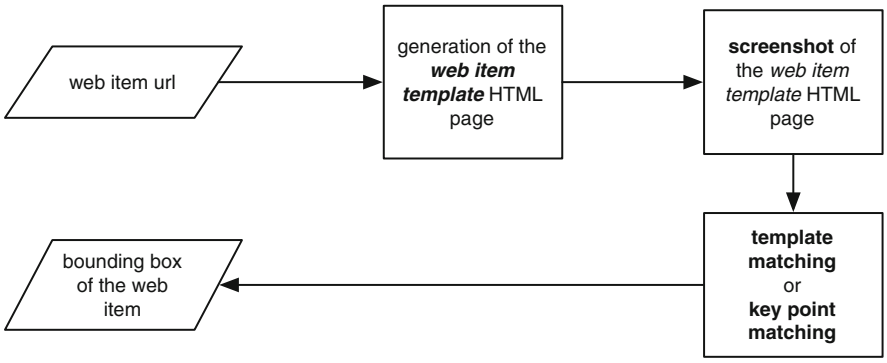


Fig. 1 General workflow of the localization system

2 Proposed Work

Given a web page, first we take its screenshot in order to have the image version of the HTML page for further operations; once we have done so, the pipeline for the article localization is used. The work flow of the system is synthetically sketched in Fig. 1.

2.1 Web Item Template Generation

The most delicate and important phase of the whole system is the generation of a HTML page containing only the sought web item which contains the news we want to localize. We call such HTML page "web item template", as it represents the item we are looking for in the parent page. In order to succeed with the following step of template (or keypoint) matching, it is crucial to have a template which is consistent with the layout of the web item: we can localize web items in their parent web pages only if the generated templates are arranged in the same manner of the original page from which are extracted. We use the JSoup Library[1] to produce web item templates. To generate the template of an article given its URL, the system seeks the item in the parent page and, for each element pointing to such URL, it runs a cleaning procedure which generates the corresponding template by removing from the page everything except the item to localize.

[1] Hedley, J.: Jsoup java html parser. http://jsoup.org.

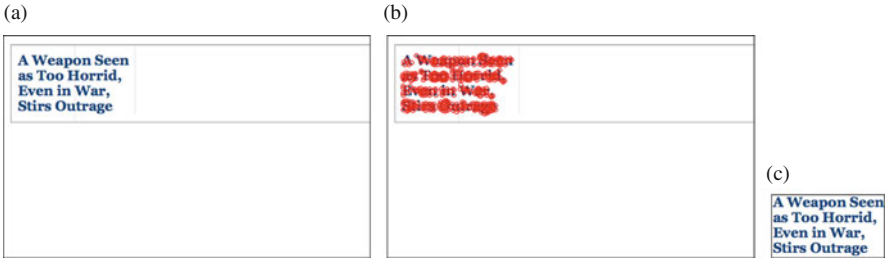


Fig. 2 Web item template cut. Since FAST keypoints are localized only upon non blank zones, we can remove void areas in order to have a tight-fitting picture representing the web item. (**a**) Web item template. (**b**) Web item template with keypoints. (**c**) Web item template after the cut

2.2 Web Item Template Screenshot

Once we have generated the web item we take its screenshot, as we did at the beginning with the main web page we are working on.

By taking the screenshot of the HTML page we obtain the image version of the HTML document and we can then extract the keypoints of such image using the OpenCV Library.[2] Since the images we have to deal with are all synthetics (no skew or rotation problems), and given that the images of the web item templates are almost blank pictures mainly composed of text, we decided to use the Fast Corner Detector [7]. This detector fits well for our purpose because the extracted keypoints in the screenshot image are localized only over text and pictures, excluding layout lines and uniform zones (i.e. the background of the page).

2.2.1 Keypoint Extraction

The extraction of the keypoints of the template images has two aims: the generation of the set of keypoints of the web item, to be used for keypoint matching; the cutting of the screenshot image, in the case of the template matching localization method.

Taking the screenshot of a web item template HTML page we obtain a mostly blank image with some content in it (our web item); if we want to localize the item via template matching, we need an image containing only the content of the item. To obtain a tight-fitting template image we cut conveniently the image, by selecting the region where the keypoints are concentrated and discarding remaining areas. Zones which contain only sparse keypoints are removed too. In Fig. 2 we can observe an example of a web item image cut. In Fig. 3 we have the keypoints extracted for the headline web item of the Italian newspaper "La Repubblica" (www.repubblica.it).

[2]Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000).

Fig. 3 FAST keypoint extracted for the headline web item template

2.3 Localizing the Web Item

The last step of our flow is the image-based localization phase. We provide two methods: keypoints based matching and template matching.

The general task performed by the two methods is the following: we have an image I, representing our main web page we are working on, and an image T representing our web item template, extracted from the main page. We know that T is contained in I, and we want to know the position of T in I. Both methods do the same thing, but differ from how they do that. Both provide the bounding box of the sought web item (the coordinates of the top left corner, the width and the height), giving hence the location of the item relative to the page in which is contained.

2.3.1 Template Matching Method

The template matching is a method to localize a pattern T (a $w \times h$ image) in an image I. The method compares the template T against overlapped regions of the image I. To do so, it slides through I, compares the overlapped patches of size $w \times h$ against T using the Normalized Correlation Coefficient (NCC) method and stores the comparison results. The NCC method compares two images as follows:

$$R(x,y) = \frac{\sum_{x',y'} (T'(x',y') \cdot I'(x+x',y+y'))}{\sqrt{\sum_{x',y'} T'(x',y')^2 \cdot \sum_{x',y'} I'(x+x',y+y')^2}} \tag{1}$$

where

$$T'(x',y') = T(x',y') - \frac{1}{w \cdot h} \cdot \sum_{x'',y''} T(x'',y'') \tag{2}$$

$$I'(x+x',y+y') = I(x+x',y+y') - \frac{1}{w \cdot h} \cdot \sum_{x'',y''} I(x+x'',y+y'')$$

$R(x, y)$ is the result of the comparison calculated in $I(x \dots x + w, y \dots y + h)$; the summation is done over the template and the image patch: $x' = 0, \dots, w - 1,\ y' = 0, \dots, h - 1$.

The point (x, y) where R reaches its maximum is the location of the top left corner of the template T inside the image I, and therefore the position of our web item relative to the web page in which is contained.

2.3.2 Keypoints Matching

To localize a pattern T in an image I, we first need the keypoints of both images. As described formerly, at this point of the work flow we already hold the keypoints of the main web page, extracted at system start time, and the keypoints of the web item template, extracted during the screenshot phase.

To localize T in I we use the FLANN (Fast Library for Approximate Nearest Neighbors) library [6] implemented in OpenCV.

We provide four descriptors for the matching: SIFT [5], SURF [1], BRIEF [2], BRISK [4]. Our tests (see Sect. 3) prove that SIFT descriptor is the one which best performs the correct bounding box extraction of the web items.

Once we have got the correspondences between the keypoints sets, we are able to obtain the location in which the matching takes place. The keypoints matching method is the default localizing method of the system, because it is much faster than the template matching method. This is due to the fact that the dimension of the pattern is much smaller than the one of the template matching method, as we have to deal only with a subset of points (the keypoints) of the template image, instead of the whole set of points (namely each pixel) of the image.

In Fig. 4 we can observe the localization of a news item via keypoint matching.

3 Experimental Results

In this section we report some of the results of the experimental tests we executed on several online newspapers. For each test, we compare the localization results obtained via template matching against those obtained via keypoint matching using the four aforementioned descriptors: SIFT, SURF, BRIEF, BRISK. We have then a total of five localization method.

The ground truth of each test is provided by a Javascript algorithm which, given in input the URL of a web item, returns its bounding box by analysing the HTML source code.

The test of a website has been developed as follows:

1. Extraction of all of the news item in the web page;
2. Generation of the bounding box $B_{gt}(a)$ (the box provided by the ground truth) for each web item a;

Fig. 4 Keypoint matching localization. The *coloured lines* track the correspondences between the keypoints, represented with *thin circles*. The *green rectangle* represents the bounding box of the news item

3. For each localization method L_m, generation of the bounding box $B_m(a)$ for each web item a;
4. Comparison of the boxes $B_{gt}(a)$ with the corresponding boxes $B_m(a)$.

The bounding boxes are compared using a rectangles overlap measure proposed in [3], which is defined as follows:

$$p\Big(B_{gt}(a), B_m(a)\Big) = \frac{\text{area}\Big(B_{gt}(a) \cap B_m(a)\Big)}{\text{area}\Big(B_{gt}(a) \cup B_m(a)\Big)} \tag{3}$$

Where the range of p values is $[0 \ldots 1]$. We use the p value to determine if a box obtained with our system is a hit or a miss: if p is greater or equal to 0.10, the overlapping area of the ground truth box with the area of the test rectangle is large enough to state that the test box is located on the correct region of the page. We have a hit also when $B_{gt}(a)$ contains entirely the box $B_m(a)$, regardless of the value of p.

The definition of hit and miss is eventually the following:

$$B_m(a) = \begin{cases} \text{HIT} & p\Big(B_{gt}(a), B_m(a)\Big) \geq 0.10 \\ \text{HIT} & B_m(a) \subset B_{gt}(a) \\ \text{MISS} & \text{otherwise} \end{cases} \tag{4}$$

where $B_m(a) \subset B_{gt}(a)$ indicates that the ground truth box contains entirely the box $B_m(a)$. For each test we report the mean of the localization error of each localization

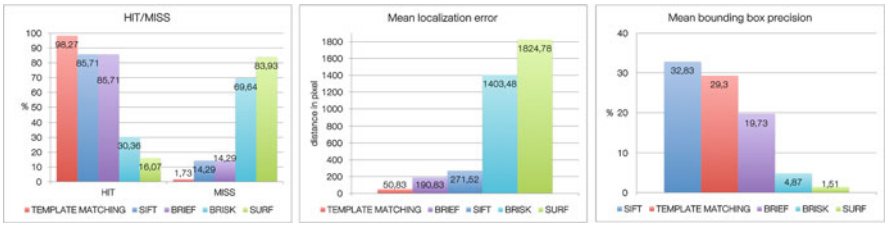


Fig. 5 Test results on the homepage of "Corriere della Sera" (www.corriere.it)

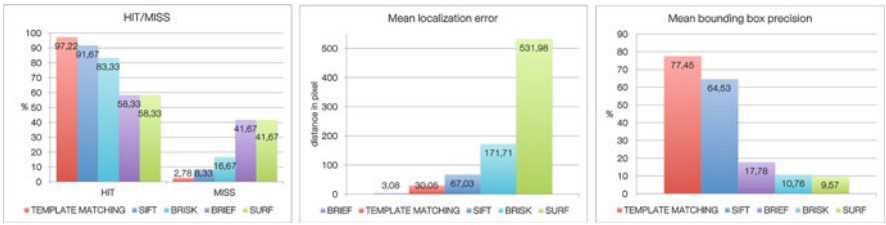


Fig. 6 Test results on the homepage of "National Geographic" (www.nationalgeographic.it)

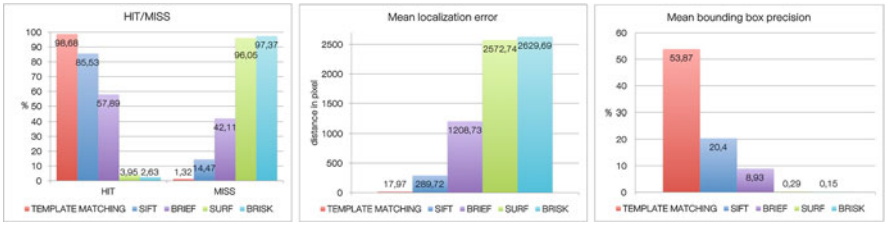


Fig. 7 Test results on the homepage of "Huffington Post" (www.huffingtonpost.it)

method, along with hit/miss percentages and mean precision of the boxes (that is, mean of p values). The localization error is defined as the Euclidean distance between the top left corners of the boxes $B_m(a)$ and $B_{gt}(a)$. In Figs. 5, 6, and 7 we report the results of three tests we executed.

4 Conclusions

The work presented in this paper aimed to provide a new point of view about web sites; if we look at a web page not only as a HTML document, but also as an image, we can imagine several Computer Vision instruments to be built for several scopes.

An important instrument is the tracking of the news items positions over time, since the location of a news in the homepage is a critical factor for the audience interest. Many other factors influence the visibility and the appealing of a news:

for instance, if an image is attached to an item, then the corresponding news has statistically a greater chance to be red. The dimension and the style of the news title, and in general the size of the bounding box containing the item is important too.

Considering web pages as images, layout analysis tools could be developed: for instance, a density map which draws the interest of the news could be helpful to properly dispose the web items. Also, one could develop a system which automatically detect page layout changes over time, or a tool which gives a score of the quality of the design of the page, according to some psychological study. In such a scenario, Computer Vision can hence offer a useful set of tools that can help web editors to improve the quality of the provided service.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features. Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
2. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: Brief: binary robust independent elementary features. In: Computer Vision—ECCV 2010, pp. 778–792. Springer, Berlin (2010)
3. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
4. Leutenegger, S., Chli, M., Siegwart, R.Y.: Brisk: Binary robust invariant scalable keypoints. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2548–2555. IEEE, New York (2011)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
6. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP (1), pp. 331–340 (2009)
7. Rosten, E., Porter, R., Drummond, T.: Faster and better: a machine learning approach to corner detection. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 105–119 (2010). doi: 10.1109/TPAMI.2008.275

MS 2
MINISYMPOSIUM: BAYESIAN AND APPROXIMATIVE SAMPLING METHODS FOR UNCERTAINTY QUANTIFICATION

Organizers

Heikki Haario[1] and Marko Laine[2]

Speakers

Marko Laine[2], Heikki Haario[1]
Parameter Estimation in Large Scale State Space Models Using Ensembles of Model Runs

Vladimir Shemyakin[3], Heikki Haario[1]
Tuning Parameters of Ensemble Prediction System and Optimization with Differential Evolution Approach

Alexander Bibov[4] Heikki Haario[1]
Stabilizing Correction for Approximative Low-Memory Kalman Filtering. Extensions and Generalizations

[1]Heikki Haario, Lappeenranta University of Technology and Finnish Meteorological Institute, Finland.

[2]Marko Laine, Finnish Meteorological Institute, Finland.

[3]Vladimir Shemyakin, Lappeenranta University of Technology, Finland.

[4]Alexander Bibov, Lappeenranta University of Technology, Finland.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_5

Keywords

Chaotic systems
Data assimilation
Filtering
Parameter estimation
State estimation

Short Description

The aim of Uncertainty Quantification is to estimate the reliability of model simulations. In addition to the statistical uncertainties due to noisy measurement data, one wants to estimate the impact of model bias and numerical approximations necessary due to high CPU demands or high model state dimensions.
The motive for the methods presented in this mini symposium comes from weather and climate models. We present methods that enable MCMC sampling for high-CPU systems, as well as approximative filtering methods that enable state estimation of very high-dimensional models. The approaches are applied to parameter estimation of chaotic systems. On the other hand, a no-cost parameter estimation approach is discussed, that is based on monitoring of operational weather predictions. The results are compared to those we obtain by differential evolution algorithms.

Numerical Modelling of Wind Flow over Hills

O. Agafonova, A. Koivuniemi, B. Conan, A. Chaudhari, H. Haario, and J. Hamalainen

Abstract The paper demonstrates when the Wind Atlas Analysis and Application Program (WAsP) is comparable to Computational Fluid Dynamics (CFD) in order to use the WAsP wind prediction later for time consuming CFD simulations. Three different numerical methods (WAsP, RANS, LES) for observation of wind flow over the hills are described and compared with the wind-tunnel experiment. The paper shows that WAsP provides reasonably realistic results for the flow over the commonly found in nature shallow hills.

Keywords Flow over hills • LES • RANS • Turbulence • WAsP • Wind tunnel experiment

1 Introduction

Numerical modelling of atmospheric flows over a complex terrain is an important problem for wind energy applications, since it helps in arrangement, installation and control of the on-shore wind farms. The research subject is a justification of possible use of WAsP as prediction for the precise and time consuming CFD for a complex terrain. This is our second paper devoted to WAsP and CFD comparison for a wind flow over two-dimensional hills. In the first study [2], two hills (Hill3 and Hill5) from Castro and Apsley [5] were studied numerically and compared with the well known RUSHIL wind tunnel experiment [9]. The expected agreement was obtained for those hills [2]. In the present work, we continue to study the wind flow over two-dimensional hills as it represents the simplest flow over a complex terrain. Earlier, experiments for the hill flow were conducted using the hot-wired anemometry methodology. This time, we study the same shape of hills but with different slopes (Hill2 and Hill4). Meanwhile, computational results for these particular hills are compared with PIV measurements. At the same time, these hills are interesting for comparison because Hill2 has a very high slope and Hill4 is a borderline case

O. Agafonova (✉) • A. Koivuniemi • B. Conan • A. Chaudhari • H. Haario • J. Hamalainen
Lappeenranta University of Technology, P.O. Box 20, FI-53851 Lappeenranta, Finland
e-mail: oxana.agafonova@lut.fi

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_6

because of the reattachment zone beyond the hill. In addition, these two hills have not been studied enough neither numerically nor experimentally.

The equations, which describe the shape of the hills, are given in [1, 5]. The average slope of hills $n\left(n = \frac{H}{a}, a \text{ is half length and } H\text{—the height of the hill}\right)$ are equal to 0.5 and 0.25. In present study, these hills are named as Hill2 and Hill4 respectively. The experiment for Hill2 and Hill4 was conducted in the VKI-L2 wind tunnel of the von Karman Institute for Fluid Dynamics using the Particle Image Velocimetry (PIV) methodology. The hills were fabricated in wood in similarity with the RUSHIL experiment. The Reynolds number, based on the hill height, used in experiments is around 17,000. The inlet velocity and turbulence profiles of the experiment are detailed in Conan [8].

2 Mathematical Modelling

The equations of motion for a viscous incompressible liquid (Navier-Stokes Equations) without body forces are obtained from the integral laws of mass and momentum conservation and are written in the form [1]:

$$\begin{cases} \nabla \cdot \mathfrak{v} = 0 \\ \rho \dfrac{d\mathfrak{v}}{dt} = -\nabla p + \mu \nabla^2 \mathfrak{v} \end{cases}$$

In scalar-tensor form the system of Navier-Stokes equations takes the following appearance:

$$\frac{\partial u_i}{\partial x_i} = 0; \tag{1}$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \frac{1}{\rho}\frac{\partial}{\partial x_j} t_{ij}. \tag{2}$$

Then, using Eqs. (1), (2) can be rewritten in the form (3):

$$\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j}(u_j u_i) = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \frac{1}{\rho}\frac{\partial}{\partial x_j} t_{ij}. \tag{3}$$

The components t_{ij} of the viscous stress tensor are equal to $t_{ij} = 2\mu s_{ij}$, where s_{ij} are the components of the strain-rate tensor $s_{ij} = \frac{1}{2}(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j})$.

Using the Reynolds averaging [12], the velocity component u_i can be represented in the form $u_i = U_i + u_i'$, where U_i and u_i' are mean and fluctuating components

respectively. Therefore, Eq. (1) leads to:

$$\frac{\partial U_i}{\partial x_i} = 0; \quad \frac{\partial u_i'}{\partial x_i} = 0. \tag{4}$$

Applying the averaging operation to Eq. (3), we obtain the Reynolds-averaged Navier-Stokes (RANS) equation:

$$\frac{\partial}{\partial t}(\rho U_i) + \frac{\partial}{\partial x_j}(\rho U_j U_i) = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}(\overline{t_{ij}} - \rho\overline{u_j' u_i'}), \tag{5}$$

where $\overline{t_{ij}} = 2\mu S_{ij}$ and $\rho\overline{u_j' u_i'}$ are the components of the viscous stress tensor and the Reynolds-stress tensor respectively.

Large Eddy Simulation (LES) is a computational technique in which the large eddies are computed and the smallest eddies are modelled. Using the filtration concept that is applying the volume-average filter to the original Navier-Stokes equations, the velocity component u_i can be written in the form $u_i = \overline{u}_i + u_i'$, where $\overline{u}_i$, u_i' denote the resolvable-scale filtered and subgrid scale (SGS) components respectively. For an incompressible flow, the continuity and Navier-Stokes equations assume the following form [12] :

$$\frac{\partial \overline{u}_i}{\partial x_i} = 0; \tag{6}$$

$$\frac{\partial \overline{u}_i}{\partial t} + \frac{\partial \overline{u_i u_j}}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x_i} + \nu \frac{\partial^2 \overline{u}_i}{\partial x_k \partial x_k}. \tag{7}$$

The WAsP equations are based on a linearization of the Navier-Stokes equations of motion. The second order terms of the Navier-Stokes equations are ignored in the solution, leading to the simple steady equations [10]:

$$\frac{\partial u_i}{\partial x_i} = 0; \tag{8}$$

$$u_{j0}\frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \frac{1}{\rho}\frac{\partial}{\partial x_j} t_{ij}, \tag{9}$$

where the velocities $(u_{10}, u_{20}, 0)$ are the components of the initial wind vector $\mathfrak{v}_0$. The flow is modeled as horizontal flow perturbations of u and is independent of the real wind velocity. This can be done using potential flow

$$u = \nabla \chi. \tag{10}$$

The boundary conditions for the model are such that the potential flow is zero far away from the point of interest

$$\nabla \chi|_{r=R} = 0, \tag{11}$$

and the flow is parallel to the ground surface

$$u_{30} = \left.\frac{\partial \chi}{\partial z}\right|_{z=0} = \overline{u}_0 \cdot \nabla h(r, \phi). \tag{12}$$

In the WAsP program the solution is derived by expressing the equation in cylindrical coordinates and writing the solution in a series of J-Bessel functions.

3 Numerical Modelling

The steady state problem is solved numerically using the CFD technique for different hill slopes. The finite-volume meshes for RANS simulations are created using the Gambit software and the total number of cells is 300,800 elements in each case. The minimum control volume dimension is $0.05\,\mathrm{H} \times 0.00205\,\mathrm{H}$ at the hill summit and is extended upwind and downwind uniformly, and vertically with expansion ratio of 1.025 using a geometric progression method. RANS equations are solved numerically using the ANSYS FLUENT software by applying the SST $k - \omega$ turbulence model with the so-called "Low Reynolds Corrections" treatment near the wall [3]. The inlet velocity profile used for the computation is logarithmic (for details see [5]). The inflow boundary conditions for the turbulent kinetic energy and the specific dissipation rate are obtained from the periodic flow simulations over the flat terrain of the same size as the computation domain for hill simulations.

In addition to RANS simulations, several LES calculations are also carried out for both Hill2 and Hill4. For this purpose, the 2D hill geometry is extruded in the spanwise direction in order to perform 3D LES (see Fig. 2). The finite volume mesh consisting of 7,875,000 cells is used for the LES simulation. The minimum control volume size at the hill summit is $0.125\,\mathrm{H} \times 0.0031\,\mathrm{H} \times 0.137\,\mathrm{H}$ in x, y and z directions, respectively. The standard Smagorinsky sub-grid scale model is used to model the smaller eddies. The mapping technique was used at the inlet boundary in order to generate the realistic upstream boundary layer for LES, as explained in [6]. Recently, we have used the recycling technique in LES for Hill3 [2, 7].

The LES computations for both hills are run with the automatic time-step by fixing the maximum Courant number to $Cu = 0.25$ until the physical time $t = 40\,\mathrm{s}$ is reached. Then, they are time averaged with all quantities over the last 30 s. In addition to time averaging, the results are also space averaged over the span-wise direction.

WAsP maps that describe similar hills are created for comparison with CFD. WAsP simulations are performed at real scale. The height of the hills is taken as

117 m and the computational domain is extended up to 1 km in vertical direction. Surface roughness is similar to RANS modelling. The flow-model parameters are configured in order to model a neutrally stratified situation. The actual flow is examined with reference sites along the stream-wise axis of the hill.

4 Results and Discussions

In this study the finite-volume method based open source OpenFOAM 2.2.2 and commercial ANSYS FLUENT 13.0 software are employed for LES and RANS calculations, correspondingly.

Figure 1 shows the mean velocity in stream-wise direction from the experiment for the steep hill. The size of the recirculation x/H can be estimated between 3.5 and 4.5.

Experimental results for Hill4 did not show the reverse flow beyond the hill. At the same time, the separation region for this hill was detected by both RANS and LES approaches (see Figs. 2 and 3b). In the case of RANS, the reattachment point is $x_r = 5.216\,\mathrm{H}$. The reattachment point predicted by LES is $x_r = 6.58\,\mathrm{H}$. In fact,

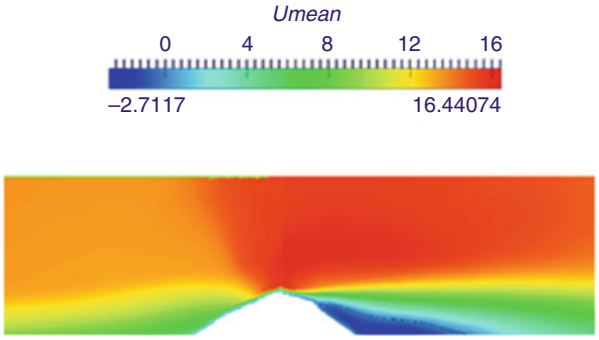


Fig. 1 Mean velocity [m/s] in stream-wise direction from the experiment [8] for Hill2

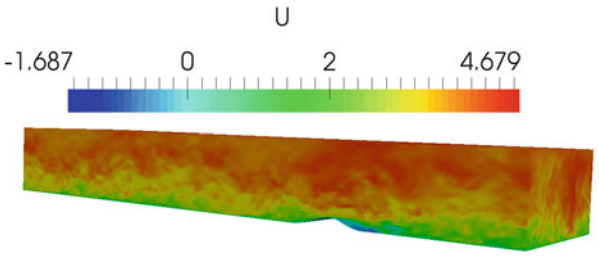


Fig. 2 Instantaneous velocity [m/s] in stream-wise direction from the simulation for Hill4

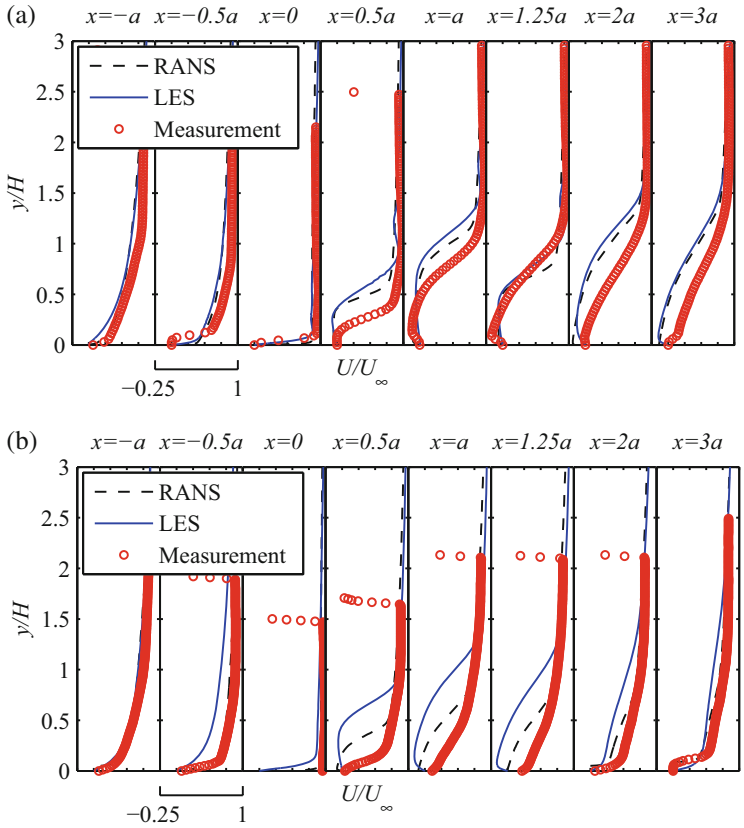


Fig. 3 Vertical profiles of the mean stream-wise velocity (U/U_∞) compared with measurements for the flow over Hill2 (**a**) and Hill4 (**b**) at certain longitudinal locations

the reattachment length in the experiment by Loureiro et al. [11] for a shallower hill (average slope is 0.2) than Hill4 equals 6.67 H.

Figure 3a, b show the vertical profiles of the stream-wise velocity compared with the measurements. The current RANS agrees with the experiment well enough for Hill2 (see Fig. 3a). Both RANS and LES underestimate the wind velocity in the separation region of Hill2.

It can be seen in Fig. 4a that WAsP significantly differs and overestimates the wind velocity in reattachment region of steep Hill2. However, Fig. 4b shows that WAsP agrees well enough for the shallow hill (Hill4) except the separation region, in which WAsP slightly overestimates the velocity.

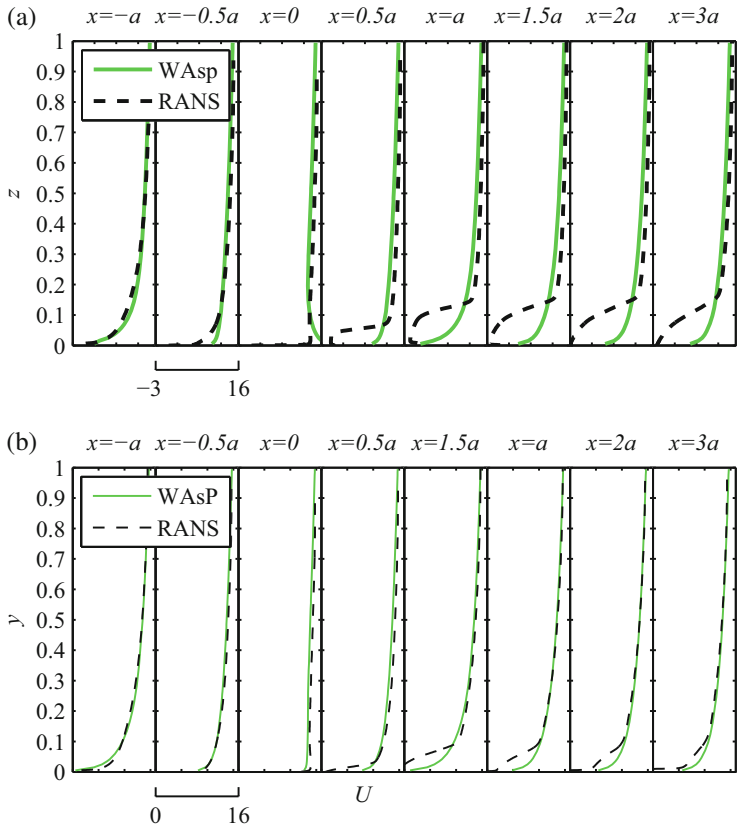


Fig. 4 Vertical profiles of the mean stream-wise velocity (U, [m/s]) compared with RANS for the flow over Hill2 (**a**) and Hill4 (**b**) at certain longitudinal locations

5 Conclusions

In the current study, we carried out WAsP, RANS and LES to investigate the turbulent flows over two-dimensional hills with different slopes. The results of RANS and LES simulations were described and compared with available experimental data. The obtained RANS results agree relatively well with LES and measurements and might be used for later validation of WAsP results.

WAsP produces reasonably realistic results for the flow over the shallow hills which are commonly found in nature (see Fig. 4b) and corresponds well with the other studies [2, 4].

This research work shows that the WAsP wind prediction agrees relatively well with RANS and can be used as an inflow for performing later RANS simulations over the real-scale and complex wind park terrain with shallow hills, forest and wind turbines.

Acknowledgements The authors would like to thank the CSC-IT Center for Science Ltd, Espoo (Finland) for their valuable computation facilities and support. The work presented in this paper is connected to the RENEWTECH project which was aimed for the development of wind power technology and business in South Finland (2012–2014) and was funded by the European Regional Development Fund (ERDF).

References

1. Agafonova, O.: Computational and experimental study of an air flow over a hill. Master thesis, Lappeenranta (2011)
2. Agafonova, O., Koivuniemi, A., Chaudhari, A., Haario, H., Hamalainen J.: Limits of WAsP modelling in comparison with CFD for wind flow over two-dimensional hills. In: Proceedings of EWEA 2014 Annual Event, Europe's Premier Wind Energy Event, Barcelona, Spain (2014)
3. Ansys Fluent Theory Guide. Online document (2011)
4. Bowen, A.J., Mortensen, N.G.: Exploring the limits of WAsP. In: Proceedings of European Union Wind Energy Conference, Göteborg (1996)
5. Castro, I.P., Apsley, D.D.: Flow and dispersion over topography: a comparison between numerical and laboratory data for two-dimensional flows. In: Atmospheric Environment, vol. 31, pp. 839–850. Elsevier B.V., Amsterdam (1997)
6. Chaudhari, A.: Large-eddy simulations of wind flows over complex terrains for wind energy applications. Ph.D. thesis, Lappeenranta University of Technology (2014)
7. Chaudhari, A., Hellsten, A., Agafonova, O., Hamalainen, J.: Large eddy simulation of boundary-layer flows over two-dimensional hills. In: Fontes, M., Gunther, M., Marheineke, N. (eds.) Progress in Industrial Mathematics at ECMI 2012. Mathematics in Industry, pp. 211–218. Springer International Publishing, Berlin (2014)
8. Conan, B.: Wind resource assessment in complex terrain by wind tunnel modelling. Ph.D. thesis, Orleans University (2012)
9. Khurshudyan, L.H., Snyder, W.H., Nekrasov, L.V.: Flow and Dispersion of pollutants over two-dimensional hills. United States Environmental Protection Agency Report, EPA-600/4-8 I-067 (1981)
10. Koivuniemi, A.: Identifying and addressing sources of uncertainty in modeling wind power production. Master thesis, Lappeenranta University of Technology (2011)
11. Loureiro, J., Pinho, F., Freire, S.A.: Near wall characterization of the flow over a two-dimensional steep smooth hill. Exp. Fluids **42**(3), 441–457 (2007)
12. Wilcox, D.: Turbulence Modelling for CFD. DCW Industries Inc, La Canada (1994)

Tuning Parameters of Ensemble Prediction System and Optimization with Differential Evolution Approach

Vladimir Shemyakin and Heikki Haario

Abstract Ensemble Prediction System (EPS) is the approach used in present day weather predictions to estimate the uncertainty of predictions. Along with the main prediction an ensemble of simulations is launched with perturbed initial values. Recently, the EPS with simultaneous parameter estimation approach (EPPES) has been proposed to tune model parameters online, without additional computational costs, by perturbing the parameter values and monitoring the respective performances. The key point of EPPES is the estimation of the parameter covariance by sequentially updating the covariance as hyperparameters by aid of importance weights. Here, we study the Differential Evolution (DE) optimization approach as a new way to solve the problem as a stochastic optimization task. We show that the convergence is improved using DE, especially in case when initial values of model parameters are far enough from the true ones.

Keywords DE • EPPES • EPS • Importance weights

1 Introduction

A number of mathematical models have been proposed to forecast the weather by taking into account its current state and range of measured data. Models differ in applicability for specific purpose, complexity and forecast power. The most efficient models make possible to continuously provide reliable predictions, estimate its uncertainty and adopt model parameters with new available data. The ability to tune the parameters with minimal computational cost is a crucial requirement for such complex models.

The core idea of EPS is to launch several predictions with slightly perturbed initial conditions to generate possible future states of the model. The main sources of uncertainties are the chaotic nature of the system and model bias caused by

V. Shemyakin (✉) • H. Haario
Lappeenranta University of Technology, Lappeenranta, Finland
e-mail: vladimir.shemyakin@lut.fi; heikki.haario@lut.fi

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_7

simplifications, approximations in calculations and rounding errors. EPS is designed to take into account these issues.

EPPES [1, 2] extends the basic function of EPS by simultaneous tuning the model parameters online, without additional computational cost. The crucial part of EPPES is the sequential estimation of hyperparameters of the underlying parameter distribution. The estimation of hyperparameters in EPPES is done using resampling with importance weights. When the cost function is specified, the importance weights for each proposed ensemble member are calculated as relative goodness with respect to other members. Further, these importance weights are used to update the distribution of hyperparameters from which the next ensemble will be drawn to continue prediction and estimation processes.

Here, we study the performance of the DE algorithm to enhance the convergence of EPPES style parameter estimation problems. DE [3, 4] belongs to the class of Evolutionary Algorithms (EA) and is based on the vector differences. Although the original DE approach is primarily designed for numerical optimization problems, we modify it in order to apply it for estimation of chaotic and stochastic optimization tasks. According to the provided test case, modified DE demonstrates clearly improved convergence from poor initial values with respect to EPPES.

2 Background

2.1 Lorenz-95 System

A conventional test case for the estimation of the chaotic behaviour is Lorenz-95 system. In order to explicitly represent parameters involved in the system the linear parametrization is added and the target model has the form:

$$\frac{dx_k}{dt} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - g_U(x_k), \tag{1}$$

where $k = 1, \ldots, 40$, $F = 8.2$, $g_U = \theta_0 + \theta_1 x_k$, $\hat{\theta} = (\theta_0, \theta_1)$ is parameter vector to be estimated.

Figure 1a, b demonstrate that the chaotic nature appears when the system is run both with small initial perturbations and small changes in underlying parameters values. Nevertheless, it can be seen that there is a time interval wherein the system behaves deterministically. This interval is called assimilation or time window. Further, the states of the system are divided into such sequential assimilation windows with corresponding measured data [5].

In order to simulate a real life parameter estimation procedure, we assume that a data assimilation mechanism provided. For the test purposes this is done by generating the synthetic data with known parameters and adding noise to it. It also makes possible the a measuring the reliability of the estimating process by comparison the resulting parameters with the original ones.

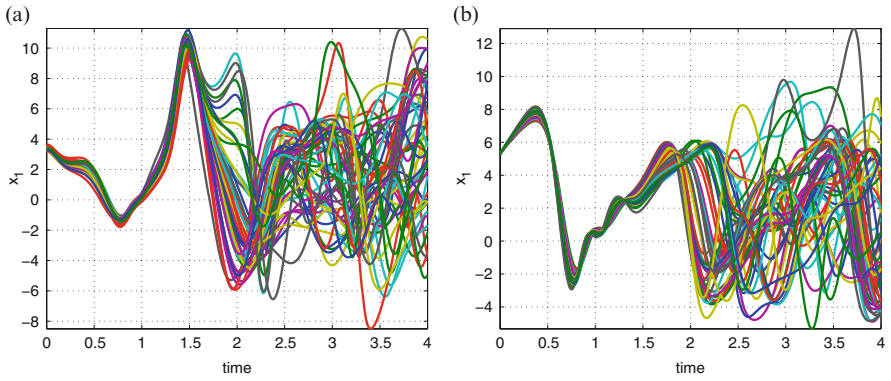


Fig. 1 Chaotic nature caused by (**a**) initial perturbations and (**b**) perturbations in parameters values

Hence, the common assumption is that there are several measurements available for each time window g and they serve to calculate a cost function, which can be computed in the least square way:

$$ss_g(\theta) = \sum_{k=1}^{K} \sum_{j=1}^{J} (x_k(t_{g,j}, \theta) - Y_k(t_{g,j}))^2, \tag{2}$$

where k is a state number, j is a measurement number within a time window g, K is a total number of states, J is a total number of measurements within a time window g, x_k is a state value for a given time point $t_{g,j}$ and a parameter value θ, Y_k is a measurement of a corresponding state for a given time point $t_{g,j}$.

The following set-up is used both for DE and EPPES cases:

- A parameter value $\hat{\theta} = (0, 0)$ is used for generating the synthetic data.
- A normally distributed noise drawn from a distribution $\mathcal{N}(0, \sigma_a^2)$ with $\sigma_a = 0.1$ is added to the generated data to simulate the measurement errors.
- The noisy data is divided into sequential assimilation windows of length 1.6 time units.
- There are 40 observable states.
- There are four measurements available within a time window for each state of the system, thus $J = 4$ in (2).
- A population/ensemble size is 51 elements, thus $K = 51$ in (2).
- Small perturbations in initial values following the normal distribution $\mathcal{N}(0, \sigma_{ip}^2)$ with $\sigma_{ip} = 0.01$ are used.

The last two items directly correspond to the EPS ideas. Since the Lorenz-95 system is highly sensitive to initial values, the ensemble with small perturbations of initial values with respect to assimilated ones are used in order to obtain the possible distribution of future states of the system. Such approach is designed to take into account different sources of possible errors involved into dynamic of the

system. Then, decision of goodness of a particular element in ensemble is based on a corresponding cost function value.

2.2 EPPES

The original concept and the method implementation have been presented by Järvinen et al. [1, 2]. The main idea of the EPPES is a hierarchical model of parameters. It means that these parameters are considered to be normally distributed with unknown hyperparameters which are being estimated and sequentially updated at each assimilation window g during EPPES run:

$$\theta_g \sim \mathcal{N}(\mu, \Sigma),\ \mu \sim \mathcal{N}(\mu_{g-1}, W_{g-1}),\ \Sigma \sim \mathcal{W}^{-1}(\Sigma_{g-1}, n_{g-1}). \tag{3}$$

The key part of EPPES is the importance weights idea. Importance weights provide information about relative goodness of a particular ensemble member with respect to the other members. With a cost function ss, given, for example, by (2), the importance weight for the n-th ensemble member θ_g^n in a Np-member population can be calculated as follows:

$$w_n = e^{-\frac{1}{2}ss(\theta_g^n)} / \sum_{i=1}^{Np} e^{-\frac{1}{2}ss(\theta_g^i)} \tag{4}$$

At each time window g, the estimation of the hyperparameters consists of several steps. Firstly, the proposal values for the parameters are sampled from the multivariate normal distribution $\mathcal{N}(\mu_{i-1}, \Sigma_{i-1})$ with the hyperparameters calculated in the previous time window $g-1$. Then, for each proposal value of the parameters the importance weight is calculated as stated in (4). These importance weights are used to resample the ensemble of parameters generated earlier. Finally, this resampled ensemble θ_g provides the information for updating the hyperparameters using a following set of formulas:

$$W_g = (W_{g-1}^{-1} + \Sigma_{g-1}^{-1})^{-1},\ \mu_g = W_g(W_{g-1}^{-1}\mu_{g-1} + \Sigma_{g-1}^{-1}\theta_g),$$

$$n_g = n_{g-1} + 1,\ \Sigma_g = (n_{g-1}\Sigma_{g-1} + (\theta_g - \mu_g)(\theta_g - \mu_g)')/n_g. \tag{5}$$

It has been shown that the EPPES approach is applicable for an on-line estimation of chaotic problems with changing data, see [6–8]. However, its ability to converge depends on specified ensemble size and provided prior distributions for hyperparameters. The key property is that it converges to the distribution of parameter vectors, not to a single parameter vector.

2.3 Differential Evolution

Differential evolution, introduced by Price and Storn [3], belongs to the class of Evolutionary algorithms. This algorithm has been shown as a powerful tool for solution of nonlinear and complex optimization problems due to its performance and simplicity in implementation. DE is a population-based optimizer as well as others EAs. The population consists of predefined number (Np) of D-dimensional vectors, where D is the dimension of the parameter space. One full evolution step contains 4 main stages: initialization, mutation, crossover and selection.

DE operates with three population within one full step:

- **Current population** ($P_{\mathbf{x},g}$) is the population after initialization step.
- **Intermediate population** ($P_{\mathbf{v},g}$) is the population of mutant vectors.
- **Trial population** ($P_{\mathbf{u},g}$) is the population of mutant vectors after crossover step of the algorithm.

All mentioned populations have the similar structure, for example, the current population $P_{\mathbf{x},g}$ consists of vectors $(\mathbf{x}_{i,g}) = (x_{j,i,g})_{j=1,\ldots,D}$, where g is a generation number, i is a population member number.

The description of evolutions steps [4] can be stated as follows:

1. **Initialization.** The initial population for the first generation ($g = 0$) is usually drawn uniformly from the specified searching domain:

$$(x_{j,i,0}) = r_j \cdot (b_{j,U} - b_{j,L}) + b_{j,L}, \tag{6}$$

   where $r_j \sim \mathcal{U}(0,1)$, $b_{j,U}$ and $b_{j,L}$ are upper and lower boundaries for the corresponding dimension j. The initial population for the next generation is the population that survived after the previous selection step.
2. **Mutation.** The most common mutation scheme is called "DE/rand/1/bin" which corresponds to a mutant vector $\mathbf{v}_{i,g}$ calculation as follows:

$$\mathbf{v}_{i,g} = \mathbf{x}_{r_1,g} + F \cdot (\mathbf{x}_{r_2,g} - \mathbf{x}_{r_3,g}), \tag{7}$$

   where $\mathbf{x}_{r_1,g}$, $\mathbf{x}_{r_2,g}$ and $\mathbf{x}_{r_3,g}$ are mutually different members of the current population, $F \in (0, \infty)$ is a scale factor which controls the evolution rate.
3. **Crossover.** Trial population is calculated according to the following equation:

$$\mathbf{u}_{i,g} = u_{i,j,g} = \begin{cases} v_{i,j,g}, & \text{if } r_j \leq Cr \text{ or } j = j_r \\ x_{i,j,g}, & \text{otherwise.} \end{cases} \tag{8}$$

   where $Cr \in [0,1]$ is a crossover probability, $x_{i,j,g}$ is a target vector, $r_j \sim \mathcal{U}(0,1)$. Additional requirement $j = j_r$, where j_r is a random index, ensures that the trial vector differs from the target vector at least in a one component.

4. **Selection.** The selection is based on comparison of a provided cost function f values of target and trial vectors:

$$\mathbf{x}_{i,g+1} = \begin{cases} \mathbf{u}_{i,g} & \text{if } f(\mathbf{u}_{i,g}) \leq f(\mathbf{x}_{i,g}) \\ \mathbf{x}_{i,g}, & \text{otherwise.} \end{cases} \tag{9}$$

After one entire evolution step of DE, the process is repeated until the desired criteria for the population is met. For more details about classical DE, see [4].

3 Parameter Estimation with Single Cost Function

3.1 DE Modification for Stochastic Cost Function

In order to adopt the original DE to work with the estimation of chaotic dynamics, we introduce a number of significant modifications to the original algorithm. Also, several known improvements has been utilized [9, 10].

Since each assimilation window has new data, there is no fixed cost function. In this case, each generation is considered only within corresponding assimilation window and produces a descendant population for the next window. If one applies the usual scheme of DE to this problem, he can face with two main problems. Firstly, for every time window it is necessary to calculate the cost function both for the current and trial populations in order to make the selection, what doubles the number of evaluation of the computational costly function. Secondly, by doing so, we entirely lose the information between the previous and present time windows, which causes the risk of rejecting, by chance, well performing parameters. That is why it is essential to preserve both the parameters and corresponding values of the cost function to be able to compare sequential populations and reduce computational cost of the algorithm. Therefore, the modified population structure of DE for a generation/assimilation window g has the following form:

- $\mathbf{x}_{i,g} = (x_{i,j,g})$ is a population member,
- $ss_{i,g}$ is the corresponding cost function value,
- $P_{\mathbf{x},s,g} = (\mathbf{x}_{i,g}, ss_{i,g})$ is current population information.

Thus, in the time window g we have the current population $\mathbf{x}_{i,g}$ with corresponding cost function values $ss_{i,g}$, which may come from one of the previous selection steps. Then, by mutation and crossover procedures we obtain the trial population $\mathbf{u}_{i,g}$. In order to compare current and trial populations we utilize the cost function for present assimilation window denoted as f_g, which is calculated in the least square way (2). Thereby, the selection decision is made according to the comparison of just calculated cost function values for the trial population $f_g(\mathbf{u}_{i,g})$ and stored cost

function values $ss_{i,g}$:

$$P_{\mathbf{x},s,g+1} = (\mathbf{x}_{i,g+1}, ss_{i,g+1}) = \begin{cases} (\mathbf{u}_{i,g}, f_g(\mathbf{u}_{i,g})) & \text{if } f_g(\mathbf{u}_{i,g})) \leq ss_{i,g} \\ (\mathbf{x}_{i,g}, ss_{i,g}) & \text{otherwise} \end{cases} \tag{10}$$

Note the stochastic nature of the cost function: the values $ss_{i,g}$ and $f_g(\mathbf{u}_{i,g}))$ are computed with different data. Hence, it may happen that the specific population member with the corresponding cost function value survives during several sequential assimilation windows until being compared to a more promising element of the trial population.

3.2 Comparison of DE and EPPES for Parameter Estimation Problem

The DE approach with discussed modification and improvements is applied to the parameter estimation of Lorenz-95 system. The description and set-up for this problem is explained earlier in Sect. 2.1.

An application of the DE to this test case demonstrates behaviour of the algorithm in details starting from the initialization and continuing until the estimated parameters in the last generation. Figure 2a–d depict this process. Here, each figure consists of upper and lower plots. The upper ones show the actual data for a fixed state (the first state out of 40 in this case) and a specific assimilation window by circles and behaviour of the population within this window by a set of continuous curves. The lower plots depict a distribution of parameters within the population for the current time window. Each generation of DE corresponds to particular time window.

We can see that DE is able to reasonable accurately estimate the true values of $\hat{\theta} = (0, 0)$ which were used to generate the data. However, it should be emphasised that DE is optimizer, hence, it tends to converge to one parameter vector with non-significant oscillation around unlike EPPES which is devoted to find the distribution of parameter vectors. Further, it is important to compare the convergence of the algorithm with EPPES. Figure 3 is devoted to make the comparison. The initial population for DE is uniformly drawn from the interval $[-10;10]$ for both parameters. Solid lines in the plot correspond to population/ensemble means for target parameters while dashed lines are intervals of $\pm 3\sigma$, where σ is population/ensemble standard deviations.

We can conclude that although both DE and EPPES are able to estimate the parameters of the problem, DE has faster convergence and, moreover, is able to succeed even from the poor initial population drawn from the huge interval.

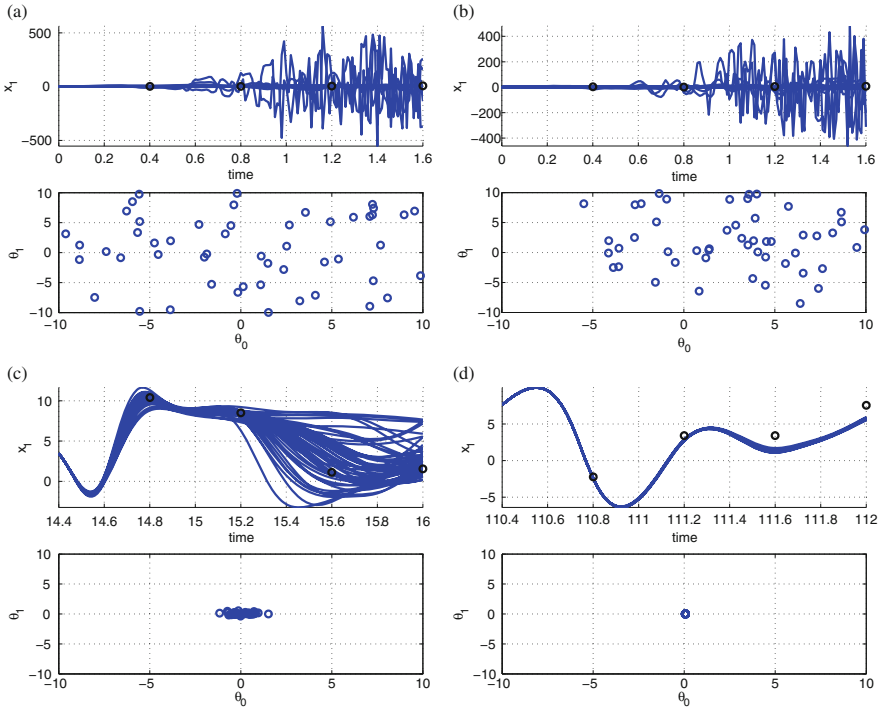


Fig. 2 Conformity of real and estimated data; distribution of parameters at different time windows. (**a**) The 1st time window after initialization of DE population. (**b**) The 1st time window and the 1st DE population. (**c**) The 10th time window and the 10th DE population. (**d**) The 70th time window and the 70th DE population

4 Results

We have presented the Lorenz-95 system with linear parametrization in order to use as a test case for estimation purposes. Also, the core ideas of original algorithms of both EPPES and DE have been explained. Further, we have modified original DE method to be applicable to the problem of the parameter estimation of chaotic dynamics and tested it on the described Lorenz-95 system. Test runs have proved the applicability of the suggested method. Moreover, the comparison between EPPES and DE has been demonstrated in order to show the convergence properties of each approach. Combining the benefits from both approaches becomes a valuable field for the further investigation of the algorithms together as building elements for a more general method.

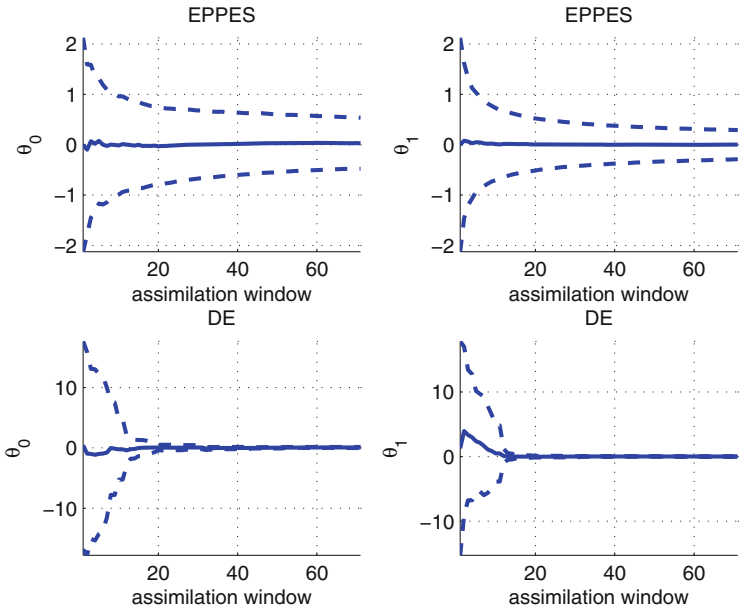


Fig. 3 EPPES vs DE

References

1. Järvinen, H., Laine M., Solonen A., Haario H.: Ensemble prediction and parameter estimation system: the concept. Q. J. R. Meteorol. Soc. **138**(663), 281–288 (2011)
2. Laine, M., Solonen, A., Haario, H., Järvinen, H.: Ensemble prediction and parameter estimation system: the method. Q. J. R. Meteorol. Soc. (2011). doi:10.1002/qj.922
3. Price, K.V., Storn, R.M.: Differential evolution a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report TR-95-012, ICSI, March (1995)
4. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential Evolution: Practical Approach to Global Optimization, pp. 1–130. Springer, Berlin (2005)
5. Lahoz, W., Khattatov, B., Menard, R.: Data Assimilation: Making Sense of Observations. Springer Science & Business Media, Berlin (2010)
6. Ollinaho, P., Bechtold, P., Leutbecher, M., Laine, M., Solonen, A., Haario, H., Järvinen, H.: Parameter variations in prediction skill optimization at ECMWF. Nonlinear Process. Geophys. **20**(6), 1001–1010 (2013). doi: 10.5194/npg-20-1001-2013
7. Ollinaho, P., Laine, M., Solonen, A., Haario, H., Järvinen, H.: NWP model forecast skill optimization via closure parameter variations. Q. J. R. Meteorol. Soc. **139**(675), 1520–1532 (2013). doi:10.1002/qj.2044
8. Ollinaho, P., Järvinen, H., Bauer, P., Laine, M., Bechtold, P., Susiluoto, J., Haario, H.: Total energy norm in NWP closure parameter optimization. Geosci. Model Dev. Discuss. **6**(4), 6717–6740 (2013). doi:10.5194/gmdd-6-6717-2013. Available: http://www.geosci-model-dev-discuss.net/6/6717/2013/
9. Chakraborty, U.K.: Advances in Differential Evolution, pp. 1–15. Springer, Berlin
10. Qing, A.: Differential Evolution: Fundamentals and Applications in Electrical Engineering, pp. 62–64. Wiley, New York (2009)

MS 3
MINISYMPOSIUM: COMPUTATIONAL FINANCE

Organizers

Matthias Ehrhardt[1], Jörg Kienitz[2] and Jan ter Maten[3]

Speakers

Alfio Borzi[4]
A Fokker-Planck Strategy to Control Stochastic Processes

Maria Grossinho[5]
A Class of Nonlinear Boundary Value Problems for a Black-Scholes Type Equation

Christoph Reisinger[6]
Numerical Solution of Stochastic PDEs Arising in Financial Engineering

[1]Matthias Ehrhardt, University of Wuppertal, Wuppertal, Germany.
[2]Jörg Kienitz, Postbank AG, Bonn, Germany.
[3]Jan ter Maten, University of Wuppertal, TU Eindhoven, The Netherlands.
[4]Alfio Borzi, University of Würzburg, Würzburg, Germany.
[5]Maria Grossinho, Instituto Superior de Economia e Gestão (ISEG), Lisbon, Portugal.
[6]Christioph Reisinger, University of Oxford, Oxford, UK.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_8

José Germán López Salas[7]
Efficient Calibration and Pricing in LIBOR Market Models with SABR Stochastic Volatility Using GPUs

Christof Heuer[8]
High-Order Compact Finite Difference Schemes for Parabolic Differential Equations with Mixed Derivative Terms in a Space Dimensions and Application to Basket Options

Marta Pou Bueno[9]
Extension of a Fourier-Cosine Method to Solve BSDEs with Higher Dimensions

Chiara Guardasoni[10]
A Boundary Element Method for Pricing Barriers Options

Matthias Ehrhardt[11]
Modelling Stochastic Correlation

Zuzana Bučková[12]
Fichera Theory and Its Application in Finance

Pedro Polvora[13]
Derivative Pricing Under Transaction Costs Using a Stochastic Utility Maximization Model

Vera Egorova[14]
A Positive, Stable and Consistent Front-Fixing Numerical Scheme for American Options

Walter Mudzimbabwe[15]
An Efficient Monte Carlo Algorithm for Pricing Arithmetic Asian Options Under a Jump Diffusion Process

[7]José Germán López Salas, University of A Coruña, A Coruña, Spain.
[8]Christof Heuer, University of Wuppertal, Wuppertal, Germany.
[9]Marta Pou Bueno, TU Delft, Delft, The Netherlands.
[10] Chiara Guardasoni, University of Parma, Parma, Italy.
[11]Matthias Ehrhardt, University of Wuppertal, Wuppertal, Germany.
[12]Zuzana Bučková, University of Wuppertal, Wuppertal, Germany.
[13]Pedro Polvora, Comenius University Bratislava, Bratislava, Slovakia.
[14]Vera Egorova, Universitat Politècnica de València, Valencia, Spain.
[15]Walter Mudzimbabwe, University Angel Kanchev, Ruse, Bulgaria.

Nicola Cantarutti[16]
Option Pricing in Exponential Levy Models with Transaction Costs

Ivan Yamshchikov[17]
Portfolio Optimization in the Case of an Asset with a Given Liquidation Time Distribution

Lara Trussardi[18]
Analysis of a Cross-Diffusion Herding Model in Social Economics

Mohamed El-Fakharany[19]
Numerical Solution of Partial-Integro Differential Option Pricing Models with Cross Derivative Term

Keywords

Computational finance
Financial mathematics
Option pricing

Short Description

In recent years the computational complexity of mathematical models employed in financial mathematics has witnessed a tremendous growth. Advanced numerical techniques are imperative for the most present-day applications in financial industry.

The aim is to deeper understand complex (mostly nonlinear) financial models and to develop effective and robust numerical schemes for solving linear and nonlinear problems arising from the mathematical theory of pricing financial derivatives and related financial products.

The motivation for this minisymposium is to exchange and discuss current insights and ideas, and to lay groundwork for future collaborations. Finally, it should serve as a kickoff for the ECMI special interest group (SIG) Computational Finance.

[16]Nicola Cantarutti, Instituto Superior de Economia e Gestão (ISEG), Lisbon, Portugal.
[17]Ivan Yamshchikov, Hochschule Zittau/Görlitz, Zittau, Germany.
[18]Lara Trussardi, TU Wien, Vienna, Austria.
[19]Mohamed El-Fakharany, Universitat Politècnica de València, Valencia, Spain.

An Efficient Monte Carlo Algorithm for Pricing Arithmetic Asian Options Under a Jump Diffusion Process

Walter Mudzimbabwe

Abstract We develop a Monte Carlo algorithm to price an Asian option whose underlying price is driven by a jump diffusion process. By conditioning on the number of jumps, we characterise the underlying asset process as lognormally distributed from which a control variate for the generic Monte Carlo algorithm is derived. Numeric results confirm that the control variate method is an effective variance reduction method.

Keywords Computational finance • Option pricing

1 Introduction

One of the most traded options is the Asian option whose payoff depends on the average price of the underlying. Due to their averaging nature, they have lower volatilities than the underlying asset, hence are cheaper that vanilla European options. They are also less prone to price manipulations.

A fixed-strike Asian option with maturity T, strike price $K > 0$ and whose underlying asset price S_t has payoff $\max(\frac{1}{T}\int_0^T S(t) - K, 0)$. To price this option, an expectation under a risk neutral measure of the discounted payoff is found (see [3]), i.e.,

$$e^{-rT}\mathbb{E}\left(\max\left(\frac{1}{T}\int_0^T S(t) - K, 0\right)\right), \tag{1}$$

The usual assumption is that price of the underlying asset $S(t)$ is lognormal. The pricing problem then hinges on the distribution of a sum of lognormal variables which is unknown (see e.g., [2]). This is the reason why a closed form solution has not been developed.

W. Mudzimbabwe (✉)
Ruse University, Studentska str. 8, 7017 Ruse, Bulgaria
e-mail: wmudzimbabwe@uni-ruse.bg

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_9

In this article, we generalise the price dynamics so that the price follows a jump-diffusion process The jumps could be interpreted as arrival of information such as mergers, takeovers etc. Under this generalisation, the distribution of $S(t)$ is intractable. Among the several methods that are used to price Asian options, is the Monte Carlo method. We develop an efficient control variate that can be used to make the Monte Carlo method efficient.

There is a vast literature on the Asian options under jump-diffusion. In [4], they develop a double-Laplace inversion method. Albrecher [1], derives an algorithm to calculate moments of the sum of the underlying then replacing it by a tractable one such as lognormal model. In [10], the Asian option is characterised by a partial integro-differential equation (PIDE). Of interest is the work by Schoutens and Symens [9] where a Monte-Carlo method based on control variate technique is explored. We develop our control variate based on a related Asian option.

The paper is organised as follows. Section 2 describes the price process model. In Sect. 3, we develop the control variate and Sect. 4 derives the Monte Carlo method. We demonstrate the control variate for different levels of parameters in Sect. 5. Conclusions are given in Sect. 6.

2 The Price Model

In this section we describe the price of the underlying dynamics $S(t)$. We assume that the $S(t)$ is driven by the following jump-diffusion stochastic differential equation (SDE):

$$dS(t) = S(t-)\left(\mu dt + \sigma dB(t) + (\Pi - 1)dY(t)\right), \tag{2}$$

where $S(\tau-) = \lim_{t\uparrow\tau} S(t)$; $B(t)$ is a Weiner process; $dY(t)$ is a Poisson process with intensity λ; Π is the jump size with expected value of $\nu + 1$; $dB(t)$, $dY(t)$ and $\Pi(t)$ are assumed to be mutually independent. We assume that Π is i.i.d lognormally distributed, i.e., $\ln(\Pi) \sim N(\mu_\pi, \sigma_\pi^2)$. This implies $\nu := \mathbb{E}(\Pi - 1) = \exp(\mu_\pi + \sigma_\pi^2/2) - 1$. These dynamics can also be found in [11], where they are used to model value of the firm assets in a structural Merton model of credit risk. See also [5] or [10].

Assuming a finite number of jumps and jump sizes, we can apply a general Ito formula (see, e.g., [7]) for semi-martingale processes such as (2). In this case, the solution for (2) is (assuming $S(0) = S$)

$$S(t) = S\exp\{(\mu - \sigma^2/2)t + \sigma B(t) + \ln(\Pi)Y(t)\}, \tag{3}$$

where $\ln(\Pi)Y(t) = \sum_{j=1}^{Y_t} \ln(\Pi_j)$, which is zero if $Y_t \equiv 0$.

Although the distribution of S_t is intractable, under certain assumptions it can be written explicitly. We also assume that in a small interval $\Delta t_i = t_i - t_{i-1}$, the number of jumps is at most one, i.e., Y_t is either 0 or 1. Using a uniform spacing, i.e., $\Delta t_i = \Delta t = T/N$, where N is the number of partitions of $[0, T]$. We may write (see, e.g. [11])

$$\ln(S(t_i)/S(t_{i-1})) = x_i + \Delta Y(t_i)\ln(\Pi)$$
$$x_i \sim N((\mu - \sigma_v^2/2)T/N, \sigma_v^2 T/N)\,;\, \Delta Y(t_i) = \begin{cases} 0, & \text{w.p } 1 - \lambda \cdot T/N \\ 1, & \text{w.p } \lambda \cdot T/N. \end{cases} \tag{4}$$

3 The Control Variate

In this section, we develop the control variate which will be used in the Monte Carlo algorithm. Similar to [8], the control variate will be based on the geometric average:

$$G_N = \left(\prod_{i=1}^{N} S_{t_i}\right)^{1/N}.$$

The price of a geometric average Asian option C^e is given by $C^e = e^{-rT}\mathbb{E}\left(\max(G_N - K, 0)\right)$. By conditioning on the number of jumps, we will show how a closed expression for the price of geometric average Asian option may be found. Under this assumption, S_t is lognormally distributed, just as in the case of pure Brownian motion, as we will see shortly. We may write

$$\begin{aligned}\log(G_N) = \frac{1}{N}\Bigg[&\log\left(\frac{S(t_N)}{S(t_{N-1})}\right) + 2\log\left(\frac{S(t_{N-1})}{S(t_{N-2})}\right) + 3\log\left(\frac{S(t_{N-2})}{S(t_{N-3})}\right) \\ &+ \ldots + (N-2)\log\left(\frac{S(t_3)}{S(t_2)}\right) + (N-1)\log\left(\frac{S(t_2)}{S(t_1)}\right) \\ &+ N\log\left(\frac{S(t_1)}{S}\right) + N\log(S)\Bigg].\end{aligned}$$

Conditioning on $\Delta Y_{t_i} = \Delta Y_t = k$, S_t is lognormally distributed, i.e., $\log\left(S(t_i)/S(t_{i-1})\right)|\,\Delta Y_t = k \sim N\left(\widehat{\mu}, \widehat{\sigma}^2\right)$, $\forall i = 1, \cdots, N$, where $\widehat{\mu} = (r - \sigma^2/2 - \lambda\nu)T/N + k\mu_\pi$ and $\widehat{\sigma}^2 = \sigma^2 T/N + k\sigma_\pi^2$. Consequently,

$$S_k(t_i) = S_k(t_{i-1})\exp\left\{\left(r - \frac{\sigma^2}{2} - \lambda\nu\right)\frac{T}{N} + \sigma B\left(\frac{T}{N}\right) + \sum_{j=1}^{k}\ln(\Pi_j)\right\}. \tag{5}$$

The subscript k, indicates the dependence of the price formula on the deterministic number of jumps. The expectation and variance of G_N can be found as

$$\begin{aligned}\mathbb{E}(\log(G_N)|\Delta Y_t = k) &= \frac{1}{N}\left(\widehat{\mu} + 2\widehat{\mu} + 3\widehat{\mu} + \cdots + N\widehat{\mu} + N\log(S)\right)\\ &= \log(S) + \left(\left(r - \frac{\sigma^2}{2} - \lambda\nu\right)\frac{T}{N} + k\mu_\pi\right)\frac{N+1}{2}.\\ \mathrm{Var}(\log(G_N)|\Delta Y_t = k) &= \frac{1}{N^2}\left(\widehat{\sigma}^2 + 4\widehat{\sigma}^2 + 9\widehat{\sigma}^2 + \cdots + N^2\widehat{\sigma}^2\right)\\ &= \left(\sigma^2\frac{T}{N} + k\sigma_\pi^2\right)\frac{(N+1)(2N+1)}{6N}, \end{aligned} \tag{6}$$

The following theorem characterises a closed form formula for the geometric average Asian option.

Theorem 1 *Assuming that $\Delta Y(t)$ follows (4), then*

$$C^e = \left(1 - \lambda\frac{T}{N}\right)C_{BS}\left(Se^{(\widetilde{r_0}-r)T}, T; K, r, \tilde{\sigma}_0^2\right) + \lambda\frac{T}{N}C_{BS}\left(Se^{(\widetilde{r_1}-r)T}, T; K, r, \tilde{\sigma}_1^2\right),$$

where $C_{BS}(S, T; K, r, \sigma^2)$ is the Black-Scholes formula for the price of a European call option.

Proof By conditioning on $\Delta Y(t) = k$, where $k = 0, 1$ and using the law of iterated expectation, we can write down the value of geometric average Asian option as

$$\begin{aligned} C^e &= e^{-rT}\mathbb{E}\left[\mathbb{E}\left((G_N - K)^+ |\Delta Y(t) = k\right)\right]\\ &= e^{-rT}\sum_k \mathbb{E}\left(\left(\exp(\log(G_N|\Delta Y(t) = k)) - K\right)^+\right)P(\Delta Y(t) = k), \end{aligned} \tag{7}$$

where $P(\Delta Y(t) = k)$ is calculated using (4) and $x^+ = \max(x, 0)$.

We can compare this with a synthetic European call whose log of the price of the underlying asset $\log(\widetilde{S}(T))$ is normal with mean $\log(S) + (\tilde{r}_k - \tilde{\sigma}_k^2/2)T$ and variance $\tilde{\sigma}_k^2 T$. Therefore

$$\begin{aligned}\tilde{\sigma}_k^2 &= \sigma^2\left(\frac{1}{N} + \frac{k\sigma_\pi^2}{T\sigma^2}\right)\frac{(N+1)(2N+1)}{6N},\\ \tilde{r}_k &= \frac{1}{2}\tilde{\sigma}_k^2 + \left(\left(r - \frac{\sigma^2}{2} - \lambda\nu\right)\frac{1}{N} + \frac{k\mu_\pi}{T}\right)\frac{N+1}{2}.\end{aligned}$$

Using properties of the Lognormal distribution it follows that

$$e^{-rT}\mathbb{E}\left((\exp(\log(G_N|\Delta Y(t_i)=k))-K)^+\right) = Se^{(\tilde{r}_k-r)T}N(\tilde{d}_1)-Ke^{-rT}N(\tilde{d}_2),$$
$$= C_{BS}\left(Se^{(\widetilde{r_k}-r)T},T;K,r,\tilde{\sigma}_k^2\right),$$

where $C_{BS}(S,T;K,r,\sigma^2)$ is the Black-Scholes formula for the price of a European call option, $N(.)$ is the cumulative normal distribution function and

$$\tilde{d}_1 = \frac{\log(S/K)+\left(\tilde{r}_k+\tilde{\sigma}_k^2/2\right)T}{\tilde{\sigma}_k\sqrt{T}},\quad \tilde{d}_2=\tilde{d}_1-\tilde{\sigma}_k\sqrt{T}.$$

The result follows from (7).

4 Monte Carlo Pricing

In this section, we consider Monte Carlo algorithm to price the fixed-strike Asian option. As in the case of the geometric Asian option we first condition on the number of jumps. In practice, the continuous sum in (1) is usually an approximation of the arithmetic average $A_N = \frac{1}{N}\sum_{i=1}^N S(t_i)$ and so the price is given by $C^a = e^{-rT}\mathbb{E}((A_N-K)^+)$. It is worth remarking that, the distribution of $S(t)$ is not known since it is the exponentiation of a sum of compound Poisson process and Brownian motion. Likewise, the distribution A_N is also intractable. We may improve tractability of $S(t)$ by considering the random variable $(B(t),\Pi(t))$ *in leu* of $(B(t),Y(t),\Pi(t))$ by conditioning on the number of jumps. Similar to (7),

$$C^a = e^{-rT}\left(\left(1-\lambda\frac{T}{N}\right)\mathbb{E}\left(A_{0,N}-K\right)^+ + \lambda\frac{T}{N}\mathbb{E}\left(A_{1,N}-K\right)^+\right) \tag{8}$$

where $A_{k,N} = \frac{1}{N}\sum_{i=1}^N S_k(t_i)$, $k=0,1$. Though $S(t)$ is now lognormal, the corresponding sum A_N does not have an explicit distribution. This is the reason why a closed form formula for the price the arithmetic Asian option has not be developed, see e.g., [2].

One method that has been used in such cases is the Monte Carlo method, see [6]. Due to the computational demands variance reduction methods are used to improve accuracy for moderate number of Monte-Carlo simulations. To illustrate the control variate Monte Carlo technique (see also [9]), let

$$X_j^k = e^{-rT}\left(\frac{1}{N}\sum_{i=1}^N S_k(t_i)-K\right)^+, Y_j^k = e^{-rT}\left(\left(\prod_{i=1}^N S_k(t_i)\right)^{\frac{1}{N}}-K\right)^+,$$

$$Z_j^k(b) := X_j^k + b^k\left(Y_j^k-\mathbb{E}\left[Y^k\right]\right),\quad b^k\in\mathbb{R},\ k=0,1.$$

As a consequence of Theorem 1, $\mathbb{E}[Y^0] = C\left(Se^{(\widetilde{r_0}-r)T}, T; K, r, \tilde{\sigma}_0^2\right)$ and $\mathbb{E}[Y^1] = C\left(Se^{(\widetilde{r_1}-r)T}, T; K, r, \tilde{\sigma}_1^2\right)$, so that $Y^k - \mathbb{E}^Q(Y^k)$ serves as control in estimating $\mathbb{E}^Q(X^k)$, our variable of interest. The control variate Monte Carlo then becomes the computation of the sample mean $\overline{Z^k(b)} = \frac{1}{M}\sum_{i=1}^{M} Z_j^k(b)$.

5 Numerical Results

In this section we perform numeric calculation based on the control variate. We compare an ordinary Monte Carlo and the control variate Monte Carlo. We also include a 95 % confidence interval (CI) for each calculation that we make. The table shows that the later method is efficient in reducing the variance.

In Table 1, we tabulate a 95 % confidence interval (CI) for both ordinary Monte Carlo method (MC) and control variate Monte Carlo method (CMC) for $T = 1$. It is apparent that our control variate method is effective as a variance reduction method by considering the ratio of confidence interval lengths i.e., the ratio of standard deviation of ordinary Monte Carlo method (MC) $\hat{\sigma}_{\text{MC}}$ to standard deviation

Table 1 Comparison of ordinary Monte Carlo method (MC) with control variate Monte Carlo method (CMC) for $T = 1$

σ	K	σ_π	MC	CI	CMC	CI	$\hat{\sigma}_{\text{MC}}/\hat{\sigma}_{\text{CMC}}$
0.05	95	0.01	8.8474	[8.8300, 8.8647]	8.8428	[8.8415, 8.8442]	12.9859
		0.1	8.8476	[8.8301, 8.8650]	8.8307	[8.8290, 8.8324]	10.2720
	100	0.01	4.3493	[4.3326, 4.3659]	4.3468	[4.3454, 4.3481]	12.3486
		0.1	4.3491	[4.3325, 4.3657]	4.3352	[4.3335, 4.3369]	9.7379
	105	0.01	0.9852	[0.9754, 0.9950]	0.9851	[0.9837, 0.9866]	6.8677
		0.1	0.9999	[0.9902, 1.0097]	0.9796	[0.9778, 0.9813]	5.5096
0.1	95	0.01	8.9415	[8.9079, 8.9752]	8.9474	[8.9453, 8.9496]	15.9013
		0.1	8.9483	[8.9146, 8.9820]	8.9364	[8.9341, 8.9388]	14.2763
	100	0.01	4.9543	[4.9252, 4.9834]	4.9556	[4.9534, 4.9577]	13.4980
		0.1	4.9510	[4.9220, 4.9801]	4.9461	[4.9437, 4.9485]	12.1640
	105	0.01	2.1161	[2.0954, 2.1368]	2.1051	[2.1029, 2.1073]	9.3769
		0.1	2.1189	[2.0984, 2.1395]	2.0993	[2.0969, 2.1018]	8.4081
0.8	95	0.01	21.5786	[21.3298, 21.8274]	21.4896	[21.4726, 21.5066]	14.6339
		0.1	21.4544	[21.2082, 21.7006]	21.4522	[21.4354, 21.4690]	14.6497
	100	0.01	19.4092	[19.1668, 19.6516]	19.4374	[19.4203, 19.4546]	14.1610
		0.1	19.3115	[19.0740, 19.5491]	19.3970	[19.3802, 19.4138]	14.1123
	105	0.01	17.5957	[17.3620, 17.8293]	17.5750	[17.5578, 17.5921]	13.6104
		0.1	17.5330	[17.3015, 17.7645]	17.5371	[17.5201, 17.5541]	13.6143

The rest of the parameters are $S = 100.0$, $r = 0.09$, $T = 1$, $\lambda = 0.05$ and $\mu_\pi = 0.0$. The other variables: σ, K and σ_π are varied in the table

Table 2 Comparison of ordinary Monte Carlo method (MC) with control variate Monte Carlo method (CMC) for longer maturities

T	σ_π	K	MC	CI	CMC	CI	$\hat{\sigma}_{\text{MC}}/\hat{\sigma}_{\text{CMC}}$
3	0.01	90	15.5453	[15.4877, 15.6029]	15.5070	[15.5033, 15.5106]	15.7303
		100	7.9934	[7.9440, 8.0427]	7.9812	[7.9775 7.9849]	13.2396
		110	2.9882	[2.9551, 3.0213]	2.9870	[2.9832, 2.9908]	8.6956
	0.1	90	15.5727	[15.5154, 15.6301]	15.4376	[15.4331, 15.4421]	12.7462
		100	8.0797	[8.0302, 8.1291]	7.9262	[7.9216, 7.9308]	10.8107
		110	3.0741	[3.0411, 3.1071]	2.9566	[2.9520, 2.9613]	7.1114
5	0.01	90	18.5849	[18.5129, 18.6568]	18.5239	[18.5188, 18.5290]	14.1504
		100	11.5713	[11.5061, 11.6366]	11.5568	[11.5517, 11.5619]	12.7443
		110	6.1737	[6.1212, 6.2262]	6.1536	[6.1484, 6.1589]	10.0625
	0.1	90	18.7349	[18.6619, 18.8078]	18.4099	[18.4035, 18.4162]	11.4871
		100	11.8572	[11.7908, 11.9235]	11.4615	[11.4551, 11.4679]	10.3574
		110	6.4608	[6.4050, 6.5166]	6.0797	[6.0730, 6.0864]	8.2976

The rest of the parameters are $S = 100.0$, $r = 0.05$ and $\sigma = 0.1$; the jump sensitivities are $\lambda = 0.1$ and $\mu_\pi = 0.0$

of control variate Monte Carlo method (CMC) $\hat{\sigma}_{\text{CMC}}$. The ratio is at least 5 and at most 15. Table 2 shows that the control variate is also effective even for longer maturities.

6 Conclusions

We have developed a control variate to price Asian options using Monte Carlo method. We considered the price of the underlying asset to be driven by a jump diffusion model resulting in a process with exponential Brownian motion a compound Poisson process. The results show that our method is an efficient speed-up for the Monte Carlo method.

References

1. Albrecher, H.: The valuation of asian options for market models of exponential lévy type. In: Vanmaele, M., et al. (eds.) Proceedings of the 2nd Actuarial and Financial Mathematics Day, Brussels, pp. 11–20 (2004)
2. Alziary, B., Déscamps, J.P., Koehl, P.F.: A P.D.E approach to Asian options: analytical and numerical evidence. J. Bank. Financ. **21**, 613–640 (1997)
3. Baxter, M., Rennie, A.: Financial Calculus An Introduction to Derivative Pricing, 1st edn. Cambridge University Press, Cambridge (1996)
4. Cai, N., Kou, S.: Pricing Asian options under a hyper-exponential jump diffusion model. Oper. Res. **60**(1), 64–77 (2012)

5. Cyganowski, S., Kloeden, P.: MAPLE schemes for jump-diffusion stochastic differential equations. In: Proc. 16th IMACS World Congress, Lausanne, pp. 216–219 (2000)
6. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Stochastic Modelling and Applied Probability, 1st edn. Springer, Berlin (2003)
7. Hanson, F.B.: Applied Stochastic Processes and Control for Jump-diffusion: Modeling, Analysis and Computation. SIAM, Philadelphia (2007)
8. Kemna, A.G.Z., Vorst, A.C.F.: A pricing method for options based on average asset values. J. Bank. Financ. **14**(1), 113–129 (1990)
9. Schoutens, W., Symens, S.: The pricing of exotic options by Monte-Carlo simulations in a Lévy Market with stochastic volatility. Int. J. Theor. Appl. Finance **8**(6), 839–864 (2003)
10. Vecer, J., Xu, M.: Pricing asian options in a semimartingale model. Quant. Finance **4**, 170–175 (2004)
11. Zhou, C.: The term structure of credit spreads with jump risk. J. Bank. Financ. **25**, 2015–2040 (2001)

A Positive, Stable and Consistent Front-Fixing Numerical Scheme for American Options

R. Company, V.N. Egorova, and L. Jódar

Abstract In this paper we propose an explicit finite-difference scheme to solve the American option pricing problem. It is based on front-fixing transformation that involves unknown free boundary to the equation. The proposed stable and consistent numerical scheme preserves positivity and monotonicity of the solution in accordance with the behavior of the exact solution. Numerical examples and comparison with other methods are included. This technique can be applied to some types of two-asset options after reducing the dimension. In the paper the front-fixing method is applied to exchange option pricing.

Keywords Computational finance • Option pricing

1 Front-Fixing Method

American call option price model is given by Wilmott [9] as the moving free boundary PDE

$$\frac{\partial C}{\partial \tau} = \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + (r-q)S\frac{\partial C}{\partial S} - rC, \qquad 0 \le S < B(\tau), \qquad 0 < \tau \le T, \tag{1}$$

together with the boundary and initial conditions

$$\frac{\partial C}{\partial S}(B(\tau),\tau) = 1, \quad C(B(\tau),\tau) = B(\tau) - E, \quad C(0,\tau) = 0, \tag{2}$$

$$C(S,0) = \max(S-E,0), \quad B(0) = \begin{cases} E, & r \le q, \\ \frac{r}{q}E, & r > q. \end{cases} \tag{3}$$

R. Company • V.N. Egorova (✉) • L. Jódar
Universitat Politécnica de Valéncia, Camino de Vera s/n, 46022 Valencia, Spain
e-mail: rcompany@imm.upv.es; egorova.vn@gmail.com; ljodar@imm.upv.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_10

where $\tau = T - t$ denotes the time to maturity T, S is the asset's price, $C(S, \tau)$ is the option price, $B(\tau)$ is the unknown early exercise boundary, σ is the volatility of the asset, r is the risk free interest rate, q is the continuous dividend yield and E is the strike price.

Note that if there is no any dividends ($q = 0$), then the optimal strategy is to exercise option at the maturity (see [3]). In that case the American call becomes European one. Because of that we consider problem (1)–(3) with $q > 0$.

Let us consider the dimensionless transformation

$$c(x, \tau) = \frac{C(S, \tau)}{E}, \quad S_f(\tau) = \frac{B(\tau)}{E}, \qquad x = \ln \frac{B(\tau)}{S}. \tag{4}$$

Under transformation (4) the problem (1)–(3) can be rewritten in normalized form

$$\frac{\partial c}{\partial \tau} = \frac{1}{2}\sigma^2 \frac{\partial^2 c}{\partial x^2} - \left(r - q - \frac{\sigma^2}{2} + \frac{S_f'}{S_f} \right) \frac{\partial c}{\partial x} - rc, \qquad x \geq 0, \qquad 0 < \tau \leq T, \tag{5}$$

with new boundary and initial conditions

$$c(x, 0) = \begin{cases} 0, & r \leq q, \\ g(x), & r > q, \end{cases} \qquad x \geq 0; \quad g(x) = \max\left(\frac{r}{q} e^{-x} - 1, 0 \right), \tag{6}$$

$$\frac{\partial c}{\partial x}(0, \tau) = -S_f(\tau), \tag{7}$$

$$c(0, \tau) = S_f(\tau) - 1, \tag{8}$$

$$\lim_{x \to \infty} c(x, \tau) = 0, \tag{9}$$

$$S_f(0) = \begin{cases} 1, & r \leq q, \\ \frac{r}{q}, & r > q. \end{cases} \tag{10}$$

Following the ideas of [10] and in order to solve the numerical difficulties derived from the discretization at the numerical boundary, we assume that (5) holds true at $x = 0$,

$$\frac{\sigma^2}{2} \frac{\partial^2 c}{\partial x^2} - \left(q + \frac{\sigma^2}{2} \right) S_f + r = 0. \tag{11}$$

Equation (5) is a non-linear differential equation on the domain $[0, \infty) \times (0, T]$. In order to solve numerically problem (5)–(10), one has to consider a bounded numerical domain. Let us introduce $x_{\max}$ large enough to translate the boundary condition (9). Then the problem (5)–(10) can be studied on the fixed domain $[0, x_{\max}] \times (0, T]$. The value x_{max} is chosen following the criterion pointed out in [4].

1.1 Finite-Difference Scheme

Let us introduce the computational grid of $M+2$ space points and N time levels with respective stepsizes h and k

$$h = \frac{x_{max}}{M+1}, \quad k = \frac{T}{N}, \tag{12}$$

$$x_j = hj, \quad j = 0, .., M+1, \quad \tau^n = kn, \quad n = 0, .., N. \tag{13}$$

The approximate value of $c(x,\tau)$ at the point x_j and time τ^n is denoted by $c_j^n \approx c(x_j, \tau^n)$ and the approximate value of the free boundary is denoted by $S_f^n \approx S_f(\tau^n)$. Then a forward two-time level and centred in a space explicit scheme is constructed for internal spacial nodes as follows

$$\frac{c_j^{n+1} - c_j^n}{k} = \frac{1}{2}\sigma^2 \frac{c_{j-1}^n - 2c_j^n + c_{j+1}^n}{h^2} -$$

$$\left(r - q - \frac{\sigma^2}{2} + \frac{S_f^{n+1} - S_f^n}{kS_f^n}\right) \frac{c_{j+1}^n - c_{j-1}^n}{2h} - rc_j^n. \tag{14}$$

Equation (14) can be rewritten on form

$$c_j^{n+1} = ac_{j-1}^n + bc_j^n + fc_{j+1}^n + \frac{S_f^{n+1} - S_f^n}{2hS_f^n}\left(c_{j+1}^n - c_{j-1}^n\right), \quad 1 \le j \le M. \tag{15}$$

The second order discretization of the boundary conditions (7), (8) and (11) is as follows

$$c_0^n = S_f^n - 1, \quad \frac{c_1^n - c_{-1}^n}{2h} = -S_f^n, \tag{16}$$

$$\frac{\sigma^2}{2}\frac{c_{-1}^n - 2c_0^n + c_1^n}{h^2} - \left(q + \frac{\sigma^2}{2}\right) S_f^n + r = 0, \tag{17}$$

where c_{-1}^n means the value of the solution at the fictitious point $x = -h$, that should be eliminated later.

The connection of the free boundary S_f^n and option value c_1^n on the same time level n is presented in form

$$c_1^n = \alpha - \beta S_f^n = -1 - \frac{rh^2}{\sigma^2} - \left(-1 + h - \left(\frac{q}{\sigma^2} + \frac{1}{2}\right)h^2\right) S_f^n, \quad n \ge 1. \tag{18}$$

We use together (15) for $j = 1$ and (18) to obtain the nonlinear law of the free boundary motion

$$S_f^{n+1} = d^n S_f^n = \frac{ac_0^n + bc_1^n + fc_2^n + \frac{c_2^n - c_0^n}{2h} - \alpha}{\frac{c_2^n - c_0^n}{2h} - \beta S_f^n} S_f^n. \tag{19}$$

1.2 Numerical Analysis

In this section we will show qualitative scheme properties such as the free boundary non-decreasing monotonicity as well as the positivity and non-increasing spacial monotonicity of the numerical option price under transformation. These properties preserve the behaviour of the theoretical solution of the American option price problem as it is shown in [5].

Note, that using expressions (15) it is easy to obtain, that the constants of the scheme a, b and f are positive for both cases: $r \leq q$ and $r > q$ under following conditions

$$h < \frac{\sigma^2}{\left| r - q - \frac{\sigma^2}{2} \right|}, \quad r \neq q + \frac{\sigma^2}{2}, \quad k < \frac{h^2}{\sigma^2 + rh^2}, \tag{20}$$

If $r = q + \frac{\sigma^2}{2}$, then under the condition (20), coefficients a, b and f are positive. Then the following result can be established:

Theorem 1 *Let $\{c_j^n, S_f^n\}$ be the numerical solution of scheme (15) for a transformed American call option problem (5) and let d^n be defined by (19). Then the numerical scheme (15) guarantees the following properties of the numerical solution:*

1. *Increasing monotonicity and positivity of values S_f^n, $n = 0, \ldots, N$;*
2. *Non-negativity and non-increasing monotonicity of the vectors $c^n = (c_0^n, \ldots c_{M+1}^n)$ with respect to space indexes for each fixed $n = 0, \ldots, N$;*
3. *The scheme (15) is stable with respect to the norm $||\cdot||_\infty$;*
4. *The numerical solution computed by the scheme is consistent of order two in space and order one in time with Eq. (5) and boundary conditions (2), (11) .*

Remark In the case $r > q$, when $6q$ is close to σ^2, we couldn't guarantee that $d^0 > 1$. It means that monotonicity of the free boundary is not preserved for the first time step.

2 Numerical Experiments

Free boundary obtained of both examples below is compared with analytical approximation closed to maturity presented in [2].

Example 1 We consider the problem with the parameters [7],

$$r = 0.03, \quad q = 0.07, \quad \sigma = 0.2, \quad T = 0.5. \tag{21}$$

The proposed method is compared with other approaches presented in [7]. As we can see from Table 1 difference between proposed method (FF) and "True" value becomes smaller with increasing asset price S. All methods: Gauss-Leguere (GL), Lower and upper bound approximation (LUBA), Han and Wu method (HW), operator splitting method (OS) are compared in [7]. The results are presented in Table 1. The root-mean-square error (RMSE) is used to measure the accuracy of the scheme.

Example 2 We can compare the front-fixing method with transformation presented by Ševčovič [8]. There is a problem with the following parameters

$$r = 0.1, \quad q = 0.05, \quad \sigma = 0.2, \quad T = 1, \quad E = 10. \tag{22}$$

The position of the free boundary at $\tau = T$ is $B(T) = 22.3754$ (in [8]) and it was computed by the proposed method as $S_f(T) = 2.2375$, since the transformed problem is dimensionless. In Table 2 we present a comparison of

Table 1 Comparison of option price calculated by proposed method (FF) for Example 1 with other methods

Asset price	True value	GL	LUBA	HW	OS	FF
80	0.2194	0.2185	0.2195	0.2193	0.2193	0.2196
90	1.3864	1.3851	1.3862	1.3858	1.3858	1.3868
100	4.7825	4.7835	4.7821	4.7816	4.7817	4.7827
110	11.0978	11.1120	11.0976	11.0969	11.0971	11.0981
120	20.0004	20.0000	20.0000	20.0005	20.0000	20.0006
	RMSE	6.4078-3	2.8636-4	6.3246-4	5.7619-4	2.5391-4

Table 2 Comparison of the proposed method with other methods for parameters (22)

Method/the asset value S	15	18	20	21	22.375
Ševčovič's method	5.15	8.09	10.03	11.01	12.37
Trinomial tree	5.15	8.09	10.03	11.01	12.37
Finite differences	5.49	8.48	10.48	11.48	12.48
Analytical approximation	5.23	8.10	10.04	11.02	12.38
Proposed method	5.21	8.09	10.03	11.01	12.37

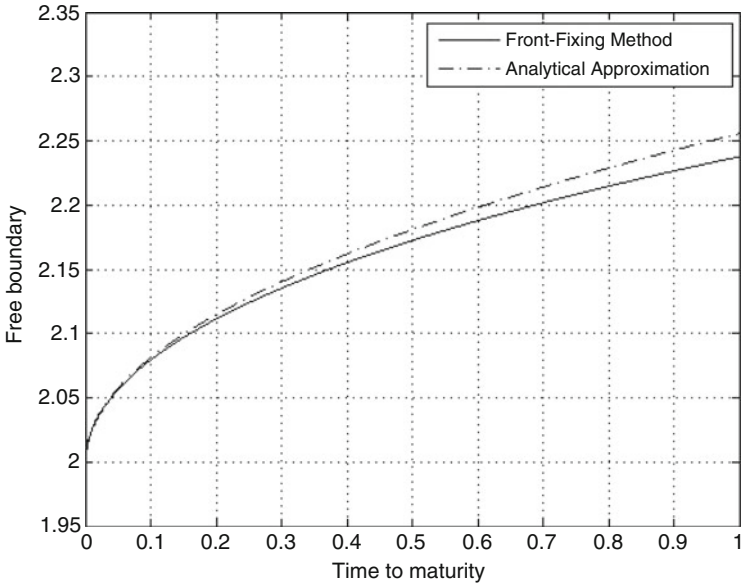


Fig. 1 Comparison free boundary from Example 2 with analytical approximation on the whole domain

results obtained by the proposed front-fixing method and by other methods such as semi-explicit formula, presented in [8], trinomial tree, finite difference approximation and analytical approximation of Barone-Adesi and Whaley [1]. The free boundary motion for that problem is compared with analytical approximation [2]. Results are presented on Fig. 1.

Example 3 Some types of two asset American Option problems can be transformed into a one-spatial dimensional equation by a suitable change of variables. It is important to realize that a reduction in the number of dimensions can contribute greatly to efficiency of the finite difference implementation. After appropriate transformation we obtain the Black-Scholes equation with the different parameters.

As a numerical example, we consider the American exchange option (see [6]) for correlated assets with $\rho = 0.5$, and $\sigma_1 = \sigma_2 = 0.5$, then $\sigma^2 = 0.25$. The results are presented on Fig. 2.

3 Conclusion

We proposed a front-fixing method for American call option with dividends pricing problem. An explicit finite difference scheme is constructed for numerical solution. It has several proved advantages such as conditional stability and consistency with

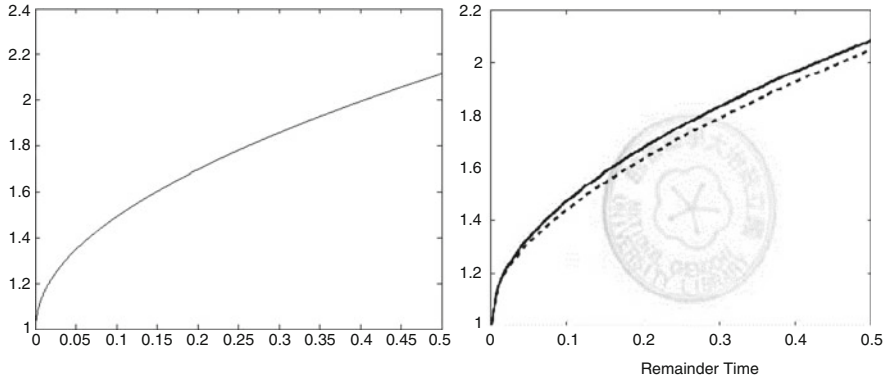


Fig. 2 Optimal exercise ratio in time: calculated by proposed method (*left*) and presented in [6]

the differential equation. Moreover, the explicit finite difference scheme guarantees the positivity and monotonicity of the solution. By numerical experiments the stability and accuracy are proved.

Moreover, several types of multi-asset option after appropriate transformation can be presented as an American call option with dividends. Therefore, the proposed method can be used for such options. Results of applying this technique are compared with known results.

Acknowledgements This paper has been partially supported by the European Union in the FP7-PEOPLE-2012-ITN program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN STRIKE-Novel Methods in Computational Finance).

References

1. Barone-Adesi, G., Whaley, R.: Efficient analytic approximation of American option values. J. Financ. **42**, 301–320 (1987)
2. Evans, J., Kuske, R., Keller, J.: American options on assets with dividends near expiry. J. Math. Financ. **12**(3), 219–237 (2002)
3. Hull, J., White, A.: Valuing derivative securities using the explicit finite difference method. J. Financ. Quant. Anal. **25**(1), 87–100 (1990)
4. Kangro, R., Nicolaides, R.: Far field boundary conditions for Black-Scholes equations. SIAM J. Numer. Anal. **38**(4), 1357–1368 (2000)
5. Kim, I.J.: The analytic valuation of American options. Rev. Financ. Stud. **3**, 547–572 (1990)
6. Liu, H.K.: The valuation of American options on single asset and multiple assets. Ph.D. thesis, National Chengchi University, Taiwan (2007)
7. Saib, A., Tangman, Y., Thakoor, N., Bhuruth, M.: On some finite difference algorithms for pricing American options and their implementation in mathematica. In: Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering (2011)

8. Ševčovič, D.: Transformation methods for evaluating approximations to the optimal exercise boundary for linear and nonlinear Black-Scholes equations. In: Nonlinear Models in Mathematical Finance: New Research Trends in Option Pricing, pp. 173–218. Nova Science Publishers Inc., New York (2008)
9. Wilmott, P., Howison, S., Dewynne, J.: The Mathematics of Financial Derivatives. Cambridge University Press, Cambridge (1995)
10. Wu, L., Kwok, Y.K.: A front-fixing method for the valuation of American option. J. Financ. Eng. **6**, 83–97 (1997)

Efficient Calibration and Pricing in LIBOR Market Models with SABR Stochastic Volatility Using GPUs

A.M. Ferreiro, J.A. García, J.G. López-Salas, and C. Vázquez

Abstract In order to overcome the drawbacks of assuming deterministic volatility coefficients in the standard LIBOR market models, several extensions of LIBOR models to incorporate stochastic volatilities have been proposed. The efficient calibration to market data of these more complex models becomes a relevant target in practice. The main objective of the present work is to efficiently calibrate some recent SABR/LIBOR market models to real market prices of caplets and swaptions. For the calibration we propose a parallelized version of the simulated annealing algorithm for multi-GPUs. The numerical results clearly illustrate the advantages of using the proposed multi-GPUs tools when applied to real market data and popular SABR/LIBOR models.

Keywords Computational finance • Market model calibration

1 SABR/LIBOR Market Models

This work is mainly concerned with three extensions of the LIBOR market model (LMM) that incorporate the volatility smile by means of the SABR stochastic volatility model. The SABR model has become the market standard for interpolating and extrapolating prices of plain vanilla caplets and swaptions [6]. It is widely used because it involves a closed-form formula for the implied volatility which allows an easy calibration of the model. In the more standard LIBOR market model [1] the dynamics of each LIBOR forward rate under the corresponding terminal measure are assumed to be martingales with constant volatility. When adding the SABR stochastic volatility model, the forward rates and volatility processes satisfy the

A.M. Ferreiro • J.A. García • J.G. López-Salas (✉) • C. Vázquez
Department of Mathematics, Faculty of Informatics, University of Coruna, Campus Elviña s/n, 15071 A Coruña, Spain
e-mail: aferreiro@udc.es; jagrodriguez@udc.es; jose.lsalas@udc.es; carlosv@udc.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_11

following coupled dynamics

$$dF_i(t) = V_i(t)F_i(t)^{\beta_i}dW_i(t),$$
$$dV_i(t) = \sigma_i V_i(t)dZ_i(t).$$

We note that if the interest rate derivative only depends on one particular forward rate then it is convenient to use the corresponding terminal measure. However when derivatives depend on several forward rates, a common measure needs to be used. Thus, in the case of pricing complex derivatives a change of measure produces the appearance of drift terms in both dynamics. The main drawback of classical LMM comes from considering constant volatilities. SABR/LIBOR market models combine the advantages of these two models. In this paper we consider the different SABR/LIBOR models proposed by Hagan [5], Mercurio and Morini [8] and Rebonato [10]. Hereafter, for sake of brevity we only present the Rebonato model. Interested readers on the other two models are referred to [4].

For each $i = 1, \ldots, M$ let F_i and V_i be the ith forward rate that matures at time T_i and its corresponding stochastic volatility, respectively. Then, under a common measure their dynamics are given by (see [10])

$$dF_i(t) = \mu^{F_i}(t)dt + V_i(t)F_i(t)^{\beta_i}dW_i(t), \tag{1}$$
$$V_i(t) = \kappa_i(t)g_i(t), \tag{2}$$
$$d\kappa_i(t) = \mu^{\kappa_i}(t)dt + \kappa_i(t)h_i(t)dZ_i(t), \tag{3}$$

where

$$g_i(t) = \big(a+b(T_i-t)\big)\exp\big(-c(T_i-t)\big)+d, \quad h_i(t) = \big(\alpha+\beta(T_i-t)\big)\exp\big(-\gamma(T_i-t)\big)+\delta,$$

with the associated correlations denoted by

$$\mathbb{E}[dW_i(t)\cdot dW_j(t)] = \rho_{i,j}dt, \quad \mathbb{E}[dW_i(t)\cdot dZ_j(t)] = \phi_{i,j}dt, \quad \mathbb{E}[dZ_i(t)\cdot dZ_j(t)] = \theta_{i,j}dt,$$

and the initial given values $\kappa_i = \kappa_i(0)$ and $F_i(0)$. Thus, the correlation structure is given by the block-matrix

$$\mathbf{P} = \begin{bmatrix} \boldsymbol{\rho} & \boldsymbol{\phi} \\ \boldsymbol{\phi}^\top & \boldsymbol{\theta} \end{bmatrix},$$

where the submatrix $\boldsymbol{\rho} = (\rho_{i,j})$ represents the correlations between the forward rates F_i and F_j, the submatrix $\boldsymbol{\phi} = (\phi_{i,j})$ includes the correlations between the forward rates F_i and the instantaneous volatilities V_j, and the submatrix $\boldsymbol{\theta} = (\theta_{i,j})$ contains the correlations between the instantaneous volatilities V_i and V_j.

More precisely, if we introduce the bank-account numeraire $\beta(t)$, defined by

$$\beta(t) = \prod_{j=0}^{i-1} \big(1 + \Delta t F_j(T_j)\big) \quad \text{if } t \in [T_i, T_{i+1}],$$

then, under the associated spot probability measure, the drift terms of the processes defined in (1) and (3) are

$$\mu^{F_i}(t) = V_i(t)F_i(t)^{\beta_i} \sum_{j=h(t)}^{i} \frac{\tau_j \rho_{i,j} V_j(t) F_j(t)^{\beta_j}}{1 + \tau_j F_j(t)},$$

$$\mu^{\kappa_i}(t) = \kappa_i(t)h_i(t) \sum_{j=h(t)}^{i} \frac{\tau_j \phi_{i,j} V_j(t) F_j(t)^{\beta_j}}{1 + \tau_j F_j(t)},$$

where $h(t)$ denotes the index of the first unfixed F_i, i.e.,

$$h(t) = j, \text{ if } t \in [T_{j-1}, T_j). \tag{4}$$

The implied volatility for this model can be computed from Hagan second order approximation formula [9]:

$$\begin{aligned}\sigma\big(K, F_i(0)\big) \approx \frac{\alpha_i}{F_i(0)^{(1-\beta_i)}} \times \Bigg\{ & 1 - \frac{1}{2}(1 - \beta_i - \phi_{i,i}\sigma_i\omega_i) \cdot \ln\Big(\frac{K}{F_i(0)}\Big) \\ & + \frac{1}{12}\Big((1-\beta_i)^2 + (2 - 3\phi_{i,i}^2)\sigma_i^2\omega_i^2 + 3\big((1-\beta_i) - \phi_{i,i}\sigma_i\omega_i\big)\Big) \\ & \cdot \left[\ln\Big(\frac{K}{F_i(0)}\Big)\right]^2 \Bigg\},\end{aligned} \tag{5}$$

where $\omega_i = \alpha_i^{-1} F_i(0)^{(1-\beta_i)}$, by using the following parameters denoted with *SABR* superindexes,

$$\begin{aligned}& \beta_i^{SABR} = \beta_i, \quad \phi_{i,i}^{SABR} = \phi_{i,i}, \quad \alpha_i^{SABR} = \kappa_i(0)\left(\frac{1}{T_i}\int_0^{T_i} g_i(t)^2 dt\right)^{\frac{1}{2}}, \\ & \sigma_i^{SABR} = \frac{\kappa_i(0)}{\alpha_i^{SABR} T_i}\left(2\int_0^{T_i} g_i(t)^2 \hat{h}_i(t)^2 t dt\right)^{\frac{1}{2}}, \text{ where } \hat{h}_i(t) = \sqrt{\frac{1}{t}\int_0^t (h_i(s))^2 ds}.\end{aligned} \tag{6}$$

For the correlations, we consider the following function parameterizations:

$$\rho_{i,j} = \eta_1 + (1-\eta_1)\exp[-\lambda_1|T_i - T_j|], \tag{7}$$

$$\theta_{i,j} = \eta_2 + (1-\eta_2)\exp[-\lambda_2|T_i - T_j|], \tag{8}$$

$$\phi_{i,j} = \text{sign}(\phi_{i,i})\sqrt{|\phi_{i,i}\phi_{j,j}|}\exp\left[-\lambda_3(T_i - T_j)^+ - \lambda_3(T_j - T_i)^+\right], \tag{9}$$

where the terms $\phi_{i,i}$ are previously calibrated using Hagan formula (5) for the whole volatilities surface.

In this work we propose an efficient calibration strategy to some market prices for the parameters appearing in the three previous models. More precisely, we consider the market prices of caplets and swaptions and we pose the corresponding global optimization problems to calibrate the model parameters. In order to speed up the optimization algorithm we use an implementation in GPUs.

2 Model Calibration

Model parameters are calibrated in two stages, firstly to caplets and secondly to swaptions. We note that model parameters can be classified into two categories (volatility and correlation parameters). The volatility parameters are $\boldsymbol{x} = (\phi_{ii}, \kappa_i,$ parameters of the volatility functions g and $h)$ and the correlation ones $\boldsymbol{y} = (\eta_1, \lambda_1, \eta_2, \lambda_2, \lambda_3)$. According to this classification, the cost functions to be minimized in the calibration process are the following:

- Function to calibrate the market prices of caplets:

$$f_c(\boldsymbol{x}) = \sum_{i=1}^{M}\sum_{j=1}^{numK}\left(\sigma\left(K_j, F_i(0)\right) - \sigma_{market}\left(K_j, F_i(0)\right)\right)^2(\boldsymbol{x}),$$

  where σ is given by Hagan formula (5) with the parameters (6), σ_{market} are the market smiles and $\boldsymbol{x}$ is the vector containing the volatility parameters of the model. Moreover, M and $numK$ denote the number of maturities and strikes of the caplets, respectively.
- Function to calibrate the market prices of swaptions:

$$f_s(\boldsymbol{y}) = \sum_{i=1}^{numSws}\left(S_{\text{Black}}(swaption_i) - S_{MC}(swaption_i)\right)^2(\boldsymbol{y}),$$

where $swaption_i$ denotes the ith swaption, S_{Black} is the Black formula for swaptions and $S_{MC}(swaption_i)$ is the value of the ith swaption computed using Monte Carlo method. Moreover $\boldsymbol{y}$ denotes the vector containing the correlation parameters of the model and $numSws$ is the number of swaptions.

In this work, the calibration of the parameters has been performed with a Simulated Annealing (SA) global optimization algorithm [7]. The algorithm consists in an external decreasing temperature loop. At each fixed temperature a Metropolis process, that can be seen as a Markov chain, is performed to compute the equilibrium state at this temperature level. This Markov chain consists of randomly choosing points in the search domain: if the value of the cost function at a new point decreases, the point is accepted; otherwise the point is randomly accepted following the Boltzman criterion, where the probability of accepting points with higher cost function value decreases with temperature. This process is repeated at each temperature level until temperature is low enough. As it is well known in the literature, SA involves a great computational cost.

In [3], the parallelization of the SA algorithm has been performed with GPUs. The idea is that at each temperature level the Markov chains are distributed among the GPU threads. Among all the final reached points of the threads, the one with the minimum cost function value is selected, thus performing a reduction operation. The selected point is the starting one for all the threads in the next temperature level. The process is repeated until reaching a certain value of temperature.

The previous implementation can also be improved using multi-GPUs. In this case, the Markov chains are distributed among GPUs (for example, if we have two GPUs, half of the chains are computed by each GPU, see Fig. 1) and at each GPU the chains are distributed among the threads of this particular GPU. Before advancing to the next temperature level the best point must be computed in each GPU and then the best point of all GPUs is computed and used as starting point for all the upcoming threads of the new temperature level (see Fig. 1). This multi-GPU algorithm was presented in [2], where it was used to calibrate some SABR models to a volatility surface.

In order to calibrate models with many parameters, as the Rebonato one, the multi-GPU version becomes more suitable, since the minimization process is very costly.

In the SABR/LIBOR market models, for the calibration to swaption market prices there is not an explicit formula to price swaptions. Therefore, we use a Monte Carlo simulation technique to price swaptions, thus leading to two nested Monte Carlo loops: one for the SA and the other one for the swaption pricer. So, as the Monte Carlo swaption pricer is carried out inside the GPU, the SA minimization

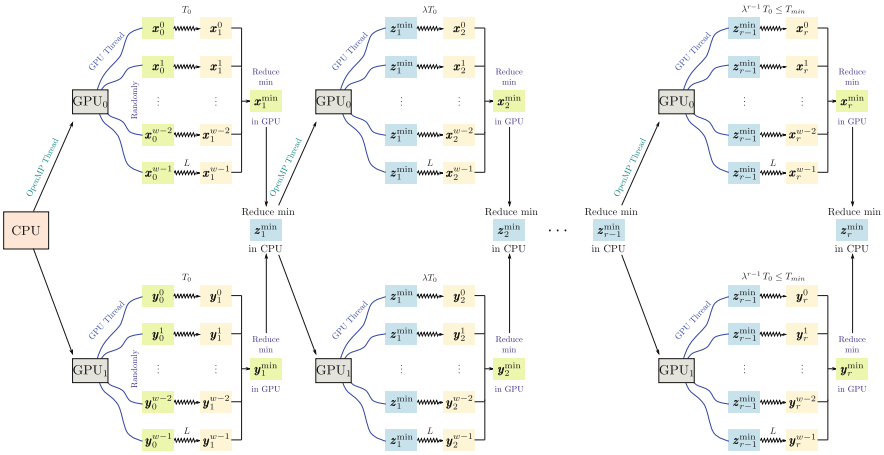


Fig. 1 Sketch of the parallel SA algorithm using two GPUs and OpenMP

algorithm is run on CPU. In order to use all available GPUs in the system, we propose a CPU SA parallelization using OpenMP. So, each OpenMP SA thread uses a GPU to evaluate the Monte Carlo objective function.

3 Numerical Results

Market data correspond to the 6 month EURIBOR rate. In this section, for sake of brevity, we only present the results of the calibration of the model to the smiles of the swap rates shown in Table 1. The results of the previous calibration to the smiles of the forward rates presented in Table 2 are detailed in the article [4].

The calibrated correlation parameters are $\eta_1 = 0.650997$, $\lambda_1 = 3.617546$, $\eta_2 = 0.999000$, $\lambda_2 = 0.380984$ and $\lambda_3 = 0.001000$. Using two GPUs the execution time was approximately 2 h (by using a cluster of GPUs time could be substantially reduced). In Table 3, some market vs. model swaption prices are shown. The mean absolute error considering all market swaptions is 6.30×10^{-2}. Figure 2 shows the model fitting to the first four swaption market prices.

Table 1 Smiles of swap rates

		−80 %	−60 %	−40 %	−20 %	0 %	20 %	40 %	60 %	80 %
1 year	21/05/2012	122.30 %	102.40 %	87.12 %	76.45 %	70.40 %	66.47 %	64.20 %	63.03 %	62.56 %
	21/11/2012	102.86 %	89.97 %	79.85 %	72.49 %	67.90 %	64.58 %	62.16 %	60.39 %	59.19 %
	21/05/2013	95.64 %	83.17 %	73.42 %	66.40 %	62.10 %	59.03 %	56.84 %	55.26 %	54.18 %
	21/11/2013	88.11 %	76.06 %	66.69 %	60.00 %	56.00 %	53.18 %	51.22 %	49.84 %	48.87 %
2 years	21/05/2012	111.50 %	91.60 %	76.32 %	65.65 %	59.60 %	55.67 %	53.40 %	52.23 %	51.76 %
	21/11/2012	89.66 %	76.77 %	66.65 %	59.29 %	54.70 %	51.38 %	48.96 %	47.19 %	45.99 %
	21/05/2013	82.94 %	70.47 %	60.72 %	53.70 %	49.40 %	46.33 %	44.14 %	42.56 %	41.48 %
	21/11/2013	77.81 %	65.76 %	56.39 %	49.70 %	45.70 %	42.88 %	40.92 %	39.54 %	38.57 %
3 years	21/05/2012	106.40 %	86.50 %	71.22 %	60.55 %	54.50 %	50.57 %	48.30 %	47.13 %	46.66 %
	21/11/2012	83.66 %	70.77 %	60.65 %	53.29 %	48.70 %	45.38 %	42.96 %	41.19 %	39.99 %
	21/05/2013	78.34 %	65.87 %	56.12 %	49.10 %	44.80 %	41.73 %	39.54 %	37.96 %	36.88 %
	21/11/2013	73.61 %	61.56 %	52.19 %	45.50 %	41.50 %	38.68 %	36.72 %	35.34 %	34.37 %
4 years	21/05/2012	101.90 %	82.00 %	66.72 %	56.05 %	50.00 %	46.07 %	43.80 %	42.63 %	42.16 %
	21/11/2012	80.26 %	67.37 %	57.25 %	49.89 %	45.30 %	41.98 %	39.56 %	37.79 %	36.59 %
	21/05/2013	75.24 %	62.77 %	53.02 %	46.00 %	41.70 %	38.63 %	36.44 %	34.86 %	33.78 %
	21/11/2013	70.91 %	58.86 %	49.49 %	42.80 %	38.80 %	35.98 %	34.02 %	32.64 %	31.67 %
5 years	21/05/2012	96.15 %	74.25 %	58.83 %	49.88 %	47.40 %	45.74 %	44.61 %	43.76 %	43.05 %
	21/11/2012	89.58 %	68.82 %	54.14 %	45.54 %	43.00 %	39.36 %	37.33 %	36.15 %	35.37 %
	21/05/2013	83.91 %	64.51 %	50.71 %	42.51 %	39.90 %	36.48 %	34.59 %	33.50 %	32.76 %
	21/11/2013	79.13 %	61.09 %	48.17 %	40.37 %	37.70 %	34.50 %	32.74 %	31.75 %	31.05 %

Maturities (first column) and moneyness (first row)

Table 2 Smiles of forward rates

	−80 %	−60 %	−40 %	−20 %	0 %	20 %	40 %	60 %	80 %
21-05-12	142.61 %	117.05 %	97.26 %	82.58 %	72.29 %	70.89 %	69.49 %	68.08 %	66.67 %
21-11-12	112.74 %	99.23 %	88.27 %	79.62 %	73.03 %	71.95 %	70.87 %	69.77 %	68.69 %
21-05-13	91.55 %	83.75 %	77.09 %	71.50 %	67.93 %	67.10 %	66.41 %	65.88 %	65.49 %
21-11-13	64.82 %	60.95 %	57.08 %	53.21 %	52.49 %	51.34 %	50.61 %	50.30 %	50.46 %
21-05-14	66.96 %	61.84 %	56.69 %	52.43 %	50.32 %	48.72 %	47.70 %	47.14 %	46.97 %
21-11-14	69.20 %	62.75 %	56.30 %	51.65 %	48.19 %	46.19 %	44.91 %	44.12 %	43.66 %
21-05-15	71.49 %	63.67 %	55.92 %	50.89 %	46.19 %	43.83 %	42.32 %	41.35 %	40.64 %
21-11-15	73.89 %	64.61 %	55.54 %	50.13 %	44.25 %	41.56 %	39.84 %	38.71 %	37.78 %
21-05-16	76.34 %	65.56 %	55.16 %	49.39 %	42.40 %	39.43 %	37.54 %	36.26 %	35.15 %
21-11-16	78.90 %	66.53 %	54.78 %	48.65 %	40.61 %	37.38 %	35.34 %	33.94 %	32.68 %
21-05-17	81.50 %	67.50 %	54.41 %	47.94 %	38.93 %	35.47 %	33.30 %	31.81 %	30.42 %
21-11-17	84.24 %	68.50 %	54.03 %	47.22 %	37.29 %	33.63 %	31.36 %	29.78 %	28.28 %
21-05-18	87.02 %	69.50 %	53.67 %	46.53 %	35.74 %	31.92 %	29.55 %	27.90 %	26.32 %

Fixing dates (first column) and moneyness (first row)

Table 3 Calibration to swaptions, S_{Black} vs. S_{MC}, prices in %

	0.5×1 swaptions			1×1 swaptions		
Moneyness	S_{Black}	S_{MC}	$\lvert S_{Black} - S_{MC}\rvert$	S_{Black}	S_{MC}	$\lvert S_{Black} - S_{MC}\rvert$
−40 %	0.4866	0.4870	4.00×10^{-4}	0.5917	0.5839	7.80×10^{-3}
−20 %	0.3562	0.3669	1.07×10^{-2}	0.4661	0.4693	3.20×10^{-3}
0 %	0.2356	0.2477	1.21×10^{-2}	0.3467	0.3546	7.90×10^{-3}
20 %	0.1363	0.1441	7.80×10^{-3}	0.2394	0.2488	9.40×10^{-3}
40 %	0.0680	0.0699	1.90×10^{-3}	0.1517	0.1606	8.90×10^{-3}
	1.5×1 swaptions			2×1 swaptions		
Moneyness	S_{Black}	S_{MC}	$\lvert S_{Black} - S_{MC}\rvert$	S_{Black}	S_{MC}	$\lvert S_{Black} - S_{MC}\rvert$
−40 %	0.7357	0.6902	4.55×10^{-2}	0.8184	0.7465	7.19×10^{-2}
−20 %	0.5908	0.5612	2.96×10^{-2}	0.6603	0.6028	5.75×10^{-2}
0 %	0.4536	0.4339	1.97×10^{-2}	0.5118	0.4620	4.98×10^{-2}
20 %	0.3277	0.3171	1.06×10^{-2}	0.3754	0.3354	4.00×10^{-2}
40 %	0.2213	0.2188	2.50×10^{-3}	0.2587	0.2308	2.79×10^{-2}

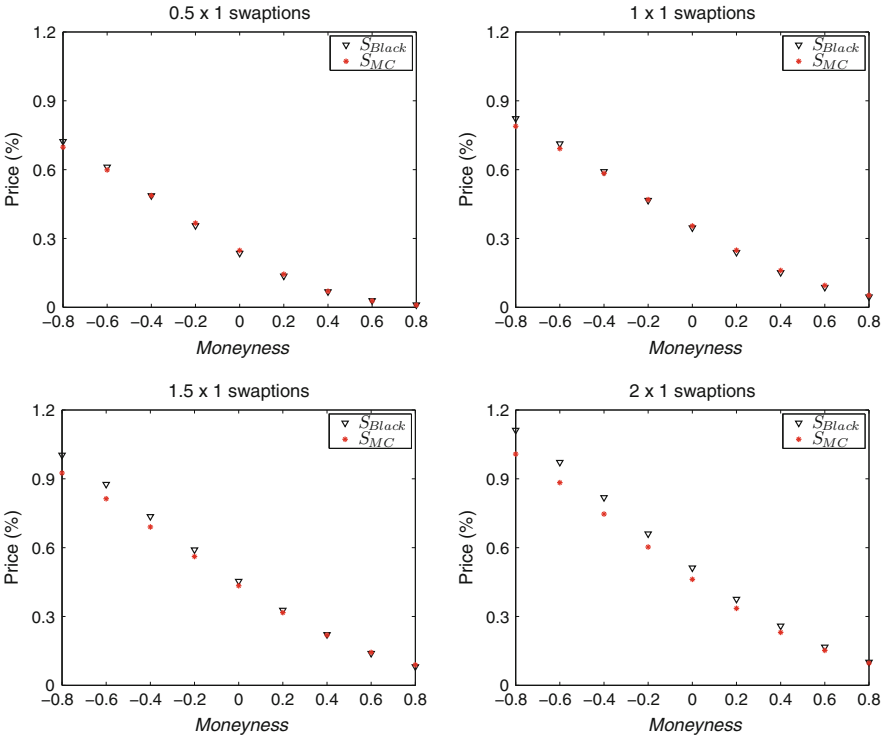


Fig. 2 S_{Black} vs. S_{MC}, $\{0.5, \ldots, 2\} \times 1$ swaptions

References

1. Brace, A., Gatarek, D., Musiela, M.: The Market model of interest rate dynamics. Math. Finance **7**(2), 127–155 (1997)
2. Fernández, J.L., Ferreiro, A.M., García, J.A., López-Salas, J.G., Vázquez, C.: Static and dynamic SABR stochastic volatility models: calibration and option pricing using GPUs. Math. Comput. Simul. **94**, 55–75 (2013)
3. Ferreiro, A.M., García, J.A., López-Salas, J.G., Vázquez, C.: An efficient implementation of parallel simulated annealing algorithm in GPUs. J. Glob. Optim. **57**(3), 863–890 (2013)
4. Ferreiro, A.M., García, J.A., López-Salas, J.G., Vázquez, C.: SABR/LIBOR market models: pricing and calibration for some interest rate derivatives. Appl. Math. Comput. (2014). http://dx.doi.org/10.1016/j.amc.2014.05.017
5. Hagan, P., Lesniewski, A.: LIBOR market model with SABR style stochastic volatility. Working paper (2008). Available at http://lesniewski.us/papers/working/SABRLMM.pdf
6. Hagan, P.S., Kumar, D., Lesniewski, A.S., Woodward, D.E.: Managing smile risk. In: The Best of Wilmott, vol. 1, pp. 249–296. Wiley, Hoboken (2002)
7. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**, 671–680 (1983)
8. Mercurio, F., Morini, M.: No-arbitrage dynamics for a tractable SABR term structure Libor model. In: Modeling Interest Rates: Advances in Derivatives Pricing. Risk Books, London (2009)
9. Oblój, J.: Fine-tune your smile: correction to Hagan et al. Wilmott Mag. (2007/2008). Preprint, arXiv:0708.0998
10. Rebonato, R.: A time-homogeneous SABR-consistent extension of the LMM. Risk **20**, 102–106 (2007)

Extension of a Fourier-Cosine Method to Solve BSDEs with Higher Dimensions

M. Pou, M.R. Ruijter, and C.W. Oosterlee

Abstract A Backward Stochastic Differential Equation (BSDE) is a stochastic differential equation for which a terminal condition has been specified. In Ruijter and Oosterlee (A Fourier-cosine method for an efficient computation of solutions to BSDEs, 2013) a Fourier-cosine method to solve BSDEs is developed. This technique is known as BCOS method and consists of the approximation of the BSDE's solution backwards in time by the use of the COS method developed in Fang and Oosterlee (SIAM J Sci Comput 31(2):826–848, 2008) to compute the conditional expectations that rise after the discretization by means of a θ-method for the time-integration.

In this work, the methodology is extended to the case in which there are more than one source of uncertainty or the terminal condition depends on more than one process, allowing the pricing of derivatives contracts such as rainbow options. The extension of the BCOS technique can be done taking into account some ideas developed in Ruijter and Oosterlee (SIAM J Sci Comput 34(5):B642–B671, 2012). We present some results concerning to derivatives on two processes without jumps. We also apply our extended method to solve the BSDEs that rise with the use of quadratic hedging techniques for pricing in incomplete markets without or with jumps (Lim, Math Oper Res 29(1):132–161, 2004; Lim, SIAM J Sci Comput 44(5):1893–1922, 2005). Problems in which the randomness of the terminal condition depends not only on the risky asset but also on the insurance risk or the counterparty default risk can be introduced in this framework (Delong, Backward Stochastic Differential Equations with Jumps and Their Actuarial and Financial Applications. Springer, London, 2013).

Keywords Backward stochastic differential equations • Computational finance • Option pricing

M. Pou • C.W. Oosterlee (✉)
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Delft University of Technology, Delft, The Netherlands
e-mail: m.pou@cwi.nl; c.w.oosterlee@cwi.nl

M.R. Ruijter
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
e-mail: m.j.ruijter@cwi.nl

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_12

1 Introduction

A Backward Stochastic Differential Equation (BSDE) is a Stochastic Differential Equation (SDE) for which a terminal condition has been specified. This terminal condition is given by a random variable, instead of by a fixed value. Specifically, we consider the case in which the terminal condition is a function of a stochastic process whose initial value and dynamic are known. By this way a Forward-Backward SDE (FBSDE) is considered. We allow the general framework of a forward stochastic process with jumps arising FBSDE with jumps (FBSDEJ). These kind of equations appear in the pricing and hedging of derivative contracts with payoff equals to a function of one or more risky assets whose prices have known dynamics.

In this paper we do not pay special attention to the theory of BSDEs that has been widely studied. See [1–4, 11, 12, 14]. In this work we extend the BCOS methodology presented in [16] to solve FBSDEs with higher dimensions, i.e., to solve FBSDEs in which more than one source of uncertainty is involved and/or in which the terminal condition depends on more than one stochastic process. This technique consists of the approximation of the solution backwards in time. After the necessary discretization of the equation, some conditional expectations have to be computed. The BCOS technique approximates these conditional expectations by the use of the close relation of the characteristic function of the forward process with the coefficients of the Fourier-cosine expansion of its density function. Therefore the simulation of the forward stochastic process is avoided. The extension of the technique to the case of higher dimensions takes some ideas from [15].

The paper is organized as follows. Section 2 presents the setting and notation on BSDEJs. In Sect. 3 several numerical examples in which these kind of equations appear are presented. Section 4 is devoted to the explanation of the BCOS technique. In Sect. 5 some of the obtained results are shown. Section 6 summarizes the conclusions.

2 Notation and Definitions on BSDEJs

In this section we introduce some definitions and theorems on BSDEJs and FBSDEJs. In order to that, according to [9, 16], we introduce the necessary notation.

We assume that all stochastic processes which we consider are defined on a finite time horizon $[0, T]$ and let $(\Omega, \mathbb{F}, \mathbb{P})$ be a complete probability space equipped with following independent stochastic processes:

- A standard d-dimensional Brownian motion $W_t = (W_t^1, \ldots, W_t^d)'$, with each component defined on $\Omega \times [0, T]$.
- A real-valued c-dimensional Poisson point process $q = (q^1, \ldots, q^c)'$, with each component defined on $\Omega \times [0, T] \times E^p$, where $E^p := \mathbb{R}\backslash\{0\}$. For $p = 1, \ldots, c$, we denote by $N^p(d\Gamma^p, dt)$ the Poisson random measure associated to q^p, whose compensator is assumed to be of the form $\nu^p(d\Gamma^p)dt$, where $\nu^p(d\Gamma^p)$ stands for

the Levy measure, which is positive and satisfies

$$\nu^p(\{0\}) = 0 \quad \text{and} \quad \int_{E^p} (1 \wedge |\Gamma^p|)^2 \nu^p(d\Gamma^p) < \infty. \tag{1}$$

$N^p(B,t)$, with $B \subset \mathbb{R}$, represents the number of jumps with size in set B which occur before or at time t, and $\nu^p(B)$ counts the expected number of jumps in a unit time interval. We denote with $\hat{N}^p$, $p = 1, \ldots, c$, the compensated Poisson random measure, that is given by:

$$\hat{N}^p(d\Gamma^p, dt) = N^p(d\Gamma^p, dt) - \nu^p(d\Gamma^p)dt. \tag{2}$$

$\mathscr{F}$ is the completed filtration generated by the processes $W = (W^1, \ldots, W^d)'$ and $N = (N^1, \ldots, N^c)'$. We assume independence between the different components in W and N.

Moreover, E^p, $p = 1, \ldots, c$, is assumed to be a finite set, $E^p = \{\gamma_1^p, \ldots, \gamma_{\tau^p}^p\}$, with Levy measure $\nu^p(\{\gamma_l^p\}) = \lambda^p P_l^p$, where $\lambda^p = \nu^p(\mathbb{R})$ is the intensity rate, i.e., P_l^p is the probability of jump size γ_l^p and

$$\int_{E^p} \Gamma^p \hat{N}^p(d\Gamma^p, dt) = \sum_{l=1}^{\tau^p} \gamma_l^p \hat{N}^p(\{\gamma_l^p\}, dt), \quad p = 1, \ldots, c. \tag{3}$$

We consider the BSDEJ:

$$dY_t = -f(t, Y_t, Z_t, U_t)dt + Z_t' dW_t + \sum_{p=1}^{c} \int_{E^p} U_t^p(\Gamma^p)\hat{N}^p(d\Gamma^p, dt), \quad 0 \leqslant t \leqslant T, \tag{4}$$

$$Y_T = \xi, \tag{5}$$

or, equivalently,

$$\begin{aligned} Y_t = \xi &+ \int_t^T f(s, Y_s, Z_s, U_s)ds \\ &- \int_t^T Z_s' dW_s - \sum_{p=1}^{c} \int_t^T \int_{E^p} U_s^p(\Gamma^p)\hat{N}^p(d\Gamma^p, ds), \quad 0 \leqslant t \leqslant T, \end{aligned} \tag{6}$$

where the terminal condition $\xi : \Omega \to \mathbb{R}$ is an $\mathscr{F}_T$-measurable random variable and the generator $f : \Omega \times [0,T] \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^c \to \mathbb{R}$ is $\mathscr{P} \otimes \mathscr{B} \otimes \mathscr{B}^d \otimes \mathscr{B}^c$-measurable, being $\mathscr{P}$ the set of $\mathscr{F}_t$-progressively measurable scalar processes on $\Omega \times [0,T]$.

A solution of (4) is a triplet (Y, Z, U), with $Z = (Z^1, \ldots, Z^d)'$, $U = (U^1, \ldots, U^c)'$, such that $\{Y_t, t \in [0,T]\}$ is a $\mathscr{F}$-adapted process in $\mathscr{L}^\infty(\mathscr{F}, \mathbb{R})$

and $\{Z_t, t \in [0,T]\}$ and $\{U_t, t \in [0,T]\}$ are $\mathscr{F}$-predictable processes in $\mathscr{P}^2(\mathscr{F}, \mathbb{R}^d)$ and $\mathscr{P}^2(\mathscr{F}, \mathbb{R}^c)$, respectively, being:

- $\mathscr{L}^\infty(\mathscr{F}, \mathbb{R}^m)$ the set of $\mathscr{F}$-adapted $\mathbb{P}$-essentially bounded $\mathbb{R}^m$-valued processes on $[0,T]$,
- $\mathscr{P}^2(\mathscr{F}, \mathbb{R}^m)$ the set of $\mathscr{F}$-predictable $\mathbb{R}^m$-valued processes ϕ on $[0,T]$ under $\mathbb{P}$ with norm

$$\|\phi\|_2 := \left(\mathbb{E}^{\mathbb{P}}[\int_0^T |\phi_t|^2 dt]\right)^{\frac{1}{2}} < \infty.$$

See [14] for conditions of existence and uniqueness of solution.

Now we suppose that the randomness of the parameters (f, ξ) of the BSDEJ comes from the state variable $X_t = (X_t^1, \ldots, X_t^n)' \in \mathbb{R}^n$, i.e.:

$$\begin{aligned} dY_t = &-f(t, X_t, Y_t, Z_t, U_t)dt + Z_t' dW_t \\ &+ \sum_{p=1}^{c} \int_{E^p} U_t^p(\Gamma^p)\hat{N}^p(d\Gamma^p, dt), \quad 0 \leqslant t \leqslant T, \end{aligned} \tag{7}$$

$$Y_T = g(X_T), \tag{8}$$

where the functions $f : [0,T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^c \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are $\mathbb{R}$-valued Borel functions and $\{X_t, t \in [0,T]\}$ is the solution to

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t + \varepsilon(t, X_t) \int_E \Gamma N(d\Gamma, dt), \quad 0 \leqslant t \leqslant T, \tag{9}$$

$$X_0 = x_0, \tag{10}$$

with $x_0 \in \mathbb{R}^n$, $\mu : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n \times d}$ and $\epsilon : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n \times r}$:

$$x_0 = \begin{pmatrix} x_0^1 \\ \vdots \\ x_0^n \end{pmatrix}, \quad \mu(t, X_t) = \begin{pmatrix} \mu^1(t, X_t) \\ \vdots \\ \mu^n(t, X_t) \end{pmatrix}, \quad \sigma(t, X_t) = \begin{pmatrix} \sigma^{11}(t, X_t) & \cdots & \sigma^{1d}(t, X_t) \\ \vdots & & \vdots \\ \sigma^{n1}(t, X_t) & \cdots & \sigma^{nd}(t, X_t) \end{pmatrix},$$

$$\varepsilon(t, X_t) = \begin{pmatrix} \varepsilon^{11}(t, X_t) & \cdots & \varepsilon^{1c}(t, X_t) \\ \vdots & & \vdots \\ \varepsilon^{n1}(t, X_t) & \cdots & \varepsilon^{nc}(t, X_t) \end{pmatrix}, \quad \int_E \Gamma N(d\Gamma, dt) = \begin{pmatrix} \int_{E^1} \Gamma^1 N^1(d\Gamma^1, dt) \\ \vdots \\ \int_{E^c} \Gamma^c N^c(d\Gamma^c, dt) \end{pmatrix}.$$

The coupled system (7)–(9) is said to represent a FBSDEJ system.

In the next section we show interesting numerical examples in which FBSDEs appear.

3 Numerical Examples with FBSDEs: Pricing and Hedging

Following [9], in this section we present the problem of pricing and hedging derivative contracts depending on more than one asset.

We consider a financial market with $n+1$ assets, consisting of 1 bond and n risky assets with price processes B and $(S^1,\dots,S^n)$, respectively. We assume that such price processes are solutions to the FSDEs:

$$\begin{cases} dB_t = rB_t dt, \quad B_0 = 1, \\ \dfrac{dS_t^j}{S_{t-}^j} = \bar{\mu}^j dt + \displaystyle\sum_{i=1}^{d} \bar{\sigma}^{ji} dW_t^i \\ \qquad\qquad + \displaystyle\sum_{p=1}^{c} \bar{\varepsilon}^{jp} \int_{E^p} \bar{\Gamma}^p N^p(d\Gamma^p, dt), \quad S_0^j = s_0^j, \quad j = 1,\dots,n. \end{cases} \tag{11}$$

Remark 3.1 In [9], the interest rate r, the terms $\bar{\mu}^j$, $\bar{\sigma}^{ji}$ and $\bar{\varepsilon}^{jp}$ are assumed to be random variables. However we consider constant values r, $\bar{\mu}^j$, $\bar{\sigma}^{ji}$ and $\bar{\varepsilon}^{jp}$. Notice that we will be in the constant parameters framework by considering the FSDE of the processes $X^j := \ln(S^j)$.

Consider an investor in this financial market who faces at time T some liability ξ, whose uncertain value depends on $(W^1,\dots,W^d)$ and on $(N^1,\dots,N^c)$. The investor would like to reduce the uncertainty. One method to minimize his/her risk is to invest in assets that depend on the same sources of uncertainty as ξ.

Let ω_t be the agent wealth at time $t \in [0,T]$ and suppose that the initial agent wealth is $\tilde{\omega}_0$. Let π_t^j denote the invested amount in each asset S^j at time t, with $j = 1,\dots,n$. The amount invested in the bond B will be $\omega_t - \sum_{j=1}^n \pi_t^j$. Therefore the wealth process associated to this strategy is the solution of the SDE:

$$\begin{cases} d\omega_t = \left(\omega_t - \displaystyle\sum_{j=1}^{n} \pi_t^j\right) \frac{dB_t}{B_t} + \sum_{j=1}^{n} \pi_t^j \frac{dS_t^j}{S_t^j} = \\ \qquad = \left[r\omega_t + \displaystyle\sum_{j=1}^{n} (\bar{\mu}^j - r)\pi_t^j \right] dt + \sum_{j=1}^{n}\sum_{i=1}^{d} \pi_t^j \bar{\sigma}^{ji} dW_t^i \\ \qquad\quad + \displaystyle\sum_{j=1}^{n} \pi_t^j \sum_{p=1}^{c} \bar{\varepsilon}^{jp} \int_{E^p} \bar{\Gamma}^p N^p(d\Gamma^p, dt), \\ \omega_0 = \tilde{\omega}_0. \end{cases} \tag{12}$$

The class of admissible portfolios is the set:

$$\mathscr{U} = \left\{ \pi = (\pi^1,\dots,\pi^n) : [0,T]\times\Omega \to \mathbb{R}^n / \pi_t \text{ is } \mathscr{F}\text{-predictable and} \right.$$
$$\left. \mathbb{E}\left[\int_0^T |\pi_t|^2 dt\right] < +\infty \right\}. \tag{13}$$

The objective is to find a hedging portfolio π_t such that the terminal value of this investment ω_T is as close as possible to the value ξ. We have the following stochastic control problem:

$$\begin{cases} \min\limits_{\pi\in\mathscr{U}} \mathbb{E}[\xi-\omega_T]^2, \\ \text{subject to} \\ \quad d\omega_t = \left[r\omega_t + \sum\limits_{j=1}^{n}(\bar{\mu}^j - r)\pi_t^j\right]dt + \sum\limits_{j=1}^{n}\sum\limits_{i=1}^{d}\pi_t^j\bar{\sigma}^{ji}dW_t^i \\ \qquad\quad + \sum\limits_{j=1}^{n}\pi_t^j\sum\limits_{p=1}^{c}\bar{\varepsilon}^{jp}\int_{E^p}\bar{\Gamma}^p N^p(d\Gamma^p, dt), \\ \quad \omega_0 = \tilde{\omega}_0. \end{cases} \tag{14}$$

The solution to this problem is different depending on the kind of market in which the investor is. In the next sections we study how to address (14) in the case of a complete market without jumps and in the cases of an incomplete market without and with jumps.

3.1 Pricing and Hedging Without Jumps (Case $c = 0$)

3.1.1 Pricing and Perfect Hedging in Complete Markets Without Jumps (Case $n = d, c = 0$)

In a complete market, that means a market with the same number of sources of uncertainty than risky assets, an investor with the appropriate initial wealth $\tilde{\omega}_0$ can eliminate all the risk by replicating ξ, that is, there is a unique value ω_0 and an unique associated trading strategy π such that an investor with initial wealth ω_0 and investing according to π will have a terminal wealth satisfying $\omega_T = \xi$, $\mathbb{P}$-a.s. Therefore a perfect hedging is possible in this case.

Taking (12) in matrix form, $Y_t = \omega_t$ and $Z_t' = (Z_t^1, \ldots, Z_t^n) = \pi_t\bar{\sigma}$, we obtain

$$dY_t = \left[rY_t + Z_t'\bar{\sigma}^{-1}(\bar{\mu} - r\mathbf{1})\right]dt + Z_t'dW_t, \quad 0 \leqslant t \leqslant T, \quad Y_T = \xi, \tag{15}$$

where $\mathbf{1} \in \mathscr{M}^{n\times 1}$ and Y_0 corresponds to the value of the portfolio at time 0 that matches the price of the claim ξ at that time.

Remark 3.2 The completeness condition of the market (number of assets equal to number of uncertainty sources) makes invertible the matrix $\bar{\sigma}$, allowing introduce $\pi = Z_t'\bar{\sigma}^{-1}$ in the driven function.

If the terminal condition is a function of $S_T^1, \ldots, S_T^n$, the BCOS methodology explained in the next section can be applied to approximate the solution to (15).

3.1.2 Pricing and Quadratic Hedging in Incomplete Markets Without Jumps (Case $n < d, c = 0$)

In a incomplete market, that means a market with the less sources of uncertainty than risky assets, perfect replication is usually not possible. Super-replication (find a portfolio such that $\omega_T \geqslant \xi$, $\mathbb{P}$-a.s.) may be possible, but it is unfeasible since the required initial wealth is usually too large to be of practical use. Then, an investor in an incomplete market (or in a complete market but with insufficient initial capital to replicate the claim) needs to solve, for instance, the problem (14) with $c = 0$. In [8] can be seen that such problem is equivalent to solve the following BSDE:

$$dY_t = \left[rY_t + (\bar{\mu} - r\mathbf{1})'(\bar{\sigma}\bar{\sigma}')^{-1}\bar{\sigma}Z_t\right]dt + Z_t'dW_t, \quad 0 \leqslant t \leqslant T, \quad Y_T = \xi. \tag{16}$$

Remark 3.3 See [8] for the derivation of and specific results of existence and uniqueness of solution to (16). Since random parameters are considered in [8], then to solve (14) is equivalent to solve one system consisting of one Stochastic Riccati Equation (SRE) and one BSDE. For simplicity, we consider constant parameters avoiding the SRE.

With (Y, Z) solution to (16), the solution of (14) is given by:

$$\pi_t = (\bar{\sigma}\bar{\sigma}')^{-1}\left[\bar{\sigma}Z_t + (\bar{\mu} - r\mathbf{1})(Y_t - \omega_t)\right]. \tag{17}$$

Another problem that an investor may be interested in solving is the problem of pricing the claim ξ. In the case of a complete market, the unique arbitrage-free price at time 0 is the value $\tilde{\omega}_0$ such that the objective function $J(\omega_0) = \mathbb{E}[\xi - \omega_T]^2$ is equal to zero. In the case of an incomplete market we can consider the mean-variance price at time 0 given by:

$$\arg\min_{\omega_0} J(\omega_0) = Y_0. \tag{18}$$

Under which conditions Eqs. (17) and (18) are satisfied can be seen in [8].

If the terminal condition is a function of $S_T^1, \ldots, S_T^n$, the BCOS methodology explained in the next section can be applied to approximate the solution to (16).

3.1.3 Note on Terminal Condition in BSDEs Without Jumps

The solution of a BSDE without jumps is given by the pair (Y, Z). In this case is possible to have an expression of both components of the solution at the terminal time thanks to the following extension of the usual Feynman-Kac theorem, that was stated in [12] and gives a relation between BSDE and semilinear PDEs.

Theorem 3.4 *Let v be a classical solution to the semilinear PDE*

$$\frac{\partial v}{\partial t}(t,x) + \mathscr{L}v(t,x) + f(t,x,v(t,x),[D_x v(t,x)\sigma(t,x)]') = 0, \quad (t,x) \in [0,T) \times \mathbb{R}^n, \tag{19}$$

$$v(T,x) = g(x), \quad x \in \mathbb{R}^n, \tag{20}$$

where D_x is the gradient of v, $D_x v(t,x) = (\partial_{x^1} v(t,x), \ldots, \partial_{x^n} v(t,x))$, and $\mathscr{L}$ denotes a second order differential operator. Assume that a constant $C \geqslant 0$ exists such that, for all (t,x), $|v(t,x)| + |D_x v(t,x)\sigma(t,x)| \leqslant C(1+|x|)$. Then the pair (Y_t, Z_t) defined by $(v(t,X_t), [D_x v(t,X_t)\sigma(t,X_t)]')$, $s \leqslant t \leqslant T$, is the solution to the BSDE

$$dY_t = -f(t,X_t,Y_t,Z_t)dt + Z_t' dW_t, \quad 0 \leqslant t \leqslant T, \quad Y_T = g(X_T), \tag{21}$$

where the functions $f : [0,T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are $\mathbb{R}$-valued Borel functions and $\{X_t, t \in [0,T]\}$ is the solution to

$$dX_t = \mu(t,X_t)dt + \sigma(t,X_t)dW_t, \quad 0 \leqslant t \leqslant T, \quad X_0 = x_0, \tag{22}$$

with $x_0 \in \mathbb{R}^n$, $\mu : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n \times d}$.

The converse results states: Suppose (Y_t, Z_t) is the solution to the BSDE (21), then the function defined by $v(t,x) = Y_t$ is the solution to (19)–(20).

3.2 Pricing and Quadratic Hedging with Jumps (Case $c > 0$)

Since the number of uncertainty sources $(d + c)$ is bigger than the number of assets (n) we have an incomplete market. Therefore, as in the previous case perfect replication is usually not possible. We need to solve the problem (14). In [8] can be seen that such problem is equivalent to solve the following BSDE:

$$\begin{aligned} dY_t = &\left[rY_t + A'\Sigma^{-1}\bar{\sigma}Z_t + A'\Sigma^{-1}\bar{\varepsilon}D\bar{\Gamma}U_t\right]dt + Z_t' dW_t \\ &+ \sum_{p=1}^{c} \int_{E^p} U_t^p(\Gamma^p)\hat{N}^p(d\Gamma^p, dt), \quad 0 \leqslant t \leqslant T, \end{aligned} \tag{23}$$

$$Y_T = \xi, \tag{24}$$

where

$$A = (A_1, \ldots, A_n)', \quad \text{with } A_j = \bar{\mu}^j + \sum_{p=1}^{c} \bar{\varepsilon}^{jp} \int_{E^p} \bar{\Gamma}^p \nu^p(d\Gamma^p) - r, \tag{25}$$

$$\Lambda = \bar{\varepsilon}\mathrm{diag}\Big(\sum_{l=1}^{\tau^1} \bar{\gamma}_l^1 P_l^1, \ldots, \sum_{l=1}^{\tau^c} \bar{\gamma}_l^c P_l^c\Big), \tag{26}$$

$$D = \text{diag}(\lambda^1, \ldots, \lambda^c), \tag{27}$$

$$\Sigma = \bar{\sigma}\bar{\sigma}' + \Lambda D \Lambda', \tag{28}$$

$$\bar{\Gamma} U_t = \left(\sum_{l=1}^{\tau^1} \bar{\gamma}_l^1 P_l^1 U^1(\gamma_l^1), \ldots, \sum_{l=1}^{\tau^c} \bar{\gamma}_l^c P_l^c U^c(\gamma_l^c)\right)'. \tag{29}$$

Remark 3.5 See [9] for the derivation of and specific results of existence and uniqueness of solution to (23). As in the no jumps case, since random parameters are considered in [9], then to solve (14) is equivalent to solve one system consisting of one SRE and one BSDE. For simplicity, we consider constant parameters avoiding the SRE.

With (Y, Z, U) solution to (23), the solution of (14) is given by:

$$\pi_t = \Sigma^{-1}\left[\bar{\sigma} Z_t + \bar{\varepsilon} D \bar{\Gamma} U_t + A(Y_{t_-} - \omega_t)\right]. \tag{30}$$

Under which conditions Eq. (30) is satisfied can be seen in [9].

If the terminal condition is a function of $S_T^1, \ldots, S_T^n$, the BCOS methodology explained in the next section can be applied to approximate the solution to (23).

In the next section we propose a technique to approximate the solution of the BSDE (7) that is an extension to higher dimensions of the method developed in [16]. This technique can be used to solve the different FBSDEs presented in this section.

4 Fourier-Cosine Method to Solve the BSDEJ

In [16] a Fourier-cosine method to solve BSDEs in which the terminal condition depends on one process is developed. This technique consists of two steps:

1. the discretization of the FBSDE by a general θ-method for the time-integration,
2. (BCOS method) the approximation of the conditional expectations that arise after the corresponding discretization by the use of the COS method developed in [5].

In this work, the methodology is extended to the case in which the terminal condition depends on more than one process. The extension of the BCOS technique follows the same steps as in [16] and takes into account ideas developed in [15].

4.1 Discretization of the FBSDEJ

We wish to discretize the coupled system (7)–(9) that is equivalent to:

$$X_t^j = X_0^j + \int_0^t \mu^j(s, X_s)ds + \sum_{i=1}^{d} \int_0^t \sigma^{ji}(s, X_s)dW_s^i + \sum_{p=1}^{c} \int_0^T \varepsilon^{jp}(s, X_s) \int_{E^p} \Gamma^p N^p(d\Gamma^p, ds), \quad 0 < t \leqslant T, \tag{31}$$

$$X_0^j = x_0^j, \tag{32}$$

with $j = 1, \ldots, n$, and:

$$Y_t = Y_T + \int_t^T f(s, \Upsilon_s)ds - \sum_{i=1}^{d} \int_t^T Z_s^i dW_s^i - \sum_{p=1}^{c} \int_t^T \int_{E^p} U_s^p(\Gamma^p)\hat{N}^s(d\Gamma^p, ds), \quad 0 \leqslant t < T, \tag{33}$$

$$Y_T = g(X_T), \tag{34}$$

where $\Upsilon_s = \big((X_s^1, \ldots, X_s^n), Y_s, (Z_s^1, \ldots, Z_s^d), (U_s^1, \ldots, U_s^c)\big)'$.

In order to obtain the discretization of the coupled system (31)–(33) we consider the partition $\{t_m\}_{m=0}^M$ with $t_0 = 0$, $t_M = T$ and $\Delta t = t_{m+1} - t_m > 0$. We use the simplification of notation:

$$\Upsilon_m \equiv \Upsilon_{t_m}, \quad X_m^j \equiv X_{t_m}^j, \quad Y_m \equiv Y_{t_m}, \quad Z_m^i \equiv Z_{t_m}^i, \quad U_m^p \equiv U_{t_m}^p,$$
$$\Delta W_m^i \equiv W_{t_{m+1}}^i - W_{t_m}^i, \quad \hat{N}^p(d\Gamma^p, \Delta t) \equiv \hat{N}^p(d\Gamma^p, t_{m+1}) - \hat{N}^p(d\Gamma^p, t_m).$$

Therefore, we obtain:

$$X_0^j = x_0^j, \tag{35}$$

$$X_{m+1}^j = X_m^j + \int_{t_m}^{t_{m+1}} \mu^j(s, X_s)ds + \sum_{i=1}^{d} \int_{t_m}^{t_{m+1}} \sigma^{ji}(s, X_s)dW_s^i + \sum_{p=1}^{c} \varepsilon^{jp}(s, X_s) \int_{t_m}^{t_{m+1}} \int_{E^p} \Gamma^p N^p(d\Gamma^p, ds), \tag{36}$$

$$m = 0, \ldots, M-1,$$

with $j = 1, \ldots, n$, and:

$$Y_M = g(X_M), \tag{37}$$

$$\begin{aligned} Y_m = Y_{m+1} &+ \int_{t_m}^{t_{m+1}} f(s, \Upsilon_s) ds - \sum_{i=1}^{d} \int_{t_m}^{t_{m+1}} Z_s^i dW_s^i \\ &- \sum_{p=1}^{c} \int_{t_m}^{t_{m+1}} \int_{E^p} U_s^p(\Gamma^p) \hat{N}^s(d\Gamma^p, ds), \\ m &= M-1, \ldots, 0. \end{aligned} \tag{38}$$

Then, the following approximation of the solution to (31) is derived by applying the Euler discretization scheme:

$$X_0^j = x_0^j, \tag{39}$$

$$\begin{aligned} X_{m+1}^j = X_m^j &+ \mu^j(t_m, X_m)\Delta t + \sum_{i=1}^{d} \sigma^{ji}(t_m, X_m)\Delta W_m^i \\ &+ \sum_{p=1}^{c} \varepsilon^{jp}(t_m, X_m) \int_{E^p} \Gamma^p N^p(d\Gamma^p, \Delta t), \\ m &= 0, \ldots, M-1, \end{aligned} \tag{40}$$

with $j = 1, \ldots, n$. However, it is not possible to apply the same discretization scheme backwards in time to (38), because it would not take into account the adaptability constraint on the solution. Following [16], conditional expectations are taken to go backwards in time. Let $\mathbb{E}_m[\cdot]$ denote $\mathbb{E}[\cdot|\mathscr{F}_{t_m}]$. Taking conditional expectations at both sides of (38), at both sides of (38) multiplied by ΔW_m^i, $i = 1, \ldots, d$, and at both sides of (38) multiplied by $\hat{N}(\{\gamma_l^p\}, \Delta t)$, $p = 1, \ldots, c$, $l = 1, \ldots, \tau^p$, we obtain:

$$Y_m \approx \mathbb{E}_m[Y_{m+1}] + \Delta t \theta_Y f(t_m, \Upsilon_m) + \Delta t(1-\theta_Y)\mathbb{E}_m[f(t_{m+1}, \Upsilon_{m+1})], \quad \theta_Y \in [0,1], \tag{41}$$

$$\begin{aligned} 0 \approx\ & \mathbb{E}_m[Y_{m+1}\Delta W_m^i] + \Delta t(1-\theta_Z^i)\mathbb{E}_m[f(t_{m+1}, \Upsilon_{m+1})\Delta W_m^i] - \\ & - \Delta t \theta_Z^i Z_m^i - \Delta t(1-\theta_Z^i)\mathbb{E}_m[Z_{m+1}^i], \quad \theta_Z^i \in [0,1], \quad i = 1, \ldots, d \end{aligned} \tag{42}$$

$$\begin{aligned} 0 \approx\ & \mathbb{E}_m[Y_{m+1}\hat{N}^p(\{\gamma_l^p\}, \Delta t)] + \Delta t(1-\theta_U^p)\mathbb{E}_m[f(t_{m+1}, \Upsilon_{m+1})\hat{N}^p(\{\gamma_l^p\}, \Delta t)] - \\ & - P_l^p \lambda^p \Delta t \theta_U^p U_m^p(\gamma_l^p) - P_l^p \lambda^p \Delta t(1-\theta_U^p)\mathbb{E}_m[U_{m+1}^p(\gamma_l^p)], \quad \theta_U^p \in [0,1], \\ & p = 1, \ldots, c, \\ & l = 1, \ldots, \tau^p. \end{aligned} \tag{43}$$

These equations allow the following approximation for the solution of (7):

1. $Y_M = g(X_M)$.
2. For $m = M-1, \ldots, 0$:

$$\begin{aligned} Z_m^i &= -(\theta_Z^i)^{-1}(1-\theta_Z^i)\mathbb{E}_m[Z_{m+1}^i] + \Delta t^{-1}(\theta_Z^i)^{-1}\mathbb{E}_m[Y_{m+1}\Delta W_m^i] + \\ &\quad + (\theta_Z^i)^{-1}(1-\theta_Z^i)\mathbb{E}_m[f(t_{m+1}, \Upsilon_{m+1})\Delta W_m^i], \quad i = 1, \ldots, d, \end{aligned} \tag{44}$$

$$\begin{aligned} U_m^p(\gamma_l^p) &= -(\theta_U^p)^{-1}(1-\theta_U^p)\mathbb{E}_m[U_{m+1}^p(\gamma_l^p)] \\ &\quad + (P_l^p\lambda^p\Delta t)^{-1}(\theta_U^p)^{-1}\mathbb{E}_m[Y_{m+1}\hat{N}^p(\{\gamma_l^p\}, \Delta t)] + \\ &\quad + (P_l^p\lambda^p)^{-1}(\theta_U^p)^{-1}(1-\theta_U^p)\mathbb{E}_m[f(t_{m+1}, \Upsilon_{m+1})\hat{N}^p(\{\gamma_l^p\}, \Delta t)], \\ &\quad p = 1, \ldots, c, \\ &\quad l = 1, \ldots, \tau^p, \end{aligned} \tag{45}$$

$$Y_m = \mathbb{E}_m[Y_{m+1}] + \Delta t\theta_Y f(t_m, \Upsilon_m) + \Delta t(1-\theta_Y)\mathbb{E}_m[f(t_{m+1}, \Upsilon_{m+1})]. \tag{46}$$

The terminal condition is a deterministic function of X_M and X is a Markov process. Therefore, there are deterministic functions $y, z^1, \ldots, z^d, u^1, \ldots, u^c$ so that

$$Y_m = y(t_m, X_m) = y(t_m, (X_m^1, \ldots, X_m^n)), \tag{47}$$

$$Z_m^i = z^i(t_m, X_m) = z^i(t_m, (X_m^1, \ldots, X_m^n)), \quad i = 1, \ldots, d, \tag{48}$$

$$\begin{aligned} U_m^p(\gamma_l^p) &= u^p(t_m, X_m, \gamma_l^p) = u^p(t_m, (X_m^1, \ldots, X_m^n), \gamma_l^p), \quad p = 1, \ldots, c, \\ &\quad l = 1, \ldots, \tau^p. \end{aligned} \tag{49}$$

Then $\mathbb{E}_m[\cdot]$ can be replaced by $\mathbb{E}_m^x[\cdot]$, where $\mathbb{E}_m^x[\cdot]$ is the expectation conditioned to $X_m = x \in \mathbb{R}^n$. We have a scheme to solve the BSDE backwards in time in which the approximation of conditional expectations is necessary. Following [15], we propose the approximation of these expectations by Fourier-cosine series expansions.

4.2 BCOS Method

In this section we propose a methodology to approximate the conditional expectations that appear after the discretization. This technique is called BCOS method and is based on the close relation of the characteristic function of the process X with the coefficients of the Fourier-cosine expansion of its density function.

We present the case $n = 2$, however the methodology can be extended to higher dimensions (see [15, Sect. 8] and [13]). For simplicity, we assume constant parameters μ, σ and ε in (9), although the use of more general terms is possible.

Then we obtain:

$$X_0^j = x_0^j, \quad j = 1, 2 \tag{50}$$

$$\begin{pmatrix} X_{m+1}^1 \\ X_{m+1}^2 \end{pmatrix} = \begin{pmatrix} X_m^1 \\ X_m^2 \end{pmatrix} + \begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix} \Delta t + \begin{pmatrix} \sigma^{11} \cdots \sigma^{1d} \\ \sigma^{21} \cdots \sigma^{2d} \end{pmatrix} \begin{pmatrix} \Delta W_m^1 \\ \vdots \\ \Delta W_m^d \end{pmatrix}$$
$$+ \begin{pmatrix} \varepsilon^{11} \cdots \varepsilon^{1c} \\ \varepsilon^{21} \cdots \varepsilon^{2c} \end{pmatrix} \begin{pmatrix} \int_{E^1} \Gamma^1 N^1(d\Gamma^1, \Delta t) \\ \vdots \\ \int_{E^c} \Gamma^c N^c(d\Gamma^c, \Delta t), \end{pmatrix},$$
$$m = 0, \ldots, M-1, \tag{51}$$

and the characteristic function in $u = (u^1, u^2)'$ of the process $X = (X^1, X^2)'$ given $x = (x^1, x^2)'$ is

$$\varphi(u|x) = \varphi(u|(0,0)')e^{iu'x} = \phi(u)e^{iu'x}, \tag{52}$$

with

$$\phi(u) = \exp\left(i\mu' u \Delta t - \frac{1}{2} u'\sigma\sigma' u \Delta t\right) \prod_{p=1}^{c} \exp\left(\lambda^p \Delta t(\varphi_{\Gamma^p}(u) - 1)\right), \tag{53}$$

where $\varphi_{\Gamma^p}(u) = \sum_{l=1}^{\tau^p} P_l^p e^{i(\varepsilon^{1p}, \varepsilon^{2p})u\gamma_l^p}$.

The characteristic function allows the approximation of the conditional expectations $\mathbb{E}_m^x[\cdot]$, $\mathbb{E}_m^x[\cdot \Delta W_m^i]$ and $\mathbb{E}_m^x[\cdot \hat{N}^p(\{\gamma_l^p\}, \Delta t)]$ by the following formulas[1]:

- $I^A := \mathbb{E}_m^x[v(t_{m+1}, X_{m+1}^1, X_{m+1}^2)]$:

$$I^A \approx \sum_{k^1=0}^{N^1-1} {}' \sum_{k^2=0}^{N^2-1} {}' \mathscr{V}_{k^1k^2}(t_{m+1}) \Pi_{k^1k^2}^A(x^1, x^2), \tag{54}$$

with

$$\Pi_{k^1k^2}^A(x^1, x^2) := \frac{b^1 - a^1}{2} \frac{b^2 - a^2}{2} \frac{1}{2} \left[\Phi_{k^1k^2}^+(x^1, x^2) + \Phi_{k^1k^2}^-(x^1, x^2)\right], \tag{55}$$

[1] $\sum'$ denotes that the first term in the sum is multiplied by $\frac{1}{2}$.

where

$$\Phi^{\pm}_{k^1k^2}(x^1,x^2) := \frac{2}{b^1-a^1}\frac{2}{b^2-a^2} Re\left\{\exp\left(ik^1\pi\frac{x^1-a^1}{b^1-a^1} \pm ik^2\pi\frac{x^2-a^2}{b^2-a^2}\right)\right.$$
$$\left.\times\phi\left(\frac{k^1\pi}{b^1-a^1}, \pm\frac{k^2\pi}{b^2-a^2}\right)\right\}. \tag{56}$$

- $I^{B(i)} := \mathbb{E}^x_m[v(t_{m+1}, X^1_{m+1}, X^2_{m+1})\Delta W^i_m]$ $(i = 1, \ldots, d)$:

$$I^{B(i)} \approx \sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \mathscr{V}_{k^1k^2}(t_{m+1}) \Pi^{B(i)}_{k^1k^2}(x^1,x^2), \tag{57}$$

with

$$\Pi^{B(i)}_{k^1k^2}(x^1,x^2) = \frac{b^1-a^1}{2}\frac{b^2-a^2}{2}\frac{1}{2}(\sigma^{1i},\sigma^{2i})\left[\nabla\Phi^{+}_{k^1k^2}(x^1,x^2) + \nabla\Phi^{-}_{k^1k^2}(x^1,x^2)\right]. \tag{58}$$

- $I^{C(p,l)} := \mathbb{E}_m[v(t_{m+1}, X^1_{m+1}, X^2_{m+1})\hat{N}^p(\{\gamma^p_l\}, \Delta t)]$ $(p = 1, \ldots, c,\ l = 1, ..., \tau^p)$:

$$I^{C(p,l)} \approx \sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \mathscr{V}_{k^1k^2}(t_{m+1}) \frac{1}{2}\left[\Pi^{C(p,l)+}_{k^1k^2}(x^1,x^2) + \Pi^{C(p,l)-}_{k^1k^2}(x^1,x^2)\right], \tag{59}$$

with

$$\Pi^{C(p,l)\pm}_{k^1k^2}(x^1,x^2) = Re\left\{\exp\left(ik^1\pi\frac{x^1-a^1}{b^1-a^1} \pm ik^2\pi\frac{x^2-a^2}{b^2-a^2}\right)\right.$$
$$\times\phi\left(\frac{k^1\pi}{b^1-a^1}, \pm\frac{k^2\pi}{b^2-a^2}\right)\cdot$$
$$\left.\cdot\, P^p_l\lambda^p\Delta t\left[\exp\left(i(\varepsilon^{1p}\frac{k^1\pi}{b^1-a^1} \pm \varepsilon^{2p}\frac{k^2\pi}{b^2-a^2})\lambda^p\right) - 1\right]\right\}. \tag{60}$$

These formulas can be obtained adapting the derivation in [15, 16]. First we put the conditional expectation as a double integer of the density function multiplied by the function v. Secondly, the infinite integration range of the expectations are truncated and a sufficiently large interval $[a^1, b^1] \times [a^2, b^2] \subset \mathbb{R}^2$ is considered. Thirdly, the density function and the function v are replaced by the Fourier-cosine series expansions on $[a^1, b^1]\times[a^2, b^2]$ and the series summations in these expansions are truncated by taking finite summations with N^1 and N^2 terms. Finally, the Fourier-cosine coefficients of the density function are approximated following [5]. Then, the terms $\mathscr{V}_{k^1k^2}(t_{m+1})$ denote the Fourier-cosine coefficients of the function v.

Briefly, all derivations carried out in [16] can be adapted to the 2 dimensional case taking into account the equality $\cos(\alpha)\cos(\beta) = \frac{1}{2}[\cos(\alpha+\beta)+\cos(\alpha-\beta)]$.

4.2.1 Approximation of $y(t_m,(x^1,x^2)), z^i(t_m,(x^1,x^2)), i=1,\ldots,d$, and $u^p(t_m,(x^1,x^2),\gamma_l^p), p=1,\ldots,c, l=1,\ldots,\tau^p$

Taking into account the formulas (54)–(59) and (47)–(48), we can approximate the conditional expectations in (44)–(46) obtaining:

$$z^i(t_m,(x^1,x^2)) = \sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \left(-\frac{1-\theta_Z^i}{\theta_Z^i}\mathscr{Z}^i_{k^1k^2}(t_{m+1})\right)\Pi^A_{k^1k^2}(x^1,x^2)+$$
$$+\sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \left(\frac{1}{\Delta t\theta_Z^i}\mathscr{Y}_{k^1k^2}(t_{m+1})+\frac{1-\theta_Z^i}{\theta_Z^i}\mathscr{F}_{k^1k^2}(t_{m+1})\right)\Delta t\Pi^{B(i)}_{k^1k^2}(x^1,x^2), \tag{61}$$

$$u^p(t_m,(x^1,x^2),\gamma_l^p) = \sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \left(-\frac{1-\theta_U^p}{\theta_U^p}\mathscr{U}^{p,l}_{k^1k^2}(t_{m+1})\right)\Pi^A_{k^1k^2}(x^1,x^2)+$$
$$+\sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \left(\frac{1}{P_l^p\lambda^p\Delta t\theta_U^p}\mathscr{Y}_{k^1k^2}(t_{m+1})\right.$$
$$\left.+\frac{1-\theta_U^p}{P_l^p\lambda^p\theta_U^p}\mathscr{F}_{k^1k^2}(t_{m+1})\right)$$
$$\times\frac{1}{2}\left[\Pi^{C(p,l)+}_{k^1k^2}(x^1,x^2)+\Pi^{C(p,l)-}_{k^1k^2}(x^1,x^2)\right], \tag{62}$$

$$y(t_m,(x^1,x^2)) = \Delta t\theta_Y f(t_m,x^1,x^2,y(t_m,(x^1,x^2)),\{z^i(t_m,(x^1,x^2))\}_{i=1}^d,$$
$$\{\{u^l(t_m,(x^1,x^2),\gamma_l^p)\}_{l=1}^{\tau^p}\}_{p=1}^r)+$$
$$+h(t_m,(x^1,x^2)), \tag{63}$$

with

$$h(t_m,(x^1,x^2)) = \mathbb{E}^x_m[Y_{m+1}]+\Delta t(1-\theta_Y)\mathbb{E}^x_m[f(t_{m+1},\Upsilon_{m+1})]\approx$$
$$\approx\sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \left(\mathscr{Y}_{k^1k^2}(t_{m+1})+\Delta t(1-\theta_Y)\mathscr{F}_{k^1k^2}(t_{m+1})\right)\Pi^A_{k^1k^2}(x^1,x^2). \tag{64}$$

$\mathscr{F}_{k^1k^2}(t_{m+1})$, $\mathscr{Y}_{k^1k^2}(t_{m+1})$, $\mathscr{Z}^i_{k^1k^2}(t_{m+1})$ and $\mathscr{U}^{p,l}_{k^1k^2}(t_{m+1})$ denote the Fourier-cosine coefficients of f, y, z^i and u^p, respectively. Notice that we use the function h to

represent the explicit part in Eq. (46). Due to the implicit part in this equation we will need an iterative method to obtain the value $y(t_m,(x^1,x^2))$. We propose $y(t_m,(x^1,x^2)) \approx y^{PIt-1}(t_m,(x^1,x^2))$ where $y^{PIt-1}(t_m,(x^1,x^2))$ is the result obtained after *PIt* Picard iterations (see [7]), starting with an initial value:

$$y^0(t_m,(x^1,x^2)) = \mathbb{E}_m^x[Y_{m+1}] \approx \sum_{k^1=0}^{N^1-1}{}' \sum_{k^2=0}^{N^2-1}{}' \mathscr{Y}_{k^1k^2}(t_{m+1}) \Pi^A_{k^1k^2}(x^1,x^2). \tag{65}$$

4.2.2 Recovery of Fourier-Cosine Coefficients

In this section we explain an efficient algorithm to compute the Fourier-cosine coefficients backwards in time.

For this purpose, following [15], we define:

$$\mathscr{M}^{\pm}_{k^1k^2}(a,b) = \frac{2}{b-a}\int_a^b e^{\pm ik^2\pi\frac{x-a}{b-a}} cos\left(k^1\pi\right)\frac{x-a}{b-a}dx. \tag{66}$$

Taking into account this notation and the expressions (61), (62) and (64) for the functions $z^i(t_m,(x^1,x^2))$, $u^p(t_m,(x^1,x^2),\gamma_l^p)$ and $h(t_m,(x^1,x^2))$, respectively, we obtain:

$$\begin{aligned}
\mathscr{Z}^i_{k^1k^2}(t_m) = {} & \frac{2}{b^1-a^1}\frac{2}{b^2-a^2}\int_{a^1}^{b^1}\int_{a^2}^{b^2} z^i(t_m,(x^1,x^2))\cos\left(k^1\pi\frac{x^1-a^1}{b^1-a^1}\right) \\
& \times\cos\left(k^2\pi\frac{x^2-a^2}{b^2-a^2}\right)dx^1dx^2 \approx \\
\approx {} & Re\left\{\sum_{\hat{k}^1=0}^{N^1-1}{}'\sum_{\hat{k}^2=0}^{N^2-1}{}' \mathbf{V}^{Z^i_+}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\mathscr{M}^+_{k^1\hat{k}^1}(a^1,b^1)\mathscr{M}^+_{k^2\hat{k}^2}(a^2,b^2)\right\} + \\
& + Re\left\{\sum_{\hat{k}^1=0}^{N^1-1}{}'\sum_{\hat{k}^2=0}^{N^2-1}{}' \mathbf{V}^{Z^i_-}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\mathscr{M}^+_{k^1\hat{k}^1}(a^1,b^1)\mathscr{M}^-_{k^2\hat{k}^2}(a^2,b^2)\right\},
\end{aligned} \tag{67}$$

where

$$\begin{aligned}
\mathbf{V}^{Z^i_\pm}_{\hat{k}^1,\hat{k}^2}(t_{m+1}) = {} & \frac{1}{2}\phi\left(\frac{\hat{k}^1\pi}{b^1-a^1},\pm\frac{\hat{k}^2\pi}{b^2-a^2}\right)\left[-\frac{1-\theta^i_Z}{\theta^i_Z}\mathscr{Z}^i_{\hat{k}^1,\hat{k}^2}(t_{m+1}) + \right. \\
& + \Delta t\left(\sigma^{1i}\frac{i\hat{k}^1\pi}{b^1-a^1} \pm \sigma^{2i}\frac{i\hat{k}^2\pi}{b^2-a^2}\right)\left(\frac{1}{\Delta t\theta^i_Z}\mathscr{Y}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\right. \\
& \left.\left. + \frac{1-\theta^i_Z}{\theta^i_Z}\mathscr{F}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\right)\right].
\end{aligned} \tag{68}$$

$$\mathscr{U}^{p,l}_{k^1k^2}(t_m) = \frac{2}{b^1-a^1}\frac{2}{b^2-a^2}\int_{a^1}^{b^1}\int_{a^2}^{b^2} u^p(t_m,(X^1_m,X^2_m),\gamma^p_l)\cos\left(k^1\pi\frac{x^1-a^1}{b^1-a^1}\right)$$
$$\times\cos\left(k^2\pi\frac{x^2-a^2}{b^2-a^2}\right)dx^1dx^2 \approx$$
$$\approx Re\left\{\sum_{\hat{k}^1=0}^{N^1-1}{}'\sum_{\hat{k}^2=0}^{N^2-1}{}'\mathbf{V}^{U^{p,l}_+}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\mathscr{M}^+_{k^1\hat{k}^1}(a^1,b^1)\mathscr{M}^+_{k^2\hat{k}^2}(a^2,b^2)\right\}+$$
$$+Re\left\{\sum_{\hat{k}^1=0}^{N^1-1}{}'\sum_{\hat{k}^2=0}^{N^2-1}{}'\mathbf{V}^{U^{p,l}_-}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\mathscr{M}^+_{k^1\hat{k}^1}(a^1,b^1)\mathscr{M}^-_{k^2\hat{k}^2}(a^2,b^2)\right\}, \tag{69}$$

where

$$\mathbf{V}^{U^{p,l}_\pm}_{\hat{k}^1,\hat{k}^2}(t_{m+1}) = \frac{1}{2}\phi\left(\frac{\hat{k}^1\pi}{b^1-a^1},\pm\frac{\hat{k}^2\pi}{b^2-a^2}\right)\Bigg[-\frac{1-\theta^p_U}{\theta^p_U}\mathscr{U}^{p,l}_{\hat{k}^1,\hat{k}^2}(t_{m+1})+$$
$$+P^p_l\lambda^p\Delta t\left[\exp\left(i(\varepsilon^{1p}\frac{\hat{k}^1\pi}{b^1-a^1}\pm\varepsilon^{2p}\frac{\hat{k}^2\pi}{b^2-a^2})\lambda^p\right)-1\right]\cdot$$
$$\cdot\left(\frac{1}{P^p_l\lambda^p\Delta t\theta^p_U}\mathscr{Y}_{\hat{k}^1,\hat{k}^2}(t_{m+1})+\frac{1-\theta^p_U}{P^p_l\lambda^p\theta^p_U}\mathscr{F}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\right)\Bigg]. \tag{70}$$

$$\mathscr{H}_{k^1k^2}(t_m) = \frac{2}{b^1-a^1}\frac{2}{b^2-a^2}\int_{a^1}^{b^1}\int_{a^2}^{b^2} h(t_m,(x^1,x^2))\cos\left(k^1\pi\frac{x^1-a^1}{b^1-a^1}\right)$$
$$\times\cos\left(k^2\pi\frac{x^2-a^2}{b^2-a^2}\right)dx^1dx^2 \approx$$
$$\approx Re\left\{\sum_{\hat{k}^1=0}^{N^1-1}{}'\sum_{\hat{k}^2=0}^{N^2-1}{}'\mathbf{V}^{H_+}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\mathscr{M}^+_{k^1\hat{k}^1}(a^1,b^1)\mathscr{M}^+_{k^2\hat{k}^2}(a^2,b^2)\right\}+$$
$$+Re\left\{\sum_{\hat{k}^1=0}^{N^1-1}{}'\sum_{\hat{k}^2=0}^{N^2-1}{}'\mathbf{V}^{H_-}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\mathscr{M}^+_{k^1\hat{k}^1}(a^1,b^1)\mathscr{M}^-_{k^2\hat{k}^2}(a^2,b^2)\right\}, \tag{71}$$

where

$$\mathbf{V}^{H_\pm}_{\hat{k}^1,\hat{k}^2}(t_{m+1}) = \frac{1}{2}\phi\left(\frac{\hat{k}^1\pi}{b^1-a^1},\pm\frac{\hat{k}^2\pi}{b^2-a^2}\right)\Big[\mathscr{Y}_{\hat{k}^1,\hat{k}^2}(t_{m+1})+\Delta t(1-\theta_Y)\mathscr{F}_{\hat{k}^1,\hat{k}^2}(t_{m+1})\Big]. \tag{72}$$

Therefore, in each case we need compute $\mathscr{C}_{k^1k^2}(t_m) = \mathscr{C}^+_{k^1k^2}(t_m) + \mathscr{C}^-_{k^1k^2}(t_m)$, with

$$\mathscr{C}^{\pm}_{k^1k^2}(t_m) = Re\left\{ \sum_{\hat{k}^1=0}^{N^1-1}{}' \sum_{\hat{k}^2=0}^{N^2-1}{}' \mathbf{V}^{\pm}_{\hat{k}^1,\hat{k}^2}(t_{m+1}) \mathscr{M}^{+}_{k^1\hat{k}^1}(a^1,b^1) \mathscr{M}^{\pm}_{k^2\hat{k}^2}(a^2,b^2) \right\}. \tag{73}$$

Notice that the coefficients $\mathscr{C}_{k^1k^2}(t_m)$ are representing the Fourier-cosine coefficients $\mathscr{Z}^i_{k^1k^2}(t_m)$, $\mathscr{U}^{p,l}_{k^1k^2}(t_m)$ or $\mathscr{H}_{k^1k^2}(t_m)$ and the values $\mathbf{V}^{\pm}_{\hat{k}^1,\hat{k}^2}(t_{m+1})$ are representing $\mathbf{V}^{Z^i_{\pm}}_{\hat{k}^1,\hat{k}^2}(t_{m+1})$, $\mathbf{V}^{U^{p,l}_{\pm}}_{\hat{k}^1,\hat{k}^2}(t_{m+1})$ or $\mathbf{V}^{H_{\pm}}_{\hat{k}^1,\hat{k}^2}(t_{m+1})$.

The computation of these coefficients consists of two matrix-vector multiplications that can be efficiently obtained by an Fast Fourier Transform (FFT) algorithm, thanks to the desirable properties of the matrix with entries $\mathscr{M}^{\pm}_{k^1k^2}(a,b)$. See [6].

Finally,

$$\mathscr{Y}_{k^1k^2}(t_m) \approx \Delta t \theta_Y \mathscr{F}_{k^1k^2}(t_m) + \mathscr{H}_{k^1k^2}(t_m), \tag{74}$$

where the coefficients

$$\begin{aligned} \mathscr{F}_{k^1k^2}(t_m) = \frac{2}{b^1-a^1}\frac{2}{b^2-a^2}\int_{a^1}^{b^1}\int_{a^2}^{b^2} & f(t_m,(x^1,x^2),y(t_m,(x^1,x^2)), \\ & \{z^i(t_m,(x^1,x^2))\}_{i=1}^d, \{\{u^p(t_m,(x^1,x^2),\gamma_l^p)\}_{l=1}^{\tau^p}\}_{p=1}^c) \cdot \\ \cdot \cos\left(k^1\pi\frac{x^1-a^1}{b^1-a^1}\right)\cos\left(k^2\pi\frac{x^2-a^2}{b^2-a^2}\right) & dx^1dx^2 \end{aligned} \tag{75}$$

are approximated by a Discrete Fourier-Cosine Transform (DCT).

4.2.3 Algorithm

Briefly, the algorithm we propose to approximate the solution of the FBSDE is:

1. Compute $\mathscr{Y}_{k^1k^2}(t_M)$ from the terminal condition.[2]
2. For $m = M-1,\ldots,1$, compute:
   a. on an equidistant (x^1,x^2)-grid with $Q^1 \times Q^2$ grid points, $\{z^i(t_m,(x^1,x^2))\}_{i=1}^d$, $\{\{u^p(t_m,(x^1,x^2),\gamma_l^p)\}_{l=1}^{\tau^p}\}_{p=1}^c$, $h(t_m,(x^1,x^2)), y(t_m,(x^1,x^2))$ and $f(t_m,(x^1,x^2),y(t_m,(x^1,x^2)),\{z^i(t_m,(x^1,x^2))\}_{i=1}^d,\{\{u^p(t_m,(x^1,x^2),\gamma_l^p)\}_{l=1}^{\tau^p}\}_{p=1}^c)$, with (61)–(65),

[2] Since the terminal conditions Z_M and U_m are not known, we set $\theta_Y = \theta_Z^1 = \ldots = \theta_Z^d = \theta_U^1 = \ldots = \theta_U^c = 1$ in the first time iteration.

b. $\{\mathscr{Z}^i_{k^1k^2}(t_m)\}_{i=1}^d$, $\{\{\mathscr{U}^{p,l}_{k^1k^2}(t_m)\}_{l=1}^{\tau^p}\}_{p=1}^c$, $\mathscr{H}_{k^1k^2}(t_m)$, $\mathscr{F}_{k^1k^2}(t_m)$ and $\mathscr{Y}_{k^1k^2}(t_m)$, by using the Fourier-cosine coefficients of the time t_{m+1} and (67)–(75).

3. Compute $\{z^i(t_0,(x_0^1,x_0^2))\}_{i=1}^d$, $\{\{u^p(t_0,(x_0^1,x_0^2),\gamma_l^p)\}_{l=1}^{\tau^p}\}_{p=1}^c$, $y(t_0,(x_0^1,x_0^2))$ from (61), (62) and (65).

5 Results

In this section we present the obtained results for the case $n = 2$, i.e., two forward stochastic processes.

5.1 No Jumps Case in Complete Market

Following the numerical example presented in Sect. 3.1.1, we consider two assets whose prices S^1 and S^2 are driven by

$$\begin{pmatrix} dS_t^1 \\ dS_t^2 \end{pmatrix} = \begin{pmatrix} \bar{\mu}^1 S_t^1 \\ \bar{\mu}^2 S_t^2 \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}^{11} S_t^1 & \bar{\sigma}^{12} S_t^1 \\ \bar{\sigma}^{21} S_t^2 & \bar{\sigma}^{22} S_t^2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix}, \tag{76}$$

with $\bar{\sigma}^{11} = \hat{\sigma}^1$, $\bar{\sigma}^{12} = 0$, $\bar{\sigma}^{21} = \rho\hat{\sigma}^2$ and $\bar{\sigma}^{22} = \sqrt{1-\rho^2}\hat{\sigma}^2$. Notice that the processes are correlated by the parameter ρ. Taking $X_t^j = \ln(S_t^j), j = 1, 2$, we obtain:

$$\begin{pmatrix} dX_t^1 \\ dX_t^2 \end{pmatrix} = \begin{pmatrix} \bar{\mu}^1 - \frac{1}{2}(\hat{\sigma}^1)^2 \\ \bar{\mu}^2 - \frac{1}{2}(\hat{\sigma}^2)^2 \end{pmatrix} dt + \begin{pmatrix} \hat{\sigma}^1 & 0 \\ \rho\hat{\sigma}^2 & \sqrt{1-\rho^2}\hat{\sigma}^2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix}. \tag{77}$$

We consider a derivative contract with payoff $\xi = g(S_T^1, S_T^2) = g(e^{X_T^1}, e^{X_T^2})$. The BSDE for pricing and hedging such derivative contract is given by:

$$\begin{aligned} dY_t &= \left[rY_t + (Z_t^1, Z_t^2)\bar{\sigma}^{-1}(\bar{\mu} - r\mathbf{1})\right] dt + (Z_t^1, Z_t^2) \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix}, \\ 0 \leqslant t \leqslant T, &\quad Y_T = g(e^{X_T^1}, e^{X_T^2}), \end{aligned} \tag{78}$$

with

$$\bar{\mu} - r\mathbf{1} = \begin{pmatrix} \bar{\mu}^1 - r \\ \bar{\mu}^2 - r \end{pmatrix}, \quad \bar{\sigma} = \begin{pmatrix} \bar{\sigma}^{11} & \bar{\sigma}^{12} \\ \bar{\sigma}^{21} & \bar{\sigma}^{22} \end{pmatrix} = \begin{pmatrix} \hat{\sigma}^1 & 0 \\ \rho\hat{\sigma}^2 & \sqrt{1-\rho^2}\hat{\sigma}^2 \end{pmatrix}. \tag{79}$$

5.1.1 Method Data

According to [15], we take $a^1 = a^2 = a$ and $b^1 = b^2 = b$ for the range of integration in (75) and further, where

$$a := \min_{j=1,2}\left[X_0^j + \mathscr{K}_1^j - L\sqrt{\mathscr{K}_2^j}\right], \quad b := \max_{j=1,2}\left[X_0^j + \mathscr{K}_1^j + L\sqrt{\mathscr{K}_2^j}\right], \quad L = 10, \tag{80}$$

being $\mathscr{K}_i^j$ the ith cumulant of the variable X^j:

$$\mathscr{K}_1^j = [\bar{\mu}^j - \frac{1}{2}(\hat{\sigma}^j)^2]T, \quad \mathscr{K}_2^j = (\hat{\sigma}^j)^2 T. \tag{81}$$

We consider two different schemes for the discretization:

- Scheme A: $\Theta_Y = 0$, $\Theta_Z^1 = \Theta_Z^2 = 1$,
- Scheme B: $\Theta_Y = 0.5$, $\Theta_Z^1 = \Theta_Z^2 = 0.5$.

$PIt = 5$ Picard iterations are taken in the Scheme B, with which $y(t_m, (X_m^1, X_m^2))$ is solved implicitly.

We will take the values $N^1 = N^2 = N$ and $Q^1 = Q^2 = Q$.

5.1.2 Model Data

The different sets of model parameters that we consider are shown in Table 1.

5.1.3 Payoff Functions

We take into account derivatives whose payoffs are given by:

- Spread Option (SO): $g(S_T^1, S_T^2) = \max\left(S_T^1 - S_T^2, 0\right)$,
- Geometric Basket Call Option (Geo): $g(S_T^1, S_T^2) = \max\left(\sqrt{S_T^1}\sqrt{S_T^2} - K, 0\right)$,
- Arithmetic Basket Call Option (Arith): $g(S_T^1, S_T^2) = \max\left((S_T^1 + S_T^2)/2 - K, 0\right)$,
- Call-On-Maximum Option (COM): $g(S_T^1, S_T^2) = \max\left(\max(S_T^1, S_T^2) - K, 0\right)$,
- Put-On-Minimum Option (POM): $g(S_T^1, S_T^2) = \max\left(K - \min(S_T^1, S_T^2), 0\right)$.

Table 1 Model parameters

	S_0^1	S_0^2	$\bar{\mu}^1$	$\bar{\mu}^2$	$\hat{\sigma}^1$	$\hat{\sigma}^2$	ρ	r	T	K
Set I	100	100	0.1	0.1	0.25	0.3	0.3	0.05	0.1	100
Set II	90	110	0.1	0.1	0.2	0.3	0.25	0.04	1	100
Set III	40	40	0.1	0.1	0.2	0.3	0.5	0.048790	7/12	40

5.1.4 Results

Tables 2 and 3 show the results with different data and payoff functions. Table 1 shows the expected behavior respect to the convergence in the variable Y when M is increased: first order with Scheme A and second order with Scheme B. As we expected, Scheme B provides better results in the pricing of derivatives than Scheme A. Table 2 shows small errors in the pricing of several contracts depending on two assets. We can conclude that the extension of the BCOS method works in the two-dimensional problem.

5.2 *No Jumps Case in Incomplete Market*

Following the numerical example presented in Sect. 3.1.2, we consider two assets whose prices S^1 and S^2 are driven by

$$\begin{pmatrix} dS_t^1 \\ dS_t^2 \end{pmatrix} = \begin{pmatrix} \bar{\mu}^1 S_t^1 \\ \bar{\mu}^2 S_t^2 \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}^{11} S_t^1 & \bar{\sigma}^{12} S_t^1 & \bar{\sigma}^{13} S_t^1 \\ \bar{\sigma}^{21} S_t^2 & \bar{\sigma}^{22} S_t^2 & \bar{\sigma}^{23} S_t^2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \\ dW_t^3 \end{pmatrix}, \tag{82}$$

with $\bar{\sigma}^{11} = \bar{\sigma}^{12} = \hat{\sigma}^1/\sqrt{2}$, $\bar{\sigma}^{13} = 0$, $\bar{\sigma}^{21} = \bar{\sigma}^{22} = \rho\hat{\sigma}^2/\sqrt{2}$ and $\bar{\sigma}^{23} = \sqrt{1-\rho^2}\hat{\sigma}^2$. Notice that the processes are correlated by the parameter ρ. Taking $X_t^j = \ln(S_t^j)$, $j = 1, 2$, we obtain:

$$\begin{pmatrix} dX_t^1 \\ dX_t^2 \end{pmatrix} = \begin{pmatrix} \bar{\mu}^1 - \frac{1}{2}(\hat{\sigma}^1)^2 \\ \bar{\mu}^2 - \frac{1}{2}(\hat{\sigma}^2)^2 \end{pmatrix} dt + \begin{pmatrix} \hat{\sigma}^1/\sqrt{2} & \hat{\sigma}^1/\sqrt{2} & 0 \\ \rho\hat{\sigma}^2/\sqrt{2} & \rho\hat{\sigma}^2/\sqrt{2} & \sqrt{1-\rho^2}\hat{\sigma}^2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \\ dW_t^3 \end{pmatrix}. \tag{83}$$

We consider a derivative contract with payoff $\xi = g(S_T^1, S_T^2) = g(e^{X_T^1}, e^{X_T^2})$. The BSDE for pricing and hedging such derivative contract is given by:

$$\begin{aligned} dY_t &= \left[rY_t + (\bar{\mu} - r\mathbf{1})'(\bar{\sigma}\bar{\sigma}')^{-1}\bar{\sigma}(Z_t^1, Z_t^2, Z_t^3)' \right] dt \\ &\quad + (Z_t^1, Z_t^2, Z_t^3) \begin{pmatrix} dW_t^1 \\ dW_t^2 \\ dW_t^3 \end{pmatrix}, \quad Y_T = g(e^{X_T^1}, e^{X_T^2}), \end{aligned} \tag{84}$$

with

$$\bar{\mu} - r\mathbf{1} = \begin{pmatrix} \bar{\mu}^1 - r \\ \bar{\mu}^2 - r \end{pmatrix}, \quad \bar{\sigma} = \begin{pmatrix} \bar{\sigma}^{11} & \bar{\sigma}^{12} & \bar{\sigma}^{13} \\ \bar{\sigma}^{21} & \bar{\sigma}^{22} & \bar{\sigma}^{23} \end{pmatrix} = \begin{pmatrix} \hat{\sigma}^1/\sqrt{2} & \hat{\sigma}^1/\sqrt{2} & 0 \\ \rho\hat{\sigma}^2/\sqrt{2} & \rho\hat{\sigma}^2/\sqrt{2} & \sqrt{1-\rho^2}\hat{\sigma}^2 \end{pmatrix}. \tag{85}$$

Table 2 Geo with Set II model data

	4	8	16	32	64	128	256
Sch A							
Error Y_0	4.0604×10^{-1}	2.1279×10^{-1}	1.0505×10^{-1}	5.1794×10^{-2}	2.5674×10^{-2}	1.2777×10^{-2}	6.3737×10^{-3}
Error Z_0^1	5.9984×10^{-1}	2.0644×10^{-1}	4.0377×10^{-2}	3.2775×10^{-2}	6.6842×10^{-2}	8.3260×10^{-2}	9.1321×10^{-2}
Error Z_0^2	6.3359×10^{-1}	2.1806×10^{-1}	4.2648×10^{-2}	3.4619×10^{-2}	7.0603×10^{-2}	8.7941×10^{-2}	9.6447×10^{-2}
Sch B							
Error Y_0	2.1687×10^{-3}	1.7059×10^{-4}	6.8241×10^{-5}	1.7645×10^{-5}	4.2948×10^{-6}	4.4580×10^{-7}	1.3074×10^{-7}
Error Z_0^1	3.5008×10^{-1}	1.5286×10^{-1}	1.1130×10^{-1}	1.0210×10^{-1}	9.9941×10^{-2}	9.9241×10^{-2}	9.8658×10^{-2}
Error Z_0^2	3.6977×10^{-1}	1.6146×10^{-1}	1.1736×10^{-1}	1.0785×10^{-1}	1.0558×10^{-1}	1.0503×10^{-1}	1.0485×10^{-1}

Reference values (see [15]): $Y_0 = 8.8808$, $Z_0^1 = 7.9910$, $Z_0^2 = 8.4407$. We take $N = Q = 256$ and $M \in \{4, 8, 16, 32, 64, 128, 256\}$

Table 3 SO with Set I model data, Arith with Set II model data and COM and POM with Set III model data

Sch B	SO	Arith	COM	POM
Reference value	4.1345×10^{0}	1.0173×10^{1}	5.4879×10^{0}	3.7986×10^{0}
Error	4.7091×10^{-4}	7.1873×10^{-5}	6.8658×10^{-4}	5.3795×10^{-4}

For the reference values see [10, 15] and [17]. We take $M = N = Q = 128$

Table 4 Mean-variance price and quadratic hedging of a Geometric Basket Call Option

Y_0	Z_0^1	Z_0^2	Z_0^3
8.8810	5.5807	5.5807	8.3358

5.2.1 Method Data

The range of integration is taken as in the previous example in the complete market.

We consider the following scheme for the discretization: $\Theta_Y = 0.5$, $\Theta_Z^1 = \Theta_Z^2 = \Theta_Z^3 = 0.5$.

$PIt = 5$ Picard iterations are taken.

We will take the values $N^1 = N^2 = Q^1 = Q^2 = M = 256$.

5.2.2 Model Data

The model parameters that we consider are the corresponding to the Set II in Table 1.

5.2.3 Payoff Functions

We apply the methodology for the pricing and hedging of a Geometric Basket Call Option whose payoff was previously indicated.

5.2.4 Results

Table 4 shows the results of the mean-variance price and the quadratic hedging of a Geometric Basket Call Option. As we expected, the mean-variance price equals to the price in the case of a complete market and the hedging is the same in the first two sources of uncertainty and equal to the complete case in the third one. We can conclude that the extension of the BCOS method works also in incomplete markets.

5.3 Jumps Case

Following the numerical example presented in Sect. 3.2, we consider two assets whose prices S^1 and S^2 are driven by

$$\begin{pmatrix} dS_t^1 \\ dS_t^2 \end{pmatrix} = \begin{pmatrix} b^1 S_t^1 \\ b^2 S_t^2 \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}^{11} S_t^1 & \bar{\sigma}^{12} S_t^1 \\ \bar{\sigma}^{21} S_t^2 & \bar{\sigma}^{22} S_t^2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix} + \begin{pmatrix} \bar{\varepsilon}^{11} S_t^1 & \bar{\varepsilon}^{12} S_t^1 \\ \bar{\varepsilon}^{21} S_t^2 & \bar{\varepsilon}^{22} S_t^2 \end{pmatrix} \begin{pmatrix} \sum_{l=1}^{2} (e^{\gamma_l^1} - 1) \hat{N}^1(\{\gamma_l^1\}, dt) \\ \sum_{l=1}^{2} (e^{\gamma_l^2} - 1) \hat{N}^2(\{\gamma_l^2\}, dt) \end{pmatrix}, \tag{86}$$

with $\bar{\sigma}^{11} = \hat{\sigma}$, $\bar{\sigma}^{12} = 0$, $\bar{\sigma}^{21} = \rho\hat{\sigma}^2$, $\bar{\sigma}^{22} = \sqrt{1-\rho^2}\hat{\sigma}^2$, $\varepsilon^{11} = \varepsilon^{22} = 1$ and $\varepsilon^{12} = \varepsilon^{21} = 0$. Notice that the processes are correlated by the parameter ρ. If we denote $\bar{\mu}^j = b^j - \sum_{l=1}^{2} (e^{\gamma_l^j} - 1)\lambda^j P_l^j, j = 1, 2$, we have:

$$\begin{pmatrix} dS_t^1 \\ dS_t^2 \end{pmatrix} = \begin{pmatrix} \bar{\mu}^1 S_t^1 \\ \bar{\mu}^2 S_t^2 \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}^{11} S_t^1 & \bar{\sigma}^{12} S_t^1 \\ \bar{\sigma}^{21} S_t^2 & \bar{\sigma}^{22} S_t^2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix} + \begin{pmatrix} \bar{\varepsilon}^{11} S_t^1 & \bar{\varepsilon}^{12} S_t^1 \\ \bar{\varepsilon}^{21} S_t^2 & \bar{\varepsilon}^{22} S_t^2 \end{pmatrix} \begin{pmatrix} \sum_{l=1}^{2} (e^{\gamma_l^1} - 1) N^1(\{\gamma_l^1\}, dt) \\ \sum_{l=1}^{2} (e^{\gamma_l^2} - 1) N^2(\{\gamma_l^2\}, dt) \end{pmatrix}. \tag{87}$$

Taking $X_t^j = \ln(S_t^j), j = 1, 2$, we obtain:

$$\begin{pmatrix} dX_t^1 \\ dX_t^2 \end{pmatrix} = \begin{pmatrix} \bar{\mu}^1 - \frac{1}{2}(\hat{\sigma}^1)^2 \\ \bar{\mu}^2 - \frac{1}{2}(\hat{\sigma}^2)^2 \end{pmatrix} dt + \begin{pmatrix} \hat{\sigma}^1 & 0 \\ \rho\hat{\sigma}^2 & \sqrt{1-\rho^2}\hat{\sigma}^2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix} + \begin{pmatrix} \bar{\varepsilon}^{11} & \bar{\varepsilon}^{12} \\ \bar{\varepsilon}^{21} & \bar{\varepsilon}^{22} \end{pmatrix} \begin{pmatrix} \sum_{l=1}^{2} \gamma_l^1 N^1(\{\gamma_l^1\}, dt) \\ \sum_{l=1}^{2} \gamma_l^2 N^2(\{\gamma_l^2\}, dt) \end{pmatrix}. \tag{88}$$

We consider a derivative contract with payoff $\xi = g(S_T^1, S_T^2) = g(e^{X_T^1}, e^{X_T^2})$. The BSDE for pricing and hedging such derivative contract is given by (23) with $Y_T = g(e^{X_T^1}, e^{X_T^2})$.

5.3.1 Method Data

According to [15], we take $a^1 = a^2 = a$ and $b^1 = b^2 = b$ for the range of integration in (75) and further, where

$$a := \min_{j=1,2}\left[X_0^j + \mathscr{K}_1^j - L\sqrt{\mathscr{K}_2^j + \sqrt{\mathscr{K}_4^j}}\right],$$
$$b := \max_{j=1,2}\left[X_0^j + \mathscr{K}_1^j + L\sqrt{\mathscr{K}_2^j + \sqrt{\mathscr{K}_4^j}}\right], \quad L = 12, \tag{89}$$

being $\mathscr{K}_i^j$ the ith cumulant of the variable X^j:

$$\mathscr{K}_1^j = [\bar{\mu}^j - \frac{1}{2}(\hat{\sigma}^j)^2 + \lambda^j \mu^{\Gamma^j}]T \tag{90}$$
$$\mathscr{K}_2^j = [(\hat{\sigma}^j)^2 + \lambda^j(\mu^{\Gamma^j})^2 + \lambda^j(\sigma^{\Gamma^j})^2]T, \tag{91}$$
$$\mathscr{K}_4^j = [(\mu^{\Gamma^j})^4 + 6(\mu^{\Gamma^j})^2(\sigma^{\Gamma^j})^2 + 3\lambda^j(\sigma^{\Gamma^j})^4]\lambda^j T, \tag{92}$$

where $\mu^{\Gamma^j} = \sum_{l=1}^{2} \gamma_l^j P_l^j$ and $\sigma^{\Gamma^j} = \left(\sum_{l=1}^{2}(\gamma_l^j)^2 P_l^j - (\mu^{\Gamma^j})^2\right)^{1/2}$.

We consider the following scheme for the discretization: $\Theta_Y = 0.5$, $\Theta_Z^1 = \Theta_Z^2 = \Theta_Z^3 = 0.5$.

$PIt = 5$ Picard iterations are taken.

We will take the values $N^1 = N^2 = Q^1 = Q^2 = 128$.

5.3.2 Model Data

The model parameters that we consider are represented in Table 5 and we take $\rho = 0.25$, $r = 0.04$, $T = 1$ and $K = 100$.

Table 5 Model parameters

j	S_0^j	b^j	$\hat{\sigma}^j$	λ^j	P^j	γ^j
1	90	0.1	0.2	0.0228	$(-0.1338, -0.9838)$	$(0.5, 0.5)$
2	110	0.1	0.3	0.03	$(-0.1, -0.8)$	$(0.5, 0.5)$

Table 6 Mean-variance price and quadratic hedging of a Geometric Basket Call Option on jump diffusion processes

M	Y_0	Z_0^1	Z_0^2	$U_0^1(\gamma_1^1)$	$U_0^1(\gamma_2^1)$	$U_0^2(\gamma_1^2)$	$U_0^2(\gamma_2^2)$
4	9.2600	7.8108	8.2503	0.6650	0.6650	0.8816	0.8816
8	9.2623	7.9959	8.4455	0.6798	0.6798	0.9011	0.9011
16	9.2622	8.0360	8.4873	0.6828	0.6828	0.9049	0.9049
32	9.2622	8.0453	8.5019	0.6833	0.6833	0.9061	0.9061
64	9.2621	8.0484	8.5136	0.6832	0.6832	0.9072	0.9072
128	9.2621	8.0498	8.5209	0.6831	0.6831	0.9079	0.9079

We take $M \in \{4, 8, 16, 32, 64, 128\}$

5.3.3 Payoff Functions

We apply the methodology for the pricing and hedging of a Geometric Basket Call Option whose payoff was previously indicated.

5.3.4 Results

Table 6 shows the results of the mean-variance price and the quadratic hedging of a Geometric Basket Call Option on jump diffusion processes.

6 Conclusions

The technique developed was applied to price and hedge derivatives contracts on two risky assets. We obtained results in the no jumps case (for complete and incomplete markets) and in the jumps case. We can conclude that the extension of the BCOS method works to solve the BSDEs that arise in the different cases.

Acknowledgements This work has been funded by the ITN Research Project STRIKE. The ITN Research Project STRIKE is supported by the European Union in the FP7-PEOPLE-2012-ITN Program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN STRIKE—Novel Methods in Computational Finance).

References

1. Barles, G., Buckdahn, R., Pardoux, E.: Backward stochastic differential equations and integral-partial differential equations. Stoch. Int. J. Probab. Stoch. Process. **60**(1–2), 57–83 (1997)
2. Bouchard, B., Ellie, R.: Discrete-time approximation of decoupled forward-backward SDE with jumps. Stoch. Process. Appl. **118**(1), 53–75 (2008)

3. Delong, L.: Backward Stochastic Differential Equations with Jumps and Their Actuarial and Financial Applications. Springer, London (2013)
4. El Karoui, N., Peng, S., Quenez, M.C.: Backward stochastic differential equations in finance. Math. Financ. **7**(1), 1–71 (1997)
5. Fang, F., Oosterlee, C.W.: A novel pricing method for European options based on Fourier-cosine series expansions. SIAM J. Sci. Comput. **31**(2), 826–848 (2008)
6. Fang, F., Oosterlee, C.W.: Pricing early-exercise and discrete barrier options by Fourier-cosine series expansions. Numer. Math. **114**(1), 27–62 (2009)
7. Gobet, E., Lemor, J.P., Warin, X.: A regression-based Monte Carlo method to solve backward stochastic differential equations. Ann. Appl. Probab. **15**(3), 2172–2202 (2005)
8. Lim, A.E.B.: Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market. Math. Oper. Res. **29**(1), 132–161 (2004)
9. Lim, A.E.B.: Mean-variance hedging when there are jumps. SIAM J. Sci. Comput. **44**(5), 1893–1922 (2005)
10. Margrabe, W.: The value of an option to exchange one asset for another. J. Financ. **33**(1),177–186 (1997)
11. Pardoux, E., Peng, S.: Adapted solution of a backward stochastic differential equation. Syst. Control Lett. **14**, 55–61 (1990)
12. Pardoux, E., Peng, S.: Backward stochastic differential equations and quasilinear parabolic partial differential equations. In: Stochastic Partial Differential Equations and Their Applications. Lectures Notes in Control and Information Sciences, vol. 176, pp. 200–217. Springer, Berlin (1992)
13. Pellegrino, T., Sabino, P.: Pricing and hedging multi-asset spread options by a three-dimensional Fourier cosine series expansion method. Available at SSRN: http://ssrn.com/abstract=2410176orhttp://dx.doi.org/10.2139/ssrn.2410176 (2014)
14. Rong, S.: On solutions of backward stochastic differential equations with jumps. Stoch. Process. Appl. **66**, 209–236 (1997)
15. Ruijter, M.J., Oosterlee, C.W.: Two-dimensional Fourier cosine series expansion method for pricing financial options. SIAM J. Sci. Comput. **34**(5), B642–B671 (2012)
16. Ruijter, M.J., Oosterlee, C.W.: A Fourier-cosine method for an efficient computation of solutions to BSDEs. Available at SSRN: http://ssrn.com/abstract=2233823 or http://dx.doi.org/10.2139/ssrn.2233823 (2013)
17. Stulz, R.M.: Options on the minimum or the maximum of two risky assets: analysis and applications. J. Financ. Econ. **10**(2), 161–185 (1982)

Fichera Theory and Its Application in Finance

Zuzana Bučková, Matthias Ehrhardt, and Michael Günther

Abstract The Fichera theory was first proposed in 1960 by Gaetano Fichera and later developed by Olejnik and Radkevič in 1973. It turned out to be very useful for establishing the well-posedness of initial boundary value problems for parabolic partial differential equations degenerating to hyperbolic ones at the boundary.

In this paper we outline the application of the Fichera theory to interest rates models of Cox-Ingersoll-Ross (CIR) and Chan-Karolyi-Longstaff-Sanders (CKLS) type. For the one-factor CIR model the obtained results are consistent with the corresponding Feller condition.

Keywords Computational finance • Interest rate model

1 Introduction

The *Fichera theory* focus on the question of appropriate *boundary conditions* (BCs) for parabolic partial differential equations (PDEs) degenerating at the boundary. According to the sign of the *Fichera function* one can separate the outflow or inflow part of the solution at the boundary. Thus, this classical theory indicates whether one has to supply a BC at the degenerating boundary.

In this paper we illustrate the application of the Fichera theory to the Cox-Ingersoll-Ross (CIR) interest rate model and its generalisation, the *Chan-Karolyi-Longstaff-Sanders* (CKLS) model [2]. Here, at the left boundary the interest rate tends to zero and thus the parabolic PDE degenerates to a hyperbolic one. For further applications of Fichera theory to other current models in financial mathematics we refer the interested reader to [4].

Z. Bučková (✉) • M. Ehrhardt • M. Günther
Lehrstuhl für Angewandte Mathematik und Numerische Analysis, Fachbereich C Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany
e-mail: buckova@math.uni-wuppertal.de; ehrhardt@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_13

2 The Boundary Value Problem for the Elliptic PDE

We consider an elliptic second order linear differential operator

$$Lu = a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + b_i \frac{\partial u}{\partial x_i} + cu, \qquad x \in \Omega \subset \mathbb{R}^n, \tag{1}$$

where $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is symmetric and induces a semi-definite quadratic form $\xi^\top A\xi \geq 0$ for all $\xi \in \mathbb{R}^n$. Σ denotes a piecewise smooth boundary of the domain Ω. The subset of Σ where the quadratic form vanishes, $\xi^\top A\xi = 0$, will be denoted as Σ_h (hyperbolic part) and the set of points of Σ where the quadratic form remains positive, $\xi^\top A\xi > 0$, is denoted as a Σ_p (parabolic) part. For Σ_h, the hyperbolic part of the boundary Σ_h, we introduce the *Fichera function*

$$b = \sum_{i=1}^{n}\left(b_i - \sum_{k=1}^{n}\frac{\partial a_{ik}}{\partial x_k}\right)\nu_i, \tag{2}$$

where ν_i is the direction cosine of the inner normal to Σ, i.e. it is $\nu_i = \cos(x_i, \mathbf{n_i})$, where $\mathbf{n_i}$ is the inward normal vector at the boundary.

On the *hyperbolic part* of the boundary Σ_h we define according to the sign of the Fichera function the three subsets Σ_0 ($b = 0$ tangential flow), Σ_+ ($b > 0$, outflow) and Σ_- ($b < 0$, inflow), i.e. the boundary $\Sigma = \Sigma_p \cup \Sigma_h$ can be written as a unification of four boundary parts: $\Sigma = \Sigma_p \cup \Sigma_0 \cup \Sigma_+ \cup \Sigma_-$.

Olejnik and Radkevič [7, Lemma 1.1.1] showed that the sign of the Fichera function b at the single points Σ_h does not change under smooth nondegenerate changes of independent variables in a given elliptic operator (1). In [7, Theorem 1.1.1] it is stated that the subsets Σ_0, Σ_+, Σ_- remain invariant under a smooth nonsingular changes of independent variables in the elliptic operator (1).

The *parabolic boundary* Σ_p can be rewritten as a unification of two sets Σ_p^D (Dirichlet BC) and Σ_p^N (Neumann BC). Let us state one simple example.

Example 1 ([6]) The boundary value problem for an elliptic PDE reads

$$\begin{aligned} Lu &= f && \text{on} \quad \Omega \subset \mathbb{R}^n, \\ u &= g && \text{on} \quad \Sigma_- \cup \Sigma_p^D \\ a_{ij}\frac{\partial u}{\partial x_i}n_j &= h && \text{on} \quad \Sigma_p^N \end{aligned}$$

If Σ_p^N is an empty set, we obtain a Dirichlet problem; if Σ_p^D is an empty set, a Neumann problem; if Σ_p^D and Σ_p^N are not empty, the problem is of mixed Dirichlet-Neumann type. Recall that for hyperbolic PDEs one must not supply BCs for outflow boundaries (Σ_+) or boundaries where the characteristics are tangential to the boundary (Σ_0), since this may violate the information that is transported from the interior of the domain.

3 Application to One-Factor Interest Rate Models of CKLS Type

We start with an interest rate model in the form of a stochastic differential equation

$$dr = \kappa(\theta - r)\,dt + \sigma r^{\gamma}\,dW, \tag{3}$$

where κ, θ are positive constants, and γ non-negative. This CKLS model [2] is a mean-reversion process with non-constant volatility σr^{γ}. Using the Itô formula for a duplicating portfolio in a risk neutral world one can derive a PDE for the zero-coupon bond price $P(r, \tau)$:

$$\frac{\partial P}{\partial \tau} = \alpha(r,\tau)\frac{\partial^2 P}{\partial r^2} + \beta(r,\tau)\frac{\partial P}{\partial r} - rP, \qquad r > 0,\ \tau > 0, \tag{4}$$

where $\alpha(r,\tau) = \frac{1}{2}\sigma^2 r^{2\gamma}$, $\beta(r,\tau) = \kappa(\theta - r)$. A closed form formula for this model can be given in special cases, cf. [1]:

(a) if $\gamma = 0$, this is the classical Vašíček model with constant volatility.
(b) for $\gamma = 0.5$, we get the Cox-Ingersoll-Ross (CIR) model (CIR), [3].

For general γ (CKLS model) there is no closed form formula for the bond price $P(r, \tau)$ and the PDE (4) has to be solved numerically.

The volatility term in (4), for a short rate r tending to zero, is $\alpha(0, \tau) = 0$. Thus the parabolic PDE (4) reduces at $r = 0$ to the hyperbolic PDE

$$\frac{\partial P}{\partial \tau} = \kappa\theta\frac{\partial P}{\partial r}, \qquad \tau > 0. \tag{5}$$

Next, the Fichera function (2) for our model reads

$$b(r) = \beta(r,\tau) - \frac{\partial \alpha(r,\tau)}{\partial r}, \tag{6}$$

and we check the sign of (6) for $r \to 0+$:

- if $\lim\limits_{r\to 0^+} b(r) \geq 0$ (outflow boundary) we must not supply any BCs at $r = 0$.
- if $\lim\limits_{r\to 0^+} b(r) < 0$ (inflow boundary) we have to define BCs at $r = 0$.

Especially for the proposed model we get $b(r) = \kappa(\theta - r) - \sigma^2\gamma r^{2\gamma-1}$ and we can distinguish the following situations:

(a) for $\gamma = 0.5$ (CIR model) $\Rightarrow$ if $\kappa\theta - \sigma^2/2 \geq 0$, we do not need any BCs.
(b) for $\gamma > 0.5 \Rightarrow$ if $\kappa\theta \geq 0$, we do not need any BCs.
(c) for $\gamma \in (0, 0.5) \Rightarrow$ if $\lim\limits_{r\to 0^+} b(r) = -\infty$, we need BCs.

Remark 1 (Feller Condition) The Feller condition guaranteeing a positive interest rate defined by (3) for the one-factor CIR model is $2\kappa\theta > \sigma^2$ and is equivalent with the condition derived from the Fichera theory. If the Feller condition holds, then the Fichera theory states that one must not supply any BC at $r = 0$.

4 A Two-Factor Interest Rate Model

We consider a general two-factor model given by the set of two SDEs

$$\mathrm{d}x_1 = (a_1 + a_2x_1 + a_3x_2)\,\mathrm{d}t + \sigma_1 x_1^{\gamma_1}\,\mathrm{d}W_1, \tag{7}$$

$$\mathrm{d}x_2 = (b_1 + b_2x_1 + b_3x_2)\,\mathrm{d}t + \sigma_2 x_2^{\gamma_2}\,\mathrm{d}W_2, \tag{8}$$

$$Cov[\mathrm{d}W_1, \mathrm{d}W_2] = \rho\,\mathrm{d}t, \tag{9}$$

containing as special cases the Vašíček model ($\gamma_1 = \gamma_2 = 0$) and the CIR model ($\gamma_1 = \gamma_2 = 0.5$). The drift functions are defined as linear functions of the two variables x_1 and x_2. Choosing $a_1 = b_1 = b_2 = 0$ we get two-factor convergence model of CKLS type (in case of general $\gamma_1, \gamma_2 \geq 0$). The variable x_1 models the interest rate of a small country (e.g. Slovakia) before entering the monetary EURO union and the variable x_2 represents the interest rate of the union of the countries (such as the EU).

Applying the standard Itô formula one can easily derive a parabolic PDE

$$\frac{\partial P}{\partial \tau} = \tilde{a}_{11}\frac{\partial^2 P}{\partial x_1^2} + \tilde{a}_{22}\frac{\partial^2 P}{\partial x_2^2} + \tilde{a}_{12}\frac{\partial^2 P}{\partial x_1 \partial x_2} + \tilde{a}_{21}\frac{\partial^2 P}{\partial x_2 \partial x_1} + \tilde{b}_1\frac{\partial P}{\partial x_1} + \tilde{b}_2\frac{\partial P}{\partial x_2} + \tilde{c}P, \tag{10}$$

where $P(x, y, \tau)$ represents the bond price at time τ for interest rates x and y, and

$$\tilde{a}_{11} = \frac{\sigma_1^2 x_1^{2\gamma_1}}{2}, \qquad \tilde{a}_{22} = \frac{\sigma_2^2 x_2^{2\gamma_2}}{2}, \qquad \tilde{a}_{12} = \tilde{a}_{21} = \frac{1}{2}\rho\sigma_1 x_1^{\gamma_1}\sigma_2 x_2^{\gamma_2}$$

$$\tilde{b}_1 = a_1 + a_2x_1 + a_3x_2, \qquad \tilde{b}_2 = b_1 + b_2x_1 + b_3x_2, \qquad \tilde{c} = -x_1,$$

for $x_1, x_2 \geq 0$, $\tau \in (0, T)$, with initial condition $P(x_1, x_2, 0) = 1$ for $x_1, x_2 \neq 0$.

Now, the Fichera function (2) in general reads

$$b(x_1, x_2) = \left[a_1 + a_2x_1 + a_3x_2 - \left(\sigma_1^2\gamma_1 x_1^{2\gamma_1 - 1} + \frac{1}{2}\rho\sigma_1 x_1^{\gamma_1}\sigma_2\gamma_2 x_2^{\gamma_2 - 1}\right)\right]\frac{x_1}{\sqrt{1 + x_1^2}}$$

$$+ \left[b_1 + b_2x_1 + b_3x_2 - \left(\frac{1}{2}\rho\sigma_1\gamma_1 x_1^{\gamma_1 - 1}\sigma_2 x_2^{\gamma_2} + \sigma_2^2\gamma_2 x_2^{2\gamma_2 - 1}\right)\right]\frac{x_2}{\sqrt{1 + x_2^2}}.$$

Depending on γ_1 and γ_2, we get the following results:

- For $\gamma_1 = \gamma_2 = 0$ (classical Vašíček model), the Fichera function simplifies to

$$b(x_1, x_2) = (a_1 + b_1) + (a_2 + b_2)x_1 + (a_3 + b_3)x_2,$$

and boundary conditions must be supplied, if

$$\begin{cases} x_1 \leq -\frac{a_1+b_1+(a_3+b_3)x_2}{a_2+b_2} & \text{for} \quad a_2 + b_2 \neq 0 \\ x_2 \leq -\frac{a_1+b_1}{a_3+b_3} & \text{for} \quad a_2 + b_2 = 0, a_3 + b_3 \neq 0 \\ a_1 + b_1 \leq 0 & \text{for} \quad a_2 + b_2 = 0, a_3 + b_3 = 0 \end{cases}.$$

- For $\gamma_1 = \gamma_2 = 0.5$ (CIR model), the Fichera function simplifies to

$$b(x_1, x_2) = \left[a_1 + a_2x_1 + a_3x_2 - \left(\sigma_1^2\gamma_1 + \frac{1}{4}\rho\sigma_1\sigma_2\sqrt{\frac{x_1}{x_2}}\right)\right]\frac{x_1}{\sqrt{1+x_1^2}} + \left[b_1 + b_2x_1 + b_3x_2 - \left(\frac{1}{4}\rho\sigma_1\sigma_2\sqrt{\frac{x_2}{x_1}} + \sigma_2^2\gamma_2\right)\right]\frac{x_2}{\sqrt{1+x_2^2}}$$

We must supply boundary conditions for $\rho > 0$, and must not for $\rho < 0$. For $\rho = 0$, BCs at $x_2 = 0$ must be posed if $x_1 \leq \sigma_1^2\gamma_1/(2a_2) - a_1/a_2$ (assuming $a_2 > 0$, and for $x_1 = 0$, if $x_2 \leq \sigma_2^2\gamma_2/(2b_2) - b_1/b_2$ (assuming $b_2 > 0$), otherwise not.

- For the general case $\gamma_1, \gamma_2 > 0$, we discuss the boundary $x_2 = 0, x_1 > 0$; due to symmetry, the case $x_2 = 0, x_1 > 0$ follows then by changing the roles of x_1 and x_2, as well as γ_1 and γ_2. For $x_2 = 0$ the Fichera function simplifies to

$$\lim_{x_2 \to 0^+} b(x_1, x_2) = \left[a_1 + a_2x_1 - \sigma_1^2\gamma_1 x_1^{2\gamma_1 - 1} - \frac{1}{2}\rho\sigma_1 x_1^{\gamma_1}\sigma_2\gamma_2 0^{\gamma_2 - 1}\right]\frac{x_1}{\sqrt{1+x_1^2}}$$

$$= \begin{cases} \left[a_1 + a_2x_1 - \sigma_1^2\gamma_1 x_1^{2\gamma_1-1}\right]\frac{x_1}{\sqrt{1+x_1^2}} & \rho = 0 \\ -\infty & 0 < \gamma_2 < 1, \rho \neq 0 \\ \left[a_1 + a_2x_1 - \sigma_1^2\gamma_1 x_1^{2\gamma_1-1} - \frac{1}{2}\rho\sigma_1 x_1^{\gamma_1}\sigma_2\right]\frac{x_1}{\sqrt{1+x_1^2}} & \gamma_2 = 1, \rho \neq 0 \\ \left[a_1 + a_2x_1 - \sigma_1^2\gamma_1 x_1^{2\gamma_1-1}\right]\frac{x_1}{\sqrt{1+x_1^2}} & \gamma_2 > 1, \rho \neq 0 \end{cases}$$

For $0 < \gamma_2 < 1$ and $\rho \neq 0$, BCs are needed, if ρ is positive, and BCs must not be posed, if ρ is negative. In all other cases, the sign of b, which defines whether BCs must be supplied or not, depends on a_1, a_2, σ_1, σ_2 and γ_1, see Fig. 1.

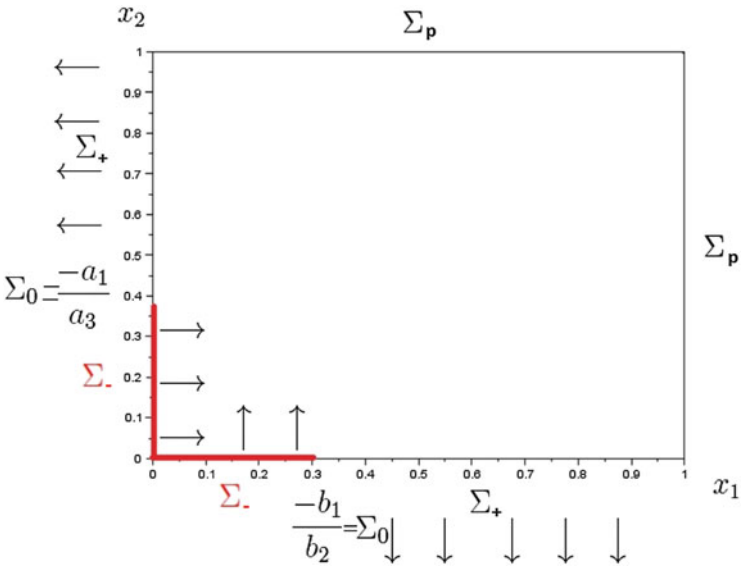


Fig. 1 Boundary decomposition in two-factor CIR model

5 Numerical Results

Choosing set of parameters $\kappa = 0.5$, $\theta = 0.05$, $\sigma = 0.1$, $\gamma = 0.5$ (CIR), we get at $r = 0$ a positive Fichera function $b = \kappa\theta - \sigma^2/2 = 0.02 > 0$. This is equivalent with the statement that the Feller condition is satisfied. According to the Fichera theory, as soon as it is outflow part of boundary, we must not supply BCs. In this example in Figs. 2 and 3 and Table 1, we intentionally supplied BCs in an 'outflow' situation when we should not in order to illustrate what might happen if one disregards the Fichera theory. In the evolution of the solution we can observe a peak and oscillations close to the boundary. In Fig. 3 we plot with the relative error, which is reported also in Table 1.

In our example we used the same parameters, but with or without defining Dirichlet BC. Here, "without BC" means that we used for the numerical BC the limit of the interior PDE for $r \to 0$. The corresponding results are shown in Figs. 4 and 5 and the relative errors are recorded in Table 2.

For the numerical solution we used the implicit finite difference method from [5]. The reference solution is obtained either as the analytic solution for the CIR model ($\gamma = 0.5$, if Feller condition is satisfied), cf. [1] or in all other cases using a very fine resolution (and suitable BCs). The conditions at outflow boundaries are obtained by studying the limiting behaviour of the interior PDE or simply by horizontal extrapolation of appropriate order. Recall that negative values of the Fichera function (i.e. an inflow boundary) corresponds to a not satisfied Feller condition and may destroy the uniqueness of solutions to the PDE.

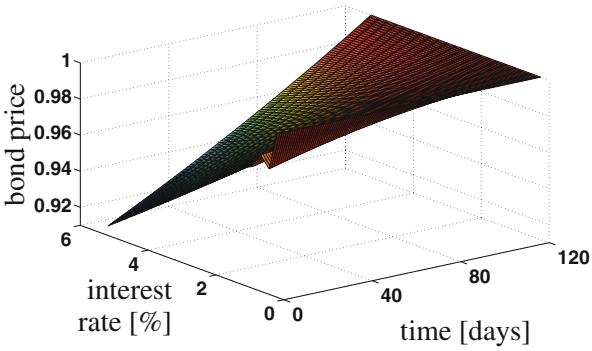


Fig. 2 Numerical solution, Dirichlet BC

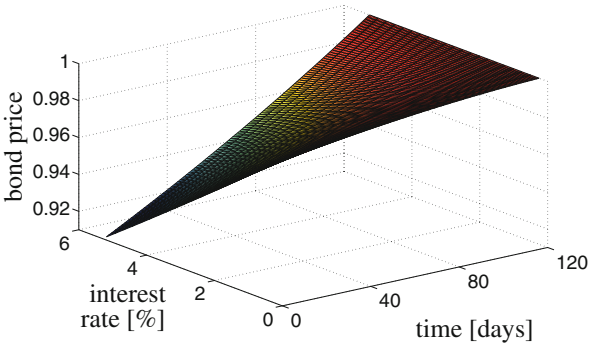


Fig. 3 Relative error, case with Dirichlet BC

Table 1 Relative error, case with BC

Time [days]	Relative error
1	0.0147
40	0.0079
80	0.0029
120 (maturity)	0

6 Conclusion

We discussed one and two factor interest rate models and applied the classical Fichera theory to the resulting degenerate parabolic PDEs. This theory provides highly relevant information how to supply BCs in these applications.

As a next step, we will investigate multi-factor models, which are coupled only via the correlation of the Brownian motion:

$$\mathrm{d}x_i = (a_i + b_i x_i)\mathrm{d}t + \sigma_i x_i^{\gamma_i}\,\mathrm{d}W_i,$$
$$Cov[\mathrm{d}W_i, \mathrm{d}W_j] = \rho_{ij}\,\mathrm{d}t, \qquad i,j = 1,\ldots,n.$$

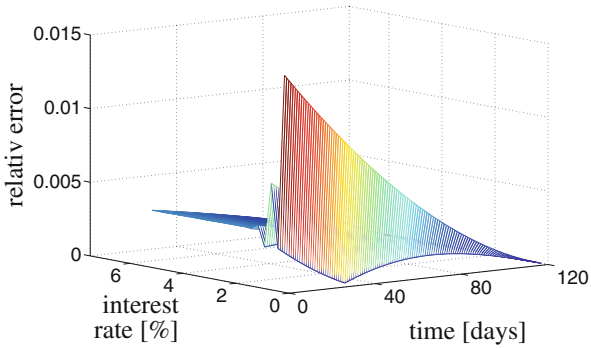


Fig. 4 Numerical solution, without BC

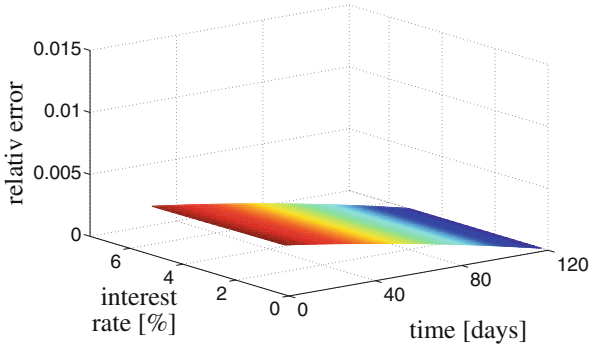


Fig. 5 Relative error, case without BC

Table 2 Relative error, case without BC

Time [days]	Relative error
1	0.0039
40	0.0029
80	0.0015
120 (maturity)	0

Acknowledgements The authors were partially supported by the European Union in the FP7-PEOPLE-2012-ITN Program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN *STRIKE—Novel Methods in Computational Finance*) and by the bilateral German-Slovakian Project *NL-BS—Numerical Solution of Nonlinear Black-Scholes Equations*, financed by the DAAD (01/2013-12/2014).

References

1. Brigo, D., Mercurio, F.: Interest Rate Models – Theory and Practice, With Smile, Inflation and Credit, 2nd edn. Springer Finance. Springer, Berlin (2006)
2. Chan, K.C., Karolyi, G.A., Longstaff, F., Sanders, A.: The volatility of short-term interest rates: an empirical comparison of alternative models of the term structures of interest rates. J. Financ. **47**, 1209–1227 (1992)

3. Cox, J.C., Ingersoll, J.E., Ross, S.A.: A theory of the term structure of interest rates. Econometrica **53**, 385–408 (1985)
4. Duffy, D.J.: Unconditionally stable and second order accurate explicit finite difference schemes using domain transformation: Part I One-factor equity problems. SSRN Electron. J. (2009). doi:10.2139/ssrn.1552926
5. Ekstroem, E., Loetstedt, P., Tysk, J.: Boundary values and finite difference methods for the single factor term structure equation. Appl. Math. Financ. **16**, 253–259 (2009)
6. Fichera, G.: On a Unified theory of boundary value problems for elliptic-parabolic equations of second order in boundary value problems. In: Langer, R.E. (eds.) Boundary Problems. Differential Equations. University of Wisconsin Press, Madison Press (1960)
7. Olejnik, O.A., Radkevič, E.V.: Second Order Equations with Nonnegative Characteristic Form. American Mathematical Society, Providence, RI (1973)

Modelling Stochastic Correlation

Long Teng, Matthias Ehrhardt, and Michael Günther

Abstract It is well known that the correlation between financial products, financial institutions, e.g., plays an essential role in pricing and evaluation of derivatives. Using a constant or deterministic correlation may lead to correlation risk, since market observations give evidence that the correlation is hardly a deterministic quantity.

Here, the approach of Teng et al. (A versatile approach for stochastic correlation using hyperbolic functions. Preprint 13/14. University of Wuppertal, 2013) for modelling the correlation as a hyperbolic function of a stochastic process is generalized to derive stochastic correlation processes (SCP) from a hyperbolic transformation of the modified Ornstein-Uhlenbeck process. We determine a transition density function of this SCP in closed form which could be used easily to calibrate SCP models to historical data.

As an example we compute the price of a quantity adjusting option (Quanto) and discuss concisely the effect of considering stochastic correlation on pricing the Quanto.

Keywords Computational finance • Option pricing • Stochastic correlation process

1 General Instructions

Correlation is an established concept for quantifying the relationship between assets. It plays an essential role in several financial applications, e.g. in portfolio credit models, the default correlation is one fundamental factor of risk evaluation [1, 9].

L. Teng (✉) • M. Ehrhardt • M. Günther
Lehrstuhl für Angewandte Mathematik und Numerische Analysis, Fachbereich C – Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany
e-mail: teng@math.uni-wuppertal.de; ehrhardt@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_14

For two random variables X_1, X_2 with finite *variances*, the *correlation* of them is

$$\rho_{1,2} = \mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1 \sigma_2}, \tag{1}$$

with the *covariance*

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}\big[(X_1 - \mu_1)(X_2 - \mu_2)\big], \tag{2}$$

where μ_i and σ_i are the *expectation* and *standard deviation* of X_i, $i = 1, 2$. Here $\rho_{1,2}$ denotes a coefficient number in the interval $[-1, 1]$. The boundaries -1 and 1 will be reached if and only if X_1 and X_2 are indeed linearly related. The greater the absolute value of $\rho_{1,2}$ the stronger the dependence between X_1 and X_2 is.

The *observability* is a problem of using a correlation concept in financial markets. Unlike price, exchange rate, etc., the correlation cannot be observed directly in the market and can only be measured in the context of a model. The easiest estimator of the correlation is the sample correlation coefficient. Given a series of N measurements of the observable quantities X_1, X_2, and denoting the measurements by $x_{1,j}, x_{2,j}, j = 1, 2, \ldots, N$, the *sample coefficient correlation* reads

$$\hat{\rho}_{12} = \frac{\sum_{j=1}^{N} (x_{1,j} - \bar{\mu}_1)(x_{2,j} - \bar{\mu}_2)}{\sqrt{\sum_{j=1}^{N} (x_{1,j} - \bar{\mu}_1)^2 \sum_{j=1}^{N} (x_{2,j} - \bar{\mu}_2)^2}}, \tag{3}$$

where $\bar{\mu}_1$ and $\bar{\mu}_2$ are the *sample means* of X_1 and X_2.

In financial models, stochastic processes are used quite often to model data series, like price, interest rate and exchange rate. The dependence between the series is given by the *correlated Brownian motions* W_1 and W_2 (in symbolic notion)

$$dW_{1,t}\, dW_{2,t} = \rho_{1,2}\, dt. \tag{4}$$

However, the dependence can be hardly modelled by a fixed constant, i.e. the constant correlation may not be an appropriate measure of co-dependence. Using constant ("wrong") correlation may result some 'correlation risk'. There exist already some works which show that the correlation should not be constant and even changes over a small time interval as the volatility, see e.g. [7]. Several approaches generalize the constant correlation to a time-varying and stochastic concept, like the *Dynamic Conditional Correlation model* [2], the *Local correlation models* see e.g. [4] and the *Wishart autoregressive process* proposed by Gourieroux [3] that guarantees the positive definiteness of the variance-covariance matrix. Furthermore, modelling correlation as stochastic process was proposed, see [5, 10] and [8].

2 Stochastic Correlation with a Modified Ornstein-Uhlenbeck Process

In this section, we specify a SCP by a hyperbolic transformation of the modified Ornstein-Uhlenbeck process. The derivation of the transition density function of this SCP is provided in a closed form which could be used easily for calibration.

2.1 The Transformed Modified Ornstein-Uhlenbeck Process

The *Ornstein-Uhlenbeck process* is defined by the SDE

$$dX_t = \kappa(\mu - X_t)\,dt + \sigma\,dW_t, \tag{5}$$

where $\kappa, \sigma > 0$ and $X_0, \mu \in \mathbb{R}$. If we want to restrict the mean value μ to be only in $(-1, 1)$, it is reasonable to modify the Ornstein-Uhlenbeck process (5) as

$$dX_t = \kappa\big(\mu - \tanh(X_t)\big)\,dt + \sigma\,dW_t, \tag{6}$$

where $\kappa, \sigma > 0$ and $X_0, \mu \in (-1, 1)$.

Lemma 1 *Applying Itô's Lemma with* $\rho_t = \tanh(X_t)$,

$$d\rho_t = \frac{\partial \tanh(X_t)}{\partial x}\,dX_t + \frac{1}{2}\frac{\partial^2 \tanh(X_t)}{\partial x^2}\sigma^2\,dt \tag{7}$$

gives the stochastic correlation process as

$$d\rho_t = (1-\rho_t^2)\big(\kappa(\mu-\rho_t) - \sigma^2\rho_t\big)\,dt + (1-\rho_t^2)\sigma\,dW_t, \tag{8}$$

where $t \geq 0$, $\rho_0 \in (-1, 1)$, $\kappa, \sigma > 0$ *and* $\mu \in (-1, 1)$.

Proof

$$\begin{aligned}(7) &= \operatorname{sech}^2(X_t)\kappa\big(\mu - \tanh(X_t)\big)dt - \operatorname{sech}^2(X_t)\frac{\sinh(X_t)}{\cosh(X_t)}\sigma^2 dt + \operatorname{sech}^2(X_t)\sigma dW_t \\ &= (1-\rho_t^2)\kappa(\mu-\rho_t)dt - (1-\rho_t^2)\rho_t\sigma^2 dt + (1-\rho_t^2)\sigma dW_t\end{aligned}$$

We define $\kappa^* = \kappa + \sigma^2$, $\mu^* = \frac{\kappa\mu}{\kappa+\sigma^2}$, $\sigma^* = \sigma$ and rewrite (8) as

$$\frac{d\rho_t}{1-\rho_t^2} = \kappa^*(1-\mu^*)dt + \sigma^* dW_t, \tag{9}$$

where $t \geq 0$, $\rho_0 \in (-1, 1)$, $\kappa^*, \sigma^* > 0$ and $\mu^* \in (-1, 1)$.

2.2 Transition Density Function

For calibration purposes, we need to determine the *transition density function* of (9) with the aid of the *Fokker-Planck equation* [6]. Then, we obtain the parameters of the correlation process (9) by fitting the density function to the market data.

Let us assume that the SCP (9) possesses a transition density $f(t, \tilde{\rho}|\rho_0)$ which satisfies the following *Fokker-Planck equation*

$$\frac{\partial}{\partial t} f(t,\tilde{\rho}) + \frac{\partial}{\partial \tilde{\rho}}\big(\hat{a}(t,\tilde{\rho}) f(t,\tilde{\rho})\big) - \frac{1}{2}\frac{\partial^2}{\partial \tilde{\rho}^2}\big(\hat{b}(t,\tilde{\rho})^2 f(t,\tilde{\rho})\big) = 0, \tag{10}$$

with

$$\hat{a}(t,\tilde{\rho}) = \kappa^*(1-\mu^*)(1-\tilde{\rho}^2), \qquad \hat{b}(t,\tilde{\rho}) = (1-\tilde{\rho}^2)\sigma^*. \tag{11}$$

For the calibration purpose we consider the stationary density (for $t \to \infty$)

$$f(\tilde{\rho}) := \lim_{t\to\infty} f(t,\tilde{\rho}|\rho_0). \tag{12}$$

With the above construction the SCP (9) is a mean-reverting process. Every two solutions of (10) are equal for $t \to \infty$, i.e. a unique stationary solution $f(\tilde{\rho})$ exists [6].

Next we show how to determine the analytical stationary density function $f(\tilde{\rho})$ of the SCP (9). First, the stationary density function $f(\tilde{\rho})$ obviously satisfies

$$\frac{\partial}{\partial \tilde{\rho}}\Big((1-\tilde{\rho}^2)\big(\kappa^*(1-\mu^*)\big) f(\tilde{\rho})\Big) = \frac{1}{2}\frac{\partial^2}{\partial \tilde{\rho}^2}\big((1-\tilde{\rho}^2)\sigma^*\big)^2 f(\tilde{\rho}). \tag{13}$$

By solving the elliptic equation (13) we obtain the stationary density $f(\tilde{\rho})$ as

$$f(\tilde{\rho}) = \frac{m}{2^{\frac{\kappa^*}{\sigma^*}}}(1+\tilde{\rho})^{\frac{\kappa^*-2\sigma^{*2}}{\sigma^{*2}}+\frac{\kappa^*\mu^*}{\sigma^{*2}}}(1-\tilde{\rho})^{\frac{\kappa^*-2\sigma^{*2}}{\sigma^{*2}}-\frac{\kappa^*\mu^*}{\sigma^{*2}}} + \frac{n}{\tilde{\rho}^2-1}\left(\frac{1}{2}\right)^{\frac{2\sigma^{*2}-\kappa^*}{\sigma^{*2}}} F\left(1, \frac{2(\sigma^{*2}-\kappa^*)}{\sigma^{*2}}, \frac{(-\mu^*-1)\kappa^*+2\sigma^{*2}}{\sigma^{*2}}, \frac{\tilde{\rho}+1}{2}\right) \tag{14}$$

with the constants $m, n \in \mathbb{R}$, the *hypergeometric function* F and the *Pochhammer symbol* $(\cdot)_k$. Now we need to fix the constants m and n in (14) to obtain the stationary density. Due to the mean reversion the stationary density $f(\tilde{\rho})$ must satisfy

$$\int_{-1}^{1} \tilde{\rho} f(\tilde{\rho})\, d\tilde{\rho} = \mu^*. \tag{15}$$

If we choose $\mu^* = 0$, we observe that the first term in (14) becomes

$$\frac{m}{2^{\frac{\kappa^*}{\sigma^{*2}}}}(1+\tilde{\rho})^{\frac{\kappa^*-2\sigma^{*2}}{\sigma^{*2}}}(1-\tilde{\rho})^{\frac{\kappa^*-2\sigma^{*2}}{\sigma^{*2}}}, \tag{16}$$

which is obviously symmetric around $\tilde{\rho} = 0$, i.e. the condition (16) will be fulfilled for $n = 0$. In the sequel we assume that $n \equiv 0$ for all general $\mu^* \in (-1, 1)$ such that the transition density function (14) can be rewritten as

$$f(\tilde{\rho}) = \frac{m}{2^{\frac{\kappa^*}{\sigma^{*2}}}}(1+\tilde{\rho})^{\frac{\kappa^*-2\sigma^{*2}}{\sigma^{*2}}+\frac{\kappa^*\mu^*}{\sigma^{*2}}}(1-\tilde{\rho})^{\frac{\kappa^*-2\sigma^{*2}}{\sigma^{*2}}-\frac{\kappa^*\mu^*}{\sigma^{*2}}}. \tag{17}$$

To determine the value of m we can employ the basic property of a density function

$$\int_{-1}^{1} f(\tilde{\rho})\, d\tilde{\rho} = 1. \tag{18}$$

The constant m in (17) must be chosen such that the normalization condition (18) is always fulfilled. We set $a = \frac{\kappa^*-2\sigma^{*2}}{\sigma^{*2}}$, $b = \frac{\kappa^*\mu^*}{\sigma^{*2}}$, and substitute it into (17) to obtain

$$f(\tilde{\rho}) = \frac{m}{2^{\frac{\kappa^*}{\sigma^{*2}}}}(1+\tilde{\rho})^{a+b}(1-\tilde{\rho})^{a-b}. \tag{19}$$

The fact, as long as the *condition (A)*: $a \pm b > -1$, is fulfilled, the integral

$$\int_{-1}^{1} (1+\tilde{\rho})^{a+b}(1-\tilde{\rho})^{a-b}\, d\tilde{\rho} \tag{20}$$

has the solution

$$M := \frac{\Gamma(1+a-b)F(1,-a-b,2+a-b,-1)}{\Gamma(2+a-b)} + \frac{\Gamma(1+a+b)F(1,-a+b,2+a+b,-1)}{\Gamma(2+a+b)}, \tag{21}$$

with the hypergeometric function F and the *gamma function* Γ. Straightforward calculations show that the condition (A) will always hold due to $\mu \in (-1, 1)$. Thus, the constant m can be determined as $m = 2^{\frac{\kappa^*}{\sigma^{*2}}}/M$. Finally, we obtain the transition density function in a closed form as

$$f(\tilde{\rho}) = \frac{(1+\tilde{\rho})^{a+b}(1-\tilde{\rho})^{a-b}}{M}, \tag{22}$$

with M defined in (21). The parameters κ^*, μ^* and σ^* can be obtained by fitting (22) to the historical correlation, if we assume that the correlation is itself observable.

We could generalize the correlation process (9) with the same definition but directly with the arbitrary parameter coefficients $\kappa > 0$, $\mu \in (-1, 1)$ and $\sigma > 0$, like

$$\frac{d\rho_t}{1-\rho_t^2} = \kappa(1-\mu)dt + \sigma\, dW_t. \tag{23}$$

For this case, we obtain $a = \frac{\kappa-2\sigma^2}{\sigma^2}$, $b = \frac{\kappa\mu}{\sigma^2}$. We perform a similar calculation for checking the condition (A) as above:

$$a+b > -1 \Leftarrow \tfrac{\kappa-2\sigma^2}{\sigma^2} + \tfrac{\kappa\mu}{\sigma^2} > -1 \Leftarrow \kappa(1+\mu) > \sigma^2 \Leftarrow \kappa > \tfrac{\sigma^2}{1+\mu},$$

$$a-b > -1 \Leftarrow \tfrac{\kappa-2\sigma^2}{\sigma^2} - \tfrac{\kappa\mu}{\sigma^2} > -1 \Leftarrow \kappa(1-\mu) > \sigma^2 \Leftarrow \kappa > \tfrac{\sigma^2}{1-\mu}.$$

The process (23) could be employed for the stochastic correlation if the condition $\kappa > \frac{\sigma^2}{1\pm\mu}$ is fulfilled. We find that this condition dovetails nicely with that condition in [10], which ensures that the boundaries -1 and 1 are unattainable.

3 Pricing Quantos with Stochastic Correlation

Quanto options hedge the exchange rate risk when investing in products not valued in the domestic currency. One has to consider the correlation between the currency exchange rate R_t between domestic and foreign currencies, and the price S_t of the

underlying. We assume that S_t and R_t follow the coupled stochastic process by

$$\begin{cases} dS_t &= \mu_S S_t\, dt + \sigma_S S_t\, dW_t^S \\ dR_t &= \mu_R R_t\, dt + \sigma_R R_t\, dW_t^R, \end{cases} \tag{24}$$

where W_t^S and W_t^R are correlated using the SCP (23) as:

$$\frac{d\rho_t}{1-\rho_t^2} = \kappa(1-\mu)\, dt + \sigma\, dW_t. \tag{25}$$

W_t is assumed to be independent of W_t^S and W_t^R. From [8] we know that the price of a Quanto Put-Option in the extended BS model incorporating the SCP reads

$$\begin{aligned} P_{\text{Quanto}} &= P_{\text{Quanto}}(S_0, K, r_e, r_d, D(\rho_t), \sigma_S, \sigma_R, T) \\ &= R_0\big(K \exp^{-r_d T}\, \mathscr{N}(-d_2) - S_0 \exp^{-D(\rho_t)T}\, \mathscr{N}(-d_1)\big) \end{aligned} \tag{26}$$

$$d_1 = \frac{\log(\frac{S_0}{K}) + ((r_d - D(\rho_t)) + \frac{\sigma_S^2}{2})/T}{\sigma_S\sqrt{T}}, \qquad d_2 = d_1 - \sigma_S\sqrt{T}. \tag{27}$$

Using the conditional Monte-Carlo approach the fair price can be approximated by

$$\mathbb{P}_0 = \mathrm{e}\left[\mathrm{e}[P_{\text{Quanto}}(S_0, K, r_d, D(\rho_t), \sigma_S, T)|\mathscr{F}(\rho_{\{0\le s\le t\}})]\right] \approx \frac{\sum_{i=1}^{M} P^i_{\text{Quanto}}}{M} \tag{28}$$

We assume the parameters for the pricing formula (28) up to correlation. Then, we estimate both the constant correlation coefficient and the SCP parameters using the same market data, namely historical data of S&P 500 and Euro/US-Dollar exchange rate (January 2003–March 2013). The estimated constant correlation is 0.025 and SCP parameters are provided in Fig. 1. We let the SCP starting from the first correlation in the historical correlations and compare the Quanto option prices between using constant and stochastic correlation, the plot of price differences is on the right side.

We can observe, whilst the maturity T is shorter than 3 years, the price with constant correlation is lower than the price with stochastic correlation. Then, from nearly $T = 3$, the price calculated with constant correlation becomes higher than the corresponding price calculated with stochastic correlation. The reason for this, before the time $T = 3$, the SCP provides the correlations which are closed to the initial correlation $\rho_0 = 0.3$ which is larger than the constant correlation $\rho = 0.025$.

We conclude that our numerical results give strong evidence that the correlation risk caused by using a wrong (constant) correlation model cannot be neglected.

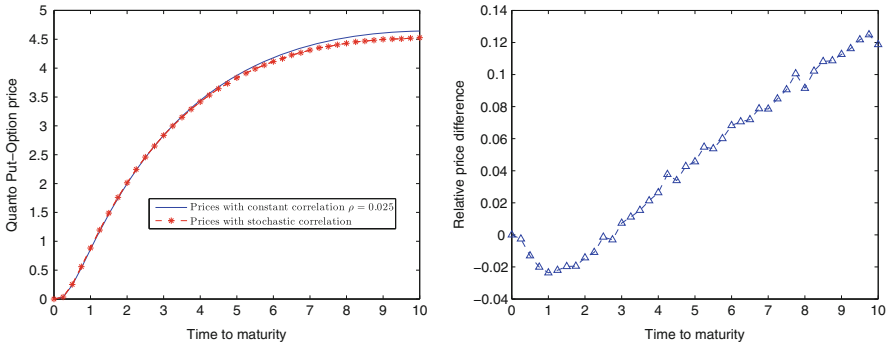


Fig. 1 Black-Scholes parameters: $K = 80$, $S_0 = 100$, $R_0 = 1$, $r_d = 0.05$, $r_e = 0.03$, $\sigma_S = 0.2$, $\sigma_R = 0.4$, Correlation process parameters: $\kappa = 7.937$, $\mu = 0.003$, $\sigma = 1.186$ and $\rho_0 = 0.3$

Acknowledgements The authors were partially supported by the European Union in the FP7-PEOPLE-2012-ITN Programme under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN *STRIKE—Novel Methods in Computational Finance*).

The authors acknowledge partial support from the bilateral German-Spanish Project *HiPeCa—High Performance Calibration and Computation in Finance*, Programme Acciones Conjuntas Hispano-Alemanas financed by DAAD.

References

1. Brigo, D., Chourdakis, K.: Counterparty risk for credit default swaps: impact of spread volatility and default correlation. Int. J. Theor. Appl. Finance **12**, 1007–1026 (2009)
2. Engle, R.F.: Dynamic conditional correlation: a simple class of multivariate GARCH. J. Bus. Econ. Stat. **17**, 425–446 (2002)
3. Gourieroux, C., Jasiak, J., Sufana, R.: The Wishart autoregressive process of multivariate stochastic volatility. J. Econ. **150**, 167–181 (2009)
4. Langnau, A.: Introduction into "Local Correlation" modelling. Quantitative Finance Papers (2009). 0909.3441. Arxiv.org
5. Ma, J.: Pricing foreign equity options with stochastic correlation and volatility. Ann. Econ. Finance **10**, 303–327 (2009)
6. Risken, H.: The Fokker-Planck Equation. Springer, Berlin (1989)
7. Schöbel, R., Zhu, J.: Stochastic volatility with an ornstein uhlenbeck process: an extension. Eur. Finan. Rev. **3**, 23–46 (1999)
8. Teng, L., van Emmerich, C., Ehrhardt, M., Günther, M.: A versatile approach for stochastic correlation using hyperbolic functions. Preprint 13/14. University of Wuppertal (2013)
9. Teng, L., Ehrhardt, M., Günther, M.: Bilateral counterparty risk valuation of cds contracts with simultaneous defaults. Int. J. Theor. Appl. Finance **16**(7), 1350040 (2013)
10. van Emmerich, C.: Modelling correlation as a stochastic process. Preprint 06/03, University of Wuppertal (2006)

Numerical Solution of Partial Integro-Differential Option Pricing Models with Cross Derivative Term

M. Fakharany, R. Company, and L. Jódar

Abstract The aim of this paper is to construct a reliable and efficient finite difference scheme for American option pricing under Bates model. First, we transform the associated partial-integro differential equation for this model into another suitable one without the cross derivative. Thereafter, a finite difference discretization has been used for the partial derivatives while the integral part is discretized using the four-points open type formula. The obtained finite difference scheme is solved using PSOR method. Several examples are included showing the advantage of the proposed approach.

Keywords Computational finance • Option pricing • Partial-integro-differential equations

1 Problem Formulation

In this paper we study American options under Bates model. This model is dedicated to describe the behavior of the underlying asset S and its variance ν by the following coupled stochastic differential equations:

$$dS(t) = (r - q - \lambda\xi)S(t)dt + \sqrt{\nu(t)}S(t)dW_1 + (\eta - 1)S(t)dZ(t),$$

$$d\nu(t) = \kappa(\theta - \nu(t))dt + \sigma\sqrt{\nu(t)}dW_2,$$

such that

$$dW_1 dW_2 = \rho dt,$$

M. Fakharany (✉) • R. Company • L. Jódar
Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
e-mail: fakharany@aucegypt.edu; rcompany@imm.upv.es; ljodar@imm.upv.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_15

where W_1 and W_2 are standard Brownian motions, Z is the poisson process. The parameter r is the risk free interest rate, q is the continuous dividend yield, λ is the jump intensity, η is the jump amplitude of the jump diffusion process and ξ is the expected relative jump size ($\xi = E[\eta - 1]$). Whereas κ is the mean reversion rate, θ is the long-run variance, σ is the volatility of the variance ν and ρ is the Wiener correlation parameter. By using Itô calculus and standard arbitrage arguments one gets the partial integro-differential equation (PIDE) for European option case [1, 2]

$$L(U) = \frac{\partial U}{\partial \tau} - \frac{1}{2}\nu S^2 \frac{\partial^2 U}{\partial S^2} - \rho\sigma\nu S \frac{\partial^2 U}{\partial S \partial \nu} - \frac{1}{2}\nu\sigma^2 \frac{\partial^2 U}{\partial \nu^2} - (r - q - \lambda\xi) S \frac{\partial U}{\partial S}$$

$$- \kappa(\theta - \nu)\frac{\partial U}{\partial \nu} + (r + \lambda)U - \lambda \int_0^\infty U(S\eta, \nu, \tau) f(\eta) d\eta = 0, \tag{1}$$

$$f(\eta) = \frac{1}{\sqrt{2\pi}\hat{\sigma}\eta} \exp[-\frac{(\ln \eta - \mu)^2}{2\hat{\sigma}^2}], \tag{2}$$

where $\tau = T - t$ is the time to maturity, μ is the mean of the jump and $\hat{\sigma}$ is the standard deviation. The problem is subjected to the initial condition given by the payoff function $g_1(S, \nu)$ for call options

$$U(S, \nu, 0) = g_1(S, \nu) = \max\{S - E, 0\}. \tag{3}$$

On the other hand the corresponding linear complementarity problem (LCP) for American option pricing is given by

$$L(U) \geq 0,\ U \geq g_1,\ L(U)(U - g_1) = 0, \tag{4}$$

associated with the following boundary conditions

$$U(0, \nu, \tau) = g_1(0, \nu),\ \lim_{S \to \infty} U(S, \nu, \tau) = \lim_{S \to \infty} g_1(S, \nu),\ \lim_{\nu \to \infty} \frac{\partial U}{\partial \nu}(S, \nu, \tau) = 0. \tag{5}$$

2 Problem Transformation and Discretization

2.1 Problem Transformation

The discretization of the cross spatial derivative generates negative coefficients which lead to several problems such as poor accuracy and slow convergence [3]. In [4], the mixed spatial derivative is discretized using nine point compact stencil, while it is discretized by seven point stencil associated with a condition on the

correlation parameter ρ in [2, 5, 6]. A finite difference scheme with cross derivatives correctors for multidimensional parabolic systems is investigated in [7] and this idea is used in [8] to solve the Heston model PDE problem for the case of European options.

Our aim is to remove the cross spatial derivative using a suitable transformation. To achieve this aim, first we use the discriminant quantity

$$\Delta(v, S) = B^2 - 4AC = \sigma^2 v^2 S^2(\rho^2 - 1), \tag{6}$$

where $A = \frac{1}{2}vS^2$, $B = \rho\sigma vS$ and $C = \frac{1}{2}v\sigma^2$. Since $-1 < \rho < 1$, implies that $\Delta(v, S) < 0$, consequently, (1) is of elliptic type. Then by using the canonical form for the elliptic equation, see [9], the suitable substitution is obtained by solving the following ordinary differential equation

$$\frac{dv}{dS} = \frac{\sigma(\rho + i\sqrt{1-\rho^2})}{S} = \frac{\sigma}{S}(\rho + i\tilde{\rho}),\ i = \sqrt{-1}, \tag{7}$$

where $\tilde{\rho} = \sqrt{1-\rho^2}$. From (7), we have $x = \sigma\tilde{\rho}\ln S$ and $y = \sigma\rho\ln S - v$, i.e., $(y = mx - v)$ such that $m = \frac{\rho}{\tilde{\rho}}$. Hence by using the following transformation

$$x = \sigma\tilde{\rho}\ln S,\ y = mx - v,\ V(x, y, \tau) = e^{(r+\lambda)\tau}U(S, v, \tau), \tag{8}$$

the operator $L(U)$ is transformed into

$$\mathscr{L}(V) = \frac{\partial V}{\partial \tau} - \frac{\tilde{\rho}^2 v\sigma^2}{2}\Big(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2}\Big) - \hat{a}\frac{\partial V}{\partial x} - \tilde{a}\frac{\partial V}{\partial y} - I(V), \tag{9}$$

with

$$I(V) = \lambda\int_0^\infty V(x + \sigma\tilde{\rho}\ln\eta,\ y + \rho\sigma\ln\eta, \tau)f(\eta)d\eta, \tag{10}$$

where

$$\hat{a} = \sigma\tilde{\rho}(\hat{\xi} - \frac{v}{2}),\ \tilde{a} = \sigma\rho(\hat{\xi} - \frac{v}{2}) - \kappa(\theta - v) \text{ and } \hat{\xi} = r - q - \lambda\xi. \tag{11}$$

In order to match the discretization of the differential and integral parts of (9), we consider the following change of variable in the integral part

$$\phi = x + \sigma\tilde{\rho}\ln\eta. \tag{12}$$

Hence from (2) and (10) one gets

$$I(V) = \frac{\lambda}{\sqrt{2\pi}\hat{\sigma}\tilde{\rho}\sigma}\int_{-\infty}^{\infty} V(\phi, y + m(\phi - x), \tau)\exp\left[\frac{-1}{\hat{\sigma}^2}\left(\frac{\phi - x}{\sigma\tilde{\rho}} - \mu\right)^2\right]d\phi. \tag{13}$$

In light of the transformation (8), the payoff (3) is converted into

$$V(x, y) = g_2(x, y) = \max\{e^{\frac{x}{\sigma\rho}} - E, 0\}.$$

Therefore, the transformed LCP (4) for American options pricing takes the following form

$$\mathscr{L}(V) = \frac{\partial V}{\partial \tau} - D(V) - I(V) \geq 0,\ V \geq g_2,\ \mathscr{L}(V)(V - g_2) = 0, \tag{14}$$

where,

$$D \equiv \frac{\tilde{\rho}^2\nu\sigma^2}{2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + \hat{a}\frac{\partial}{\partial x} + \tilde{a}\frac{\partial}{\partial y}. \tag{15}$$

The transformed boundary conditions are

$$\lim_{x\to\pm\infty} V(x, y, \tau) = \lim_{x\to\pm\infty} g_2(x, y),\ \frac{\partial V}{\partial y} = 0, mx - y \to \infty. \tag{16}$$

Under the transformation (8), a numerical bounded rectangular domain of the form $[S_1, S_2]\times[\nu_1, \nu_2]$ is converted into a rhomboid $ABCD$ as shown in Fig. 1, where the sides are given by

$$\overline{AB}=\{(x,y)\in\mathbb{R}^2|\ a\leq x\leq b,\ y=mx-\nu_2\},$$

$$\overline{BC}=\{(x,y)\in\mathbb{R}^2|\ x=b,\ y=mb-\nu,\ \nu_1\leq\nu\leq\nu_2\},$$

$$\overline{CD}=\{(x,y)\in\mathbb{R}^2|\ a\leq x\leq b,\ y=mx-\nu_1\},$$

$$\overline{DA}=\{(x,y)\in\mathbb{R}^2|\ x=a,\ y=ma-\nu,\ \nu_1\leq\nu\leq\nu_2\},$$

where $a=\sigma\tilde{\rho}\ln S_1$ and $b=\sigma\tilde{\rho}\ln S_2$.

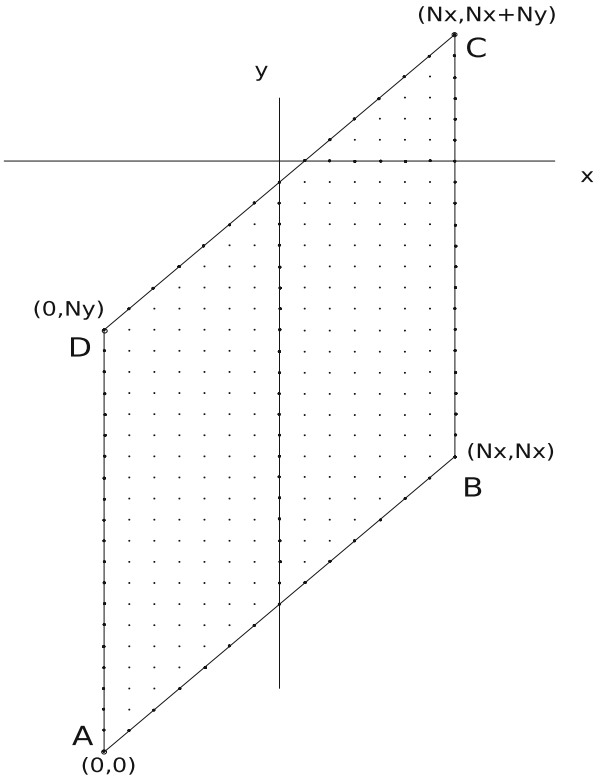


Fig. 1 Rhomboid numerical domain *ABCD*

2.2 Numerical Scheme Construction

Here the rhomboid computational domain is discretized by a uniform mesh points (x_i, y_j) such that $x_i = a + ih,\ 0 \le i \le N_x$ and $y_j = y_0 + j|m|h,\ i \le j \le N_x + i$ where $h = (b-a)/N_x$, $y_0 = ma - \nu_2$ and $N_y = (\nu_2 - \nu_1)/|m|h$. The first and second derivatives of the spatial variables of the operator D are discretized using the central finite difference approximation as follow

$$\begin{array}{ll} \frac{\partial V}{\partial x} \approx \frac{V_{i+1,j}-V_{i-1,j}}{2h} & \frac{\partial V}{\partial y} \approx \frac{V_{i,j+1}-V_{i,j-1}}{2|m|h} \\ \frac{\partial^2 V}{\partial x^2} \approx \frac{V_{i+1,j}-2V_{i,j}+V_{i-1,j}}{h^2} & \frac{\partial^2 V}{\partial y^2} \approx \frac{V_{i,j+1}-2V_{i,j}+V_{i,j-1}}{m^2h^2}. \end{array} \tag{17}$$

Thus the discretization of the differential operator is given by

$$D(V_{i,j}) \approx \breve{B}(i,j)V_{i-1,j} + \breve{C}(i,j)V_{i,j-1} - B(i,j)V_{i,j} + \hat{B}(i,j)V_{i+1,j} + \hat{C}(i,j)V_{i,j+1}, \tag{18}$$

where

$$\begin{aligned}\breve{B}(i,j) &= \left(\frac{\tilde{\rho}^2\sigma^2\nu_{i,j}}{2h^2} - \frac{\hat{a}_{i,j}}{2h}\right),\ B(i,j) = \frac{\sigma^2\nu_{i,j}}{m^2h^2},\ \hat{B}(i,j) = \left(\frac{\tilde{\rho}^2\sigma^2\nu_{i,j}}{2h^2} + \frac{\hat{a}_{i,j}}{2h}\right)\\ \breve{C}(i,j) &= \left(\frac{\tilde{\rho}^2\sigma^2\nu_{i,j}}{2m^2h^2} - \frac{\tilde{a}_{i,j}}{2|m|h}\right),\ \hat{C}(i,j) = \left(\frac{\tilde{\rho}^2\sigma^2\nu_{i,j}}{2m^2h^2} + \frac{\tilde{a}_{i,j}}{2|m|h}\right),\end{aligned} \tag{19}$$

$\hat{a}_{ij}$ and $\tilde{a}_{ij}$ are obtained from (11) by replacing ν with $\nu_{i,j}$. It is worth mention that based on the foregoing transformation (8) we obtain a five point discretization stencil for the spatial differential operator D which leads to minimize the computational cost. Moreover, there is no restriction on the correlation parameter ρ.

Since the improper integral part of the Bates model contains the Gaussian probability density function, thus the unbounded domain of this integration can be truncated into a finite domain (a, b) with a suitable tolerance error $\varepsilon > 0$, see [10]

$$b = \sqrt{-2\hat{\sigma}^2 \ln(\varepsilon\hat{\sigma}\sqrt{2\pi})},\ \ a = -b. \tag{20}$$

There are various kinds of approximations for the integration. The approximation is said to be of closed type if the integrand function is evaluated at the end points of the interval and it is of open type when these end points are omitted. In order to discretize the integral part (13) using a suitable method, here the four point open type approximation has been used. Derives its accuracy via extrapolating the integrand function based on four interior points and excluding the end points of each subinterval [11]. Furthermore, for functions whose derivatives have singularities at the end points, open type formulas are more efficient than the corresponding closed formulas.

Let us denote the approximation of the $I(V)$ by $I_{i,j}$. Here each subinterval is selected such that it contains four interior points, so N_x should be chosen as a multiple of 5. Then we have

$$\begin{aligned}I_{i,j} = \hat{\lambda} \sum_{\ell=0}^{N_x/5-1} \big(&11\hat{g}_{i,5\ell+1}V_{5\ell+1,5\ell+1+j-i} + \hat{g}_{i,5\ell+2}V_{5\ell+2,5\ell+2+j-i}\\ &+\hat{g}_{i,5\ell+3}V_{5\ell+3,5\ell+3+j-i} + 11\hat{g}_{i,5\ell+4}V_{5\ell+4,5\ell+4+j-i}\big),\end{aligned} \tag{21}$$

where $\hat{\lambda} = \frac{5h\lambda}{24\sqrt{2\pi}\hat{\sigma}\tilde{\rho}\sigma}$ and

$$\hat{g}_{i,\ell} \equiv \hat{g}(x_i, \phi_\ell) = \exp\left[\frac{-1}{2\hat{\sigma}^2}\left(\frac{\phi_\ell - x_i}{\sigma\tilde{\rho}} - \mu\right)^2\right],\ 0 \le \ell \le N_x. \tag{22}$$

Hence we have the following semi-discrete LCP

$$\frac{\partial \mathbf{V}}{\partial\tau} + A\mathbf{V} \ge 0,\ \mathbf{V} \ge \mathbf{g_2},\ (\frac{\partial \mathbf{V}}{\partial\tau} + A\mathbf{V})^T(\mathbf{V} - \mathbf{g_2}) = 0, \tag{23}$$

where A is a matrix of size $(N_x + 1)(N_y + 1) \times (N_x + 1)(N_y + 1)$ involving the differential and integral parts.

Finally, the time variable τ is discretized using the Rannacher scheme [12], such that the first four time levels are implemented using the implicit Euler while the rest of the time levels are obtained using Crank-Nicolson. The time variable is discretized in this manner to avoid the oscillation of the solution, see [5].

$$\tau^n = \begin{cases} (\frac{n}{2N_\tau})^2 T, & n = 0, 1, 2, 3, \\ (\frac{n-2}{N_\tau - 2})^2 T, & n = 4, 5, \ldots, N_\tau. \end{cases} \tag{24}$$

The time step size is given by $k_n = \tau^{n+1} - \tau^n$, $n = 0, 1, \ldots, N_\tau - 1$. Hence we obtain the following sequence of LCPs

$$LCP(\tilde{A}^{(n+1)}, \mathbf{V}^{(n+1)}, \tilde{\mathbf{V}}^{(n+1)}, \mathbf{g_2}), \tag{25}$$

where

$$\begin{aligned} \tilde{A}^{(n+1)} &= \begin{cases} I + k_n A & n = 0, 1, 2, 3, \\ I + \frac{1}{2} k_n A & n = 4, 5, \ldots, N_\tau - 1, \end{cases}, \\ \tilde{\mathbf{V}}^{(n+1)} &= \begin{cases} \mathbf{V}^{(n)} & n = 0, 1, 2, 3, \\ (I - \frac{1}{2} k_n A)\mathbf{V}^{(n)} & n = 4, 5, \ldots, N_\tau - 1. \end{cases} \end{aligned} \tag{26}$$

3 Numerical Results

The following examples have been done using a CPU with Microprocessor 3.4 GHz Intel Core i7 and implemented in Matlab. The option prices are obtained using the **PSOR** method with the relaxation parameter $\omega = 1.5$. The first example investigates the error for Bates model with negative correlation, meanwhile the second one investigates the error with positive correlation.

Example 1 Consider an American call option under Bates model with the following parameters $T = 0.5$, $E = 100$, $r = 0.03$, $q = 0.05$, $\theta = 0.04$, $\kappa = 2$, $\sigma = 0.25$, $\hat{\sigma} = 0.4$, $\mu = -0.5$, $\lambda = 0.2$ and $\rho = -0.5$, the computational domain for $x \in [a, b]$, where a and b are obtained by (20) and $[\nu_1, \nu_2] = [0.1, 1]$. Table 1 shows the variation of the root mean square relative error (RMSRE) of the option value at $S = \{80, 90, 100, 110, 120\}$ for several values of domain discretizations (N_x, N_y, N_τ). The reference values for the prices U are given in [5]. The ratio in Table 1 is the ratio of the each two successive root mean square relative errors.

Table 1 Results for example 1

PSOR(N_x, N_y, N_τ)	RMSRE	Ratio	CPU (s)
(25,15,10)	0.1377		0.024
(50,28,25)	0.0571	2.4116	0.185
(100,55,50)	0.0103	5.5269	5.596
(125,69,75)	0.0068	1.5117	17.08
(150,82,75)	0.0033	2.0430	51.74

Table 2 Results for example 2

PSOR(N_x, N_y, N_τ)	RMSRE	Ratio	CPU (s)
(25,53,10)	0.1405		0.24
(40,85,25)	0.0687	2.0452	3.56
(60,162,50)	0.0136	5.0676	57.42
(80,216,75)	0.0078	1.7439	200.73
(100,270,100)	0.0042	1.8525	415.26

Example 2 The parameters for an American call option under Bates model are selected as follow $T = 0.5$, $E = 100$, $r = 0.03$, $q = 0.05$, $\theta = 0.04$, $\kappa = 2$, $\sigma = 0.4$, $\hat{\sigma} = 0.1$, $\mu = 0$, $\lambda = 5$ and $\rho = 0.5$ with a tolerance error $\varepsilon = 10^{-4}$. The reference values are in [2], Table 2 reveals the associated RMSRE, ratio and CPU time for several step sizes discretization.

Acknowledgements This work has been partially supported by the European Union in the FP7-PEOPLE-2012-ITN program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN STRIKE-Novel Methods in Computational Finance).

References

1. Bates, D.S.: Jumps and stochastic volatility: exchange rate processes implicit Deutsche mark options. Rev. Financ. Stud. **9**, 69–107 (1996)
2. Chiarella, C., Kang, B., Mayer, G.H., Ziogas, A.: The evaluation of American option prices under stochastic volatility and jump-diffusion dynamics using the method of lines. Research Paper Quantitative Finance Research Centre University of Technology, Sydney, vol. 219, pp. 1–43 (2008).
3. Zvan, R., Forsyth, P.A., Vetzal, K.R.: Negative coefficients in two-factor option pricing models. J. Comput. Finance **7**, 37–73 (2003)
4. Düring, B., Fournié, M.: High-order compact finite difference scheme for option pricing in stochastic volatility models. J. Comput. Appl. Math. **236**, 4462–4473 (2012)
5. Toivanen, J.: A componentwise splitting method for pricing American options under the Bates model. In: Computational Methods in Applied Sciences, vol. 15, pp. 213–227. Springer, Berlin (2010)
6. Salmi, S., Toivanen, J., Von Sydow, L.: Iterative methods for pricing American options under the bates model. Procedia Comput. Sci. **18**, 1136–1144 (2013)
7. Bouchut, F., Frid, H.: Finite difference schemes with cross derivatives correctors for multidimensional parabolic systems. J. Hyperbolic Differ. Equ. **3**(1), 27–52 (2006)

8. Briani, M., Natalini, R., Papi, M., Paris, C.: Finite differences approximations for multidimensional models of pricing. PREMIA Report (version 14), pp. 1–14 (2014)
9. Farlow, S.J.: Partial Differential Equations for Scientists and Engineers. Dover, New York (1993)
10. Briani, M., La Chioma, C., Natalini, R.: Convergence of numerical schemes for viscosity solutions to integro-differential degenerate parabolic problems arising in financial theory. Numer. Math. **98**(4), 607–646 (2004)
11. Davis, P.J., Rabinowitz, P.: Methods of Numerical Integration, 2nd edn. Academic, New York (1984)
12. Rannacher, R.: Finite element solution of diffusion problems with irregular data. Numer. Math. **43**, 309–327 (1982)

MS 4
MINISYMPOSIUM: CURRENT CHALLENGES IN COMPUTATIONAL FINANCE

Organizers

Claudio Albanese[1]

Speakers

Giacomo Pietronero[2]
Finite Difference Methods with Level Algebra

Claudio Albanese[1], Giacomo Pietronero[2]
Portfolio Simulations for the "Mega-Models"

Paolo Regondi[3], Claudio Albanese[1], Mohammad Zubair[4]
Speeding Algebraic Pricing Methods Using Matrix Operations on GPU

Sebastian Del Bano Rollin[5], Claudio Albanese[1], Giacomo Pietronero[2]
Probabilistic Interpretation of Finite Difference Methods

[1]Claudio Albanese, Global Valuation Ltd., London, United Kingdom.
[2]Giacomo Pietronero, Global Valuation Ltd., London, United Kingdom.
[3]Paolo Regondi, Global Valuation Ltd., London, United Kingdom.
[4]Mohammad Zubair, Old Dominion University, Norfolk, United States.
[5]Sebastian Del Bano Rolllin, Department of Mathematic, University College London, United Kingdom.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_16

Keywords

Computational finance
Finite difference
Portfolio simulation

Short Description

In the aftermath of the crisis, the computational challenges in Finance have been shifting from exotic derivative pricing to the simulation of large complex portfolios of simpler instruments. This symposium focuses on emerging problems and challenges in this area.

Recasting Finite Difference Methods in Finance to Exploit GPU Computing

Claudio Albanese, Sebastian del Baño Rollin, and Giacomo Pietronero

Abstract Finite difference methods (FDM) have been developed and optimized in a technology context that has radically changed. When FDMs became a standard it used to be that memory was a scarce resource and that algorithms were either memory or compute bound. As a consequence traditional FDMs have been designed to minimize the number of operations and the memory footprint given a certain level of accuracy. In this paper we describe how the potential of GPU computing can be exploited to rethink the way FDM are implemented in the context of financial applications.

Keywords Computational finance • Finite difference • GPU computing

1 Introduction

Finite difference methods (FDM) have been developed and optimized in a technology context that is no longer current. When FDMs affirmed themselves as a standard it used to be that memory was a scarce resource and that algorithms were either memory or compute bound. As a consequence traditional FDMs have been designed to minimize the number of operations and the memory footprint to the detriment of accuracy.

The potential of computers in terms of floating point operations per seconds (FLOPS) and memory available increased dramatically over the last years offering an opportunity for a rethink of the way finite difference methods are implemented.

C. Albanese
Global Valuation, London, UK
e-mail: claudio.albanese@global-valuation.com

S. del Baño Rollin (✉)
Computer Science, University College London, London, UK
e-mail: Sebastian.delbanorollin@ucl.ac.uk

G. Pietronero
Global Valuation, London, UK

Computer Science, University College London, London, UK
e-mail: giacomo.pietronero@global-valuation.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_17

Nowadays only a small class of algorithms are able to exploit the computational power of modern hardware. Most algorithms, including the (sparse) matrix-vector multiplication that is a cornerstone of traditional finite difference methods, are bounded by the number of memory reads and therefore are unable to exploit the computational power of parallel computing. To highlight these limitations we test the performance of FDMs on GPUs and analyze how GPU computing can be exploited to improve the way FDM are traditionally implemented.

As an alternative to the FDM approach we also describe a framework proposed by Albanese in [1] that is based on matrix multiplication (BLAS 3) and aims at retaining modelling flexibility while using compute bound algorithms.

2 Traditional Finite Difference Methods

Finite difference methods are numerical methods that approximate the solution of differential equations using finite difference equation to approximate derivatives.

In finance often problems are formulated in terms of partial differential equations (PDE), where the price of a derivative over time is a function of the first and second order derivative of the price of the option with respect to the underlying asset. When the PDE does not admit a closed form solution one can resort to finite difference methods to approximate the derivatives and solve the PDE iteratively on a lattice, [5].

In this paper we use the following parallelism between the different finite difference approximations and Padé approximants described in [6] and [7]. A pricing partial differential equation can be represented in terms of:

$$V_t + \mathrm{L} \cdot V = 0 \tag{1}$$

With L being a Sturm-Liuville operator. Assuming a time-homogeneous setting the solution to Eq. (1) is formally given by iterating:

$$V(t - \Delta t) = \mathrm{e}^{\Delta t \cdot \mathrm{L}} V(t) \tag{2}$$

FDMs can be seen as ways to approximate the matrix exponential $\mathrm{e}^{\Delta t \cdot \mathrm{L}}$, for the methods listed below the approximation corresponds with a Padé approximant.

Given a function $f(x)$ the Padé approximant is defined as the rational polinominal of order (m,n) where the coefficients a_i and b_j are chosen so to match the first $m + n$ derivatives of the exponential:

$$\mathrm{R}_{(m,n)}(x) = \frac{\sum_{i=1}^{m} a_i x^i}{1 + \sum_{j=1}^{n} b_j x^j} \tag{3}$$

In line with this parallelism the traditional explicit method is given by approximating the exponential in (2) with its Padé of order $(1, 0)$:

$$\mathrm{R}_{(1,0)}(\Delta t \cdot \mathrm{L}) = I + \Delta t \cdot \mathrm{L} \tag{4}$$

The implicit method instead approximates the exponential in (2) with the Padé of order $(0, 1)$:

$$\mathrm{R}_{(1,0)}(\Delta t \cdot \mathrm{L}) = (I - \Delta t \cdot \mathrm{L})^{-1} \tag{5}$$

and the Crank-Nicolson method corresponds to the Padé approximant of order $(1, 1)$:

$$\mathrm{R}_{(1,0)}(\Delta t \cdot \mathrm{L}) = \left(I + \frac{1}{2}\Delta t \cdot \mathrm{L}\right)\left(I - \frac{1}{2}\Delta t \cdot \mathrm{L}\right)^{-1} \tag{6}$$

In Eq. (2), given the choice of the method implemented, we substitute the appropriate Padé approximant to obtain:

$$V(t - \Delta t) = R_{m,n}(\Delta t \cdot \mathrm{L})V(t) \tag{7}$$

In general this iteration based on a matrix-vector multiplication is repeated many times until $V(0)$ is computed. So far the research in the field of FDM has focused mainly on finding approximations that allow to achieve a certain degree of accuracy while minimizing the memory usage and number of floating point operations needed to execute the algorithm. This translate in a search for algorithms that are well behaved for large Δt steps, so to minimize the number of iterations. Therefore important properties are the stability and the degree of accuracy of the approximation as Δt increases. Of the three methods presented here Crank-Nicolson is the only one that is second order accurate and it is also A_0 stable, meaning that the approximation converges for all choices of the time step. Nevertheless above a certain threshold for the choice of the time-step the eigenvalues of the operator R computed using CN can become negative, giving rise to unwanted oscillations in the solution as we show in Fig. 1.

3 GPU Computing

In the last decade the performance of a single processor has been increasing at a rate that has been much slower then in the previous decades. As a consequence the focus has shifted towards building multi-processor machines able to keep up for the need for more powerful machines. The result of this trend is that nowadays servers can have more then one central processing unit (CPU) and each CPU can have tens of

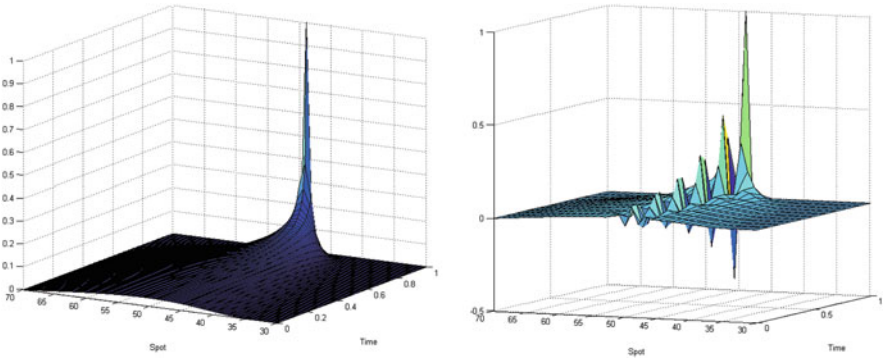


Fig. 1 Comparison of the solution of a PDE with a delta function as the final condition, on the *left* the solution has been obtained using CN with a step of 0.01 while on the *right* the step is 0.1

cores that execute instructions independently. In addition to that there is a wide range of co-processors that have been developed to provide additional computational power for bespoke problems. Among co-processors, Graphics Processing Units (GPUs) have initially been developed to accelerate computer graphics problems but are now used for a wide range of problems.

Nowadays a GPU can have up to 2880 cores, 12 GB of memory, and can achieve a performance of up to 4.27 Teraflops. The type of parallelism that can be implemented on GPUs is restricted to a Single Instruction Multiple Data (SIMD) paradigm, meaning that in order to compute a parallel calculation the same instruction set is applied by each core to different data.

Although the performance of multithreaded systems still continues to growth at a high rate, it becomes more and more difficult to achieve the peak performance of a machine. Many algorithms are not parallelizable and therefore cannot execute in a multithreaded fashion. Further the speed of memory access has been growing at a very low level and that means that the performance of algorithms is bounded by the number of memory reads rather than the number of floating point operations.

As it has been pointed out by Dongarra and van der Steen [4] the algorithms that are better suited to exploit parallel hardware, in particular GPUs, are the ones that show a high ratio of floating point operations (FLOP) per memory read. Let d be the dimension of a square matrix, a Matrix-Vector multiplication requires $O(d^2)$ FLOP and $O(d^2)$ memory reads, the ratio between the two is a constant. Matrix-matrix multiplication instead requires $O(d^3)$ FLOP and $O(d^2)$, so the ratio between the two depends on the dimension d. As shown in Fig. 2 the performance measured in terms of FLOP per second on a GPU is very different for the two algorithms.

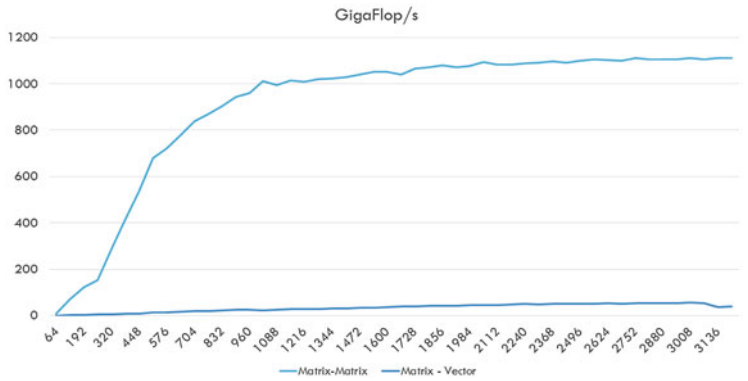


Fig. 2 Comparison of the performance measured in Floating point operations per second of the Matrix-Matrix multiplication (GEMM) vs Matrix-Vector multiplication (GEMV) for different dimensions of the matrix

4 How to Adapt FDM to Parallel Hardware

We have discussed how GPU computing favors algorithm that rely on matrix-matrix multiplications.

Under the assumption of time-homogeneity the iterative scheme introduced in (7) can be restructured using fast exponentiation. Instead of $\frac{T}{\Delta t}$ matrix-vector operations we can re-write the algorithm in terms of $\log_2\left(\frac{T}{\Delta t}\right)$ matrix squarings (for an appropriate choice of Δt), with only one final matrix-vector operation:

$$V(0) = \left(R^{\Delta t}_{m,n}(\theta)\right)^{\frac{T}{\Delta t}} V(T) \tag{8}$$

where the exponential of the operator can be computed using a fast exponentiation algorithm that allows to double the time step of the operator by squaring the operator itself:

$$\begin{aligned} \left(R^{\Delta t}_{m,n}(\theta)\right) \cdot \left(R^{\Delta t}_{m,n}(\theta)\right) &= \left(R^{\Delta t}_{m,n}(\theta)\right)^2 \\ \left(R^{\Delta t}_{m,n}(\theta)\right)^2 \cdot \left(R^{\Delta t}_{m,n}(\theta)\right)^2 &= \left(R^{\Delta t}_{m,n}(\theta)\right)^4 \end{aligned} \tag{9}$$

The solution has the advantage that the Δt can be halved with an additional matrix-matrix operation, while the traditional iterative algorithm would require to double the number of iterations. Nevertheless the assumption of time-homogeneity is not realistic, in financial applications typically one or more parameters have a term structure so to fit the complexity of market data.

4.1 A Case Study: Crank-Nicolson

The Crank-Nicolson (CN) method is one of the most popular among practitioners because it is second order accurate and A_0 stable, meaning that the approximation converges for all choices of the time step. On the other side above a certain threshold for the time-step the eigenvalues of the operator R computed with CN can become negative, creating unwanted oscillations in the solution.

To avoid this issue we implement a backward induction where the approximation is computed over an interval Δt small enough so that the matrix $R^{\Delta t}_{m,n}(\theta)$ has only positive eigenvalues. This operator is then exponentiated in order to obtain the operator over ΔT so that the backward induction is executed over longer time steps.

$$V(T - \Delta T) = \left[R^{\Delta t}_{m,n}(\theta)\right]^{\frac{\Delta T}{\Delta t}} V(T) \tag{10}$$

We implemented a case study on a nVidia K10 GPU using the cuBLAS library. The set up of the backward induction is $T = 1$ year, $\Delta T = \frac{T}{10}$ and Δt is chosen to be equal to $2^{-10} \cdot \Delta T$, as a result Δt is less then 1 h. The aim of the study is to compare the performance of three different implementations of a backward induction using a Crank-Nicolson approximation. This is how the backward induction algorithm has been structured in the different runs:

- Run 1: this is the classic backward induction algorithm based on matrix-vector operations executed with $\frac{T}{\Delta t}$ steps of length Δt
- Run 2: in this run we allow for time dependencies in the PDE, as a consequence $R^{\Delta t}_{m,n}(\theta)$ is computed and exponentiated for each interval ΔT
- Run 3: in this run we assume that the PDE is time-homogeneous, therefore the operator $R^{\Delta T}_{m,n}(\theta)$ can be computed only once.

In the different runs the Crank Nicolson algorithm is always implemented over a step Δt, hence the final results differ only because of floating point errors. Figures 3

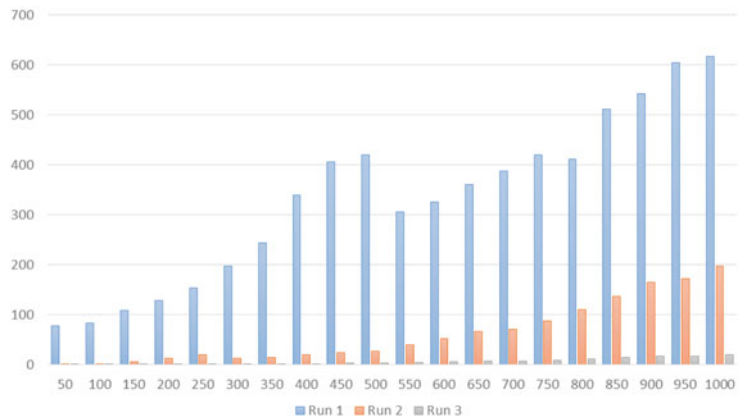


Fig. 3 Time to execute the backward induction in MS for different choices in the number of points for the spot discretization

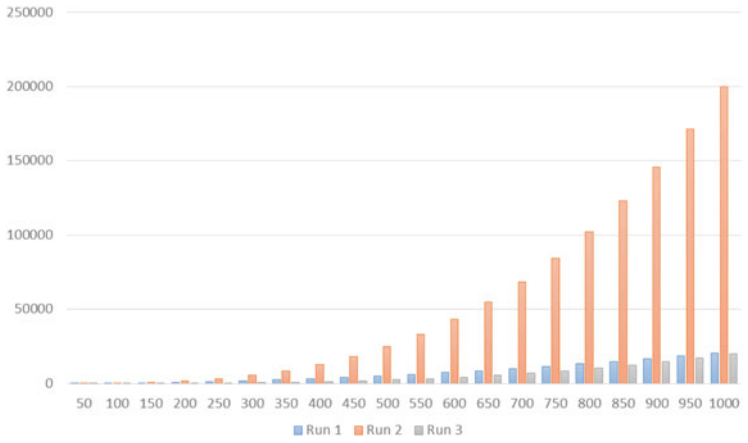


Fig. 4 Number of floating point operations for different choices in the number of points for the spot discretization

and 4 show how despite Run 1 requires considerably less number of floating point operations compared to Run 2 it takes more time to execute it.

The fast exponentiation can be even more useful to speed up higher-order Padé approximants, where the approximation is stable and free of oscillations only below a very small Δt.

4.2 Operator Methods

In finance numerical schemes are typically used to approximate processes that are continuous in time and space. The calibration of the model can be carried out exploiting the analytical properties of the exact dynamic while numerical schemes can be used for pricing, when analytical shortcuts are not available for a given payoff.

In [2], [1] and [3] Albanese et al., taking advantage of the performance of BLAS 3 algorithms on GPUs, has developed a mathematical framework that allows to specify a dynamic in discrete time and space.

A process is defined by a parametrized set of rules that fill a Markov generator L, this generator is used to build a matrix u of transition probabilities over an interval δt:

$$u_{\delta t}\left(y, y'\right) = I + \delta t \cdot L\left(y, y'\right) \tag{11}$$

where y is a generic state of the underlying and

$$\delta t \leq \frac{1}{max_y |L(y,y)|} \tag{12}$$

Although this formula is the same we used for the explicit method approximation in Eq. (4), here the formula is used to define a process, not as approximation of a continuous-time process with generator L.

As this set-up doesn't require any analytical tractability of the dynamic, one can implement a wider range of models to achieve a better degree of realism in the process.

Also calibration and pricing can be executed consistently using the same numerical schemes.

5 Conclusions

GPUs strongly favor algorithms based on BLAS 3 routines, in particular matrix-matrix multiplication. This opens an opportunity to rethink the way finite difference methods are implemented. Instead of looking for methods that prove satisfactory for a large time-step, fast exponentiation allows to efficiently use higher order methods with a small time-step so to avoid instabilities and oscillations.

References

1. Albanese, C.: Operator methods, Abelian processes and dynamic conditioning. Online document (2007)
2. Albanese, C., Li, H.: Monte Carlo pricing using operator methods and measure changes (2009). Available at SSRN: http://ssrn.com/abstract=1484556
3. Albanese, C., Bellaj, T., Gimonet, G., Pietronero, G.: Coherent global market simulations and securitization measures for counterparty credit risk. In: Quantitative Finance, vol. 11, pp. 1–20. Taylor and Francis, London (2011)
4. Dongarra, J.J., van der Steen, A.J.: High-performance computing systems: status and outlook. Acta Numer. **21**, 379–474 (2012)
5. Duffy, D.: Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach. Wiley Finance. Wiley, New York (2006)
6. Jäckel, P.: Finite discretizing schemes as Padé approximants. Online document (2011)
7. Smith, G.D.: Numerical Solution of Partial Differential Equations: Finite Difference Methods. Oxford University Press, New York (1978)

BLAS Extensions for Algebraic Pricing Methods

Claudio Albanese, Paolo Regondi, and Mohammad Zubair

Abstract PDE pricing methods such as backward and forward induction are typically implemented as unconditionally marginally stable algorithms in double precision for individual transactions. In this paper, we reconsider this strategy and argue that optimal GPU implementations should be based on a quite different strategy involving higher level `BLAS` routines. We argue that it is advantageous to use conditionally strongly stable algorithms in single precision and to price concurrently sub-portfolios of similar transactions. To support these operator algebraic methods, we propose some `BLAS` extensions. CUDA implementations of our extensions turn out to be significantly faster than implementations based on standard `cuBLAS`. The key to the performance gain of our implementation is in the efficient utilization of the memory system of the new GPU architecture.

Keywords Computational finance • GPU computing

1 Introduction

Recently, one of the authors of this paper developed a method for combined value-risk analysis within a global market using mathematical formalism around matrix operations making it amenable to efficient implementation on massively parallel architectures like GPUs [1]. One of the matrix operations that dominates the overall execution time of the calibration component of the combined value-risk analysis framework is a BLAS-like operation, which we refer to it as `Sgemv4`.

The `Sgemv4` computation can be viewed as a multiplication type operation of a vector of matrices $A^1, A^2, \ldots A^m$ with a matrix B, where A^i is multiplied by a number of columns $\{B_{j_1}, B_{j_2}, B_{j_3}, \ldots, B_{j_{q_i}}\}$, not necessarily contiguous, of matrix B. Note that q_i, which is the number of columns of B that need to be used for

C. Albanese • P. Regondi
Global Valuation Limited, 9 Devonshire Square, London EC2M 4YF, UK
e-mail: claudio.albanese@global-valuation.com; paolo.regondi@global-valuation.com

M. Zubair (✉)
Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA
e-mail: zubair@cs.odu.edu

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_18

Table 1 Performance comparison of `Sgemm8` with `cublasSgemv` and `cublasSgemm` for matrix size 1024 × 1024

Number of vectors	`Sgemm8` (GFLOPS/s)	`cublasSgemv` (GFLOPS/s)	`cublasSgemm` (GFLOPS/s)
1	52	30	14
2	95	30	27
3	128	30	31
4	161	30	54
5	175	30	68
6	188	30	82
7	202	30	94
8	200	30	107

multiplication, varies and is a function of i. In this paper, we discuss different ways of implementing `Sgemv4` on Kepler GPU with varying level of performance. A naive implementation is to use a level-2 BLAS [2], for this operation. CUDA Toolkit 5.5 is shipped with CUDA Basic Linear Algebra Subroutines (`cuBLAS`) library [5]. We make q_i calls of `cublasSgemv` (part of `cuBLAS`) for multiplying A_i by q_i columns of B. We improve on the naive implementation by collecting columns $\{B_{j_1}, B_{j_2}, B_{j_3}, \ldots, B_{j_{q_i}}\}$ to form a matrix and using a `cublasSgemm` (part of `cuBLAS`), a level-3 BLAS [7] that replaces q_i calls of `cublasSgemv`. We observed that for small values of q_i, the performance of `cublasSgemm` is worse than repeated calls of `cublasSgemv`. The value of q_i in our application ranges from 1 to 50. In our experimentation with matrices of size 1024 × 1024, we found that for value of $q_i < 8$ it is better not to use `cublasSgemm`. This required that we develop our own BLAS-like extension, `Sgemm8`, that is optimized for multiplying a matrix with one to eight columns. NVIDIA `cuBLAS` library does not support such an extension.[1]

We developed an optimized implementation of `Sgemm8` for Tesla Kepler architecture. The performance of `Sgemm8` is significantly better than that of `cublasSgemv` or `cublasSgemm` for matrix multiplication with one to eight vectors. Table 1 summarizes the performance of `Sgemm8` as compared to that of `cuBLAS` library on a GK104 device. `Sgemm8` is 40 % better than the `cublasSgemv` for multiplying a matrix with a single vector; and is two times better than `cublasSgemm` for multiplying a matrix with eight vectors for matrix sizes that occur in our financial application. The key to the performance gain of `Sgemm8` is due to a novel algorithm that utilizes the new intrinsic function shuffle, which allows sharing of data between threads of a warp and helps in reducing memory latency. Note that `Sgemm8` is close to a level-2 BLAS and the execution time is dominated by the time to access the data from device memory.

[1] Intel MKL library does support, `Sgem2vu`, an extension of `Sgemv` for multiplying a matrix with two vectors [3].

We incorporated `Sgemm8` in the `Sgemv4` computation and observed an overall speedup of 7 as compared to a base level implementation of `Sgemv4` that uses `cublasSgemv`.

The rest of the paper is organized as follows. In the next section, we briefly discuss CUDA and Kepler architecture. We discuss implementation of `Sgemv4` using CUDA BLAS library in Sect. 3. In Sect. 4, we give details of the kernel `Sgemm8` that is key to enhancing the performance of `Sgemv4`. We discuss performance results of `Sgemv4` that utilizes the kernel `Sgemm8` and `cublasSgemm` in Sect. 5. Finally we conclude in Sect. 6.

2 CUDA and NVIDIA Kepler

A typical program on a system with a single GPU device is a C/C++ program with CUDA APIs to move data between system memory and GPU device memory, and to launch computation kernels on GPU [4]. The data between system memory and the device memory is moved using the PCI Express (PCIe) bus. These transfers are costly and therefore applications that have a higher computation to I/O ratio are suitable for GPU computing. Also, if possible these transfers should be minimized and it is desirable to leave the data on GPU if a subsequent kernel is going to use the same data. A GPU device uses several memory spaces that differ in their size, access latency, and read/write restrictions. These memory spaces include global, local, shared, texture, and registers. Global, local, and texture memory have the greatest access latency, followed by constant memory, registers, and shared memory.

CUDA provides an abstraction of thread hierarchy to allow computation from different domain to nicely map to different cores of the underlying hardware. The GPU hardware consists of a number of streaming multiprocessor which in turn consists of multiple cores. Threads are organized in blocks, where one or more block runs on a streaming multiprocessors. The threads in a block are further partitioned into subgroups of 32 threads referred as Warps. A Warp, that is a sub block of 32 threads, runs on eight or sixteen cores of a streaming multiprocessor in multiple clock cycles. Typically, data sharing between threads of a block is facilitated by the shared memory. The Kepler architecture supports another way of sharing data between threads of a warp, namely by using an intrinsic shuffle function. In this paper, we focus our implementation on the NVIDIA Kepler architecture.

The Kepler architecture comes in two models K10 (GK104) and K20 (GK110). We limit our discussion to K10 as we ran all our experiments on this model. Note that the K10 board has two GK104 devices. However, our performance results will hold on K20 also. All our performance results reported in this paper are on a single GK104 device. The Kepler GK104 model has eight streaming multiprocessors (SMX) with 192 cores on each SMX for a total of 1536 cores. Kepler has a larger register file per multiprocessor as compared to earlier models, which helps in improving occupancy [6].

3 Performance Result Using cuBLAS

We tested the performance of `Sgemv4` that occurs in the calibration of US dollar interest rate model that is part of the counter party credit risk analysis framework. For this scenario, the number of A matrices that is the value of $m = 57$, size of A matrix is 1024×1024, and size of B matrix is 1024×1213. Recall that in `Sgemv4` computation, we are multiplying a matrix A^i with q_i columns of B, for $i = 1$–57.

3.1 Base Level Using `cublasSgemv`

The base level implementation of `Sgemv4` was done using `cublasSgemv`. We make q_i calls of `cublasSgemv` to multiply A^i with q_i columns of B, for $i = 1$–57. We collected frequency distribution for values of q_is to capture how often we need to multiply a matrix with multiple vectors. The result is shown in Fig. 1. We also calculated the percentage of time spent for $q_i \leq 8$ and $q_i > 8$, see Fig. 2. The reason for this was to explore whether using `cublasSgemm` can help in improving performance. The choice of 8 was based on our experimentation that indicated that `cublasSgemm` has reasonable performance beyond eight columns. In this figure we also included the percentage of time we spent in data movement. After multiplying a vector of B with the matrix, we need to store the result vector back in B at the same location. This requires that we use a temporary buffer in device memory to hold the output vector before moving it back to the storage area of B. These results indicate that we are spending around 60 % of the time on multiplying matrices with $q_i > 8$. This suggest it may be better to use `cublasSgemm` when multiplying A^i with multiple vectors, that is for $q_i > 8$.

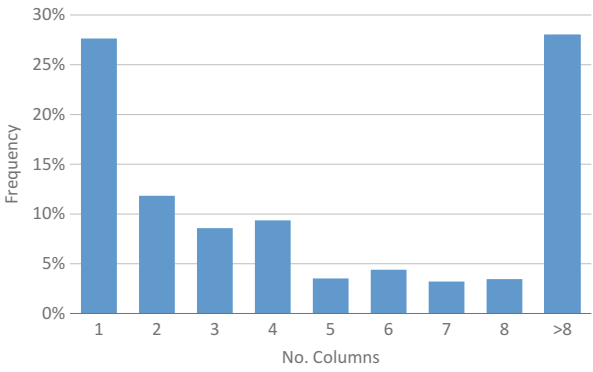


Fig. 1 Frequency distribution for values of q_i for USD model

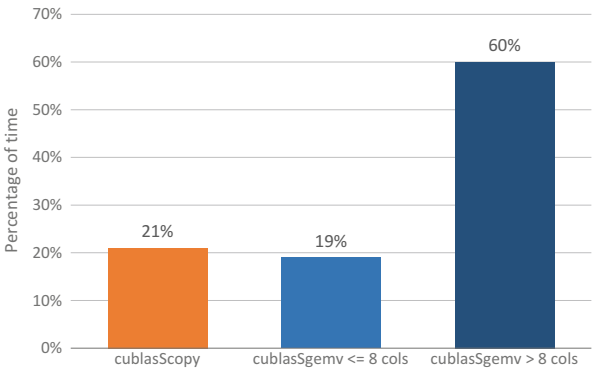


Fig. 2 Percentage of time spent by `cublasSgemv` for $q_i \leq 8$, and $q_i > 8$. The cost of data movement required to support `cublasSgemv` is also shown in the figure

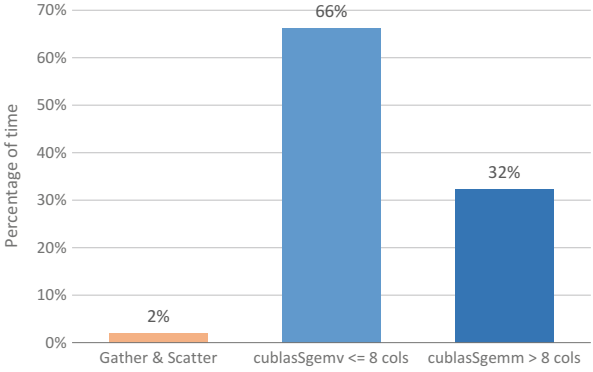


Fig. 3 Percentage of time spent by `cublasSgemv` for $q_i \leq 8$, and `cublasSgemm` for $q_i > 8$. The cost of data movement using efficient gather and scatter is also shown in the figure

3.2 *Using `cublasSgemv` and `cublasSgemm`*

For using `cublasSgemm`, we need to gather q_i vectors of B into a matrix. We make q_i calls of `cudaCopy` to collect q_i vectors into a matrix, before making a call to `cublasSgemm`. Though the execution time for matrix operations has significantly reduced by a factor of 50 %, the overhead of moving data within device memory has significantly increased. The major reason of this overhead is that we are making $q = \sigma_{i=1}^{m} q_i$ calls of `cudaCopy` and in each call we are copying a small amount of data. To address this issue, we wrote two kernel programs: gather and scatter. The gather program copies all q columns of B scattered in the memory to a temporary contiguous area of memory; and the scatter program does the reverse. This reduced the data movement time by a factor of 100. We also observed that for small values of q_i in the range from 2 to 4, the `cublasSgemm` is more expensive than calling `cublasSgemv` q_i times. The performance results where we call `cublasSgemm` for $q_i > 8$ along with the efficient gather and scatter is shown in Fig. 3.

4 Sgemm8

The `Sgemm8` kernel is an extension of BLAS to support matrix multiplication with one to eight vectors. For the case of one vector the `Sgemm8` is essentially a standard `Sgemv` BLAS routine.

The key idea of the algorithm is to use the new intrinsic shuffle function available on devices of compute capability 3.x that enables data sharing between threads of a warp without going to the shared memory. This has two benefits: (a) the sharing of data between threads happen with low latency, and (b) use of shared memory reduces, which in turn helps in improving occupancy. The shuffle function comes in four flavors [4].

- `_shfl()`: Copy from a specified source lane
- `_shfl_up()`: Copy from a lane with lower ID by a specified delta relative to caller
- `_shfl_down()`: Copy from a lane with higher ID by a specified delta relative to caller
- `_shfl_xor()`: Copy from a lane based on bitwise XOR of own lane ID

In the proposed implementation, we use the `_shfl()` function to share a register value of a thread with other threads in a warp. Figure 4 illustrates the working of `_shfl(xt, j)`.

4.1 *Performance Results of `Sgemm8`*

We compare the performance of `Sgemm8` for various matrix sizes with that of `cublasSgemv` and `cublasSgemm`, and the results are summarized in Figs. 5 and 6. These results indicate that the performance of `Sgemm8` is 40 % better than the

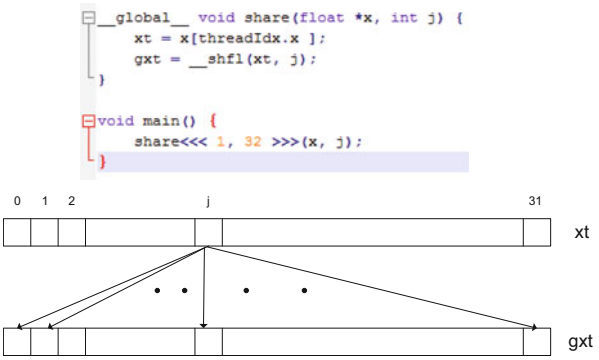


Fig. 4 A thread *th* reads *j* element of *x* into *xt*. The shuffle function as shown in the code segment broadcast value of *xt* at thread *j* to all threads in the warp

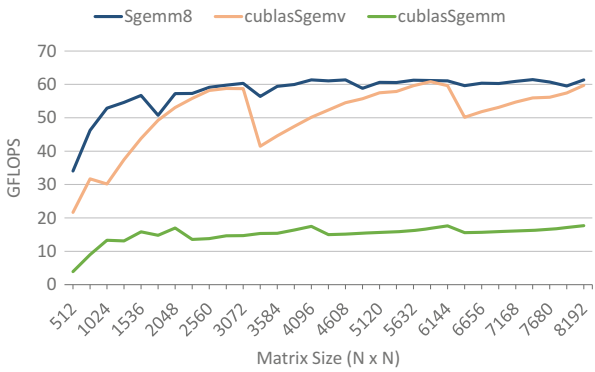


Fig. 5 Performance comparison of `Sgemm8` with `cuBlas` routines for multiplying a matrix with one column

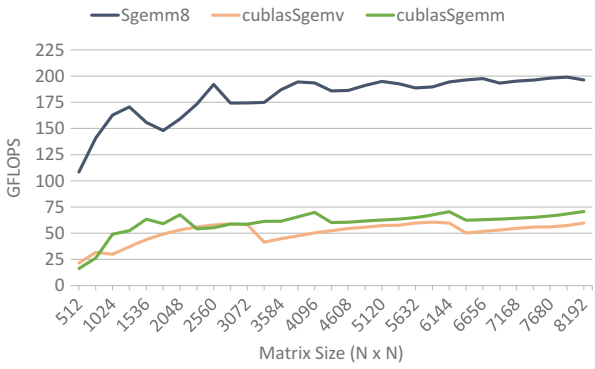


Fig. 6 Performance comparison of `Sgemm8` with `cuBlas` routines for multiplying a matrix with four columns

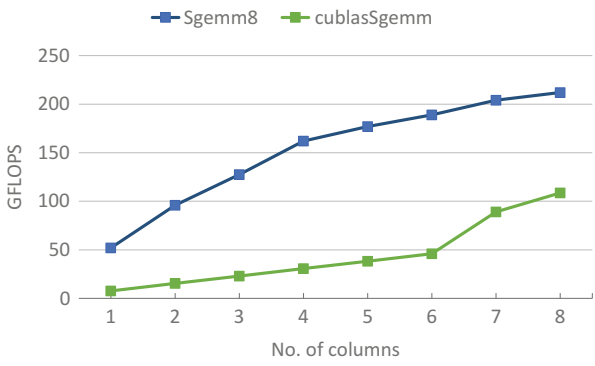


Fig. 7 Performance results for `Sgemm8` with varying number of columns

`cublasSgemv` for multiplying a matrix with a single vector for matrix sizes up to 1500. For larger matrices `Sgemm8` has a consistent performance for a single vector case, and is slightly better than `cublasSgemv`. The performance of `Sgemm8` is up to two times better compared to `cublasSgemm` or `cublasSgemv`, for multiplying matrix with multiple columns. We also summarize the performance results for `Sgemm8` with varying number of columns for matrix size 1024×1024 that occur in our financial application, see Fig. 7.

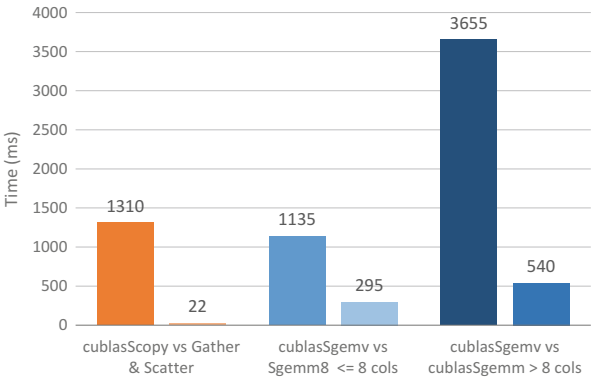


Fig. 8 The performance comparison of the base level with the final implementation that has all the optimizations like efficient gather/scatter, use of `Sgemm8` for columns 1–8, and `cublasSgemm` for columns greater than 8

5 Performance Results for `Sgemv4` Using `Sgemm8` Kernel

The performance comparison of the base level with the final implementation that has all the optimizations like efficient gather/scatter, use of `Sgemm8` for columns 1–8, and `cublasSgemm` for columns greater than 8 is shown in Fig. 8. Observe that the total execution time of the base level implementation is 6100 ms compared to that of 857 ms for the final implementation using all the optimizations, giving an overall speedup of over 7.

6 Conclusion

In this paper, we give an optimized implementation of a financial calibration application that relies heavily on matrix operations. We developed a BLAS-like kernel, `Sgemm8`, in support of the financial application that is used for multiplying a matrix with one to eight vectors. We demonstrated that the performance of `Sgemm8` is significantly better than that of `cublasSgemv` or `cublasSgemm` for matrix multiplication with one to eight vectors. `Sgemm8` is 40 % better than the `cublasSgemv` for multiplying a matrix with a single vector; and is two times better than `cublasSgemm` for multiplying a matrix with eight vectors for matrix sizes that occur in our financial application. The key to the performance gain of `Sgemm8` is due to a novel algorithm that utilizes the new intrinsic function shuffle, which allows sharing of data between threads of a warp and helps in reducing latency. We demonstrated that the calibration model speeds up by a factor of over 7 when we use `Sgemm8` compared to a base level implementation using `cuBLAS` routines.

Acknowledgements The author Mohammad Zubair would like to thank Prof. Philip Treleaven, who provided him the opportunity to spend time in Spring 2014 at University College of London and interact with companies working on high performance computing for financial applications.

References

1. Albanese, C., Bellaj, T., Gimonet, G., Pietronero, G.: Coherent global market simulations and securitization measures for counterparty credit risk. Quant. Finance **11**(1), 1–20 (2011). http://EconPapers.repec.org/
2. Anderson, E., Bai, Z., Bischof, C.H., Blackford, S., Demmel, J., Dongarra, J.J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.C.: LAPACK Users' Guide, 3rd edn. SIAM, Philadelphia, PA (1999). http://www.netlib.org/lapack/lug/
3. Corporation, I.: Computes two matrix-vector products using a general matrix (2013). http://software.intel.com/en-us/node/468648
4. Corporation, N.: Cuda c programming guide (2013). http://docs.nvidia.com/cuda/cuda-c-programming-guide/
5. Corporation, N.: Nvidia cuda basic linear algebra subroutines (2013). https://developer.nvidia.com/cublas
6. Corporation, N.: Tuning cuda applications for kepler (2013). http://docs.nvidia.com/cuda/kepler-tuning-guide/
7. Dongarra, J.J., Du Croz, J., Duff, I.S., Hammarling, S.: A set of Level 3 basic linear algebra subprograms. ACM Trans. Math. Softw. **16**, 1–28 (1990) (Algorithm 679)

MS 5
MINISYMPOSIUM: EU-MATHS-IN: A EUROPEAN NETWORK OF MATHEMATICS FOR INDUSTRY AND INNOVATION

Organizers

Peregrina Quintela Estévez[1] and Antonino Sgalambro[2]

Speakers

EU-MATHS-IN: A European Network of Mathematics for Industry and Innovation

Wil Schilders[3]
EU-MATHS-IN: An Introduction

Hans Georg Bock[4]
KoMSO – The German Strategic Committee for Mathematical Modeling, Simulation and Optimization

[1]Peregrina Quintela Estévez, Spanish Network for Mathematics and Industry (math-in), and Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

[2]Antonino Sgalambro, SM[i]2-Sportello Matematico per l'industria italiana, Roma, Italy.

[3]Wil Schilders, Platform Wiskunde Nederland, Amsterdam, The Netherlands.

[4]Hans Georg Bock, Committee for Mathematical Modeling, Simulation and Optimization (KoMSO), Heidelberg, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_19

[5]Guadalupe Parente, Spanish Network for Mathematics and Industry (math-in), Santiago de Compostela, Spain.

[6]Evgeny Verbitskiy, Platform Wiskunde Nederland (PWN), Amsterdam, The Netherlands.

[7]Michiel Bertsch, IAC - National Research Council of Italy, Roma, Italy.

[8]Maurizio Ceseri, National Research Council of Italy, Roma, Italy.

[9]Roberto Natalini, National Research Council of Italy, Roma, Italy.

[10]Mario Santoro, National Research Council of Italy, Roma, Italy.

[11]Francesco Visconti, National Research Council of Italy, Roma, Italy.

[12]Georges-Henri Cottet, Agence pour les mathématiques en interaction avec l'entreprise et la société (AMIES), Paris, France.

[13]Zoltán Horváth, Széchenyi István University, Gyõr, Hungary.

[14]Ioannis Bozas, European Patent Office, Munich, Germany.

[15]János Jósvai, Széchenyi István University, Gyõr, Hungary.

[16]Tamás Hajba, Széchenyi István University, Gyõr, Hungary.

[17]Sándor Kálmán, Audi Hungária Motors Ltd., Gyõr, Hungary.

EU-MATHS-IN: European Success Stories with Industry Part II. Traffic Management and Sustainable Energy

[18]Matteo Pozzi, Optit srl, Bologna, Italy.
[19]Daniele Vigo, Alma Mater Università di Bologna, Bologna, Italy.
[20] Angelo Gordini, Optit srl, Bologna, Italy.
[21] Claudio Caremi, Optit srl, Bologna, Italy.
[22] Sandro Bosso, Heracomm srl, Bologna, Italy.
[23]Giuseppe D'Aleo, Heracomm srl, Bologna, Italy.
[24]Beatrice Beleggia, Heracomm srl, Bologna, Italy.
[25]Valerio Vannini, Heracomm srl, Bologna, Italy.
[26]Giulia Biancardi, Heracomm srl, Bologna, Italy.
[27]Alpar Juttner, Eötvös University of Sciences, Budapest, Hungary.
[28]Stephen O'Brien, University of Limerick, Limerick, Ireland.
[29]William Lee, University of Limerick, Limerick, Ireland.
[30]Joanna Mason, University of Limerick, Limerick, Ireland.
[31]Ekaterina Kostina, University of Marburg, Marburg, Germany.
[32]Fabrice Gamboa, Institut de Mathématiques de Toulouse, Toulouse, France.
[33]Philippe Goudal, Mediamobile, Ivry-sur-Seine, France.
[34]Jean-Michel Loubes, Institut de Mathématiques de Toulouse, Toulouse, France.

Emilio Carrizosa[35], Carmen Ana Domínguez-Bravo[36], Enrique Fernández-Cara[37] and Manuel Quero[38]
Optimal Design of Solar Power Tower Systems

David Aller[39], Alfredo Bermúdez[40], Maria Teresa Cao-Rial[41], Pedro Fontan[42], Francisco Pena[43], Andrés Prieto[44], Jerónimo Rodríguez[45] and Jose Francisco Rodríguez-Calo[46]
Automatic Analysis of Floating Offshore Structures

EU-MATHS-IN: European Success Stories with Industry Part III. Biomedical Imaging, Electronics and Telecommunications

Vittoria Bruni[47], Raino Ceccarelli[48], Vincenzo Ricco[49] and Domenico Vitulano[50]
Efficient Image Processing Tools for Confocal Microscopy: A Case Study

Peter Maass[51]
MALDI Imaging: Sparsity Concepts for Analyzing and Visualizing Metabolic Information

[35]Emilio Carrizosa, Universidad de Sevilla, Sevilla, Spain.
[36]Carmen Ana Domínguez-Bravo, Universidad de Sevilla, Sevilla, Spain.
[37]Enrique Fernández-Cara, Universidad de Sevilla, Sevilla, Spain.
[38]Manuel Quero, Abengoa Solar Nt, Sevilla, Spain.
[39]David Aller, Centro de Tecnología Repsol, Madrid, Spain.
[40]Alfredo Bermúdez, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[41]Maria Teresa Cao-Rial, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[42]Pedro Fontan, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[43]Francisco Pena, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[44]Andrés Prieto, Universidad de A Coruña, A Coruña, Spain.
[45]Jerónimo Rodríguez, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[46]Jose Francisco Rodríguez-Calo, Centro de Tecnología Repsol, Madrid, Spain.
[47]Vittoria Bruni, University of Roma 'La Sapienza', Roma, Italy.
[48]Raino Ceccarelli, CrestOptics, Roma, Italy.
[49]Vincenzo Ricco, CrestOptics, Roma, Italy.
[50]Domenico Vitulano, Istituto per le Applicazioni del Calcolo-CRN, Bari, Italy.
[51]Peter Maass, ZeTeM-University of Bremen, Bremen, Germany.

Antoine Girard[52], Pierre Lebeaut[53], Roland Hildebrand[54], Nicolas Peltier[55] and Sylvian Kaiser[56]
MaiMoSiNE/DOCEA Power: Assessing the Quality of Reduced Order Models of Heat Transfer in Electronic Devices

Dee Denteneer[57]
Mathematical Models for Wireless Mesh Networks

Wil Schilders[58]
Model Order Reduction Within the Electronics Industry

Keywords

Biomedical imaging, electronics devices and telecommunications
European service network
Industrial mathematics
Innovation
Modelling, simulation and optimization in manufacturing and service management
Sustainable energy and traffic management models

Short Description

The development of new products or production processes today is dominated by the use of simulation and optimization methods that, based on a detailed mathematical modeling, support or even replace the costly production of prototypes and classical trial-and-error methods. To address this development and following the Recommendations of the Forward Look *Mathematics and Industry* published by the European Science Foundation, several European research networks have established a new organization to increase the impact of mathematics on innovations in key technologies and to foster the development of new modeling, simulation and optimization tools.

[52]Antoine Girard, University of Grenoble Alpes – CNRS, Grenoble, France.
[53]Pierre Lebeaut, DOCEA Power, Moirans, France.
[54]Roland Hildebrand, University of Grenoble Alpes – CNRS, Grenoble, France.
[55]Nicolas Peltier, DOCEA Power, Moirans, France.
[56]Sylvian Kaiser, DOCEA Power, Moirans, France.
[57]Dee Denteneer, Philips Research, Eindhoven, The Netherlands.
[58]Wil Schilders, TU Eindhoven, Eindhoven, The Netherlands.

In order to present the goals and strategies of EU-MATHS-IN and the operational background of its founding members, some mini-symposia were organized to collect and present the industrial experiences of these organizations in cooperation with their research partners, highlighting, when possible, the benefit for the interested firms.

The first session of mini symposium, entitled *EU-MATHS-IN: a European Network of Mathematics for Industry and Innovation* was devoted to introduce the goals and strategies of the new European network, together with the organization and the role of each national node member of EU-MATHS-IN, including their legal organization, their way of working to promote the relationship between researchers and businesses, how they spread the skills and experience of their groups, how they act as a one stop shop for their groups, which research groups/entities are represented in the node or how to become a member.

Then, three sessions of 2 h were followed, entitled: *EU-MATHS-IN: European success stories with Industry* by collecting contributions from several national networks. In particular, these sessions were organized by sorting on the base of the industrial sector they refer to: *Manufacturing and Service Management; Traffic Management and Sustainable Energy and Biomedical Imaging, Electronics and Telecommunications*.

Automatic Analysis of Floating Offshore Structures

David Aller, Alfredo Bermúdez, María Teresa Cao-Rial, Pedro Fontán, Francisco Pena, Andrés Prieto, Jerónimo Rodríguez, and José Francisco Rodríguez-Calo

Abstract In the coming years offshore wind energy will be one of the most promising areas in the renewable power generation field. Achieving the optimum design of floating platforms requires a rigorous analysis chain to establish the response of the whole platform under different scenarios. With this aim, we have developed a software package that automatically analyzes the feasibility of a floating structure. The structure of the platform is defined according to a very general set of parameters, allowing us to consider a wide range of designs. The package calls some commercial applications and some own codes, to complete the analysis process. Returned results include the hydrostatic equilibrium position, hydrodynamic pressure, RAOs (response-amplitude operators), material costs and static stresses.

Keywords Optimal design • Sustainable energy

1 Introduction

Offshore wind power is one of the most promising fields in renewable energy generation in the coming years. More than 90 % of the world's offshore wind power is currently installed in Europe. According to Global Wind Energy Council, offshore wind represents today about 2 % of the global wind power installed capacity, and

D. Aller • J.F. Rodríguez-Calo
Centro de Tecnología Repsol, Ctra. Extremadura km 18, 28935 Móstoles, Spain
e-mail: jfrodriguezc@repsol.com

A. Bermúdez • M.T. Cao-Rial • P. Fontán • F. Pena (✉) • J. Rodríguez
Department of Applied Mathematics, Faculty of Mathematics, Campus Vida s/n, 15782 Santiago de Compostela, Spain
e-mail: fran.pena@usc.es

A. Prieto
Department of Mathematics, Faculty of Computer Science, Campus Elviña s/n, 15071 A Coruña, Spain
e-mail: andres.prieto@udc.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_20

this figure will increase to 10 % by 2020, with many ongoing projects mainly in Europe, United States, China and Japan.

Offshore wind has a number of advantages compared to on land such as higher wind speeds and less turbulence, thus generating more energy from fewer turbines, and usually fewer environmental constraints. Offshore is particularly suitable for large scale developments near major demand centers represented by large coastal cities, avoiding the need for long transmission lines to bring the power to these demand centers, as is the usual case onshore.

The main areas for exploitation are however found far off the coast, in deep waters, where fixed supporting structures similar to the ones installed on land are no longer economical. These distant sites mean more difficult sea bottom operations and higher waves and thus floating platforms are more suitable in these conditions.

Floating platform designs were initially conceived for Oil and Gas industry operations. Therefore these designs were associated to huge safety factors due to implications to human safety and to the environment of the failure of such installations. But floating wind requirements are completely different and thus the major challenge for offshore wind development today is to continue to bring down costs, developing designs aimed at minimizing the capital expenditure and operating expenses while guaranteeing structural integrity and providing suitable operating conditions for the turbine.

For a formal optimization cycle to achieve the optimal design that minimizes the cost of produced energy (balancing the produced power and the expenses), it is first necessary to develop an analysis chain able to:

- Provide the response of the set platform-tower-turbine to different load scenarios (wind and waves spectra) in a fast and rigorous way.
- Robustly handle changes in design variables.

The response analysis of the whole set to different wave and wind scenarios is usually quite complex. The basis of floating structures optimization can be consulted in [1–3, 7]. Unfortunately these tools are not widely available for the companies working in this field. With this aim, we have developed a software package that automatically analyzes the feasibility of a floating structure. The structure of the platform is defined according to a very general set of parameters, allowing us to consider a wide range of designs. The package calls some commercial applications and some own codes, to complete the analysis process. The main steps are:

- Generation of a CAD file of the floating structure.
- Estimation of material cost for the whole structure.
- Calculation of the hydrostatic equilibrium position, subject to moorings and wind force at the top of the tower.
- Calculation of hydrodynamic pressure and RAOs (response-amplitude operators) considering moorings and wave interaction.
- Structural analysis of the platform, using the previous calculations.

This analysis tool can be used into a multi-objective optimization strategy. This can help us find optimal designs depending on the placement of future exploitation fields.

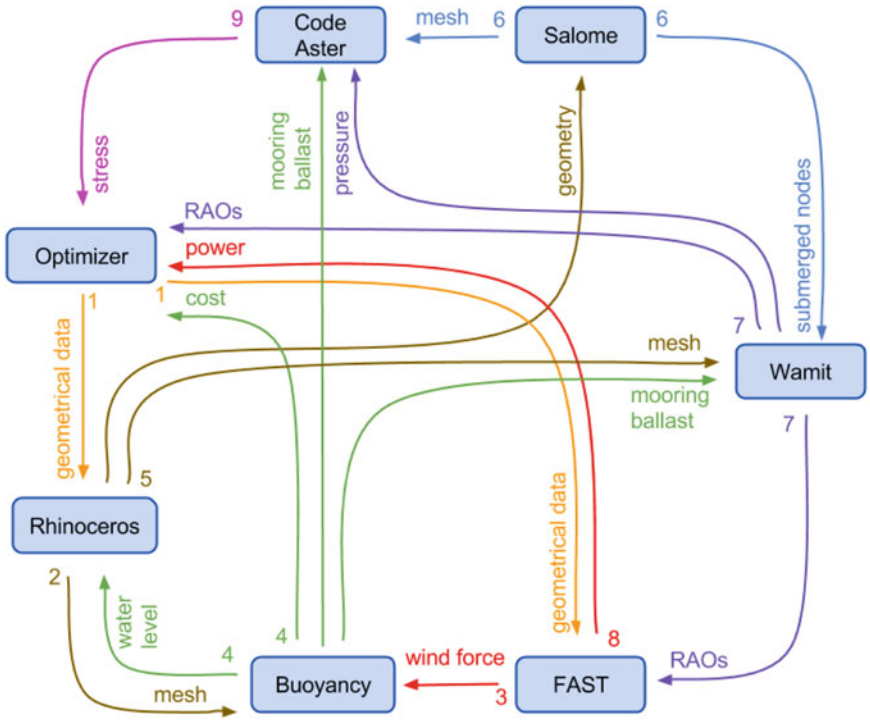


Fig. 1 Flow chart for the analyzer program: (*1*) process geometry data, (*2*) mesh for the buoyancy program, (*3*) wind force, (*4*) buoyancy program, (*5*) geometry in the equilibrium state, (*6*) conformal mesh for structural study, (*7*) RAOs and hydrostatic pressure, (*8*) power calculation and (*9*) structural study

Figure 1 shows a flowchart of the analyzer program. The rest of the paper is organized as follows: Sect. 2 describes how to encode geometry and to create a CAD model of the platform. Section 3 deals with the calculation of the equilibrium state. The numerical procedures to compute aerodynamic and hydrodynamic loads are described in Sect. 4. Section 5 details the structural study.

2 Geometry Encoding

In order to consider a geometry encoding flexible enough to model a wide variety platform designs, we assume that platforms are mainly composed of empty bodies made of metal sheets, that is, their internal structures are neglected. We distinguish between three type of objects: pillars, connectors and towers.

- Pillars are bodies of cylindrical section that give the platform the ability to float. They can have a rectangular or elliptical base and their dimensions and position

in the space are parametrized, as well as their lateral profile, thickness and anchor points. If they contain water acting as ballast, water height is also a parameter.
- Connectors have also cylindrical section; their geometry is parametrized in the same way than pillars. They can connect pillars or other connectors; contact points with the connected objects are also parameters.
- Towers are cylindrical objects on the top of some pillars; they are intended to hold wind generators.

The previous information is stored in a file keeping the structure of the three object types. Thus, it is very natural not only to change a specific parameter in the file, but also to remove or include a complete pillar, connector or tower. Such actions are among the first rules that an optimization algorithm based on grammatical evolution could need to be implemented (see [9]).

The first step in the analyzer program is to create a CAD model of the platform from the geometry encoding. To this end, a Python script was programmed to take advantage of the Python scripting for Rhinoceros [10]. The resulting geometry is composed of NURBS surfaces that can be exported in several formats. Figure 2 shows the resulting CAD file for the a semisubmersible platform designed by Mitsui Engineering and Shipbuilding Co. [4].

3 Buoyancy Position

The hydrodynamic behavior of the platform is modelled with WAMIT [8], which assumes that the structure is given in the equilibrium state. We have implemented the calculation of such equilibrium state for a rigid body subjected to its weight, buoyancy forces, moorings, wind forces applied at the top of the tower and ballasts. We remark that the movement of a rigid body can be decomposed into the movement of the center of mass and the movement induced by the rotation respect to the center of mass. Besides, the total force applied to the body produces a change in the linear moment, whereas the total moment respect to the center of mass changes the angular moment of the body. When the body is balanced, both linear and angular moments are null as well as both the sum of forces and the sum of moments.

To find out the equilibrium state requires to solve a nonlinear system: the condition of the vertical alignment of the center of mass and the buoyancy center gives two equations; the balance between total forces and weight gives another one. Among all possible solutions, only those which are stable are relevant. A position is stable when the body recovers its position subjected to small perturbations. To calculate the stable positions, the time-dependent dynamic problem is solved, integrating the equations of the rigid body with frictional force in a time interval long enough.

We assume that moorings are composed of chains or cables that partially lay on the seabed. They are modeled with a nonlinear uni-element model based on catenary (see [11]). Both flexural rigidity and friction with seabed are neglected.

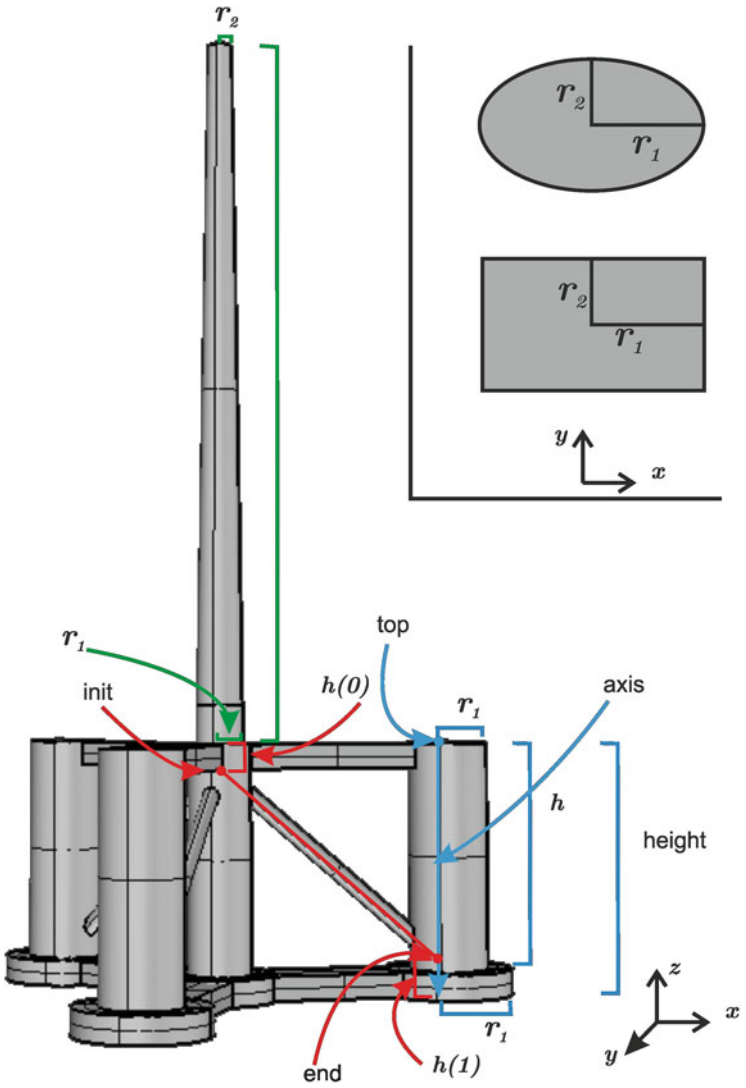


Fig. 2 Example of a semisubmersible platform designed by Mitsui Engineering and Shipbuilding Co. Some of the geometry parameters are detailed: pillars in *blue*, connectors in *red* and towers in *green*

4 Hydrodynamic and Aerodynamic Modelling

The hydrodynamic interaction between surface waves and the platform has been computed using the software package WAMIT [8]. Its implemented model is based on a linear model where a potential representation is applied to the fluid velocity field. Once this potential is split taking into account the radiation and diffraction contributions, the hydrodynamic loads on the wetted body surface are computed. The numerical solution involves the discretization of an integral equation whose Green function satisfies the free-surface boundary condition. The high-order implementation of this numerical method, the so-called *panel method*, represents the surface body geometry by means of continuous B-splines. This geometric setting is accomplished since NURBS surfaces are approximated by B-splines when the original structure representation is exported from Rhinoceros.

These numerical simulations allow to evaluate physical quantities such as the total force and total moment acting on the rigid solid and also fluid fields (pressure, velocity, and free-surface elevation). However, only the RAOs and the hydrodynamic pressure computed by WAMIT are relevant for our analyzer. The six RAOs are transfer functions associated to each degree of freedom (DOF) of the platform motion. They depend on the heading angle and the frequency of the incident plane-wave excitations.

For the aerodynamic modelling, the software package FAST [6] has been used to compute forces and moments induced by wind at the top of the tower. Since only static wind loads have been considered, only the module Aerodyn was used. This computational code requires two kind of input data: those ones related to the platform dynamics (such as the turbine configuration, its weight, characteristics of its mechanical components, the tower dimensions, its vibration modes, etc.), and those data related to the aerodynamic setting, which include the physical parameters of air, wind speed and direction, airfoil profiles and blade configuration. This code has been used twice in the analysis process (see Fig. 1): Firstly, forces and moments are computed at the equilibrium position of the platform, which have been taken into account to determine the buoyancy position; Second, for each frequency considered, the aerodynamic forces and moments are computed and used as input data in the structural analysis performed by Code_Aster [5].

5 Structural Study

The structural analysis is done using Code_Aster [5], a finite element code which includes a wide variety of mathematical models. Pillars are modeled as shells, while connectors and towers are modeled as beams with a shell transition at the end. Since Code_Aster can be executed through Python scripts, it is suitable to be integrated in the analyzer.

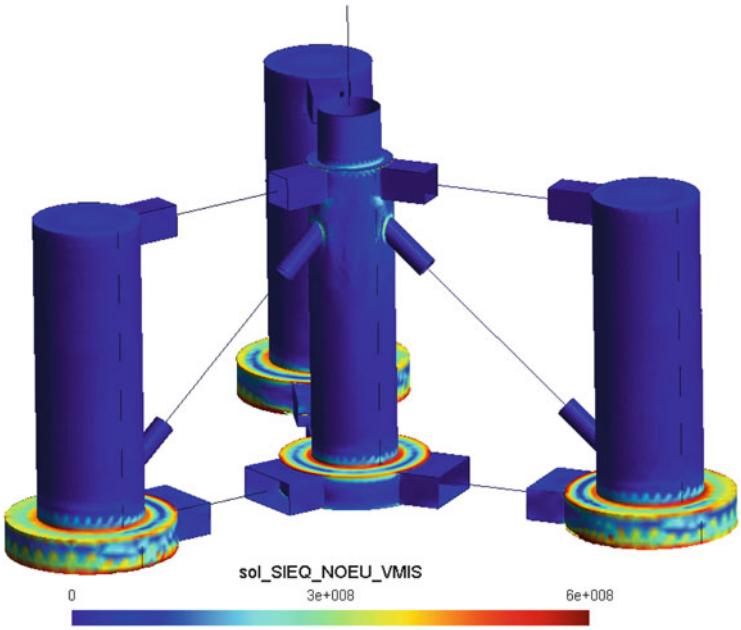


Fig. 3 Von Mises norm of the stress calculated with Code_Aster

Mesh produced by Rhinoceros is not a conformal one; we created an appropriate mesh from the CAD file using SALOME. The forces and loads calculated through the process are provided to Code_Aster to define two problems:

- The dynamic problem considers time dependent forces and it is solved in the frequency domain. Its solution is added to the solution of the next problem.
- The static problem takes into account the hydrostatic pressure, moorings and the wind force. This is a pure Neumann problem because static forces are balanced and there are no fixed nodes. In order to remove rigid movements and to have a well posed problem, assembly matrices are modified and transferred to the external solver UMFPACK. The result is injected again in Code_Aster.

The resulting stress is processed to detect critical values in the structure (see Fig. 3).

6 Conclusions

- The increasing development of offshore wind power requires tools to evaluate the validity of the platforms.
- An analyzer program is presented in this paper, being the result of a project conducted by the Repsol Technology Center.

- This program combines commercial and specifically developed software to calculate power, RAOs and stress for each structure.
- Inputs and outputs have been designed to easily integrate analyzer in an optimization code.

Acknowledgements Authors would like to thank the support of Repsol in this project and, in particular, the enthusiasm showed by Rosana Plaza Baonza and Jesús García San Luis. We want to recognize the work of Ibán Constenla Rozados in the early part of the project. Finally, we appreciate the opportunity given by the math-in network of presenting this paper in ECMI 2014.

References

1. Aubault, A., Cermelli, C., Roddier, D.: Parametric optimization of a semi-submergible platform with heave plates. In: Proceedings of the 26th International Conference on Offshore Mechanics and Arctic Engineering: OMAE2007 (2007).
2. Birk, L., Clauss, G.F.: Automated hull optimisation of offshore structures based on rational seakeeping criteria. In: Proceedings of the Eleventh (2001) International Offshore and Polar Engineering Conference Stavanger, Norway, pp. 382–389 (2001).
3. Birk, L., Clauss, G.F.: Optimization of offshore structures based on linear analysis of wave-body interaction. In: ASME Conference Proceedings, pp. 275–289. ASME, Estoril (2008).
4. Bossler, A.: Japan's floating offshore wind projects: an overview. Maine Ocean and Wind Industry Initiative: MOWII Webinar (May 2013).
5. Code_Aster: analysis of structures and thermomechanics for studies and research. http://www.code-aster.org (2014)
6. Jonkman, J.: FAST - NWTC Computer-Aided Engineering Tools. http://wind.nrel.gov/designcodes/simulators/fast/. Last modified 28 Oct 2013. Accessed 30 Mar 2014
7. Lee, J.Y., Clauss, G.F.: Automated development of floating offshore structures in deepwater with verified global performances by coupled analysis. In: The International Society of Offshore and Polar Engineers (ISOPE-1-07-208) (2007)
8. Lee, C.H., Newman, N.: WAMIT - Wave Analysis at MIT. User manual version 7.0. http://www.wamit.com (2013)
9. O'Neill, M., McDermott, J., Swafford, J.M., Byrne, J., Hemberg, E., Brabazon, A., Shotton, E., McNally, C., Hemberg, M.: Evolutionary design using grammatical evolution and shape grammars: designing a shelter. Int. J. Des. Eng. **3**(1), 4–24 (2010)
10. Rhinoceros: modeling tools for designers. http://www.rhino3d.com/ (2014)
11. Sheng, W., Lewis, T., Alcorn, R.: Numerical investigation into hydrodynamics of moored floating wave energy converters. In: Proceedings of the 9th European Wave and Tidal Energy Conference (EWTEC) (2011)

Math-in: A Structure Created to Improve the Transfer of Mathematical Technology to Industry

G. Parente and P. Quintela

Abstract Since 2007, a group of Spanish mathematicians has been promoting mathematical knowledge transfer through the Ingenio Mathematica (i-MATH) Project's 'Consulting Platform'. The outcome of this groundwork has been the creation of the national network, **math-in**. In this paper we will briefly present its aims, management, main activities and workspace.

Keywords EU-maths-in • Network of industrial mathematics

1 The Organization

The Spanish Network for Mathematics and Industry (**math-in**)[1] is a private non-profit organization focused on transferring mathematical technology to businesses and industrial sectors, thus stimulating competitiveness not only within the research groups involved, but industry itself. It arises as the result of the work of a group of Spanish mathematicians who decided to take on the challenge of industrial mathematics, by proposing a new means of mathematical knowledge in which the role of researchers would be fully proactive.

One of the key differentiating features of **math-in** is its network structure that facilitates access to nearly 40 research groups, which include 428 highly skilled researchers spread throughout Spain, for companies. Research activities within the groups are aimed at specific issues of their own areas of expertise, paying particular attention to development and innovation in companies. **Math-in** has created its own corporate image to be easily identified with (see Fig. 1).

[1] www.math-in.net.

G. Parente
Spanish Network for Mathematics and Industry (math-in), Santiago de Compostela, Spain

P. Quintela (✉)
Spanish Network for Mathematics and Industry, Santiago de Compostela, Spain
e-mail: peregrina.quintela@math-in.net

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_21

Fig. 1 **Math-in**'s corporate image

According to **math-in**'s principles, its aims in the field of Mathematics to industry are:

- Promote and facilitate strategic relationships with researchers.
- Increase the presence of mathematical methods and techniques in the productive sector by encouraging the participation of researchers in collaborative strategic projects.
- Realize the potential of existing knowledge via training.
- Facilitate the internationalization of research results by promoting partnerships with other entities through R and D projects.
- Promote and lead collaborative projects of national and international interest.
- Ensure competitive advantage of researchers through the registration and exploitation of their research results.
- Create a favorable environment for the creation of technology-based companies arising from research results.
- Reinforce the confidence and the interest of the industry in the mathematical community.
- Strengthen the technological image of the mathematical community in Spain.

math-in main aim is to provide solutions and transfer mathematical technology to the productive sectors of society, by introducing innovations and improvements using the most demanded mathematical technologies nowadays.

1.1 Partnerships

Thirty seven research groups in industrial mathematics spread throughout Spain, one legal entity which represents research groups and two sponsors constitute the **math-in** network on June, 2014. Members pay an annual fee to be part of this network.

The research groups vary in size and number, but are linked together by their common interest in research. Each group has a coordinator who acts as an interlocutor with **math-in**.

The General Assembly and the Board of Directors constitute **math-in** governing bodies.

The Board of Directors manages and represents **math-in**. It is composed of eight members, selected by the General Assembly, who belong to the different research groups. Its mandate is for 4 years, changing half of its members every 2 years.

All members in the Board of Directors are appointed by the General Assembly following elections.

The General Assembly is the main governing body and comprises 40 Spanish research groups. Each research group has a vote in the assembly regardless of the number of its members. The following research groups belong to the **math-in** network:

- A2M_ICMAT | Mathematical Analysis Applications, Institute of Mathematical Sciences.
- D-Time | Development and Transfer of Mathematical Innovation in Business, University of Almería.
- DEMACOM | Challenges in Computational Mathematics, University of Sevilla.
- EDA_ICMAT | Differential Equations and Applications, Institute of Mathematical Sciences.
- EDANYA | Differential Equations, Numerical Analysis and Applications, University of Málaga.
- EDNL | Non Linear Differential Equations, University of Santiago de Compostela.
- EIO_ICMAT | Statistics and Operations Research, Institute of Mathematical Sciences.
- EOPT | Statistics and Optimization, University of the País Vasco.
- GATNA_ICMAT | Algebraic Geometry, Number Theory and Applications, Institute of Mathematical Sciences.
- GEUVA | UVa Statistical Applications Group, University of Valladolid.
- GIOPTIM | Optimization, University of Sevilla.
- GNOM | Numerical Optimization and Modelling, Polytechnic University of Catalunya.
- GRID[ECMB] | Interdisciplinary Group in Statistics, Computing, Medicine and Biology, University of Santiago de Compostela.
- GSC | Simulation and Control Group, University of Vigo.
- GSO | Optimization Solutions Group, University of Valladolid.
- INFERES | Statistical Inference, Decision and Operations Research, University of Vigo.
- INTERTECH | Interdisciplinary Modelling Group, Polytechnic University of Valencia.
- LOGRO | Location, University of Sevilla.
- M2NICA | Numerical Models and Methods in Engineering and Applied Sciences, University of A Coruña.
- M2S2M | Mathematical Modelling and Simulation of Environmental Systems, University of Sevilla.
- M3A | Mathematical Modelling with Multidisciplinary Applications, Basque Center for Applied Mathematics.
- MAI | Differential Equations and Numerical Simulation Group, University of Vigo.
- mat+i | Mathematical Engineering, University of Santiago de Compostela.
- MathCUD | Interdisciplinary Group of Mathematics, University Center for the Defense of Zaragoza.

- MCS-UAB | Mathematical Consulting Service, Autonomous University of Barcelona.
- MGC_ICMAT | Geometric Mechanics, Control and Applications, Institute of Mathematical Sciences.
- MODES | Modelling and Statistical Inference, University of A Coruña.
- MODESTYA | Optimization Modelling, Decision, Statistics and Applications, University of Santiago de Compostela.
- MOSISOLID | Mathematical Modelling and Numerical Simulation in Solid Mechanics, University of Santiago de Compostela.
- OPTECO | Multicriteria Optimization and Econometric Modelling applied to the Socio-Economic Sphere, University of Málaga.
- PSYCOTRIP | Programming and Symbolic Computation, University of La Rioja.
- RiTO | Risk, Time and Optimization, Rey Juan Carlos University.
- SCT-CRM | Consultancy and Transfer Service, Centre de Recerca Matemática.
- SOR | Stochastic and Operations Research, Basque Center for Applied Mathematics.
- TAMI | Processing and Mathematical Analysis of Digital Images, University of Les Illes Balears.
- TEBADM | Bayesian and Decision Statistical Techniques in Economy and Enterprise, University of Las Palmas de Gran Canaria.
- TTM | Mathematical Technology Transfer, University of the País Vasco.

Another type of partner is the legal entity:

- ITMATI | Technological Institute for Industrial Mathematics

which represents nine partner groups, and finally the following companies as sponsors:

- BSH | BSH Electrodomésticos España, S.A.
- REPSOL | Repsol, S.A.

Math-in has signed (or is in process to signing) agreements with the 20 universities or research centers which legally support the procedures between **math-in** and its partners. Furthermore, the network has also signed (or is in process of signing) a framework agreements with the companies that are partners of **math-in**.

2 Management

The **math-in** Transfer Office is located in Santiago de Compostela. This office is the point of contact of all its members and is also in charge of improving the relationships between research groups and companies. The **math-in** website, also

managed from the office in Santiago, contains information classified in 23 sectors of economic activity and linked to one of the following icons:

The **math-in** website contains information about the groups expertise in industrial contracts, partially supported by companies research projects, training courses for technicians, experience in the use of free or commercial software and software development for companies. Also in the website there is extensive information about the masters on the subject of Industrial Mathematics in which some of the **math-in** groups are involved.

An important feature of the website is its public on-line database which is an easy way to look for information on research groups, research lines, projects, contracts, training courses, software by industrial sector and keywords.

The Transfer Office handles all international relations with institutions which deal with mathematical technology transfer to the productive sector. As such, **math-in** is the Spanish node of the European network EU-MATHS-IN. Other nodes of this Network are: AMIES (France), EU-MATHS-IN.se (Sweden), HU-MATHS-IN (Hungary), IMNA (Austria), KoMSO (Germany), PL-MATHS-IN (Poland), PWN (The Netherlands), SM[i]^2 (Italy), and The Smith Institute (UK). In addition, **math-in** belongs to the European Consortium for Mathematics in Industry (ECMI) since 2012.

The Transfer Office is also a single point of contact for companies. In this sense, **math-in** aims to improve the industrial processes and needs of companies by always selecting the most appropriate research groups for each industrial challenge.

3 Areas of Technology Transfer

Math-in network provides all business sectors and Administrations with a wide range of technological solutions based on the application of Mathematics. The services provided by **math-in** can be grouped into three major areas:

- Computer-Aided Engineering (CAD/CAE)

  Computer-Aided Engineering (CAE) uses the results obtained in computer-aided design (e.g. parts, plans, images or graphics design) and in calculation computer programmes to simulate, predict or study the performance of products or processes (e.g. for thermic, mechanical stress, manufacturing process studies, etc.), thus achieving significant improvement in cost, time and, in general, the monitor research, development and innovation processes.

  Computer-Aided Engineering is applied to all types of fields: mechanical or structural, thermal or thermodynamic, manufacturing processes (injection, stamping, forging, etc.), electronic and electromagnetic, fluid (gases and liquids), acoustic or vibroacoustic, environmental, multiphysics, etc.

  Within the CAD/CAE field, **math-in** meets all companies needs concerning information or advice on its possible applicability, selection, initial implementation or validation of the tools to be used, training, definition or calculation of the processes to be improved, development of customized software or interfaces between programmes, etc.
- Statistical, data analysis or decision support techniques

  These mathematical techniques include different methods used such as to improve customer analysis, markets, products, quality, planning, risks, logistics, allocation or optimization of resources and processes, etc.

  These methods cover business needs in several areas: quality control; stock control and optimization, manufacturing process control and optimization; risk or financial product analysis; business strategy, decision, logistics and planning; customer analysis, and market research or product studies; exploitation of internal information (data mining, business intelligence); design of experiments; clinical trials, etc.
- Other mathematical techniques

  Several mathematical techniques can be applied to areas such as geographical location; image or signal processing; geometry, design or visualization; bio-informatics or biomathematics; search and coding information or computing.

  Applying such techniques provides solutions in several areas: digital image processing (graphics, video, animation, image recognition); geometric analysis (computational geometry, visualization, CAD development, symbolic methods); digital signal processing; design of geographic information systems such as GIS or GPS; communication networks; information coding, cryptography and computer security; computing, computer algebra; language processors; symbolic and numeric algorithms; information and knowledge processing and search

(semantic web, algorithms for Internet); bio-informatics, genomic and proteomic, biomathematics (applications in life and health sciences such as diagnostic techniques, medical prescription, drug administration, growth and spread of diseases, pest control, systems biology), etc. (See [1–3]).

References

1. Quintela, P., González, W., Alonso, M.T., Ginzo, M.J., López, M.: i-MATH Map of Company Demand for Mathematical Technology. TransMATH. Nino-Centro de Impresión Digital, Santiago de Compostela (2010)
2. Quintela, P., Fernández, A.B., Martínez, A., Parente, G., Sánchez, M.T.: TransMath. Innovative Solutions from Mathematical Technology. Springer, New York (2012)
3. Quintela, P., Parente, G., Sánchez, M.T., Fernández, A.B.: Soluciones Matemáticas Para Empresas Innovadoras. Catálogo de Servicios Ofertados Por Investigadores Españoles. McGraw-Hill, New York (2012)

On the Italian Network of Industrial Mathematics and Its Future Developments: Sportello Matematico per l'Industria Italiana

Michiel Bertsch, Maurizio Ceseri, Roberto Natalini, Mario Santoro, Antonino Sgalambro, and Francesco Visconti

Abstract Sportello Matematico per l'Industria Italiana is a project developed by the National Research Council of Italy to build an effective and high-quality network of research groups in Industrial Mathematics in Italy. Here we will recall the objectives and the main actions taken by the project team during its first year of activities.

Keywords EU-maths-in • Network of industrial mathematics

1 Introduction

Mathematics is a key enabling factor for scientific and industrial innovation since it offers flexible, cheap, and highly reliable techniques [6–8]. Mathematical skills will be more and more important to innovate production systems and to face the hard challenges set by the international competitiveness. Mathematics is a driving force of economic growth, and this is why the efforts to introduce mathematical research into the industry have grown worldwide in the last 20 years; they have been sustained by governments in conjunction with Universities and Applied Mathematical Societies such as SIAM in the USA and ECMI in Europe. In North America, there exists a strong cooperation between public research bodies and the productive sector. In Europe, institutions such as AMIES (France), Matheon (Germany), MATH-IN (Spain), and Smith Institute (Great Britain) offer mathematical consultancy for innovation to private firms [1]. These cooperations between Applied Mathematics and Industry have generally been fostered by different kind of actions. Master degree and PhD programs or PostDoctoral Fellowships with an application focus are organized by Universities in collaboration with an industrial partner (that partly funds the program and supervises the research). Modeling weeks, Study Groups and internships are other way to foster industry academia contacts. The objective

M. Bertsch • M. Ceseri • R. Natalini • M. Santoro • A. Sgalambro (✉) • F. Visconti
Istituto per le Applicazioni del Calcolo "M. Picone", via dei Taurini 19, 00185 Roma, Italy
e-mail: m.bertsch@iac.cnr.it; m.ceseri@iac.cnr.it; r.natalini@iac.cnr.it; m.santoro@iac.cnr.it; a.sgalambro@iac.cnr.it; f.visconti@iac.cnr.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_22

is to train a professional figure called "Technology Translator" or Facilitator: a mathematician that can communicate with both University and Industry thanks to the so called "Soft Skills" [8]: business experience, team work, the ability to meet time constraints just to mention a few.

These efforts have started to pay off. A recent report commissioned by the Engineering and Physical Sciences Research Council (EPSRC) to Deloitte [2] states that Mathematical Research accounts for about 2.8 million jobs and £208 billion on Gross Value Added in UK. Similar figures are reported for The Netherlands [3].

In Italy, the situation has been way less vibrant so far, and this is caused by several factors (see for example [1] and references therein). One is the industrial landscape dominated by small enterprises: in 2010, Italian companies were 4,372,143 and 94.9 % of them employed less than 10 workers. A second factor is the poor investment in research and development: 1.26 % of Gross Domestic Product, well below the 3 % stated in the Lisbon Strategy [4]. Lack of cooperation between industry and academia is another reason: just 12.1 % of the Italian innovative enterprises are involved in some kind of collaboration (with other enterprises and/or Universities). However, Italian enterprises show a natural tendency to innovation: 56.3 % of companies has innovated their products and/or processes in 2010—a proportion that compares favorably with respect to the European average, 52.9 % [5]. The Italian Applied Mathematics community can easily catch this trend. This is why Sportello Matematico per l'Industria Italiana (Sportello Matematico or SM$[I]^2$ from now on) is born (http://www.sportellomatematico.it): to connect research groups and companies and activate collaborations between them towards innovation. There are several examples of success cases of Academia-Industry cooperation: the aim of Sportello Matematico is to make these collaborations systematic in the Italian landscape. In the remainder of this paper the main activities carried on by SM$[I]^2$ during its first year will be described.

2 Sportello Matematico per l'Industria Italiana

Sportello Matematico is a project of the Istituto per le Applicazioni del Calcolo "Mauro Picone" (IAC) of the National Research Council of Italy and has been funded by the Italian Ministry of Education, University and Research in 2012 for a starting period of 3 years. The project is run in collaboration with the Italian Society of Applied and Industrial Mathematics (SIMAI) and the Italian Association of Operations Research (AIRO).

The project has the following objectives:

1. to promote the industrial mathematics towards the productive sector;
2. to put into contact industries with Applied Mathematics Research Groups in order to deal with innovation problems requiring the use of mathematical models and numerical simulation tools;

3. to give industry, and especially SMEs, a unique and qualified center for consultancy in the field of applied mathematics;
4. to create an Italian network of excellence in the field of Industrial Mathematics and to integrate it in the European context;
5. to stimulate the future engagement of young mathematicians in private enterprises.

The first two objectives are clearly connected since enterprises are often not aware of what mathematics can offer to them in terms of product innovation and process optimization. On the other hand, research groups have the tendency to present their job and skills in a too much technical way; this gives private enterprises the impression that mathematics is useless for their practical problems. The team of Sportello Matematico addressed them by simplifying the message towards the industry on the real benefit of mathematics to product innovation and process optimization through a capillary and intense promotional activity: a marketing campaign of Mathematics towards the productive sector based on a high number of contacts among companies.

Even in case an enterprise wants to collaborate with the Industrial Mathematics Community, it often does not know which research group is suitable for its needs. This issue is the focus of the third and fourth objectives. They aim at creating a network of research groups acting as a *one-stop-shop* for industrial mathematics: a company can address its problems to this structure; the structure will in turn support the company in finding the qualified research group to deal with their specific needs. Such an activity is particularly important for SMEs that produce most part of the innovative technology but have few contacts with academia. The network shall interact with other European Institutes sharing the same mission to coordinate the efforts of each Country in the Continent, especially in the framework of Horizon 2020 program.

Finally, the last objective is known to be one of the major way to foster connections between Industry and Academia: a young mathematician from a Master or PhD program remains likely in contact with his former research group.

Given the above challenges, it came naturally that Sportello Matematico was founded as an initiative of the Istituto per le Applicazioni del Calcolo—the first Institute fully devoted to Applied Mathematics ever created in the world. Mauro Picone (1885–1977) founded IAC in 1927 to promote the interactions between Applied Mathematics and Industry (http://www.iac.cnr.it). His vision was based on the strong conviction that coupling Mathematical abstraction and simulation tools is successful to solve real life problems and, thus, to the advance of both Industry and Society. The strategic project of SM$[I]^2$ is a strong bid to carry on Picone's intuition while updating it to present days.

To promote this ideal, Sportello Matematico quickly realized that it was necessary to turn to three actors: the Industrial Mathematics community, the world of enterprises and the young mathematicians.

2.1 Towards the Research Community in Applied Mathematics

Sportello Matematico promoted a partnership with a growing number of Italian research centers in Applied Mathematics and Operations Research, all of them sharing a common interest towards concrete collaborations with the industrial sector. About thirty high-quality research groups in Industrial Mathematics have been already involved in the SM$[I]^2$ project as partners: they represent the real core of the Industrial Mathematics network described above and will be involved in collaborations with Italian enterprises through the intermediary services of SM$[I]^2$. Once an enterprise contacts Sportello Matematico about a business issue, the team translates this problem in a request of service and forward it to the network. The interested research groups communicate their potential willing to collaborate with the company and are involved in a technical meeting with the enterprise. After the meeting, the interested research groups send to the enterprise their individual offers of service detailing the project, the timeline and the economic plan. Finally, the enterprise will evaluate the received proposals and choose the offer that considers the most suitable, eventually giving rise to a collaboration with the related research group. To provide visibility to each group in the network, SM$[I]^2$ team developed a questionnaire for its partners to collect success stories of collaboration with Italian Enterprises. Such stories are summarized according to the following scheme: the industrial problem, the scientific approach and the final benefits for the company involved in the collaboration.

2.2 Towards the Productive Sector

Most part of the work of SM$[I]^2$ has been to contact industries possibly interested in a collaboration with the Italian Applied Mathematics Community. To this aim, Sportello Matematico participates at events where it can meet enterprises. For example, the team has been involved in a series of events organized together by National Research Council and Confindustria—the Italian Industry Association. They were organized in different Italian cities with the participation of several Italian enterprises: at any workshop, SM$[I]^2$ presented its activities and a list of success stories from its network of scientific partners, selected to match the specific interests of the present companies. The latter were successively contacted by SM$[I]^2$ and many of them showed an interest in the activities of the project: following these contacts, a few meetings have been arranged. Although, most of these companies are not fully aware of the potential benefits arising from Industrial Mathematics, a lot of work can be done in this sense and in the long run thanks to the action of the Sportello Matematico project.

A survey for the productive sector has been developed by SM$[I]^2$. The survey collects information on the activities of the firms and their past collaboration with research centers: such information will be useful to have a first idea on the needs

of the company—especially if this company requests a meeting with the team of Sportello Matematico—and to improve the communication strategy. Furthermore, this is another way to promote the activities of the network. Indeed, such a survey will be at the center of a marketing campaign: a sample of Italian enterprises will be chosen and contacted to fill the survey. From this campaign, SM$[I]^2$ expects to find a number of contacts that can be translated in active cooperations. Finally, survey's results will be included in future reports about the state of the art of Industry-Academic cooperation in the field of Industrial Mathematics in Italy.

2.3 Towards Young Mathematicians

The last objective of SM$[I]^2$ concerns the employment of young mathematicians. Being Mathematics a driving force for innovation, a simple way for an enterprise to innovate its processes and/or products is to hire young mathematicians. This is important for cooperation purposes as well: mathematicians can play a relevant role in the future contacts with research centers. In order to make more systematic the connections between Industry and Applied Mathematics research groups, this objective becomes crucial. SM$[I]^2$ developed a survey for young mathematicians with questions about their Academic degrees, their mathematical expertise, their past experiences with enterprises and so on; finally, they can submit their whole CV as well and express their willingness to be contacted by the team of Sportello Matematico in case a possible matching with a job is identified.

2.4 Sportello Matematico in Europe

Sportello Matematico is in contact with other organizations with similar mission and interests in order to share experience and improve the quality and effectiveness of its activities. Collaborations with other European entities have started during SM$[I]^2$ first year of activity and will be further developed in the participation in EU-MATHS-IN.

In November 2013, the European network of networks in Industrial Mathematics EU-MATHS-IN has been founded (http://www.eu-maths-in.eu). It currently collects major Industrial Mathematics networks from ten Countries: AMIES (France), HU-MATHS-IN (Hungary), IMNA (Austria), KoMSO (Germany), EU-MATHS-IN.se (Sweden), MATH-IN (Spain), PWN (The Netherlands), PL-MATHS-IN (Poland), Smith Institute for Industrial Mathematics and System Engineering (UK) and SM$[I]^2$ (Italy). The European Mathematical Society (EMS) and the European Consortium for Mathematics in Industry (ECMI) promoted the foundation of EU-MATHS-IN and Sportello Matematico joined the project from the beginning as Italian member of this network of networks. EU-MATHS-IN was born in the wake of the experience reported in "Success Stories in Industrial Mathematics" [7]: to

better coordinate the efforts of the National networks in the whole Continent. Moreover, it will have a crucial role in the Horizon 2020 framework program to facilitate the formation of transnational partnerships among research centers in Industrial Mathematics and private companies.

References

1. Bertsch, M., Ceseri, M., Felici, G., Natalini, R., Santoro, M., Sgalambro, A., Visconti, F.: Mathematical desk for Italian industry: an applied and industrial mathematics project. Proc. Soc. Behav. Sci. **108**(0), 79–95 (2014). doi:http://dx.doi.org/10.1016/j.sbspro.2013.12.822
2. DELOITTE: Measuring the Economic Benefits of Mathematical Sciences Research in the UK. http://www.epsrc.ac.uk/newsevents/news/2012/Pages/mathsciresearch.aspx (2012)
3. DELOITTE: Mathematical sciences and their value for the Dutch economy. http://www.eu-maths-in.eu/download/generalReports/20140115%20Mathematical%20sciences%20v6%20Web.pdf (2014)
4. EUROSTAT: Science, technology and innovation. http://epp.eurostat.ec.europa.eu/portal/page/portal/science_technology_innovation/introduction (2012)
5. EUROSTAT: Innovation statistics. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Innovation_statistics (2013)
6. Lery, T., Primicerio, M., Esteban, M.J., Fontes, M., Maday, Y., Mehrmann, V., Quadros, G., Schilders, W., Schuppert, A., Tewkesbury, H.: Forward Look Mathematics and Industry. European Science Foundation, Strasbourg (2010)
7. Lery, T., Primicerio, M., Esteban, M.J., Fontes, M., Maday, Y., Mehrmann, V., Quadros, G., Schilders, W., Schuppert, A., Tewkesbury, H.: European Success Stories in Industrial Mathematics. Springer, New York (2011)
8. SIAM: SIAM report on mathematics in industry. http://www.siam.org/reports/mii/2012/ (2012)

Optimal Design of Solar Power Tower Systems

E. Carrizosa, C. Domínguez-Bravo, E. Fernández-Cara, and M. Quero

Abstract In this paper we review the recent research done by the authors in Solar Power Tower systems design focusing on the heliostat field problem. We first analyze the basic problem, in which all heliostats have the same size, as commonly addressed in the literature. A brief review of the problem itself and the pattern-free procedure proposed to solve it is given. The algorithm proposed, a greedy-based heuristic procedure, provides a new way to solve the problem different from previous algorithms in the literature. Our methodology consists of a pattern-free heliostat location and therefore it can be easily (even though carefully) adapted to solve other issues such as multi-size or multiple-receiver heliostat field.

The multi-size heliostat fields design is also reviewed given the similarities of the problem. This algorithm is tested using two different heliostat sizes. Some ideas about the application of the procedure to more general settings, such as multiple-receiver field, are given as further work.

Keywords Optimal design • Sustainable energy

E. Carrizosa
Department of Statistics and Operations Research, University of Seville, Sevilla, Spain
e-mail: ecarrizosa@us.es

C. Domínguez-Bravo (✉)
Institute of Mathematics, University of Seville, Sevilla, Spain
e-mail: carmenanadb@us.es

E. Fernández-Cara
Department of Differential Equations and Numerical Analysis, University of Seville, Sevilla, Spain
e-mail: cara@us.es

M. Quero
Abengoa Solar N.T., Seville, Spain
e-mail: manuel.quero@solar.abengoa.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_23

1 Introduction

Solar Power Tower (SPT) system is one of the most promising technologies for producing solar electricity because of the high thermodynamic performances reached, see review [8] and the references therein. Since much of this technology is recent, there is still room for improving designs and emerging concepts are often proposed and analyzed.

An SPT system is here considered of two elements: a tower-receiver and a field of heliostats, see [5]. Operation of the system is as follows: direct solar radiation is reflected by the heliostat field onto a receiver placed at the top of the tower. In the receiver, the thermal energy is transferred to a heat transfer fluid to produce electricity through a thermodynamic cycle. The heliostats field is considered of a group of rectangular mirrors having two-axis movement, they move around to follow the position of the sun in order to correctly reflect solar radiation.

In this paper we give a review on the heliostat field design problem, some variants and the procedures we have applied to address this problem when the tower-receiver variables are fixed. This is the critical part of the process, since, once the heliostat field is optimized, the full SPT optimization problem can be tackled by means of an alternating approach, as shown in [2].

The remainder of the paper is organized as follows. The heliostat field design problem and the heuristic algorithm to solve it are presented in Sect. 2. Section 3 explains our methodology to design multi-size heliostats fields. The last section is intended to present some results, the work in progress and perspectives for further work.

2 Heliostat Field Design

When addressing the heliostat field design problem two challenging issues appear, namely, the dimensionality of the optimization problem (with hundreds or thousands of variables a priori unknown), and the evaluation of the objective function (many local optima, hard to compute and no apparent mathematical structure which can help).

The process we study, presented in [2], with the goal to find an optimal heliostat field is different from others in the literature in three aspects: no initial field is needed, no parametric form is used (commonly used in the literature: radial-stagger [7], spiral [9], grid [10]) and no oversize procedure is applied.

2.1 Variables and Functions

We will assume that the receiver consists of a cylinder pointing to the North with circular aperture, as explained in [5]. As we have said we are going to focus on the heliostat field design problem, we suppose the tower-receiver parameters fixed. In what concerns the heliostat field, the variables to be used are the heliostats locations, given by the coordinates (x, y) of their centers. From now on we will denote by $\mathscr{S}$ the collection of coordinates of the centers of the heliostats. It is expressed as follows, where N denotes the total amount of heliostats: $\mathscr{S} = \{(x^1, y^1), (x^2, y^2), \ldots, (x^N, y^N)\}$, observe that N is also a decision variable.

Two criteria are taken into account for the optimization: the total investment cost and the generated annual energy. The cost function C takes into account the investment in equipment (tower, receiver and heliostats) and it depends on the number of heliostats in the field, as we can see in (1):

$$C(|\mathscr{S}|) = K + \Psi(|\mathscr{S}|), \tag{1}$$

where K is a constant including all fixed costs, $|\mathscr{S}|$ stands for the cardinality of $\mathscr{S}$ and Ψ is the linear heliostat cost function.

With this notation, the annual energy input function E generated by the plant takes the form:

$$E(\mathscr{S}) = \int_O^T \tilde{\Pi}_t(\mathscr{S})\, dt\,, \tag{2}$$

where the function $\tilde{\Pi}_t$ denotes the polynomial fitting of the power input reached by the plant at each time instant t. The power input values of the system at each time instant are calculated by adding the values reached by each heliostat as follows:

$$\Pi_t(\mathscr{S}) = I(t)\eta(t)f_{ref} \sum_{i=1}^{N} \varphi(t, x^i, y^i, \mathscr{S})\,. \tag{3}$$

Here $I(t)$ is the so-called instantaneous direct solar radiation, $\eta(t)$ measures the radiation losses, f_{ref} is the heliostat reflectance factor and φ is the product of the efficiency factors. The efficiency factors are detailed as follows: $\varphi = f_{at} f_{cos} f_{sb} f_{sp}$. In particular, f_{at} is the atmospheric efficiency, f_{cos} is the cosine efficiency; f_{sb} is the shadowing and blocking efficiency, and, finally f_{sp} is the interception efficiency or spillage factor. We refer the reader to [1, 4, 6] and the references therein, for further details.

2.2 Optimization Problem

As mentioned above, the two criteria involved are the total investment cost and the annual energy produced. No common optimum can be found for both criteria, so they are aggregated into one single objective, namely, the maximization of generated energy per unit cost $F(\mathscr{S})$. The $(\mathscr{P})$ problem, that is, the optimization of the heliostat field for a given tower-receiver, can be written as follows:

$$(\mathscr{P}) \quad \begin{cases} \max\limits_{\mathscr{S}} & F(\mathscr{S}) = E(\mathscr{S})/C(\mathscr{S}) \\ \text{subject to} & \Pi_{T_d}(\mathscr{S}) \geq \Pi_0 \\ & \mathscr{S} \subseteq \mathscr{S}_0 \\ & ||(x^i, y^i) - (x^j, y^j)|| \geq \delta \quad \text{for } i \neq j\,, \end{cases}$$

where the first constraint sets a minimal power Π_0 that has to be achieved at T_d. As explained in [4, 10], a fixed instant, named *design point* T_d, is used to size the SPT system. The second constraint defines the feasible region $\mathscr{S}_0$ where heliostats must be located. And finally, for the proper operation the heliostats, they have to rotate freely avoiding collisions consequently, we have to consider the last constraints, where $\delta > 0$ is the given safety distance.

2.3 Procedure: Greedy Algorithm

The *Greedy Algorithm*, presented in [2] and designed to solve problem $(\mathscr{P})$, operates as follows: firstly it locates the heliostats one by one at the best feasible position, that is, the location where the annual energy input is highest. Once a new heliostat is located and the shading and blocking effects are incorporated, the process is repeated until no improvement is reached. Only the two geometrical constraints have to be taken into account: the field shape constraint and safety distance constraints to avoid collisions.

The optimization problem of locating heliostat k when there are already located $k-1$ heliostats in the field is equivalent to maximizing the energy generated by the new heliostat location, because the cost function is fixed at each step as all the heliostats have the same size and cost in this problem. In order to solve the problem, since the functions involved are highly multimodal, and the output strongly depends on the starting points, a multistart procedure is used to avoid local minima. The heliostat position that reaches the best objective value will become part of the heliostat field solution. See Fig. 1 as a graphic example, two heliostat fields left-side (radial-stagger distribution) and right-side (*Greedy Algorithm*).

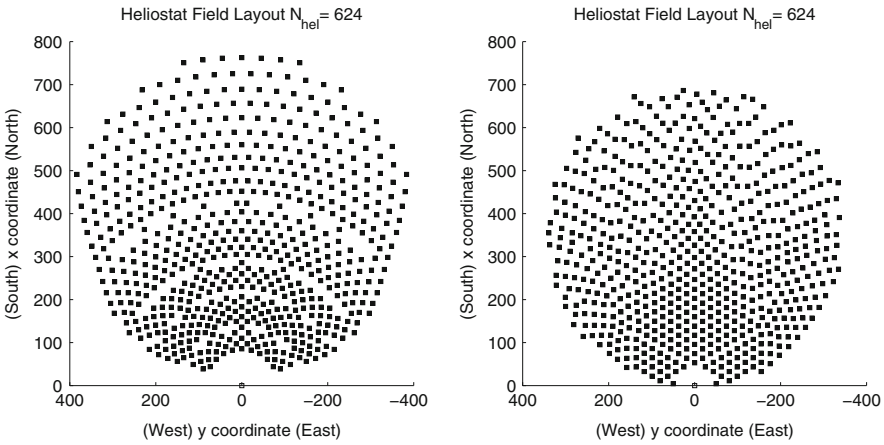


Fig. 1 Radial-Stagger and *Greedy Algorithm*

The annual energy is modified at each step as well as the shading and blocking effects that the new heliostat is causing in the field. This is the main reason of increase of the computing time. Given that the energy function is hard to compute [10], we approximate it by a much simpler function: Instead of computing E as in (2), the power input (3) at the design point T_d is used.

3 Multi-Size Heliostat Field Design

All the papers we are aware of address problem ($\mathscr{P}$) assuming all heliostats with identical size. However choosing all heliostats of one single size may not lead to optimal fields. We address problem ($\mathscr{P}$) using two heliostat sizes, called *large-size* and *small-size*, in [3]. The choice of the size of the heliostats may dramatically affect the performance of the field and also its cost.

The variables involved in the optimization problem are the same as in the previous section adding the heliostat size as a variable. We denote by d, the new variable to be included i.e. the heliostat size. In order to continue with the same notation the coordinates of the centers and sizes of the heliostats, denoted by (x, y, d), are the variables to be used and the collection of them is denoted by Ω. The set Ω is described as follows: $\Omega = \left\{(x^i, y^i, d^i) \text{ for } i \in [1, N] \text{ with } (x^i, y^i) \in \mathscr{S} \text{ and } d^i \in \mathscr{D}\right\}$, where $\mathscr{D}$ denotes the set of the possible sizes, large-size and small-size.

In order to properly calculate the solar efficiencies values the heliostat size must be carefully taken into account as a variable in this problem. The optimization

problem can be rewritten to collect all these changes as follows:

$$\left(\mathscr{P}^{\mathscr{D}}\right) \begin{cases} \max\limits_{\Omega} & F(\Omega) \\ \text{subject to} & \Pi_{T_{\mathrm{d}}}(\Omega) \geq \Pi_0 \\ & \Omega \subset \mathscr{S}_0 \times \mathscr{D} \\ & ||(x^i, y^i) - (x^j, y^j)|| \geq \delta_i + \delta_j \text{ for } i \neq j\,. \end{cases}$$

note that in this case the security distance must be dependant on the heliostat size in order to properly avoid collisions.

3.1 Procedure: Expansion-Contraction Algorithm

We briefly describe in this paragraph the heuristic algorithm to solve problem $\left(\mathscr{P}^{\mathscr{D}}\right)$, called *Expansion-Contraction Algorithm* and studied in [3]. The algorithm starts generating a large-size heliostat field following the *Greedy Algorithm* already explained in Sect. 2. Then two phases, *Expansion* and *Contraction*, are applied into this initial field and repeated until no improvement is obtained.

In the *Expansion Phase* the field is filled with small-size heliostats until a certain power input value $\Pi_0^+ > \Pi_0$ is reached. Small-size heliostats are more versatile, they are expected to fill-in holes between large-size heliostats and to reduce spillage losses, reaching higher energy values. Once the mixed-field is calculated, the heliostats are arranged according to their annual energy values per unit area. The best heliostats are sequentially selected in the *Contraction Phase* and the final number of heliostats of the mixed-field is given by Π_0.

As an example, see Fig. 2, where an intermediate stage of Expansion-Contraction process is shown.

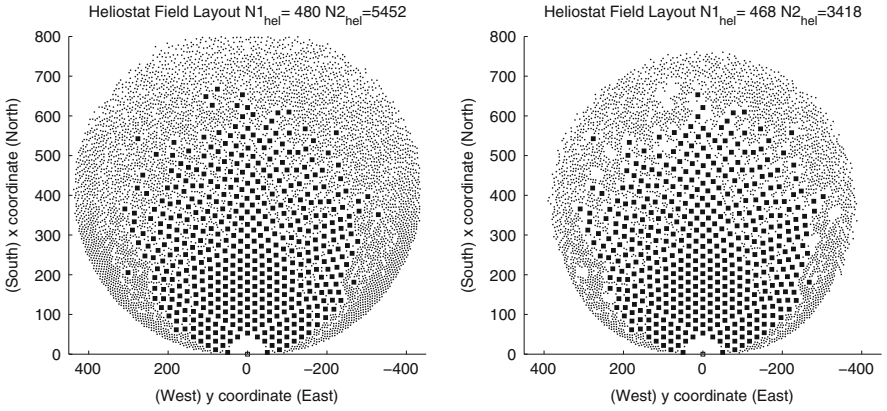


Fig. 2 *Expansion Phase* and *Contraction Phase*

4 Concluding Remarks and Extensions

In this paper we have reviewed our findings on heliostats fields design. Our *Greedy Algorithm* provides competitive results against the standard results in the literature as can be seen in [2] and is more versatile since it is not based on geometrical patterns which may be valid only under certain physical conditions. Moreover, our algorithm can be extended to address many other situations such as ground irregularities. Moreover, the procedure is extended to a Multi-Size heliostat field using the *Expansion-Contraction Algorithm*.

The results shown in [3] give better efficiency values than single-size heliostats fields. The numerical experiments show the advantages of combining heliostats of different sizes, if different prices apply, which is a very reasonable assumption.

Our approach can be extended to more challenging problems, now under research. One of them is the possibility of considering multiple receivers. Higher conversion efficiency of solar energy to electricity can be achieved only at high temperatures, and SPT systems with multiple receivers are required to achieve them. In *Multiple Receiver* systems heliostats are allow to be located all around the tower, because they can reflect the solar rays to one of the receivers at almost any position of the field.

When designing the heliostat field with multiple receivers, if we consider that the aiming strategy is fixed (heliostats aim the same receiver for all time instant) the main problems are the calculation of the feasible region for each receiver and the properly location of the heliostats in order to maximize the SPT annual efficiency. As our algorithm allows a pattern-free heliostat location, the feasible region associated with each receiver do not have any shape limitation. A polynomial fitting can be used to approximate the boundaries of each region taking into account the annual efficiency values obtained with the empty field. An iterative procedure is now being tested, see Fig. 3 as an example with three receivers.

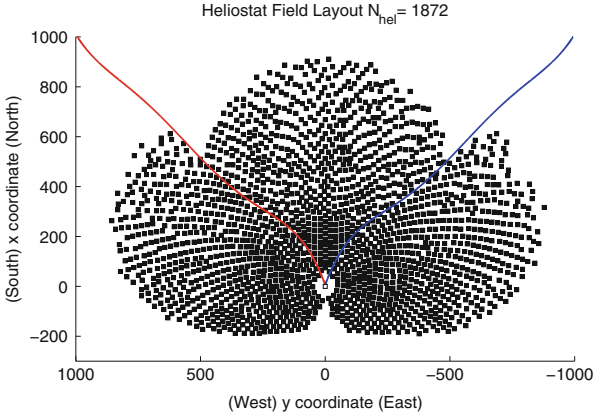


Fig. 3 Three-receivers field

Acknowledgements This research has been supported by Abengoa Solar N.T. and Institute of Mathematics of University of Seville (IMUS), through the research contract "CapTorSol". The authors would also like to acknowledge the support from the Government of Spain (Grants MTM2010-15992, MTM2012-36136), Andalucía (Grant P11-FQM-7603) and EU COST Action TD1207.

References

1. Biggs, F., Vittitoe, C.: The HELIOS model for the optical behavior of reflecting solar concentrators. Technical Report SAND76–0347, Sandia National Labs (1976). URL http://prod.sandia.gov/techlib/access-control.cgi/1976/760347.pdf
2. Carrizosa, E., Domínguez-Bravo, C., Fernández-Cara, E., Quero, M.: A global optimization approach to the design of solar power plants. Technical Report, IMUS (2013). URL \http://www.optimization-online.org/DB_FILE/2014/04/4305.pdf
3. Carrizosa, E., Domínguez-Bravo, C., Fernández-Cara, E., Quero, M.: An optimization approach to the design of multi-size heliostat fields. Technical Report, IMUS (2014). URL \http://www.optimization-online.org/DB_FILE/2014/05/4372.pdf
4. Collado, F.J.: Quick evaluation of the annual heliostat field efficiency. Sol. Energy **82**(4), 379–384 (2008)
5. Collado, F.J., Guallar, J.: A review of optimized design layouts for solar power tower plants with campo code. Renew. Sust. Energ. Rev. **20**, 142–154 (2013)
6. Crespo, L., Ramos, F.: NSPOC: a new powerful tool for heliostat field layout and receiver geometry optimizations. In: Proceedings of SolarPACES 2009 (2009)
7. Lipps, F.: Theory of cellwise optimization for solar central receiver systems. Technical Report SAND-85-8177, Houston University, TX, Energy Lab (1981). URL \http://www.osti.gov/scitech/biblio/5734792
8. Mills, D.R.: Advances in solar thermal electricity technology. Sol. Energy **76**, 19–31 (2004)
9. Noone, C.J., Torrilhon, M., Mitsos, A.: Heliostat field optimization: a new computationally efficient model and biomimetic layout. Sol. Energy **86**, 792–803 (2012)
10. Sánchez, M., Romero, M.: Methodology for generation of heliostat field layout in central receiver systems based on yearly normalized energy surfaces. Sol. Energy **80**(7), 861–874 (2006)

MS 6
MINISYMPOSIUM: EUROPEAN STUDY GROUPS WITH INDUSTRY

Organisers

Charles Holland[1] and Hilary Ockendon[2]

Speakers

William Lee[3]
Study Groups in Ireland

Andreas Münch[4]
Two Industry Workshops in Berlin

Poul Hjorth[5]
A Flying Corps of Applied Mathematicians

Stefka Dimova[6]
ESGI95: The First Study Group with Industry in Bulgaria

[1]Hilary Ockendon, University of Oxford, UK.
[2]Charles Holland, Office of Naval Research Global Prague, Czech Republic.
[3]William Lee, University of Limerick, Ireland.
[4]Andreas Münch, University of Oxford, UK.
[5]Poul Hjorth, Danish Technical University, Lyngby, Denmark.
[6]Stefka Dimova, Sofia University, Bulgaria.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_24

Adérito Araújo[7]
European Study Groups with Industry: The Portuguese Experience

Isabel Cristina Lopes[8]
A Mathematical Model for Supermarket Order Picking

Charles Holland[1](chair)
Discussion Session

Keyword

Study group

Short Description

In the year in which the 100th ESGI is held, it is timely to assess the effectiveness of this popular mechanism for facilitating the interaction of industry and academic mathematicians. In this minisymposium, six speakers describe their own experience of Study Groups. Some talks give an overview of how Study Groups work in a particular country and others describe one or two problems that have been successfully solved. The unique strength of mathematics as a technology transfer tool, where the same mathematics may apply in very different application areas, and the unreasonable effectiveness of a relatively simple mathematical model are highlighted. Finally, a discussion session was held which addressed the pros and cons of the Study Group concept.

[7] Adérito Araújo, University of Coimbra, Portugal.
[8] Isabel Cristina Lopez, Politecnico do Porto, Portugal.

A Mathematical Model for Supermarket Order Picking

Eliana Costa e Silva, Manuel Cruz, Isabel Cristina Lopes, and Ana Moura

Abstract Order picking consists in retrieving products from storage locations to satisfy independent orders from multiple customers. It is generally recognized as one of the most significant activities in a warehouse (Koster et al, Eur J Oper Res 182(2):481–501, 2007). In fact, order picking accounts up to 50 % (Frazelle, World-class warehousing and material handling. McGraw-Hill, New York, 2001) or even 80 % (Van den Berg, IIE Trans 31(8):751–762, 1999) of the total warehouse operating costs. The critical issue in today's business environment is to simultaneously reduce the cost and increase the speed of order picking. In this paper, we address the order picking process in one of the Portuguese largest companies in the grocery business. This problem was proposed at the 92nd European Study Group with Industry (ESGI92). In this setting, each operator steers a trolley on the shop floor in order to select items for multiple customers. The objective is to improve their grocery e-commerce and bring it up to the level of the best international practices. In particular, the company wants to improve the routing tasks in order to decrease distances. For this purpose, a mathematical model for a faster open shop picking

E. Costa e Silva
CIICESI/ESTGF-IPP - Center for Research and Innovation Business Sciences and Information Systems, School of Management and Technology of Felgueiras, Polytechnic of Porto, Felgueiras, Portugal
e-mail: eos@estgf.ipp.pt

M. Cruz
LEMA/ISEP/IPP - Mathematical Engineering Lab, School of Engineering, Polytechnic of Porto, Porto, Portugal
e-mail: mbc@isep.ipp.pt

I.C. Lopes (✉)
LEMA/CIEFGEI/ESEIG-IPP - Mathematical Engineering Lab, School of Management and Industrial Studies, Polytechnic of Porto, Porto, Portugal
e-mail: cristinalopes@eseig.ipp.pt

A. Moura
LEMA/ISEP/IPP - Mathematical Engineering Lab, School of Engineering, Polytechnic of Porto and CMUP - Center of Mathematics, University of Porto, Porto, Portugal
e-mail: aim@isep.ipp.pt

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_25

was developed. In this paper, we describe the problem, our proposed solution as well as some preliminary results and conclusions.

Keywords Order picking • Study group

1 Introduction

This paper addresses the problem of the order picking process on large grocery e-commerce business. This is a case study of one of the major players in the Portuguese grocery business, SonaeMC, which was presented to the 92nd European Study Group with Industry (ESGI92). It regards the optimization of the order picking process that is conducted in an open shop, to fulfill orders from online customers. The challenge proposed to the participants of the ESGI92 was to boost the efficiency in the picking rate in 10 %. Optimizing the order picking process for faster picking and 100 % accuracy (as a goal) is crucial to increase the competitiveness of the company and to preserve customers. This work includes several ideas held by the contributors of the ESGI92, as well as the modeling, implementation, some tests and improvements made afterwards by the authors of this paper.

Order picking may be defined as the process of retrieving products from storage in response to a specific customer request. It is generally recognized as one of the most significant activities in a warehouse. According to different authors, order picking accounts up to 50 % [5], or even 80 % [9], of the total warehouse operating costs. The picker's time distribution has been estimated to be around 50 % for traveling, 20 % for searching and 15 % for picking, with the remaining 15 % being divided between setups and other minor activities [7]. In this paper, we focus our efforts on the third and fourth of the following main issues of the order picking activity [2]:

1. Storage Assignment—The assignment of articles to storage locations.
2. Zoning—The establishment of work zones to which pickers are confined.
3. Order Consolidation—The transformation of customer orders into picking orders.
4. Routing—The determination of sequences (routes) according to which the items have to be picked.

We propose a mathematical model to be solved with integer linear programming (ILP) techniques. The goal is to minimize the distance traveled by the pickers inside the shop. At first sight, this may resembles a capacitated vehicle routing problem (CVRP). However, for a given vehicle (picking trolley), the items from each customer order must be grouped in the proper boxes, which poses additional constraints to the problem. The CVRP is already a NP-hard problem. Furthermore, given the large number of orders processed by SonaeMC daily, the optimization problem is a large scale one. The need to get the solution in a small period of time and to deal with the additional constraints, led us to propose a tailored model (see Sect. 3).

2 An Overview of the Company's Picking Process

In this section we present an overview of the company's picking process at the time of the ESGI92. SonaeMC holds, all over Portugal, stores with different dimensions and architectures. However, they all share the topology depicted in Fig. 1, where five areas can be found:

- The *shop*—where the customers that visit the store pick their products.
- The *warehouse*—where the products are stored for replacement.
- The *depot*—an area inside the warehouse where the products of each client are gathered and packed.
- The *HRP area*—an area inside the warehouse where the High Rotation Products are stored.
- The *dock*—where the packed orders are loaded into vans that will deliver them to their final destination (the customer's address).

The process of assembling an order may be described as follows. An online customer makes an order on the company's website. The e-commerce team partitions the order in: *High Rotation Products* (HRP)—top selling articles or seasonal products; *Fresh Products* (FP)—products with special temperature and/or conservation requirements; and *Regular Products* (RP)—all the other products. The RP items are collected inside the shop by pickers who take them to the depot, where they are gathered with the HRP items that were previously packed for that customer. The remaining products—FP—are only delivered (by another picker) at the dock station when the order is taken to the van that will make the deliver.

We devoted our attention to the Regular Products (RP) which, according to the company, is the largest subset of items and where the pickers spend most of the time.

Unlike a warehouse, a supermarket is not designed to be efficient for the picking process (the goal is the opposite: *make the customer see as much as possible*). Therefore, changing the location of the products was not allowed.

At that time, pickers collected items for a single customer at a time. The purpose for this was to minimize the probability of swapping products between different orders. However, the picking trolleys have several separate boxes, which allows picking for multiple clients, and as they are already equipped with barcode scanners, the swapping can be avoided with a simple software update (checking in which box the item was placed).

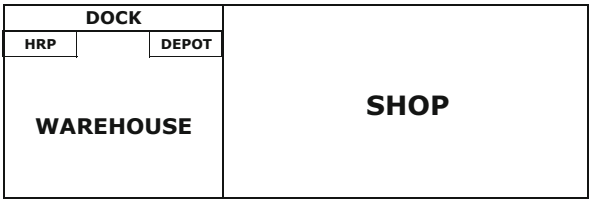


Fig. 1 Each store consists of five different areas: the shop, the warehouse, the depot, high rotation products area (HRP) and the dock

3 The Model

The problem may be stated as: Given a set of orders of different customers, the objective is to minimize the traveling distance needed to pick all the items, satisfying the weight and volume capacity constraints of the trolley, and imposing that in each box there are items of only one customer.

For the capacity constraints, let R be the maximum number of routes allowed, B the number of boxes per trolley, W the maximum weight allowed per box and S the maximum size allowed per box.

Let I_c be the set of items requested by customer $c \in C$ and $I = \cup_{c \in C} I_c$ the set of items to be picked. An *item* is understood as the quantity of a specific product ordered by one customer, i.e., in our model, the same product requested by different customers produces different items. For all $i \in I$, let w_i and s_i be, respectively, the weight and the size of item i, and c_i the customer who requested it. We suppose that $w_i \leq W$ and $s_i \leq S$, for all $i \in I$. If not, we previously divide the whole quantity of a specific product by a sufficient number of items in order to satisfy the conditions.

Our problem can be viewed as a variant of the CVRP Problem (see [8] for some variants), where:

- the depot d and each aisle v_i in the supermarket belong to the set of vertices V;
- each circuit (route) visits the depot vertex d exactly once;
- each vertex is visited at least once in the total of the circuits, and at most once in each circuit;
- each vertex v_i needs to be visited while there are items i to collect there;
- the total weight and the total volume of the items picked in a circuit do not exceed the vehicle and boxes capacity;
- the items are separated according to clients as they are being picked and put in the boxes of the trolley respecting the assignment of boxes to clients.

We construct a directed weighted graph $G = (V_I, A, \rho)$ as follows (see Fig. 2). For every $i \in I$, the set V_I is the subset of V consisting of vertices $v_i \in V$ corresponding to the aisles where items i are stored, together with the depot. Note that while V consists of all aisles of the shop (and the depot), V_I contains only the aisles with the items to be picked (and the depot). To define the set of arcs A, we consider a total ordering on the set of aisles, without the depot, $V_I \backslash \{d\}$, defined as:

$$u < v \text{ if and only if } u_x < v_x \vee (u_x = v_x \wedge u_y > v_y),$$

where (u_x, u_y) and (v_x, v_y) are the coordinates of the center of the aisles $u, v \in V_I \backslash \{d\}$. Now, the arcs A of the graph are:

$$A = \{(u, v) \in (V_I \backslash \{d\})^2 \mid u < v\} \cup \{(u, v) \in V_I^2 \mid u = d \ \vee \ v = d\}.$$

Finally, the function $\rho : A \to \mathbb{R}$ defines the weight of every arc $(u, v) \in A$ as the minimum distance that the picker needs to travel throughout the aisles of the

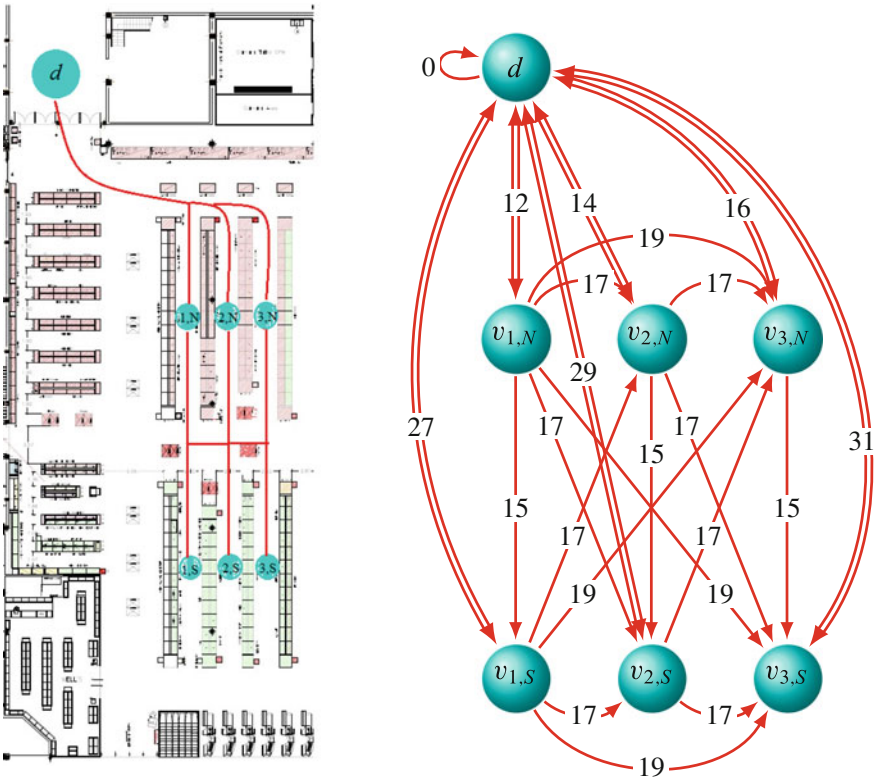


Fig. 2 An example of the directed graph G of part of the store, with the depot d, and aisles $v_{1,N}$, $v_{1,S}$, $v_{2,N}$, $v_{2,S}$, $v_{3,N}$ and $v_{3,S}$. The labels of the arcs are the distances

store, to get from vertex u, located at the center of an aisle, to vertex v, at the center of another aisle (it is not the usual Euclidean distance). To formulate the problem we use the binary variables $x^r_{uv} = 1$ if arc $(u, v) \in A$ is selected in route r, with $r = 1, \ldots, R$, and $y^{rb}_i = 1$ if item $i \in I$ is picked in route r into box b, with $r = 1, \ldots, R$, and $b = 1, \ldots, B$. The model is the following.

$$\text{Min} \sum_{1 \le r \le R} \sum_{(u,v) \in A} \rho_{uv} x^r_{uv} \tag{1}$$

$$s.t. \quad \sum_{(v,u) \in A} x^r_{vu} \le 1 \quad \forall r = 1, \ldots, R, \ \forall v \in V_I \backslash \{d\} \tag{2}$$

$$\sum_{(u,v) \in A} x^r_{uv} \le 1 \quad \forall r = 1, \ldots, R, \ \forall v \in V_I \backslash \{d\} \tag{3}$$

$$\sum_{(v,u)\in A} x^r_{vu} = \sum_{(u,v)\in A} x^r_{uv} \quad \forall r = 1,\ldots,R,\ \forall v \in V_I\backslash\{d\} \tag{4}$$

$$\sum_{r=1,\ldots,R} \sum_{(v,u)\in A} x^r_{vu} \geq 1 \quad \forall v \in V_I\backslash\{d\} \tag{5}$$

$$\sum_{r=1,\ldots,R} \sum_{(u,v)\in A} x^r_{uv} \geq 1 \quad \forall v \in V_I\backslash\{d\} \tag{6}$$

$$\sum_{(d,u)\in A} x^r_{du} = 1 \quad \forall r = 1,\ldots,R \tag{7}$$

$$\sum_{(u,d)\in A} x^r_{ud} = 1 \quad \forall r = 1,\ldots,R \tag{8}$$

$$\sum_{r=1,\ldots,R} \sum_{b=1,\ldots,B} y^{rb}_i = 1 \quad \forall i \in I \tag{9}$$

$$\sum_{b=1,\ldots,B} y^{rb}_i \leq \sum_{(u,v_i)\in A} x^r_{uv_i} \quad \forall r = 1,\ldots,R\ \forall i \in I \tag{10}$$

$$y^{rb}_i + y^{rb}_j \leq 1 \quad \forall i \in I, \forall j \in I\backslash I_{c_i}, \forall r = 1,\ldots,R, \forall b = 1,\ldots,B \tag{11}$$

$$\sum_{i\in I} w_i y^{rb}_i \leq W \quad \forall r = 1,\ldots,R,\ \forall b = 1,\ldots,B \tag{12}$$

$$\sum_{i\in I} s_i y^{rb}_i \leq S \quad \forall r = 1,\ldots,R,\ \forall b = 1,\ldots,B \tag{13}$$

$$x^r_{uv} \in \{0,1\} \quad \forall (u,v) \in A,\ \forall r = 1,\ldots,R \tag{14}$$

$$y^{rb}_i \in \{0,1\} \quad \forall i \in I,\ \forall r = 1,\ldots,R,\ \forall b = 1,\ldots,B \tag{15}$$

The objective function in (1) deems to minimize the total traveling distance. Constraints (2) and (3) ensure that there will be at most one arc leaving every aisle v, and there will be at most one arc entering v in each route r, for all vertices in $V_I\backslash\{d\}$. Equations (4) ensure that, for each route, the number of arcs entering an aisle is equal to the number of arcs leaving it. Constraints (5) and (6) guarantee that each aisle is visited at least once in the total of routes. There are also similar constraints for the depot: (7) and (8). The next constraints concern the boxes where the items are placed. In (9) we force that each item is picked to exactly one box in exactly one route. Also, if an item is picked in a route, its aisle needs to be visited in such route; this is given by (10). Inequalities (11) guarantee that each box in a route does not have items of different customers. Finally, we add the weight and volume constraints concerning the dimensions of the boxes (12) and (13). The range of the variables are established in (14) and (15).

4 Results

The model was implemented using AMPL [4] and tested with real data provided by SonaeMC, regarding the orders placed at one of their main stores in a given period. The dataset was composed of several orders, with an overall weight of $\approx$600 kg and volume of $\approx$1.6 m^3. In our implementation we set $R = 10$, $B = 6$, $W = 12$ kg and $S = 46{,}400\,cm^3$. The model was submitted to Gurobi solver on NEOS online server [1, 3, 6].

As the complexity of this kind of model is non-polynomial, we used an heuristic to reduce the dimension of the problem. This heuristic computes a matrix of distances between customers, based on their similarity in terms of number of products in common aisles, and gathers the orders in clusters, so that the available computational resources may solve the integer programming model, within these clusters. The heuristic splits our set of orders in 7 subsets: 3 of them with products in a single aisle (no routing needed), and 4 clusters with products in several aisles. For each one of these 4 clusters, we solved the IP model in (1)–(15). In this 4 subsets the optimal solutions were found, in a total time of 435 s.

We compared the results of our model with a simulation of SonaeMC's current picking. We were able to reduce the total distance by 39 %, from 1341.5 to 818 m. The number of routes and the number of boxes is similar in both. With our routing solution, the picker would increase the picking rate in 24 %.

5 Conclusions

We focused on a case study of a large Portuguese grocery company, regarding the open shop order picking process for online customers. We developed an integer programming model for a variant of the Capacitated Vehicle Routing Problem, with additional constraints to deal with the specificity of the trolley and the company's requirements. The model integrates the batching and the routing problems with very promising first results. Also, an heuristic was designed for reducing the dimension of the problem, in order to obtain reasonable computational times. On a real dataset, the model reduced the total traveling distance by 39 %, and increased the picking rate in 24 %, a result that exceeded the company's expectations.

Acknowledgements A. Moura was partially supported by CMUP (UID/MAT/00144/2013), which is funded by FCT (Portugal) with national (MEC) and European structural funds through the programs FEDER, under the partnership agreement PT2020.

References

1. Czyzyk, J., Mesnier, M., Moré, J.: The NEOS server. IEEE J. Comput. Sci. Eng. **5**(3), 68–75 (1998)
2. De Koster, R., Le-Duc, T., Roodbergen, K.J.: Design and control of warehouse order picking: a literature review. Eur. J. Oper. Res. **182**(2), 481–501 (2007)
3. Dolan, E.: The NEOS server 4.0 administrative guide. Technical Memorandum ANL/MCS-TM-250. Mathematics and Computer Science Division. Argonne National Laboratory, Argonne (2001)
4. Fourer, R., Gay, D.M., Kernighan, B.W.: AMPL - A Modeling Language for Mathematical Programming. Thomson/Brooks/Cole, Pacific Grove (2003)
5. Frazelle, E.: World-Class Warehousing and Material Handling. McGraw-Hill, New York (2001)
6. Gropp, W., Moré, J.: Optimization environments and the NEOS server. In: M.D. Buhmann, A. Iserles (eds.) Approximation Theory and Optimization, pp. 167–182. Cambridge University Press, Cambridge (1997)
7. Tompkins, J., White, J., Bozer, Y., Tanchoco, J.: Facilities Planning Wiley, New Jersey (2003)
8. Toth, P., Vigo, D.: The Vehicle Routing Problem. Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics, Philadelphia (2002)
9. Van den Berg, J.P.: A literature survey on planning and control of warehousing systems. IIE Trans. **31**(8), 751–762 (1999)

Study Groups in Ireland: A Reflection

William Lee, Joanna Mason, and Stephen O'Brien

Abstract Study groups were first introduced to Ireland in 2008 by MACSI. We present an overview of MACSI study groups focusing on the role study groups play in initiating longer term interactions between industry and academia.

Keywords Study group

1 Introduction

A study group is a week long workshop at which academic mathematicians work on problems brought to the group by industry. Typically there are 6–8 problems, which are introduced by the industrial representatives on the first day. Following these presentations the mathematicians break up into groups to work on the problems. The format works best if the industrial representatives stay with the group to provide information as needed and to ensure that the group stays focussed on the most important aspects of the problem. On the final day of the workshop preliminary results are shared with all academic and industrial participants. Following the study group a report of progress made during the study group is written.

Study groups in Ireland were introduced to Ireland in 2008 by MACSI, the Mathematics Applications Consortium for Science and Industry based at the University of Limerick. In this paper we report progress towards developing a mutually beneficial interaction between academic applied mathematicians and industry. In particular we consider the role that study groups with industry play in initiating these interactions.

An idealised vision of the interaction between academic mathematicians and industry is shown in Fig. 1. In this vision a study group acts as the gateway to an ongoing interaction leading to outputs beneficial to both the academics and the industrial partner.

Following the study group the company, having seen the value in a mathematical approach to their problem, commissions a small amount of consultancy work to be

W. Lee (✉) • J. Mason • S. O'Brien
MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
e-mail: william.lee@ul.ie; joanna.mason@ul.ie; stephen.obrien@ul.ie

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_26

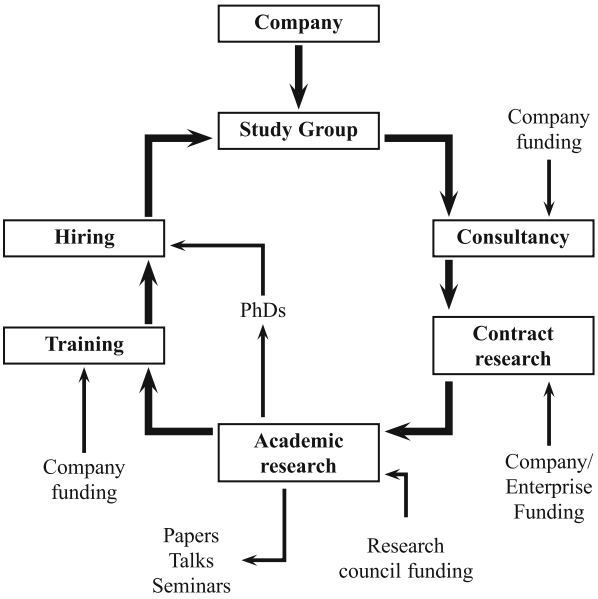


Fig. 1 Idealised study group workflow

carried out by academic or research staff already in employment. This allows work to begin immediately, without recruitment delays, and, since it is fully funded by the industrial partner, entitles them to full ownership of any IP generated. Any of the company's IP they were not able to bring to an open forum such as a study group can now be disclosed since an NDA will form part of the consulting agreement.

This short burst of consultancy work will be followed by a program of contract or collaborative research, funded either entirely by the company or with some portion of the funding coming from enterprise support agencies e.g. Enterprise Ireland. This contract research includes any original research needed to develop a viable solution to the company's needs. At the conclusion of the contract research stage there is a handover of the solution to the company. The solution may be in the form of software, documented research results or something else.

During the course of the industrially focussed research, avenues for academic research will have opened up, maybe leading to more accurate or more complete solutions to the industrial problem, maybe only tangentially related to it. Thus, a contract research program will be followed by academic research, funded by an academic funding body such as Science Foundation Ireland. This research will often be supported by the company, whose interaction with the project will strengthen the case for the research satisfying impact criteria that are becoming an increasingly prominent feature of the academic landscape, particularly in Ireland and the UK [2]. This phase of the research will lead to standard academic deliverables: papers, seminars, invited talks and trained PhDs.

Having found the mathematical tools developed during the project of interest to themselves the company will request training, either in the specifics of the tools developed in this project or more generally in the mathematical modelling toolkit and approach to problem solving.

Finally the company will enhance its own research and development capability by hiring graduating PhDs trained by this research program. Also, benefitting from an enhanced ability to identify problems susceptible to a mathematical approach, the company will participate in another study group, restarting the cycle.

As described above this is an idealised vision of the role study groups play in a thriving interaction between academics and industry. It should be noted that this vision may be specific to an Irish context where due to severe economic pressures on the government academic funding is predicated on being able to demonstrate "impact" which is often translated to mean industry cash contribution. With a large number of groups chasing smaller and smaller pots of money alternative income generation strategies are highly desirable.

In the rest of this paper we consider how closely reality approaches this vision. In Sect. 2 we give a brief overview of Irish study groups. The next two sections consider two case studies of the interaction of MACSI with companies which have attended multiple study groups: Analog Devices in Sect. 3 and Aughinish Alumina in Sect. 4. We discuss how to get the most from study groups in Sect. 5, and give our conclusions in Sect. 6.

2 Overview of Irish Study Groups

Table 1 gives an overview of study groups with industry in Ireland, including outputs subsequent to the study group: papers, undergraduate, MSc and PhD thesis, company or enterprise funding following the study group and media exposure.

Table 1 Irish study groups in numbers

Study groups	6
Study group problems	42
Companies returning	4
Papers	13
Theses	9
Funding	9

Study groups held in Ireland have proved successful in generating and funding subsequent academic activity, such as papers and theses (PhD, MSc or undergraduate)

The table shows a number of important facts. Firstly a number of companies have attended multiple study groups. Secondly a number of study group projects have stimulated further engagement as illustrated by followup funding. (It is important to note that lack of further funding may also indicate success: a problem completely solved during the study group itself.) Thirdly a number of study groups problems have opened up avenues of academic enquiry as shown by resulting publications. Nevertheless these still only form a small fraction of the total number of study group problems. It is vital to ascertain whether this indicates that the study group format is acting as it should in giving academics and industrialists a rapid way of determining if there is scope for a mutually beneficial outcomes, or if there is a "valley of death" following a study group in which academically or industrially viable projects are not being managed effectively.

3 Case Study: Analog Devices

Analog Devices are a multinational electronics company with a significant presence in Limerick. They have brought problems to five out of the six MACSI study groups, one of which is illustrated in Fig. 2. The problem of modelling the blowing of polysilicon fuses was brought by Analog Devices to ESGI62 and a model initially developed there was developed further by MACSI academics and Analog electrical engineers working together [1].

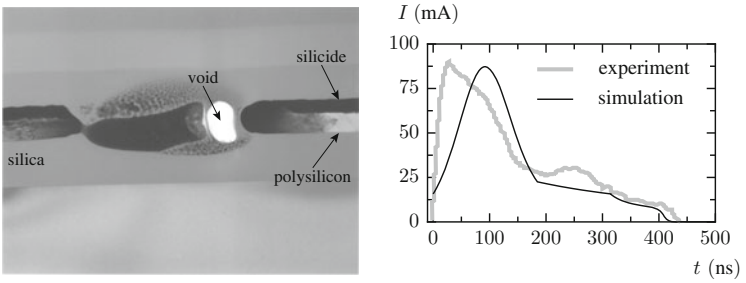


Fig. 2 Mathematical model of blowing polysilicon fuses [1]. Developed following ECMI62

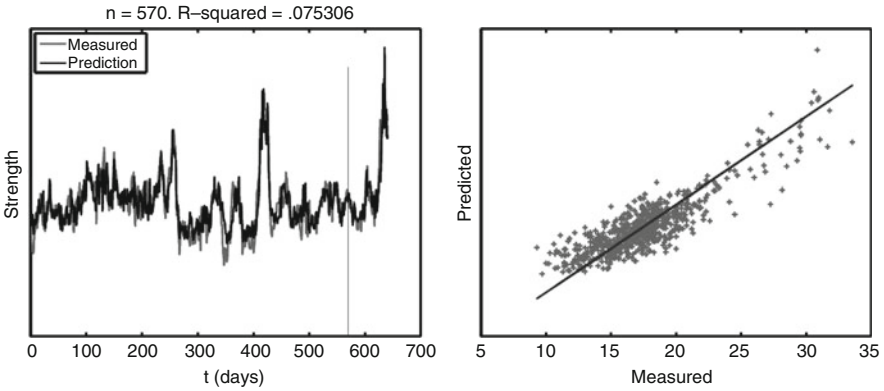


Fig. 3 Modelling product quality in the Bayer Process. Developed preceding ESGI82

4 Case Study: Aughinish Alumina

Rusal Aughinish Alumina is Europe's largest alumina refinery, situated on the Shannon Estuary in Ireland. Aughinish Alumina has also participated in multiple study groups, and has both funded research coauthored papers with MACSI coauthors. However their interaction also doesn't conform to the idealised workflow shown in Fig. 1 since in their case a contract research project whose results are illustrated in Fig. 3 acted as a gateway to study groups rather than the reverse.

5 Discussion

As the case studies show, the idealised study group driven workflow shown in Fig. 1 remains an ideal, with no real interaction between academic mathematicians and industry conforming to the template. While this in itself is not grounds for concern, the failure of most study group projects to lead to interactions of any kind should be examined carefully. As discussed above, in some cases this is an indicator of success: the problem posed is completely solved during the study group. In other cases it is hard for the academics to make meaningful progress on the project because of lack of access to the industrial representative—this form of interaction is inherently collaborative and cannot be expected to succeed without both sides working together. However, we there are several steps that can be taken by academics, both before and after the study group to maximise the chances of a meaningful engagement.

Before the Study Group:

- Work with the company to formulate the problem statement and to identify information that is likely to be needed during the study group. We have found it best to write the problem statement ourselves, getting feedback from industry on each draft of the problem statement until they are satisfied it accurately captures their requirements. This usually results in us having a much better understanding of the problem.
- Manage expectations. Make sure industry understands the likely outcomes. Is a complete solution to the problem during the study group possible? Is it already known that their problem is very hard (e.g. granular flow and segregation)?

After the Study Group:

- Quick turnaround of reports is essential. In practice this means the host institution must take charge of the report writing.
- Going to the site to present reports. This is especially important if the main contacts require buy in from more senior colleagues to continue the engagement.
- Options for follow up work must be explicitly discussed.

6 Conclusions

In the current higher education landscape in which available academic funding is diminishing and increasingly tied to industry co-funding or demonstrable impact Study Groups can play an important role demonstrating the continuing relevance of applied mathematics, while also unlocking alternative, non-exchequer revenue streams. However, this cannot be achieved by study groups in isolation. A study group must be seen as only one step of a pathway leading to mutually beneficial outcomes for academics and industry. Our own analysis of study groups in Ireland has highlighted the importance of correctly managing the subsequent step which currently acts a something of a "valley of death" for industrial-academic collaborations.

Acknowledgements We acknowledges the support of the Mathematics Applications Consortium for Science and Industry (http://www.macsi.ul.ie) funded by the Science Foundation Ireland Investigator Award Grant 12/IA/1683. In addition we would like to thank all participants in MACSI study groups, both from academia and industry.

References

1. Lee, W.T., Fowler, A.C., Power, O., Healy, S., Browne, J.: Blowing of polycrystalline silicon fuses. Appl. Phys. Lett. **97**(2), 023502 (2010)
2. Science Foundation Ireland: Agenda 2020. www.sfi.ie (2012)

MS 7
MINISYMPOSIUM: HIGH PERFORMANCE COMPUTATIONAL FINANCE

Organizers

José Pedro Silva[1]

Speakers

Alvaro Leitao[2] and Cornelis W. Osterlee[3]
On a GPU Acceleration of the Stochastic Grid Bundling Method

Jinzhe Yang[4]
Parallel Implementation of Particle Filtering-Based Maximum Likelihood Estimation of Jump-Diffusion Model

J.P. Silva[1], J. ter Maten[5], M. Günther[6] and M. Ehrhardt[7]
Proper Orthogonal Decomposition in Option Pricing: Basket Options and Heston Model

[1]José Pedro Silva, Bergische Universität Wuppertal, Wuppertal, Germany.
[2]Alvaro Leitao, TU Delft, Delft, The Netherlands.
[3]Cornelis W. Osterlee, TU Delft, Delft, The Netherlands.
[4]Jinzhe Yang, Aberdeen Asset Management, Aberdeen, UK.
[5]Jan W. ter Maten, Bergische Universität Wuppertal, Wuppertal, Germany.
[6]Michael Günther, Bergische Universität Wuppertal, Wuppertal, Germany.
[7]Matthias Ehrhardt, Bergische Universität Wuppertal, Wuppertal, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_27

Keywords

Computational finance
High performance computing
Model order reduction

Short Description

New computing paradigms and platforms have introduced a new range of previously unfeasible strategies and algorithms to solve computationally intensive problems in finance. On the other hand, it allows current schemes and algorithms to tackle more complex problems. This symposium will bring together young researchers from two research networks, ITN-Strike and ITN-HPCFinance, who are currently developing methods to apply in computationally demanding problems both from the point of view of using new architectures, i.e., GPUs, FPGAs, as well as strategies, namely, Model Order Reduction. The interchange of ideas will certainly create the basis for a collaboration between these two networks.
A parallel GPU version of the Monte Carlo based Stochastic Grid Bundling Method (SGBM) for pricing multi-dimensional Bermudan options is presented, as well as parameter estimation using Field-Programmable Gate Arrays and the use of Proper Orthogonal Decomposition as a technique to drastically reduced the computational time needed to solve option pricing PDE.

On a GPU Acceleration of the Stochastic Grid Bundling Method

Alvaro Leitao and Cornelis W. Oosterlee

Abstract Pricing early-exercise financial options under multi-dimensional stochastic processes is a challenge in the financial sector. For this purpose, the authors in Jain and Oosterlee (The stochastic grid bundling method: efficient pricing of Bermudan options and their Greeks, 2015) proposed a practical simulation-based algorithm called Stochastic Bundling Grid Method (SGBM). SGBM is a Monte Carlo based method for pricing multi-dimensional Bermudan options. The method is based in a combination of dynamic programming, simulation, regression and bundling of paths. In the present work, the SGBM method is taken to the extreme with as a purpose a near-future extension of the method, for example, to Credit Value Adjustment (CVA) calculations. Here, the number of Monte Carlo paths, the problem dimensions, the amount of bundles are increased drastically. As a consequence, the SGBM method becomes significantly more (almost impractically) expensive. Overall, with the increase of the number of bundles, the iterative bundling process used in the original method would take too much computing time. In addition, the algorithm needs a huge storage because many bundles contain many more Monte Carlo paths. In order to make the method affordable, the General-Purpose computing on Graphics Processing Units (GPGPU) paradigm is used to parallelize the algorithm. More specifically, the Nvidia CUDA platform (CUDA webpage: URL http://www.nvidia.com/object/cuda_home_new.html) is chosen to reach this aim, taking advantage of its latest features. Two steps of parallelization are performed, one for the Monte Carlo path simulation and another one for the bundling calculations. Furthermore, a new way to make the bundles is proposed, which is efficient and overcomes the drawbacks caused by the increasing number of bundles and the problem dimensionality.

Keywords Computational finance • High performance computing

A. Leitao (✉)
TU Delft, Delft Institute of Applied Mathematics, Delft, The Netherlands

CWI-Centrum Wiskunde and Informatica, Amsterdam, The Netherlands
e-mail: A.LeitaoRodriguez@tudelft.nl

C.W. Oosterlee
CWI-Centrum Wiskunde and Informatica, Amsterdam, The Netherlands
e-mail: c.w.oosterlee@cwi.nl

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_28

1 Introduction

In recent years, different Monte Carlo simulation techniques for pricing *high-dimensional early-exercise option contracts* appearing in computational finance were developed. In the wake of the recent financial crisis, accurately modeling and pricing these kinds of options gains importance due to the incorporation of a so-called *counterparty risk premium* to the option value, which can be seen as an option in which a counterparty may go into default before the end of a financial contract, and thus cannot pay possible contractual obligations. One of the recent Monte Carlo pricing techniques is the Stochastic Bundling Grid Method (SGBM), proposed by Jain and Oosterlee in [5] for pricing Bermudan options with several underlying assets. The method is a hybrid of *regression-* and *bundling*—based approaches, and uses regressed value functions, together with bundling of the state space to approximate continuation values at different time steps. In this paper, we extend the method's applicability by increasing the number of bundles and the problem dimensionality, which, together, also imply a drastic increase of the number of Monte Carlo paths. As the method becomes much more time-consuming then, we propose to parallelize the SGBM method taking advantage of the General-Purpose computing on Graphics Processing Units (GPGPU) paradigm. For this purpose, the CUDA [2] tool developed by Nvidia for their GPUs is used. In order to get a significant improvement, we also present a new bundling technique for SGBM which is much more efficient on parallel hardware.

The paper is organized as follows. In Sect. 2, the particular Bermudan option pricing problem is introduced. Section 3 describes briefly the Stochastic Grid Bundling Method. Section 4 gives details about the choices made in the GPU implementation process. In Sect. 5, some results and time comparisons are shown. Finally, we conclude in Sect. 6.

2 Problem Formulation

This section defines the Bermudan option pricing problem and sets up the notations used in this paper. A Bermudan option is an option where the buyer has the right to exercise at a set number of times, $t \in [t_0 = 0, \ldots, t_m, \ldots, t_M = T]$, before the end of the contract, T. $\mathbf{S}_t = (S_t^1, \ldots, S_t^d) \in \mathbb{R}^d$ defines the d-dimensional underlying process. Let $h_t := h(\mathbf{S}_t)$ be an adapted process representing the intrinsic value of the option, i.e. the holder of the option receives $\max(h_t, 0)$, if the option is exercised at time t. With the risk-less savings account process $B_t = \exp(\int_0^t r_s\, ds)$, where r_t denotes the instantaneous risk-free rate of return, we define

$$D_{t_m} = \frac{B_{t_m}}{B_{t_{m+1}}}.$$

We consider the special case where r_t is constant. The problem is then to compute

$$V_{t_0}(\mathbf{S}_{t_0}) = \max_{\tau} \mathbb{E}\left[\frac{h(\mathbf{S}_\tau)}{B_\tau}\right],$$

where τ is a stopping time, taking values in the finite set $\{0, t_1, \ldots, T\}$. The value of the option at the terminal time T is equal to the option's payoff,

$$V_T(\mathbf{S}_T) = \max(h(\mathbf{S}_T), 0).$$

The conditional continuation value Q_{t_m}, i.e. the expected payoff at time t_m, is given by:

$$Q_{t_m}(\mathbf{S}_{t_m}) = D_{t_m}\mathbb{E}\left[V_{t_{m+1}}(\mathbf{S}_{t_{m+1}})|\mathbf{S}_{t_m}\right].$$

The Bermudan option value at time t_m and state $\mathbf{S}_{t_m}$ is then given by

$$V_{t_m}(\mathbf{S}_{t_m}) = \max(h(\mathbf{S}_{t_m}), Q_{t_m}(\mathbf{S}_{t_m})).$$

We are interested in finding the value of the option at the initial state $\mathbf{S}_{t_0}$, i.e. $V_{t_0}(\mathbf{S}_{t_0})$.

3 Stochastic Grid Bundling Method

The Stochastic Grid Bundling Method (SGBM) [5] is a simulation-based Monte Carlo dynamic programming method, which first generates Monte Carlo paths forward in time, followed by determining the optimal early-exercise policy moving backwards in time in a dynamic programming framework, based on the Bellman principle of optimality. The steps involved in the SGBM algorithm are briefly described in the following paragraphs:

Step I: Generation of Stochastic Grid Points

The grid points in SGBM are generated by simulating independent copies of sample paths, $\{\mathbf{S}_{t_0}(n), \ldots, \mathbf{S}_{t_M}(n)\}$, $n = 1, \ldots, N$, of the underlying process $\mathbf{S}_t$, all starting from the same initial state $\mathbf{S}_{t_0}$. The n-th grid point at time step t_m is then denoted by $\mathbf{S}_{t_m}(n)$, $n = 1, \ldots, N$. Depending upon the underlying stochastic process an appropriate discretization scheme, e.g. the Euler scheme, is used to generate sample paths. Sometimes the diffusion process can be simulated directly, essentially because it appears in closed form, like for the regular multi-dimensional Black-Scholes model.

Step II: Option Value at Terminal Time

The option value at the terminal time $t_M = T$ is given by:

$$V_{t_M}(\mathbf{S}_{t_M}) = \max(h(\mathbf{S}_{t_M}), 0),$$

with $\max(h(\mathbf{S}_{t_M}), 0)$ a multi-dimensional payoff function. This relation is used to compute the option value for all grid points at the final time step.

The following steps are subsequently performed for each time step, t_m, $m \leq M$, recursively, moving backwards in time, starting from t_M.

Step III: Bundling

The grid points at t_{m-1} are *bundled* into $\mathscr{B}_{t_{m-1}}(1), \ldots, \mathscr{B}_{t_{m-1}}(\nu)$ non-overlapping sets or partitions. Different approaches for partitioning can be considered. Due to its importance, this decision is discussed in more detail in Sect. 3.1.

Step IV: Mapping High-Dimensional State Space to a Low-Dimensional Space

Corresponding to each bundle $\mathscr{B}_{t_{m-1}}(\beta)$, $\beta = 1, \ldots, \nu$, a parameterized value function $Z : \mathbb{R}^d \times \mathbb{R}^K \mapsto \mathbb{R}$, which assigns values $Z(\mathbf{S}_{t_m}, \alpha_{t_m}^{\beta})$ to states $\mathbf{S}_{t_m}$, is computed. Here $\alpha_{t_m}^{\beta} \in \mathbb{R}^K$ is a vector of free parameters. The objective is then to choose, for each t_m and β, a parameter vector $\alpha_{t_m}^{\beta}$ so that

$$Z(\mathbf{S}_{t_m}, \alpha_{t_m}^{\beta}) \approx V_{t_m}(\mathbf{S}_{t_m}).$$

After some approximations, $Z(\mathbf{S}_{t_m}, \alpha_{t_m}^{\beta})$ can be computed using ordinary least squares regression.

Step V: Computing the Continuation and Option Values at t_{m-1}

The continuation values for $S_{t_{m-1}}(n) \in \mathscr{B}_{t_{m-1}}(\beta)$, $n = 1, \ldots, N$, $\beta = 1, \ldots, \nu$, are approximated by

$$\widehat{Q}_{t_{m-1}}(\mathbf{S}_{t_{m-1}}(n)) = \mathbb{E}[Z(\mathbf{S}_{t_m}, \alpha_{t_m}^{\beta}) | \mathbf{S}_{t_{m-1}}(n)]$$

The option value is then given by:

$$\widehat{V}_{t_{m-1}}(\mathbf{S}_{t_{m-1}}(n)) = \max(h(\mathbf{S}_{t_{m-1}}(n)), \widehat{Q}_{t_{m-1}}(\mathbf{S}_{t_{m-1}}(n))).$$

3.1 Bundling

One of the techniques proposed to partition the data into ν non-overlapping sets is the *k-means* clustering technique. The algorithm uses an iterative refinement algorithm, where, given an initial guess of clusters means, first of all the algorithm assigns each data item to one specific set and subsequently updates the clusters. This process is repeated until some stop criterion is satisfied (see [5] for more details).

Since our goal is to drastically increase the number of bundles used and, in particular, the problem dimensionality, the k-means algorithm becomes too expensive in terms of computational time and memory usage. In addition, it may happen that some bundles do not contain enough data to compute an accurate regression function when the number of bundles is increased. In order to overcome these two problems of the k-means clustering, we propose a new bundling technique which is more efficient taking into account our goal: it does not involve an iterative process and distributes the data equally. The details are shown in the next subsection.

3.1.1 Equal-Partitioning Technique

This bundling technique is particularly well-suited for parallel processing; it involves two steps: sorting and splitting. The general idea is to sort the data first under some convenient criterion and then split the sorted data items in sets of equal size. A schematic representation of this technique is shown in Fig. 1.

With this simple approach, the drawbacks of the iterative bundling for very high dimensions and an enormous amount of paths are avoided. The sorting process is more efficient and less time-consuming than an iterative search and, furthermore, it is highly parallelizable. The split stage assigns directly the portions of data to bundles which will contain the same amount of *similar* (following some criterion) data items. Hence, the regression can be performed even though the number of bundles increases in a significant way.

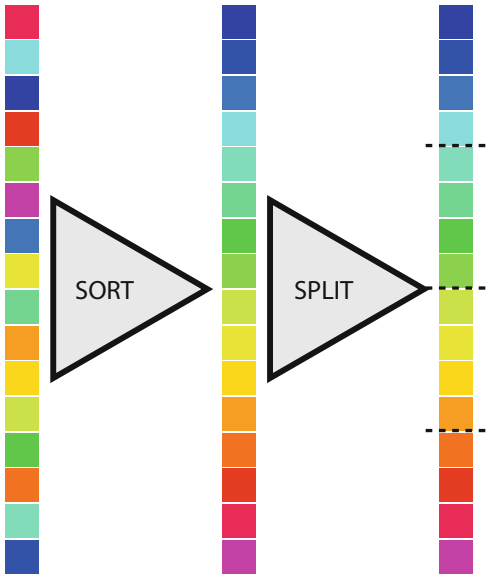


Fig. 1 Equal partitioning scheme

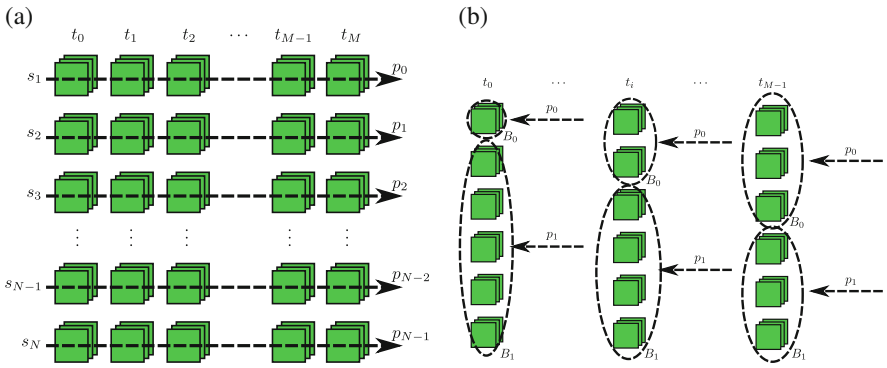


Fig. 2 SGBM Monte Carlo and bundling stages. (**a**) SGBM Monte Carlo stage. (**b**) SGBM bundling stage

4 Implementation Details

The original SGBM implementation was done in Matlab. The first step of our implementation is to code an efficient version of the method in the C programming language, because it eases the subsequent parallel coding in CUDA. In addition, we can use both implementations to compare results and execution times. Once we obtained the C-version, we coded the CUDA-version aiming to parallelize the suitable parts of the method. In addition, we also carry out the implementation of SGBM with the equal-partitioning bundling technique in C and CUDA.

Since the SGBM method is based on two clearly separated parts, we perform the parallelization separately. First of all, the Monte Carlo grid generation is parallelized (step I). As is well-known, Monte Carlo methods are very suitable for parallelization, because of characteristics like a high number of simulations and data independence. In Fig. 2a, we see how the parallelization is done. The second main stage of SGBM is the regression and the computation of the continuation and option values (Steps IV and V) in each bundle, backwards in time. Due to the data dependency between time steps, the only way to parallelize this part of the method is by performing the calculations in each bundle in parallel. A schematic representation is given in Fig. 2b.

We will focus on the CUDA implementation and its main details. The memory transfers between CPU and GPU are key since we have to move huge amounts of data. The increase of the number of bundles implies an increase of the number of Monte Carlo paths which we need. For example, in case of the Monte Carlo scenario generator, we need to store in and move from GPU *MC simulations* · *Time Steps* · *Dimension* doubles[1] (8 bytes). In order to improve the performance in this aspect, we take advantage of the CUDA feature Unified Virtual Addressing (UVA) which

[1]The number in double-precision floating-point format.

allows asynchronous transfers and page-locked memory accesses using a single address space for CPU and GPU.

In the following subsections, we will show more specific details of the CUDA implementations of the two SGBM versions, the original one (with k-means bundling) and new one (with equal-partitioning).

4.1 Parallel SGBM

4.1.1 Monte Carlo paths

In a GPU very large amount of threads can be managed, so we launch one thread per Monte Carlo simulation. The necessary random numbers are obtained "on the fly" in each thread by means of cuRAND library [3]. In addition, the intrinsic value of the option is calculated by the Monte Carlo generator decreasing the launched loops and improving the GPU memory accesses. In the original implementation, we need to store all Monte Carlo data because it will be used in the bundling stage. This is a limiting factor since the maximum memory storage is easily reached. Furthermore, we have to move each obtained value from GPU to CPU and, with a significant amount of paths and dimensions, this transfer could become really expensive.

As in the case of the k-means clustering version, one GPU thread per Monte Carlo scenario is launched when using *equal-partitioning clustering*. The main difference between the two schemes is that we now perform calculations for the sorting criterion to avoid storage of the complete Monte Carlo simulation and to transfer only a small amount of data between GPU and CPU. This approach gives us a considerable performance improvement and allows us to increase drastically the number of simulations (depending on the number of bundles) and thus the dimensionality.

4.1.2 Bundling Schemes

After the Monte Carlo simulation, the bundling process using k-means clustering is performed in one test, and the equal-partitioning scheme in another test. For the *k-means clustering* the computations of the differences between the cluster centers and all of the scenarios have been parallelized. However, the very large amount of bundles makes this very expensive, since the bundling must be done in each time step. The other parts of the algorithm have to be done in a sequential way because of data dependency.

As mentioned, equal-partitioning bundling involves two stages: sorting and splitting. For the sorting stage, we employ the sort functions of the Thrust library. After the sorting, the splitting stage is immediate since the size of the bundles is known, i.e. *Paths*/*Number Bundles*. Each GPU thread manages a pointer which points at the beginning of each bundle.

4.1.3 Estimator

When the bundling stage is done, the exercise policy and the final time option values can be computed. In order to use the GPU memory efficiently, the obtained scenarios (once bundled by k-means clustering) are sorted with respect to the bundles and stored in that way. We minimize the memory accesses and the amount of used memory. For this purpose, we use the Thrust library [6] which allows, among other features, sorting of data on the GPU. Note that the data is already sorted in the case of equal-partitioning bundling.

At each time step, one GPU thread per bundle is launched. For each bundle, the regression and option values are calculated on GPU. All threads collaborate in order to compute the continuation value which determines the early-exercise policy.

In the final stage, a summation is necessary to determine the option price. Again, we take advantage of the Thrust library to perform this reduction on the GPU.

5 Results

Experiments were performed on the Accelerator Island system of the Cartesius Supercomputer (more information in [1]) with the following main characteristics:

- Intel Xeon E5-2420 (Sandy Bridge).
- NVIDIA Tesla K20m.
- C-compiler: GCC 4.4.6.
- CUDA version: 5.5.

All computations are made in double precision, because a high accuracy is needed in the k-means clustering and in the regression computations. In each performed test, we consider the d-dimensional problem of pricing a geometric basket Bermudan put option with the following characteristics: $\mathbf{S}_{t_0} = (40, \ldots, 40) \in \mathbb{R}^d$, $K = 40$, $r_t = 0.06$, $\sigma = (0.2, \ldots, 0.2) \in \mathbb{R}^d$, $\rho_{ij} = 0.25$, $T = 1$ and $M = 10$. As the stochastic asset model, we choose the multi-dimensional Geometric Brownian motion (GBM), and for the discretization scheme we employ the Euler discretization.

In the original SGBM paper, the authors have shown the convergence of SGBM using k-means bundling, in dependence of the number of bundles. For the equal-partitioning bundling, we make a similar convergence study by pricing the previously mentioned option. In Fig. 3, we show the convergence of the calculated option price to the reference price (obtained by the COS method [4]) for different dimensionalities, i.e. $5d$, $15d$ and $50d$.

Once the convergence of the equal-partitioning technique is shown numerically, we increase drastically the amount of bundles and, hence, the number of Monte Carlo paths. For the two presented bundling techniques, we perform a time comparison between the C and CUDA implementations. The results are shown in Table 1. We observe a significant acceleration for both cases, with a special

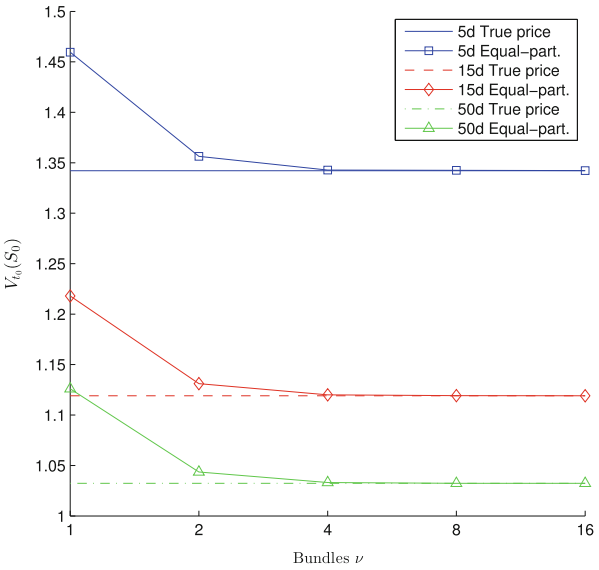


Fig. 3 Convergence with equal-partitioning bundling technique. Test configuration: $N = 2^{16}$ and $\delta t = T/M$

Table 1 Time (s) for the C and CUDA versions

	k-means			Equal-partitioning		
	$5d$	$10d$	$15d$	$5d$	$10d$	$15d$
C	676.25	1347.07	2008.16	157.83	234.60	320.59
CUDA	38.77	145.28	307.64	13.85	14.78	15.42
Speedup	17.44	9.27	6.52	11.40	15.87	20.79

Test configuration: $N = 2^{22}$, $\delta t = T/M$ and $\nu = 2^{11}$

Table 2 Time (s) for a high-dimensional problem with equal-partitioning

	$\nu = 2^{12}$			$\nu = 2^{13}$			$\nu = 2^{14}$		
	$30d$	$40d$	$50d$	$30d$	$40d$	$50d$	$30d$	$40d$	$50d$
C	570.83	787.36	989.15	573.88	787.51	986.16	571.97	787.22	984.51
CUDA	18.06	21.44	25.09	18.43	21.93	25.42	19.25	22.60	26.06
Speedup	31.61	36.72	39.42	31.14	35.91	38.79	29.71	34.83	37.78

Test configuration: $N = 2^{22}$ and $\delta t = T/M$

improvement in the case of the equal-partitioning technique. This is because the iterative process of k-means bundling penalizes parallelism and memory transfers, while equal-partitioning handles these issues in a more efficient way.

The second goal is to increase the problem dimensionality. In the case of the k-means bundling algorithm, this is not possible because of memory usage. However, we save memory using the equal-partitioning technique which enables increasing dimensions. In Table 2, the execution times of pricing the geometric basket Bermudan put option in different dimensions and with different numbers of bundles, ν, are shown. Note that the number of bundles hardly influences the

execution times. With the equal-partitioning technique, the performance is mainly dependent on the number of paths and the dimensionality. For that reason, we can thus exploit the GPU parallelism reaching a speedup of around 40 times for the 50-dimensional problem.

6 Conclusions

In this work, we have presented an efficient implementation of the Stochastic Grid Bundling Method on a GPU architecture. Through the GPU parallelism, we can speed up the execution times when the number of bundles and the dimensionality increase. In addition, we have proposed a new bundling technique which is more efficient in terms of memory usage and parallelism. These two improvements enable the use of SGBM for more involved problems, like, for example, counterparty risk and CVA computations.

Acknowledgements The first author is supported by the European Union in the FP7-PEOPLE-2012-ITN Program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN STRIKE—Novel Methods in Computational Finance). The authors would like to thank Shashi Jain, ING Bank, for providing support and the original codes of the Stochastic Grid Bundling Method.

References

1. Cartesius webpage (2014) https://www.surfsara.nl/systems/cartesius
2. CUDA webpage (2014) http://www.nvidia.com/object/cuda_home_new.html
3. cuRAND webpage (2014) https://developer.nvidia.com/curand
4. Fang, F., Oosterlee, C.W.: Pricing early-exercise and discrete Barrier options by Fourier-cosine series expansions. Numer. Math. **114**(1), 27–62 (2009)
5. Jain, S., Oosterlee, C.W.: The stochastic grid bundling method: efficient pricing of Bermudan options and their Greeks. Appl. Math. Comput. **269**, 412–431 (2015)
6. Thrust webpage (2014) http://thrust.github.io/

Proper Orthogonal Decomposition in Option Pricing: Basket Options and Heston Model

J.P. Silva, E.J.W. ter Maten, M. Günther, and M. Ehrhardt

Abstract The finance world, relying more and more on mathematical models, also expects them to be fast, robust and cheap, especially for calibration purposes. The recent revolution in Graphical Processing Units (GPU) and Field-Programmable Gate Array (FPGA) has helped to reduce time and costs but it is the algorithms that ultimately prevail. In this respect, Model Order Reduction (MOR) seems to be especially suited to financial problems as it can reduce extremely computational costs (Achdou and Pironneau, Computational methods for option pricing. SIAM frontiers in applied mathematics, vol 30. Society for Industrial and Applied Mathematics, Philadelphia, 2005). We present two cases when MOR can be extremely useful and how Proper Orthogonal Decomposition (POD) stands out as a valid MOR technique in finance (Volkwein, Proper orthogonal decomposition: theory and reduced-order modelling. Lecture notes, Universität Konstanz, 2013). We show the validity of its application to pricing of basket options, as well as to stochastic volatility models (Heston, Rev Financ Stud 6:327–343, 1993), through the solution of a reduced Black-Scholes PDE. Finally, its computational efficiency when compared with some extensively used numerical methods, as well as some of its limitations are discussed.

Keywords Computational finance • Model order reduction • Option pricing • Proper orthogonal decomposition

1 Introduction

Model Order Reduction (MOR) [2] emerged at the end of the twentieth century as an answer to the increasing complexity of models being developed. Higher and higher resolution schemes lead to bigger problems which, in turn, lead to the development of new accurate schemes (non-uniform and refined grids, higher-order schemes, sparse schemes, parallelization, problem-specific hardware, etc.). The goal of MOR is to generate smaller models, faster to solve and, if not with

J.P. Silva • E.J.W. ter Maten (✉) • M. Günther • M. Ehrhardt
Bergische Universität Wuppertal, Gauß Straße, 42219 Wuppertal, Germany
e-mail: silva@math.uni-wuppertal.de; termaten@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de; ehrhardt@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_29

similar, with high enough precision with respect to the original Full Order Model (FOM). The Reduced Order Model (ROM) is then a cheaper and faster proxy of the FOM, making it ideal for multi-query problems: parameter studies, parameter optimization, inverse problems, control problems. In finance, and particularly option pricing, inverse problems arise when calibrating model parameters to market data, with volatility being one of the parameters, for example [1].

Among the different MOR techniques, c.f. [2], Proper Orthogonal Decomposition (POD) stands out as a fairly robust technique as it is one of the few techniques able to tackle general non-linear problems. Due to its data-driven approach, it generates ROM in a tailored way [15].

Similar problems have been tackled before in a different setting. For example, in [7], Reduced Basis are used for the American Options pricing with parameter dependency, in [4] tailored basis are generated based on the Black-Scholes operator applied to the Black-Scholes and Merton models. We present the results for basket options and Heston Model using numerical methods used in practice to compare the numerical advantage of ROM.

In Sect. 2 we describe briefly Proper Orthogonal Decomposition and in Sect. 3 we introduce the models used and present numerical results for these models. Firstly a 2-dimensional European-type basket option with two underlyings which originates a 2d PDE and then a particular model belonging to the class of stochastic volatility models, namely the Heston Model.

2 Proper Orthogonal Decomposition

In practice, most reduced models are generated in a two step approach. In a first step, information from the full order model is retrieved and with that information a basis of a subspace is generated. In a second step, the original model is projected onto the same (different) subspace space spanned by this new basis, a procedure called Galerkin projection (Petrov-Galerkin projection). In that sense, POD is no different, the big difference being the basis generation, as it is generated solely from data.

The POD is a mathematical procedure that, given an ensemble of data, constructs a basis for the ensemble that is optimal in the following sense. Let X be a real Hilbert space, with inner product $(\cdot,\cdot)_X$, and $Y = \left[y_1\ y_2\ \ldots\ y_n \right]$ an ensemble of n snapshots $y_i \in X$. The snapshots contain the solution for different configurations of the problem, i.e., it may contain the solution at different time instances for an evolution problem, it may contain the solution for different parameter values or any other configuration, which we will try to reproduce with our ROM. Then, for some $l \ll n$ a POD basis is an optimal orthonormal basis $\psi_j, j = 1, \ldots, l$ such that the square error between the elements y_i and its l-partial sum of the decomposition of y_i in the space spanned by ψ_j, is minimized, i.e.

$$\min_{\{\psi_k\}_{k=1}^{l}} \mathscr{J}(\boldsymbol{\psi}) = \min_{\{\psi_k\}_{k=1}^{l}} \sum_{i=1}^{n} \left\| y_i - \sum_{j=1}^{l} \left(y_i, \psi_j\right)_X \psi_j \right\|_X^2, \tag{1}$$

subject to $\left(\psi_i, \psi_j\right)_X = \delta_{ij}$. It can be proved that the above minimization problem is equivalent to the eigenvalue problem

$$YY^\top \psi = \lambda \psi.$$

If we factorize Y using a Singular Value decomposition (SVD), we can see that the resulting left-singular vectors form a POD basis, where $\lambda = \sigma^2$, with σ the singular values of Y. For the POD basis, $\mathscr{J}(\psi) = \sum_{i=l+1}^{n} \lambda_i = \sum_{i=l+1}^{n} \sigma_i^2$. The size of the basis l necessary for a good approximation is problem dependent, but we can take the relative error as a good indicator. As we are minimizing a sum of squares, this criterion guarantees that we are maximizing the information on the reconstructed snapshots in the least-squares sense,

$$\mathscr{E}(l) = \frac{\sum_{i=1}^{l} \sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2}. \tag{2}$$

As the singular values are ordered and reflect the relevance of each dimension in the state space, it is sometimes called *relative information measure*.

The second step in constructing a ROM is to project the PDE onto the space spanned by the POD basis. Rewriting our PDEs as

$$\frac{\partial}{\partial t} V = \mathscr{L} V, \tag{3}$$

where $\mathscr{L}$ is a linear operator, we project in a Galerkin fashion, i.e.

$$\left(\psi_i, \frac{\partial V}{\partial t}\right)_X = \left(\psi_i, \mathscr{L} V\right)_X, \qquad i = 1, \ldots, l.$$

Substituting V by its representation in the POD basis of size l , $V = \sum_j^l a_j(t)\, \psi_j(\mathbf{s})$ and bearing in mind the orthogonality of the basis, we obtain the explicit system of ODEs

$$\dot{a}_i = \sum_{j=1}^{l} a_j(t) \left(\psi_i, \mathscr{L} \psi_l\right)_X \qquad i = 1, \ldots, l.$$

The inner product exhibits two roles in the construction of the ROM. First by defining the POD basis optimality and secondly in the projection step of the PDE. Besides, there are two ways in which we can treat our projection step, before or after semi-discretizing our original, continuous PDE. In our numerical results we will use the former, where we use the method of lines (MOL) to discretize our PDE in space.

3 POD in Option Pricing

The first FOM we want to reduce is an European-type basket option. The basket option is the generalization of the usual option contract with one underlying to n underlyings, usually n correlated assets, see [10]. So, assuming a financial contract with n underlyings following geometrical Brownian motions (GBM) we obtain its price $V = V(t, \mathbf{s})$ as the solution of the PDE

$$\frac{\partial V}{\partial t}(t, \mathbf{s}) + \sum_{i}^{n} (r - q_i) s_i \frac{\partial V}{\partial s_i}(t, \mathbf{s}) + \frac{1}{2} \sum_{i,j}^{n} \rho_{ij}^{\star} \sigma_i \sigma_j s_i s_j \frac{\partial^2 V}{\partial s_i \partial s_j}(t, \mathbf{s}) - rV(t, \mathbf{s}) = 0 \tag{4}$$

with $s_i \in [0, \infty), t \in [0, T], \rho_{ij}^{\star} = 2\rho_{ij}, i \neq j$. r is the risk-free interest rate, ρ_{ij} is the correlation between stochastic processes S_i and S_j, q_i is the dividend yield of S_i and σ_i is the annualized standard deviation of logarithmic returns of S_i.

As this parabolic PDE is, most of the time, supplied with a terminal condition at $t = T$, we will integrate it backwards in time. Depending on the characteristics of the financial contract, we supply (4) with appropriate boundary and terminal conditions.

The second model comes as a result of GBM being a very restrictive model in what concerns the paths of the underlying. A possible extension is to assume that volatility is not constant but follows its own stochastic process, giving origin to the class of Stochastic Volatility Models. Introducing a square root variance model with a mean reverting process for variance one obtains the *Heston PDE* [9]

$$\begin{aligned} \frac{\partial V}{\partial t}(t, \nu, S) = &- \frac{1}{2} \nu S^2 \frac{\partial^2 V}{\partial S^2}(t, \nu, S) - \rho \sigma \nu S \frac{\partial^2 V}{\partial \nu \partial S}(t, \nu, S) - \frac{1}{2} \sigma^2 \nu \frac{\partial^2 V}{\partial \nu^2}(t, \nu, S) \\ &- (r_d - r_f) S \frac{\partial V}{\partial S}(t, \nu, S) - \kappa(\theta - \nu) \frac{\partial V}{\partial \nu} + rV(t, \nu, S), \end{aligned} \tag{5}$$

with ρ the correlation between Wiener processes, κ the reversion speed of the volatility to its long-term mean, θ the long-term mean of variance and σ^2 the variance of variance. Here we denote the risk-free interest rate r as r_d as the dividend rate q as r_f as they take the role of domestic and foreign interest rates, respectively.

We used the MOL in space to discretize Eqs. (4) (for $n = 2$) and (5) obtaining a system of ODEs, with second order approximations for both first and second derivatives. We took n_i discretization points in direction x_i, resulting in a grid of size $N = \prod_i n_i$, including Dirichlet boundary conditions. Our PDE becomes then a system of ODEs with size equal to the total number of (interior) discretization points $N_{int} = \prod_i (n_i - 2)$, which can easily be written in a state-space formulation, common to most MOR techniques,

$$\dot{v} = Av + b \qquad v, b \in \mathbb{R}^{N_{int}}, \quad A \in \mathbb{R}^{N_{int} \times N_{int}} \tag{6}$$

where A has a sparse structure.

In this setting, we have $X = \mathbb{R}^{N_{int}}$ with the Euclidean inner product, $(x_1, x_2)_X = x_1^T x_2$. Setting $v = \psi a$ and projecting Eq. (6) we obtained the reduced ODE system

$$\underbrace{I_N}_{\psi^T\psi}\dot{a} = \underbrace{\tilde{A}}_{\psi^T A\psi} a + \underbrace{\tilde{b}}_{\psi^T b}. \tag{7}$$

We first solve (6), whose solution we will call truth solution, and then proceed to solve (7) using the same integration scheme as in (6). There is no need to ensure we use the same integration scheme however that will be generally the case either in third party software or for ease of implementation.

3.1 2D Basket Option

We solve (4) for two underlyings (2D PDE) in a uniform grid with n_1 points in s_1-direction and n_2 points in s_2-direction for the spatial domain $\Omega = [0, 6K] \times [0, 6K]$, where K is a problem dependent parameter. Although different terminal conditions are possible (geometric average, arithmetic average, max, etc...), we chose the standard weighted average with weights w_i, reflecting the weight of each underlying in the portfolio, which the basket option comprises. So with a put option payoff with strike price K as terminal condition, i.e.

$$V(T, S_1, S_2) = \max(K - \omega_1 S_1 - \omega_2 S_2, 0), \quad \omega_1 + \omega_2 = 1, \ \omega_i > 0$$

and following boundary conditions (V^* is a 1D put option with a rescaled strike price, $K^* = K/\omega_1$)

$$V(t, S_{1min} = 0, S_2) = \omega_2 V^\star(t, S_2) \quad V(t, S_{1max} = 6K, S_2) = 0$$
$$V(t, S_1, S_{2min} = 0) = \omega_1 V^\star(t, S_1) \quad V(t, S_1, S_{2max} = 6K) = 0.$$

We used the following set of parameters (Table 1)

We proceed to solve the FOM with a trapezoidal integration in time with 100 time steps, retrieve our snapshots at each time level $t_i = i\Delta t$, $\Delta t = \frac{T}{100}$, generate the basis, project and solve the ROM with the same trapezoidal integration in time. We used all equally time spaced snapshots available to generate our basis. In Fig. 1, we display on the left axis the maximum absolute error between the FOM and the ROM at the final time $t = 0$ for increasing number of basis vectors Ψ_j and the corresponding squared singular values, σ^2, on the right axis.

Table 1 2D parameters

ρ	σ_1	σ_2	r	K	T	ω_1	n_1	n_2
0.5	0.1	0.2	0.025	100	1	0.25	20	40

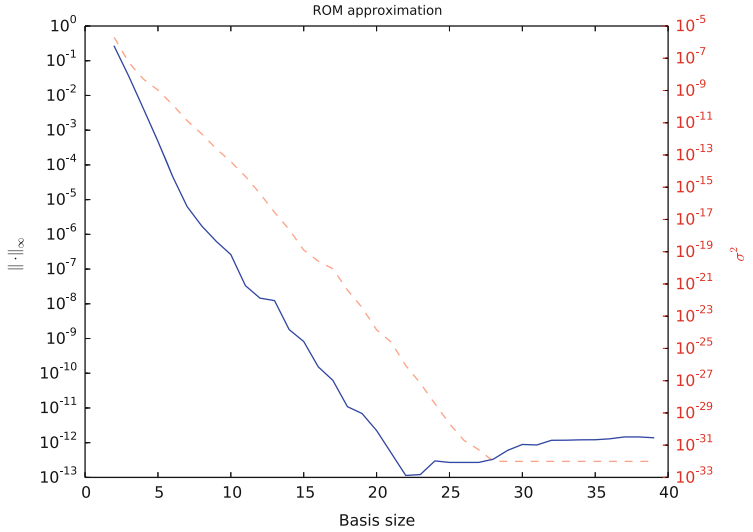


Fig. 1 Absolute Error $\|V_{FOM}(0, S_1, S_2) - V_{ROM}(0, S_1, S_2)\|_\infty$ at final time $t = 0$ for reduced 2D Basket Option

First of all, we can observe an exponential decay in the singular values, a condition necessary for our FOM to possess the so-called *sparse representation property* [4]. Secondly, we can observe that only 20 basis vectors are enough to achieve a 10^{-12} precision.

3.2 *Heston Model*

In our second case, following [11], we applied an Alternating Direction Implicit (ADI) type of scheme, Modified Craig-Sneyd (MCS), to solve the ODE system resulting from the spatial discretization of (5). More details regarding ADI schemes with mixed derivatives and the MCS scheme in particular can be found in [5, 12, 13]. Contrary to the Basket Option case, we used a non-uniform spatial grid based on the hyperbolic sine with focus around $\nu = 0$ and $S = K$ [11]. We used a spatial domain $\Omega = [0, 15] \times [0, 30K]$ with a discretization consisting of 25 points in the ν direction and 50 in the S direction. All the following results are for $n_t = 1000$.

To supply our PDE with the appropriate conditions, we define for our terminal condition as a *call* option payoff $V(T, \nu, S) = \max(S - K, 0)$ and for the boundary conditions

$$V(t, 15, S) = Se^{-r_f t}, \quad V(t, \nu, 0) = 0, \quad \frac{\partial V}{\partial S}(t, \nu, 30K) = e^{-r_f t}.$$

Note that we do not impose any boundary condition at $v = 0$, a degenerate point of our PDE, as numerically we just use the degenerate PDE along $v = 0$, cf. [6]. As in [11], we tested our reduced models with four different set of parameters, originally taken from [3] (Table 2).

Figure 2 presents the results for the absolute error for the solution at final time $t = 0$ for each of the four cases. All the numerical integration is done in the full grid, we decided to evaluate the error only in a region of interest $[0, 1] \times [0, 6K]$, instead of on the original grid domain $[0, 15] \times [0, 30K]$, as those range of values would be of little meaning in financial markets. Even though we are in the realm of numerical analysis, we should note one thing about applying these methods in finance. With some exceptions (American-type options), we are mostly interested

Table 2 Heston model parameters [11]

	ρ	σ	r_d	r_f	θ	κ	K	T
Case 1	−0.13	0.49	0.02	0.04	0.02	6.02	100	0.25
Case 2	−0.67	0.62	0.01	0.02	0.02	1.50	100	1
Case 3	−0.55	1.26	0.01	0.06	0.09	0.38	100	4
Case 4	0.78	0.15	0.1	0.02	0.06	0.3	100	5

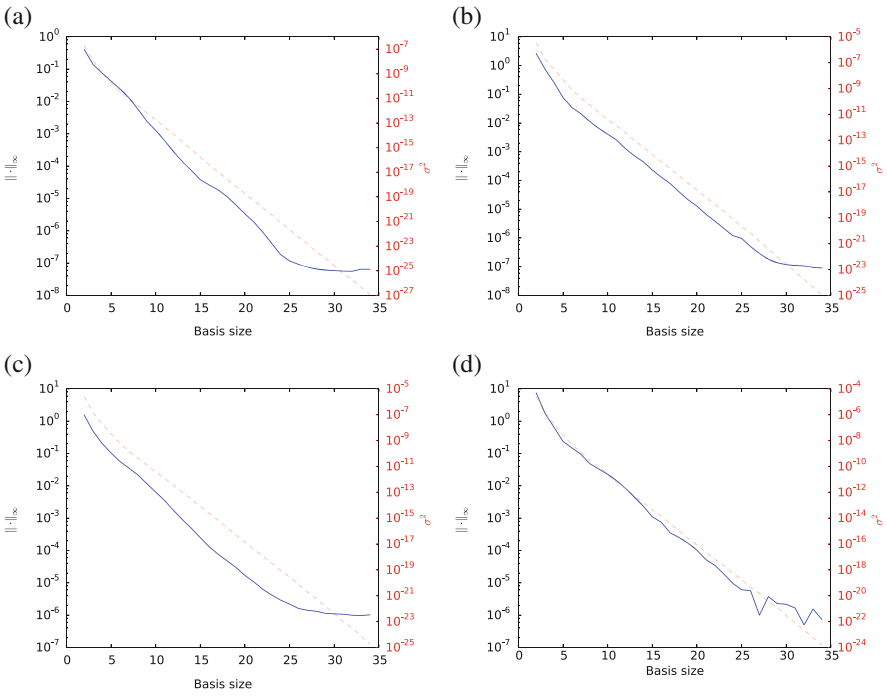


Fig. 2 Absolute Error $\|V_{FOM}(0, v, S) - V_{ROM}(0, v, S)\|_\infty$ at time $t = 0$ for reduced Heston Model and $(v, S) \in [0, 1] \times [0, 6K]$. (**a**) Case 1. (**b**) Case 2. (**c**) Case 3. (**d**) Case 4

in the solution of our PDE at the final time $t = 0$ or at a few selected times. This provides an opportunity to optimize our choice of snapshots in order to minimize the error at these selected times, procedure that was not taken into account in this paper.

Comparing these results with the FOM in [11], we can see that we have a very good approximation with an error of similar magnitude to the temporal discretization error in the FOM. We also would like to observe that in cases 1, 2 and 4, the maximum error is attained near the focus of the grid while on case 3 it happens at the corner of the domain, (1, 6 K). The error of case 3 might then be even smaller if a smaller region of interest is considered, according to the values of ν and S desired.

3.3 ADI and MOR for Higher-Dimensional Problems

Splitting methods, and ADI in particular, have recently been used in finance [8] as it lightens the weight of the curse of dimensionality by having to solve, at each time step, only tridiagonal systems implicitly, some of them time-independent. Following conventions in [14], we evaluate and compare the computational cost of using an ADI MCS scheme to solve a full model and the cost for of solving the corresponding reduced one. In what follows, we take d as the number of dimensions in our problem, n_t the number of time-steps in our time-stepping scheme and n_d the number of discretization points in each dimension. We will assume n_d is the same in all dimensions.

After space discretization, the operator A is decomposed into A_0, which contains all the mixed derivatives terms, and A_j, which contains the spatial operator in dimension j. Therefore, we observe that the MCS scheme consists, at each time step, of

$$
\begin{aligned}
Y_0 &= U_{n-1} + \Delta t F\left(t_{n-1}, U_{n-1}\right) \\
Y_j &= Y_{j-1} + \theta \Delta t \left(F_j\left(t_n, Y_j\right) - F_j\left(t_{n-1}, U_{n-1}\right)\right) \qquad j = 1, 2, \ldots, d \\
\hat{Y}_0 &= Y_0 + \Delta t A_0 \left(F_0\left(t_n, Y_3\right) - F_j\left(t_{n-1}, U_{n-1}\right)\right) \\
\tilde{Y}_0 &= \hat{Y}_0 + \left(\frac{1}{2} - \theta\right) \Delta t \left[F\left(t_n, Y_3\right) - F\left(t_{n-1}, U_{n-1}\right)\right] \\
\tilde{Y}_j &= \tilde{Y}_{j-1} + \theta \Delta t \left(F_j\left(t_n, \tilde{Y}_j\right) - F_j\left(t_{n-1}, U_{n-1}\right)\right) \qquad j = 1, 2, \ldots, d \\
V_n &= \tilde{Y}_d
\end{aligned}
$$

- two implicit integrations corresponding to A (predictor step)
- one explicit integration of the mixed derivative term A_0
- two implicit integrations per direction/dimension corresponding to A_j

Dealing with a PDE of up to second order and with second order approximation for the derivative operators will result in tridiagonal systems for each A_j. For the mixed derivatives, the PDE may contain up to $\frac{d(d-1)}{2}$ mixed derivatives and each mixed derivative discretization will originate 4 new points so, in total, we have $2d\,(d-1)$ diagonals in A_0. This situation occurs in financial PDEs unless the correlation between each stochastic process is zero, hence we will always have all mixed derivatives terms. In the following calculation we assume time independence of our matrices, which allows for a LU decomposition a priori and corresponds to time-independent parameters in our PDE, as happens in both our examples. The computational cost will then be:

1. Once

$$d\frac{3^2 n_d^d}{2}$$

   as for the LU decomposition for the tridiagonal matrices $A_j, j = 1, \ldots, d$
2. At each time step

   a. Two Explicit steps for A

$$2\,(2d\,(d-1) + 2d + 1)\,n_d^d$$

   b. Two Implicit steps per dimension

$$2d5n_d^d = 10dn_d^d$$

   c. One explicit step for A_0

$$2d\,(d-1)\,n_d^d$$

Hence, the total cost is

$$\begin{aligned} f(\cdot) &= d\frac{3^2 n_d^d}{2} + n_t\left((6d\,(d-1) + 4d + 2)\,n_d^d + 10dn_d^d\right) \\ &= \frac{9}{2}dn_d^d + \left(6d^2 + 8d + 2\right)n_t n_d^d \\ &= dn_d^d\left(\frac{9}{2} + 8n_t + 6dn_t\right) + 2n_t n_d^d \end{aligned}$$

We now represent graphically the computational cost of ADI vs a reduced model generated with basis of different sizes. Note that the reduced model will have much less sparsity than the FOM. The ADI method exploits this sparsity and, therefore, only when the dense ROM can compete with the ADI, can it be of practical use.

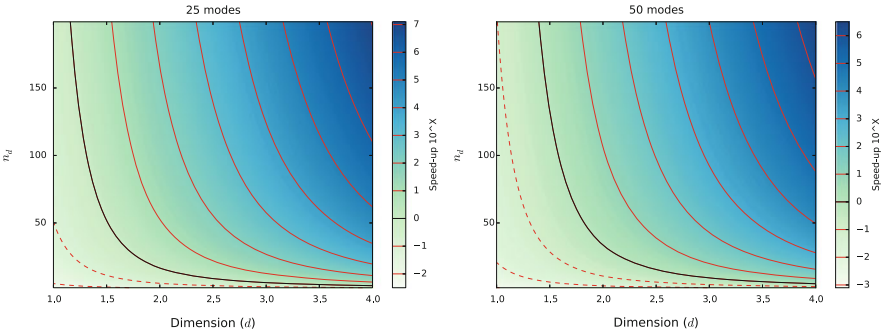


Fig. 3 Computational advantage of MOR

We see that we can achieve a significant reduction in the number of operations (speed-up) already for a two-dimensional problem even in the 50 modes case. Due to the exponential dependence on dimension for the ADI method and the respective independence for the reduced model, we can theoretically obtain better and better results the higher the dimension of the problem. Although higher dimensional ADI methods still lack some rigorous proofs of their properties (stability, consistency), in practice they have been applied with success [8] and so, at least for up to four dimensional problems, we can regard Fig. 3 as showing realistic cases of application.

4 Conclusion

We generated reduced models using POD for two of the most common mathematical models in finance: Basket Options and Heston Model. In both cases it was shown that 25 basis elements at most are needed to obtain the best approximation. We also showed that even for numerical schemes regarded as computationally efficient (ADI) we can obtain significant gains already on 3 and 4 dimensional problems [8]. The advantage is even more clear in a multi-query problem as the cost of SVD is diluted over each online calculation. We expect that to be the case in parametric ROM, which will be subject to future work.

Acknowledgements The work of the authors was partially supported by the European Union in the FP7-PEOPLE-2012-ITN Program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN STRIKE—Novel Methods in Computational Finance). The authors would like to thank Prof. Karel in 't Hout for providing the ADI MCS code for the Heston Model.

References

1. Achdou, Y., Pironneau, O.: Computational Methods for Option Pricing. SIAM Frontiers in Applied Mathematics, vol. 30. Society for Industrial and Applied Mathematics, Philadelphia, PA (2005)
2. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Advances in Design and Control. Society for Industrial and Applied Mathematics, Philadelphia, PA (2005)
3. Clark, I.J.: Foreign Exchange Option Pricing. Wiley, Chichester (2011)
4. Cont, R., Lantos, N., Pironneau, O.: A reduced basis for option pricing. SIAM J. Financ. Math. **2**(1), 287–316 (2011)
5. Craig, I., Sneyd, A.: An alternating-direction implicit scheme for parabolic equations with mixed derivatives. Comput. Math. Appl. **16**(4), 341–350 (1988). doi:10.1016/0898-1221(88)90150-2
6. Daskalopoulos, P., Feehan, P.M.N.: Existence, uniqueness, and global regularity for degenerate elliptic obstacle problems in mathematical finance. arXiv e-prints (2011)
7. Haasdonk, B., Salomon, J., Wohlmuth, B.: A reduced basis method for the simulation of American options. In: ENUMATH 2011 Proceedings (2012)
8. Haentjens, T.: ADI schemes for the efficient and stable numerical pricing of financial options via multidimensional partial differential equations. Ph.D. thesis, University of Antwerp (2013)
9. Heston, S.L.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)
10. Hull, J.C.: Options, Futures, and Other Derivatives. Pearson Studium, Upper Saddle River (2009)
11. in't Hout, K.: ADI schemes for pricing options under the Heston model. In: WBS Quants Hub Workshop (2013)
12. in 't Hout, K., Welfert, B.: Unconditional stability of second-order ADI schemes applied to multi-dimensional diffusion equations with mixed derivative terms. Appl. Numer. Math. **59**(3–4), 677–692 (2009). doi:10.1016/j.apnum.2008.03.016
13. McKee, S.: Alternating direction methods for parabolic equations in two space dimensions with a mixed derivative. Comput. J. **13**(1), 81–86 (1970). doi:10.1093/comjnl/13.1.81
14. Rathinam, M., Petzold, L.R.: A new look at proper orthogonal decomposition. SIAM J. Numer. Anal. **41**, 1893–1925 (2003)
15. Volkwein, S.: Proper Orthogonal Decomposition: Theory and Reduced-Order Modelling. Lecture Notes, Universität Konstanz (2013)

MS 8
MINISYMPOSIUM: IMAGING AND INVERSE PROBLEMS

Organizers

A. Carpio[1] and M.L. Rapun[2]

Speakers

P. González-Rodríguez[3]
The Inverse Source Problem in Mesoscopic Scattering Regimes Using Angle-Resolved Measurements

Antonio Marquina[4]
Variational Models and Numerical Algorithms for One-Dimensional Signal Reconstruction of Noisy and Blurry Signals: Application to the Signal Recovery for Future Detection of Gravitational Waves

Susana Serna[5]
The Regularized Split Bregman Method Based on Rational Approximations of the Absolute Value Function for Total Variation Image Restoration

[1]Departamento de Matemática Aplicada, Universidad Complutense de Madrid, Madrid, Spain
e-mail: ana_carpio@mat.ucm.es

[2]Universidad Politécnica de Madrid, Instituto Gregorio Millán, Madrid, Spain mariaaluisa.rapun@upm.es

[3]Universidad Carlos III

[4]Universidad de Valencia

[5]Universidad Autónoma de Barcelona

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_30

Ana Carpio[6] and María-Luisa Rapún[7]
Domain and Parameter Reconstruction in Photothermal Imaging

Peter Monk and Virginia Selgas[8]
Transmission Eigenvalues for a Dielectric Object Resting on a Perfect Conductor

B. Tomas Johansson[9]
Source Reconstruction from Final Data in the Heat Equation

Keywords

Imaging problem
Inverse problem

Short Description

Imaging problems appear in many real life applications, for example, medical imaging, material testing, geophysical or astrophysical studies. Depending on the context, these problems may adopt different mathematical forms. Often, information about objects is to be found from knowledge of the influence that those obstacles have on propagating waves. A different situation arises when an existing image or signal has to be restored to remove noise. In this minisymposium, we present a summary of recent results in both frameworks.

P. Gonzalez will discuss optical imaging of tissues. A. Marquina analyzes the reconstruction of noisy and blurry signals by means of L^1 regularizations. S. Serna pursues this topic resorting to total variation image restoration. M.L. Rapun considers photothermal imaging. V. Selgas presents some results on scattering by dielectric objects. Finally, T. Johansson will address source reconstruction from final data in diffusion problems.

The Inverse Source Problem in Mesoscopic Scattering Regimes Using Angle-Resolved Measurements
P. González-Rodríguez, Universidad Carlos III

We study optical imaging of tissues in the mesoscopic scattering regime in which light multiply scatters in tissues, but is not fully diffusive. Our purpose is to show

[6]Universidad Complutense
[7]Universidad Politécnica de Madrid
[8](University of Delaware) and (Universidad de Oviedo)
[9](Linköping University)

that, in this regime, using angle-resolved data improves the results considerably. To prove this we compare the solution of two similar inverse source problems that are solved using the same ℓ^1-optimization method. In the first problem we find a solution of a linear system $Ax = b$ in which the initial entries $A_{i,j}$ of the matrix are the scalar flux response at a boundary grid point ρ_i due to an isotropic point source r_i governed by the radiative transport equation. Once the matrix is calculated, we normalize the columns. The right hand side b is the angle averaged data. The second problem $Bx = c$ is similar, but in this case the entries $B_{i,j}$ of the matrix are the directional response at a boundary grid point ρ_i due to an isotropic point source r_i also governed by the radiative transport equation. As with A, we normalize the columns of B. The right hand side c represents the angle-resolved measurements. We show that recovering the location and strength of several point-like sources is not possible when using angle-averaged measurements, while just using two angled-resolved measurements the results are improved radically. In both problems, the matrices A and B are calculated efficiently computing the RTE Green's functions as an expansion of plane wave solutions.

Variational Models and Numerical Algorithms for One-Dimensional Signal Reconstruction of Noisy and Blurry Signals: Application to the Signal Recovery for Future Detection of Gravitational Waves

Antonio Marquina, Universidad de Valencia

In this research work we examine the one dimensional variational models for reconstruction of signals using L^1-regularizations. We present an analysis of the variational models based on L^1-regularization and we implement numerical algorithms that allow to recover noisy and blurry signals, using direct methods and regularization procedures. We shall present an application for the recovery of one-dimensional signals to be observed in the near future in different gravitational wave detectors.

The Regularized Split Bregman Method Based on Rational Approximations of the Absolute Value Function for Total Variation Image Restoration

Susana Serna, Universidad Autónoma de Barcelona

We explore the regularization of the "shrinkage" function based on an approximation of the absolute value function to design a class of split Bregman methods for total variation image restoration. We introduce a hierarchy of regularizations depending on a positive parameter that determines the accuracy in the approximation of the absolute value function by rational functions. We present a set of numerical tests involving the restoration of signals and synthetic images contaminated with noise and blur.

Domain and Parameter Reconstruction in Photothermal Imaging
Ana Carpio (Universidad Complutense) and María–Luisa Rapún (Universidad Politécnica de Madrid)

Photothermal imaging aims to reconstruct the inner structure of materials by heating their surface using a laser beam and recording the surface temperature. The goal is to detect structural defects or inclusions (determine their location, size, shape, orientation) and their nature (physical parameters).

In this work we propose an iterative descent method that combines topological derivative computations to reconstruct the geometry of the defects with gradient iterations to approximate the material parameters.

Some numerical experiments showing the ability of the method to obtain reasonable reconstructions in a few iterations will be shown. Furthermore, we numerically corroborate that a small number of sampling points and source points allow for reliable reconstructions if we record the temperature during a time interval.

Transmission Eigenvalues for a Dielectric Object Resting on a Perfect Conductor
Peter Monk (University of Delaware) and Virginia Selgas (Universidad de Oviedo)

We introduce a new transmission eigenvalue problem with mixed boundary conditions that arises when a dielectric scatterer is mounted on a metal structure.

Indeed, we describe the forward problem and show that it has a unique solution using a reflection principle. We also formulate the inverse problem of identifying the shape of the dielectric from near field measurements. To solve numerically this inverse problem, we propose the standard near field Linear Sampling Method (LSM); notice that the equations involved in the LSM and in the approximation of transmission eigenvalues from measurements are one and the same.

Moreover, we reformulate the mixed transmission eigenvalue problem as a fourth order partial differential equation. Then we show that there exist infinitely many transmission eigenvalues and derive monotonicity as well as a lower bound estimate for the first eigenvalue. Our analysis mainly uses techniques from [1, 2, 4], and requires us to prove suitable density and compactness properties.

We also provide numerical examples for the LSM; and finally demonstrate that, for the cases we have considered, mixed transmission eigenvalues can be approximated from near field data; see [3] for a study of the corresponding far field problem for standard transmission eigenvalues.

1. Cakoni, F., Haddar, H.: A variational approach for the solution of the electromagnetic interior transmission problem for anisotropic media. Inverse Probl. Imag. **1**, 443–456 (2007)
2. Cakoni, F., Haddar, H.: On the existence of transmission eigenvalues in an inhomogeneous media. Appl. Anal. **88**, 475–493 (2009)
3. Cossonnière, A.: Valeurs propres de transmission et leur utilisation dans l'identification d'inclusions a partir de mesures électromagnétiques. Ph.D. thesis, Université de Toulouse (2011)

4. Haddar, H.: The interior transmission problem for anisotropic Maxwell's equations and its applications to the inverse problem. Math. Methods Appl. Sci. **27**, 2111–2129 (2004)

Source Reconstruction from Final Data in the Heat Equation

B. Tomas Johansson (Linköping University)

We consider the inverse ill-posed problem of determining an unknown source term in the linear heat conduction equation from final time data (together with known boundary and initial conditions) having applications in pollutant source identification and in the design of melting and freezing processes. We shall review a recent extension of a uniqueness result for source reconstruction. Moreover, an iterative method together with some numerical results for the reconstruction of a source from final data will be given.

Domain and Parameter Reconstruction in Photothermal Imaging

Ana Carpio and María-Luisa Rapún

Abstract In this work we address the inverse problem of reconstructing inclusions and their thermal parameters given temperature measurements at the accessible side of a material. We describe an iterative descent method that combines topological derivative computations to reconstruct the geometry of the defects with gradient iterations to approximate the material parameters. A numerical experiment showing the ability of the method to obtain reasonable reconstructions in a few iterations is presented.

Keywords Inverse problem • Photothermal imaging

1 Statement of the Problem

Photothermal imaging techniques are suitable means of inspecting composite materials with nondestructive tests. In this work we develop techniques to detect defects Ω buried in a medium by surface thermal measurements. We are interested in a photothermal technique that consists in heating the surface of a semi–infinite medium by a laser beam and recording the temperature at several receptors located on the same surface during a time interval, see Fig. 1. Recent physical experiments using this kind of technique can be found in [9, 18].

The forward problem is modelled by a heat diffusion equation in the half plane $\mathbb{R}^2_- := \{(x, y) \in \mathbb{R}^2, \ \ y < 0\}$. The surface of the sample $\Pi := \{(x, 0), \ \ x \in \mathbb{R}\}$ is thermically excited with a delta–pulse located at a source point $\mathbf{x}_0 \in \Pi$, generating a thermal wave of the form

$$U_{inc}(\mathbf{x}, t) = (1/t)\, \exp\left(-\rho_e |\mathbf{x} - \mathbf{x}_0|^2/(4\kappa_e t)\right), \qquad \mathbf{x} \in \mathbb{R}^2, \ t > 0. \tag{1}$$

A. Carpio
Departamento de Matemática Aplicada, Fac. Matemáticas, Universidad Complutense de Madrid, 28040 Madrid, Spain
e-mail: ana_carpio@mat.ucm.es

M.-L. Rapún (✉)
Departamento de Fundamentos Matemáticos, ETSI Aeronáuticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain
e-mail: marialuisa.rapun@upm.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*, Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_31

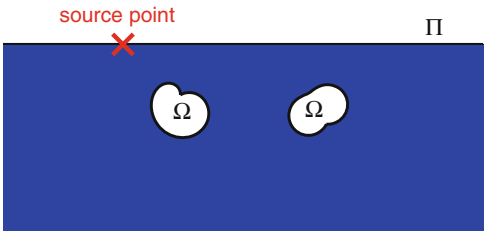


Fig. 1 Geometrical configuration. The source and observation points are located on the boundary Π of the medium. The objects Ω and their physical parameters are the unknowns of the photothermal imaging problem

Here κ_e is the thermal conductivity of the exterior medium $\Omega_e := \mathbb{R}^2_- \setminus \Omega$ and ρ_e is the density multiplied by its specific heat. The corresponding thermal parameters inside the inclusions Ω are κ_i and ρ_i. The temperature distribution

$$U(\mathbf{x},t) := \begin{cases} U_+(\mathbf{x},t), & \text{in } \Omega_e \times (0,\infty), \\ U_-(\mathbf{x},t), & \text{in } \Omega \times (0,\infty), \end{cases}$$

satisfies the heat equations

$$\rho_e \partial_t U_+ = \kappa_e \Delta U_+, \ \text{ in } \Omega_e \times (0,\infty), \qquad \rho_i \partial_t U_- = \kappa_i \Delta U_-, \ \text{ in } \Omega \times (0,\infty). \tag{2}$$

In the exterior domain Ω_e, the total temperature $U_{total} = U_+ + U_{inc}$ is the superposition of U_+ and the incident wave defined in (1). The temperature satisfies the following transmission conditions at the common interface:

$$U_- - U_+ = U_{inc}, \quad \kappa_i \partial_{\mathbf{n}} U_- - \kappa_e \partial_{\mathbf{n}} U_+ = \kappa_e \partial_{\mathbf{n}} U_{inc}, \qquad \text{on } \partial\Omega \times (0,\infty). \tag{3}$$

The forward problem is completed imposing an adiabatic boundary condition on the upper boundary Π, and homogeneous initial conditions:

$$\partial_{\mathbf{n}} U_+ = 0, \quad \text{on } \Pi \times (0,\infty), \qquad U_+(\mathbf{x},0) = U_-(\mathbf{x},0) = 0, \quad \forall \mathbf{x} \in \mathbb{R}^2_-. \tag{4}$$

The solution U of the forward problem can be numerically approximated using the following strategy [12, 14]: if we consider the Laplace transform of U and U_{inc}, $u(\mathbf{x},s) = \int_0^\infty e^{-st} U(\mathbf{x},t)dt$ and $u_{inc,s}(\mathbf{x}) = \int_0^\infty e^{-st} U_{inc}(\mathbf{x},t)dt$, then, for each value of s the function $u_s(\mathbf{x}) := u(\mathbf{x},s)$ is a radiating solution of the stationary problem

$$\begin{cases} \kappa_e \Delta u_s - s\rho_e u_s = 0, & \text{in } \Omega_e, \\ u_s^- - u_s^+ = u_{inc,s}, & \text{on } \partial\Omega \\ \partial_{\mathbf{n}} u_s = 0, & \text{on } \Pi. \end{cases} \qquad \begin{cases} \kappa_i \Delta u_s - s\rho_i u_s = 0, & \text{in } \Omega, \\ \kappa_i \partial_{\mathbf{n}} u_s^- - \kappa_e \partial_{\mathbf{n}} u_s^+ = \kappa_e \partial_{\mathbf{n}} u_{inc,s}, & \text{on } \partial\Omega, \end{cases} \tag{5}$$

To invert the Laplace transform we choose the hyperbolic paths of the form [17]: $\gamma(\theta) := \mu(1-\sin(\pi/4+\imath\theta)),\ \theta \in \mathbb{R}$, where $\mu > 0$. Then, the solution of (2)–(4) is

$$U(\mathbf{x},t) = \frac{1}{2\pi\imath}\int_{-\infty}^{\infty} e^{t\gamma(\theta)}u(\mathbf{x},\gamma(\theta))\gamma'(\theta)d\theta.$$

Numerical approximations of U can be calculated using a truncated trapezoidal rule

$$U(\mathbf{x},t) \approx \sum_{\ell=-L}^{L} c_\ell e^{ts_\ell}u(\mathbf{x},s_\ell),$$

$s_\ell = \gamma\left(\frac{\log(L)}{L}\,\ell\right)$ and weights $c_\ell = \frac{\log(L)}{2\pi\imath L}\gamma'\left(\frac{\log(L)}{L}\ell\right)$.

The inverse problem consists in finding the objects Ω and the parameters κ_i, ρ_i such that the solution of the forward transmission problem (2)–(4) equals the measured values of the total wave $U_{meas}(\mathbf{x}_i,t_j)$ at the detector locations $\mathbf{x}_1,\ldots,\mathbf{x}_M \in \Pi$ at the time instants $t_1,\ldots,t_N$. Since this problem is ill–posed, we consider a weaker variational reformulation: find Ω,κ_i,ρ_i minimizing the functional

$$J(\mathbb{R}^2_- \setminus \Omega,\kappa_i,\rho_i) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} f(t_j)\left(U_{total}(\mathbf{x}_i,t_j) - U_{meas}(\mathbf{x}_i,t_j)\right)^2, \tag{6}$$

where U_{total} is the solution of the forward problem (2)–(4) when the object is Ω and the interior thermal parameters are κ_i and ρ_i. The weight function $f(t)$ normalizes the time decay of the solutions of the heat equation. For our numerical experiment in Sect. 3 we select $f(t) = \max_{\mathbf{x}\in\{\mathbf{x}_1,\ldots,\mathbf{x}_M\}}|U_{meas}(\mathbf{x},t)|^{-1}$. Other possibilities were explored in [2, 6].

Based on the Laplace-transform strategy described above, we proposed in [2, 6] to substitute the cost functional (6) by the approximated functional

$$J(\mathbb{R}^2_- \setminus \Omega,\kappa_i,\rho_i) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} f(t_j)\left(\sum_{\ell=-L}^{L} c_\ell e^{t_j s_\ell}u(\mathbf{x}_i,s_\ell) - U_{meas}(\mathbf{x}_i,t_j)\right)^2, \tag{7}$$

having now $2L+1$ stationary constraints.

2 Iterative Method to Reconstruct Inclusions and Parameters

To solve the optimization problem we combine gradient and topological derivative (TD in the sequel) methods to generate sequences of parameters and objects in such a way that the cost functional decreases throughout the iterative procedure. Our choice of a TD strategy is based on the following advantageous features:

- Without a priori information, the TD provides a good first guess of the number, size and location of the inclusions. This has been tested in a wide range of physical settings, including acoustics, electromagnetism, elastodynamics, electrical impedance tomography, fluorescence optical tomography, and photothermal imaging [1, 5, 7, 8, 10, 15, 20].
- Iterative TD methods allow for topological changes during the iterations, in contrast to classical shape deformation strategies [11, 13, 19] that require knowledge of the number of objects from the start. Using iterative TD based methods, new objects may be created in the course of the iterations, existing contours may merge and holes inside existing objects may be detected, see [3, 4]. Furthermore, even if the number of inclusions is known (assumption that in most practical applications is not realistic), TD-iterative methods are a powerful alternative to these classical methods, providing accurate reconstructions at a low computational cost, as extensively checked by the authors in different contexts (see [2–6] and references therein).
- In comparison with other strategies allowing for topological changes (as i.e. level set methods [16, 21]), the number of iterations with respect to the domain is usually much smaller.

In our previous papers [2, 6] we used a non–standard formulation of the photothermal problem (2)–(4), involving two interior parameters related with $\kappa_e, \kappa_i, \rho_e$, and ρ_i with no physical meaning. In this paper we adapt the results in [2, 6] to deal with the reconstruction of defects and of their physical parameters κ_i and ρ_i.

The TD of a shape functional $\mathscr{J}(\mathscr{R})$ is a pointwise function defined as [22]:

$$D_T(\mathbf{x}, \mathscr{R}) = \lim_{\varepsilon \to 0} \frac{\mathscr{J}(\mathscr{R} \setminus B_\varepsilon(\mathbf{x})) - \mathscr{J}(\mathscr{R})}{\pi \varepsilon^2}, \quad \mathbf{x} \in \mathscr{R}, \tag{8}$$

where $B_\varepsilon(\mathbf{x})$ is a ball centered at $\mathbf{x}$ with radius ε. Then, it follows the expansion:

$$\mathscr{J}(\mathscr{R} \setminus B_\varepsilon(\mathbf{x})) = \mathscr{J}(\mathscr{R}) + D_T(\mathbf{x}, \mathscr{R}) \pi \varepsilon^2 + o(\varepsilon^2), \quad \text{as } \varepsilon \to 0.$$

This motivates the key idea for the reconstruction technique: if we locate small objects $B_\varepsilon(\mathbf{x})$ at the points $\mathbf{x} \in \mathscr{R}$ where $D_T(\mathbf{x}, \mathscr{R})$ is negative, then $\mathscr{J}(\mathscr{R} \setminus B_\varepsilon(\mathbf{x})) < \mathscr{J}(\mathscr{R})$, that is, the value of the functional decreases. Hence we will identify the points where the TD attains the larger negative values with the regions where it is more likely to have an object.

The next result can be proved following Theorem 3.2 in [2].

Theorem 1 *The TD of the functional $J(\mathbb{R}^2_- \setminus \Omega, \kappa_i, \rho_i)$ defined in (7) is*

$$D_T(\mathbf{x}) = Re\left(\sum_{\ell=-L}^{L} \frac{2\kappa_e(\kappa_e - \kappa_i)}{\kappa_e + \kappa_i} \nabla u_{total,s_\ell}(\mathbf{x}) \nabla p_{s_\ell}(\mathbf{x}) + (\rho_e - \rho_i) s_\ell u_{total,s_\ell}(\mathbf{x}) p_{s_\ell}(\mathbf{x}) \right) \tag{9}$$

for $\mathbf{x} \in \mathbb{R}^2_- \setminus \Omega$, *where* $u_{total,s_\ell} = u_{inc,s_\ell} + u_{s_\ell}$ *and* u_{s_ℓ} *is the solution of (5) for* $s = s_\ell$. *The adjoint fields* p_{s_ℓ} *are solutions of:*

$$\begin{cases} \kappa_e \Delta p_{s_\ell} - s_\ell \rho_e p_{s_\ell} = g_{s_\ell}, & \text{in } \Omega_e, \\ p^-_{s_\ell} - p^+_{s_\ell} = 0, & \text{on } \partial\Omega \\ \partial_{\mathbf{n}} p_{s_\ell} = 0, & \text{on } \Pi, \end{cases} \qquad \begin{array}{l} \kappa_i \Delta p_{s_\ell} - s_\ell \rho_i p_{s_\ell} = 0, \quad \text{in } \Omega, \\ \kappa_i \partial_{\mathbf{n}} p^-_{s_\ell} - \kappa_e \partial_{\mathbf{n}} p^+_{s_\ell} = 0, \quad \text{on } \partial\Omega, \end{array} \tag{10}$$

with $g_{s_\ell}(\mathbf{x}) := \sum_{i=1}^{M} \sum_{j=1}^{N} f(t_j) c_{s_\ell} e^{t_j s_\ell} \left(\overline{U_{meas}(\mathbf{x}_i, t_j) - \sum_{k=-L}^{L} c_k e^{t_j s_k} u_{s_k}(\mathbf{x}_i)} \right) \delta_{\mathbf{x}_i}(\mathbf{x})$.

By iteratively applying Theorem 1, we construct a monotone sequence of approximate domains $\Omega_d \subset \Omega_{d+1}$ adding to the current approximation Ω_d the points where the TD attains pronounced negative values. To be able to remove points from Ω_d, we need to compute the TD inside the inclusions. The definition (8) can be extended to the points inside Ω [4], and an analogous expression to (9) can be found for $\mathbf{x} \in \Omega$ [5, 6]. This extension is the basis to develop iterative strategies able to correct an approximation of Ω by removing the points where the TD attains pronounced positive values.

To determine the thermal parameters we proceed as follows. If $\widetilde{\kappa}_i$, $\widetilde{\rho}_i$ are approximate values of κ_i and ρ_i, and $\widetilde{\Omega}$ is an approximation of Ω, then we correct the values $\widetilde{\kappa}_i, \widetilde{\rho}_i$ by a gradient method. The idea is to define $\kappa_i = \widetilde{\kappa}_i + \eta\phi$, $\rho_i = \widetilde{\rho}_i + \eta\psi$, where $\eta > 0$ is small and ϕ, ψ are selected calculating the derivative of $J(\eta) := J(\widetilde{\Omega}, \widetilde{\kappa}_i + \eta\phi, \widetilde{\rho}_i + \eta\psi)$ with respect to η to ensure that $J'(0) < 0$. A procedure to obtain explicit formulae in terms of forward and adjoint fields for this kind of functionals is explained in [4, 6]. In our case, it can be proven that the choice

$$\phi = \mathrm{Re}\left(\int_{\widetilde{\Omega}} \sum_{\ell=-L}^{L} \nabla u_{s_\ell} \nabla p_{s_\ell} \right), \qquad \psi = \mathrm{Re}\left(\int_{\widetilde{\Omega}} \sum_{\ell=-L}^{L} s_\ell \, u_{s_\ell} p_{s_\ell} \right), \tag{11}$$

makes $J'(0) < 0$. Here u_{s_ℓ}, p_{s_ℓ} are solutions of (5) and (10), respectively, with $\Omega = \widetilde{\Omega}$, $\kappa_i = \widetilde{\kappa}_i$ and $\rho_i = \widetilde{\rho}_i$.

Finally, our procedure is as follows. In a first step we consider initial guesses of the parameters $\kappa_i = \kappa_i^0$, $\rho_i = \rho_i^0$ and compute the TD in $\mathbb{R}^2_-$ for these parameters, that is, the TD of $J(\mathbb{R}^2_-, \kappa_i^0, \rho_i^0)$. We find then a first approximation Ω_1 of Ω as the union of all the points where the TD is smaller than a negative constant (see [2, 4] for guidelines of the selection of such constant). Once Ω_1 is set, we update the values of the parameters performing Q iterations of the gradient method ($Q = 8$ in our numerical example in Sect. 3) as explained above:

$$\kappa_i^q = \kappa_i^{q-1} + \eta\phi^q, \quad \rho_i^q = \rho_i^{q-1} + \eta\psi^q, \quad q = 1, \ldots, Q,$$

with ϕ^q, ψ^q defined as in (11) with $\widetilde{\Omega} = \Omega_1$, $\widetilde{\kappa}_i = \kappa_i^{q-1}$ and $\widetilde{\rho}_i = \rho_i^{q-1}$. Once the parameters are corrected, we compute the TD of $J(\mathbb{R}^2_- \setminus \Omega_1, \kappa_i^Q, \rho_i^Q)$ to update the domain Ω_1 by adding to it the points $\mathbf{x} \in \mathbb{R}^2_- \setminus \Omega_1$ where the TD attains the

larger negative values, and removing from Ω_1 the points inside it, if any, where the TD attains the larger positive values. Once the approximation of the domains is improved we perform further gradient iterations to update the parameters and so on. The algorithm stops if any of the following stopping criteria is satisfied:

- $\text{meas}(\Omega_d \setminus \Omega_{d-1})$ is small,
- $|\kappa_i^q - \kappa_i^{q-1}| + |\rho_i^q - \rho_i^{q-1}|$ is small and $\|U_{meas} - U_{total}\|$ is small,
- $J(\mathbb{R}^2_- \setminus \Omega_d, \kappa_i^q, \rho_i^q)$ is small.

3 A Numerical Example

In this section we present a numerical example to illustrate the feasibility of our reconstruction algorithm. We consider a simple geometry where Ω is the ellipse $\Omega = \{(x, y) \in \mathbb{R}^2,\ x^2/0.55^2 + (y + 1)^2/0.35^2 < 1\}$, with thermal parameters $\kappa_i = 1/2$ and $\rho_i = 1$. In the exterior medium the values of the parameters are $\kappa_e = 1$ and $\rho_e = 1/5$.

Synthetic data are created solving (2)–(4) by means of the Laplace transform with respect to time and a boundary element formulation in space (see [2, 12] for details). A relative 1 % Gaussian error was added at each observation point to both avoid inverse crimes and to simulate measurement errors. We have considered six incident waves of the form (1) generated at the uniformly distributed source points represented in all the plots in Fig. 2 by '•' marks. Measurements of the temperature were taken at the seven observation points represented by '×' marks at 10 uniformly distributed times in the time interval [0.05,0.5].

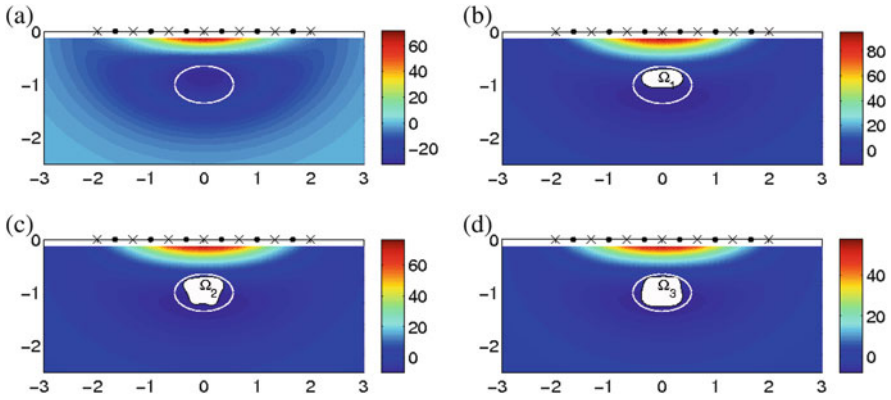


Fig. 2 (**a**) Topological derivative when $\kappa_i^0 = 3/4$, $\rho_i^0 = 1/3$ and $\Omega = \emptyset$. (**b**)–(**d**) Approximated domains Ω_d, $d = 1, 2, 3$, superimposed to the TD computed for $\Omega = \Omega_d$ and the corresponding updated values of the thermal parameters κ_i and ρ_i

We started the algorithm by choosing the initial values $\kappa_i^0 = 3/4$ and $\rho_i^0 = 1/3$. To obtain an initial guess for the domain, we computed the TD in the sampling region $[-3, 3] \times [-2.5, 0]$. This yields the colormap represented in Fig. 2a. Dark blue colors indicate the regions where the TD takes large negative values, at which the objects should be located. The boundary of the true defect is represented by a solid white line in all plots in Fig. 2. Our initial guess Ω_1 is represented in Fig. 2b. We set now $\Omega = \Omega_1$ and perform eight iterations with the gradient method to correct the values of κ_i and ρ_i. In the next step, the values of the parameters are fixed and the TD is again computed. In Fig. 2c we represent the updated object Ω_2. The hybrid algorithm alternates eight iterations with the gradient method with one TD computation. It stopped at the tenth iteration with respect to the domain. The first three approximated domains are represented in Fig. 2. In Fig. 3a we show the true object (solid blue line), the initial guess Ω_1 (dashed green line), and the final reconstruction Ω_{10} (dashed red line). The values of κ_i and ρ_i throughout the iterations are given in Fig. 3b. Two identical values mark a TD computation to update the domain. The final approximations were $\kappa_i^{final} = 0.5703$ (recall that the true value is 0.5), and $\rho_i^{final} = 0.8266$ (while the true value is 1). We have obtained a satisfactory reconstruction taking into account that no a priori knowledge about the number, size or location of the objects is assumed, and that few data were available.

In our example, we found a sequence of enlarging sets, i.e., satisfying $\Omega_d \subset \Omega_{d+1}$. An example where the TD provides a sequence of defects where $\Omega_d \not\subseteq \Omega_{d+1}$ for some values of $d \in \mathbb{N}$ can be found in [6]. The interested reader may find some reconstructions with other geometries, multiple objects and different weight functions $f(t)$ in [2, 6]. Furthermore, a gallery of comparisons varying the different parameters of the problem, namely, the number of source points and/or observation points, the number of time observations, etc, can be found in [2] for a simplified situation where the interior parameters are assumed to be known and in [6] for a related problem with unknown domains and parameters.

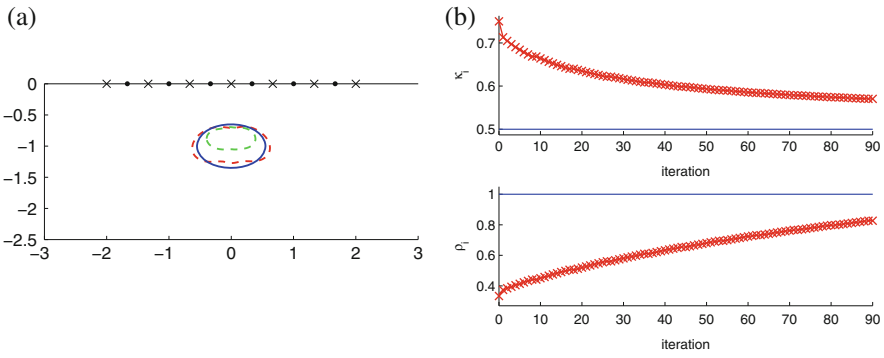


Fig. 3 Final reconstruction after 10 iterations with respect to the domain. (**a**) Initial (*green dashed line*), predicted (*red dashed line*) and true (*blue solid line*) objects. (**b**) Values of the thermal parameters κ_i and ρ_i versus the number of iterations

Acknowledgements The authors are partially supported by the Spanish Government research project TRA2010–18054 and the Spanish Ministerio de Economía y Competitividad Grants No. FIS2011–222888–C02–02 and FIS2010–22438–E.

References

1. Bonnet, M., Constantinescu, A.: Inverse problems in elasticity. Inverse Prob. **21**, 1–50 (2005)
2. Carpio, A., Rapún, M.-L.: Domain reconstruction using photothermal techniques. J. Comput. Phys. **15**, 8083–8106 (2008)
3. Carpio, A., Rapún, M.-L.: Topological derivatives for shape reconstruction. Inverse Problems and Imaging. Lecture Notes in Mathematics, pp. 85–134. Springer, Berlin (2008)
4. Carpio, A., Rapún, M.-L.: Solving inhomogenous inverse problems by topological derivative methods. Inverse Prob. **24**, art. num. 045014 (2008)
5. Carpio, A. Rapún, M.-L.: Hybrid topological derivative and gradient based methods for non-destructive testing. Abstr. Appl. Anal. **2013**, art. num. 816134 (2013)
6. Carpio, A., Rapún, M.-L.: Parameter identification in photothermal imaging. J. Math. Imaging Vision **49**, 273–288 (2014)
7. Feijoo, G.R.: A new method in inverse scattering based on the topological derivative. Inverse Prob. **20**, 1819–1840 (2004)
8. Garreau, S., Guillaume, P., Masmoudi, M.: The topological asymptotic for PDE systems: the elasticity case. SIAM J. Control Optim. **39**, 1756–1778 (2001)
9. Garrido, F., Salazar, A.: Thermal wave scattering by spheres. J. Appl. Phys. **95**, 140–149 (2004)
10. Guzina, B.B., Bonnet, M.: Small–inclusion asymptotics of misfit functionals for inverse problems in acoustics. Inverse Prob. **22**, 1761–1785 (2006)
11. Hettlich, F.: Fréchet derivatives in inverse scattering problems. Inverse Prob. **11**, 371–382 (1995)
12. Hohage, T., Sayas, F.-J.: Numerical approximation of a heat diffusion problem by boundary element methods using the Laplace transform. Numer. Math. **102**, 67–92 (2005)
13. Kirsch, A.: The domain derivative and two applications in inverse scattering theory. Inverse Prob. **9**, 81–96 (1993)
14. Laliena, A., Sayas, F.-J.: LDBEM in diffusion problems. In: Proceedings of XIX CEDYA/IX CMA (electronic version) (2005)
15. Laurain, A., Hintermüller, M., Freiberger, M., Scharfetter, H.: Topological sensitivity analysis in fluorescence optical tomography. Inverse Prob. **29**, art. num. 025003 (2013)
16. Litman, A., Lesselier, D., Santosa, F.: Reconstruction of a two-dimensional binary obstacle by controlled evolution of a level-set. Inverse Prob. **14**, 685–706 (1998)
17. López-Fernández, M., Palencia, C.: On the numerical inversion of the Laplace transform of certain holomorphic mappings. Appl. Numer. Math. **51**, 289–303 (2004)
18. Mendioroz, A., Castelo, A., Celorrio, R., Salazar, A.: Characterization of vertical buried defects using lock-in vibrothermography: I direct problem. Meas. Sci. Technol. **24**, art. num. 065601 (2013)
19. Potthast, R.: Fréchet differentiability of the solution to the acoustic Neumann scattering problem with respect to the domain. J. Inverse Ill-Posed Prob. **4**, 67–84 (1996)
20. Samet, B., Amstutz, S., Masmoudi, M.: The topological asymptotic for the Helmholtz equation. SIAM J. Control Optim. **42**, 1523–1544 (2003)
21. Santosa, F.: A level-set approach for inverse problems involving obstacles. Optimisation et Calcul des Variations **1**, 17–33 (1996)
22. Sokolowski, J., Zolésio, J.P.: Introduction to Shape Optimization. Shape Sensitivity Analysis. Springer, Heidelberg (1992)

Fast Backprojection Operator for Synchrotron Tomographic Data

Eduardo X. Miqueles and Elias S. Helou

Abstract Reduction of computational time in high resolution image reconstruction is essential in basic research and applications as well. This reduction is important for different types of traditional non diffractive tomography in medical diagnosis as well as for applications in nanomaterials research, related to modern technologies. Alternatives to alleviate the computationally intense part of each iteration of iterative methods in tomographic reconstruction have all been based on interpolation over a regular grid in the Fourier domain or in fast nonuniform Fourier transforms. Both approaches speed up substantially the computation of each iteration of classical algorithms, but are not suitable for being used in a large class of more advanced faster algorithms: incremental methods such as OS-EM, BRAMLA or BSREM, among others, cannot benefit from these techniques. The backprojection is a stacking operator, known to be the adjoint of the Radon transform. As a mapping $\mathscr{B}$, the backprojection can be recast as a convolution operator, in a different coordinate system, which is an improvement in accelerating the computation of $\mathscr{B}$. In this work, we propose several analytical representations for the operator $\mathscr{B}$, in order to find a fast algorithm.

Keywords Imaging problem • Image reconstruction

E.X. Miqueles (✉)
CNPEM/LNLS - Brazilian Synchrotron Light Source, Rua Máximo Giuseppe Scolfaro 10.000, Polo de Alta Tecnologia 13083-970, Campinas, SP, Brazil
e-mail: edu.miqueles@gmail.com

E.S. Helou
ICMC - University of São Paulo, Avenida Trabalhador São-carlense, 400 Centro 13566-590, São Carlos, SP, Brazil
e-mail: elias@icmc.usp.br

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_32

1 Introduction

Image reconstruction from projections depends on the computation of the so called stacking operators *Radon transform* and *Backprojection transform*. They are defined, respectively, as

$$g(t,\theta) = \int_{\Omega(t,\theta)} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \equiv \mathscr{R}f(t,\theta), \quad b(\boldsymbol{x}) = \mathscr{B}g(\boldsymbol{x}) = \int_{[0,2\pi]} g(\boldsymbol{x}\cdot\boldsymbol{\xi}_\theta,\theta)\mathrm{d}\theta \qquad (1)$$

with $\Omega(t,\theta) = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x}\cdot\boldsymbol{\xi}_\theta = t\}$ a straight line parameterized by $(t,\theta) \in [-1,1]\times[0,2\pi]$. Here $\xi_\theta = (\cos\theta, \sin\theta)^T$ is the normal vector to Ω and $\boldsymbol{x}$ lies in the unit ball $\|\boldsymbol{x}\|_\infty \leq \frac{\sqrt{2}}{2}$. The pair (b,g) is *dual* in the sense that $\mathscr{B}$ is the adjoint transform of $\mathscr{R}$ at an appropriate functional space, see [4, 8, 12]. Let us denote U as the *feature space*, wherein lie all the map functions of the form $z = z(\boldsymbol{x})$, and denote V as the *Radon space*, gathering all the sinogram functions $g = g(t,\theta)$.

The goal is: find f given the function g over the set $\{\Omega(t,\theta)\}$ satisfying $g = \mathscr{R}f$. Usually $b = b(\boldsymbol{x})$ is used as a first order approximation of $f = f(\boldsymbol{x})$, see [15]. A typical algorithm to find an approximation of f, either analytical or iterative, depends on the computation of $\mathscr{B}g$. This has been recognized in the literature as the main time consuming bottleneck in image reconstruction. Indeed, as the integral of $\mathscr{B}$ has to be computed for each *pixel* $\boldsymbol{x}$, the average cost for reconstruction increases proportionally to the dimensions of the scanning geometry.

Our aim in this paper is to present a fast computation of the backprojection image $b = \mathscr{B}g$ for a given sinogram $g \in V$. It has a great impact at iterative methods, which generally produce better results for noisy data compared to the analytical ones. In fact, a typical iterative procedure is given by $f^{(k+1)} = f^{(k)} + \mathscr{D}\left(f^{(k)}, \mathscr{B}g^{(k)}, \mathscr{R}f^{(k)}\right)$ where $\mathscr{D}$ is an operator defining a direction towards the solution. If the cost per iteration is high, and the reconstruction time is required to be low, such iterations are non-recommended, unless high computing performance is involved. This is one of the reasons why the celebrated filtered backprojection algorithm has been preferred when the data is large and the calculation of $\mathscr{B}g$ is time consuming. Nevertheless, even the filtered backprojection suffers from time delaying since $\mathscr{B}g$ is part of the analytical inversion.

For high-resolution tomographic synchrotron experiments, the amount of data for a micro-tomography setup is considered huge; for instance, in reconstructions of 2200×2200 pixel images, with sinograms having 3200 angles and 2500 rays, if a single slice reconstruction takes up to 1.5 s, a total of 3000 slices takes more than an hour. Program execution with a GPU (graphics processing unity) and naïve implementation of $\mathscr{B}g$ is becoming highly attractive for these problems, where the full reconstruction can be achieved within the tolerance of 10 min. If the mathematical model for $\mathscr{B}$ is more sophisticated, e.g. using Fourier transforms, these times can be reduced even more through the usage of available Fast Fourier Transform implementations [5, 9].

A Fourier approach was already established in [1], where the computation of $\mathscr{B}g$ depends on a change from cartesian to log-polar coordinates. This approach, although elegant, suffer from the ill-conditioning of the Log-polar transform at the "fovea". Nevertheless, it is possible to translate the fovea to different regions of the cartesian plane, in order to enclose the reconstruction region. This leads to the concept of *partial-backprojection* which can be easily implemented in a parallel form. See also [10, 11, 13, 14].

Other methods for fast computation of $\mathscr{B}g$ were recently discovered in [2, 3, 6], using an hierarchical approach based on the old computing strategy *divide and conquer*. In this paper, we propose another fast method for $\mathscr{B}g$, also based on Fourier transform. We claim that the backprojection of $g \in V$ can be easily done by filtering the lines of the $\tilde{g}$ one by one, where $\tilde{g}$ is the polar representation of g in $S_+ = \mathbb{R} \times [0, \pi]$.

2 Integral Representations

From now on, we will use the following representation for path integrals, the proof of which can be found in [8]. *For a continuously differentiable* $m\colon \mathbb{R}^m \to \mathbb{R}$ *such that* $\|\nabla m\| \neq \mathbf{0}$:

$$\int_{\mathbb{R}^m} h(\boldsymbol{y})\delta(m(\boldsymbol{y}))\mathrm{d}\boldsymbol{y} = \int_{C^{-1}(0)} \frac{h(\boldsymbol{y}(s))}{\|\nabla m(\boldsymbol{y}(s))\|}\mathrm{d}s(\boldsymbol{x}). \tag{2}$$

where $\mathrm{d}s(\boldsymbol{x})$ is the arclength measure along curve $C^{-1}(0)$. Let $g \in V$ a given sinogram and $\boldsymbol{x}$ a pixel in the reconstruction region. The backprojection (1) of g is defined as the contribution of all possible straight lines, parameterized by the angle θ, and passing through $\boldsymbol{x}$. Using the sifting property of the Delta distribution, we have

$$\mathscr{B}g(\boldsymbol{x}) = \int_0^\pi g(\boldsymbol{x}\cdot\boldsymbol{\xi}_\theta, \theta)\mathrm{d}\theta = \int_0^\pi \int_{\mathbb{R}} g(t,\theta)\delta(t - \boldsymbol{x}\cdot\boldsymbol{\xi}_\theta)\mathrm{d}t\mathrm{d}\theta \tag{3}$$

Switching the above integral from (t, θ) coordinates to cartesian coordinates $\boldsymbol{y} \in \mathbb{R}^2$ we have $|t|\mathrm{d}t\mathrm{d}\theta = \mathrm{d}\boldsymbol{y}$; where $\mathscr{B}g$ now becomes

$$\mathscr{B}g(\boldsymbol{x}) = \int_{\mathbb{R}^2} [g]_{\mathrm{c}}(\boldsymbol{y})\delta(m(\boldsymbol{y}))\frac{1}{\|\boldsymbol{y}\|}\mathrm{d}\boldsymbol{y} \tag{4}$$

with $[g]_c(\boldsymbol{y}) = g(t(\boldsymbol{y}), \theta(\boldsymbol{y}))$ referring to the sinogram g in cartesian coordinates. In fact, $|t| = \|\boldsymbol{y}\|$ is the unsigned distance to the origin and $\theta = \arctan(\frac{y_2}{y_1}) \in [0, \pi]$ is the angle with respect to the $\boldsymbol{y}_1$-axis. Function m reads

$$m(\boldsymbol{y}) = t - \boldsymbol{x} \cdot \boldsymbol{\xi}_\theta = \|\boldsymbol{y}\| - \boldsymbol{x}_1 \cos\theta(\boldsymbol{y}) - \boldsymbol{x}_2 \sin\theta(\boldsymbol{y}) \tag{5}$$

$$= \|\boldsymbol{y}\| - \boldsymbol{x}_1 \frac{\boldsymbol{y}_1}{\|\boldsymbol{y}\|} - \boldsymbol{x}_2 \frac{\boldsymbol{y}_2}{\|\boldsymbol{y}\|} = \|\boldsymbol{y}\| - \frac{(\boldsymbol{x}_1\boldsymbol{y}_1 + \boldsymbol{x}_2\boldsymbol{y}_2)}{\|\boldsymbol{y}\|} = \frac{\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x})}{\|\boldsymbol{y}\|} \tag{6}$$

From (6), (4) and the property $\delta(au) = \frac{1}{|a|}\delta(u)$ for all $u \in \mathbb{R}$, the backprojection now follows:

$$\mathscr{B}g(\boldsymbol{x}) = \int_{\mathbb{R}^2} [g]_c(\boldsymbol{y})\delta(\kappa_{\boldsymbol{x}}(\boldsymbol{y}))\mathrm{d}\boldsymbol{y}, \quad \kappa_{\boldsymbol{x}}(\boldsymbol{y}) = \boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x}) \tag{7}$$

It should be noted that, for a fixed $\boldsymbol{x} \in \mathbb{R}^2$, the set $\kappa_{\boldsymbol{x}}^{-1}(0) = \{\boldsymbol{y} \in \mathbb{R}^2 : \kappa_{\boldsymbol{x}}(\boldsymbol{y}) = 0\}$ is defined as a circle in the plane. Indeed, since $\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x}) = \boldsymbol{y} \cdot \boldsymbol{y} - 2\boldsymbol{y} \cdot \left(\frac{\boldsymbol{x}}{2}\right) = \left\|\boldsymbol{y} - \frac{\boldsymbol{x}}{2}\right\|^2 - \left\|\frac{\boldsymbol{x}}{2}\right\|^2$, it follows that $\kappa_{\boldsymbol{x}}^{-1}(0)$ is a circle passing through the origin $\boldsymbol{y} = \boldsymbol{0}$, centered at $\frac{1}{2}\boldsymbol{x}$ and with radius $\frac{1}{2}\|\boldsymbol{x}\|$, see Fig. 1. Since $\kappa_{\boldsymbol{x}}^{-1}(0) = \{\frac{1}{2}\boldsymbol{x} + r\boldsymbol{\xi}_\theta : \theta \in [0, 2\pi],\ r = \frac{1}{2}\|\boldsymbol{x}\|\}$ is a parametric representation of the circle, the backprojection operator also reads, in an alternative form:

$\mathscr{B}$ is a stacking operator through circles $\kappa_{\boldsymbol{x}}^{-1}(0)$:

$$\mathscr{B}g(\boldsymbol{x}) = \int_{\kappa_{\boldsymbol{x}}^{-1}(0)} \frac{[g]_c(\boldsymbol{y})}{\|2\boldsymbol{y} - \boldsymbol{x}\|}\mathrm{d}s = \frac{1}{2}\int_0^{2\pi} [g]_c\left(\frac{1}{2}\boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x}\|\boldsymbol{\xi}_\theta\right)\mathrm{d}\theta \tag{8}$$

The above representation follows from $ds = \frac{1}{2}\|x\|d\theta$, (7) and (2) with $\nabla\kappa_{\boldsymbol{x}}(\boldsymbol{y}) = 2\boldsymbol{y} - \boldsymbol{x}$. Last equality comes from $\boldsymbol{y} = \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x}\|\xi_\theta \in \kappa_{\boldsymbol{x}}^{-1}(0)$ for some θ. Therefore, in cartesian coordinates, the backprojection contribution for a ball $\{\boldsymbol{z} \in \mathbb{R}^2 : \|\boldsymbol{z} - \boldsymbol{x}\| \leq \epsilon\}$ comes from a family of circles passing through the ball and the origin, see dashed region $\boldsymbol{S}$ in Figs. 1 and 2. This fact is closely related to the "comet-tail region" mentioned by [7].

Andersson's formula: The integral form of the classical backprojection operator, either from (1) or (7), is not suitable for a fast implementation. Andersson's approach [1], using a convolution kernel, is an elegant and analytical alternative for a rapid execution of $\mathscr{B}g$. For completeness, we rederive his formula using four

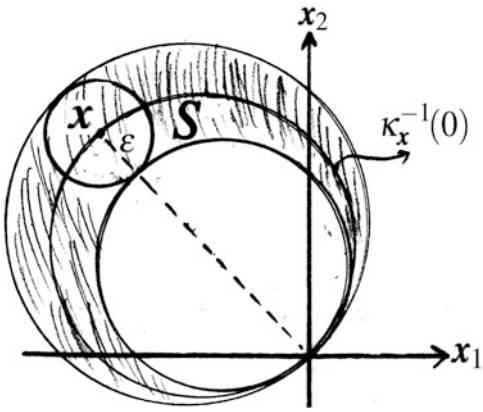


Fig. 1 Only *dashed region* $\boldsymbol{S}$ contributes for the backprojection of the ϵ-ball centered at $\boldsymbol{x}$

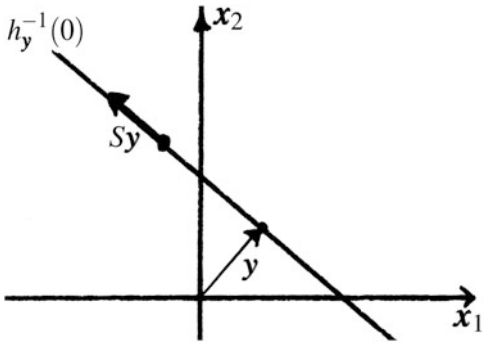


Fig. 2 Support set $h_y^{-1}(0)$ for operator $\mathscr{T}$ in (18). S is a two dimensional $\frac{\pi}{2}$ rotation matrix

steps:

1. Changing the integral (7) from cartesian coordinates $\boldsymbol{y} \in \mathbb{R}^2$ to Prüfer coordinates,[1] i.e., $\boldsymbol{y} \equiv \boldsymbol{y}_{\mu,\theta} = p(\mu)\boldsymbol{\xi}_\phi$ we get $\mathrm{d}\boldsymbol{y} = |p'(\mu)p(\mu)|\mathrm{d}\mu\mathrm{d}\theta$ and

$$\mathscr{B}g(\boldsymbol{x}_{\rho,\theta}) = \int_{S_+} g(\boldsymbol{y}_{\mu,\phi})\delta\left(\kappa_{\boldsymbol{x}_{\rho,\theta}}(\boldsymbol{y}_{\mu,\phi})\right)|p'(\mu)p(\mu)|\mathrm{d}\mu\mathrm{d}\phi \tag{9}$$

2. The support of the Delta distribution in (9) is

$$\kappa_{\boldsymbol{x}_{\rho,\theta}}(\boldsymbol{y}_{\mu,\phi}) = p(\mu)^2\left[1 - \frac{p(\rho)}{p(\mu)}\boldsymbol{\xi}_\phi \cdot \boldsymbol{\xi}_\theta\right] = p(\mu)^2\left[1 - \frac{p(\rho)}{p(\mu)}\cos(\phi - \theta)\right] \tag{10}$$

[1] A generalized polar coordinate system $\{(\mu,\theta);(\mu,\theta) \in S_+\}$, where p is invertible and $p(\mu) = \|\boldsymbol{y}\|$ for cartesian coordinates $\boldsymbol{y} \in \mathbb{R}^2$. Using $p(\mu) = \mu$ we arrive at the polar system and $p(\mu) = e^\mu$ to the log-polar system.

3. Let $[\cdot]_{\mathsf{G}}$ be the representation in Prüfer coordinates. From (10) and (9) we arrive at

$$\begin{aligned}[\mathscr{B}g]_{\mathsf{G}}(\rho,\theta) &= \int_{S_+} [g]_{\mathsf{G}}(\mu,\phi)\delta\left(p(\mu)^2\left[1-\frac{p(\rho)}{p(\mu)}\cos(\phi-\theta)\right]\right)|p'(\mu)p(\mu)|\mathrm{d}\mu\mathrm{d}\phi \\ &= \int_{S_+} [g]_{\mathsf{G}}(\mu,\phi)\delta\left(1-\frac{p(\rho)}{p(\mu)}\cos(\phi-\theta)\right)\frac{|p'(\mu)p(\mu)|}{p(\mu)^2}\mathrm{d}\mu\mathrm{d}\phi\end{aligned} \tag{11}$$

4. A convolution is obtained in (11) only if p is such that $p(\rho) = p(\mu)p(\rho-\mu)$, which in turn implies that p is an exponential function. Hence, Prüfer coordinates reduce to log-polar coordinates; which we denote by $[\cdot]_{\mathsf{L}}$. Finally,

$$[\mathscr{B}g]_{\mathsf{L}}(\rho,\theta) = \int_{S_+} [g]_{\mathsf{L}}(\mu,\phi)\delta\left(1-e^{\rho-\mu}\cos(\phi-\theta)\right)\mathrm{d}\mu\mathrm{d}\phi \tag{12}$$

Andersson's convolution formula for backprojection:

$$[\mathscr{B}g]_{\mathsf{L}}(\rho,\theta) = \left([g]_{\mathsf{L}} \star [\boldsymbol{K}]_{\mathsf{L}}\right)(\rho,\theta), \quad [\boldsymbol{K}]_{\mathsf{L}}(\rho,\theta) = \delta\left(1-e^{\rho}\cos\theta\right) \tag{13}$$

The above formula is particularly good for implementation through the use of Fast fourier transforms. Figure 3 presents a real sinogram, in different coordinate systems.

3 Fourier Analysis

Our analysis starts from the two-dimensional Fourier transform of (7), i.e., $\mathscr{F}\colon \mathscr{B}g \mapsto \widehat{\mathscr{B}g}$:

$$\widehat{\mathscr{B}g}(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} \mathscr{B}[g]_{\mathsf{c}}(\boldsymbol{x})e^{-i\boldsymbol{\omega}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^2}\mathrm{d}\boldsymbol{x}\int_{\mathbb{R}^2}\mathrm{d}\boldsymbol{y}\,[g]_{\mathsf{c}}(\boldsymbol{y})\delta\left(\boldsymbol{y}\cdot(\boldsymbol{y}-\boldsymbol{x})\right)e^{-i\boldsymbol{\omega}\cdot\boldsymbol{x}} \tag{14}$$

$$= \int_{\mathbb{R}^2}\mathrm{d}\boldsymbol{y}[g]_{\mathsf{c}}(\boldsymbol{y})\int_{\mathbb{R}^2}\mathrm{d}\boldsymbol{x}\,\delta\left(\boldsymbol{y}\cdot(\boldsymbol{y}-\boldsymbol{x})\right)e^{-i\boldsymbol{\omega}\cdot\boldsymbol{x}} \tag{15}$$

$$\equiv \int_{\mathbb{R}^2}\mathrm{d}\boldsymbol{y}[g]_{\mathsf{c}}(\boldsymbol{y})\mathscr{T}(\boldsymbol{y},\boldsymbol{\omega}) \tag{16}$$

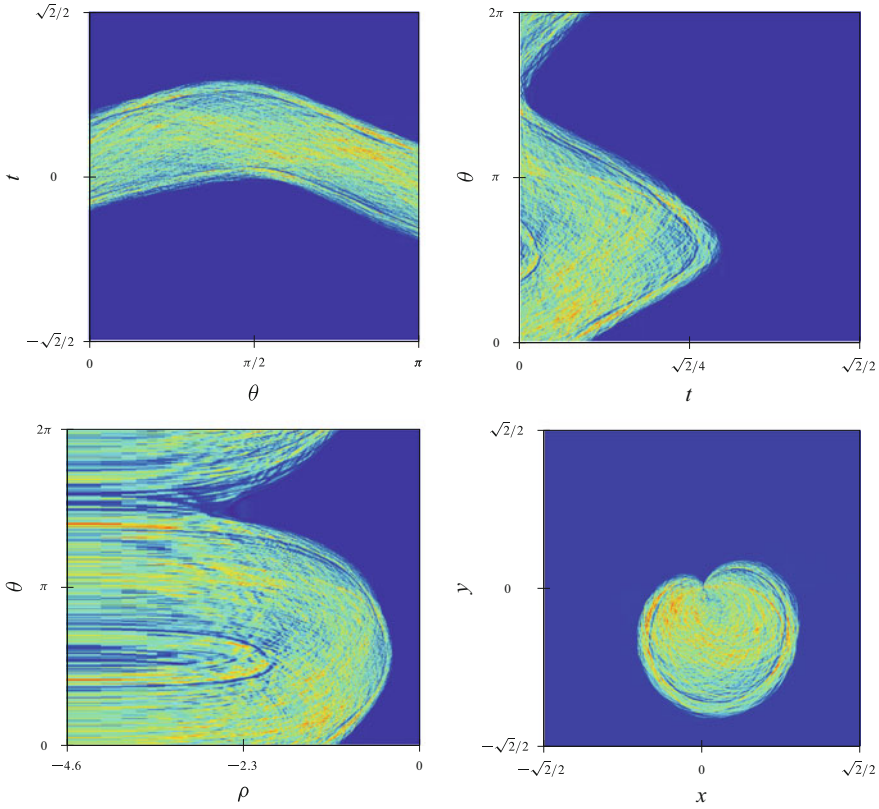


Fig. 3 Synchroton data for a mustard seed. Sinogram in different coordinate systems. *Clockwise from the top*: g in usual sinogram coordinates, $[g]_{\mathsf{p}}$ in polar, $[g]_{\mathsf{L}}$ in log-polar and $[g]_{\mathsf{c}}$ in cartesian coordinates

Keeping $\boldsymbol{y}, \boldsymbol{\omega} \in \mathbb{R}^2$ fixed, we evaluate $\mathscr{T}$ now:

$$\mathscr{T}(\boldsymbol{y}, \boldsymbol{\omega}) = \int_{\mathbb{R}^2} \mathrm{d}\boldsymbol{x}\, \delta\left(h_{\boldsymbol{y}}(\boldsymbol{x})\right) e^{-i\boldsymbol{\omega}\cdot\boldsymbol{x}}, \quad h_{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{y}\cdot(\boldsymbol{y}-\boldsymbol{x}) \tag{17}$$

Since the above distribution is supported in the set $h_{\boldsymbol{y}}^{-1}(0) = \{\boldsymbol{x} \in \mathbb{R}^2 : h_{\boldsymbol{y}}(\boldsymbol{x}) = 0\}$, it follows from the integral representation (2) and $\nabla h_{\boldsymbol{y}} = -\boldsymbol{y}$ that

$$\mathscr{T}(\boldsymbol{y}, \boldsymbol{\omega}) = \frac{1}{\|\boldsymbol{y}\|} \int_{h_{\boldsymbol{y}}^{-1}(0)} e^{-i\boldsymbol{\omega}\cdot\boldsymbol{x}} \mathrm{d}s(\boldsymbol{x}) = \int_{h_{\boldsymbol{y}}^{-1}(0)} e^{-i\boldsymbol{\omega}\cdot\boldsymbol{x}(q)} \mathrm{d}q \tag{18}$$

The set $h_{\boldsymbol{y}}^{-1}(0)$ determines a straight line passing through $\boldsymbol{y}$ and with normal vector $\boldsymbol{y}$, see Fig. 2. Hence, $h_{\boldsymbol{y}}^{-1}(0) = \boldsymbol{y} + \mathrm{span}\{S\boldsymbol{y}\}$, being $S\boldsymbol{y} \perp \boldsymbol{y}$ and S a $\frac{\pi}{2}$-rotation matrix. Therefore, $\boldsymbol{x}(q) \in h_{\boldsymbol{x}}^{-1}(0)$ is on the form $\boldsymbol{x}(q) = \boldsymbol{y} + qS\boldsymbol{y}$ and the integral in (18)

yields

$$\mathscr{T}(\boldsymbol{y},\boldsymbol{\omega}) = \int_{\mathbb{R}} e^{-i\boldsymbol{\omega}\cdot[\boldsymbol{y}+qS\boldsymbol{y}]}\mathrm{d}q = e^{-i\boldsymbol{\omega}\cdot\boldsymbol{y}}\int_{\mathbb{R}} e^{-iq\boldsymbol{\omega}\cdot(S\boldsymbol{y})}\mathrm{d}q = e^{-i\boldsymbol{\omega}\cdot\boldsymbol{y}}\,\delta\,(\boldsymbol{\omega}\cdot S\boldsymbol{y}) \tag{19}$$

At this point, the Fourier transform of $\mathscr{B}g$ becomes

$$\widehat{\mathscr{B}g}(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} [g]_{\mathsf{c}}(\boldsymbol{y})\delta\,(\boldsymbol{\omega}\cdot S\boldsymbol{y})\,e^{-i\boldsymbol{\omega}\cdot\boldsymbol{y}}\,\mathrm{d}\boldsymbol{y} \tag{20}$$

For $\boldsymbol{\omega}$ fixed, $\{\boldsymbol{y}\in\mathbb{R}^2 : \boldsymbol{\omega}\cdot(S\boldsymbol{y}) = 0\} = \mathrm{span}\{\boldsymbol{\omega}\}$. Indeed, since $S\boldsymbol{y}\perp\boldsymbol{w}$ and $S\boldsymbol{y}\perp\boldsymbol{y}$, it follows $\boldsymbol{\omega}\parallel\boldsymbol{y}$. Once again, using the representation (2) for (20) we arrive at

$$\widehat{\mathscr{B}g}(\boldsymbol{\omega}) = \int_{\mathbb{R}} \frac{[g]_{\mathsf{c}}(q\boldsymbol{\omega})}{\|S\boldsymbol{\omega}\|} e^{-i\boldsymbol{\omega}\cdot(q\boldsymbol{\omega})}\,\mathrm{d}s(\boldsymbol{\omega}) \tag{21}$$

Since $\|S\boldsymbol{\omega}\| = \|\boldsymbol{\omega}\|$ and $\mathrm{d}s(\boldsymbol{\omega}) = \|\boldsymbol{\omega}\|dq$, we finally obtain

$$\widehat{\mathscr{B}g}(\boldsymbol{\omega}) = \int_{\mathbb{R}} [g]_{\mathsf{c}}(q\boldsymbol{\omega})e^{-iq\|\boldsymbol{\omega}\|^2}\,\mathrm{d}q \tag{22}$$

We claim that the backprojection is a polar convolution. Indeed, switching the frequency domain to polar coordinates, i.e., $\boldsymbol{\omega} = \sigma\boldsymbol{\xi}_\theta$ (with $\sigma\in\mathbb{R}_+$ and $\theta\in[0,2\pi]$) we get

$$\widehat{\mathscr{B}g}(\sigma\boldsymbol{\xi}_\theta) = \int_{\mathbb{R}} [g]_{\mathsf{c}}(q\sigma\boldsymbol{\xi}_\theta)e^{-iq\|\sigma\boldsymbol{\xi}_\theta\|^2}\mathrm{d}q = \int_{\mathbb{R}} \frac{[g]_{\mathsf{c}}(u\boldsymbol{\xi}_\theta)}{\sigma}e^{-iu\sigma}\mathrm{d}u. \tag{23}$$

Now, letting $[\cdot]_{\mathsf{s}}$ be the representation in semi-polar coordinates, it is true that $[g]_{\mathsf{c}}(u\boldsymbol{\xi}_\theta) = [g]_{\mathsf{s}}(u,\theta)$ is the input sinogram $g(u,\theta)$. From (22) and (23), using semi-polar coordinates

$$[\widehat{\mathscr{B}g}]_{\mathsf{p}}(\sigma,\theta) = \widehat{\mathscr{B}g}(\sigma\boldsymbol{\xi}_\theta) = \frac{1}{\sigma}\int_{\mathbb{R}} g(u,\theta)e^{-iu\sigma}\mathrm{d}u \tag{24}$$

Identity (24) is our backprojection-slice theorem for computing the operator $\mathscr{B}$. Indeed, at each radial line θ in the frequency domain, the two-dimensional Fourier transform of $\mathscr{B}$ equals the one-dimensional radial Fourier transform of the projection $g(t,\theta)$ multiplied by the kernel $1/\sigma$.

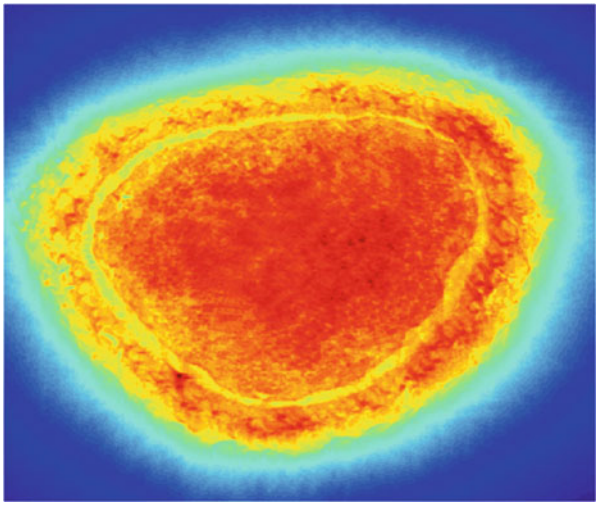


Fig. 4 Full backprojection with the sinogram presented in Fig. 3

Backprojection Slice-Theorem:

$$g \in V \quad \Rightarrow \quad \widehat{\mathscr{B}g}(\sigma \boldsymbol{\xi}_\theta) = \frac{\hat{g}(\sigma, \theta)}{\sigma}, \quad \sigma \in \mathbb{R}_+ \;\; \theta \in [0, 2\pi] \tag{25}$$

It is evident that a signal function and a derivative operator is present in (25), i.e., the backprojection $\mathscr{B}g$ is closely related to the Hilbert transform. A full backprojection from synchrotron data is depicted in Fig. 4.

Acknowledgements This work was supported by FAPESP 2013/16508-3.

References

1. Andersson, F.: Fast inversion of the radon transform using log-polar coordinates and partial back-projections. SIAM J. Appl. Math. **65**(3), 818–837 (2005)
2. Basu, S., Bresler, Y.: $O(N^2 \log_2 N)$ filtered back projection reconstruction algorithm for tomography. IEEE Trans. Med. Imag. **9**(10), 1760–1763 (2000)
3. Brandt, A., Mann, J., Brodski, M., Galun, M.: A fast and accurate multilevel inversion of the radon transform. SIAM J. Appl. Math. **60**(2), 437–462 (1999)
4. Deans, S.: The Radon Transform and Some of Its Applications. Dover, New York (1983)
5. Frigo, M., Johnson, S.G.: The design and implementation of FFTW3. Proc. IEEE **93**(2), 216–231 (2005). Invited paper, Special Issue on Program Generation, Optimization, and Platform Adaptation

6. George, A., Bresler, Y.: Fast tomographic reconstruction via rotation based hierarchical back projection. SIAM J. Appl. Math. **68**(2), 574–597 (2007)
7. Hass, R., Faridani, A.: Regions of backprojection and comet tail artifacts for pi-line reconstruction formulas in tomography. SIAM J. Imag. Sci. **5**(4), 1159–1184 (2012)
8. Helgason, S.: Radon Transform, 2nd edn. Cambridge University Press, Cambridge (1999)
9. Marone, F., Stampanoni, M.: Regridding reconstruction algorithm for real time tomographic imaging. Synchrotron Radiat. **19**, 1029–1037 (2012)
10. Metz, C.E., Pan, X.: A unified analysis of exact methods of inverting the 2-D exponential radon transform, with implications for noise control in SPECT. IEEE Trans. Med. Imag. **14**(4), 643–658 (1995)
11. Miqueles, E.X., Helou, E.S., De Pierro, A.R.: Generalized backprojection operator: fast calculation. J. Phys. Conf. Ser. **490**, 012148 (2014)
12. Natterer, F., Wubbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
13. Tretiak, O., Metz, C.E.: The exponential radon transform. SIAM J. Appl. Math. **39**, 341–354 (1980)
14. Van Loan, C.F.: The ubiquitous Kronocker product. J. Comput. Appl. Math. **123**, 85–100 (2000)
15. Wei, Y., Wang, G., Hsieh, J.: Relation between the filtered backprojection algorithm and the backprojection algorithm in CT. IEEE Signal Process. Lett. **12**(9), 633–636 (1995)

MS 9
MINISYMPOSIUM: INDUSTRIAL PARTICLE AND INTERFACE DYNAMICS

Organizers

Tuoi T. N Vo[1]

Speakers

William Lee[2] and Ellen Murphy[3]
Bubble Dynamics in Stout Beers

Michael Vynnycky[4], Sarah Mitchell[5], Brendan Florio[6] and Stephen O'Brie[7]
Decoupling the Interaction of Solid and Fluid Mechanics in the Modelling of Continuous Casting Processes

[1]Tuoi Vo, MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland.

[2]William T. Lee, MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland.

[3]Ellen Murphy, Bath Institute for Mathematical Innovation, University of Bath, Bath, UK.

[4]Michael Vynnycky, Department of Materials Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden.

[5]Sarah Mitchell, MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland.

[6]Brendan Florio, MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland.

[7]Stephen B. G. O'Brien, MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_33

Kevin M. Moroney[8], William T. Lee[2], Stephen B. G. O'Brien[7], Freek Suijver[9] and Johan Marra[10]
Mathematical Modelling of the Coffee Brewing Process

Tuoi T. N. Vo[1], William Lee[2], Simon Kaar[11], Jan Hazenberg[12], and James Power[13]
Modelling Particle-Wall Interaction in Dry Powder Inhalers

José Ramón Fernández García[14], Piotr Kalita[15] Stanislaw Migórski[16], María Del Carmen Muñiz Castiñeira[17] and Cristina Núñez García[18]
Numerical Analysis and Simulations of a Mixed Kinetic-Diffusion Model for Surfactant Solutions

Dana Mackey[19], Paul OReilly[20] and Izabela Naydenova[21]
Optimising Copying Accuracy in Holographic Patterning

Brendan Florio[6]
The Generation and Interaction of Convection Modes in a Box of a Saturated Porous Medium

Keywords

Interface dynamics
Particle dynamics
Particulate system

[8]Kevin M. Moroney, MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland.

[9]Freek Suijver, Philips Research, Eindhoven, The Netherlands.

[10]Johan Marra, Philips Research, Eindhoven, The Netherlands.

[11]Simon Kaar, MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland.

[12]Jan Hazenberg, Teva Pharmaceuticals, Unit 301, Waterford Industrial Estate, Waterford, Ireland.

[13]James Power, Teva Pharmaceuticals, Unit 301, Waterford Industrial Estate, Waterford, Ireland.

[14]José Ramón Fernández García, Universidade de Vigo, Vigo, Spain.

[15]Piotr Kalita, Jagiellonian University, Krakow, Poland.

[16]Stanislaw Migórski, Jagiellonian University, Krakow, Poland.

[17]María Del Carmen Muñiz Castiñeira, Universidade de Santiago de Compostella, Santiago de Compostella, Spain.

[18]Cristina Núñez García, Universidade de Santiago de Compostella, Santiago de Compostella, Spain.

[19]Dana Mackey, Dublin Institute of Technology, Dublin, Ireland.

[20]Paul OReilly, Dublin Institute of Technology, Dublin, Ireland.

[21]Izabela Naydenova, Dublin Institute of Technology, Dublin, Ireland.

Short Description

Particles and interfaces are ubiquitous in industrial processes. This mini-symposium will focus on mathematical modelling of processes occurring at interfaces and in particulate systems which arise in industry (for example, food and drink, pharmaceutical, and other sectors). These challenging industrial problems are currently under investigation at the Mathematics Application Consortium for Science and Industry (MACSI) in Ireland. The talks will cover the development of innovative mathematical models to help industry optimise and improve processes, increase scientific understanding, and meet regulatory requirements. Models are developed for chemical extraction from powdered substrates, and particle motion and adhesion in circulating flows. Asymptotic, analytical and numerical techniques are used to investigate the mathematical models.

Bubble Dynamics in Stout Beers

W.T. Lee and E. Murphy

Abstract Technology for promoting nucleation is important in a number of contexts, for instance degassing carbon dioxide lakes, designing champagne glasses and stout beer widgets. A new design of stout beer widget has recently been proposed which makes use cellulose fibres to initiate foaming in canned stout beers. However, our current scientific understanding of the nucleation of bubbles by cellulose fibres is incomplete, making it impossible to optimise this technology. One particularly poorly understood aspect is the detachment of bubbles from a gas pocket in the fibre. We report experimental and theoretical results towards a model of the detachment based on a model of Rayleigh-Plateau instability including a disjoining pressure.

Keywords Disjoining pressure • Nucleation • Stout beer

1 Introduction

Technology for promoting bubble nucleation is important for numerous applications. These include:

- Siphons for removing carbon dioxide from lakes near volcanoes, such as Lake Kivu in Rwanda [1]. Uncontrolled degassing of these lakes can release a large volume of carbon dioxide which can asphyxiate nearby livestock and communities.
- Engraved champagne glasses. By adding artificial nucleation sites to champagne glasses by engraving the rate of effervescence of the champagne can be controlled, leading to a more appealing appearance and bouquet [2].
- Widgets for stout beers. Unlike most beers which foam spontaneously, stout beers require special technology to promote foaming. In canned stout beer this technology takes the form of a widget [3].

W.T. Lee (✉) • E. Murphy
MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
e-mail: william.lee@ul.ie

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_34

The reason for a difference in foaming properties in stout beers is ultimately traceable to the dissolved gases in the beer. A typical carbonated beer foams due to the presence of dissolved carbon dioxide. However most stout beers contain both dissolved carbon dioxide and nitrogen—the latter being a much less soluble gas than carbon dioxide. This gas mixture imparts a number of desirable properties to stout beers. It affects the taste since carbon dioxide is acidic in solution, replacing it with inert nitrogen gives stout beers a smoother, less acidic taste. The low solubility of nitrogen also means that bubbles in stout beers are smaller, this leads to the famous sinking bubble of Guinness phenomenon, but in addition is responsible for the long lasting creamy head. However these desirable properties come at a price. The low solubility of nitrogen also makes it difficult to initiate foaming in stout beers. While carbonated beers foam (apparently) spontaneously, stout beers will not foam without an external trigger.

In canned stout beers this trigger takes the form of a widget, a small plastic ball containing compressed gasses from the headspace of the can. When the can is opened the headspace depressurises, leading to a pressure imbalance between the gas in the widget and the headspace. This pressure difference expels the gas in the widget through a small opening into the beer. The turbulent jet of gas released breaks up into millions of bubble nuclei which grow by absorbing dissolved gasses in the beer, rise to the surface and form the head.

Widgets are not required in carbonated beers, which appear to foam spontaneously. This foaming is not however truly spontaneous. Careful investigation of the nucleation sites at glass surfaces shows that foaming in carbonated beer is due to the presence of cellulose fibres. These fibres contain trapped gas pockets which release bubble nuclei through a cyclic process first elucidated by Liger-Belair et al. Experimental and theoretical investigation of this bubble production mechanism show that it does occur in stout beers, but at a rate too slow to be observable. Rough estimates do suggest that a large concentration of cellulose fibres could potentially act as an alternative widget. This new widget design could potentially have a number of advantages over the current generation of widget technology, for instance being more environmentally friendly than the plastic widgets.

This new design of widget must be able to generate nuclei of the 10^8 bubbles needed to form the head of a pint of stout in the 3 s it takes to pour the can into a glass. While rough estimates suggest this is possible, there are a number of significant gaps in our quantitative understanding of the generation of bubble nuclei by cellulose fibres which make it difficult to design such a widget.

In this paper we report progress towards a quantitative understanding of the detachment of bubble nuclei from the gas pocket in a cellulose fibre. The remainder of the paper is arranged as follows. In Sect. 2 we describe the acquisition of the experimental data used to inform and validate our model. Section 3 describes a key ingredient of the model, namely the disjoining pressure. A model of the detachment process is presented in Sect. 4. Our conclusions are presented in Sect. 5.

2 Experiments

The first observations on bubble growth by gas pockets in cellulose fibres were carried out Liger-Belair [4]. However, data from stout beer bubbles can be captured using a much simpler setup. The much more gentle pace of the nucleation of bubbles in stout beers makes them an ideal system in which to observe the cellulose fibre nucleation mechanism. The slow rate of formation and small size of bubbles means that agitation of the fluid—which would affect the quality of the images—is small. Furthermore bubble formation can be observed from just above the surface of the fluid. This would be impossible in champagne: the bursting of large champagne bubble would rapidly coat the objective lens with droplets.

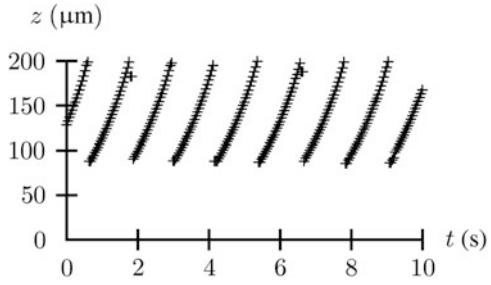


Fig. 1 Experimentally measured gas pocket size [5]

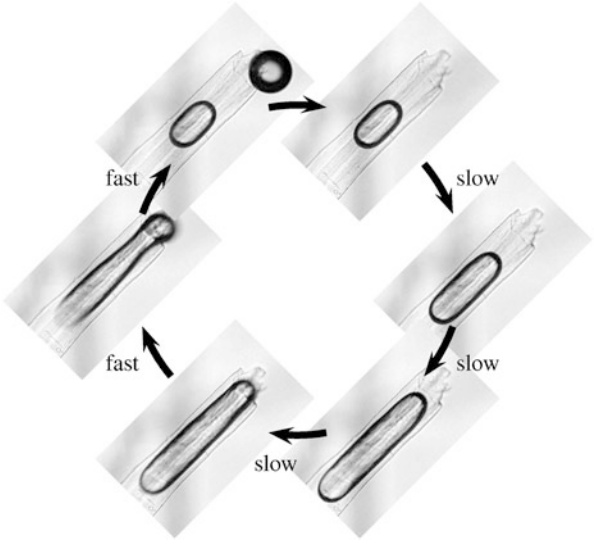


Fig. 2 Cyclic process by which a gas pocket in a cellulose fibre creates bubbles [5]

The results of these observations are shown in Figs. 1 and 2. As was the case with champagne experiments we have not been able to systematically observe the detachment of bubbles. However in one case we were able to observe a bubble apparently in the act of breaking off, as shown in Fig. 2. This image is suggestive of a Plateau-Rayleigh mechanism for bubble detachment, a possibility we investigate in more detail below.

The detachment time would be an important parameter to measure since it could be used to check our models. As has been noted above, it cannot be measured directly. It can however be estimated via a statistical argument. The dataset presented in Fig. 1 contains the detachment of 20 bubbles (the full dataset is not shown in the graph). In only a single case is the bubble caught in the act of detaching. A first estimate of the time taken for a bubble to detach would be to equate the fraction of times a bubble is observed detaching with the ratio of the detachment time to the interval between frames. Since the time between frames is 40 ms this suggests a detachment time of approximately 2 ms.

3 Disjoining Pressure

To model the detachment process we need to assemble the ingredients of a model. The fluid flow can be modelled using the Navier Stokes equations, while the interface conditions can be modelled using the Young-Laplace law. It is clear from Fig. 2 that such a model will be incomplete. At the tip of the gas pocket the interface is approximately spherical, thus by the Young Laplace law the pressure in the gas pocket must be higher than the fluid pressure by approximately $2\gamma/r$ where γ is the surface tension and r is the radius. Applying the Young-Laplace law to the pressure difference in the shaft of the tube gives a pressure difference of γ/r. Thus there is a pressure difference of γ/r as yet unaccounted for. The small growth rate of the gas pocket, We $= 10^{-15}$ means it is impossible to account for this difference by dynamic effects.

The most plausible explanation of this discrepancy is the existence of a disjoining pressure [6] in the thin film of beer between the cellulose fibre and the shaft of the bubble. In theory the disjoining pressure can be reconstructed from its effect on the shape of the gas pocket. By extracting the coordinates of the interface and using finite difference approximations to the derivatives we can roughly extract the form of the disjoining pressure as shown in Fig. 3. Although this reconstruction fails for small and large values of the film thickness there is a significant range of viable data. This is used to estimate an exponential form of the disjoining pressure.

To test the accuracy of this disjoining pressure it is used to reconstruct the original bubble. A comparison of the actual bubble profile and the reconstructed bubble is shown in Fig. 4.

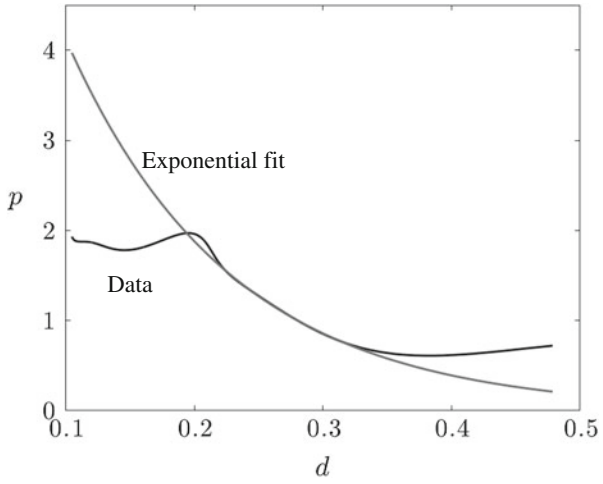


Fig. 3 Exponential fit to disjoining pressure

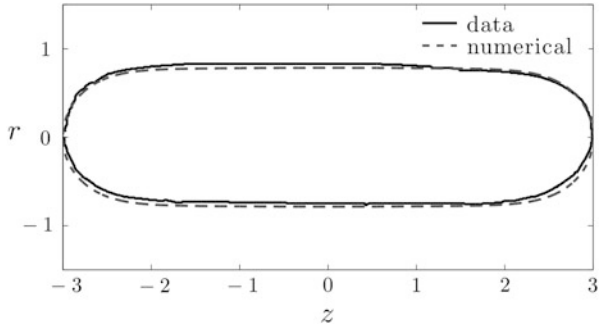


Fig. 4 Comparison of experimental gas pocket shape with theoretical reconstruction based on an exponential disjoining pressure. r and z are scaled by the radius of the fibre

4 Detachment

With the ingredients of the model assembled, we can now proceed with the main aim of this paper: to determine if the detachment of bubble nuclei from the gas pocket can be modelled using a Plateau-Rayleigh type instability. We make the following simplifications

– For simplicity we model only the gas within the gas pocket, which we treat as an ideal fluid.
– We simplify the geometry to make the system periodic.

Neither of these assumptions is strictly justified by the parameters of the system. However these are appropriate for an initial investigation.

The geometry used in this investigation is shown in Fig. 5. A gas tube is contained with a periodic array of cellulose fibres. The stability of this system is investigated numerically with Fig. 6 shows the timescales corresponding to the most unstable eigenvalues as a function of the gap between fibres. One of the eigenvectors is shown superimposed on the detaching bubble in Fig. 7. The timescales shown in Fig. 6 are comparable with our estimate of the detachment time from experiments (2 ms). This suggests that the Plateau-Rayleigh instability is a viable explanation of the detachment process and more detailed modelling along these lines is appropriate.

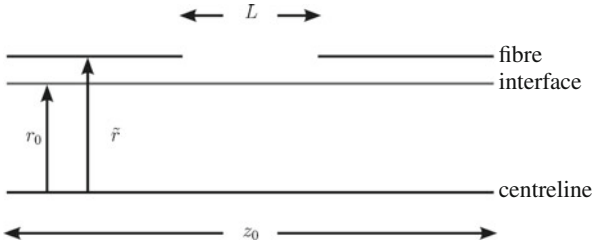


Fig. 5 Geometry used to investigate bubble detachment times. The geometry consists of a periodic array of fibres with period z_0 with gaps between the fibres of size L

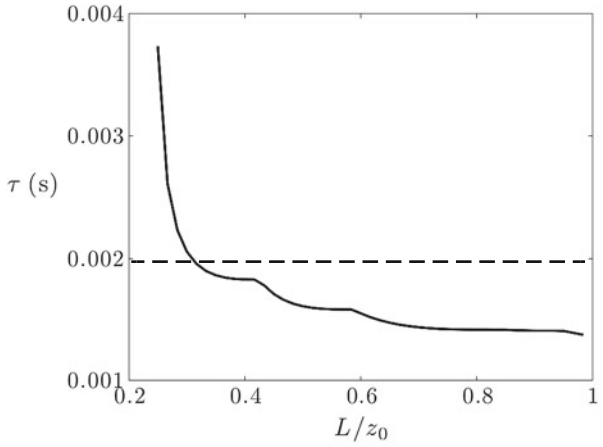


Fig. 6 Timescale corresponding to most unstable eigenvalue of the system as a function of fibre separation. *Dashed line* gives the estimate of the bubble detachment time from experiment

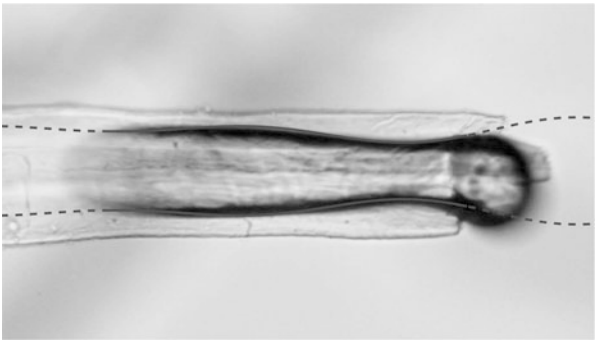


Fig. 7 Comparison with eigenvector and experiment

5 Conclusions

Stout beer widgets made from cellulose fibres offer an appealing alternative to current widget designs. However to optimise the design of such widgets a complete mathematical model of the generation of bubbles by cellulose fibres is needed. Here we have focused on one aspect of bubble generation: the detachment of a bubble nucleus from the gas pocked in the cellulose fibre. A simplified model based on the introduction of a disjoining pressure and the Rayleigh Plateau instability is consistent with the experimental results.

Acknowledgements WTL acknowledges the support of the Mathematics Applications Consortium for Science and Industry (http://www.macsi.ul.ie) funded by the Science Foundation Ireland Mathematics Investigator Award 12/IA/1683.

References

1. Nayar, A.: Nature **460**, 321 (2009)
2. Polidori, G., Beaumont, F., Jeandet, P., Liger-Belair, G.: J. Vis. **11**(4), 279 (2008). doi:10.1007/BF03182193. http://dx.doi.org/10.1007/BF03182193
3. Lee, W.T., McKechnie, J.S., Devereux, M.G.: Phys. Rev. E **83**(5, 1), 051609 (2011). doi:10.1103/PhysRevE.83.051609
4. Liger-Belair, G., Vignes-Adler, M., Voisin, C., Robillard, B., Jeandet, P.: Langmuir **18**, 1294 (2002)
5. Lee, W.T., Devereux, M.G.: Am. J. Phys. **79**(10), 991 (2011). doi:10.1119/1.3620416
6. Derjanguin, B.V., Churaev, N.V.: J. Colloid Interface Sci. **49**, 249 (1974)

Decoupling the Interaction of Solid and Fluid Mechanics in the Modelling of Continuous Casting Processes

M. Vynnycky, S.L. Mitchell, B.J. Florio, and S.B.G. O'Brien

Abstract The modelling of the continuous casting of metals is known to involve the complex interaction of non-isothermal fluid and solid mechanics. However, using asymptotic methods and an earlier numerical result obtained via computational fluid dynamics, we demonstrate how the motion of the liquid metal can be systematically decoupled from the stresses induced in the solidified shell. The resulting asymptotically reduced model can then serve as a computationally efficient module for stress mechanics models that aim to predict segregation and crack formation in the solid metal.

Keywords Computational fluid dynamics • Metal casting • Solid-fluid-interactions

1 Introduction

Continuous casting has been developed industrially worldwide since the 1950s as a high throughput method for producing, amongst other things, metal billets, blooms and slabs. In a continuous casting process, such as the strip casting of copper, jets of molten metal enter into the top of a water-cooled mould, where intense cooling causes a solidified metal shell to form; subsequently, the metal is withdrawn at a uniform casting speed. The industrial importance of the process has led to interest amongst mathematicians and engineers, with a view to obtaining an improved understanding of the factors that influence product quality and process productivity. Central to these is the coupled heat and momentum transfer that occurs

M. Vynnycky
Division of Casting of Metals, Department of Materials Science and Engineering, Royal Institute of Technology, Brinellvägen 23, 100 44 Stockholm, Sweden
e-mail: michaelv@kth.se

S.L. Mitchell (✉) • B.J. Florio • S.B.G. O'Brien
Mathematics Applications Consortium for Science and Industry (MACSI), Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
e-mail: sarah.mitchell@ul.ie; brendan.florio@ul.ie; stephen.obrien@ul.ie

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_35

during solidification, particularly as regards determining the location of the interface between molten and solidified metal and solid.

Typically, the flow in the molten metal is turbulent, and it is generally believed that a computational fluid dynamics (CFD) approach, based around the Reynolds-averaged Navier Stokes equations, is necessary in order to correctly capture the heat transfer characteristics [1–8, 13–15]. In this contribution, we consider an asymptotically reduced version of the CFD-based model for the continuous casting of copper presented in [7, 8] and demonstrate that, even by neglecting the details of the turbulence in the molten pool, the reduced model gives predictions for the pool depth, local temperature profiles and mould wall heat flux that agree very well with the results of the original CFD model.

2 Model Formulation

As in [7, 8], we consider a steady state 2D problem, as shown in Fig. 1, in which pure liquid metal at temperature T_{cast}, which is greater than the melting temperature, T_{melt}, enters a mould region via a jet at $z = 0$, first solidifies at the inner mould

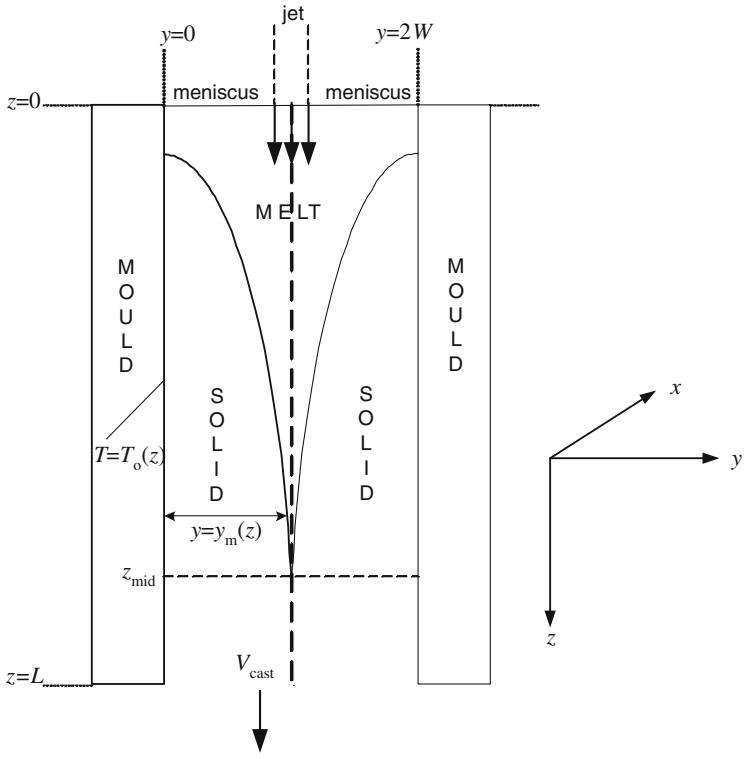


Fig. 1 Schematic of vertical continuous casting

surface at $z = z_{\text{melt}}$ and is withdrawn at a casting speed V_{cast}. For $z_{\text{melt}} < z < L$, solidification occurs in the region $0 < y < y_{\text{m}}(z)$. Eventually, after complete solidification has occurred at $z = z_{\text{mid}}$, the solid region occupies $0 < y < W$. Due to symmetry, we need only consider the left-hand side of Fig. 1.

Typically, the width of the jet is much less than width of the mould region, $2W$; mass conservation therefore suggests that the velocities in the upper part of the molten metal region should be much greater than the casting speed. Nevertheless, the fact that the molten metal velocity must eventually reduce to V_{cast} in the vicinity of the, as yet unknown, solid-liquid interface, suggests that its location might be adequately determined by approximating the liquid velocity by V_{cast}. Therefore, we proceed on this basis.

For $0 < z < z_{\text{melt}}$ and $0 < y < W$, and then $z > z_{\text{melt}}$ and $y_{\text{m}}(z) < y < W$, we have

$$\rho_{\text{l}} c_{\text{pl}} V_{\text{cast}} \frac{\partial T_{\text{l}}}{\partial z} = k_{\text{l}} \frac{\partial^2 T_{\text{l}}}{\partial y^2}, \tag{1}$$

where k_{l} is the thermal conductivity of the liquid metal, c_{pl} is its specific heat capacity and ρ_{l} its density. In Eq. (1), we use the fact that casting geometries are often slender, which motivates us to assume that $\partial^2/\partial z^2 \ll \partial^2/\partial y^2$. For $z_{\text{melt}} < z < L$ and $0 < y < y_{\text{m}}(z)$, we have

$$\rho_{\text{s}} c_{\text{ps}} V_{\text{cast}} \frac{\partial T_{\text{s}}}{\partial z} = k_{\text{s}} \frac{\partial^2 T_{\text{s}}}{\partial y^2}, \tag{2}$$

where k_{s} is the thermal conductivity of the solid metal, c_{ps} is its specific heat capacity and ρ_{s} its density; for simplicity, we will henceforth take $\rho_{\text{s}} = \rho_{\text{l}} = \rho$. Typically, however, $\rho_{\text{s}} > \rho_{\text{l}}$, but the underlying reason for taking $\rho_{\text{s}} = \rho_{\text{l}}$ is to be able to assume to a common streaming velocity, V_{cast}, for both phases; otherwise, we would be required to solve momentum transfer equations in the liquid phase.

For boundary conditions at $y = y_{\text{m}}(z)$, we have

$$T_{\text{s}} = T_{\text{l}} = T_{\text{melt}}, \tag{3}$$

and the Stefan condition,

$$k_{\text{s}} \frac{\partial T_{\text{s}}}{\partial y} - k_{\text{l}} \frac{\partial T_{\text{l}}}{\partial y} = \rho \Delta H_{\text{f}} V_{\text{cast}} \frac{\mathrm{d}y_{\text{m}}}{\mathrm{d}z}, \tag{4}$$

where ΔH_{f} is the latent heat of fusion; this form for (4) also makes use of the fact that the geometry is slender. However, once solidification is complete, at $z = z_{\text{mid}}$, we treat $y = W$ as a symmetry axis, so that (3) and (4) are replaced by

$$\frac{\partial T_{\text{s}}}{\partial y} = 0. \tag{5}$$

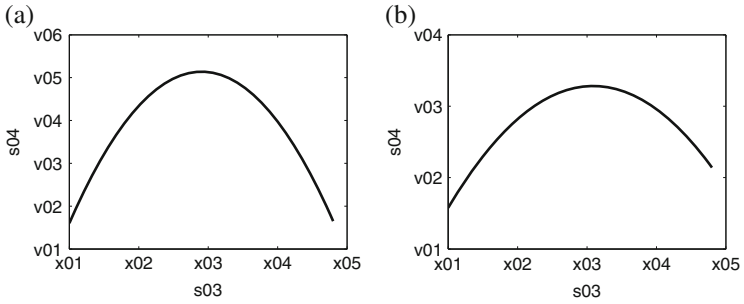


Fig. 2 (**a**) Heat transfer coefficient, h, used in the computations; (**b**) mould temperature, T_{o}, used in the computations

For the mould wall at $y = 0$,

$$\begin{cases} k_{\mathrm{l}} \frac{\partial T_{\mathrm{l}}}{\partial y} = h\,(z)\,(T_{\mathrm{l}} - T_{\mathrm{o}}(z))\,, \text{ for } 0 \le z \le z_{\mathrm{melt}}, \\ k_{\mathrm{s}} \frac{\partial T_{\mathrm{s}}}{\partial y} = h\,(z)\,(T_{\mathrm{s}} - T_{\mathrm{o}}(z))\,, \text{ for } z_{\mathrm{melt}} \le z \le L. \end{cases} \tag{6}$$

In (6), $T_{\mathrm{o}}(z)$ is an experimentally measured temperature, $h\,(z)$ is an experimentally determined heat transfer coefficient; we take the profiles shown in Fig. 2, which are very similar to those used in an earlier full CFD model [7, 8].

Since (1) and (2) are both parabolic partial differential equations, initial conditions for T_{l} and T_{s} are necessary at $z = 0$ and $z = z_{\mathrm{melt}}$, respectively. For T_{l}, we take

$$T_{\mathrm{l}} = T_{\mathrm{cast}} \quad \text{at } z = 0, \tag{7}$$

whereas for T_{s}, we take

$$T_{\mathrm{s}} = T_{\mathrm{melt}} \quad \text{at } y = 0,\ \ z = z_{\mathrm{melt}}; \tag{8}$$

in addition, we must have

$$y_{\mathrm{m}}\,(z_{\mathrm{melt}}) = 0. \tag{9}$$

3 Nondimensionalization

We write

$$Y = \frac{y}{W}, \quad Y_{\mathrm{m}} = \frac{y_{\mathrm{m}}}{W}, \quad Z = \frac{z}{L}, \quad H = \frac{h}{[h]},$$

$$\theta_{\mathrm{l}} = \frac{T_{\mathrm{melt}} - T_{\mathrm{l}}}{\Delta T}, \quad \theta_{\mathrm{s}} = \frac{T_{\mathrm{melt}} - T_{\mathrm{s}}}{\Delta T}, \quad \theta_{\mathrm{o}} = \frac{T_{\mathrm{melt}} - T_{\mathrm{o}}(z)}{\Delta T},$$

where ΔT is a temperature scale and $[h]$ is a heat transfer coefficient scale to be specified; these can be chosen from $T_o(z)$ and $h(z)$, respectively, as

$$\Delta T = T_{\text{melt}} - \min\left(T_o(z) \middle| z \geq 0\right), \quad [h] = \max\left(h(z) \middle| z \geq 0\right). \tag{10}$$

Equations (1) and (2) then become, respectively,

$$\widetilde{Pe}_l \frac{\partial \theta_l}{\partial Z} = \frac{\partial^2 \theta_l}{\partial Y^2}, \quad \widetilde{Pe}_s \frac{\partial \theta_s}{\partial Z} = \frac{\partial^2 \theta_s}{\partial Y^2}, \tag{11}$$

where $\widetilde{Pe}_l$ and $\widetilde{Pe}_s$ are reduced Péclet numbers, given by

$$\widetilde{Pe}_l = \frac{\rho c_{pl} V_{\text{cast}} L}{k_l} \left(\frac{W}{L}\right)^2, \quad \widetilde{Pe}_s = \frac{\rho c_{ps} V_{\text{cast}} L}{k_s} \left(\frac{W}{L}\right)^2. \tag{12}$$

The boundary conditions for θ_l and θ_s are then: at $Y = Y_m(Z)$,

$$\theta_s = \theta_l = 0, \quad \frac{\partial \theta_s}{\partial Y} - K \frac{\partial \theta_l}{\partial Y} = -\frac{\widetilde{Pe}_s}{St} \frac{\mathrm{d}Y_m}{\mathrm{d}Z}, \tag{13}$$

where $K = k_l/k_s$. At $Y = 0$,

$$\frac{\partial \theta_l}{\partial Y} = Bi\, H(Z)\big(\theta_l - \theta_o(Z)\big) \quad \text{for } 0 \leq Z < Z_{\text{melt}}, \tag{14}$$

$$\frac{\partial \theta_s}{\partial Y} = Bi\, K\, H(Z)\big(\theta_s - \theta_o(Z)\big) \quad \text{for } Z_{\text{melt}} \leq Z < 1, \tag{15}$$

where $Z_{\text{melt}} = z_{\text{melt}}/L$; at $Y = 1$,

$$\frac{\partial \theta_l}{\partial Y} = 0 \quad \text{at } Y = 1 \text{ for } 0 \leq Z \leq Z_{\text{mid}}, \tag{16}$$

$$\frac{\partial \theta_s}{\partial Y} = 0 \quad \text{at } Y = 1 \text{ for } Z_{\text{mid}} \leq Z \leq 1. \tag{17}$$

where $Z_{\text{mid}} = z_{\text{mid}}/L$. In addition, St and Bi are the Stefan and Biot numbers, respectively, and are given by

$$St = \frac{c_{ps} \Delta T}{\Delta H_f}, \quad Bi = \frac{[h]\, W}{k_l}. \tag{18}$$

The initial conditions (7)–(9) are, respectively,

$$\theta_{\mathrm{l}} = \theta_{\mathrm{cast}} \quad \text{at } Z = 0, \tag{19}$$

$$\theta_{\mathrm{s}} = 0 \quad \text{at } Z = Z_{\mathrm{melt}}, \quad Y = 0 \tag{20}$$

$$Y_{\mathrm{m}}(Z_{\mathrm{melt}}) = 0, \tag{21}$$

where $\theta_{\mathrm{cast}} = (T_{\mathrm{melt}} - T_{\mathrm{cast}})/\Delta T$.

With typical model parameters as

$$c_{\mathrm{pl}} \sim 495\,\mathrm{J\,kg^{-1}\,K^{-1}}, \quad c_{\mathrm{ps}} \sim 485\,\mathrm{J\,kg^{-1}\,K^{-1}}, \quad k_{\mathrm{l}} \sim 165\,\mathrm{W\,m^{-1}\,K^{-1}},$$

$$k_{\mathrm{s}} \sim 335\,\mathrm{W\,m^{-1}\,K^{-1}}, \quad L \sim 0.38\,\mathrm{m}, \quad T_{\mathrm{cast}} \sim 1396\,\mathrm{K}, \quad T_{\mathrm{melt}} \sim 1356\,\mathrm{K},$$

$$V_{\mathrm{cast}} \sim 0.018\,\mathrm{m\,s^{-1}}, \quad W \sim 0.0135\,\mathrm{m}, \quad \rho \sim 8000\,\mathrm{kg\,m^{-3}}, \quad \Delta H_{\mathrm{f}} \sim 205{,}000\,\mathrm{J\,kg^{-1}},$$

we have

$$Bi \approx 0.14, \quad K \approx 0.49, \quad \widetilde{Pe}_{\mathrm{l}} \approx 0.21, \quad \widetilde{Pe}_{\mathrm{s}} \approx 0.1, \quad St \approx 2.31, \quad \theta_{\mathrm{cast}} \approx -0.04.$$

It now remains to determine a numerical solution of the remaining equations.

4 Results

The numerical solution is not entirely straightforward, because the location of z_{melt} is not known a priori, and because the solid is initially of zero thickness; more details of how to overcome these difficulties are given in [9, 10, 12] and we proceed instead to a presentation of the key results. Figure 3a compares the location of the solidification front, as predicted by the two models, whereas Fig. 3b compares the heat flux Q at $y = 0$. This is defined as

$$Q = \begin{cases} -k_{\mathrm{l}} \frac{\partial T_{\mathrm{l}}}{\partial y} & \text{at } y = 0, \quad z < z_{\mathrm{melt}}, \\ -k_{\mathrm{s}} \frac{\partial T_{\mathrm{s}}}{\partial y} & \text{at } y = 0, \quad z > z_{\mathrm{melt}}. \end{cases}$$

Figure 3c and d compare the temperature profiles at $y = W$ and $y = 0$, respectively. Apart from minor differences near the start and end of solidification, the results of the two models agree very well, indicating that the asymptotic model captures the key features of the full model.

As a corollary to these results, we note that the form of the reduced model is similar to that of a recent thermoelastic model for the process that takes into account the air gap between the solidified metal shell and the mould wall [11, 12]; the air gap arises as a result of shrinkage when the metal solidified. In summary,

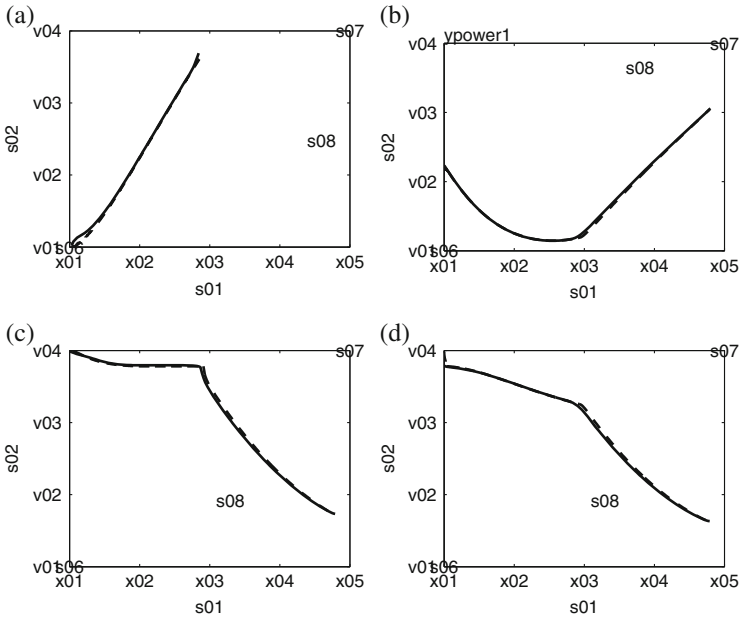


Fig. 3 Comparison of: (**a**) the location of the solidification front, y_m; (**b**) the heat flux at the outer edge of the copper strip ($y = 0$); (**c**) the temperature at the centreline ($y = W$); (**d**) the temperature at the outer edge of the copper strip ($y = 0$)

this suggests that solid-mechanical and fluid-mechanical effects can be decoupled, simply by replacing the molten metal velocity by V_{cast}; the resulting model can then be used for more detailed studies on segregation and crack formation within the solidified metal.

Acknowledgements SLM, BJF and SBGO'B acknowledge the support of the Mathematics Applications Consortium for Science and Industry (http://www.macsi.ul.ie), funded by the Science Foundation Ireland grant 12/IA/1683.

References

1. Aboutalebi, M.R., Hasan, M., Guthrie, R.I.L.: Coupled turbulent flow, heat and solute transport in continuous casting processes. Metall. Trans. B **26B**, 731–744 (1995)
2. Aboutalebi, M.R., Hasan, M., Guthrie, R.I.L.: Numerical study of coupled turbulent flow and solidification for steel slab casters. Numer. Heat Transfer A **28**, 279–297 (1995)
3. Amin, M.R., Greif, D.: Conjugate heat transfer during two-phase solidification process in a continuously moving metal using average heat capacity method. Int. J. Heat Mass Transfer **42**, 2883–2985 (1999)

4. Das, S.K., Sarkar, A.: Computational modelling of thermal transport phenomena in continuous casting process based on non-orthogonal control volume approach. Commun. Numer. Methods Eng. **12**, 657–671 (1996)
5. Lee, J.E., Han, H.N., Oh, K.H., Yoon., J.K.: A fully coupled analysis of fluid flow, heat transfer and stress in continuous round billet casting. ISIJ Int. **39**, 435–444 (1999)
6. Mahmoudi, J.: Mathematical modelling of fluid flow, heat transfer and solidification in a strip continuous casting process. Int. J. Cast Met. Res. **19**, 223–236 (2006)
7. Mahmoudi, J., Vynnycky, M., Fredriksson, H.: Modelling of fluid flow, heat transfer, solidification in the strip casting of copper base alloy. Part 3. Solidification – a theoretical study. Scand. J. Metall. **30**(3), 136–145 (2001)
8. Mahmoudi, J., Vynnycky, M., Sivesson, P., Fredriksson, H.: An experimental and numerical study on the modelling of fluid flow, heat transfer and solidification in a copper continuous strip casting process. Mater. Trans. **44**(9), 1741–1751 (2003)
9. Mitchell, S.L., Vynnycky, M.: An accurate finite-difference method for ablation-type stefan problems. J. Comput. Appl. Math. **236**, 4181–4192 (2012)
10. Mitchell, S.L., Vynnycky, M.: On the numerical solution of two-phase Stefan problems with heat-flux boundary conditions. J. Comput. Appl. Math. **264**, 49–64 (2014)
11. Vynnycky, M.: An asymptotic model for the formation and evolution of air gaps in vertical continuous casting. Proc. R. Soc. A **465**, 1617–1644 (2009)
12. Vynnycky, M.: On the onset of air-gap formation in vertical continuous casting with superheat. Int. J. Mech. Sci. **73**, 69–76 (2013)
13. Yao, M., Yin, H.B., Fang, D.C.: Real-time analysis on non-uniform heat transfer and solidification in mould of continuous casting round billets. ISIJ Int. **44**, 1696–1704 (2004)
14. Yao, M., Yin, H., Wang, J.C., Fang, D.C., Liu, X., Yu, Y., Liu, J.J.: Monitoring and analysis of local mould thermal behaviour in continuous casting of round billets. Ironmaking Steelmaking **32**, 359–368 (2005)
15. Zhang, L.F.: Computational fluid dynamics modeling: application to transport phenomena during the casting process. JOM **64**, 1059–1062 (2012)

Mathematical Modelling of the Coffee Brewing Process

K.M. Moroney, W.T. Lee, S.B.G. O'Brien, F. Suijver, and J. Marra

Abstract The drip filter coffee market is a multi-billion euro industry. Despite this, although the chemistry of coffee brewing has been investigated in great detail, the physics of the process has received relatively little attention. In order to explain in scientific terms correlations between the coffee quality and the process variables, a physical model is required. In this study, flow through a static, saturated coffee bed, under the influence of a pressure gradient, is described using a double porosity model. The model is parametrised using experimentally obtained data from a cylindrical flow-through cell containing a coffee bed. Mass transfer from the coffee grains to the interstitial water is modelled using two mechanisms; mass transfer from the surface of the grains and mass transfer from the interior (bulk) of the grains. Mass transfer resistances are estimated by fitting experimental data. Initially coffee extraction is dominated by mass transfer from the grain surface, while transfer from the kernel of the grain is the rate limiting mechanism once the surface coffee has been exhausted.

Keyword Coffee brewing

1 Introduction

Coffee, made from the seeds (beans) of the coffee plant, is among the most popular beverages consumed worldwide. The beans are roasted and ground and then some of their soluble content is extracted by hot water. The resulting solution of hot water and coffee solubles is called coffee. Despite its widespread consumption, coffee quality, even in coffee brewed by professional baristas, is very inconsistent. This difficulty arises from the dependency of coffee quality on a large number of process variables. Some of these include brew ratio, grind size and distribution, brewing

K.M. Moroney (✉) • W.T. Lee • S.B.G. O'Brien
MACSI, Department of Mathematics, University of Limerick, Limerick, Ireland
e-mail: kevin.moroney@ul.ie; william.lee@ul.ie; stephen.obrien@ul.ie

F. Suijver • J. Marra
Philips Research, Eindhoven, The Netherlands
e-mail: freek.suijver@philips.com; johan.marra@philips.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_36

time, water temperature, agitation, water quality and uniformity of extraction [4, 5]. One of the numerous ways of making coffee is to use a drip filter coffee machine. In this machine coffee is placed in a filter and hot water is poured over it. As the water percolates down through the bed the solubles are extracted and any undissolved solids in the solution are filtered out to give the final coffee brew. The aim of this study is to formulate mathematical models to describe coffee extraction and use them to investigate the influence of different brewing parameters on coffee quality with a particular focus on drip filter brewing. This paper outlines the formulation of a system of equations to model coffee extraction. Numerical solutions of the system are compared with experimental data. To the best knowledge of the authors, the model presented here, is one of the first experimentally validated models of coffee extraction from a coffee bed. This process is covered in much more detail in [3].

1.1 Measuring Coffee Quality

Coffee is composed of over 1800 different chemical components [4]. Such a complex chemistry makes it very difficult to find correlations between these different chemical components and the quality of a coffee beverage. In general, when required, coffee quality is evaluated by professional coffee tasters. A simple, but useful, alternative measure of coffee quality is defined in terms of the brew strength and extraction yield percentages of a coffee beverage. The perfect beverage has the optimum balance of brew strength and extraction yield. Brew strength or concentration is the ratio of mass of dissolved coffee in the beverage to volume. Extraction yield is the percentage of dry coffee grind mass that has extracted as solubles into the water . The desirable ranges of strength and extraction are specified by coffee associations throughout the world. International standards consider extraction yields of 18–22 % and brew strength of 1.15–1.55 % optimal. The difficulty in achieving these, lies in the fact that the extraction process depends on a number of factors including brew ratio, brewing time, water temperature, grind size and uniformity, water quality, coffee bed geometry and brewing method. Brew strength and extraction yield are related by the brew ratio (dry coffee grind mass to water volume used). Thus, in theory, for a given extraction yield, choosing the correct brew ratio will allow a given brew strength to be achieved. The ideal ranges specified by the Speciality Coffee Association of Europe (SCAE) are summarised by the Coffee Brewing Control Chart in Fig. 1. It should be noted that this measure does not take into account extraction uniformity within the bed and uneven extraction can have a negative effect on coffee quality. Given that this is the most widely accepted coffee quality standard, it will be used in this study. Thus, when focusing on quality, it seems reasonable to model extraction of coffee as one entity rather than consider the extraction of its various components, since the quality measure does not distinguish between them. This will be the approach followed here.

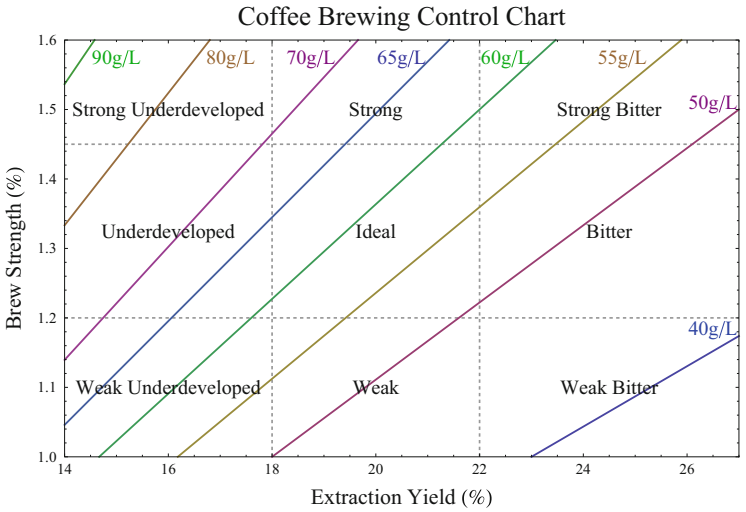


Fig. 1 Coffee Brewing Control Chart: each brew ratio determines a linear relationship between brew strength and extraction yield. The final position on the line is determined by the process parameters

2 Mathematical Modelling

The brewing process consists of three stages. Initially in the filling stage, hot water is poured on the dry coffee grounds and begins to fill the filter, but doesn't leave. In the steady state stage the bed is saturated, water is still entering the filter, but also leaving at the same rate. Finally in the draining stage no more water enters the bed but still drains out. Only the steady state stage is considered in this study. In this stage, the coffee bed is modelled as a static, saturated porous medium with the flow driven by a pressure gradient, which can be hydrostatic or mechanically applied. The bed is doubly porous since the coffee grains themselves have a porous cellular structure. Thus there are three distinct length scales in the coffee bed, the cell diameter, the (average) grain diameter and the bed depth. Leveraging the techniques of volume averaging, the transport of coffee solubles and water in the bed can be described by a system of partial differential equations. At a macroscopic level the coffee bed can be considered to consist of two phases. A highly permeable phase (h-phase) consisting of the pores between the coffee grains and a low permeability phase (l-phase) consisting of the coffee grains. At a microscopic level the coffee grains consist of two further phases. The pore or void space within the grains is called the v-phase, while the solid coffee cellular matrix is called the s-phase. Conservation equations for the h-phase, v-phase and s-phase can be formed at the microscale and macroscale. Using volume averaging the macroscopic quantities can be written in terms of the microscopic (measurable) quantities. Thus, at a macroscale, the system is represented as three overlapping continua, representing the intergranular

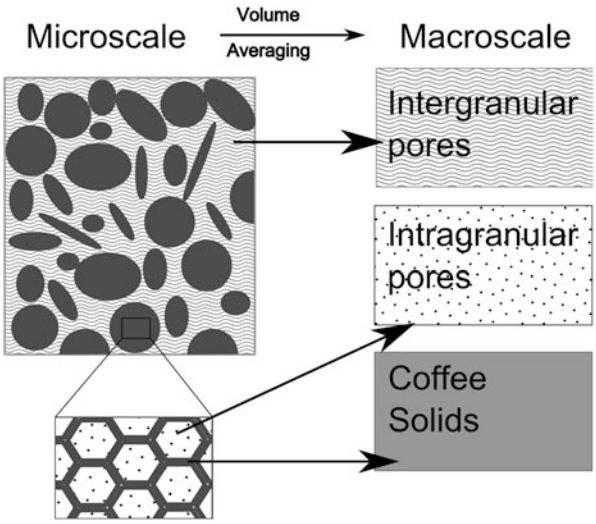


Fig. 2 Macroscopic equations are matched to microscopic equations using volume averaging. At the macroscopic level the system is represented by three overlapping continua for the intergranular pores (h-phase), intragranular pores (v-phase) and solid coffee (s-phase)

pores (pores between grains), the intragranular pores (pores within grains) and the coffee solids. Coffee and water transfers between phases on the microscale appear as source terms on a macroscale. The volume averaging procedure used here is outlined in [1, 2]. The complete model derivation is outlined in [3]. The averaging procedure results in five conservation equations. These represent conservation of coffee and water in the h-phase and v-phase, and conservation of coffee solid (s-phase). A schematic of the volume averaging procedure is shown in Fig. 2. The transfer terms between phases are represented by constitutive relations. A schematic of the transfer terms in the coffee bed is shown in Fig. 3.

Mass transfer from the coffee grains to the interstitial water is modelled using two mechanisms; mass transfer from the surface of the grains and mass transfer from the interior (bulk) of the grains. Transfer from the grains' surfaces also includes mass transfer from coffee fines which consist of single cells or broken cell walls. The two transfer mechanisms are used because coffee extracts from the surface layers of the grain much faster than from the grain kernel due to surface washing and proximity to the pores. Surface cells on a coffee grain also tend to be damaged from grinding, leading to reduced mass transfer resistances. This fast initial extraction has been observed in other studies such as [6] where 90 % of extraction occurred within 1 min. The volume averaging process and modelling of the transfer terms lead to Eqs. (1)–(7) . Initial and boundary conditions are specified depending on the brewing process used.

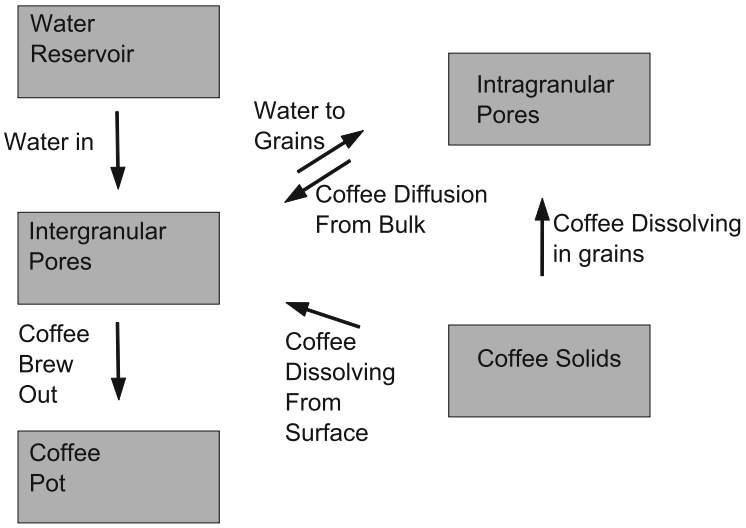


Fig. 3 Representation of transfer terms in coffee bed

2.1 Model Equations: Intergranular Pores

The following equations of the model are given below. The model equations assume isothermal conditions in the coffee bed. They are described in more detail in [3].

Coffee Conservation:

$$\begin{aligned}\phi_h \frac{\partial c_h}{\partial t} &= \frac{k_{sv1}^2 \phi_h^3}{36\kappa\mu(1-\phi_h)^2}\nabla\cdot(c_h(\nabla p_h + \rho g)) + \phi_h^{\frac{4}{3}} D_h^a \nabla^2 c_h \\ &+ \phi_h D_h^b \nabla^2 c_h - \alpha(1-\phi_h)\phi_v^{\frac{4}{3}}(c_h - c_v) \\ &+ \beta(1-\phi_h)(c_{\text{sat}} - c_h)\psi_s - \frac{6(1-\phi_h)m^2}{180\mu k_{sv2} l_l}\frac{\phi_v^3}{(1-\phi_v)^2}(p_h - p_v)c_h, \end{aligned} \tag{1}$$

Water Conservation:

$$\frac{k_{sv1}^2 \phi_h^3}{36\kappa\mu(1-\phi_h)^2}\nabla\cdot(\nabla p_h + \rho g) = \frac{6(1-\phi_h)m^2}{180\mu k_{sv2} l_l}\frac{\phi_v^3}{(1-\phi_v)^2}(p_h - p_v). \tag{2}$$

2.2 Model Equations: Intragranular Pores

Coffee Conservation:

$$\frac{\partial c_v}{\partial t} = \alpha \phi_v^{\frac{1}{3}}(c_h - c_v) + \frac{12D_v}{m^2}\frac{(1-\phi_v)}{\phi_v}(c_{\text{sat}} - c_v)(1 - \frac{c_v}{c_s})\psi_v + \frac{6m^2}{180\mu k_{sv2} l_l}\frac{\phi_v^2}{(1-\phi_v)^2}(p_h - p_v)c_h, \tag{3}$$

Water Conservation:

$$\frac{12D_v\phi_{c0}}{m^2}\left(\frac{c_{\text{sat}} - c_v}{c_s}\right)\psi_v = \frac{6m^2}{180\mu k_{sv2} l_l}\frac{\phi_v^3}{(1-\phi_v)^2}(p_h - p_v). \tag{4}$$

2.3 Model Equations: Coffee Solids

Coffee Conservation:

$$\frac{\partial \phi_v}{\partial t} = \beta\left(\frac{c_{\text{sat}} - c_h}{c_s}\right)\psi_s + \frac{12D_v\phi_{c0}}{m^2}\left(\frac{c_{\text{sat}} - c_v}{c_s}\right)\psi_v, \tag{5}$$

Fraction of Surface Coffee Remaining:

$$\frac{\partial \psi_s}{\partial t} = -\beta\left(\frac{c_{\text{sat}} - c_h}{c_s}\right) r_s \psi_s, \tag{6}$$

Fraction of Bulk Coffee Remaining:

$$\frac{\partial \psi_v}{\partial t} = -12\frac{D_v\phi_{c0}}{m^2}\left(\frac{c_{\text{sat}} - c_v}{c_s}\right) r_v \psi_v. \tag{7}$$

3 Results

The model equations are parametrised and mass transfer resistances fitted using experimental data obtained by Philips Research, Eindhoven. The data used here is from a coffee extraction experiment in a cylindrical flow-through cell containing ground coffee. Fresh water is pumped in at the top at a given pressure and flow rate and the concentration of coffee at the filter exit at the bottom and in the coffee pot is measured over the course of the extraction. Assuming that properties are

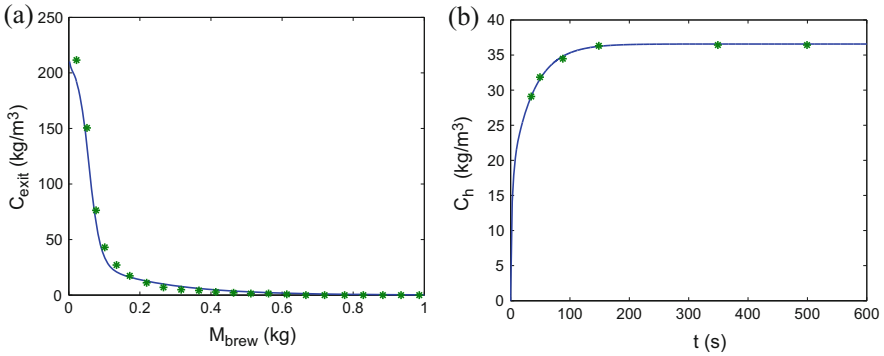


Fig. 4 (**a**) Plot of model numerical solution (*solid line*) and data (*asterisk*) for coffee concentration against mass flowed at filter exit for cylindrical flow through cell. In this experiment, fluid temperature was 90 °C, applied pressure was 2.3 bar with a flow rate of 250 ml min^{-1}. (**b**) Plot of model numerical solution (*solid line*) and data (*asterisk*) for coffee extraction profile for fixed volume experiment. Fluid temperature was 90 °C

homogeneous in any cross section allows reduction of the system to one spatial dimension. The bed porosity is fitted by matching the flow rate with the Kozeny-Carman equation. Mass transfer resistances are also fitted from batch extraction experiments where coffee grains are placed in a fixed volume of water and the coffee concentration of the water is measured at a particular time. This experiment is repeated for different times to build up an extraction profile. The comparison between numerical model simulations and experiment is shown in Fig. 4.

Acknowledgements This research was carried out in the Mathematics Applications Consortium for Science and Industry (www.macsi.ul.ie) in the University of Limerick funded by the SFI Investigator Award (MACSI) 12/IA/1683. Experimental data was obtained from experiments performed at Philips Research Laboratories in Eindhoven.

References

1. Bear, J., Cheng, A.H.D.: Modeling groundwater flow and contaminant transport. In: Theory and Applications of Transport in Porous Media. Springer, Heidelberg (2010)
2. Gray, W., Hassanizadeh, S.: Macroscale continuum mechanics for multiphase porous-media flow including phases, interfaces, common lines and common points. Adv. Water Resour. **21**(4), 261–281 (1998)
3. Moroney, K., Lee, W., O'Brien, S., Suijver, F., Marra, J.: Coffee extraction kinetics in a well mixed system. J. Math. Ind. **7**(3) (2017). doi: 0.1186/s13362-016-0024-6
4. Petracco, M.: Technology IV: Beverage Preparation: Brewing Trends for the New Millennium, pp. 140–164. Blackwell Science Ltd., New York (2008)
5. Rao, S.: Everything But Expresso. Independent Publisher (2010)
6. Voilley, A., Simatos, D.: Modeling the solubilization process during coffee brewing. J. Food Process Eng. **3**(4), 185–198 (1979)

Modelling Particle-Wall Interaction in Dry Powder Inhalers

Tuoi T.N. Vo, William Lee, Simon Kaar, Jan Hazenberg, and James Power

Abstract Dry powder inhalers deliver drugs in powdered form to the lungs. The drug is stored within the inhaler bound to an excipient. The drug-excipient conglomerate is broken apart in a vortex chamber by collisions with the walls and other conglomerates. During the initial doses, some drug adheres to the wall of the vortex chamber reducing the amount of drug delivered to the patient. We developed mathematical models for particle-wall adhesion to investigate why drug particles adhere to the wall of the vortex chamber. Two different models are developed to validate our results and a good agreement has been obtained. The first model describes the motion of particles in a turbulent flow field based on Stochastic Differential Equations (SDE). The second model is a continuum model of particle-wall adhesion based on Partial Differential Equations (PDE). This model focuses on the rate at which drug particles are captured by the wall and the time taken for drug particles to fill the wall area. Estimates of magnitudes of adhesive forces suggest that excipient particles do not adhere to the walls, while drug particles bind to the wall due to van der Waals forces when their velocity is below a critical value.

Keywords Dry powder inhalers • Mathematical modelling • Particle-wall adhesion

1 Introduction

Dry powder inhalers are devices that deliver solid drugs to the lungs. The powder particles are irregularly shaped conglomerations consisting of a lactose substrate with attached drug particles. The conglomerates are too large to enter the lungs and must be broken apart before they can be inhaled. This is accomplished by a vortex chamber in which turbulent circulatory flow causes particle-particle and particle-

T.T.N. Vo (✉) • W. Lee • S. Kaar
MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
e-mail: tuoi.vo@ul.ie

J. Hazenberg • J. Power
Teva Pharmaceuticals, Unit 301, Waterford Industrial Estate, Waterford, Ireland

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_37

wall collisions. These collisions detach drug particles which are small enough to enter the lungs, i.e. smaller than 5 μm. During the initial doses, some drug adheres to the wall of the vortex chamber, reducing the amount of drug delivered to the patient. This is thought to be due to the adhesion of particles to the wall of the chamber. In this paper, we developed mathematical models for particle-wall adhesion to investigate why drug particles adhere to the wall.

Models of particle-particle and particle-wall collisions were developed and analyzed in [4–6]. Although the effect of wall roughness was considered, particle-wall adhesion was not included. Heinl and Bohnet developed a model of particle-wall adhesion which was simulated with Computational Fluid Dynamics (CFD) [1, 2]. The particle-wall adhesion was achieved when the particle velocity before the particle-wall collision was smaller than a critical particle velocity, $v_{\mathrm{P,crit}}$, which is determined from the energy balance of a particle-wall collision. This model included most of the effects of adhesion which involved van der Waals forces and electrostatic forces.

Critical Velocity We used an energy balance argument to determine whether a particle colliding with the wall will adhere [1–3]. It is found that the excipient particles can leave the wall after impact due to their high kinetic energies. However, the drug particles may adhere to the wall after impact if their kinetic energies are not sufficient to overcome the van der Waals energies and energy loss. We have:

$$v_{\mathrm{P,crit}} = \sqrt{\frac{12E_{\mathrm{vdW}}}{\pi\rho_p d_p^3 e^2}}, \tag{1}$$

and the condition of adhesion is achieved if the particle velocity is below $v_{\mathrm{P,crit}}$, which is a critical velocity of particle. Here E_{vdW} is the van der Waals energy, ρ_p is the particle density, d_p is the particle diameter, and e is the coefficient of restitution. Estimates of the van der Waals force show that, to cause adhesion, a collision needs to occur at less than 1 m s^{-1}.

In the vortex chamber, the particles are dispersed by turbulence, are advected by flow field and centrifugal forces, collide with the wall, and are delivered to the lungs. In the flow field, the particles dynamics are influenced by drag forces, centrifugal forces and Coriolis forces. In this paper, a continuum model was developed to describe particle-wall adhesion based on partial differential equations (PDE). In this model, we assumed that drug particles are only released at the wall after collision and the Coriolis force was neglected. Results from this model predicted the rate at which drug particles are captured by the wall and the time for drug particles filling the wall area. However, they did not agree with results from an alternate model based on stochastic differential equations (SDE), which predicted a much greater particle density, especially in the centre of the vortex chamber. We demonstrate below that this is because the Coriolis force plays an important role in particle distribution and can not be neglected. Therefore the model was adjusted to take into account the effect of both the centrifugal force and Coriolis force.

2 The Model

Initially, we developed a SDE model to describe the motion of particles in a flow field in the vortex chamber. While this model rapidly gave an equilibrium distribution of particle positions and velocities in the absence of adhesion it was too computationally intensive to allow simulation of the dynamics on the wall adhesion timescale. Thus we developed a simple PDE model to describe the same phenomena. The PDE model focused on the rate at which drug particles are captured by the wall and the time taken for drug particles to fill the wall area.

2.1 PDE Model

We developed a continuum model to describe particle-wall adhesion based on partial differential equations (PDE). In this model, we assumed that drug particles are only released at the wall after collision and neglected the effect of the Coriolis force.

2.1.1 Model Equations

To illustrate particle-wall adhesion, a simple model was considered in one dimension:

$$\begin{aligned}
&\frac{\partial c}{\partial t} + v\frac{\partial c}{\partial x} = D\frac{\partial^2 c}{\partial x^2}, && 0 < x < L,\\
&\frac{dc_w}{dt} = Kc(c_{w0} - c_w), && \text{on } x = L,\\
&vc - D\frac{\partial c}{\partial x} = \frac{dc_w}{dt} - Q_c, && \text{on } x = L,\\
&c = 0, && \text{on } x = 0,\\
&c = 0,\ c_w = 0, && \text{at } t = 0,
\end{aligned} \tag{2}$$

where $x = 0$ is at the centre of the vortex chamber and $x = L$ is at the wall; $c(x, t)$ is the number of particles per unit volume in the vortex chamber; $c_w(t)$ is the number of particles per unit wall area; c_{w0} is the number of particles filling a unit wall area; v is the centrifugal velocity; D is the diffusion coefficient; K is the rate at which particles stick to the wall; and Q_c is the rate of particle release due to the collision of conglomerates with the wall. Here, we assumed that v and D are constants. In the partial differential equations:

(1) $v\frac{\partial c}{\partial x}$ represents outward motions of particles due to centrifugal forces;
(2) $D\frac{\partial^2 c}{\partial t^2}$ represents random motions of particles due to turbulent buffeting;

(3) $vc - D\frac{\partial c}{\partial x}$ represents the flux of particles, so the first boundary condition on $x = L$ describes absorption of particles by the wall;
(4) The second boundary condition represents the removal of particles at the centre, $x = 0$.

We can rewrite the system (2) in dimensionless form by using the following non-dimensional variables:

$$\bar{x} = \frac{x}{L}, \quad \bar{t} = \frac{t}{D/(KQ_cL)}, \quad \bar{c}_w = \frac{c_w}{c_{w0}}, \quad \bar{c} = \frac{c}{(Q_cL)/D},$$

to obtain (dropping overbars for convenience):

$$\begin{aligned}
&\varepsilon\frac{\partial c}{\partial t} + \mathrm{Pe}\frac{\partial c}{\partial x} = \frac{\partial^2 c}{\partial x^2}, \ 0 < x < 1,\\
&\frac{dc_w}{dt} = c(1 - c_w), \quad \text{on } x = 1,\\
&\frac{\partial c}{\partial x} = \mathrm{Pe}c - \alpha c(1 - c_w) + 1, \quad \text{on } x = 1,\\
&c = 0, \quad \text{on } x = 0,\\
&c = 0, \ c_w = 0, \quad \text{at } t = 0,
\end{aligned} \tag{3}$$

where

$$\mathrm{Pe} = \frac{vL}{D}, \quad \varepsilon = \frac{KQ_cL^3}{D^2}, \quad \alpha = \frac{KLc_{w0}}{D}$$

are dimensionless constants.

Assuming that $\varepsilon \ll 1$, the system (3) can be solved for $c(x, t)$ and $c_w(t)$ to obtain:

$$\begin{aligned}
c(x,t) &= \frac{e^{\mathrm{Pe}\,x} - 1}{\mathrm{Pe} + \alpha(e^{\mathrm{Pe}} - 1)(1 - c_w(t))},\\
c_w(t) &= 1 - \exp\left\{-\mathrm{LambertW}\left[\frac{\alpha(e^{\mathrm{Pe}} - 1)}{\mathrm{Pe}}\exp\left(\frac{(\alpha - t)(e^{\mathrm{Pe}} - 1)}{\mathrm{Pe}}\right)\right] + \frac{(\alpha - t)(e^{\mathrm{Pe}} - 1)}{\mathrm{Pe}}\right\},
\end{aligned} \tag{4}$$

where the LambertW function satisfies $\mathrm{LambertW}(t)\, e^{\mathrm{LambertW}(t)} = t$.

The dose release rate is defined as follows:

$$F(t) = \left(\frac{\partial c}{\partial x} - \mathrm{Pe}c\right)_{x=0} = \left(\frac{\partial c}{\partial x}\right)_{x=0} \qquad (c(x = 0) = 0),$$

i.e. the particle flux into the centre. From $(4)_1$, we obtain:

$$F(t) = \frac{\mathrm{Pe}}{\mathrm{Pe} + \alpha(e^{\mathrm{Pe}} - 1)(1 - c_w(t))}. \tag{5}$$

2.1.2 Results

Figures 1 and 2 show the analytical solutions of the number of particles per unit volume $c(x,t)$, the number of particles per unit wall area $c_w(t)$, and dose released $F(t)$ for various values of non-dimensional parameters (Pe, α and ε) using Eqs. (4) and (5). Parameter values used to generate these results are listed in Table 1. We see

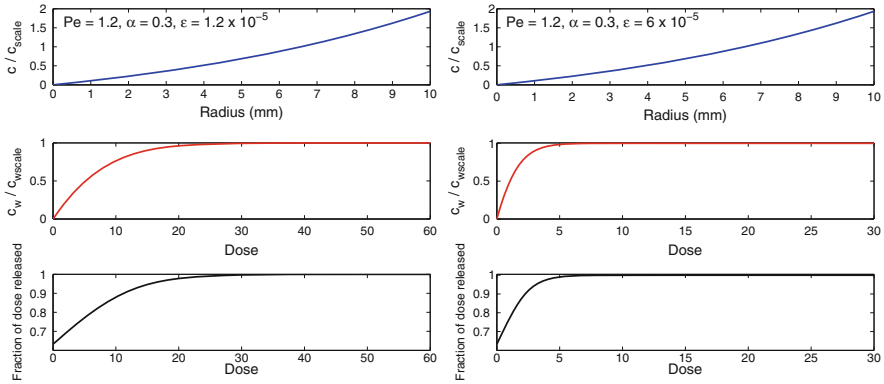


Fig. 1 Analytical solutions of the number of particles per unit volume $c(x,t)$, the number of particles per unit wall area $c_w(t)$, and dose released $F(t)$ in non-dimensional form

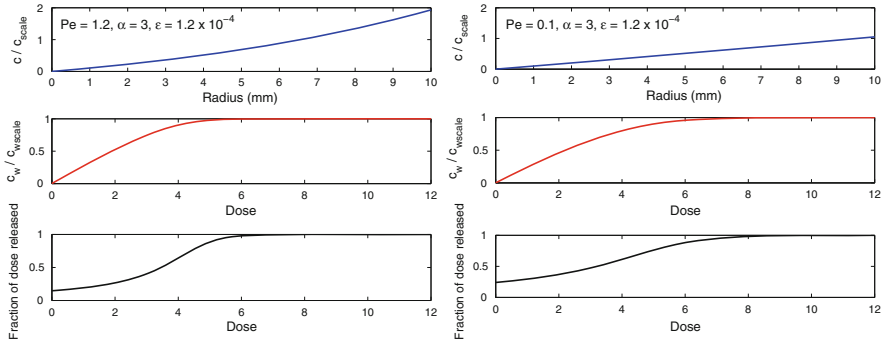


Fig. 2 Analytical solutions of the number of particles per unit volume $c(x,t)$, the number of particles per unit wall area $c_w(t)$, and dose released $F(t)$ in non-dimensional form

Table 1 Parameter values used to generate results in Figs. 1 and 2

Parameter	Value	Description
L	10 mm	Length scale
V	30 m s^{-1}	Average velocity of air flow
R_p	2.5×10^{-6} m	Radius of drug particles
$D = LV$	0.3 m^2 s^{-1}	Diffusion coefficient
$c_{w0} = \frac{1}{\pi R_p^2}$	5×10^{10} m^{-2}	1 active site per unit area

that there is an increase in particle number density, c, near the wall due to the effect of centrifugal forces and particle release at the wall after collision. At the time the wall is filled, $c_w = 1$, we obtain 100 % dose released.

The fraction of drug released in the first dose decreases from 65 % to about 20 % while increasing α from 0.3 to 3. In Fig. 1, affected doses decrease from 28 to 5 when ε increases from 1.2×10^{-5} to 6×10^{-5}. Increasing ε corresponds to increasing the rate of particle release at the wall, Q_c, or to increasing delivered doses. This means that reducing the amount of drug per dose increases the number of affected doses and the time for the drug to fill the wall area. In Fig. 2, as Pe is reduced from 1.2 to 0.1, the particle number density decreases and the affected doses increase.

The results of this model can be compared to experimental data, if available, using Pe, α and ε as fitting parameters.

2.2 Adjusted PDE Model

2.2.1 SDE Model

In this model, the equation of motion of a particle in a flow field $\mathbf{u} + \mathbf{w}$ describes its velocity $\mathbf{v}$ and position $\mathbf{x}$:

$$\begin{aligned} m\frac{d\mathbf{v}}{dt} &= \mu(\mathbf{u} + \mathbf{w} - \mathbf{v}), \\ \frac{d\mathbf{x}}{dt} &= \mathbf{v}, \end{aligned} \tag{6}$$

where $\mathbf{u}$ is the average air flow velocity, $\mathbf{w}$ is the turbulent air flow velocity modelled as a Weiner process, m is the particle mass, and μ is a friction constant.

The SDE model was solved by Monte Carlo simulation of a cloud of 1000 particles from which averaged quantities were calculated. The number density of drug particles was displayed in Fig. 3 (blue points).

2.2.2 The Adjusted Model

Although PDE model results could predict the rate at which drug particles are captured by the wall and the time taken for drug particles filling the wall area, they did not agree with the SDE model which has much greater particle density, especially in the centre of the vortex chamber. The explanation for this may be that the Coriolis effect causes an azimuthal velocity increase near the centre, retarding the rate at which particles reach the centre and exit the chamber. It means that the flow field used to derive the PDE model should be adjusted to include the effects of the Coriolis force.

The following model takes into account the effect of both the centrifugal force and Coriolis force and the sink term $c = 0$ at the chamber centre is replaced

with a reaction term in which the rate of removal of particles is proportional to the particle concentration. Restricting ourselves to two dimensions and working in polar coordinates, we have:

$$
\begin{aligned}
&\frac{dv_r}{dt} = \frac{v_\theta^2}{r} + \frac{\mu}{m}(u_r - v_r),\\
&\frac{dv_\theta}{dt} = -\frac{v_r v_\theta}{r} + \frac{\mu}{m}(u_\theta - v_\theta),\\
&\frac{\partial c}{\partial t} + \left(v_r - \frac{D}{r}\right)\frac{\partial c}{\partial r} = D\frac{\partial^2 c}{\partial r^2},\ R_1 < r < R_2,\\
&\frac{dc_w}{dt} = Kc(c_{w0} - c_w),\ \text{ on } r = R_2,\\
&\left(v_r - \frac{D}{r}\right)c - D\frac{\partial c}{\partial r} = \frac{dc_w}{dt} - Q_c,\ \text{ on } r = R_2,\\
&c = k_c\left[\left(v_r - \frac{D}{r}\right)c - D\frac{\partial c}{\partial r}\right] \text{ on } r = R_1.\\
&c = 0,\ c_w = 0,\ \text{ at } t = 0,
\end{aligned}
\tag{7}
$$

where this coordinate system introduces inertial forces: the Centrifugal force, $F_{cen} = m\frac{v_\theta^2}{r}$, and the Coriolis force, $F_{cor} = -m\frac{v_r v_\theta}{r}$. Notation description and parameter values used to generate results in Fig. 3 are listed in Table 2.

The system of PDEs was solved numerically using the open source OpenFOAM libraries [7]. These make use of a finite volume discretisation of the equations.

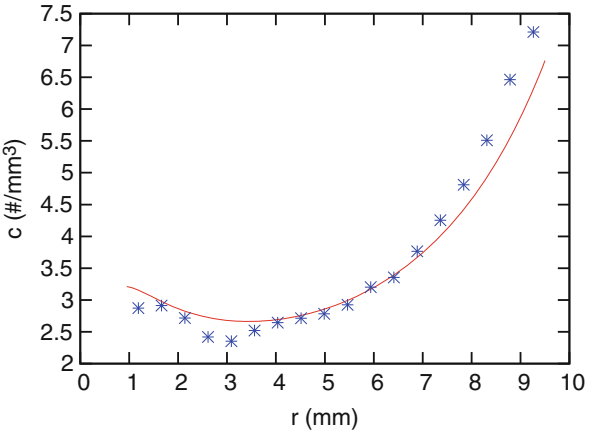


Fig. 3 Steady state concentration profile: results of the number density of drug particles, $c(r, t)$, from the SDE model (*blue*) and the adjusted PDE model (*red*)

Table 2 Notation used in the models

Notation	Unit/values	Description
$c(r,t)$	$\#\,\mathrm{m}^{-3}$	Number density of particles
$c_w(t)$	$\#\,\mathrm{m}^{-2}$	Number of particles per unit wall area
c_{w0}	$\#\,\mathrm{m}^{-2}$	Number of particles filling a unit wall area
m	kg	Particle mass
v_r	$\mathrm{m\,s}^{-1}$	Radial velocity of particles
v_θ	$\mathrm{m\,s}^{-1}$	Azimuthal velocity of particles
D	$\mathrm{m}^2\,\mathrm{s}^{-1}$	Diffusion coefficient
K	$\mathrm{m}^3\,\#^{-1}\,\mathrm{s}^{-1}$	Rate at which particles stick to the wall
Q_c	$\#\,\mathrm{m}^{-2}\,\mathrm{s}^{-1}$	Rate of particle release at the wall
k_c		Proportional constant
Q	$60\,\mathrm{L\,min}^{-1}$	Volume flux of an inhalation
H	4.6 mm	Height of vortex chamber
A	$2.93 \times 10^{-5}\,\mathrm{m}^2$	Area of both tangential inlets
R_2	9.5 mm	Radius of vortex chamber
R_1	0.95 mm	Radius of central exit
R_p	2.5×10^{-6} m	Radius of drug particles
η	$2 \times 10^{-5}\,\mathrm{kg\,m}^{-1}\,\mathrm{s}^{-1}$	Dynamic viscosity of air
u_r	$-Q/(2\pi rH)\,\mathrm{m\,s}^{-1}$	Radial velocity of air
u_θ	$Qr/(AR_2)\,\mathrm{m\,s}^{-1}$	Azimuthal velocity of air
μ	$6\pi\eta R_p$	Particle friction factor

Figure 3 shows a good agreement between results of the number density of drug particles, $c(r,t)$, from the SDE model (blue) and the adjusted PDE model (red).

3 Discussion

Our results show that the amount of drug retained by the vortex chamber is comparable with the amount of drug needed to form a monolayer on the inside of the chamber. These results suggest that the total amount of drug absorbed will remain the same irrespective of dose cup size and that only the rate at which this dose is absorbed depends on the dose cup. Estimates of magnitudes of adhesive forces suggest that excipient particles do not adhere to the walls, while drug particles bind to the wall due to van der Waals forces. Drug adhesion is primarily affected by Hamaker constant, surface roughness and coefficient of restitution. Simplified simulations of the particle motion in the chamber are consistent with this picture.

Experimental data is not available for the system of interest so a number of parameters have been extrapolated from similar systems. Measuring the parameters this paper has identified as being most relevant (i.e., Hamaker constant and coefficient of restitution) could be useful to improve the models. Combining the

model of particle adhesion developed in this paper with a full computational fluid dynamics model of the exact chamber geometry could yield insights into where exactly particle deposition is taking place. Finally, the existing models, although simple can be used to estimate the effects of changing various geometrical and aerodynamic factors.

Acknowledgements We gratefully acknowledge support from Enterprise Ireland's Innovation Partnership scheme, *Particle-wall Interaction in a Dry Powder Inhaler* IP-2012-0171, and the Mathematics Applications Consortium for Science and Industry (www.macsi.ul.ie) funded by the Science Foundation Ireland (SFI) Investigator Award 12/IA/1683. Dr Vo thanks the New Foundations Award 2013 from Irish Research Council.

References

1. Heinl, E., Bohnet, M.: Numerical simulation of particle wall adhesion in gas–solid flows. Chem. Eng. Technol. **27**(11), 1143–1146 (2004)
2. Heinl, E., Bohnet, M.: Calculation of particle-wall adhesion in horizontal gas–solids flow using CFD. Powder Technol. **159**(2), 95–104 (2005)
3. Israelachvili, J.N.: Intermolecular and Surface Forces, 3rd edn. Academic Press, Elsevier Inc., Amsterdam (2011)
4. Sommerfeld, M.: Analysis of collision effects for turbulent gas-particle flow in a horizontal channel: part I. Particle transport. Int. J. Multiphase Flow **29**, 675–699 (2003)
5. Sommerfeld, M., Huber, N.: Experimental analysis and modelling of particle-wall collisions. Int. J. Multiphase Flow **25**, 1457–1489 (1999)
6. Sun, K., Lu, L., Jiang, H.: Modelling of particle deposition and rebound behaviour on ventilation ducting wall using an improved wall model. Indoor Built Environ. **20**, 300–312 (2011)
7. Weller, H.G., Tabor, G., Jasak, H., Fureby, C.: A tensorial approach to computational continuum mechanics using object orientated techniques. Comput. Phys. **12**(6), 620–631 (1998). OpenFOAM www.openfoam.com/

Optimising Copying Accuracy in Holographic Patterning

Dana Mackey, Paul O'Reilly, and Izabela Naydenova

Abstract We propose a partial differential equations model for the formation and evolution of a holographic grating in a photopolymer system and use perturbation methods and numerical simulations in order to investigate the dynamical mechanism by which distortions of the illumination pattern arise during recording. The parameters of interest are diffusion and photopolymerization rates as well as exposure time, for which we seek to determine regimes which allow for high fidelity copying.

Keywords Holographic patterning • Photopolymerization-diffusion

1 Introduction

Holography has many applications such as holographic displays, optical elements and sensors, security holograms and holographic data storage. A hologram is essentially a recording of an interference pattern created by an object beam and a reference beam in a photosensitive material; in all applications the accuracy with which this pattern is copied is crucial for the performance of the hologram. Photopolymers are often the material of choice in holographic patterning because of qualities such as versatility, self-processing nature, good dynamic range and relatively low cost. A photopolymer system consists of one or two monomers, photoinitiator and sensitizing dye, all dispersed in a binder matrix. Following exposure to an illumination pattern, the monomer polymerizes and the recorded holographic grating is given by the spatial variation of the refractive index, which results from changes in the relative density of components.

A mathematical model was introduced in [1] and [2], which generalizes the standard monomer diffusion equation of [5] by differentiating between mobile (diffusing) polymer chains and immobile ones. This model supports the "two way

D. Mackey (✉) • P. O'Reilly
School of Mathematical Sciences, Dublin Institute of Technology, Dublin, Ireland
e-mail: dana.mackey@dit.ie; paul.oreilly8@student.dit.ie

I. Naydenova
Industrial and Engineering Optics Centre, Dublin Institute of Technology, Dublin, Ireland
e-mail: izabela.naydenova@dit.ie

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_38

diffusion theory", proposed in [4] and [3], which states that the diffusion of some polymer chains away from bright fringes leads to a reduction in the refractive index modulation and is one of the processes responsible for the experimentally observed poor diffraction efficiency at high recording frequencies.

In this paper, we use an improved version of this model to investigate regimes which allow for high accuracy photopatterning. The main parameters influencing the copying fidelity are the spatial frequency of recording, which defines the distance over which the monomer and polymer molecules have to travel, and the intensity of recording, which determines the concentration of free radicals and thus the rate at which monomer molecules are converted into polymer chains and the rate at which the chains are terminated. It has been suggested by experimental and theoretical observations [4, 5], that distortions of the illumination pattern tend to occur when the ratio of monomer diffusion to polymerization rate is small and we verify that our model reproduces this observation.

2 Photopolymerization-Diffusion Model

We study a simple grating recorded by the interference of two coherent beams of intensities I_1 and I_2, which create the illumination pattern

$$I(x) = I_0\,(1 + V\,\cos(kx)),$$

where k is the grating wavenumber, $I_0 = I_1 + I_2$ and $V = 2\sqrt{I_1 I_2}/(I_1 + I_2)$ are the overall intensity and visibility of the interference pattern, respectively. The holographic grating formation then proceeds in three steps: initiation, propagation and termination. Upon illumination, the sensitizing dye absorbs a photon and reacts with the electron donor to produce free radicals; in the presence of monomer these free radicals initiate polymerization. During the propagation step, free radicals and monomer molecules interact and produce growing polymer chains. At the termination step, two free radicals interact and the polymer chains stop growing.

The photopolymerization and diffusion processes described above can be captured by the following partial differential equations (see [1, 2])

$$\frac{\partial m}{\partial t} = D_m \frac{\partial^2 m}{\partial x^2} - \Phi(t)\,F(x)\,m \tag{1}$$

$$\frac{\partial p}{\partial t} = D_p \frac{\partial^2 p}{\partial x^2} + \Phi(t)\,\left[F(x)m - \Gamma p^2\right], \qquad 0 \le x \le \Lambda; \quad t \ge 0 \tag{2}$$

$$\frac{\partial q}{\partial t} = \Phi(t)\,\Gamma p^2 \tag{3}$$

where $m(x,t)$, $p(x,t)$, $q(x,t)$ are the concentrations of monomer, short and long polymers, respectively, $\Lambda = 2\pi/k$ is the grating period, $F(x) = F_0\,(1 + V\cos(kx))$

is the polymerization rate (assumed proportional to the light intensity) and $\Phi(t)$ is a step function equal to 1 during the light exposure and 0 afterwards. The monomer and polymer diffusion coefficients, D_m and D_p, are assumed constant, which is a reasonable simplification for short exposures. The immobilization rate, introduced in [1], depends mainly on cross-linking of short chains and is now more correctly described as proportional to p^2 (where Γ denotes the rate constant). The initial conditions are $m(x,0) = m_0,\ p(x,0) = q(x,0) = 0$ for all $0 \le x \le \Lambda$, and we assume zero-flux boundary conditions for all species. This model describes a simple mechanism for monomer and short chain polymer diffusion, coupled with photopolymerization and immobilization (i.e. transition from the diffusing to immobile state). Tracking the evolution of polymer chain lengths is beyond our scope here.

The refractive index of a material consisting of a mixture of components can be calculated with the well-known Lorentz-Lorenz equation

$$\frac{n^2-1}{n^2+2} = \sum_i \Phi_i \frac{n_i^2-1}{n_i^2+2}$$

where n is the effective refractive index of the mixture, n_i are the refractive indices of the components (monomer, polymer and binder) determined separately from spectrophotometric measurements, and Φ_i are the normalized concentrations of the components (e.g. $\Phi_m = m/(b+m+p)$, where m, p and b denote concentrations of monomer, polymer and binder, respectively). See [1] for more details and numerical values. In what follows, we use the refractive index modulation

$$\Delta n(t) = n_{max}(t) - n_{min}(t)$$

as a measure of the grating strength. With the choice of non-dimensional variables $\bar{x} = \frac{x}{\Lambda}\ \bar{t} = \frac{t}{t_0}, \bar{m} = \frac{m}{m_0}, \bar{p} = \frac{p}{m_0}, \bar{q} = \frac{q}{m_0}$ the system becomes (after dropping bars)

$$\frac{\partial m}{\partial t} mt = \alpha \frac{\partial^2 m}{\partial x^2} - \Phi(t)\,\beta f(x)\,m, \tag{4}$$

$$\frac{\partial p}{\partial t} = \alpha\varepsilon \frac{\partial^2 p}{\partial x^2} + \Phi(t)\left[\beta f(x) m - \gamma p^2\right], \qquad 0 \le x \le 1; \quad t \ge 0 \tag{5}$$

$$\frac{\partial q}{\partial t} = \Phi(t)\,\gamma p^2 \tag{6}$$

where

$$\alpha = \frac{D_m t_0}{\Lambda^2}, \quad \varepsilon = \frac{D_p}{D_m}, \quad \beta = t_0 F_0, \quad \gamma = m_0 t_0 \Gamma, \quad f(x) = 1 + \cos(2\pi x)$$

We also have $m(x,0) = 1$, $p(x,0) = q(x,0) = 0$, as well as zero-flux boundary conditions at $x = 0,\ 1$. The reference time t_0 reflects the light exposure timescale.

3 Perturbation Analysis and Numerical Simulations

We use perturbation methods to study the solutions of this system in the particular cases when $\alpha \ll \beta$ (the diffusion rate is much smaller than the polymerization rate) and $\beta \ll \alpha$ (polymerization rate smaller than diffusion rate). We introduce the small parameter $\epsilon = D_p/D_m$, as diffusion of short polymers is always much slower than that of monomers and also assume that the immobilization rate γ is slower than the polymerization rate β (reflected in all subsequent scaling choices). We use $D_m = 10^{-12}$ and $D_p = 10^{-14}$ m^2/s, consistent with the values determined in [1, 4]. We first assume infinite light exposure so that $\Phi(t) \equiv 1$ in (4)–(6) and study the long term behaviour of monomer, polymer and refractive index (the numerical solutions are shown in Figs. 1 and 2). The evolution of the refractive index modulation $\Delta n(t)$ is then investigated for different finite exposure periods, t_e and the graphs are shown in Fig. 3. In this case we let $\Phi(t) \equiv 0$ for $t > t_e$ so after exposure the model is governed by diffusion equations with initial conditions given by the concentrations at the end of exposure, $m(x, t_e)$, $p(x, t_e)$ and $q(x, t_e)$. Since the monomer and mobile polymer quickly assume spatially homogeneous steady states, the refractive index spatial variation will be completely determined by that of the immobile polymer.

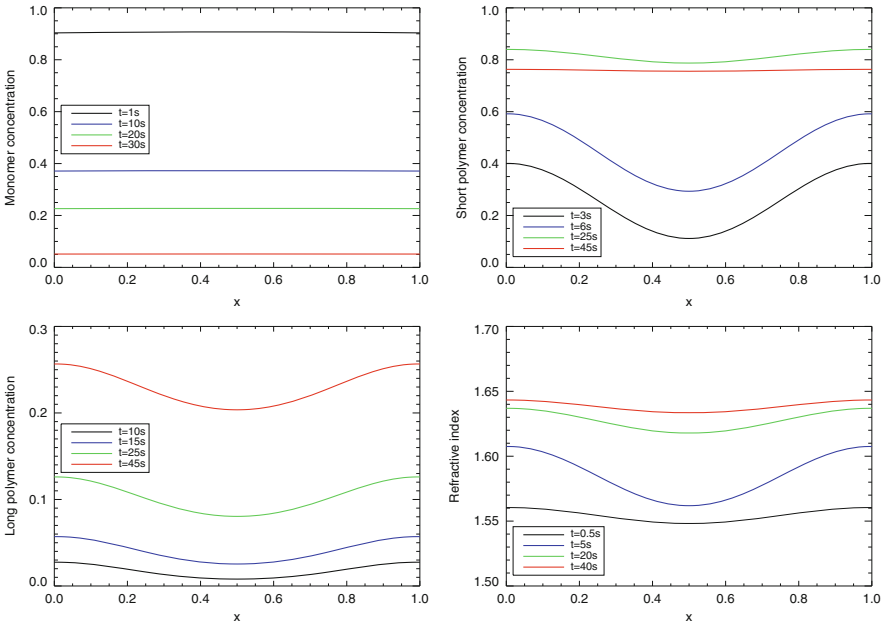


Fig. 1 Long term evolution of monomer, polymers and refractive index in the case of continuous exposure. Here, $\alpha = 1$, $\varepsilon = \beta = 0.01$ and $\gamma = 0.001$

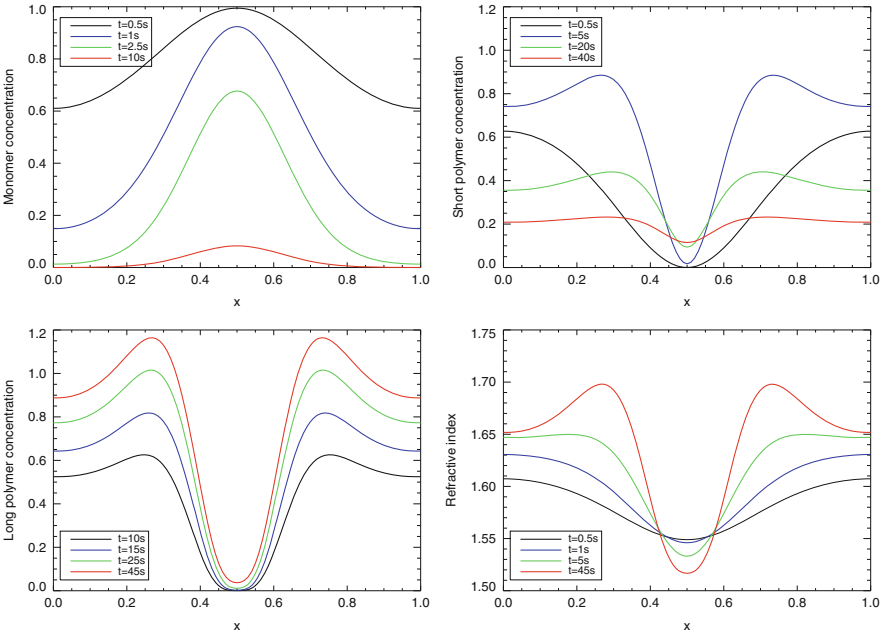


Fig. 2 Long term evolution of monomer, polymers and refractive index in the case of continuous exposure. Here, $\alpha = \varepsilon = 0.01$, $\beta = 1$ and $\gamma = 0.1$

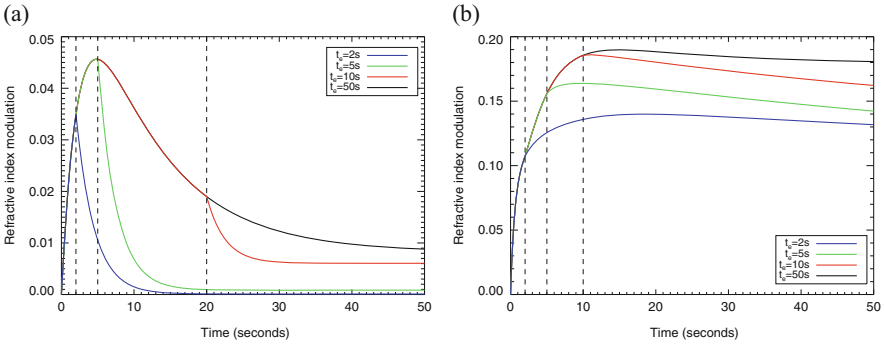


Fig. 3 The evolution of the refractive index modulation for different exposure times. (**a**) $\beta \ll \alpha$ ($\Lambda = 10^{-6}$ m, $F_0 = 0.01\,\mathrm{s}^{-1}$). (**b**) $\alpha \ll \beta$ ($\Lambda = 10^{-5}$ m, $F_0 = 1\,\mathrm{s}^{-1}$)

3.1 Polymerization Rate Less than Diffusion Rate ($\beta \ll \alpha$)

Without loss of generality we choose $\alpha = D_m t_0/\Lambda^2 = 1$ (achieved, for example, with $t_0 = 1s$, $\Lambda = 10^{-6}m$, which corresponds to a recording frequency of 1000 lines/mm) and scale $\beta = \varepsilon\beta'$, $\gamma = \varepsilon^2\gamma'$, where β', γ' are $O(1)$. Multiple scales expansions are appropriate since the polymers evolve more slowly than the monomer and can be used to approximate the initial behaviour for all species. Using the short time t and long time $\tau = \varepsilon t$, we find

$$m(x,t,\tau) = \mathrm{e}^{-\tau} + \varepsilon\left[\frac{1}{8\pi^2}\,\tau\mathrm{e}^{-\tau} + \frac{1}{4\pi^2}\,\mathrm{e}^{-\tau}\left(\mathrm{e}^{-4\pi^2 t} - 1\right)\cos(2\pi x)\right] + \cdots$$

$$p(x,t,\tau) = p_0(x,\tau) + \cdots = 1 - \mathrm{e}^{-\tau} + \frac{\mathrm{e}^{-\tau} - \mathrm{e}^{-4\pi^2\tau}}{4\pi^2 - 1}\,\cos(2\pi x) + \cdots$$

$$q(x,t,\tau) = \varepsilon\int_0^\tau p_0^2(x,s)\,ds + \cdots$$

These are valid up to $t = O(1/\varepsilon)$ (although the monomer expansion can be shown to hold as $t \to \infty$) and on this time scale it can be seen that the leading order spatial dependence is sinusoidal. This behaviour is confirmed by the numerical simulations shown in Fig. 1. We also plot the evolution of the refractive index modulation, $\Delta n(t)$, for different exposure times in Fig. 3a. Note the sharp decrease in $\Delta n(t)$ after the light exposure is stopped, which confirms the poor response of the photopolymer system at high recording frequencies, as discussed in Sect. 1.

3.2 Diffusion Rate Less than Polymerization Rate ($\alpha \ll \beta$)

When the diffusion rate is much slower than the polymerization rate, we choose $\beta = t_0 F_0 = 1$, $\alpha = \varepsilon \ll 1$ and $\gamma = \varepsilon\gamma'$. (We take $\Lambda = 10^{-5}$ m, which corresponds to a low frequency of 100 lines/mm.) A standard perturbation approach gives

$$m(x,t) = \mathrm{e}^{-(1+\cos(2\pi x))\,t} + \varepsilon\,\frac{2}{3}\pi^2 t^2\,\mathrm{e}^{-(1+\cos(2\pi x))\,t}\,(t + 3\cos(2\pi x) - t\cos(4\pi x)) \tag{7}$$

which is not valid for large t and $x \approx \frac{1}{2}$ (the point of zero illumination). This suggests introducing the variables $\xi = (x - \frac{1}{2})\varepsilon^{-1/4}$, $\tau = \varepsilon^{1/2}t$ to obtain

$$\frac{\partial m}{\partial \tau} = \frac{\partial^2 m}{\partial \xi^2} - 2\pi^2\xi^2\,m + \varepsilon\,\frac{2\pi^4}{3}\xi^4\,m, \tag{8}$$

$$\frac{\partial p}{\partial \tau}p\tau = \varepsilon\,\frac{\partial^2 p}{\partial \xi^2} + 2\pi^2\xi^2 m - \varepsilon\,\frac{2\pi^4}{3}\xi^4\,m - \varepsilon\gamma' p^2 \tag{9}$$

$$\frac{\partial p}{\partial \tau}q\tau = \varepsilon\gamma' p^2 \tag{10}$$

with $-\infty \leq \xi \leq \infty$, $\tau \geq 0$. The leading order solution for the monomer equation is

$$m_0(\xi,\tau) = \sum_{n=0}^{\infty} C_{2n} \mathrm{e}^{-\pi\sqrt{2}(4n+1)\,\tau - \frac{\pi}{\sqrt{2}}\xi^2} H_{2n}\left((2\pi^2)^{1/4}\,\xi\right), \tag{11}$$

where H_n are Hermite polynomials and $C_{2n} = \frac{\sqrt{2}}{2^n(2n)!!}$. The inner and outer expansions (11) and (7) accurately describe the behaviour of the monomer concentration for all times and can be used to determine similar approximations for the polymer functions; all these approximations match the numerical solutions shown in Fig. 2. Note that the refractive index profile no longer resembles the sinusoidal illumination as it develops multiple maxima per grating period. Figure 3b shows the refractive index modulation in this case can achieve a high saturation value.

4 Conclusions

We have investigated a possible mechanism by which distortions appear in the holographic grating created by a simple sinusoidal illumination pattern, using the ratio of diffusion to polymerization as our main parameter. When the polymerization rate is slower than the diffusion rate, the recorded grating profile resembles the sinusoidal interference pattern however, the refractive index modulation drops rapidly after exposure and the system is characterized by poor diffraction efficiency. When the diffusion is slower than polymerization we observe distortions of the illumination pattern after long exposure times ($t = O(1/\varepsilon)$. In this latter case, the refractive index modulation is much higher and keeps rising even after exposure is stopped, consistent with experimental observations [1, 4]. A more comprehensive perturbation analysis is, however, needed in order to study the interplay between diffusion rate, polymerization rate and exposure time. This study will also be generalized to more complex two-dimensional gratings produced with multiple beam holography as well as more complex photopolymer systems containing nanoparticle dopants.

Acknowledgements Many thanks to E. Benilov and S. O'Brien (MACSI) for useful comments.

References

1. Babeva, T., Naydenova, I., Mackey, D., Martin, S., Toal, V.: Two-way diffusion model for short exposure holographic grating, formation in acrylamide based photopolymers. J. Opt. Soc. Am. B **27**(2), 197–203 (2010)
2. Mackey, D., Babeva, T., Naydenova, I., Toal, V., Babeva, T., Naydenova, I., Toal, V.: A diffusion model for spatially dependent photopolymerization. In: Fitt, A.D., Norbury, J., Ockendon, H.,

Wilson, E. (eds.) Progress in Industrial Mathematics at ECMI 2008. Mathematics in Industry, vol. 15, pp. 253–259. Springer, Berlin, Heidelberg, New York (2010)
3. Martin, S., Naydenova, I., Jallapuram, R., Howard, R., Toal, V.: Two-way diffusion model for the recording mechanism in a self developing dry acrylamide photopolymer. Proc. SPIE **6252**, 62525–625217 (2006)
4. Naydenova, I., Jallapuram, R., Howard, R., Martin, S., Toal, V.: Investigation of the diffusion processes in a self-processing acrylamide-based photopolymer system. Appl. Opt. **43**, 2900–2905 (2004)
5. Zhao, G., Mouroulis, P.: Diffusion model of hologram formation in dry photopolymer materials. J. Mod. Opt. **41**, 1929–1939 (1994)

MS 10
MINISYMPOSIUM: MATHEMATICAL AND NUMERICAL MODELLING OF THE CARDIOVASCULAR SYSTEM

Organizers

Piero Colli Franzone[1], Luca F. Pavarino[2] and Simone Scacchi[3]

Speakers

Toni Lassila[4]
Computational Simulation of Heart Function with an Orthotropic Active Strain Model of Electromechanics

Martin Weiser[5]
Spectral Deferred Correction Methods for Adaptive Electro-Mechanical Coupling in Cardiac Simulation

Dorian Krause[6]
A Lightweight Approach to Parallel Adaptivity: Design, Implementation and Application in Electrophysiology

[1]Piero Colli Franzone, Università degli Studi di Pavia, Pavia, Italy.
[2]Luca F. Pavarino, Università degli Studi di Milano, Milano, Italy.
[3]Simone Scacchi, Università degli Studi di Milano, Milano, Italy.
[4]Toni Lassila, EPFL, Lausanne, Switzerland.
[5]Martin Weiser, ZIB, Berlin, Germany.
[6]Rolf Krause, Università della Svizzera Italiana, Lugano, Switzerland.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_39

Daniele Boffi[7]
Advances in the Mathematical Theory of the Finite Element Immersed Boundary Method

Joakim Sundnes[8]
Computational Models of Electro-Mechanical Interactions in the Heart

Paola Causin[9]
Impact of Blood Flow in Ocular Pathologies: Can Mathematical and Numerical Modeling Help Preventing Blindness?

Christian Vergara[10]
Inexact Schemes for the Fluid-Structure Interaction with Application to the Ascending Aorta

Naima Aissa[11]
Global Existence of Weak Solutions to an Angiogenesis Model

Keywords

Biomedical science
Cardiac fluid-mechanical and electrical activity
Cardiovascular system

Short Description

The numerical simulation of the cardiac fluidomechanical and electrical activity is a very challenging task. Different multiscale and nonlinear effects such as the propagation of the electrical activation front, the chemical reactions within the ion channels, the orthotropic fiber architecture, the constitutive laws and active tension of the cardiac tissue have to be properly taken into account. Furthermore, the numerical methods required to simulate efficiently such complex models needs to be carefully devised and adapted to properly couple/decouple the different submodels. This minisymposium aims at bringing together researchers in computational cardi-

[7]Daniele Boffi, Università degli Studi di Pavia, Pavia, Italy.
[8]Joakim Sundnes, SIMULA, Oslo, Norway.
[9]Paola Causin, Università degli Studi di Pavia, Pavia, Italy.
[10]Christian Vergara, MOX, Politecnico di Milano, Milano, Italy.
[11]Naima Aissa, Ecole Politechnique, Paris, France.

ology, focusing on the latest developments and on the new research pathways and applications.

ECMI motivation/relevance: In silico studies of the cardiovascular system are very relevant to ECMI since they are a crucial part of the ongoing efforts to bridge advanced research and clinical applications, as e.g. in the Virtual Physiological Human (VPH) project. The main goal of these studies is to develop, test and implement integrative biomedical science and technology-facilitated applications, as well as to improve current simulation techniques.

Advances in the Mathematical Theory of the Finite Element Immersed Boundary Method

Daniele Boffi, Nicola Cavallini, and Lucia Gastaldi

Abstract The Immersed Boundary Method (IBM) is an effective mathematical model and approximation scheme for the discretization of biological systems which involve the interaction of fluids and solids. The Finite Element IBM (FE-IBM) proved to be competitive with respect to the original IBM (based on finite differences and on a suitable approximation of a Dirac delta function) in several aspects: in particular, the position of the solid can be dealt with in a natural way by taking advantage of the underlying variational formulation (thus avoiding the use of the delta function); moreover, the use of finite elements allows for sharp pressure jumps when discontinuous pressure schemes are adopted. Recently [see Boffi et al. (Coupled Problems 2013. Computational Methods for Coupled Problems in Science and Engineering V, Cimne, 2013)], a fully variational approach of the FE-IBM has been introduced, which can be shown to be unconditionally stable with respect to the time discretization. The novelty consists in the treatment of the coupling between the solid and the fluid: in the standard formulation, this is given by a differential equation stating that the velocity of the solid is equal to that of the fluid, while in the new formulation this coupling is imposed in a weak form. A rigorous mathematical analysis shows the stability of the coupling and the unconditional time stability.

Keywords Finite element immersed boundary method

D. Boffi (✉) • N. Cavallini
Dipartimento di Matematica "F. Casorati", Università degli Studi di Pavia, Via Ferrata 1, 27100 Pavia, Italy
e-mail: daniele.boffi@unipv.it; nicola.cavallini@unipv.it

L. Gastaldi
DICATAM, Dipartimento di Ingegneria Civile, Architettura, Territorio, Ambiente e di Matematica, Università degli Studi di Brescia, via Branze 43, 25123 Brescia, Italy
e-mail: lucia.gastaldi@unibs.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_40

1 Introduction

When Peskin introduced the Immersed Boundary Method (IBM) he was mainly motivated by the Fluid Structure Interaction problem originated from the blood flow during the heartbeat. The methodological competition about this method is devoted to improve its conservation properties, in particular mass preservation is a key performance indicator, see [14] for a method's review.

After the introduction of the Finite Elements IBM (FE-IBM) by Boffi and Gastaldi, [1], several applicative studies explored the performances of the method (see [2–7, 13]). Moreover several theoretical and applicative works explore the mass conservation properties of enhanced pressure spaces, [8–11] conservative and approximation properties of several finite elements schemes.

A Distributed Lagrangian Multiplier (DLM) formulation of the finite element immersed boundary method has been recently introduced in [11]. This formulation is characterized by several analogies with the Fictitious Domain method [12] and shows very good results in terms of mass conservation and CFL stability. In this paper we will present numerical results for this scheme, which complement the ones reported in [11]. More precisely, while in [11] the main emphasis was on the CFL condition and on the mass conservation property, here we focus particularly on the stability of the scheme. More specifically, Table 1 shows what are the conditions

Table 1 Test cases to compare the CFL stability of IBM and DLM schemes

Δt	h_s (mesh size of the structure domain)						
	$\frac{1}{64}$	$\frac{1}{48}$	$\frac{1}{40}$	$\frac{1}{32}$	$\frac{1}{24}$	$\frac{1}{16}$	$\frac{1}{8}$
IBM *codimension one*							
1×10^{-2}	0.943	0.942	0.942	0.941	0.941	0.941	0.940
2×10^{-2}	CFL	0.942	0.941	0.941	0.940	0.940	0.940
3×10^{-2}	CFL	CFL	CFL	0.941	0.940	0.940	0.939
5×10^{-2}	CFL	CFL	CFL	0.795	0.940	0.939	0.938
1×10^{-1}	CFL	CFL	CFL	CFL	CFL	0.574	0.936
DLM *codimension one*							
1×10^{-2}	inf-sup	inf-sup	inf-sup	1.023	1.022	1.022	1.023
2×10^{-2}	inf-sup	inf-sup	inf-sup	1.022	1.022	1.022	1.022
3×10^{-2}	inf-sup	inf-sup	1.023	1.022	1.022	1.022	1.022
5×10^{-2}	inf-sup	inf-sup	inf-sup	1.021	1.021	1.021	1.022
1×10^{-1}	inf-sup	inf-sup	1.021	1.020	1.020	1.020	1.020

The test case is the one where an ellipsoidal structure aims to a circular equilibrium position. The elastic constant is $\kappa = 5$, $h_x = 1/32$, the simulated time is $T = 2$. In these tests we experience two different instabilities. The DLM scheme is bounded by an inf-sup stability condition. The IBM scheme experiences a CFL instability. Diffusivity measure is given by initial and final area: $|1 - A/A_0| < 3\,\%$, $|1 - A/A_0| < 10\,\%$, $|1 - A/A_0| < 20\,\%$, $|1 - A/A_0| \geq 20\,\%$

linking fluid and solid mesh sizes for the stability of the scheme, both in the case of FE-IBM and DLM.

2 Problem Setting

The fluid domain is $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, the structure domain is denoted by $\mathscr{B}_t \subset \Omega$ and is immersed into the fluid one. The fluid is incompressible and the structure is viscoelastic, a key assumption in IBM. In fact in IBM we assume that the stress tensor of an undifferentiated media is the sum of fluid and solid stress tensors. The fluid stress tensor acts all over Ω and the structure stress tensor is located on the structure domain $\mathscr{B}_t$. Fluid and structure are coupled applying the principle of virtual work.

At a given time t the structure lays on the time dependent domain $\mathscr{B}_t$, having codimension 0 or 1. We assume that there is no intersection between the structure and fluid boundary $\partial\mathscr{B}_t \cap \partial\Omega = \emptyset$. Moreover, the current domain $\mathscr{B}_t$ is the image of a reference domain $\mathscr{B} \subset \mathbb{R}^m$, $m = d, d-1$ through the map $\mathbf{X}$ defined as follows:

$$\mathbf{X} : \mathscr{B} \times [0, T] \to \mathscr{B}_t \quad \text{so that } \mathbf{x} = \mathbf{X}(\mathbf{s}, t)\ \forall \mathbf{x} \in \mathscr{B}_t. \tag{1}$$

The map $\mathbf{X}(\mathbf{s}, t)$ is assumed to be invertible at any time, meaning that the deformation gradient: $\mathbb{F}_{\alpha i} := \left(\nabla_s \mathbf{X}(\mathbf{s}, t)\right)_{\alpha i} = \mathbf{X}_{\alpha,i}(\mathbf{s}, t) = \frac{\partial \mathbf{X}_\alpha(\mathbf{s},t)}{\partial s_i}$ has rank m. The incompressibility condition forces $|\mathbb{F}| = 1$ during time evolution. Here $|\mathbb{F}|$ is the determinant of $\mathbb{F}$ when $m = d$. When $m = d - 1$ we set $|\mathbb{F}| = |\partial\mathbf{X}/\partial s|$ for $m = 1$ and $|\mathbb{F}| = |\partial\mathbf{X}/\partial s_1 \wedge \partial\mathbf{X}/\partial s_2|$ for $m = 2$.

Fluid and solid phases are characterized by piecewise constant densities: more precisely, $\rho = \rho_f$ in $\Omega \setminus \mathscr{B}_t$ and $\rho = \rho_s$ in $\mathscr{B}_t$. The interested reader can refer to [7] for a detailed study of stability criteria regarding ρ_s/ρ_f.

As mentioned before the key idea of IBM is in the definition of the Cauchy stress tensor $\boldsymbol{\sigma}$. For viscous Newtonian fluid we set: $\boldsymbol{\sigma}_f = -p\mathbb{I} + \mu(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)$. Then the Cauchy stress tensor is given by $\boldsymbol{\sigma} = \boldsymbol{\sigma}_f$ in $\Omega \setminus \mathscr{B}_t$ and $\boldsymbol{\sigma} = \boldsymbol{\sigma}_f + \boldsymbol{\sigma}_s$ in $\mathscr{B}_t$. Hence the fluid stress tensor lays over the whole domain, while the elastic tensor is associated to the structure position. This assumption is accepted in biological applications where viscoelasticity plays a key role (see, e.g., [15]).

In the following we shall use the first Piola–Kirchhoff stress tensor which can be derived from the elastic stress tensor $\boldsymbol{\sigma}_s$ using Lagrangian variables as: $\mathbb{P}(\mathbf{s}, t) = |\mathbb{F}(\mathbf{s}, t)|\, \boldsymbol{\sigma}_s(\mathbf{X}(\mathbf{s}, t), t)\, \mathbb{F}^{-T}(\mathbf{s}, t)$. Using the principle of virtual work and the above definitions we obtain the following formulation of the problem:

Problem 1 Given $\mathbf{u}_0 \in H^1_0(\Omega)^d$ and $\mathbf{X}_0 : \mathscr{B} \to \Omega$ such that $\mathbf{X}_0 \in W^{1,\infty}(\mathscr{B})$, for all $t \in]0, T[$, find $(\mathbf{u}(t), p(t)) \in H^1_0(\Omega)^d \times L^2_0(\Omega)$ and $\mathbf{X}(t) \in W^{1,\infty}(\mathscr{B})$, such that

$$\rho_f \frac{d}{dt}(\mathbf{u}(t), \mathbf{v}) + b(\mathbf{u}(t), \mathbf{u}(t), \mathbf{v}) + a(\mathbf{u}(t), \mathbf{v}) - (\nabla\cdot\mathbf{v}, p(t)) = \langle (t), \mathbf{v}\rangle + \langle \mathbf{F}(t), \mathbf{v}\rangle \qquad \forall \mathbf{v} \in H^1_0(\Omega)^d \tag{2a}$$

$$(\nabla\cdot\mathbf{u}(t), q) = 0 \qquad \forall q \in L^2_0(\Omega) \tag{2b}$$

$$\langle \mathbf{d}(t), \mathbf{v}\rangle = -(\rho_s - \rho_f) \int_{\mathscr{B}} \frac{\partial^2 \mathbf{X}}{\partial t^2} \mathbf{v}(\mathbf{X}(\mathbf{s}, t))\, d\mathbf{s} \qquad \forall \mathbf{v} \in H^1_0(\Omega)^d \tag{2c}$$

$$\langle \mathbf{F}(t), \mathbf{v}\rangle = -\int_{\mathscr{B}} \mathbb{P}(\mathbb{F}(\mathbf{s}, t)) : \nabla_s \mathbf{v}(\mathbf{X}(\mathbf{s}, t))\, d\mathbf{s} \qquad \forall \mathbf{v} \in H^1_0(\Omega)^d \tag{2d}$$

$$\frac{\partial \mathbf{X}}{\partial t}(\mathbf{s}, t) = \mathbf{u}(\mathbf{X}(\mathbf{s}, t), t) \qquad \forall \mathbf{s} \in \mathscr{B} \tag{2e}$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) \qquad \forall \mathbf{x} \in \Omega \tag{2f}$$

$$\mathbf{X}(\mathbf{s}, 0) = \mathbf{X}_0(\mathbf{s}) \qquad \forall \mathbf{s} \in \mathscr{B}. \tag{2g}$$

Here $a(\mathbf{u}, \mathbf{v}) = \mu(\nabla \mathbf{u}, \nabla \mathbf{v})$, $b(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \dfrac{\rho_f}{2}((\mathbf{u} \cdot \nabla \mathbf{v}, \mathbf{w}) - (\mathbf{u} \cdot \nabla \mathbf{w}, \mathbf{v}))$.

3 Distributed Lagrange Multiplier Formulation

In view of the finite element discretization of Problem 1, we write the kinematic coupling equation (2e) in weak form as follows:

$$\left\langle \boldsymbol{\mu}, \mathbf{u}(\mathbf{X}(\cdot, t), t) - \frac{\partial \mathbf{X}(\cdot, t)}{\partial t} \right\rangle = 0 \quad \forall \boldsymbol{\mu} \in (H^1(\mathscr{B})^d)^* \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H^1(\mathscr{B})^d$ and its dual space $(H^1(\mathscr{B})^d)^*$ if $m = d$ and between $H^{1/2}(\mathscr{B})^d$ and its dual space if $m = d - 1$. The notation $(\cdot, \cdot)_{\mathscr{B}}$ stands for the L^2-scalar product in $L^2(\mathscr{B})$.

Let $V_h \subseteq H^1_0(\Omega)^d$ and $Q_h \subseteq L^2_0(\Omega)$ be a pair of finite element spaces stable for the approximation of the Stokes equations. Let $S_h \subseteq H^1(\mathscr{B})^d$ contain piecewise linear vector valued functions and $\Lambda_h = S_h$. Then we introduce a Lagrange multiplier associated to the constraint (3), so that we obtain the following fully discrete Problem 1:

Problem 2 Given $\mathbf{u}_{0,h} \in V_h$ and $\mathbf{X}_{0,h} \in S_h$, for $n = 1, \ldots, N$ find

$$(\mathbf{u}^n_h, p^n_h) \in V_h \times Q_h, \quad \mathbf{X}^n_h \in S_h, \quad \boldsymbol{\lambda}^n_h \in \Lambda_h,$$

such that $\mathbf{u}_h^0 = \mathbf{u}_{0,h}$, $\mathbf{X}_h^0 = \mathbf{X}_{0,h}$ and

$$\rho_f \left(\frac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^n}{\Delta t}, \mathbf{v} \right) + b(\mathbf{u}_h^{n+1}(t), \mathbf{u}_h^{n+1}(t), \mathbf{v}) + a(\mathbf{u}_h^{n+1}, \mathbf{v})$$
$$- (\nabla \cdot \mathbf{v}, p_h^{n+1}) + \langle \boldsymbol{\lambda}_h^{n+1}, \mathbf{v}(\mathbf{X}_h^n) \rangle = 0 \qquad \forall \mathbf{v} \in V_h$$
$$(\nabla \cdot \mathbf{u}_h^{n+1}, q) = 0 \qquad \forall q \in Q_h$$
$$(\rho_s - \rho_f) \left(\frac{\mathbf{X}_h^{n+1} - 2\mathbf{X}_h^n + \mathbf{X}_h^{n-1}}{\Delta t^2}, \mathbf{Y} \right)_{\mathscr{B}} + (\mathbb{P}(\mathbb{F}_h^{n+1}), \nabla_s \mathbf{Y})_{\mathscr{B}}$$
$$- \langle \boldsymbol{\lambda}_h^{n+1}, \mathbf{Y} \rangle = 0 \qquad \forall \mathbf{Y} \in S_h$$
$$\left\langle \boldsymbol{\mu}, \mathbf{u}_h^{n+1}(\mathbf{X}_h^n) - \frac{\mathbf{X}_h^{n+1} - \mathbf{X}_h^n}{\Delta t} \right\rangle = 0 \qquad \forall \boldsymbol{\mu} \in \Lambda_h.$$

4 Numerical Experiments

In the subsequent experiments the Piola stress tensor is assumed to be proportional to the deformation gradient $\mathbb{P} = \kappa \mathbb{F}$, see [6] and references therein. The fluid convective term is neglected. The velocity and pressure spaces are the enhanced Bercovier–Pironneau elements, see [9]. In the latter we will call DLM the solution of Problem 2 and IBM the one obtained by applying the finite element discretization to Problem 1.

We first consider the scheme performances in terms of CFL stability. In Table 1 we present the results in terms of area conservation for the codimension one elastic string. We report the ratio between the initial and final internal area A/A_0. The area of the discrete region A is evaluated exactly, using the current structure triangulation The string initial position is an ellipse, with 1.4 dimensions ratio, which tends to a circular equilibrium position. IBM and DLM are affected by different nature instabilities. IBM is affected by a CFL condition extensively explored in [4, 7]. The DLM scheme needs to satisfy an inf-sup type condition that bounds the structure mesh to be sufficiently coarse with respect to the fluid one. DLM scheme is stable regardless the Δt choice.

In Fig. 1 we present a codimension zero simulation, where a rectangular structure tends to a square equilibrium position. The simulation parameters are: $h_x = 1/32$ (mesh size of the fluid domain), $h_s = 1/16$ (mesh size of the structure domain), $\kappa = 100$, $\Delta t = 10^{-3}$, the viscosity $\nu = 0.01$. It is clear that the parameters in this simulation would require an appropriate advection treatment to recover physical consistency. Since in the present paper we aim at showing the robustness of the scheme, this topic is left to further exploration. Figure 1a shows the first time step evaluated. The rectangle upper right corner generates an infinite vorticity singularity,

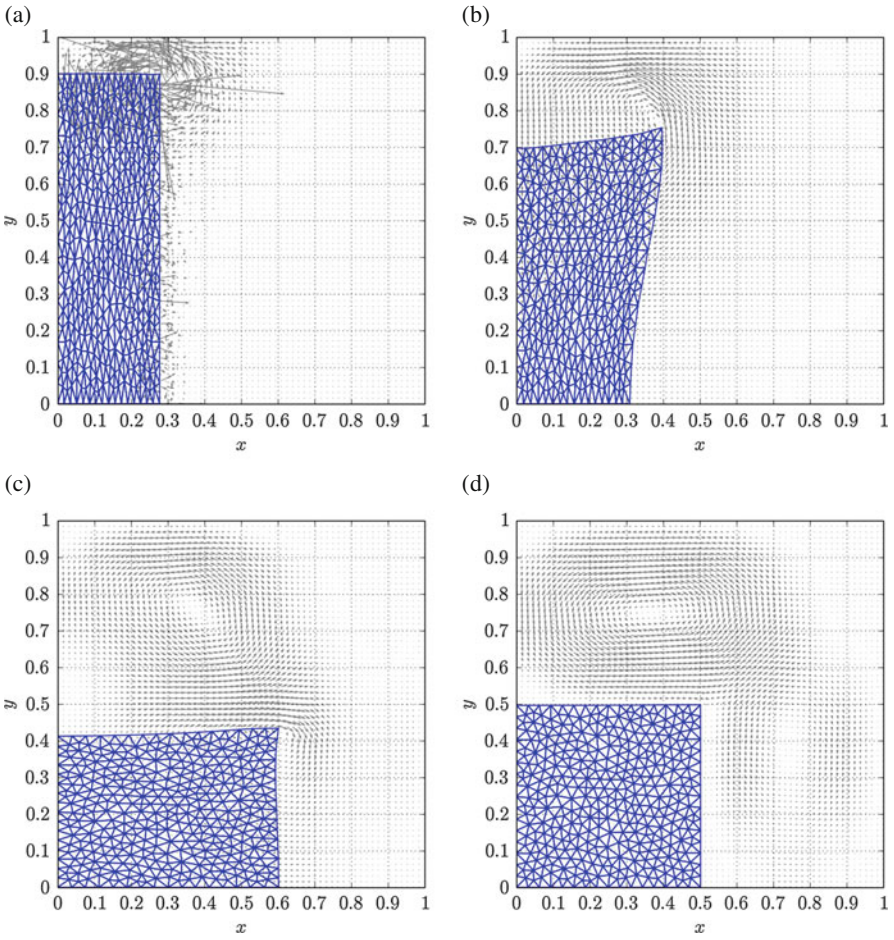


Fig. 1 Codimension zero structure. Oscillating square immersed into the fluid. (**a**) $t = 0.001$, (**b**) $t = 0.06$, (**c**) $t = 0.25$, (**d**) $t = 2$

that in this plot is represented by the greatest velocity values. In Fig. 1b the velocities are arranged to follow the structure motion. The structures oscillates and in Fig. 1c "rebounds" toward the equilibrium position. It is interesting to notice the inversion of the velocities. The structure finally gets to the equilibrium in Fig. 1d. The authors remark that due to space restrictions, the velocity detail are better represented by zooming in the electronic version of the paper.

5 Conclusions

In this paper we presented a Distributed Lagrange Multiplier FE-IBM scheme. The problem setting is improved to be fully variational; a rigorous analysis of the scheme is going to be objective of future works.

Moreover numerical tests show that there is no CFL restriction for the scheme. On the other hand an inf-sup condition forces the structure mesh to be coarser then the fluid one; a rigorous study of this phenomenon is in the authors' future plans.

The presented codimension zero simulation shows the robustness of the method. A rectangular structure introduces a singularity in the fluid vorticity. Nevertheless the scheme is stable and the solution converges to the equilibrium position.

References

1. Boffi, D., Gastaldi, L.: A finite element approach for the immersed boundary method. Comput. Struct. **81**(8–11), 491–501 (2003). In honour of Klaus-Jürgen Bathe
2. Boffi, D., Gastaldi, L., Heltai, L.: A finite element approach to the immersed boundary method. In: Topping, B.H.V., Mota Soares, C.A. (eds.) Progress in Engineering Computational Technology, pp. 271–298. Saxe-Coburg Publications, Stirling (2004)
3. Boffi, D., Gastaldi, L., Heltai, L.: Stability results and algorithmic strategies for the finite element approach to the immersed boundary method. In: Numerical Mathematics and Advanced Applications, pp. 575–582. Springer, Berlin (2006)
4. Boffi, D., Gastaldi, L., Heltai, L.: Numerical stability of the finite element immersed boundary method. Math. Models Methods Appl. Sci. **17**(10), 1479–1505 (2007)
5. Boffi, D., Gastaldi, L., Heltai, L.: On the CFL condition for the finite element immersed boundary method. Comput. Struct. **85**(11–14), 775–783 (2007)
6. Boffi, D., Gastaldi, L., Heltai, L., Peskin, C.S.: On the hyper-elastic formulation of the immersed boundary method. Comput. Methods Appl. Mech. Eng. **197**(25–28), 2210–2231 (2008)
7. Boffi, D., Cavallini, N., Gastaldi, L.: Finite element approach to immersed boundary method with different fluid and solid densities. Math. Models Methods Appl. Sci. **21**(12), 2523–2550 (2011)
8. Boffi, D., Cavallini, N., Gardini, F., Gastaldi, L.: Immersed boundary method: performance analysis of popular finite element spaces. In: Papadrakakis, M., Onate, E., Schrefler, B. (eds.) Coupled Problems 2011. Computational Methods for Coupled Problems in Science and Engineering IV, Cimne (2011)
9. Boffi, D., Cavallini, N., Gardini, F., Gastaldi, L.: Local mass conservation of Stokes finite elements. J. Sci. Comput. **52**(2), 383–400 (2012)
10. Boffi, D., Cavallini, N., Gardini, F., Gastaldi, L.: Stabilized stokes elements and local mass conservation. Bollettino Unione Mat. Ital. (9) **V**, 543–573 (2012)
11. Boffi, D., Cavallini, N., Gardini, F., Gastaldi, L.: Mass preserving distributed Lagrange multiplier approach to immersed boundary method. In: Idelsohn, M.P.S., Schrefler, B. (eds.) Coupled Problems 2013. Computational Methods for Coupled Problems in Science and Engineering V. Cimne (2013)

12. Girault, V., Glowinski, R., Pan, T.W.: A fictitious-domain method with distributed multiplier for the Stokes problem. In: Applied Nonlinear Analysis, pp. 159–174. Kluwer/Plenum, New York (1999)
13. Heltai, L.: On the stability of the finite element immersed boundary method. Comput. Struct. **86**(7–8), 598–617 (2008)
14. Peskin, C.S.: The immersed boundary method. In: Acta Numerica 2002. Cambridge University Press, Cambridge (2002)
15. Quarteroni, A., Tuveri, M., Veneziani, A.: Computational vascular fluid dynamics: problems, models and methods. Comput. Vis. Sci. **2**, 163–197 (2000)

Impact of Blood Flow on Ocular Pathologies: Can Mathematical and Numerical Modeling Help Preventing Blindness?

Paola Causin, Giovanna Guidoboni, Francesca Malgaroli, Riccardo Sacco, and Alon Harris

Abstract The pathogenesis of many blinding diseases, such as diabetic retinopathy, glaucoma or retinopathy of prematurity, is thought to be related to retinal tissue hypoxia. Yet, the mechanisms governing oxygen delivery to the retina are still poorly understood. Since it is not currently possible to disentangle the influence of all the concurring factors in retinal oxygenation during experimental and clinical measurements, mathematical models can serve as virtual laboratories to separately investigate the individual influence of different parameters. In this contribution, we propose a mathematical model which describes the oxygen profile along the whole retinal depth, including sources from blood circulation and tissue metabolic consumption. An analytical solution for the profile is computed and quantitative estimates of the sensitivity of retinal oxygen profiles to changes in geometrical and metabolic parameters of the retinal tissue are provided. In particular, this analysis highlights the important role played by the thickness of the different retinal layers and warns of potential issues when using experimental data across species.

Keywords Biomedical science • Modeling of eye retina

P. Causin (✉)
Department of Mathematics, University of Milano, Milano, Italy
e-mail: paola.causin@unimi.it

G. Guidoboni
Department of Mathematical Sciences, Indiana University, Indianapolis, IN, USA

Eugene and Marilyn Glick Eye Institute, Indiana University, Indianapolis, IN, USA

Institut de Recherche Mathématique Avancée, University of Strasbourg, Strasbourg, France

F. Malgaroli • R. Sacco
Department of Mathematics, Politecnico of Milano, Milano, Italy

A. Harris
Eugene and Marilyn Glick Eye Institute, Indiana University, Indianapolis, IN, USA

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_41

1 Introduction

Eye diseases represent a significant medical and social concern in countries undergoing population aging. One of the most frequent causes of vision loss is represented by damage to the retina, the light-sensitive tissue lining the inner surface of the eye. Alterations of blood circulation and oxygen (O_2) delivery to the retina have been identified as important factors in many retinopathies, but the pathogenic mechanisms leading to vision loss are still poorly understood [6, 20]. Recently, several models have been proposed by some of the authors of this paper to investigate the biomechanics/hemodynamics/oxygenation relationship in the retinal tissue, see [1, 3, 4, 7, 8]. However, several questions remain open. One important question concerns the relative contribution of different O_2 sources to the oxygenation of the retinal tissue. In the avascular outer region of the retina, also referred to as outer retina, O_2 is mainly provided via diffusion from the choroid, whereas in the inner portion of the retina, also referred to as inner retina, O_2 is mainly provided via transport from a network of embedded blood vessels. Oxygen profiles have been recorded in vivo in cats, rats and monkeys via microelectrode measurements (see [19] and references therein) and the profiles have been successfully fitted in the outer retina region via 1D diffusion-reaction mathematical models [5, 12, 13]. In these models, an inner retina portion is sometimes included, without accounting for blood sources, in order to check that its presence does not disrupt the computed solution in the outer portion. The inclusion of blood sources in the inner retinal layers is proposed in [15]. Fixed prescribed blood flow and arterial O_2 values from arterial microcirculation are used to model sources in the diffusion-reaction PDE system describing the O_2 profile. In [15], a numerical solution of the problem is pursued and the possible effects of retinal detachment are analyzed. The purpose of the present study is to develop an improved and more physiologically realistic, albeit still analytically solvable, model of the O_2 distribution in the retina, which combines O_2 sources from the choroid and from the retinal blood circulation and to carry out a sensitivity analysis to assess the relative importance of various physiologically-relevant factors.

2 Eye Retina Anatomy

In this section, we provide a short description of the retina anatomy useful for the following discussion. The retina is a complex matrix of neural cells lining the innermost surface of the eye globe; light travels through the thickness of the retina before striking and activating the photoreceptors (rods and cones). Histologically, the retinal tissue complex is composed of multiple layers that subserve its visual function. From the modeling viewpoint, it is useful to divide the retinal thickness into two different regions (see Fig. 1): (1) the *outer retina* (OR), proximal to the sclera, which is an avascular region that includes the retinal pigmented epithelium,

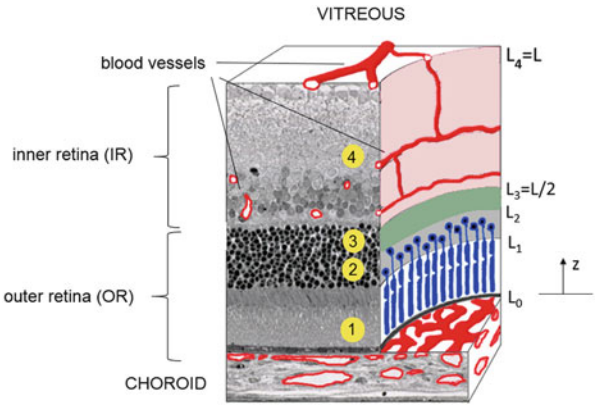


Fig. 1 Schematic representation of a portion of eye retina across its thickness. From the modeling viewpoint, it is useful to distinguish two main regions: the *outer retina* (including layer 1 to 3), which is avascular and has high oxygen consumption due to the presence of photoreceptor mitochondria (layer 2), and the *inner retina* (layer 4), which is vascularized and has an average O_2 consumption rate that is approximately one-fifth of that of the outer region. Drawing modified from [17]

the outer and inner segments of the photoreceptors layer and the outer nuclear layer. The outer part of the retina is mainly nourished by diffusion from the choroid; and (2) the *inner retina* (IR), proximal to the vitreous, which is a vascularized region that goes from the outer plexiform layer up to the nerve fiber/ganglion cell layer.

Oxygen consumption is not uniformly distributed across the retinal thickness. Nearly 100 % of the O_2 consumption of the outer retina takes place in the inner segment of the photoreceptors layer, which contains most of the photoreceptors mitochondria. The inner retina, which shares many similarities with other parts of the central nervous system, has an average O_2 consumption rate that is approximately one-fifth of that of the outer region [14].

3 Mathematical Model of O_2 Transport in the Retina

Let us consider a Cartesian coordinate system with the z axis oriented from the choroid to the vitreous across the retinal thickness (see Fig. 1). General agreement in the literature exists on the adequacy of a three-layer model to describe the 1D oxygen dynamics in the OR [2, 12, 13]. Layer 1 extends from the choroid, $z = 0 = L_0$, to $z = L_1$, layer 2 extends from $z = L_1$ to $z = L_2$ and layer 3 extends from $z = L_2$ to the interface with the IR, at $z = L_3$. The thickness L_3 of the OR is assumed to be half of the retinal thickness L (see [13] and references therein for a similar choice). The IR consists in a single layer, layer 4, which extends from $z = L_3$ to $z = L_4 = L$. The O_2 profile in the retinal tissue is studied under the assumption

that the relevant phenomena are diffusion, metabolic consumption by neural cells and O_2 delivery from blood vessels, so that, in each region $z \in (L_{j-1}, L_j) = \Omega_j$, $j = 1, \ldots, 4$, the general diffusion–reaction equation holds

$$-Dk\frac{\partial^2 p_j}{\partial z^2} = f_j - Q_j, \tag{1}$$

where p_j is the restriction of the O_2 tension $p : [0, L] \to \mathbb{R}^+$ to layer j, D and k are constant oxygen diffusivity and solubility, respectively, f_j is the O_2 source term from blood vessels and Q_j the constant O_2 metabolic rate. The choice of constant metabolic rates is a fairly simplified assumption. A more detailed analysis which considers the more complex Michaelis–Menten kinetics for the metabolic rates will be the object of a forthcoming paper.

3.1 Solution for the O_2 Profile in the Retinal Tissue

Equation (1) holds with $f_j = 0$ for $j = 1, 2, 3$, due to the absence of oxygen sources from embedded blood vessels. Moreover, based on the above considerations about the metabolic activity distribution in the OR layers , we consider $Q_1 = Q_3 = 0$ and we confine the whole metabolic consumption in the OR to layer 2, where photoreceptor mitochondria are located (see Fig. 1). In the IR, Eq. (1) holds with a nonzero f source, that we assume to be a prescribed constant distributed term. The following boundary and interface conditions are enforced:

$$\begin{cases} p_1 = p_{ch}, & & \text{at} \quad z = 0, \\ p_1 = p_2, & \dfrac{\partial p_1}{\partial z} = \dfrac{\partial p_2}{\partial z}, & \text{at} \quad z = L_1, \\ p_2 = p_3, & \dfrac{\partial p_2}{\partial z} = \dfrac{\partial p_3}{\partial z}, & \text{at} \quad z = L_2, \\ p_3 = p_4, & \dfrac{\partial p_3}{\partial z} = \dfrac{\partial p_4}{\partial z} & \text{at} \quad z = L_3, \\ \dfrac{\partial p_4}{\partial z} = 0 & & \text{at} \quad z = L, \end{cases} \tag{2}$$

where, at the interface with the vitreous, a homogeneous Neumann boundary condition has been assumed, according to what done in [16] and to the observations of [19]. The piecewise polynomial solution, belonging to $\mathscr{C}^1(\overline{\Omega})$, for the oxygen

tension in the retina is then given by:

$$\begin{cases} p_1(z) = \alpha_1 z + \beta_1, & z \in (0, L_1), \\ p_2(z) = \dfrac{Q_2}{2Dk} z^2 + \alpha_2 z + \beta_2, & z \in (L_1, L_2), \\ p_3(z) = \alpha_3 z + \beta_3, & z \in (L_2, L_3), \\ p_4(z) = \dfrac{Q_4}{2Dk} z^2 + \alpha_4 z + \beta_4, & z \in (L_3, L), \end{cases} \tag{3}$$

where:

$$\begin{cases} \alpha_1 = \dfrac{Q_2}{Dk}(L_1 - L_2) - L\dfrac{Q_4}{2Dk}, \quad \beta_1 = p_{ch}, \\ \alpha_2 = -L_2\dfrac{Q_2}{Dk} - L\dfrac{Q_4}{2Dk}, \quad \beta_2 = p_{ch} + \dfrac{Q_2}{2Dk}L_1^2, \\ \alpha_3 = -L\dfrac{Q_4}{2Dk}, \quad \beta_3 = p_{ch} + \dfrac{Q_2}{2Dk}(L_1^2 - L_2^2), \\ \alpha_4 = -L\dfrac{Q_4}{Dk}, \quad \beta_4 = p_{ch} + \dfrac{Q_2}{2Dk}(L_1^2 - L_2^2) + \dfrac{Q_4}{8Dk}L^2. \end{cases} \tag{4}$$

The values of the numerical parameters for the baseline condition are chosen according to the fitting procedure carried out in [12] in the case of the cat retina and are specified in Table 1. The value of the source term f_4 in this condition is selected in such a way that the oxygen tension at the OR/IR interface, corresponding to the value $p_3(L/2) = p_4(L/2)$, is equal to 24 mmHg, according to the experimental datum found in [12]. The O_2 profile corresponding to Eq. (3) is represented in Fig. 2 as a function of the retinal coordinate $z \in [L_0 = 0, L]$. Oxygen distribution decreases

Table 1 Parameters used in the numerical simulations resulting from the fitting procedure carried out in [12]

Name	Definition	Value	Units
D	Oxygen diffusion coefficient in tissue	10^{-5}	cm^2/s
f_4	Source from retinal circulation	6.15×10^{-3}	ml_{O_2} / (ml_{tissue}· s)
$K_{0.5}$	Half saturation constant	2	mmHg
k	Oxygen solubility coefficient in tissue	2×10^{-5}	ml_{O_2} / (ml_{tissue} · mmHg)
L	Total thickness of the retina	234	μm
L_1	Abscissa of the first interface in OR	35	μm
L_2	Abscissa of the second interface in OR	50	μm
L_3	Abscissa of the interface OR/IR	117	μm
p_{ch}	Oxygen tension at choroid	80	mmHg
Q_2	Metabolic rate in layer 2	4.5×10^6	ml_{O_2} / (ml_{tissue}· s)
Q_4	Metabolic rate in layer 4	1×10^6	ml_{O_2} / (ml_{tissue}· s)

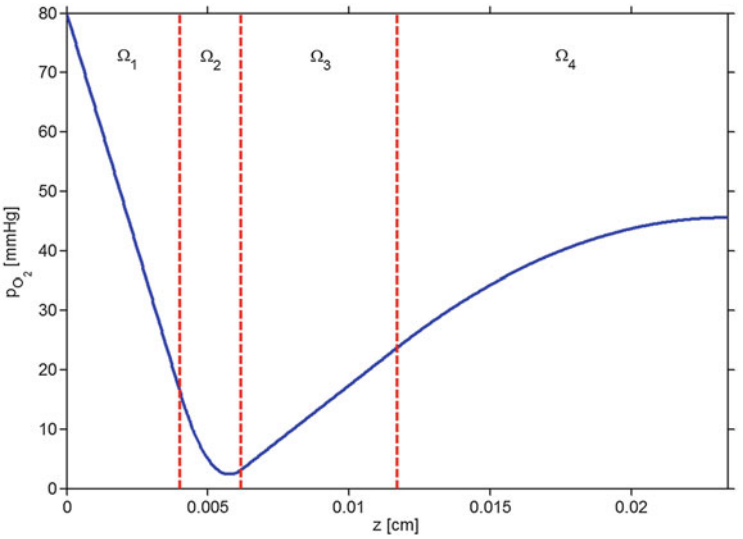


Fig. 2 Oxygen tension profile along the retinal thickness as computed from (3). The choroid is located at the *left side* ($z = 0$), while the vitreous is located at the *right side* ($z = L$). The *dashed vertical lines* indicate the location of the interfaces between retinal layers

steeply moving away from the choroid towards layer Ω_2 because of the heavy O_2 consumption by the mitochondria in the photoreceptor layer. On the other hand, O_2 tension returns to appreciably high values in the IR because of a mix between external supply and moderate consumption (much lower than in the OR). An almost flat profile in the neighborhood of the vitreous is the result of the homogeneous Neumann boundary condition at $z = L_4$, whereas an increasing slope of oxygen tension is visible in the layer of the OR proximal to the IR because of the need of restoring continuity of oxygen tension and flux at $z = L_3$.

4 Sensitivity Analysis

The analytical solution obtained in Sect. 3 is used here to explore the sensitivity of the O_2 profile in the retina with respect to changes in the thickness of layers 1 and 2, the consumption rates of layers 2 and 4, and the intensity of the O_2 blood source. More precisely, we define the set of model parameters $\mathscr{P} = \{\mathscr{P}_1, \mathscr{P}_2, \mathscr{P}_3, \mathscr{P}_4, \mathscr{P}_5\}$, with $\mathscr{P}_1 = L_1$, $\mathscr{P}_2 = \delta := L_2 - L_1$, $\mathscr{P}_3 = Q_2$, $\mathscr{P}_4 = Q_4$ and $\mathscr{P}_5 = f$, and we denote by $\mathscr{P}^* = \{\mathscr{P}_1^*, \mathscr{P}_2^*, \mathscr{P}_3^*, \mathscr{P}_4^*, \mathscr{P}_5^*\}$ the set of the corresponding baseline values as in Table 1. We define the sensitivity index $S_p^k(z)$ with respect to the model parameter $\mathscr{P}_k$ as

$$S_p^k(z) := \frac{\partial p(z;\mathscr{P})}{\partial \mathscr{P}_k}\bigg|_{\mathscr{P}=\mathscr{P}^*} \qquad k = 1, \ldots, 5.$$

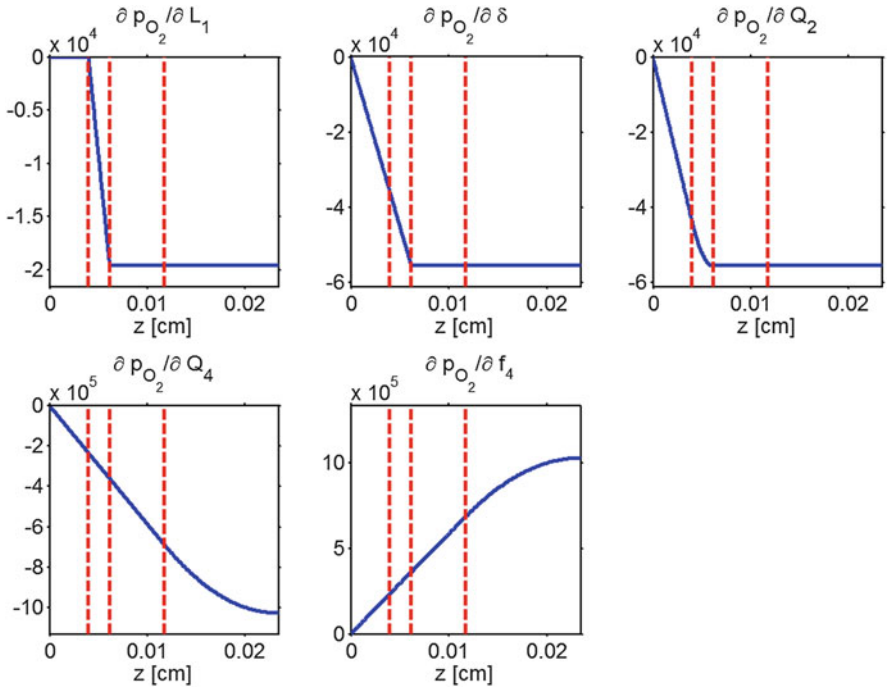


Fig. 3 Sensitivity indices of the solution as a function of the coordinate z with respect to different biophysical parameters evaluated according to the baseline values of Table 1. The *dashed vertical lines* indicate the location of the interfaces between retinal layers

Figure 3 reports the indices as a function of the coordinate z along the retinal thickness.

5 Discussion

From a clinical viewpoint, understanding how changes in parameters influence retinal O_2 is crucial to better understand the pathogenesis of various ocular diseases, including glaucoma [18, 20], diabetic retinopathy [9] and retinal artery and/or vein occlusions [11]. Currently, O_2 profiles in the retinal tissue can be measured experimentally by implanting oxygen-sensitive microelectrodes in the retina [5, 13]. However, this technique is extremely invasive and has been performed only on animals. On human subjects, O_2 levels are measured only within the larger retinal vessels via the non invasive retinal oximetry techniques [10], but these measurements cannot be directly associated with O_2 profiles across the retinal tissue layers, due to the many factors influencing such association, including thickness of retinal layers and O_2 blood sources. Moreover, it is not currently possible to disentangle

the influence of these factors during experimental and clinical measurements. The sensitivity analysis presented here and summarized in Fig. 3 constitutes a first attempt to theoretically weight out the relative contributions of various factors contributing to the O_2 profile in the retinal tissue.

The sensitivity indices experience noticeable variations with z. The O_2 levels in the OR are very sensitive to all parameters in $\mathscr{P}$, whereas the O_2 in the IR is most sensitive to the consumption rate Q_4, fairly corresponding to the consumption of the retinal ganglion cells, and to the O_2 source from blood, f_4. In addition, it is important to notice that the geometrical parameters L_1 and δ, relative to the outer retinal layers 1 and 2, strongly influence the O_2 profile in the OR. These dimensions may vary significantly among different species, and therefore this result warns of potential issues when using experimental data across species. Also worth of notice is the fact that the retinal O_2 profile is extremely sensitive to both changes in the consumption rates of the photoreceptors in the OR, Q_2, and of the retinal ganglion cells in the IR, Q_4, while it is equally sensitive to changes in the consumption rates of the photoreceptors in the OR, Q_2, and of the retinal ganglion cells in the IR, Q_4.

In conclusion, despite its numerous simplifications, the model presented in this paper has provided quantitative estimates of the sensitivity of retinal oxygen profiles to changes in geometrical and metabolic parameters of the retinal tissue. These results might help deepening the current understanding of the pathophysiology of the retina and designing novel, more effective therapeutic approaches.

References

1. Arciero, J., Harris, A., Siesky, B., Amireskandari, A., Gershuny, V., Pickrell, A., Guidoboni, G.: Theoretical analysis of vascular regulatory mechanisms contributing to retinal blood flow autoregulation. Invest. Ophthalmol. Vis. Sci. **54**, 5584–5593 (2013)
2. Birol, G., Wang, S., Budzynski, E., Wangsa-Wirawan, N.D., Linsenmeier, R.A.: Oxygen distribution and consumption in the macaque retina. Am. J. Physiol. Heart. Circ. Physiol. **293**(3), H1696–H1704 (2007)
3. Causin, P., Guidoboni, G., Harris, A., Prada, D., Sacco, R., Terragni, S.: A poroelastic model for the perfusion of the lamina cribrosa in the optic nerve head. Math. Biosci. **257**, 33–41 (2014). Multiscale models and methods in biomedicine
4. Causin, P., Malgaroli, F., Guidoboni, G., Sacco, R., Harris, A.: Blood flow mechanics and oxygen transport and delivery in retinal microcirculation: multiscale mathematical modeling and numerical simulation. Biomech. Model Mechanobiol. **15**(3), 525–542 (2016)
5. Cringle, S.J., Yu, D.Y.: A multi-layer model of retinal oxygen supply and consumption helps explain the mutated rise in inner retinal P_{O_2} during systemic hyperoxia. Comput. Biochem. Physiol. A: Mol. Integr. Physiol. **132**(1), 61–66 (2002)
6. Grieshaber, M., Flammer, J.: Blood flow in glaucoma. Am. J. Ophthalmol. **16**(2), 79–83 (2005)
7. Guidoboni, G., Harris, A., Carichino, L., Arieli, Y., Siesky, B.A.: Effect of intraocular pressure on the hemodynamics of the central retinal artery: a mathematical model. Math. Biosci. Eng. **11**(3), 523–546 (2014)
8. Guidoboni, G., Harris, A., Cassani, S., Arciero, J., Siesky, B., Amireskandari, A., Tobe, L., Egan, P., Januveliciene, I., Park, J.: Intraocular pressure, blood pressure and retinal blood flow autoregulation: a mathematical model to clarify their relationship and clinical relevance. Invest. Ophthalmol. Vis. Sci. **7**(55), 4105–4118 (2014)

9. Hardarson, S., Stefánsson, E.: Retinal oxygen saturation is altered in diabetic retinopathy. Br. J. Ophthalmol. **96**, 560–563 (2012)
10. Hardarson, S.H., Basit, S., Jonsdottir, T.E., Eysteinsson, T., Halldorsson, G.H., Karlsson, R.A., Beach, J.M., Benediktsson, J.A., Stefansson, E.: Oxygen saturation in human retinal vessels is higher in dark than in light. Invest. Ophthalmol. Vis. Sci. **50**(5), 2308–2311 (2009)
11. Hardarson, S.H., Elfarsson, A., Agnarsson, B.A., Stefánsson, E.: Retinal oximetry in central retinal artery occlusion. Acta ophthalmol. **91**(2), 189–190 (2013)
12. Haugh, L., Linsenmeier, R., Goldstick, T.K.: Mathematical models of the spatial distribution of retinal oxygen tension and consumption, including changes upon illumination. Ann. Biomed. Eng. (1), 19–36 (1989)
13. Lau, J., Linsenmeier, R.A.: Oxygen consumption and distribution in the Long-Evans rat retina. Exp. Eye Res. **102**, 50–58 (2012)
14. London, A., Benhar, I., Schwartz, M.: The retina as a window to the brain-from eye research to CNS disorders. Nat. Rev. Neurol. **9**(1), 44–53 (2013)
15. Roos, W.M.: Theoretical estimation of retinal oxygenation during retinal artery occlusion. Physiol. Meas. **25**(6), 1523–1532 (2004)
16. Roos, M.: Theoretical estimation of retinal oxygenation during retinal detachment. Comput. Biol. Med. **37**(6), 890–896 (2007)
17. Sim, D., Fruttiger, M.: Ophthalmology: keeping blood vessels out of sight. eLife **2**, e00948 (2013)
18. Vandewalle, E., Abegão Pinto, L., Olafsdottir, O.B., De Clerck, E., Stalmans, P., Van Calster, J., Zeyen, T., Stefánsson, E., Stalmans, I.: Oximetry in glaucoma: correlation of metabolic change with structural and functional damage. Acta ophthalmol. **92**(2), 105–110 (2014)
19. Wangsa-Wirawan, N.D., Linsenmeier, R.A.: Retinal oxygen: fundamental and clinical aspects. Arch. Ophthalmol. **121**(4), 547–557 (2003)
20. Weinreb, R., Harris, A. (eds.): Ocular Blood Flow in Glaucoma. Kugler Publications, Amsterdam (2009)

Spectral Deferred Correction Methods for Adaptive Electro-Mechanical Coupling in Cardiac Simulation

Martin Weiser and Simone Scacchi

Abstract We investigate spectral deferred correction (SDC) methods for time stepping and their interplay with spatio-temporal adaptivity, applied to the solution of the cardiac electro-mechanical coupling model. This model consists of the Monodomain equations, a reaction-diffusion system modeling the cardiac bioelectrical activity, coupled with a quasi-static mechanical model describing the contraction and relaxation of the cardiac muscle. The numerical approximation of the cardiac electro-mechanical coupling is a challenging multiphysics problem, because it exhibits very different spatial and temporal scales. Therefore, spatio-temporal adaptivity is a promising approach to reduce the computational complexity. SDC methods are simple iterative methods for solving collocation systems. We exploit their flexibility for combining them in various ways with spatio-temporal adaptivity. The accuracy and computational complexity of the resulting methods are studied on some numerical examples.

Keywords Biomedical science • Cardiac electro-mechanical coupling • Cardiovascular system • Multiphysics problem • Spatio-temporal adaptivity

1 Introduction

The spread of the electrical impulse in the cardiac muscle and the subsequent contraction-relaxation process is quantitatively described by a mathematical model called electro-mechanical coupling. The electrical model consists of the Monodomain system (a reduction of the Bidomain model), which is a reaction-diffusion equation describing the evolution of the transmembrane voltage. The PDE is

M. Weiser (✉)
Zuse Institute Berlin, Bereich Numerische Mathematik, Taku strasse 7, 14195 Berlin-Dahlem, Germany
e-mail: weiser@zib.de

S. Scacchi
Department of Mathematics, University of Milan, via Saldini 50, 20133 Milano, Italy
e-mail: simone.scaccchi@unimi.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_42

coupled through the reaction term with a stiff system of ordinary differential equations (ODEs), the so-called *membrane model*, describing the flow of the ionic currents through the cellular membrane. The mechanical model consists of quasi-static finite elasticity, coupled with a system of ODEs modeling the development of biochemically generated active stress.

The numerical approximation of the cardiac electro-mechanical coupling is a challenging multiphysics problem, because the space and time scales associated with the electrical and mechanical models are very different. Therefore, spatial and temporal adaptivity is a promising approach to reduce the computational complexity [2, 3]. However, spatial adaptivity by local mesh refinement incurs a substantial overhead for error estimation, grid manipulation, repeated integration until spatial accuracy is achieved, and reassembly of mass and stiffness matrices, which reduces the performance gain.

In this work, we investigate the use of spectral deferred correction (SDC) methods for time stepping and their interplay with spatial and temporal adaptivity. SDC methods are simple iterative methods for solving collocation systems. Their flexibility allows to combine them in various ways with spatio-temporal adaptivity. We explore interleaving mesh refinement with SDC iterations for improved convergence and local time stepping. In particular, we develop SDC methods for strong electro-mechanical coupling including mechano-electrical feedback and their potential for multi-rate integration. The properties of the resulting methods in terms of accuracy and computational complexity are discussed at a simple numerical example.

2 Mathematical Models

2.1 Mechanical Deformation

Let us denote the region occupied by the undeformed myocardium by Ω. For now we consider a simple two-dimensional square domain. The myocard undergoes a time-dependent deformation with displacement $u : \Omega \times (0,T) \to \mathbb{R}^2$, such that point $x \in \Omega$ is moved to $x + u(x,t)$ at time $t \in (0,T)$. As usual, $\mathbf{F} = \mathbf{I} + u_x$ denotes the deformation derivative, $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ the Cauchy-Green deformation tensor, and $\mathbf{E} = \frac{1}{2}(\mathbf{C} - \mathbf{I})$ the Green-Lagrange strain tensor, with identity matrix $\mathbf{I}$.

The cardiac tissue is modeled as a transversely isotropic nonlinear hyperelastic material with exponential strain energy function

$$W_{\text{pas}}(\mathbf{E}) = c_1 \exp(b_1 E_{11}^2 + 4b_2 E_{22}^2 + 4b_3 E_{12}^2).$$

introduced in [11], where the muscle fiber direction is just x_1. The near-incompressibility is modeled by an additive volume change penalization term

$$W_{\text{com}}(\det \mathbf{F}) = c_2((\det \mathbf{F})^2 + (\det \mathbf{F})^{-2} - 2),$$

which ensures orientation preservation. The contraction of the ventricles results from the active tension T_a generated by the myofilaments, which are activated by calcium release. We assume that the generated active stress acts only in the direction of the fibers [6, 10, 12]. This leads to a third term in the variational functional:

$$W_{\text{act}}(\mathbf{E}, T_a) = T_a E_{11}.$$

The biochemically generated active stress T_a is modeled as *stretch and stretch-rate independent*. Thus, we assume as in [5, 9] that the dynamics of T_a depends only on the transmembrane voltage v according to a simple twitch-like rule,

$$\frac{\partial T_a}{\partial t} = \epsilon(v)(k_{T_a}(v - v_r) - T_a), \tag{1}$$

where $k_{T_a} > 0$ controls the saturated value of T_a for a given voltage v and a given resting voltage v_r, see [5, 9] for details.

We assume that the time-dependent inertial term in the governing elastic wave equation may be neglected, see, e.g., [7, 12]. At any point in time, the myocard then assumes the stationary minimizer of the internal energy, subject to essential boundary conditions on the Dirichlet part of $\partial\Omega$:

$$\min_{u(t)\in H^1(\Omega)^2} \int_\Omega W_{\text{pas}}(\mathbf{E}) + W_{\text{com}}(\det \mathbf{F}) + W_{\text{act}}(\mathbf{E}, T_a(t))\, dx \quad \text{s.t.} \quad u(t)|_{\partial\Omega_D} = 0. \tag{2}$$

2.2 Electrical Excitation

The electrical excitation is described by the monodomain model using the Aliev-Panfilov membrane model [1] on the reference cardiac domain Ω [9, 10, 12]. Given an applied current per unit volume $I_{\text{app}} : \Omega \times (0, T) \to \mathbb{R}$, and initial conditions $v_0 : \Omega \to \mathbb{R}$, $w_0 : \Omega \to \mathbb{R}$, find the transmembrane potential $v : \Omega \times (0, T) \to \mathbb{R}$ and the gating variable $w : \Omega \times (0, T) \to \mathbb{R}$ such that

$$c_m \frac{\partial v}{\partial t} - \operatorname{div}(\mathbf{F}^{-1} D_m \mathbf{F}^{-T} \nabla v) + I_{\text{ion}}(v, w) = I_{\text{app}} \qquad \text{in } \Omega \times (0, T), \tag{3}$$

$$\frac{\partial w}{\partial t} = R(v, w) \qquad \text{in } \Omega \times (0, T), \tag{4}$$

holds. Note that the length changes due to tissue deformation change the diffusion tensor from D_m to $F^{-1} D_m F^{-T}$, neglecting the impact of volume changes. The

functions

$$I_{\text{ion}}(v,w) = -g_a v(v-a)(v-1) - vw$$
$$R(v,w) = -\frac{1}{4}\left(\epsilon_1 + \frac{\mu_1 w}{v+\mu_2}\right)(w + g_s v(v-a-1))$$

are given by the Aliev-Panfilov membrane model [1]. Insulating boundary conditions on v are prescribed.

3 Numerical Methods

3.1 Spatial Discretization: Finite Elements

A pure displacement discretization with P1-elements is used for computing the tissue deformation in reaction to active stress T_a. A Newton-like method is employed for minimizing (2). As hyperelastic energies can be nonconvex, the elemental matrices are modified during assembly to be positive definite. This ensures that the computed Newton step is a descent direction. Line search is applied to ensure monotone decrease of the elastic energy.

The transmembrane voltage is less smooth than the displacement, but easier to solve for. Thus, a finer spatial discretization is used. For implementation simplicity, we use P3-elements on the same mesh for transmembrane voltage, gating variables, and active stress generation. The transfer between different spatial discretizations is done by interpolation at quadrature nodes. Mesh refinement is based on an embedded energy error estimator for the transmembrane voltage, as this is the variable with dominating local dynamics.

3.2 Time Discretization: Spectral Deferred Correction Methods

Spectral deferred correction methods [4] are simple iterative methods for solving ODE collocation systems, where each iteration consists of a sequence of time steps with a low order scheme, most often an Euler scheme. For simplicity of notation, we consider an initial value problem $\dot{u} = f(u)$ with initial value $u(0) = u_0$ and exact solution u^*. On a time step $[0,\tau]$ we define a collocation time subgrid $0 = \tau_0 < \cdots < \tau_n = \tau$ and a polynomial approximate solution $u^0 \in \mathbb{P}_n$ with values $u_i^k = u^k(\tau_i)$ at the collocation points τ_i. The defect $u^* - u^k$ satisfies the Picard equation

$$\frac{d}{dt}(u^* - u^k)(t) = \int_{s=0}^{t} (f(u^*) - \dot{u}^k)\, ds. \tag{5}$$

Linearizing f around u^k, integrating the implicit term in (5) approximately with the right-looking rectangular rule and the explicit terms by a quadrature rule on the collocation time grid gives approximate defect values

$$\delta u_{i+1}^k = \delta u_i^k + (\tau_{i+1} - \tau_i) \left(\sum_{j=0}^{n} S_j f(u_j^k) + f'(u_{i+1}^k) \delta u_{i+1}^k \right) - (u_{i+1}^k - u_i^k) \tag{6}$$

at the collocation nodes, which in turn define a polynomial defect approximation δu^k by interpolation. Note that (6) is a linearly implicit Euler scheme on the collocation time grid. Updating the approximation by $u^{k+1} = u^k + \delta u^k$ yields an iteration the fixed point of which satisfies the collocation condition $f(u_i) = \dot{u}_i$. In lack of better initialization, the starting iterate is the constant initial value: $u_i^0 = u_0$.

3.3 Interleaved SDC and Mesh Refinement

Popular diagonally linearly implicit Runge-Kutta schemes, such as Rosenbrock methods, can be combined with spatial adaptivity in two different ways. Error estimation and refinement can be performed either for the final result, or for the very first stage (essentially a linearly implicit Euler step) only. The first option is more conservative, but requires the recomputation of all stages from scratch, since order and accuracy of Rosenbrock schemes deteriorate when the stages are computed on different spatial grids. The second option is more efficient, as only the first stage is recomputed on mesh refinement, but assumes a sufficient similarity of Euler step and final Rosenbrock step to produce suitable meshes for the latter. As demonstrated in Sect. 4, this assumption can be quite wrong.

In contrast to Rosenbrock methods, SDC methods compute an independent correction in every sweep, wherever the approximation error originates, may it be the SDC iteration error or a spatial discretization error. Hence, spatial mesh refinement can be performed in between any SDC sweeps, creating meshes adapted to the final SDC step, and nevertheless the previously computed values can be reused.

Applied to the electromechanical model described in Sect. 2 above, the SDC iterations are performed for the transmembrane voltage (3), the gating variables (4), and the active stress generation in turn. After each sweep, the elastic displacement is updated at all collocation points by a simplified Newton method, followed by error estimation both for the spatial discretization error and the SDC iteration error. If the spatial error exceeds the iteration error, adaptive mesh refinement is performed.

3.4 Multi-rate Integration

As the dynamics in the active stress generation and hence the mechanical displacement is slower than in the transmembrane voltage, a coarser time discretization of the displacement can be used. We exploit the continuous in time representation of approximate solutions by polynomial interpolation, using a finer collocation grid for the transmembrane voltage than for the displacement. Additionally, as after an SDC sweep the electrical state is still only an approximation, an exact solution of the nonlinear mechanic model is not required. The number of Newton steps can therefore be reduced. Finally, less than one Newton step per sweep effort can be achieved by solving for the elasticity part just every other SDC sweep. The induced inaccuracy in the displacement will have an impact on the convergence of the transmembrane voltage due to the mechano-electrical feedback.

4 Numerical Results

We study the effect of the algorithmic variants in detail at a particularly simple example, the spread of an excitation wave in the 2D domain $\hat{\Omega} =]0, 2[^2$ with an excitation current in $[0.5, 0.55]^2$ for 1 ms. For simplicity, the time step size is fixed to 1.5 ms on a Radau(4) collocation time grid, using cubic finite elements for the transmembrane voltage and linear FE for the displacement. Errors in u_h are quantified by the norm difference $\|u_h\|_{L^2(\hat{\Omega})} - \|u\|_{L^2(\hat{\Omega})}$ to the space-continuous collocation solution u, which is closely related to the error in the average conduction velocity.

First we study the performance impact of interleaving mesh refinement and SDC iterations. To this extent, we simulate the non-interleaving mode of operation by initializing the solution at all collocation points to the initial value after mesh refinement, in effect starting the SDC method only after a suitably refined grid has been constructed for the Euler solution. This mimics the approach used in some Rosenbrock schemes [8], where mesh adaptation is performed for the first stage only.

As shown in Fig. 1 left, the interleaved scheme is more efficient, roughly by a factor of two for large tolerances. The non-interleaved mode does not achieve high accuracy at all, independent of the tolerance. Figure 1 right gives an explanation for this bad performance. It turns out that at the chosen time step size the first sweep results in a rather poor approximation of the front, in particular a too slow front speed and a significant overshoot. This leads to mesh refinement behind, and an insufficient refinement at the actual front position.

Next we turn to multi-rate integration for electromechanical coupling. With a fixed tolerance for spatial discretization error and SDC iteration error, we reduce the accuracy of displacement computation in each time step by reducing the collocation nodes from 4 to 1 (lines *a*), the number of simplified Newton steps from 10 to 1

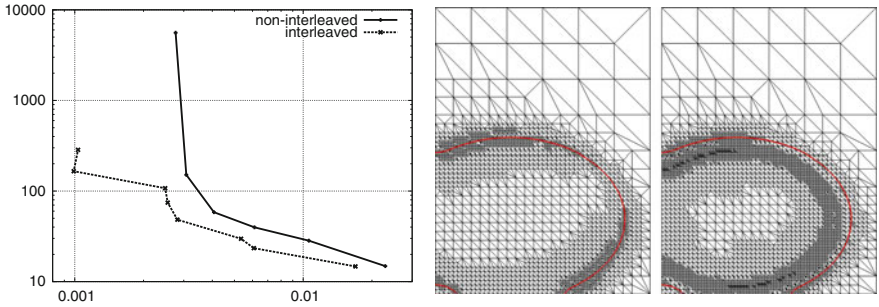


Fig. 1 *Left:* Wall clock time vs. achieved error for different tolerances. *Right:* Grid maladaptation by mesh refinement based on the first sweep. Front position is marked

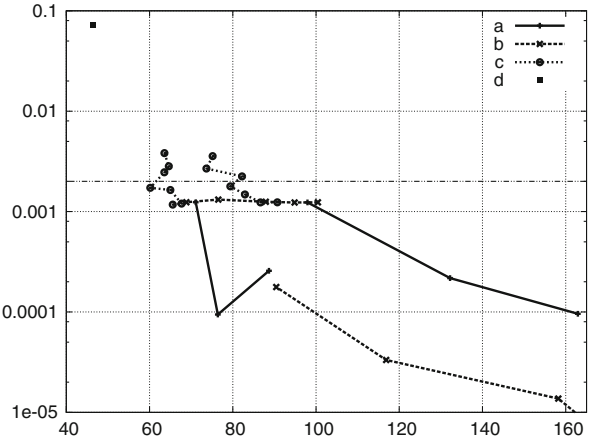


Fig. 2 Total error vs. run time for inexact solution of displacement

(lines *b*), additionally skipping the displacement computation for up to 7 SDC sweeps (lines *c*), and report the deviation from the non-reduced reference solution in Fig. 2. The error of this reference solution is roughly 2×10^{-3}. Apparently, reduction of Newton iteration count and collocation points for the displacement computation introduce a coupling error well below the overall error tolerance. Additionally omitting the displacement computation during the first SDC sweeps exceeds this limit, without substantial run time reduction. Neglecting the mechano-electrical feedback completely yields an unacceptably large error (point *d*).

References

1. Aliev, R.R., Panfilov, A.V.: A simple two-variable model of cardiac excitation. Chaos Solitons Fractals **7**, 293–301 (1996)
2. Bendahmane, M., Buerger, R., Ruiz-Baier, R.: A multiresolution space-time adaptive scheme for the bidomain model in electrocardiology. Numer. Methods Partial Differ. Equ. **26**, 1377–1404 (2010)

3. Colli Franzone, P., Deuflhard, P., Erdmann, B., Lang, J., Pavarino, L.F.: Adaptivity in space and time for reaction-diffusion systems in electrocardiology. SIAM J. Sci. Comput. **28**, 942–962 (2006)
4. Dutt, A., Greengard, L., Rokhlin, V.: Spectral deferred correction methods for ordinary differential equations. BIT Numer. Math. **40**(2), 241–266 (2000)
5. Göktepe, S., Kuhl, E.: Electromechanics of the heart – a unified approach to the strongly coupled excitation-contraction problem. Comput. Mech. **80**, 227–243 (2010)
6. Humphrey, J.D.: Cardiovascular Solid Mechanics, Cells, Tissues and Organs. Springer, New York (2001)
7. Kerckhoffs, R.C.P., Bovendeerd, P.H.M., Kotte, J.C.S., Prinzen, F.W., Smits, K., Arts, T.: Homogeneity of cardiac contraction despite physiological asynchrony of depolarization: a model study. Ann. Biomed. Eng. **31**, 536–547 (2003)
8. Lang, J.: Adaptive multilevel solution of nonlinear parabolic PDE systems. Lecture Notes in Computational Science and Engineering, vol. 16. Springer, New York (2001)
9. Nash, M.P., Panfilov, A.V.: Electromechanical model of excitable tissue to study reentrant cardiac arrhythmias. Prog. Biophys. Mol. Biol. **85**, 501–522 (2004)
10. Pathmanathan, P.J., Whiteley, J.P.: A numerical method for cardiac mechanoelectric simulations. Ann. Biomed. Eng. **37**, 860–873 (2009)
11. Vetter, F.J., McCulloch, A.D.: Three-dimensional stress and strain in passive rabbit left ventricle: a model study. Ann. Biomed. Eng. **28**, 781–792 (2000)
12. Whiteley, J.P., Bishop, M.J., Gavaghan, D.J.: Soft tissue modelling of cardiac fibres for use in coupled mechano-electric simulations. Bull. Math. Biol. **69**, 2199–2225 (2007)

MS 11
MINISYMPOSIUM: MATHEMATICAL MODELLING IN ENERGY MARKETS

Organizers

Michael Coulon[1] and Matthias Ehrhardt[2]

Speakers

Michael Coulon[3]
An Introduction to Mathematical Models in Energy Markets

Nina Lange[4]
Currency Risk in Energy Markets

Virginie Dordonnat[5]
Nonparametric Modelling and Short-Term Forecasting of French Electricity Demand and Spot Prices

Christian Jacobsson[6]
Optimizing Renewable, Hydro and Gas Assets as Realistically as Possible

[1]Michael Coulon, University of Sussex, Sussex, UK.
[2]Matthias Ehrhardt, University of Wuppertal, Wuppertal, Germany.
[3]Michael Coulon, University of Sussex, Sussex, UK.
[4]Nina Lange, Copenhagen Business School, Copenhagen, Denmark.
[5]Virginie Dordonnat, EDF, France.
[6]Christian Jacobson, Alpiq, Switzerland.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_43

Christian Hendricks[7], Matthias Ehrhardt[8] and Michael Günther[9]
Clean Spread Options in the German Electricity Market

Alexander Boogert[10]
Explaining Gas Storage Levels Using Price Information

Magnus Wobben[11]
Valuation of Storage Contracts in Incomplete Gas Markets

Tamsin Lee[12], Stephen A. Haben[13] and Peter Grindrod[14]
Modelling the Electricity Consumption of Small to Medium Enterprises

Keywords

Energy markets
Energy modeling
Energy risk management

Short Description

The rapid changes in energy trading within the last two decades have attracted many researchers in academia and industry. Their aim is to adequately model energy prices and typically also to design methods and guidelines for risk management challenges such as power plant portfolio optimization.

The well-known non-storability property of electricity (and challenges in natural gas storage) leads to major modelling differences compared to stock or bond markets. Coupled with highly inelastic demand and a variety of supply side constraints, this lack of storage can result in sudden price spikes and high, time-varying volatility. Furthermore, mean reversion rates and typical seasonal patterns exhibit a complex multi-scale nature with respect to the time variable: intra-day, weekly and annual.

[7]Christian Hendricks, University of Wuppertal, Wuppertal, Germany.
[8]Matthias Ehrhardt, University of Wuppertal, Wuppertal, Germany.
[9]Michael Günther, University of Wuppertal, Wuppertal, Germany.
[10]Alexander Boogert, EnergyQuants, Amsterdam.
[11]Magnus Wobben, d-fine AG, Frankfurt.
[12]Tamsin Lee, University of Oxford, Oxford, UK.
[13]Stephen A. Haben, University of Oxford, Oxford, UK.
[14]Peter Grindrod, University of Oxford, Oxford, UK.

The aim of this mini symposium is to discuss the latest research in industry and academia in energy modelling and energy risk management. It will cover different modelling approaches for energy prices with particular focus on applications in gas and electricity markets.

Integrated Forecasting of Day-Ahead Prices in the German Electricity Market

Christian Hendricks, Matthias Ehrhardt, and Michael Günther

Abstract Since the start of the liberalization of energy markets the energy sector has undergone major changes. Energy companies now provide electricity at variable prices and are faced to a competitive market environment. Their trading is subject to risks and uncertainty about future price developments. In this work we introduce a regularized regression approach to forecast *Phelix Peak* prices in the German electricity market. Additionally we investigate the influence of fundamental price drivers on the forecasting accuracy. Since the problem complexity grows exponentially with the dimension of the feature space, the regression problem suffers from the curse of dimensionality. To cope with this problem we apply the combination technique, which enables us to reduce the complexity while keeping a high approximation accuracy.

Keywords Electricity spot price forecasting • Energy markets • Regression approach • Sparse grid

1 Introduction

During the last two decades the energy sector has undergone major changes. Energy companies now provide electricity at variable prices and are faced to a competitive market environment. Their trading is subject to risks and uncertainty about future price developments. These risks are mainly associated with the volatile nature of input costs, like coal and gas prices, but also other factors influence energy markets. The global concern regarding the climate change has led to the introduction of an emission trading system in the European Union (EU). Today energy suppliers have to surrender *European Emission Allowances* (EUA) to offset their emission of greenhouse gases. As a conventional power plant has to burn fuel and emit CO_2 to produce electricity, these allowances can be interpreted as additional input costs.

C. Hendricks (✉) • M. Ehrhardt • M. Günther
Lehrstuhl für Angewandte Mathematik und Numerische Analysis, Fachbereich C Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany
e-mail: hendricks@math.uni-wuppertal.de; ehrhardt@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_44

The vastly changing market environment has attracted lots of researchers to develop forecasting models for electricity markets on different time frames. They range from ARMA to neural network approaches or models known from game theory. Regarding energy markets in continental Europe mainly the Spanish and German market have been focused [2, 3, 10].

In this work we present a non-linear approach to forecast electricity spot prices in the German electricity market. Additionally the influence of fundamental price drivers on the spot price and the forecasting accuracy shall be investigated. The arising high dimensional approximation problems will be solved with the combination technique on sparse grids [1, 4, 7].

2 The Data Set

The data set consists of electricity spot prices (Phelix Peak) and time series of coal (ARA coal future, front month with nearest expiry), gas (GASPOOL Spot), European Emission Allowances and day-ahead forecasts of wind and solar supplies.[1] All time series range from 2011 to 2012. The coal time series is quoted in USD per t and has been converted to EUR per t.

In order to avoid any perturbation caused by seasonal patterns, we only consider "business-as-usual" days. With the help of the Augmented Dickey-Fuller test (ADF) we check for a unit root in each of the time series. Table 1 shows the test results: the electricity, coal, gas and emission allowance time series exhibit a unit root at the common confidence level of 5 %. Therefore we differentiate the time series to eliminate the stochastic trends. To achieve variance stationarity the logarithm of all price series is taken.

3 The Modelling Framework

In order to forecast electricity prices we apply a regularized regression approach, which is able to capture non-linear relationships between the input variables and the desired output.

Table 1 Augmented Dickey-Fuller test results

	Electricity	Coal	Gas	EUA	Wind supply	Solar supply
p-Value	0.1008	0.0973	0.8590	0.1370	0.0010	0.0018

[1] Provided by www.transparency.eex.com.

3.1 Regularized Regression

In this section we formulate the forecasting problem as a regularized least square regression. These kind of models have already proven to be useful in data mining [5], foreign exchange [6] and wind time series forecasting [9]. Let $\Omega \subseteq \mathbb{R}^d$ denote a d dimensional feature space, then it is the goal to find a function $f : \Omega \to \mathbb{R}$, which maps the model's input $x_i \in \mathbb{R}^d$ to the desired output $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$ observations. The unknown function f belongs to some function space V, which we will specify later on. The resulting regularization problem can be written as

$$\inf_{f \in V} \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \left\| Pf \right\|^2_{L^2(\Omega)} \right). \tag{1}$$

The second term is a penalty term for non-smooth f. The parameter $\lambda > 0$ determines the balance between the accuracy of the fitted function and its smoothness. P is a regularization operator, we will use $P = \nabla$. In order to estimate f, a function space is needed to be specified. We will restrict ourselves to a finite dimensional space $V_m \subseteq V$ and express f with the help of basis functions $\{\phi_i\}_{i=1,\ldots,m}$ by

$$f(x) = \sum_{i=1}^{m} \alpha_i \phi_i(x). \tag{2}$$

Plugging (2) into (1) the approximation reduces to a minimization problem, which can be rewritten as a linear equation system $\left(\lambda C + B B^T\right) \alpha = By$, with matrices $C_{j,k} = n \langle \nabla \phi_j, \nabla \phi_k \rangle_{L^2(\Omega)}, j, k = 1, \ldots, m, B_{j,i} = \phi_j(x_i), j = 1, \ldots, m, i = 1, \ldots, n$ and the m dimensional vector α. The vector α contains the degrees of freedom and represents a unique solution if the minimization problem is well-posed.

If the dimension of the feature space increases, the system that has to be solved grows and the curse of dimensionality shows its effects quickly. On a uniform grid with mesh size $h_N = 2^{-N}$, and level $N \in \mathbb{N}$, this would lead to $\mathcal{O}(h_N^{-d})$ degrees of freedom. To cope with this problem we use the sparse grid combination technique. In [7] it is shown that the number of grid points can be lowered to an order of $\mathcal{O}(h_N^{-1} \log(h_N^{-1}))^{d-1})$. In the following we will briefly recall the fundamentals of this technique. For a detailed introduction to sparse grids we refer to [1, 7].

The combination technique is based on linearly combining a sequence of functions. Let $\Omega := [0, 1]^d$ be the d dimensional unit cube and let $\mathbf{l} = (l_1, \ldots, l_d) \in \mathbb{N}^d$, $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$ denote multi-indices, then we can define a family of grids $\{\Omega_l\}_{l \in \mathbb{N}}$ on Ω with mesh sizes $\mathbf{h}_l = (h_{l_1}, \ldots, h_{l_d}) = (2^{-l_1}, \ldots, 2^{-l_d})$. Each grid consists of the points $\mathbf{x}_{l,i} = (x_{l_1,i_1}, \ldots, x_{l_d,i_d})$ with $x_{l_t,i_t} = i_t h_{k_t}$, $i_t = 0, \ldots, 2^{l_t}$, $t = 1, \ldots, d$. The d dimension basis functions $\phi_{l,i}$ are given by the tensor product of

piecewise linear one dimensional basis functions ϕ_{l_t,i_t} for $x \in \Omega$

$$\phi_{l,i}(x) = \prod_{t=1}^{d} \phi_{l_t,i_t}(x_t).$$

The one dimensional basis function ϕ_{l_t,i_t} is given by a hat function $\phi_{l_t,i_t}(x_t) = \max\{1 - |\frac{x_t - i_t h_{l_t}}{h_{l_t}}|, 0\}$.

With the help of these basis functions the function space $V_l = \text{span}\{\phi_{l,i}, i_t = 1, \ldots, 2^{l_t}, t = 1, \ldots, d\}$ on grid Ω_l can be defined. The function f_l on Ω_l is represented by

$$f_l(x) = \sum_{i_1=1}^{2_1^l} \ldots \sum_{i_d=1}^{2_d^l} \alpha_{l,i} \phi_{l,i}(x).$$

If we combine linearly the solution f_l from different grids Ω_l according to the formula

$$f_N(x) = \sum_{q=0}^{d-1} (-1)^q \binom{d-1}{q} \sum_{|l|_1 = N-q} f_l(x),$$

we obtain the function f_N, which lives in the sparse grid space with $\mathcal{O}(h_N^{-1}(\log(h_N^{-1}))^{d-1})$ grid points. Provided that f fulfills certain smoothness conditions the approximation error is of order $\mathcal{O}(h_N^2 \log(h_N^{-1})^{d-1})$.

3.2 Multivariate ARMA

In order to evaluate the performance quality of the fitted function of the previous section, we apply ARMA/VARMA models as a benchmark. The model is of the form

$$\phi(B) y_t = \theta(B) e_t + \omega(B) u_t,$$

where B is the back shift operator. y_t and e_t are p dimensional vectors of observed output variables and unobserved residuals. The model order is chosen with the help of the Akaike Information Criterion and the parameters are calibrated with the R software package *DSE*.[2]

[2]Dynamic System Estimation (DSE): available at www.cran.r-project.org.

4 Forecasting Methodology and Results

The accuracy of our forecasts is quantified with the help of the following measures, where $\hat{y}_i$ is the prediction and y_i the true electricity spot price for values $i = 1, \ldots, n$: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (RMSE).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \quad \text{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

To build up our forecasting model, we use half of our data as a training set. The first half of 2012 works as a validation set to select the model parameters. Based on the selected model, we compute day-ahead forecasts for the second half of the year 2012 to test the out-of-sample behavior of our fitted models.

4.1 Forecasting Results

The easiest feature space one can think of consists of the time series itself. With the help of the validation set, we experimentally answer the question how many delayed values should be included. Experiments turn out that two values are appropriate. The training procedure of the regularized regression approach is therefore proposed to find a function f, which matches

$$y_t - y_{t-1} = f_1(y_{t-1} - y_{t-2}, y_{t-2} - y_{t-3})$$

for all t in the training set. In Table 2 we compare the results of the regularized regression approach and an ARMA model. The out-of-sample accuracy is slightly worse than in the validation set. However it can beat the ARMA model in all three accuracy measures. We now extend the feature space by coal and gas time series. The resulting six dimensional regression problem is of the form

$$\begin{aligned} y_t - y_{t-1} = f_2(&y_{t-1} - y_{t-2}, y_{t-2} - y_{t-3}, \\ &c_{t-1} - c_{t-2}, c_{t-2} - c_{t-3}, g_{t-1} - g_{t-2}, g_{t-2} - g_{t-3}), \end{aligned}$$

Table 2 Forecast results

	$\text{RegReg}_{\text{validation}}$	$\text{RegReg}_{\text{out-of-sample}}$	ARMA
MAE	4.8914	4.9398	5.0255
MAPE [%]	9.8921	9.7868	9.8946
RMSE	6.3562	6.9317	7.0979

Table 3 Forecast results including fuel prices

	$\text{RegReg}_{\text{validation}}$	$\text{RegReg}_{\text{out-of-sample}}$	VARMA
MAE	4.9617	5.0884	5.0950
MAPE [%]	9.9802	10.0623	10.1035
RMSE	6.4023	7.1392	7.1341

Table 4 Forecast results including EUAs

	$\text{RegReg}_{\text{validation}}$	$\text{RegReg}_{\text{out-of-sample}}$	VARMA
MAE	4.9617	5.0884	5.1013
MAPE [%]	9.9802	10.0436	10.1038
RMSE	6.4023	7.0247	7.1263

Table 5 Forecast results including wind and solar production forecasts

	$\text{RegReg}_{\text{validation}}$	$\text{RegReg}_{\text{out-of-sample}}$	VARMA
MAE	3.9290	3.5466	3.5231
MAPE [%]	7.5325	6.6547	6.9041
RMSE	5.7464	5.0532	5.1876

where c_t is the coal price and g_t the gas price at time t in the training set. Table 3 shows the forecasting results. The introduction of both fuel price series seem to have an adverse influence on the accuracy. In order to check, whether historical prices of EUAs can enhance the accuracy, we add them to the feature space and obtain a four dimensional problem

$$y_t - y_{t-1} = f_3(y_{t-1} - y_{t-2}, y_{t-2} - y_{t-3}, e_{t-1} - e_{t-2}, e_{t-2} - e_{t-3}).$$

In Table 4 we compare the forecasting quality of both models. The regression approach slightly outperforms its benchmark. We see that an addition of EUAs to the feature space does not improve the forecasting results. Along with coal and gas fired power plants, renewable energy sources play an important role in the German electricity market. Since the *Renewable Energy Act* (EEG) and the preferred feed-in of green energy, there is a deep impact of production capacities provided by wind and solar generators on the spot price for electricity [8]. The variable w_t denotes the wind production forecast at time t, while s_t is the solar production forecast. They are published by the *transmission system operators*[3] (TSO) and we assume that this information is available at time level $t - 1$. The fitting problem reads

$$y_t - y_{t-1} = f_4(y_{t-1} - y_{t-2}, y_{t-2} - y_{t-3}, w_t - w_{t-1}, s_t - s_{t-1}).$$

Table 5 shows the great improvement to the previous feature spaces. The error in terms of the MAE, MAPE and RMSE can be lowered by 28.20, 32.07 and 27.15 % compared to the first model.

[3] 50Hertz, Amprion, APG, TenneT, TransnetBW.

5 Conclusion

In this article we investigated the potential of a non-linear regression approach in the prediction of day-ahead electricity prices in Germany. The out-of-sample tests show that this model performs better than its benchmark ARMA/VARMA model. The strength of our technique lies in its ability to capture a big variety of relationships up to a high order of dimension. Within this work we considered problems up to dimension 6. In four different tests we evaluated the benefit of important impact factors on the prediction accuracy. The inclusion of fuel prices and CO_2 allowance prices turned out to introduce more noise. If wind and solar production forecasts are added to the feature space, the accuracy is greatly improved. These results underline the strong price effects of renewable energy sources in the German electricity market.

References

1. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numer. **13**, 147–269 (2004)
2. Catalâo, J., Mariano, S., Mendes, V., Ferreira, L.: Short-term electricity prices forecasting in a competitive market: a neural network approach. Electr. Power Syst. Res. **77**, 1297–1304 (2007)
3. García-Martos, C., Rodríguez, J., Sánchez, M.J.: Modelling and forecasting fossil fuels, CO_2 and electricity prices and their volatilities. Appl. Energy **101**, 363–375 (2013)
4. Garcke, J.: Regression with the optimised combination technique. In: Proceedings of the 23rd ICML, pp. 321–328 (2006)
5. Garcke, J., Griebel, M., Thess, M.: Data mining with sparse grids. Computing **67**, 225–253 (2001)
6. Garcke, J., Gerstner, T., Griebel, M.: Intraday foreign exchange rate forecasting using sparse grids. In: Sparse Grids and Applications. Lecture Notes in Computational Science and Engineering, vol. 88, pp. 81–105. Springer, New York (2013)
7. Griebel, M., Schneider, M., Zenger, C.: A combination technique for the solution of sparse grid problems. Iterative Methods in Linear Algebra, pp. 263–281. Elsevier, North Holland (1992)
8. Ketterer, J.: The impact of wind power generation on the electricity price in Germany. Ifo Working Paper Series Ifo Working Paper No. 143, Ifo Institute for Economic Research at the University of Munich (2012)
9. Kramer, O., Gieseke, F.: Analysis of wind energy time series with kernel methods and neural networks. In: 2011 Seventh International Conference on Natural Computation (ICNC), vol. 4, pp. 2381–2385 (2011). doi:10.1109/ICNC.2011.6022597
10. Nogales, F., Contreras, J., Conejo, A.J., Espínola, R.: Forecasting next-day electricity prices by time series models. Power Syst. **17**(2), 342–348 (2002)

Modelling the Electricity Consumption of Small to Medium Enterprises

T.E. Lee, S.A. Haben, and P. Grindrod

Abstract Estimating the demand on the low voltage network is essential for the distribution network operator (DNO), who is interested in managing and planning the network. Such concerns are particularly relevant as the UK moves towards a low carbon economy, and the electrification of heating and transport. Furthermore, small to medium enterprises (SMEs) contribute a significant proportion to network demand but are often overlooked. The smart meter roll out will provide greater visibility of the network, but such data may not be readily available to the DNOs. The question arises whether useful information about customer demand can be discerned from limited access to smart meter data? We analyse smart meter data from 196 SMEs so that one may create an energy demand profile based on information which is available without a smart meter. The profile itself comprises of simply two estimates, one for operational power and another for non-operational power. We further improve the profile by clustering the SMEs using a simple Gaussian mixture model. In both cases, the average difference between the actual and predicted operational/non-operational power is less than 0.15 kWh, and clustering reduces the range around this difference. The methods presented here out perform the flat profile (akin to current methods).

Keywords Electricity demand modeling • Energy markets

1 Introduction

Small to medium enterprises contribute a large proportion of the total energy demand in the UK but are often overlooked in research [2]. Therefore it is essential that their demand is accurately modelled so that distribution network operators (DNOs) can manage and plan the network. Such concerns are more immediate with the increase of low carbon technologies, such as electric vehicles and photovoltaics (PV), which will impact the low voltage (LV) network [8]. Currently DNOs use

T.E. Lee (✉) • S.A. Haben • P. Grindrod
Mathematical Institute, University of Oxford, Oxford, UK
e-mail: tamsin.lee@maths.ox.ac.uk; haben@maths.ox.ac.uk; grindrod@maths.ox.ac.uk

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_45

After Diversity Maximum Demand [5] to model maximum demand for SMEs. This approach does not account for time-of-day information, and is therefore becoming increasingly insufficient. For example, as more PV are installed, DNOs require the time of low demand.

The UK is in the preliminary stages of rolling-out smart meters, which measures energy demand of millions of customers every 30 min. In the UK, smart meter data is proprietary, and therefore possibly unavailable to DNOs. Nonetheless, they do have access to quarterly readings. As such, we ask if a DNO can estimate an SMEs electricity profile from quarterly readings and publicly available information. Since SMEs behaviour is regular and predictable (compared to domestic customers), one should be able to accurately recreate their demand profiles with this limited information.

In this paper we use preprocessed smart meter data from the Irish smart meter trial [4]. We consider a years worth of smart meter data for 196 SMEs starting from midnight 14th July 2009. From this data we demonstrate how one could accurately estimate customers weekly energy demand profile using only knowledge of their operating times (potentially public information) and their mean daily usage (potentially available from their quarterly meter readings available from the supplier).

We further improve this estimate by using a basic clustering of operational and non-operational power. The advantage of the method presented in this report is twofold. Firstly, very little information is required to produce the estimates. Secondly, any new customers can be assigned to the current clusters, and therefore can be modelled only knowing their mean daily demand and operational hours.

Along with the smart meter data, there is a survey completed for 138 of the SMEs. This contains replies for a number of questions including, type of business, the number of employees, the age of the building, what weekend days are operational, etc. We consider the relationship between the survey responses and the clusters.

We begin in by describing how we determine operational and non-operational times from the smart meter data. In Sect. 3 we describe how we cluster the customers, and in Sect. 4 we compare our estimates based on clustering and not clustering. Finally, we summarise in Sect. 5.

2 Identifying Operational Hours

To determine a businesses operational hours we use a data driven approach. However, potentially this information is publicly available.

We first use the smart meter data to determine operational days. To do this we state that if the median daily usage for a particular day of the week $\tilde{x}_m^{DAY}$ is less than a chosen quantile of the overall daily use, q_m, we assume the business is closed on that day. For example, suppose $\tilde{x}_m^{SUN} < q_m$ (for meter m), then we assume the business is closed on Sundays. To determine the quantile, we consider quantiles for

probabilities from 0 to 1 (intervals of 0.01). For each quantile choice, the 'closed outcomes' for Saturday and Sunday is compared to a survey response in which customers state which days the business is open on weekends. From the F-score [7], we take the 0.4 quantile since it offers the closest match to the survey responses. There does not appear to be a distinct pattern for the spread of incorrectly predicted meters.

To determine whether a business is operational in a half hour we first remove 'closed' days. Using the remaining 'open' days, the average power is calculated and compared to the average power for each half hour. If the average power for a half hour is larger than the average power over all the 'open' days, we consider the business to be operational during this segment.

3 Clustering

In order to group customers with similar attributes we use clustering based on their operational and non-operational power, and the half hourly standard deviation (as a measure of variation). We can then use these cluster groups to improve our estimate (see Sect. 4).

We model our three attributes as a finite mixture model (FMM) of uncorrelated Gaussian distributions (reference). The parameters of the mixtures and the mixing proportions are found easily through an implementation of the Expectation-Maximisation algorithm, which finds the parameters that maximise the likelihood function of the model [3]. Traditionally the k-means algorithm has been used for clustering in power systems. However, this is simply a less versatile model than the FMM [3]. We use the Matlab function `gmdistribution` to implement the algorithm.

As with many clustering algorithms, a disadvantage of the method is that the number of clusters must be defined before the clustering algorithm is implemented. A large number of clusters will provide a better fit of each meter to its cluster, but the number of model parameters increases. There are no definitive ways of choosing the number of clusters, but there are some indicators and metrics that can be used to help inform the decision. One metric is to find the maximum loglikelihood $\log L(\theta)$ for a model with k parameters fit to N data points. The Bayesian Information Criterion (BIC) for such a model is $-2\log L(\theta) + k\log(N)$, where a smaller BIC indicates a favourable balance between the number of parameters and the model fit [3]. By considering how the BIC changes with the number of clusters, we use it to choose a cluster size which obtains a good model of the observations without an excessive number of parameters. As shown in Fig. 1a, five clusters is a reasonable choice.

Figure 1c shows our clusters in terms of their usage and normalised standard deviation. Since the groups form clear partitions of daily mean, μ_m, an SME without smart meter data can be placed in a cluster using the quarterly readings alone. It was found when comparing the clusters that there is no relationship between the survey responses (such as number of employees, annual turnover, age of building, etc.) and

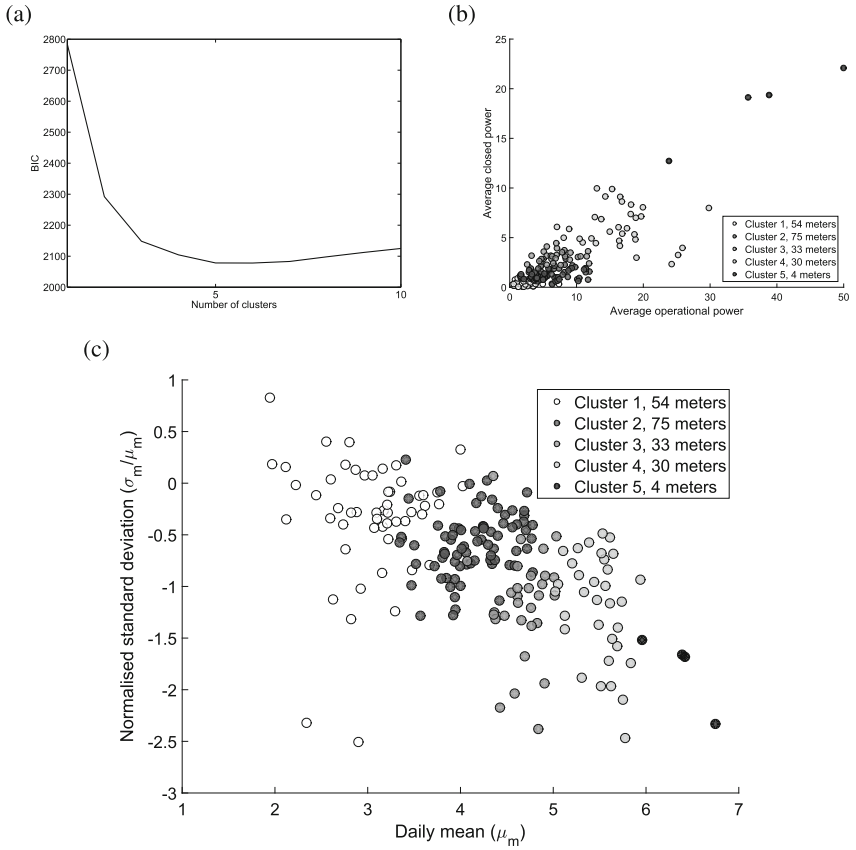


Fig. 1 Clustering when the attributes are average operational power, average non-operational power, and normalised standard deviation. Clusters are numbered by daily mean μ_m: Cluster 1 has the lowest daily mean and Cluster 5 has the highest daily mean. (**a**) The BIC indicates five clusters is most suitable. (**b**) The five clusters arranged by operational power and non-operational power. (**c**) The five clusters arranged by their daily mean μ_m and normalised standard deviation σ_m/μ_m (domestic households have upper bounds of $\log(\sigma_m/\mu_m) \approx 0.65$ and $\log(\mu_m) \approx 1$, see [1])

the clusters (therefore the daily mean). This supports previous research which also found little correlation between energy consumption and household type [6].

Here we have assumed that the attributes are uncorrelated. Clearly assuming that operational and non-operational power are not correlated is misleading (see Fig. 1b). However, clustering the data with correlation provides an optimum model of four clusters (as opposed to five). This in itself is not a problem; but under this model the last cluster (which ideally only encapsulates the high end users, say those with log mean daily use larger than 6, see Fig. 1c) encapsulates users with log mean daily use as low as 5—making it impossible to differentiate between Cluster 3 and Cluster 4 based on mean daily use alone. We recommend that future SME smart meter data sets are clustered with correlation where possible.

4 Predicting Electricity Use

We create an estimated electricity profile for each of the 196 m in our dataset. The prediction is then compared to the actual data set. For a meter m ($m = 1, 2, \ldots, 196$) let us first consider all meters as a single set (not clustered). We estimate the average operational power $\hat{e}_{m,O}$ and average non-operational power $\hat{e}_{m,N}$ (for meter m) using the mean for the corresponding variables from the remaining 195 m. Therefore, since we know the operational hours for meter m, we estimate the total predicted energy during an average week, T_m, for meter m, via

$$T_m = H_{m,O} \sum_{i=1}^{H_{m,O}} \hat{e}_{m,O} + H_{m,N} \sum_{i=1}^{H_{m,N}} \hat{e}_{m,N}, \tag{1}$$

where $H_{m,O}$ is the number of operational hours and $H_{m,N}$ is the number of non-operational hours in a week. Ideally, the estimate (1) matches the weekly energy use for the meter, $T_m = 7\mu_m$ (obtained for quarterly reading data). Therefore, we adjust the profile by setting

$$e_{m,O} = \frac{7\mu_m}{T_m}\hat{e}_{m,O} \quad \text{and} \quad e_{m,N} = \frac{7\mu_m}{T_m}\hat{e}_{m,N}, \tag{2}$$

where $e_{m,O}$ is the adjusted average operational power, and $e_{m,N}$ is the adjusted average non-operational power. Using these adjusted power values for meter m, and the operational hours (see Sect. 2), we compose the predicted profile. This process is carried out for all meters. We repeat this process for the clustered data set, using the mean values for the current meter's cluster. As expected, the adjustment from $T/7\mu$ is larger when the data is not clustered (a maximum adjustment of 16.43 compared with a maximum adjustment of 3.98). This confirms that the other members of a cluster have similar weekly usages.

As an example, Fig. 2 compares this predicted average weekly profile with the actual average weekly profile for a single meter from the dataset. We used the non-clustered data and the clustered data. The peak behaviour is captured better when using the data from the cluster (cluster 1, 54 m) compared to the whole data set (195 m). This is similar for other customers.

To test the accuracy of these two methods (non-clustered and clustered), we compare them with a flat estimate, which is the daily mean μ_m divided by 48 half hours. Two error measures are calculated, one for the operational power and one for the non-operational power,

$$E_{m,O} = \bar{e}_{m,O} - e_{m,O}, \; E_{m,N} = \bar{e}_{m,N} - e_{m,N}, \tag{3}$$

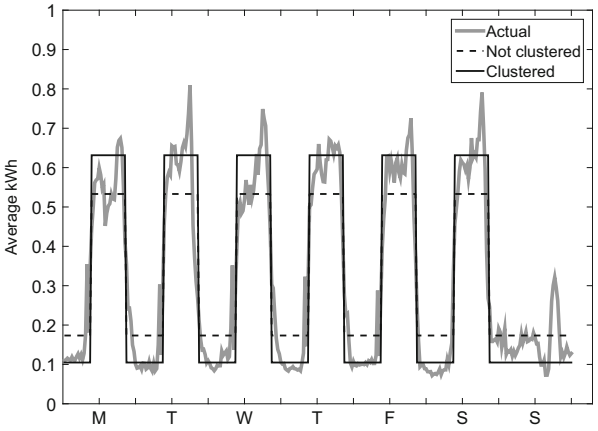


Fig. 2 The predicted average week (not clustering and clustering) and the actual average week for an example meter. The data was clustered according to average operational power, average non-operational power and normalised standard deviation

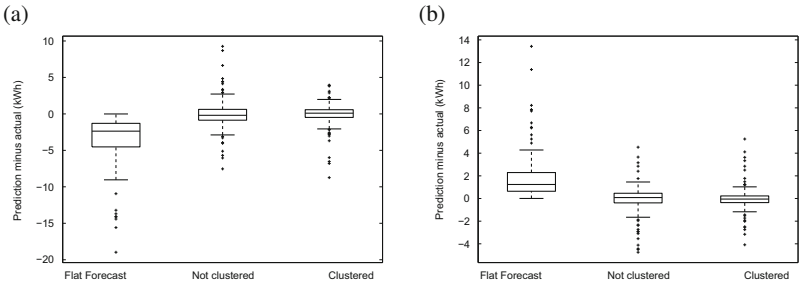


Fig. 3 Comparing the error from a flat estimate, our prediction when not clustering the data, and our prediction when clustering the data. (**a**) Average operational power. From *left to right*, the medians are 1.13, 0.08 and −0.05 kWh. (**b**) Average non-operational power. From *left to right*, the medians are −2.15, −0.13 and −0.06 kWh

where $\bar{e}_{m,O}$ is the actual average operational power, and $\bar{e}_{m,N}$ is the actual average non-operational power. We consider the prediction coming from the whole data set, from the clustered data set, and from a simple flat prediction ($e_{m,O} = e_{m,N} = \mu_m/24$). Such an estimate comes from only knowing quarterly information and thus we have no time-of-use information.

Both clustering and not clustering the data outperforms the flat estimate, see Fig. 3. It is likely that the adjustment stage, Eq. (2), ensures that the non-clustered and clustered approaches are competitive to each other. To remove the effect of seasonality, we applied our methods to the first 3 months of the data set. The difference in results from using 3 months to 1 year is negligible, and hence not presented here.

4.1 Peak Usage on an Aggregated Level

We have presented a method that predicts the average weekly profile to a high level of accuracy. As well as an accurate estimate for an electricity profile, DNOs are interested in the size of peak electricity demand on an aggregated level. Naturally peak electricity usage is always underestimated when using the mean. Furthermore, this underestimation grows as meters are aggregated. We now describe a simple adaptation to the method so that peak demand is more accurately estimated on an aggregated level. We compare peak demand for up to 25 m aggregated together, which is approximately the maximum number of SMEs on a single phase on a feeder.

The initial estimate for the average operational power $\hat{e}_{m,O}$ and average non-operational power $\hat{e}_{m,N}$ (for meter m) is no longer the mean from the cluster (or whole dataset). Instead, for each meter m in the aggregation, we sample from the Normal distribution

$$\hat{e}_{m,O} \sim N(M_O, \Sigma_O^2),$$
$$\hat{e}_{m,N} \sim N(M_N, \Sigma_N^2),$$

where M_O is the mean operational power for the appropriate cluster (or whole dataset), and Σ_N^2 is the variance for the appropriate cluster (or whole dataset). When the means (M_0, M_N) and variances (Σ_O^2, Σ_N^2) are determined by the cluster, we are using the Gaussian distributions determined by the finite mixture model.

To avoid unrealistic estimates, if a negative value is sampled, or a value much larger than the mean ($M_O + 3\Sigma_O$ or $M_N + 3\Sigma_N$ appropriately), we discard the value and re-sample. The initial estimates for an aggregate of meters is simply the sum of the $\hat{e}_{m,O}$ or $\hat{e}_{m,N}$ for the meters in the aggregate. From here, the initial estimates can be adjusted so that the weekly estimates corresponds to the daily mean from quarterly readings (see (1) and (2)). Note that we are now shifting the operational power prediction up so as to better match peak demand so that, on average, the prediction is overestimating. This risk averse practise is common for DNOs.

The estimate more closely matches the actual peak usage as the aggregate size increases, see Fig. 4a. Conveniently, the estimate converges to a value a little larger than the actual usage, which (as mentioned above) is preferred by DNOs. Taking the initial estimates, and not adjusting for the daily mean, provides an even greater overestimate—should a DNO prefer a larger overestimate, see Fig. 4b.

5 Conclusion

This paper identified when a business is open based upon their electricity use. With merely the operational hours (publicly available information) and the daily mean (available from quarterly readings), we have created a good approximation

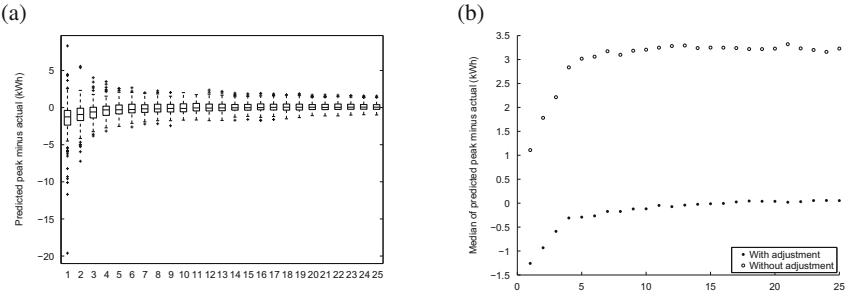


Fig. 4 The predicted peak demand compared to the actual peak demand for aggregates of 1–25 m, where the initial estimates are sampled from a Normal distribution. (**a**) The difference between prediction and actual peak usage (with adjustment according to the daily mean). (**b**) The median difference where the prediction is adjusted according to the daily mean, and where it is not

of a SME customers weekly energy demand profile. The prediction significantly outperforms the flat estimate, which is akin to current methods. This approximation can be further improved by clustering the data before making a prediction. We used operational power, non-operational power and standard deviation to cluster the meters into five categories. The operational and non-operational power are related to the daily mean. Consequently, customers without smart meters can be readily placed in a cluster when only their quarterly readings are available.

The method is flexible to match a DNOs objective. Should a DNO be more interested in peak demand on an aggregated level, then the method is easily accommodated so that peak usage is accurately estimated (or overestimated should this be favoured). Interesting future work would be examining the distribution of electricity usage during operational hours, providing DNOs with more information than simply the mean, or peak usage for a SME (or aggregate thereof).

Counter-intuitively, we did not find any correlation between data attributes (such operational power, non-operational power, daily mean, and standard deviation) and non-electricity features of the data (such as number of employees and the type of business). However, this is in line with earlier work [6].

Acknowledgements The authors thank SSEPD for support via the New Thames Valley Vision Project (SSET203 New Thames Valley Vision), funded through the Low Carbon Network Fund.

References

1. Haben, S. (2013) Categorisation of data: Technical report, University of Reading
2. Kannan, R., Boie, W.: Energy management practices in SME: case study of a bakery in Germany. Energy Convers. Manag. **44**(6), 945–959 (2003)
3. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2004)

4. McLoughlin, F., Duffy, A., Conlon, M.: Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: an Irish case study. Energy Build. **48**, 240–248 (2012)
5. McQueen, D.H., Hyland, P.R., Watson, S.J.: Monte Carlo simulation of residential electricity demand for forecasting maximum demand on distribution networks. IEEE Trans. Power Syst. **19**(3), 1685–1689 (2004)
6. Morley, J. Hazas. M.: The significance of difference: understanding variation in household energy consumption. In: ECEEE Summer Study (2011)
7. Powers, D.M.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. **2**(1), 37–63 (2011)
8. Putrus, G.A., Suwanapingkarl, P., Johnston, D., Bentley, E.C., Narayana, M.: Impact of electric vehicles on power distribution networks. In: Vehicle Power and Propulsion Conference, VPPC'09. IEEE, pp. 827–831 (2009)

MS 12
MINISYMPOSIUM: MATHEMATICAL MODELLING OF DRUG DELIVERY

Organizers

Sean McGinty[1] and Sean McKee[2]

Speakers

Giuseppe Pontrelli[3] and Sean McGinty[4]
Drug Delivery in Biological Tissues: A Two-Layer Reaction-Diffusion-Convection Model

Sean McGinty[4], Sean McKee[5], Christopher McCormick[6] and Marcus Wheel[7]
A Mathematical Model for the Estimation of Drug Binding Rates in Biological Tissue

[1]Sean McGinty, University of Strathclyde, Glasgow, UK.
[2]Sean McKee, University of Strathclyde, Glasgow, UK.
[3]Giuseppe Pontrelli, Istituto per le Applicazioni del Calcolo, CNR, Rome, Italy.
[4]Sean McGinty, University of Strathclyde, Glasgow, UK.
[5]Sean McKee, University of Strathclyde, Glasgow, UK.
[6]Christopher McCormick, University of Strathclyde, Glasgow, UK.
[7]Marcus Wheel, University of Strathclyde, Glasgow, UK.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_46

Martin Meere[8] and Tuoi Vo[9]
Mathematical Models for Drug Release from Affinity Hydrogels

Jahed Naghipoor[10], José Ferreira[11] and Paula De Oliveira[12]
A Coupled Non-Fickian Model of a Cardiovascular Drug Delivery System

Franz Bozsak[13], Francois Cornat[14], Jean-Marc Chomaz[15] and Abdul Barakat[16]
Optimization of Drug-Eluting Stents

Keywords

Biomedical science
Drug delivery

Short Description

Medical device companies spend tens of millions of euros annually developing novel and sophisticated drug delivery devices. In the majority of cases an empirical approach is adopted during the product development stage resulting in a great many experiments, often involving animals. It is not uncommon for projects to be abandoned after great expense due to unacceptable performance in laboratory or clinical trials. Mathematical modelling of drug delivery devices has the potential to not only reduce the number of costly experiments in product development, and identify products which are doomed to failure, but also to provide a better understanding of the underlying drug release mechanism(s) and drug redistribution in living tissue. This is of critical importance since an element of control is often

[8]Martin Meere, Department of Mathematics, Statistics and Applied Mathematics, NUI Galway, Republic of Ireland.

[9]Tuoi Vo, MACSI, University of Limerick, Republic of Ireland.

[10]Jahed Naghipoor, Department of Mathematics, University of Coimbra, Portugal.

[11]José Ferreira, Department of Mathematics, University of Coimbra, Portugal.

[12]Paula De Oliveira, Department of Mathematics, University of Coimbra, Portugal.

[13]Franz Bozsak, Cardiovascular Cellular Engineering Laboratory, Ecole Polytechnique, Paris, France.

[14]Francois Cornat, Cardiovascular Cellular Engineering Laboratory, Ecole Polytechnique, Paris, France.

[15]Jean-Marc Chomaz, Cardiovascular Cellular Engineering Laboratory, Ecole Polytechnique, Paris, France.

[16]Abdul Barakat, Cardiovascular Cellular Engineering Laboratory, Ecole Polytechnique, Paris, France.

required: if too much drug is delivered to the biological system then toxicity can arise and if too little drug reaches the affected area then it will not have the desired effect.

This symposium will bring together researchers from across Europe who are applying mathematical models and techniques, both analytical and numerical, to try to better understand drug delivery in living systems. The Symposium will provide a platform for new methodologies and ideas to be discussed. Since the modelling of drug delivery involves solving a mass transport problem where diffusion, dissolution, convection and binding often play important roles, it is anticipated that the techniques and ideas presented may well be complementary and applicable to a number of different drug delivery systems.

Drug Delivery in Biological Tissues: A Two-Layer Reaction-Diffusion-Convection Model

Sean McGinty and Giuseppe Pontrelli

Abstract In this paper we present a general model of drug release from a delivery device and the subsequent drug transport in biological tissue. Our model consists of a system of partial differential equations describing the solid–liquid mass transfer and diffusion in the device coating as well as the drug transport through the biological tissue via diffusion, convection and reaction. The drug release from the device depends not only on the properties within the coating and the tissue, but also on the coupling of the two layers. In order to take this into account, our model fully couples the two distinct layers through flux and permeable interface conditions. The model has a wide applicability and we point the reader towards some solution methods, noting that simplifications may be made depending on the parameter values in a given system.

Keywords Biomedical science • Drug delivery

1 Introduction

Local drug delivery devices (DDD) have received much attention in recent years, since they provide a convenient means of targeting drug at the site where it is needed most. Historically, drugs have been administered either orally, topically or hypodermically, and often by the patients themselves. The advent of local DDD has meant that drug delivery can be more controlled, with a prescribed amount of drug being delivered over the necessary time period, and with less input required from the patient. Whilst the drug delivery may in principle be monitored, it is often unclear how the DDD can be designed to achieve the level of control required for a specific

S. McGinty
Strathclyde University, Glasgow, UK
e-mail: s.mcginty@strath.ac.uk

G. Pontrelli (✉)
IAC-CNR, Rome, Italy
e-mail: giuseppe.pontrelli@gmail.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_47

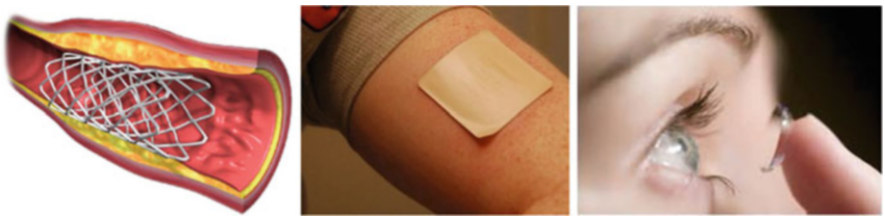


Fig. 1 Three examples of DDD: the drug-eluting stent (*left*), the transdermal patch (*center*), the therapeutic contact lens (*right*)

purpose, since there exists complex interplay between biology, polymer chemistry and pharmacology [1].

Important examples of local DDD include drug-eluting stents for the prevention of restenosis following percutaneous coronary intervention [2], therapeutic contact lenses to deliver ophthalmic drugs [3], and transdermal drug delivery [4] (Fig. 1). In each case, the drug is commonly contained in a polymeric gel platform that is in direct contact with the biological tissue. The polymeric gel acts as a reservoir of drug and provides an adjustable level of control over the rate of drug delivery to the tissue [5]. Both the polymeric gel layer and the interfaced tissue are treated as porous media from a macroscopic point of view. The therapeutic success is dependent on the extent of drug elution, the rate of release, accumulation of drug and binding to components within the tissue [6]. Furthermore, the local drug concentrations achieved are directly correlated with biological effect and local toxicity. The pharmacological effects of the drug, tissue accumulation, duration and distribution could potentially have an effect on its efficacy and a delicate balance between adequate amount of drug delivered over an extended period of time and minimal local toxicity should be found [7].

Mathematical modelling can serve as an extremely useful tool for providing insight into the important parameters in the system, and to give an indication of how the device may be modified to achieve the desired drug release profile. Many studies on DDD have been conducted regarding efficacy and optimal design either with experimental methods, or modelling/numerical simulations, or a combination of both [8–12]. Nonetheless, many questions remain unanswered for bioengineers and pharmaceutics developers who continue to explore and evaluate this technology. In particular, finding the optimum dose to be delivered in a personalized way to a specific tissue still remains a significant challenge.

In this paper we present a general model of drug release from a delivery device and the subsequent drug transport in biological tissue. Our model consists of a system of partial differential equations describing the solid–liquid mass transfer and diffusion in the device coating as well as the drug transport through the biological tissue via diffusion, convection and reaction. The controlled drug release from the device depends not only on the kinetics within the coating and the tissue, but also on the coupling of the two layers. In order to take this into account, our model fully couples the two distinct layers through flux and permeable interface conditions.

The model has a wide applicability and we point the reader towards some solution methods, noting that simplifications may be made depending on the parameter values in a given system.

2 The Mathematical Model

2.1 Model Set-Up

In a typical DDD, the mass dynamics occurs across a two-layered system composed of: (1) a polymeric coating, acting as a reservoir, where the drug is initially stored, and (2) the biological tissue where the drug is directed, and exerts a therapeutical effect (Fig. 2). Both layers are treated as porous and for the purposes of this paper we assume homogeneous properties in each layer. Layer (1) is typically a planar slab in contact with layer (2) on one side and with an impermeable backing on the other side. At the interface between the two layers, a rate-controlling membrane (*topcoat*) may exist. Due to the impermeable backing of the coating, drug is directed towards the interface where it then transported to the biological tissue. Mathematically, since most of the drug transport occurs normal to the tissue surface, we may reasonably simplify the geometry to that of a one-dimensional system. Let us define the x-axis to be normal to the surface of the coating and oriented with the positive direction outwards. We locate the interface at $x = 0$, with the coating of thickness l_1 occupying $[-l_1, 0)$ and the tissue of thickness l_2 occupying $(0, l_2]$. Typically the coating is considerably thinner than the biological tissue such that $l_2 \gg l_1$ (Fig. 2).

In what follows we adopt a continuum mechanics approach, treating the porous media as homogenous materials with variables averaged over the representative elementary volume, V. This is taken to be larger that the pore scale, but smaller than the typical length scale of the phenomenon, and comprises the void volume, V^f, and the solid volume, V^s, such that $V = V^f + V^s$. We define porosity, ϕ, as the ratio of void volume to total volume. To take account of situations where not all of the void space is accessible to solute, we introduce the partition coefficient, k, such that $k\phi$ represents the fraction of the void volume that is available to solute

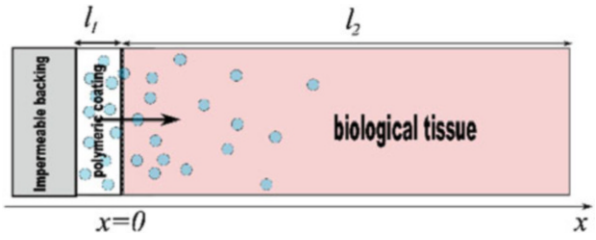


Fig. 2 Schematic of a typical DDD. The drug is initially contained in the coating layer (of thickness l_1) adjacent to the impermeable backing. After dissolving from its solid form to liquid form, drug diffuses through the coating before being transported to the adjoining tissue layer (of thickness l_2), driven by the concentration gradient across the interface (located at $x = 0$)

transport. It is convenient to introduce Φ as the ratio of accessible void volume to solid volume:

$$\Phi = \frac{k\phi}{1-\phi}. \tag{1}$$

In the literature it is common for concentrations to be defined to as either volume-averaged or *intrinsic* volume-averaged. The latter is based on the volume of each phase contained in V, that is kV^f for the accessible fluid-phase[1] and V^s for the solid-phase, while the former is based on averaging over the whole V [13]. The *intrinsic* volume-averaged fluid and solid phase drug concentrations, c^f and c^s (μg/ml), are related to the fluid and solid phase volume-averaged drug concentrations, c and c^* via:

$$c = k\phi c^f, \qquad c^* = (1-\phi)c^s. \tag{2}$$

2.2 Equations of Drug Transport in the Coating

Before implantation, the drug resides in the polymer coating in a biologically unavailable solid form (c_1^s). The drug must undergo a solid–liquid mass transfer process in order to become available for diffusion out of the coating and into the tissue. The solid–liquid mass transfer process (dissolution) is initiated by the ingress of biological fluid into the device. We assume that the rate of transfer of drug from the solid phase to the biologically available free phase (c_1^f) is proportional to the difference between c_1^s and c_1^f. Making the further assumption that the diffusion of drug in the solid phase is negligible, it can be shown that the equations for the drug transport in layer (1) in terms of volume averaged concentrations (2) are:

$$\frac{\partial c_1^*}{\partial t} = -\delta_1\left(\Phi_1 c_1^* - c_1\right) = -\beta_1 c_1^* + \delta_1 c_1 \qquad \text{in } (-l_1, 0) \tag{3}$$

$$\frac{\partial c_1}{\partial t} = D_1\frac{\partial^2 c_1}{\partial x^2} + \delta_1\left(\Phi_1 c_1^* - c_1\right) = D_1\frac{\partial^2 c_1}{\partial x^2} + \beta_1 c_1^* - \delta_1 c_1 \qquad \text{in } (-l_1, 0) \tag{4}$$

The parameter D_1 (cm^2 s^{-1}) is the effective diffusion coefficient of the solute, δ_1 (s^{-1}) is the solid–liquid transfer rate parameter [13] and $\beta_1 = \delta_1\Phi_1$.

[1] Superscripts s and f denote solid and fluid phases, respectively. Subscripts 1 and 2 indicate layers (1) and (2) respectively.

2.3 Equations of Drug Transport in the Tissue

The biological tissue is typically comprised of several layers of different thickness and composition. In what follows, we will consider tissue as a homogeneous single-layered medium. However, the model may be extended to multi-layers following the approach of [14]. Within the tissue the free drug undergoes diffusion and, in many cases, convection due to a pressure difference across the tissue. The drug may also bind to components within the tissue (*association*) and either be metabolised or subsequently unbind (*dissociation*). We model the binding/unbinding process as a first order reaction, similarly to the solid–liquid mass transfer in (3) and (4). Differently from layer (1), we allow for different rates between the forward and reverse reactions, say β_2 and $\delta_2 \geq 0$ (s^{-1}). The magnitude of these reaction parameters can be evaluated experimentally through the equilibrium dissociation constant δ_2/β_2 . Within the tissue, we denote the free and bound drug concentrations as c_2^f and c_2^s, respectively. Thus, in the biological tissue, the drug transport is governed by the following coupled linear reaction-convection-diffusion equations in terms of volume averaged concentrations (2):

$$\frac{\partial c_2}{\partial t} = D_2 \frac{\partial^2 c_2}{\partial x^2} - v \frac{\partial c_2}{\partial x} - \beta_2 c_2 + \delta_2 c_2^* \qquad \text{in } (0, l_2) \tag{5}$$

$$\frac{\partial c_2^*}{\partial t} = \beta_2 c_2 - \delta_2 c_2^* \qquad \text{in } (0, l_2) \tag{6}$$

Here v ($\mathrm{cm\,s}^{-1}$) is the magnitude of the convection while D_2 is the effective diffusivity of unbound drug. The above parameter δ_2 includes the effects of porosity (i.e. it is the unbinding rate multiplied by Φ_2), making clear the analogy with Eqs. (3) and (4). We have also assumed that the drug does not diffuse within the components to which it is bound. We note that the first order linear reaction model of the binding/unbinding process in (5) and (6) may not be the most suitable in all circumstances. In some DDD, such as in drug-eluting stents, a second-order saturable reversible binding model has been proposed to describe the binding of limus compound drugs to arterial tissue [6]: this comprehensive model includes a number of drug dependent parameters which are difficult to measure experimentally and, nevertheless, does not necessarily apply in all DDD. Even in cases where a non-linear model is generally more appropriate, the linear model with suitably chosen parameter values can be shown to suffice in certain circumstances (i.e. at early times and for sufficiently high initial drug concentrations and binding site density). Looking more closely at Eqs. (3), (4) and (5), (6) we note that modelling the full delivery process can be viewed as a set of direct-reverse reactions (*local mass non-equilibrium model*) with different coefficients: the drug dissolution-release-absorption process starts from the coating and ends at the tissue receptors, with a bidirectional phase changes in a cascade sequence as schematically represented in Fig. 3.

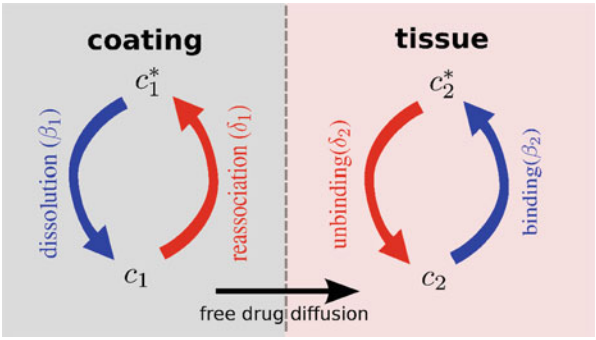


Fig. 3 A diagram sketching the cascade mechanism of drug delivery in the coating-tissue coupled system

2.4 Initial, Boundary and Interface Conditions

We assume that the drug is initially contained wholly in the solid phase at some uniform concentration C_e, leading to the following initial conditions:

$$c_1^*(x,0) = C_e \qquad c_1(x,0) = 0 \qquad c_2(x,0) = 0 \qquad c_2^*(x,0) = 0.$$

At the boundary between the coating and the impermeable backing we impose a zero flux condition, while at the external tissue boundary we propose a general Robin boundary condition:

$$D_1 \frac{\partial c_1}{\partial x} = 0 \qquad \text{at } x = -l_1,$$

$$-D_2 \frac{\partial c_2}{\partial x} + v c_2 = \gamma c_2 \qquad \text{at } x = l_2,$$

where γ is a constant. We note that the two limit cases of zero flux and infinite sink conditions may be recovered by choosing γ sufficiently small or large, respectively. Finally, we must impose two appropriate conditions at the interface to ensure the coupling between the two layers. We impose continuity of flux

$$-D_1 \frac{\partial c_1}{\partial x} = -D_2 \frac{\partial c_2}{\partial x} + v c_2 \qquad \text{at } x = 0.$$

and in addition we allow for a concentration jump across the interface:

$$-D_2 \frac{\partial c_2}{\partial x} = P\left(\frac{c_1}{k_1 \phi_1} - \frac{c_2}{k_2 \phi_2}\right) \qquad \text{at } x = 0,$$

with P (cm s^{-1}) the overall mass transfer coefficient [14].

3 Model Solution

The model we have presented is very general and may be applied to several DDD. It is often useful to write a model in non-dimensional form. A typical non-dimensionalization for a system of reaction-diffusion-convection equations leads to three important numbers: the Péclet number, the first Damköhler number and the second Damköhler number. These dimensionless groups define, respectively, the relative importance of convection to diffusion, of reaction to convection, and of reaction to diffusion. By examining their size, it is often possible to simplify the model by neglecting parameters that are unimportant. For example, in the case of drug-eluting stents, while there exists a small convective flow due to the transmural pressure gradient across the arterial wall, the Péclet number is often small, meaning that the convective term can be reasonably neglected. Depending on the particular DDD, it may also be possible to neglect the solid–liquid mass transfer terms if the timescale for this process is far shorter than that of diffusion in the polymer. Similarly, if the timescale for reaction is far shorter than that of diffusion and convection, then the reaction may be considered instantaneous, in which case the two phases in the tissue can be assumed to be in dynamic equilibrium. When the model has been simplified as far as can be, an analytical or semi-analytical solution may be obtained by using the techniques described in [14–16], or alternatively, an appropriate numerical procedure can be used. We aim to fully develop a solution method for a specific example of a DDD in a forthcoming work [17].

4 Parameter Estimation

One of the great difficulties in modelling biological systems is in obtaining reliable estimates of the various parameters governing the system. In many cases it is not possible to accurately measure the parameter in question, while in the cases of those that can be obtained with any degree of certainty, there are often significant inter-species and inter-sample variations. The advantage of mathematical modelling is that several different parameter sets can be tested, and the effects on the solution compared, thereby reducing the need for costly experiments that often involve animals. Not only is mathematical modelling useful for analysing the effect of parameter variation, but when coupled with controlled in vitro experiments, mathematical models can provide a cheap and ethical way of parameter estimation. For example, the mass of drug released from a DDD placed in an insulated release medium can be measured at various time points and compared with the solution of the equations in layer 1 with insulated boundary conditions (after integrating over the spatial domain to convert concentration to mass). By way of an inverse problem, the diffusion coefficient, D_1 and the solid–liquid mass transfer coefficient, δ_1 may be estimated. A similar procedure can be repeated to estimate the various parameters

in the biological tissue, by placing the tissue sample in a bath of release medium at a given initial drug concentration. A more extensive analysis along with several simulations of the present model is currently being developed [17].

5 Summary

In this paper we have presented a general local mass non-equilibrium model of drug release from a DDD and the subsequent drug transport in biological tissue. The model is based on a two-layer two-phase linear system of partial differential equations describing both the solid–liquid transfer and diffusion processes in the polymeric layer as well as diffusion, convection and reaction in the tissue layer. The analytical approach helps to identify and quantify the relevant concurrent phenomena in DDD and is useful for experimental design and clinical applications, providing the basis for the optimization of parameters. The model contains several parameters that need to be identified before it can be used in a predictive way and provide the significant kinetics. Having done that, the proposed model can be tuned and used to quantitatively characterize the drug delivery, showing how the release can be suited to the clinical requirements needed for therapeutical purposes.

Acknowledgement The study was partly funded by the MIUR-CNR project "Interomics", 2014.

References

1. Siepmann, J. Siepmann, F.: Mathematical modeling of drug delivery. Int. J. Pharm. **364**(2), 328–343 (2008)
2. McGinty, S., McKee, S., Wadsworth, R.M., McCormick, C.: Modelling drug-eluting stents. Math. Med. Biol. **28**, 1–29 (2011)
3. Zhang, W.E., Prausnitz, M.R., Edwards, A.U.: Model of transient drug diffusion across cornea. J. Control. Release **99**, 241–258 (2004)
4. Prausnitz, M.R., Langer, R.: Transdermal drug delivery. Nat. Biotechnol. **26**(11), 1261–1268 (2008)
5. Langer, R.: Polymeric delivery systems for controlled drug release. Chem. Eng. Commun. **6**, 1–48 (1980)
6. Tzafriri, A.R., Groothuis, A., Price, G.S., Edelman E.R.: Stent elution rate determines drug deposition and receptor-mediated effects. J. Control. Release **161**, 918–926 (2012)
7. Creel, C.J., Lovich, M.A., Edelman, E.R.: Arterial paclitaxel distribution and deposition. Circ. Res. **86**(8), 879–884 (2000)
8. Siepmann, J., Göpferich, A.: Mathematical modeling of bioerodible, polymeric drug delivery systems. Adv. Drug Deliv. Rev. **48**, 229–247 (2001)
9. Crane, M., Hurley, N.J., Crane, L., Healy, A.M., Corrigan, O.I., Gallagher, K.M., McCarthy, L.G.: Simulation of the USP drug delivery problem using CFD: experimental, numerical and mathematical aspects. Simul. Model. Pract. Theory **12**(2), 147–158 (2004)
10. Narasimhan, B.: Mathematical models describing polymer dissolution: consequences for drug delivery. Adv. Drug Deliv. Rev. **48**, 195–210 (2001)

11. Biscari, P., Minisini, S., Pierotti, D., et al.: Controlled release with finite dissolution rate. SIAM J. Appl. Math. **71**(3), 731–752 (2011)
12. Bozsak, F., Chomaz, J.M., Barakat, A.I.: Modeling the transport of drugs eluted from stents: physical phenomena driving drug distribution in the arterial wall. Biomech. Model. Mechanobiol. **13**(2), 327–347 (2014)
13. de Monte, F., Pontrelli, G., Becker, S.M.: Drug release in biological tissues. In: Becker, S.M., Kuznetsov, A.V. (eds.) Transport in Biological Media, Chap. 3, pp. 59–118. Elsevier, New York (2013)
14. Pontrelli, G., de Monte, F.: A multi-layer porous wall model for coronary drug-eluting stents. Int. J. Heat Mass Transf. **53**, 3629–3637 (2010)
15. Pontrelli, G., Di Mascio, A., de Monte, F.: Local mass non-equilibrium dynamics in multi-layered porous media: application to the drug-eluting stents. Int. J. Heat Mass Transf. **66**, 844–854 (2013)
16. McGinty, S., McKee, S., Wadsworth, R.M., McCormick, C.: Modeling arterial wall drug concentrations following the insertion of a drug-eluting stent. SIAM J. Appl. Math. **73**(6), 2004–2028 (2014)
17. Sean, M., Giuseppe, P.: On the influence of solid-liquid mass transfer in the modelling of drug release from stents. J. Coupled Syst. Multiscale Dyn. **3**(1), 47–56 (2015)

MS 13
MINISYMPOSIUM: MATHEMATICAL PROBLEMS FROM SEMICONDUCTOR INDUSTRY

Organizers

Giuseppe Alì[1] and Giovanni Mascali[2]

Speakers

Paolo G. Ferrandi[3], Stefano Micheletti[4] and Paolo Simioni[5]
Charge Transport in OLETs: Mathematical Model and Numerical Simulations

Bratislav Tasic[6], Jos J. Dohmen[7], E. Jan W. ter Maten[8], Theo G.J. Beelen[9] and Wil H.A. Schilders[10]

[1]Giuseppe Alì, Department of Physics, University of Calabria, Arcavacata di Rende, Italy.

[2]Giovanni Mascali, Department of Mathematics and Computer Science, University of Calabria, Arcavacata di Rende, Italy.

[3]Paolo G. Ferrandi, MOX Laboratorio di Modellistica e Calcolo Scientifico, Politecnico di Milano, Italy.

[4]Stefano Micheletti, Dipartimento di Matematica, Politecnico di Milano, Italy.

[5]Paolo Simioni, MOXOFF Mathematics for innovation, Politecnico di Milano, Italy.

[6]Bratislav Tasic, NXP Semiconductors, Eindhoven, The Netherlands.

[7]Jos J. Dohmen, NXP Semiconductors, Eindhoven, The Netherlands.

[8]E. Jan W. ter Maten, Department of Mathematics and Computer Science, Technische Universiteit, Eindhoven, The Netherlands.

[9]Theo G.J. Beelen, NXP Semiconductors, Eindhoven, The Netherlands.

[10]Wil H.A. Schilders, Department of Mathematics and Computer Science, Technische Universiteit, Eindhoven, The Netherlands.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_48

Fast Fault Simulation to Identify Subcircuits Involving Faulty Components

Giuseppe Alì[1], Nicodemus Banagaaya[11] and Wil H.A. Schilders[10]
Index-Aware Model Order Reduction Methods for Electrical Networks

Alfio Borzì[12] and Gabriele Ciaramella[13]
Newton Methods for the Optimal Control of Closed Quantum Spin Systems

Michael Günther[14], Christof Hatchel[15] and Adrian Sandu[16]
Multirate GARK Schemes for Coupled Problems

Roland Pulch[17], Andreas Bartel[18] and Sebastian Schöps[19]
Quadrature Methods with Adjusted Grids for Stochastic Models of Coupled Problems

Keywords

Micro- and nano-electronics
Semiconductor industry

Short Description

The minisymposium collects contributions related to problems arising in the context of coupled modeling, simulation and optimization in micro and nano-electronics. The objective is to present the latest developments, insights, methods and ideas in the above areas of research, and indications for future research directions.

[11]Nicodemus Banagaaya, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

[12]Alfio Borzì, Institut für Mathematik, Universität Würzburg, Germany.

[13]Gabriele Ciaramella, Institut für Mathematik, Universität Würzburg, Germany.

[14]Michael Günther, Institute of Mathematical Modelling, Analysis and Computational Mathematics, Bergische Universität Wuppertal, Germany.

[15]Christof Hatchel, Institute of Mathematical Modelling, Analysis and Computational Mathematics, Bergische Universität Wuppertal, Germany.

[16]Adrian Sandu, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, USA.

[17]Roland Pulch, Institut für Mathematik und Informatik, Greifswald, Germany.

[18]Andreas Bartel, Institute of Mathematical Modelling, Analysis and Computational Mathematics, Bergische Universität Wuppertal, Germany.

[19]Sebastian Schöps, Computational Electromagnetics Laboratory, TU Darmstadt, Germany.

The minisymposium also involves researchers working for industries, which have provided a more timely indication of the most relevant up-to-date problems and technique encountered in real industrial environments.

Fast Fault Simulation to Identify Subcircuits Involving Faulty Components

B. Tasić, J.J. Dohmen, E.J.W. ter Maten, T.G.J. Beelen, H.H.J.M. Janssen, W.H.A. Schilders, and M. Günther

Abstract Imperfections in manufacturing processes may cause unwanted connections (faults) that are added to the nominal, "golden", design of an electronic circuit. By fault simulation we simulate all situations: new connections and each with different values for the newly added element. We also consider "opens" (broken connections). During the transient simulation the solution of a faulty circuit is compared to the golden solution of the fault-free circuit. A strategy is developed to efficiently simulate the faulty solutions until their moment of detection. We fully exploit the hierarchical structure of the circuit in the simulation process to bypass parts of the circuit that appear to be unaffected by the fault. Accurate prediction and efficient solution procedures lead to fast fault simulation in which the golden solution and all faulty solutions are calculated over a same time step. Finally, we store a database with detectable deviations for each fault. If such a detectable output

B. Tasić (✉) • J.J. Dohmen
NXP Semiconductors, High Tech Campus 46, 5656 AE Eindhoven, The Netherlands
e-mail: Bratislav.Tasic@nxp.com; Jos.J.Dohmen@nxp.com

E.J.W. ter Maten
Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Bergische Universität Wuppertal, Gaußstraße 20, D-42219 Wuppertal, Germany
e-mail: Jan.ter.Maten@math.uni-wuppertal.de

T.G.J. Beelen
NXP Semiconductors, High Tech Campus 46, 5656 AE Eindhoven, The Netherlands

Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: Theo.G.J.Beelen@nxp.com

H.H.J.M. Janssen
NXP Semiconductors, High Tech Campus 46, 5656 AE Eindhoven, The Netherlands
e-mail: Rick.Janssen@nxp.com

W.H.A. Schilders
Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: W.H.A.Schilders@tue.nl

M. Günther
Bergische Universität Wuppertal, Gaußstraße 20, D-42219 Wuppertal, Germany
e-mail: Michael.Guenther@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_49

"matches" a measurement result of a product that has been returned because of malfunctioning it helps to identify the subcircuit that may contain the real fault.

Keywords Electronic circuit • Fault simulation • Micro-and nano-electronics

1 Time Integration of Circuit Equations

The electronic circuit equations can be written as [4, 9]

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}(\mathbf{x}) + \mathbf{j}(\mathbf{x}) = \mathbf{s}(\mathbf{x}, t). \tag{1}$$

Here $\mathbf{s}(\mathbf{x}, t)$ represents the specifications of the sources. The unknown $\mathbf{x} = \mathbf{x}(t)$ consists of nodal voltages and of currents through voltage defined elements. We assume that $\mathbf{q}(\mathbf{0}) = \mathbf{0}$, and $\mathbf{j}(\mathbf{0}) = \mathbf{0}$.

For time integration in circuit simulation we consider the BDF1, or Euler Backward method. Assuming time points $t_{k+1} = t_k + h_k$ ($k \geq 0$) with stepsizes h_k and approximation $\mathbf{x}^n$ at t_n, BDF1 calculates $\mathbf{x}^{n+1}$ by

$$\frac{\mathbf{q}^{n+1} - \mathbf{q}^n}{h_n} + \mathbf{j}^{n+1} = \mathbf{s}^{n+1}. \tag{2}$$

Here $\mathbf{q}^k = \mathbf{q}(\mathbf{x}^k)$, $\mathbf{j}^k = \mathbf{j}(\mathbf{x}^k)$, for $k = n, n+1$, and $\mathbf{s}^{n+1} = \mathbf{s}(\mathbf{x}^{n+1}, t_{n+1})$. The system is solved by a Newton-Raphson procedure. For efficient direct methods to solve the intermediate linear systems, see [1]. A fixed Jacobian can reduce the number of LU-decompositions, but, in general, will increase the number of iterations and thus the number of (costly) evaluations. Also, in case of circuit simulation, the assembly of the matrices does not need much more effort when compared to the function evaluations. A fixed decomposed Jacobian can be efficient within some Picard-iteration [8] in solving a linear system, or, more general, in using it as preconditioner within GMRES. When changing stepsizes during time integration similar remarks apply.

In case of an hierarchical linear solver one can profit from hierarchical bypassing [3], which we will also exploit in this paper. When applying it also in the time integration, it even supports a first form of multirate time-integration [13].

2 Fault Simulation

We first consider the effect of adding faulty, linear elements to the circuit. F.i., in [2, 11] we did add linear bridges (resistors) to the circuit. For each fault only one element is added to the original, golden circuit. It may mean a new connection, while also different values are considered. In [11] a novel time-integration was involved:

during each time-step, first the fault-less, golden solution was determined at the next time step. Next, all faulty problems were integrated over this time-interval. Hence, effectively, a parameter loop is placed inside the time integration. Also the hierarchical structure was enhanced such that the hierarchical solver could deal with all new elements. This enables to exploit an enhanced form of bypassing.

The golden solution at each new time point provides an estimate for the solution of a faulty problem (in addition to the one using extrapolation by Nordsieck vectors). Each faulty problem uses the stepsize of the golden solution as a maximum one. When the faulty solution really needs a stepsize that is significantly smaller then used by the golden solution, the traditional time integration is invoked, even without bypassing, until the time moment of synchronization with the golden solution.

In this paper we enhance the algorithm in also considering the case of adding linear capacitors. However, in practise, the linear resistor case is by far more important. Hence, we either have[1]

$$\mathbf{j}(\mathbf{x}(t,p),p) = \mathbf{j}_0(\mathbf{x}(t,p)) + p\,\mathbf{u}\mathbf{v}^T\mathbf{x}(t,p), \quad \text{or} \tag{3}$$

$$\mathbf{q}(\mathbf{x}(t,p),p) = \mathbf{q}_0(\mathbf{x}(t,p)) + p\,\mathbf{a}\mathbf{b}^T\mathbf{x}(t,p). \tag{4}$$

For simplicity, to reduce notation and the amount of partial derivatives further on, we use p, both in (3) and in (4). Fault Analysis consists of simulations for a large number of pairs of vectors $(\mathbf{u},\mathbf{v})$, or $(\mathbf{a},\mathbf{b})$ and various values of p, and compare the result $\mathbf{x}_p(t)$ of (1) at specific time points with the "golden" solution $\mathbf{x}(t)$ of the fault-free circuit (corresponding with $p = 0$). If the deviation exceeds some threshold, the fault triple $(\mathbf{u},\mathbf{v},p)$, or $(\mathbf{a},\mathbf{b},p)$, is marked as detectable and is taken out of the list.

Clearly, for each fault we have a new contribution $p\,\mathbf{u}\mathbf{v}^T\mathbf{x}(t,p)$, or $p\,\mathbf{a}\mathbf{b}^T\mathbf{x}(t,p)$, as low-rank modification to the system of the golden solution, either added to $\mathbf{j}_0$ or to $\mathbf{q}_0$, see (3)–(4). Here $p \geq 0$ is just a scalar, by which the $\mathbf{p}$-sensitivity 'matrix' $\hat{\mathbf{x}}_p(t,p) = \frac{\partial \mathbf{x}(t,p)}{\partial p}$ reduces to a vector.

The golden solution $\mathbf{x}(t)$ used $\mathbf{j}(\mathbf{x}(t,p),p) = \mathbf{j}_0(\mathbf{x}(t,p))$, and $\mathbf{q}(\mathbf{x}(t,p),p) = \mathbf{q}_0(\mathbf{x}(t,p))$.

Let $\mathbf{x}_p^k = \mathbf{x}^k(p) \approx \mathbf{x}(t_k,p)$ be the numerical approximations for $k = n, n+1$ of the faulty system and $\hat{\mathbf{x}}_p^k$ be the corresponding sensitivities. Then with $\mathbf{C}_p^k \equiv \frac{\partial \mathbf{q}(\mathbf{x}_p^k)}{\partial \mathbf{x}}$, $\mathbf{G}_p^k \equiv \frac{\partial \mathbf{j}(\mathbf{x}_p^k)}{\partial \mathbf{x}}$ (and including the effect of the rank-one term with the factor p) and $\mathbf{S}_p^k \equiv \frac{\partial \mathbf{s}(\mathbf{x}_p^k,t_k)}{\partial \mathbf{x}}$, by sensitivity analysis [6, 10], we obtain

$$[\frac{1}{h_n}\mathbf{C}_p^{n+1} + \mathbf{G}_p^{n+1} - \mathbf{S}_p^{n+1}]\hat{\mathbf{x}}_p^{n+1} = -\frac{1}{h_n}\mathbf{a}\mathbf{b}^T(\mathbf{x}^{n+1} - \mathbf{x}^n) - \mathbf{u}\mathbf{v}^T\mathbf{x}_p^{n+1} + \frac{1}{h_n}\mathbf{C}_p^n\hat{\mathbf{x}}_p^n. \tag{5}$$

[1] Note that for inductors and for voltage-defined resistors we need two rank-one updates to describe the total contribution.

For $p = 0$, (5) gives the limit sensitivity $\hat{\mathbf{x}}^k = \hat{\mathbf{x}}_0^k$ for the golden, fault-free, solution $\mathbf{x}^k = \mathbf{x}_0^k$ $(k = n, n+1)$

$$[\frac{1}{h_n}\mathbf{C}^{n+1} + \mathbf{G}^{n+1} - \mathbf{S}^{n+1}]\hat{\mathbf{x}}^{n+1} = -\frac{1}{h_n}\mathbf{a}\mathbf{b}^T(\mathbf{x}^{n+1} - \mathbf{x}^n) - \mathbf{u}\mathbf{v}^T\mathbf{x}^{n+1} + \frac{1}{h_n}\mathbf{C}^n\hat{\mathbf{x}}^n, \tag{6}$$

where $\mathbf{C}^k = \mathbf{C}_0^k$ $(k = n, n+1)$, $\mathbf{G}^{n+1} = \mathbf{G}_0^{n+1}$ and $\mathbf{S}^{n+1} = \mathbf{S}_0^{n+1}$. By Taylor expansion we additionally have

$$\mathbf{x}_p^k = \mathbf{x}^k + p\,\hat{\mathbf{x}}^k + \mathscr{O}(p^2) \quad (k = n, n+1). \tag{7}$$

The golden solution satisfies the linearized equations of the fault-free circuit up to a term $\mathbf{R}$ that indicates the deviation from linearity (note that in (1) we did assume that $\mathbf{q}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{j}(\mathbf{0}) = \mathbf{0}$)

$$[\frac{1}{h_n}\mathbf{C}^{n+1} + \mathbf{G}^{n+1} - \mathbf{S}^{n+1}]\mathbf{x}^{n+1} = \mathbf{r}(t_{n+1}, \mathbf{x}^n, \mathbf{x}^{n+1}), \tag{8}$$

where $\mathbf{r}(t_{n+1}, \mathbf{x}^n, \mathbf{x}^{n+1}) = \mathbf{s}^{n+1} + \frac{1}{h_n}\mathbf{C}^n\mathbf{x}^n + \mathbf{R}$. With (7) and (6) this gives

$$\begin{aligned}
&[\frac{1}{h_n}\mathbf{C}^{n+1} + \mathbf{G}^{n+1} - \mathbf{S}^{n+1}]\mathbf{x}_p^{n+1} \;= \\
&= \;[\frac{1}{h_n}\mathbf{C}^{n+1} + \mathbf{G}^{n+1} - \mathbf{S}^{n+1}]\mathbf{x}^{n+1} + p\,[\frac{1}{h_n}\mathbf{C}^{n+1} + \mathbf{G}^{n+1} - \mathbf{S}^{n+1}]\hat{\mathbf{x}}^{n+1} + \mathscr{O}(\ldots), \\
&= \;\mathbf{r}(t_{n+1}, \mathbf{x}^n, \mathbf{x}^{n+1}) - \frac{p}{h_n}\mathbf{a}\mathbf{b}^T(\mathbf{x}^{n+1} - \mathbf{x}^n) - p\mathbf{u}\mathbf{v}^T\mathbf{x}^{n+1} + \frac{p}{h_n}\mathbf{C}^n\hat{\mathbf{x}}^n + \mathscr{O}(\ldots), \\
&= -\frac{p}{h_n}\mathbf{a}\mathbf{b}^T(\mathbf{x}^{n+1} - \mathbf{x}^n) - p\mathbf{u}\mathbf{v}^T\mathbf{x}^{n+1} + \frac{1}{h_n}\mathbf{C}^n(p\hat{\mathbf{x}}^n) + \mathbf{r}(t_{n+1}, \mathbf{x}^n, \mathbf{x}^{n+1}) + \mathscr{O}(\ldots), \\
&= -\frac{p}{h_n}\mathbf{a}\mathbf{b}^T(\mathbf{x}_p^{n+1} - \mathbf{x}^n) - p\mathbf{u}\mathbf{v}^T\mathbf{x}_p^{n+1} + \frac{1}{h_n}\mathbf{C}^n(\mathbf{x}_p^n - \mathbf{x}^n) + \mathbf{r}(t_{n+1}, \mathbf{x}^n, \mathbf{x}^{n+1}) + \mathscr{O}(\ldots),
\end{aligned}$$

in which all $\mathscr{O}(\ldots)$ terms are of the form $\mathscr{O}(p^2 + \frac{p^2}{h_n})$. Hence

$$\begin{aligned}
&[\big(\frac{1}{h_n}\mathbf{C}^{n+1} + \mathbf{G}^{n+1} - \mathbf{S}^{n+1}\big) + \frac{p}{h_n}\,\mathbf{a}\mathbf{b}^T + p\,\mathbf{u}\mathbf{v}^T]\mathbf{x}_p^{n+1} \;= \\
&= \;\frac{p}{h_n}\mathbf{a}\mathbf{b}^T\mathbf{x}^n + \frac{1}{h_n}\mathbf{C}^n(\mathbf{x}_p^n - \mathbf{x}^n) + \mathbf{r}(t_{n+1}, \mathbf{x}^n, \mathbf{x}^{n+1}) + \mathscr{O}(p^2 + \frac{p^2}{h_n}), \qquad (9) \\
&= \;\mathbf{r}(t_{n+1}, \mathbf{x}^n, \mathbf{x}^{n+1}) + \mathscr{O}(p^2 + \frac{p^2}{h_n} + \frac{p}{h_n}). \qquad (10)
\end{aligned}$$

Note that (9) may be a more accurate alternative then (10). However, for simplicity, we just used (10), after ignoring the $\mathcal{O}(\cdot)$ terms at the right-hand side. This invites for applying the Sherman-Morrison formula [5]. Let $\mathbf{A} = \left(\frac{1}{h_n}\mathbf{C}^{n+1} + \mathbf{G}^{n+1} - \mathbf{S}^{n+1}\right)$, and $\mathbf{Aw} = p\mathbf{u}$, $\mathbf{Ac} = \frac{p}{h_n}\mathbf{a}$, and $\mathbf{Ay} = \frac{p}{h_n}\mathbf{C}^n\hat{\mathbf{x}}^n$. Then the sensitivity predictions for $\mathbf{x}_p^{n+1}$ become

$$\text{for (3)}: \ \mathbf{x}_p^{n+1} = \mathbf{x}^{n+1} - \frac{\mathbf{v}^T\mathbf{x}^{n+1}}{1 + \mathbf{v}^T\mathbf{w}}\,\mathbf{w}, \quad \text{or} \tag{11}$$

$$\mathbf{x}_p^{n+1} = (\mathbf{x}^{n+1} + \mathbf{y}) - \frac{\mathbf{v}^T(\mathbf{x}^{n+1} + \mathbf{y})}{1 + \mathbf{v}^T\mathbf{w}}\,\mathbf{w}, \tag{12}$$

$$\text{for (4)}: \ \mathbf{x}_p^{n+1} = \mathbf{x}^{n+1} - \frac{\mathbf{b}^T\mathbf{x}^{n+1}}{1 + \mathbf{b}^T\mathbf{c}}\,\mathbf{c}, \quad \text{or}, \tag{13}$$

$$\mathbf{x}_p^{n+1} = (\mathbf{x}^{n+1} + \mathbf{y}) - \frac{\mathbf{b}^T\mathbf{y}}{1 + \mathbf{b}^T\mathbf{c}}\,\mathbf{c}. \tag{14}$$

Note that the first term in (9) has a simplifying effect in (14) (when compared to (12)). In this case one really needs $\mathbf{y}$ to get a first estimate that is different from $\mathbf{x}^{n+1}$. The advantage of the right-hand side in (10) is that it is independent of the solution $\mathbf{x}_p^k$ at the previous time steps. Of course, when followed by further Newton-Raphson iterations, $\mathbf{x}_p^n$ is still needed. To judge the accuracy of the linear sensitivity prediction the nonlinear solver evaluates the circuit at the sensitivity solution and updates the solution. The difference in the initial sensitivity solution and the nonlinear update is a measure for the truncation error.

If we just stick to the prediction, we may calculate the prediction of the fault at a few selected time points, which significantly reduces the work load for the fault sensitivity analysis. We finally remark that in (11), (13) the sensitivity matrix $\hat{\mathbf{x}}^n$ is not explicitly calculated.

2.1 *Modeling Faulty "Opens"*

Next we consider a faulty resistor, with value R, in series with another, linear resistor, with value r. Clearly, this introduces an extra node n_e. If the golden system used $R(n_1, n_2) = r$, the faulty system uses $R(n_1, n_e) = R$, $R(n_e, n_2) = r$. The voltage at this new node can be simply eliminated by noting that $v(n_e) = (r\,v(n_1) + R\,v(n_2))/(r + R)$. Doing this directly, the remaining system can be formulated as in (3) in which $p = R/(r(R + r))$. If $R \to \infty$ we obtain an "open" between the nodes n_1 and n_e and $v(n_e) \to v(n_2)$. In [11] we did introduce an extra port to model bridges between models. This extra node can also become functional in providing the extra node.

For modeling a broken joint (or weld) at a node n, it is, mathematically, convenient to first split the node n into two nodes n_1 and n_2, with a simple voltage source in

between for the golden circuit: $E(n_1, n_2) = 0$. Clearly this satisfies our assumption $\mathbf{j}(\mathbf{0}) = \mathbf{0}$. The faulty system uses $R(n_1, n_e) = R$, $E(n_e, n_2) = 0$. We assume local coordinates that correspond with $v(n_1)$, $v(n_2)$, $i(E)$. We deduce that the faulty system perturbs the golden system with $\mathbf{u}_1\mathbf{v}_1^T(R) + \mathbf{u}_2\mathbf{v}_2^T(R)$, in which, in local coordinates, $\mathbf{u}_1^T = (1, 0, 0)$, $\mathbf{v}_1^T = (-1/R, 1/R, 1)$, $\mathbf{u}_2^T = (0, 0, 1)$, $\mathbf{v}_2^T = (1, -1, R)$. This can be treated in a similar way as before.

3 Results

The FFS-algorithm has been implemented in Pstar.[2] For fault simulations in DC-simulations a significant speed-up (> 100) was obtained by exploiting bypassing and abandoning only, but inclusion of sensitivity analysis appeared essential to get significant speed up for a broad class of problems during transient simulation. Table 1 shows the speed-up by including sensitivity prediction for a LIN Converter IP Block (first part), as well as for a nonlinear control DAC (2nd part). Clearly, the linear sensitivity estimate offers an interesting speed-up. Following nonlinear corrections do reduce this effect. For the LIN Converter IP Block the effect of more iterations remains quite bounded (with 100 iterations still a speed-up of more than 10 was found, see [11]). For the nonlinear control DAC until 5 iterations a speed-up of 10 was obtained. Further speed-up scenarios are currently considered by initiating the fault later. If one can simply skip the initial integration of the faults until $t_1 > 0$, for a large collection of faults no initial simulations have to be made. The scenarios differ in how the fault is started: suddenly, or using a smooth start-up, similar as for the source-stepping-by-transient method as described in [11]. Because of the many faults that are possible, a short start-up is a balance between efficiency and robustness.

Table 1 Speed-up by including sensitivity prediction

	LIN converter IP block			Control DAC		
Analysis	#iterations Per step Δt	CPU time [s]	Speed up	#iterations Per step Δt	CPU time [s]	Speed up
Standard AS/DOTSS	–	100,437	1	–	52,513	1
Linear sensitivity	0	458	219	0	916	58
Nonlinear correction	5	2341	43	1	4808	11

Left: a LIN Converter IP Block, #faults=412. Right: a Control DAC, #faults=100. See also [11]

[2] Pstar: in-house circuit simulator of NXP Semiconductors.

4 Relation to Uncertainty Quantification

Interchanging the time integration loop with a parameter-sweep loop for a given pair of connection nodes, also has an interesting opportunity for Uncertainty Quantification [7, 12, 14]. F.i., when considering Stochastic Collocation in which all L_2 inner-products in parameter space are replaced by quadrature, a list of deterministic parameter values p_k, $k = 1, \ldots, K$, is defined for which the solution $\mathbf{x}(t, p_k)$ has to be calculated. Then $\mathbf{x}(t, p) = \sum_{i=0}^{m} \mathbf{v}_i(t)\phi_i(p)$, in which $\mathbf{v}_i(t) = \sum_{k=1}^{K} w_k \mathbf{x}(t, p_k)\phi_i(p_k)$. This expansion is a so-called generalized Polynomial Chaos expansion, using polynomials $\phi_i(p)$ that are orthogonal with respect to some probability density function f in the parameter space for p. For FFS, where parameter values are positive, one may think about an exponential decay (for an infinite range; here one generates Laguerre polynomials), or a (α, β)-density function (when considering a finite range for p; here one generates Jacobi-polynomials). Now, first, one can simulate the K deterministic solutions (in which one can exploit the sensitivity estimate, as described before). Next, the actual FFS is done as a post-processing action in which one compares $\mathbf{x}(t, p)$ with $\mathbf{x}(t, 0)$, at specific time moments and at circuit nodes. Note that mean and variance are cheaply provided by the $\mathbf{v}_i(t)$. During the time-integration one also has at each completed time-level t $\mathbf{x}(t, p)$ and $\partial \mathbf{x}(t, p)/\partial p$ available from the expansion.Clearly, FFS is just an example for varying particular parameters. Also more general, Stochastic Collocation, can benefit by moving the parameter loop inside the time integration loop.

Acknowledgements We acknowledge the support from the EU project nanoCOPS, Nanoelectronic COupled Problems Solutions (FP7-ICT-2013-11/619166), http://www.fp7-nanoCOPS.eu/.

References

1. Davis, T.A., Natarajan, E.P.: Sparse matrix methods for circuit simulation problems. In: Michielsen, B., Poirier, J.-R. (eds.) Scientific Computing in Electrical Engineering (SCEE). Series Mathematics in Industry, vol. 16, pp. 3–14. Springer, Berlin (2012)
2. De Jonghe, D., Maricau, E., Gielen, G., McConaghy, T., Tasić, B., Stratigopoulos, H.: Advances in variation-aware modeling, verification, and testing of analog ICs. In: Proceedings of the Design, Automation and Test in Europe (DATE), pp. 1615–1620 (2012)
3. Fijnvandraat, J.G., Houben, S.H.M.J., ter Maten, E.J.W., Peters, J.M.F.: Time domain analog circuit simulation. J. Comput. Appl. Math. **185**, 441–459 (2006)
4. Günther, M., Feldmann, U., ter Maten, J.: Modelling and discretization of circuit problems. In: Schilders, W.H.A., ter Maten, E.J.W. (eds.) Handbook of Numerical Analysis, vol. XIII, Special Volume on Numerical Methods in Electromagnetics, pp. 523–659. Elsevier BV, North-Holland (2005)
5. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn.. The Johns Hopkins University Press, Baltimore (1996)

6. Ilievski, Z., Xu, H., Verhoeven, A., ter Maten, E.J.W., Schilders, W.H.A., Mattheij, R.M.M.: Adjoint transient sensitivity analysis in circuit simulation. In: Ciuprina, G., Ioan, D. (eds.) Scientific Computing in Electrical Engineering (SCEE, 2006). Series Mathematics in Industry, vol. 11, pp. 183–189. Springer, Berlin (2007)
7. Le Maître, O.P., Knio, O.M.: Spectral Methods for Uncertainty Quantification, with Applications to Computational Fluid Dynamics. Springer, Science+Business Media B.V., Dordrecht (2010)
8. Li, Z., Shi, C.-J.R.: SILCA: spice-accurate iterative linear-centric analysis for efficient time-domain simulation of VLSI circuits with strong parasitic couplings. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. (TCAD) **25**(6), 1087–1103 (2006)
9. Ogrodzki, J.: Circuit Simulation Methods and Algorithms. CRC, Boca Raton (1994)
10. Petzold, L., Li, S.T., Cao, Y., Serban, R.: Sensitivity analysis of differential-algebraic equations and partial differential equations. Comput. Chem. Eng. **30**, 1553–1559 (2006)
11. Tasić, B., Dohmen, J.J., ter Maten, E.J.W., Beelen, T.G.J., Schilders, W.H.A., de Vries, A., van Beurden, M.: Robust DC and efficient time-domain fast fault simulation. COMPEL Int. J. Comput. Math. Electr. Electron. Eng. **33**(4), 1161–1174 (2014)
12. ter Maten, E.J.W., Pulch, R., Schilders, W.H.A., Janssen, H.H.J.M.: Efficient calculation of uncertainty quantification. In: Fontes, M., Günther, M., Marheineke, N. (eds.) Progress in Industrial Mathematics at ECMI 2012. Series Mathematics in Industry, vol. 19, pp. 361–370. Springer, New York (2014)
13. Verhoeven, A.: Redundancy reduction of IC models by multirate time integration and model order reduction. Ph.D. Thesis Eindhoven University of Technology (2008). http://alexandria.tue.nl/extra2/200712281.pdf
14. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)

Quadrature Methods with Adjusted Grids for Stochastic Models of Coupled Problems

Roland Pulch, Andreas Bartel, and Sebastian Schöps

Abstract We consider coupled problems with uncertain parameters modelled as random variables. Due to the largely differing behaviour of subsystems in coupled problems, we introduce a strategy of adjusted grids defined in the parameter domain for resolving the stochastic model. This allows us to adapt quadrature grids to each subsystem. The communication between the different grids requires global approximations of coupling variables in the random space. Since implicit time integration methods are typically included, we investigate dynamic iteration schemes to realise this approach. Numerical results for a thermal-electric test circuit outline the feasibility of the method.

Keywords Coupled problems • Stochastic modeling • Thermal-electric circuit • Uncertain parameters

1 Introduction

In many applications, the simulation task addresses a coupled, multiphysical problem. Often the resulting models consist of differential algebraic equations together with partial differential equations, see [1]. Due to an inherent multirate or multiscale behaviour, a co-simulation of a coupled problem can be often efficient.

R. Pulch (✉)
Institute for Mathematics and Computer Science, Ernst-Moritz-Arndt-Universität Greifswald, Walther-Rathenau-Str. 47, D-17489 Greifswald, Germany
e-mail: pulchr@uni-greifswald.de

A. Bartel
Chair of Applied Mathematics and Numerical Analysis, Bergische Universität Wuppertal, Bendahler Str. 29, D-42285 Wuppertal, Germany
e-mail: bartel@math.uni-wuppertal.de

S. Schöps
Graduate School Computational Electromagnetics, Technische Universität Darmstadt, Dolivostr. 15, D-64293 Darmstadt, Germany
e-mail: schoeps@gsc.tu-darmstadt.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*, Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_50

Mathematically, this is also referred to as dynamic iteration, see [3]. Our application is in electrical engineering, where we consider a thermal-electric test circuit.

Physical parameters of a coupled problem may exhibit uncertainties due to measurement errors, imperfections of an industrial production or other reasons. We quantify the uncertainties by random variables for the parameters and the solution becomes a random process. Statistics like the expected value and the variance can be computed by sampling methods or quadrature rules. Alternatively, a stochastic Galerkin method or stochastic collocation techniques can be used, see [4–6]. However, the Galerkin approach results in a much larger coupled system.

We investigate quadrature formulas in this paper. If the parts of the coupled problem show a different sensitivity with respect to the dependence on the random parameters, then the usage of quadrature on grids with different refinement levels becomes favourable. Thus we introduce different grids for the subsystems of a coupled problem. The application of this approach is straightforward in case of an explicit time integration scheme. To realise an implicit time integration, we apply a dynamic iteration to the overall problem, which decouples the subsystems to some extend. It follows that communications between the different parameter grids are required in discrete time points. Arbitrary global approximations of the solution on the random space are feasible for this communication. We use truncated expansions of the solution with respect to orthogonal basis polynomials depending on the random variables, i.e., a spectral approach appears in the probability space, see [6].

Finally, we test this strategy using a problem from [2], where an electric network is combined with thermal effects. Two different grids are applied for the two parts of the coupled problem. We test grids of several resolutions and based on a reference solution we qualitatively compare the achieved accuracies.

2 Problem Definition

We consider a time-dependent coupled problem consisting of two parts

$$\begin{aligned} \mathbf{F}_1\left(\mathbf{y}_1(t,\mathbf{p}), \mathbf{y}_2^{\mathrm{cpl}}(t,\mathbf{p}), t, \mathbf{p}\right) &= \mathbf{0}, \\ \mathbf{F}_2\left(\mathbf{y}_2(t,\mathbf{p}), \mathbf{y}_1^{\mathrm{cpl}}(t,\mathbf{p}), t, \mathbf{p}\right) &= \mathbf{0}, \end{aligned} \tag{1}$$

where parameters $\mathbf{p} \in \Pi \subseteq \mathbb{R}^Q$ are included. The operators $\mathbf{F}_1, \mathbf{F}_2$ represent ordinary differential equations (ODEs), differential algebraic equations (DAEs) or partial differential equations (PDEs) after a semidiscretisation in space. Hence time derivatives are involved in each part. The operators $\mathbf{F}_i$ comprise n_i equations and the solution of the system (1) is $\mathbf{y}_i : [t_0, t_{\mathrm{end}}] \times \Pi \to \mathbb{R}^{N_i}$ for $i = 1, 2$, where initial values are given for all $\mathbf{p} \in \Pi$. The coupling variables are defined as $\mathbf{y}_i^{\mathrm{cpl}} := B_i \mathbf{y}_i$ with constant matrices $B_i \in \{0, 1\}^{R_i \times N_i}$ such that the coupling variables include just

a subset of $\mathbf{y}_i$ for each $i = 1, 2$. Typically, it holds that $R_1 \ll N_1$ and $R_2 \ll N_2$, i.e., the coupling variables represent just a small portion of the solution. Furthermore, it is allowed that just one of two subsystems in (1) includes all the parameters. Generalisations to more than two subsystems are straightforward.

In many technical applications, implicit time integration schemes have to be applied, since either DAEs or stiff ODEs are involved. Due to a multirate behaviour, a co-simulation based on a dynamic iteration becomes efficient in some cases. Moreover, co-simulation is required if the equations of a subsystem in (1) are not available directly, i.e., just a software package is given including a numerical solver. We consider a dynamic iteration, where the total time span is split into windows with a first window $[t_0, t_{\text{win}}]$. For a fixed $\mathbf{p} \in \Pi$, the iteration of Gauss-Seidel type for the coupled system (1) reads as

$$\begin{aligned} \mathbf{F}_1\Big(\mathbf{y}_1^{(\nu+1)}(t,\mathbf{p}), \mathbf{y}_2^{\mathrm{cpl}\,(\nu)}(t,\mathbf{p}), t, \mathbf{p}\Big) &= \mathbf{0}, \\ \mathbf{F}_2\Big(\mathbf{y}_2^{(\nu+1)}(t,\mathbf{p}), \mathbf{y}_1^{\mathrm{cpl}\,(\nu+1)}(t,\mathbf{p}), t, \mathbf{p}\Big) &= \mathbf{0}, \end{aligned} \qquad \text{for } \nu = 0, 1, 2, \ldots \tag{2}$$

using the starting values $\mathbf{y}_2^{(0)}(t,\mathbf{p}) \equiv \mathbf{y}_2(t_0,\mathbf{p})$. However, a numerical method outputs just the solutions $\mathbf{y}_1$, $\mathbf{y}_2$ on a discrete set of time points, which may also differ for the two subsystems. We assume that all coupling variables are interchanged in a few communication time points $\bar{t}_j$ with $t_0 \le \bar{t}_1 < \bar{t}_2 < \cdots < \bar{t}_J = t_{\text{win}}$. Interpolation yields approximations of the coupling variables $\mathbf{y}_i^{\mathrm{cpl}}(t,\mathbf{p})$ for $t \in [t_0, t_{\text{win}}]$ and $i = 1, 2$.

Now we suppose that the parameters are not known exactly. To perform an uncertainty quantification, the parameters are modelled by random variables $\mathbf{p} : \Omega \to \Pi$ on some probability space $(\Omega, \mathscr{A}, \mu)$ with a joint density $\rho : \Pi \to \mathbb{R}$. Statistical information for a function $g : \Pi \to \mathbb{R}$ is obtained by probabilistic integrals

$$\mathrm{E}(g) := \int_\Omega g(\mathbf{p}(\omega))\, \mathrm{d}\mu(\omega) = \int_\Pi g(\mathbf{p})\, \rho(\mathbf{p})\, \mathrm{d}\mathbf{p} \tag{3}$$

provided that the integral exists. For example, probabilistic integration can be applied to the solution of (1) component-wise. Crucial information consists of the expected value and the standard deviation for the solution. Furthermore, higher moments and failure probabilities also represent integrals of the type (3). Our aim is to compute statistics of the solution $\mathbf{y}_1, \mathbf{y}_2$ for either the complete time interval or just at a final time.

A quadrature scheme or a sampling method yields an approximation of a probabilistic integral (3), see [6] and the references therein. We obtain a finite sum of the form $\mathrm{E}(g) \doteq w_1 g(\mathbf{p}^{(1)}) + \cdots + w_K g(\mathbf{p}^{(K)})$ with grid points $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(K)} \in \Pi$ and weights $w_1, \ldots, w_K \in \mathbb{R}$. For $g = \tilde{g}(\mathbf{y}_1, \mathbf{y}_2)$ at some final time t_{end}, it follows that an initial value problem of the system (1) has to be resolved K times for the different realisations of the parameters.

3 Quadrature with Adjusted Grids

If the solutions of the subsystems in the coupled problem (1) behave differently with respect to the random parameters, then the application of different quadrature formulas might become advantageous. For example, a higher variance within a subsystem often indicates that a higher accuracy of the quadrature is required. Thus we introduce two grids $\mathscr{G}_i := \{\mathbf{p}_i^{(1)}, \ldots, \mathbf{p}_i^{(K_i)}\}$ with $\mathbf{p}_i^{(k)} \in \Pi$ for $i = 1, 2$ dedicated to the two parts of the coupled problem (1). In general, any grid is initiated by a quadrature scheme. The numbers of grid points K_1, K_2 may differ significantly. The subsystem for $\mathbf{F}_i$ together with its solution $\mathbf{y}_i$ is integrated in time on the grid $\mathscr{G}_i$ for each $i = 1, 2$.

Following (2), we have to solve the problems

$$\begin{aligned} \mathbf{F}_1\Big(\mathbf{y}_1^{(\nu+1)}(t, \mathbf{p}_1^{(k)}), \mathbf{y}_2^{\mathrm{cpl}\,(\nu)}(t, \mathbf{p}_1^{(k)}), t, \mathbf{p}_1^{(k)}\Big) &= \mathbf{0} \qquad \text{for } k = 1, \ldots, K_1, \\ \mathbf{F}_2\Big(\mathbf{y}_2^{(\nu+1)}(t, \mathbf{p}_2^{(k)}), \mathbf{y}_1^{\mathrm{cpl}\,(\nu+1)}(t, \mathbf{p}_2^{(k)}), t, \mathbf{p}_2^{(k)}\Big) &= \mathbf{0} \qquad \text{for } k = 1, \ldots, K_2, \end{aligned} \tag{4}$$

in each step of the dynamic iteration. The first iteration step $\nu = 0$ in (4) for $\mathbf{F}_1$ can be computed directly using the globally defined initial values. The output is $\mathbf{y}_1^{(1)}(\bar{t}_j, \mathbf{p}_1^{(k)})$ for $k = 1, \ldots, K_1$ in the communication time points $\bar{t}_1, \ldots, \bar{t}_J$ introduced in Sect. 2. To this end, we need the coupling variables $\mathbf{y}_1^{\mathrm{cpl}(1)}(\bar{t}_j, \mathbf{p}_2^{(k)})$ for $k = 1, \ldots, K_2$ and $j = 1, \ldots, J$. Likewise, the output of $\mathbf{F}_2$ is the solution $\mathbf{y}_2^{(1)}(\bar{t}_j, \mathbf{p}_2^{(k)})$ for $k = 1, \ldots, K_2$ and thus has to be transformed into the coupling variables $\mathbf{y}_2^{\mathrm{cpl}(1)}(\bar{t}_j, \mathbf{p}_1^{(k)})$ for $k = 1, \ldots, K_1$, i.e., the evaluation on the other quadrature grid is crucial. This strategy repeats in each iteration step. Hence transitions between the two grids have to be defined for a fixed time point.

For the interchange of information between the two grids, we consider global approximations in the parameter space Π. An arbitrary global approximation method, which just requires the evaluations in the grid points, is feasible like an interpolation scheme, for example. Alternatively, we apply an approximation based on orthogonal basis polynomials with respect to the L^2-inner product of the probability space induced by the integral (3). Hence a truncated sum of the polynomial chaos expansion is used, see [6]. Let the time $\bar{t}$ be fixed. The global approximation reads as

$$\tilde{\mathbf{y}}_i^{\mathrm{cpl}}(\bar{t}, \mathbf{p}) := \sum_{m=0}^{M_i} \mathbf{u}_{i,m}(\bar{t}) \Phi_m(\mathbf{p}) \tag{5}$$

for $i = 1, 2$ with known basis polynomials $\Phi_m : \Pi \to \mathbb{R}$ satisfying the orthonormality condition $\mathrm{E}(\Phi_m \Phi_n) = \delta_{mn}$. In general, all polynomials up to a certain degree are involved. The coefficient functions in (5) are determined

approximately by

$$\mathbf{u}_{i,m}(\bar{t}) := \int_{\Pi} \mathbf{y}_i^{\mathrm{cpl}}(\bar{t}, \mathbf{p}) \Phi_m(\mathbf{p}) \rho(\mathbf{p}) \, \mathrm{d}\mathbf{p} \doteq \sum_{k=1}^{K_i} w_i^{(k)} \mathbf{y}_i^{\mathrm{cpl}}(\bar{t}, \mathbf{p}_i^{(k)}) \Phi_m(\mathbf{p}_i^{(k)}) \tag{6}$$

for $i = 1, 2$, where the values $w_i^{(k)} \in \mathbb{R}$ represent the weights of quadrature formulas on the grids $\mathscr{G}_i$. Thus the sums (5) can be evaluated for an arbitrary $\mathbf{p} \in \Pi$. In particular, we obtain approximations of the coupling variables on each grid. Since the number of coupling variables is relatively low in comparison to the dimension of the coupled problem, the computational effort for the global approximation is usually negligible compared to the time integration.

After the convergence of the dynamic iteration in a time window, the same approach is repeated in the next time window. Therein, initial values can be transformed between the two grids again by the above procedure. If the approximations have been computed at the final time t_{end}, then we reconstruct statistical data by quadrature formulas using the same grid points.

4 Simulation of a Test Example

To demonstrate the feasibility of the approach described in Sect. 3, we simulate a coupled problem introduced in [2], which consists of an electric part and a thermal part illustrated by Fig. 1. A resistor as well as a diode exhibit a voltage-current-relation depending on the temperature. The electric network is modelled by a nonlinear system of DAEs with dimension $N_1 = 3$. In the thermal part, the temperature of the resistor follows from a one-dimensional linear heat equation, where a semidiscretisation yields ODEs of dimension $N_2 = 20$. The diode receives a scalar temperature from the (right-hand) boundary of the PDE. More details can be found in [2].

The electric network is supplied by a sinusoidal input signal. We compute the numerical solution in the total time interval [0 s, 0.1 s] and apply five time windows for the dynamic iteration (2). In our example, the solution of the electric part is more expensive than the thermal part, since smaller step sizes have to be used in time. The circuit part is solved first in this co-simulation. As communication time points, just

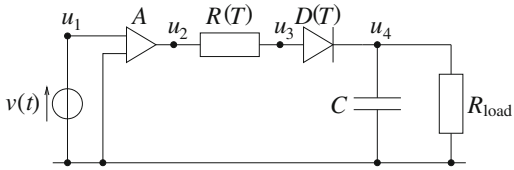


Fig. 1 Electric circuit with temperature-dependent resistor and diode

the final times of each window are involved. The time integration of the subsystems is done by an implicit multistep method based on numerical differentiation formulas.

We introduce two random variables with independent uniform distributions. In the DAE part, the (temperature-independent) load resistance is a random parameter with variations of 10 %. In the ODE part, the heat conduction coefficient becomes random with variations of 40 %. Although the PDE and its ODE discretisation are linear, the dependence of the solution on the parameters is nonlinear in each case. It follows that the parameter space represents a rectangle $\Pi \subset \mathbb{R}^2$.

As quadrature formulas, we employ the two-dimensional midpoint rule on grids of size $L_1 \times L_2$, i.e., L_1 nodes discretise the random resistance and L_2 nodes are dedicated to the random heat conduction. Two different quadrature formulas are considered, which gives a first grid $\mathscr{G}_1$ for the circuit part and a second grid $\mathscr{G}_2$ for the thermal part. In the communication between the grids, we use the global approximations defined by (5), (6), where all polynomials up to degree two are included ($M_1 + 1 = M_2 + 1 = 6$ basis functions).

To illustrate some statistics of the coupled problem, we compute the numerical solution for a combination of a first grid with size 8×6 and a second grid with size 6×8. Figures 2 and 3 depict the first and second moment for the output of the circuit part and the thermal part, respectively, which result from the quadrature formulas associated to the two grids.

We also tried several other grid sizes for comparison. If two identical grids are chosen, then the evaluations of the coupling variables are available directly. Nevertheless, we still perform the projections (6) and reconstructions (5) to investigate the accuracy. A reference solution is computed by the midpoint rule on a single grid with 40×40 nodes, where no transitions between different grids and thus no errors from global approximations occur. Table 1 demonstrates the comparison for the expected values as well as the standard deviations, where the maximum differences have been calculated for both all involved time points and all components of a subsystem.

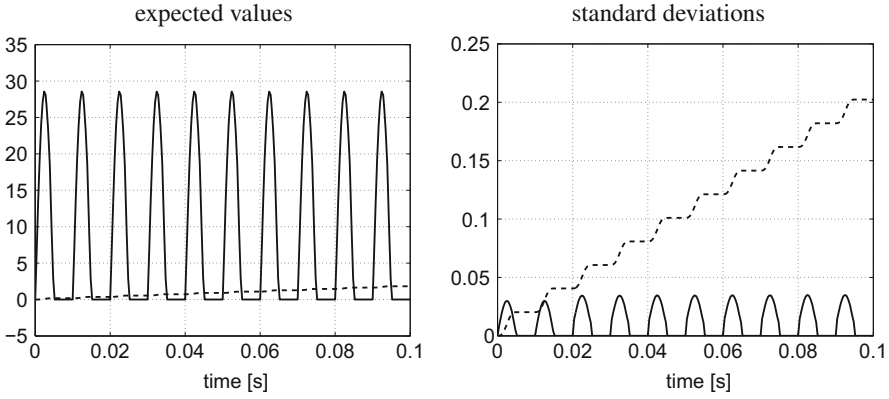


Fig. 2 Expected values as well as standard deviations for output voltage u_4 in unit [V] (*solid line*) and dissipated energy in unit [J] (*dashed line*) within circuit part

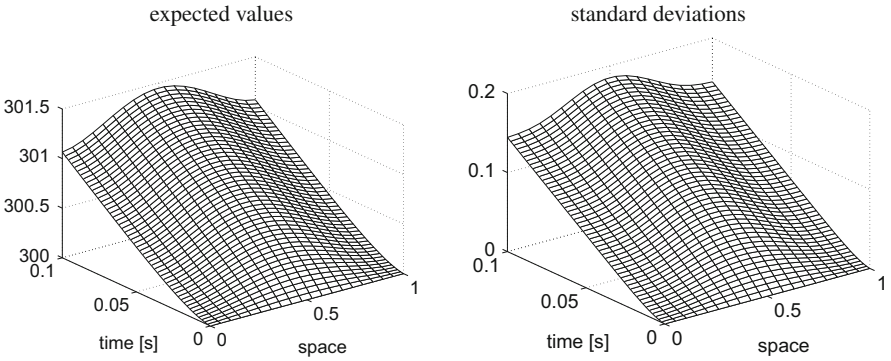


Fig. 3 Expected values as well as standard deviations for the resistor's temperature in unit [K] (spatial domain is standardised to [0, 1]) within thermal part

Table 1 Maximum differences for statistics computed using different grids with respect to reference solution separately for circuit variables and temperature

Grid sizes		Circuit variables		Temperature	
First grid	Second grid	Expected value	St. deviation	Expected value	St. deviation
10×10	10×10	4.6e−4	8.0e−3	6.2e−4	2.7e−3
5×5	10×10	1.3e−3	7.1e−3	1.9e−3	1.6e−2
5×5	5×5	4.9e−3	4.7e−2	1.1e−2	2.7e−2
8×6	6×8	1.3e−3	2.2e−2	2.7e−3	6.8e−3
8×4	4×8	2.8e−3	6.1e−2	8.3e−3	2.9e−2

5 Conclusions

We explained the need for coupled quadrature grids in uncertainty quantification and its algorithmic application within co-simulation. With a multiphysics example we showed the applicability and the prospect of the method.

Acknowledgements This work is a part of the project 'Nanoelectronic Coupled Problems Solutions' (NANOCOPS) funded by the European Union within FP7-ICT-2013 (grant no. 619166).

References

1. Bartel, A., Pulch, R.: A concept for classification of partial differential algebraic equations in nanoelectronics. In: Bonilla, L.L., Moscoso, M., Platero, G., Vega, J.M. (eds.) Progress in Industrial Mathematics at ECMI 2006. Mathematics in Industry, vol. 12, pp. 506–511. Springer, Berlin (2007)
2. Bartel, A., Günther, M., Schulz, M.: Modeling and discretization of a thermal-electric test circuit. In: Antreich, K. (eds.) Modeling, Simulation and Optimization of Integrated Circuits. ISNM, vol. 146, pp. 187–201. Birkhäuser, Boston (2003)

3. Bartel, A., Brunk, M., Günther, M., Schöps, S.: Dynamic iteration for coupled problems of electric circuits and distributed devices. SIAM J. Sci. Comput. **35**(2), B315–B335 (2013)
4. Chauvière, C., Hesthaven, J.S., Lurati, L.: Computational modeling of uncertainty in time-domain electromagnetics. SIAM J. Sci. Comput. **28**(2), 751–775 (2006)
5. Pulch, R.: Stochastic collocation and stochastic Galerkin methods for linear differential algebraic equations. J. Comput. Appl. Math. **262**, 281–291 (2014)
6. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)

MS 14
MINISYMPOSIUM: MATHEMATICS AND CAGD: INTERACTIONS AND INTERSECTIONS

Organizers

Costanza Conti[1] and Lucia Romani[2]

Speakers

Hartmut Prautzsch[3]
Rational Free Form Spline Surfaces with Linear Transitions Maps

Xiaodiao Chen[4] and Weiyin Ma[5]
Geometric Clipping Methods for Efficient Root Finding

Tomas Sauer[6]
Numeric Evaluation of Geometric Continuity in CAD Systems

[1]Costanza Conti, Université degli Studi Firenze, Firenze, Italy.
[2]Lucia Romani, Université degli Studi Milano-Bicocca, Milano, Italy.
[3]Hartmut Prautzsch, Karlsruhe Institute for Technology, Karlsruhe, Germany.
[4]Xiaodiao Chen, Hangzhou Dianzi University, Hangzhou, China.
[5]Weiyin Ma, City University of Hong Kong, Hong Kong, China.
[6]Tomas Saue, FORWISS (Institute for Software Systems in Technical Applications of Computer Science), Passau, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_51

Christian Arber[7]
The Unreasonable Effectiveness of Mathematics in the CAD/CAM TopSolid Software

Costanza Conti[1], Lucia Romani[2], Virginie Uhlmann[8] and Michael Unser[9]
Active Contours for Biomedical Images Based on Hermite Exponential Splines

Keywords

Computer aided geometric design
Theoretical and applied issues arising from technology and industry

Short Description

Computer aided geometric design (CAGD) concerns itself with the mathematical description of shapes for use, for example, in computer graphics, manufacturing, CAD/CAM, scientific visualization, or computer animation. Drawing from many areas and influencing others, CAGD is inherently interdisciplinary involving geometry, computer graphics, numerical analysis, approximation theory, data structures, linear and computer algebra. CAGD started in the 1960s, going back to efforts by Citroen and Renault in France and by Boeing and General Motors in the U.S. The term Computer Aided Geometric Design was coined after the 1972 conference at the University of Utah, organized by R. Barnhill and R. Riesenfeld, and since then it became a discipline in its own right. In recent years the interest in CAGD and its applications is beyond the mentioned contexts, not only because graphical applications are spreading their diffusion and increasing their request in different fields (industrial, medical, biological, topographic, geological applications) but also as a consequence of the relevant theoretical problems originated in the field, which are considered stimulating not only in the "contiguous disciplines" of numerical analysis and geometry, but also in different contexts. Therefore, aim of this mini-symposium is to gather academic and industrial scientists to give notice of new mathematical methods for the description of geometric objects. In particular, the selected talks will deal with both theoretical and applied issues arising from technology and industry.

[7]Christian Arber, TOPSOLIDE, Paris, France.
[8]Virginie Uhlmann, Biomedical Imaging Group - EPFL, Lausanne, Switzerland.
[9]Michael Unser, Biomedical Imaging Group - EPFL, Lausanne, Switzerland.

MS 15
MINISYMPOSIUM: MATHEMATICS IN NANOTECHNOLOGY

Organizers

Tim G. Myers[1] and Luis Bonilla[2]

Speakers

Ana Carpio[3], David Rodriguez[4] and Baldvin Einarsson[5]
Dynamics of Bacterial Aggregates in Microflows

Mariano Alvaro[6], Luis L. Bonilla[2] and Manuel Carretero[7], Evgeny Ya. Sherman[8]
Dynamics of Optically Injected Currents in Carbon Nanotubes

Francesc Font Martinez[9] and Tim G. Myers[1]
Mathematical Model for the Melting of a Nano-thin Film

[1]Tim G. Myers, Centre de Recerca Matemàtica, Barcelona, Spain.
[2]Luis Bonilla, Universidad Carlos III de Madrid, Leganes, Spain.
[3]Ana Carpio, Complutense, Madrid, Spain.
[4]David Rodriguez, Centro Nacional de Biotecnologia, CSIC, Madrid, Spain.
[5]Baldvin Einarsson of California at Santa Barbara, USA.
[6]Mariano Alvaro, Universidad Carlos III de Madrid, Leganes, Spain.
[7]Manuel Carretero, Universidad Carlos III de Madrid, Leganes, Spain.
[8]Evgeny Ya. Sherman, Ikerbasque, Universidad del Pais Vasco-EHU, Spain.
[9]Francesc Font Martinez, CRM, Barcelona, Spain.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_52

Michelle MacDevette[10] and Tim G. Myers[1]
Boundary Layer Analysis and Heat Transfer of a Nanofluid

Keywords

Nanofluid
Nanotechnology
Nano-thin film
Nanotubes

Short Description

Nanotechnology is one of the key modern research directions, with billions being invested by governments throughout the world, and in particular by the US, Europe and Japan. Nanotechnology is relevant to a vast range of practical applications, such as in medicine, electronics, biomaterials and energy production. To date the vast majority of research has focused on the experimental side, with the theory often lagging behind. However, there are a number of mathematical groups now working on topics relevant to the nano industry. In this mini-symposium we intend to bring together a selection of speakers who will discuss a broad range of topics relevant to nanoscience and who will be able to demonstrate the relevance of mathematics to this research field.

[10]Michelle MacDevette, Centre de Recerca Matematica, Barcelona, Spain.

Boundary Layer Analysis and Heat Transfer of a Nanofluid

T.G. Myers and M.M. MacDevette

Abstract Nanofluids have been hailed as a possible winner in the race to find sufficiently powerful cooling systems for emerging high-power electronic devices. There exist numerous experiments demonstrating nanofluids to have remarkable properties. However, there has been much controversy in the literature with discrepancies between results concerning the heat transfer and thermal conductivity of nanofluids. In this paper we analyse a popular model for nanofluid flow which previously has been employed to demonstrate the improved heat transfer. We find the opposite result and then move on to explain some of the reasons behind the discrepancies.

Keywords Boundary layer analysis • Heat flow • Nanofluid

1 Introduction

Modern high-performance electrical devices often produce large amounts of heat, which must somehow be removed. Nanofluids, which consist of a base fluid and a suspension of nanoparticles have been shown in many research programmes to be capable of removing large amounts of heat, even with a remarkably low particle concentration [3, 5]. However, recently, a small number of authors have questioned this property of nanofluids. Specifically, the benchmark study carried out in over 30 laboratories around the world seemed to imply no great increase in heat transfer [2].

T.G. Myers (✉)
Centre de Recerca Matemàtica, Barcelona, Spain
e-mail: tim.myerscrm@gmail.com

M.M. MacDevette
CERECAM, University of Cape Town, Cape Town, South Africa
e-mail: michelle.macdevette@uct.ac.za

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_53

In this paper we will analyse one of the standard models for nanofluid flow. This particular model has been used to show theoretically that nanofluids can increase heat transfer by a significant amount. By applying standard boundary layer theory we will find the opposite conclusion. In the final section we discuss three highly cited papers on this topic and explain why the authors reached a different conclusion to that of the present work.

2 Nanofluid Heat Transfer

Buongiorno [1] developed a model for nanofluid flow with the following main assumptions: incompressible flow; negligible external forces; dilute mixture; negligible viscous dissipation; negligible radiative heat transfer. This leads to the following set of equations:

$$\nabla \cdot \mathbf{u} = 0, \tag{1}$$

$$\rho_{nf}\left[\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right] = -\nabla p - \nabla\cdot\tilde{\tau}, \tag{2}$$

$$\frac{\partial(\chi_{nf}T)}{\partial t} + \nabla\cdot(\chi_{nf}\mathbf{u}T) = \nabla\cdot(k_{nf}\nabla T), \tag{3}$$

$$\frac{\partial\phi}{\partial t} + \nabla\cdot(\phi\mathbf{u}) = \quad \nabla\cdot\left[D_B\nabla\phi + D_T\frac{\nabla T}{T}\right], \tag{4}$$

where $\mathbf{u}$ is the velocity vector, T the temperature and ϕ the volume fraction of nanoparticles. Subscripts bf, nf and np refer to the base fluid, nanofluid and nanoparticle, respectively. The density, volumetric heat capacity and specific heat depend on the volume fraction

$$\rho_{nf} = \phi\rho_{np} + (1-\phi)\rho_{bf}, \tag{5}$$

$$\chi_{nf} = (\rho c)_{nf} = \phi\rho_{np}c_{np} + (1-\phi)\rho_{bf}c_{bf}, \tag{6}$$

$$k_{nf} = \frac{k_{bf}}{(1-\phi^{1/3})^2}\left[(1-\phi) + \phi\frac{\rho_{np}c_{np}}{\rho_{bf}c_{bf}}\right]\frac{n-1}{2(n+1)}\left[\frac{1+\phi^{1/3}}{2} - \frac{1}{n+1}\right]^{-1}, \tag{7}$$

where $n = 2.233$ [10]. In the boundary layer section of this analysis we will require only part of the stress tensor, $\mu_{nf}\partial u/\partial y$, where the viscosities

$$\mu_{nf} = (1 + 7.3\phi + 123\phi^2)\mu_{bf}\,, \tag{8}$$

$$\mu_{nf} = (1 - 0.19\phi + 306\phi^2)\mu_{bf}\,, \tag{9}$$

are for water and ethylene-glycol based nanofluids, see [8].

An interesting feature which distinguishes the above system from standard flow models is the presence of Brownian motion and thermophoresis terms (represented by D_B and D_T respectively). Thermophoresis describes a particles motion against a temperature gradient. Almost all authors who deal with this system use a definition of the diffusion coefficients that involves T and ϕ. Since these are variables in the model those authors take some form of average (rather than allowing them to vary). In [9] a different form is used for the coefficients

$$C_B = \frac{D_B}{T} = \frac{k_B}{3\pi\mu_{bf}d_p}, \qquad C_T = \frac{D_T}{\phi} = \frac{\beta\mu_{bf}}{\rho_{bf}}, \tag{10}$$

where k_B, d_p are the Boltzmann constant and particle diameter, so C_B, C_T are constant and the T, ϕ dependence is correctly formulated in the governing equations.

2.1 Boundary Layer Analysis

For flow over a flat surface, $y = 0$, standard boundary layer theory requires a set of boundary conditions of the following form. At $y = 0$

$$k_{nf}T_y = -Q \qquad u = v = 0\,, \tag{11}$$

where Q represents the energy input at the boundary. At the inlet $x = 0$

$$\phi = \phi_{in} \qquad T = T_\infty \qquad \mathbf{u} = (U, 0)\,. \tag{12}$$

In the far-field conditions $y \to \infty$

$$u = U \qquad v = 0 \qquad T = T_\infty\,. \tag{13}$$

The variables are now re-scaled in order to focus close to the boundary and eliminate negligible terms:

$$\hat{x} = \frac{x}{L} \qquad \hat{y} = \frac{y}{L}\sqrt{Re} \qquad \hat{T} = \frac{T - T_\infty}{A} \tag{14}$$

$$\hat{u} = \frac{u}{U} \qquad \hat{v} = \frac{v}{U}\sqrt{Re} \qquad \hat{p} = \frac{p - p_\infty}{\rho_{bf}U^2}, \tag{15}$$

where $U, Re = \rho_{bf}UL/\mu_{bf}, A$ are the far field velocity, the Reynolds number and the temperature scale. In [9] the velocity scale is fixed to the base fluid value, this permits a straightforward comparison of heat transfer coefficients. However, it does mean that if U is fixed, regardless of the particle loading, then the pressure drop to drive the flow must increase with the particle loading. This should be taken into

account when examining the final results. The physical properties of the fluid are also scaled with the base fluid value:

$$\hat{\mu}_{nf} = \frac{\mu_{nf}}{\mu_{bf}} \qquad \hat{\rho}_{nf} = \frac{\rho_{nf}}{\rho_{bf}} \qquad \hat{k}_{nf} = \frac{k_{nf}}{k_{bf}} \qquad \hat{\phi} = \frac{\phi}{\phi_{in}} \qquad \hat{\chi}_{nf} = \frac{\chi_{nf}}{\chi_{bf}}. \quad (16)$$

The most obvious change to the system under this scaling may be found by examining the particle concentration equation (after dropping the hat notation)

$$\nabla \cdot (\phi \mathbf{u}) = \gamma \frac{\partial}{\partial y}\left[\left(T + \frac{T_\infty}{A}\right)\frac{\partial \phi}{\partial y} + \lambda \frac{\phi}{T + T_\infty/A}\frac{\partial T}{\partial y}\right]. \quad (17)$$

This involves two non-dimensional groupings $\gamma = C_B A \rho_{bf}/\mu_{bf}$ and $\lambda = C_T/(C_B A)$. Both these numbers contain the temperature scale A. The temperature increase is caused by the heat input at the boundary, in non-dimensional form the boundary condition is

$$\frac{k_{bf} A \sqrt{Re}}{L}\frac{\partial T}{\partial y} = -Q \quad (18)$$

which then suggests the choice $A = QL/(k_{bf}\sqrt{Re})$. In [1, 9] appropriate values are provided for the physical parameters for water or ethylene-glycol based nanofluids with Al_2O_3 particles. For our analysis the appropriate list of values is provided in Table 1

In particular we note that for EG $\gamma = \mathcal{O}(10^{-5})$ and for water $\gamma = \mathcal{O}(10^{-4}) \ll 1$. Consequently we may neglect the term involving γ in (17) and so, after imposing the incompressibility condition, find

$$\nabla \cdot (\phi \mathbf{u}) = \mathbf{u} \cdot \nabla \phi \approx 0. \quad (19)$$

This result demonstrates that ϕ is constant along a streamline and so, noting that $\phi(x = 0) = 1$, we find $\phi = 1$ everywhere. That is, the temperature gradient within the fluid does not act to move particles in any significant way. It also means that physical parameters such as density, viscosity and conductivity remain constant. This finding is in contrast with the vast majority of previous studies which find significant variation in particle concentration and hence parameters values. However

Table 1 Values of coefficients in non-dimensional equations

Quantity	Ethylene glycol	Water	Quantity	Ethylene glycol	Water
C_B	4.3825×10^{-15}	7.0559×10^{-14}	C_T	3.1918×10^{-8}	5.0721×10^{-9}
Re	68.8696	10^3	A	4.6705×10^4	5.1926×10^3
γ	1.4×10^{-5}	3.6638×10^{-4}	λ	155.9362	13.8437

our findings are in line with those of Evans et al. [4] via a molecular dynamics simulation.

With the finding that concentration and associated parameter values are approximately constant we are able to apply standard boundary layer theory to the remaining equations.

The well-known Blasius solution for boundary layer flow over a flat plate involves first introducing a stream function ψ where $u = \frac{\partial \psi}{\partial y}$, $v = -\frac{\partial \psi}{\partial x}$. A similarity variable $\eta = y/\sqrt{2\nu_{nf}x}$ is then introduced where, to satisfy the momentum equation $\psi = \sqrt{2\nu_{nf}x} f(\eta)$ and f is an unknown function. The resultant boundary value problem may be simplified using Töpfer's transformation $f(\eta) = rF(r\eta)$, where $r > 0$ is a constant and the problem reduces to

$$F''' + FF'' = 0 \tag{20}$$

$$F(0) = 0 \qquad F'(0) = 0 \qquad F''(0) = 1\,, \tag{21}$$

The value of r is determined via $r = (F'(\infty))^{-1/2} \approx 0.7773$.

The above approach allows us to calculate the fluid velocity, which is then required in the heat equation (the scaled version of (3)). In [9] this is solved numerically and approximately via the Heat Balance Integral Method. In the following section we will briefly discuss the heat transfer coefficient, which is the main focus of the paper, and then show results obtained through the approach described above.

2.2 Heat Transfer Coefficient

The heat transfer coefficient (HTC) is calculated, with little thought, in many studies of heat flow and even in basic courses on boundary value problems, yet it is a very poorly defined quantity. Generally the HTC, h, is defined through the relation

$$-k_{nf} \left.\frac{\partial T}{\partial y}\right|_{y=0} = h\Delta T\,. \tag{22}$$

The problem is, what is the definition of the temperature jump ΔT? In the literature it may be described as the difference between the wall temperature and the far-field temperature. Mathematical texts often replace the wall temperature with the fluid temperature $T|_{y=0+}$. The problem with both choices is they implicitly assume a linear temperature change from wall to far-field.

If we wish to find out about the actual transfer of heat energy to the fluid we may consider what is termed the 'cup average', see [9]. This represents the temperature of the fluid that would occur if it were collected in a cup at the end of the pipe. Say a fluid enters the system at $x = 0$, with an initial temperature T_∞. A distance L

downstream of the inlet the energy flux above the initial value is given by

$$\int_0^{\delta_T} \rho c u (T - T_\infty)\, dy\,, \tag{23}$$

where $\delta_T(L)$ is the thickness of the thermal boundary layer at $x = L$. The average temperature rise in the fluid T_{av} is defined via

$$(T_{av} - T_\infty) \int_0^{\delta_T} \rho c u\, dy = \int_0^{\delta_T} \rho c u (T - T_\infty)\, dy\,. \tag{24}$$

An HTC that actually represents the energy transfer from the substrate to the fluid is then

$$h = \frac{Q}{T_{av} - T_\infty} = \frac{Q \int_0^{\delta_T} \rho c u\, dy}{\int_0^{\delta_T} \rho c u (T - T_\infty)\, dy}\,. \tag{25}$$

In [9] the solution to the boundary layer equations is substituted into (25) for various values of the particle concentration. The results are presented in Fig. 1. The base fluid and 5 and 10 % volume fraction fluids are depicted by the circle, dashed and solid lines respectively. From the results it is quite clear that the HTC decreases with volume fraction. This directly contradicts the results of hundreds of research papers. In fact this conclusion should be even more clear. As mentioned earlier we fix the velocity scale so that as the volume fraction increases we must increase pumping power, so not only does the HTC decrease with volume fraction but the energy required to move the fluid increases.

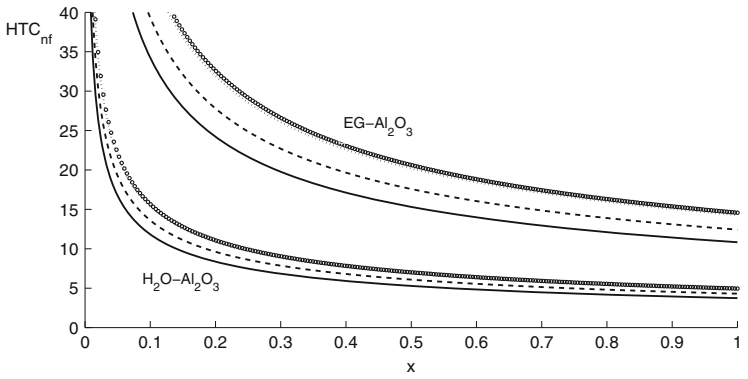


Fig. 1 Variation of HTC for ethylene glycol and water with alumina nanoparticles. *Reprinted with kind permission from Springer Science+Business Media: [9, Fig. 4]*

2.3 Discussion

Given that so many previous research papers conclude that the HTC increases with increasing volume fraction, it is worth discussing the source of the differences between ours and previous results.

In [9] three examples of previous papers are used, those of [1, 6, 7]. They were chosen due to their high citation count, (together they have almost 1400 citations on Google Scholar).

Firstly, we note that all of the above papers use the definition for the HTC $h = Q/(T_w - T_\infty)$ and so will not accurately capture the heat input into the fluid. The first paper [1] takes the governing system of equations described at the start of this paper. To make analytical progress the system is reduced in the style of a lubrication theory model, which then represents some 'laminar sublayer'. This sublayer is matched to the outer turbulent region. However, if we analyse the governing equations it turns out that the sublayer equations hold in a region three orders of magnitude smaller than the true laminar sublayer, consequently the matching is invalid. The second two papers [6, 7] deal with similarity solutions. In general they follow the model of [1] and use parameter values quoted in that paper, with the exception of the Lewis number (the ratio of thermal to mass diffusivity). In [1] the parameter values for a water-alumina system result in $Le = \mathcal{O}(10^5)$. Both sets of authors of [6, 7] take $Le = 10$. If the correct value had been used then the effect of particle motion would have dropped out of the governing equations, leading to a system similar to that in [9]. The much lower value, $Le = 10$, magnifies the particle motion, so implying the particles have a dramatic effect on fluid properties.

In conclusion then it appears that nanofluids do not greatly enhance heat transfer. This has been demonstrated through the benchmark experimental study [2]. The present paper, which summarises the work of [9], provides theoretical confirmation of this result.

Acknowledgements The research of TGM was supported by a Marie Curie International Reintegration Grant *Industrial applications of moving boundary problems* Grant no. FP7-256417 and Ministerio de Ciencia e Innovación Grant MTM2011-23789. MMM acknowledges the support of a Centre de Recerca Matemàtica PhD grant.

References

1. Buongiorno, J.: Convective transport in nanofluids. J. Heat Transf. **128**, 240–250 (2006)
2. Buongiorno, J., Venerus, D., Prabhat, N., McKrell, T., Townsend, J., et al.: A benchmark study on the thermal conductivity of nanofluids. J. Appl. Phys. **106**, 094312 (2009)
3. Das, S.K., Choi, S.U.S., Yu, W., Pradeep, T.: Nanofluids: Science and Technology. Wiley, New York (2008)
4. Evans, W., Fish, J., Keblinski, P.: Role of Brownian motion hydrodynamics on nanofluid thermal conductivity. Appl. Phys. Lett. **88**, 093116 (2006)

5. Godson, L., Rajab, B., Mohan Lala, D., Wongwises, S.: Enhancement of heat transfer using nanofluids—an overview. Renew. Sust. Energ. Rev. **14**(2), 629–641 (2010)
6. Khan, W.A., Pop, I.: Boundary-layer flow of a nanofluid past a stretching sheet. Int. J. Heat Mass Trans. **53**, 2477–2483 (2010)
7. Kuznetsov, A.V., Nield, D.A.: Natural convective boundary-layer flow of a nanofluid past a vertical plate. Int. J. Therm. Sci. **49**, 243–247 (2010)
8. Maiga, S.E.B., Nguyen, C.T., Galanis, N., Roy, G.: Heat transfer behaviors of nanofluids in a uniformly heated tube. Superlattice Microst. **35**, 543–557 (2004)
9. MacDevette, M.M., Myers, T.G., Wetton, B.R.: Boundary layer analysis and heat transfer of a nanofluid. Microfluid. Nanofluid. (2014). doi:10.1007/s10404-013-1319-1
10. Myers, T.G., MacDevette, M.M., Ribera, H.: A time-dependent model to determine the thermal conductivity of a nanofluid. J. Nanopart. Res. **15**, 1775 (201). doi:10.1007/s11051-013-1775-2

Dynamics of Bacterial Aggregates in Microflows

Ana Carpio, Baldvin Einarsson, and David R. Espeso

Abstract Biofilms are bacterial aggregates that grow on moist surfaces. Thin homogeneous biofilms naturally formed on the walls of conducts may serve as biosensors, providing information on the status of microsystems (MEMS) without disrupting them. However, uncontrolled biofilm growth may largely disturb the environment they develop in, increasing the drag and clogging the tubes. To ensure controlled biofilm expansion we need to understand the effect of external variables on their structure. We formulate a hybrid model for the computational study of biofilms growing in laminar microflows. Biomass evolves according to stochastic rules for adhesion, erosion and motion, informed by numerical approximations of the flow fields at each stage. The model is tested studying the formation of streamers in three dimensional corner flows, gaining some insight on the effect of external variables on their structure.

Keywords Biofilm • Microflows

1 Introduction

As the size of the components of technological devices diminishes, new procedures to measure their inner variables without disturbing the system must be developed. For some microdevices, cheap and environmentally friendly monitoring might be achieved exploiting the bacteria that live in them. Bioremediation policies already benefit from microorganisms. Bacteria feeding on a wide variety of toxic pollutants are deliberately released to clean up oil spills or to purify underground water in

A. Carpio (✉)
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, Madrid, Spain
e-mail: ana_carpio@mat.ucm.es

B. Einarsson
Air-Worldwide, Boston, MA, USA
e-mail: BEinarsson@AIR-Worldwide.com

D.R. Espeso
Instituto Gregorio Millán, Universidad Carlos III de Madrid, Madrid, Spain
e-mail: darodri@pa.uc3m.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_54

farming land and mines [13]. For technological purposes, the ability of bacteria to emit optic signals is more appealing. Microorganisms naturally occurring in the environment fluoresce in response to the presence of certain chemicals or certain processes. Such is the case of bioluminiscence phenomena in the southern seas.

Many bacterial species survive in moist environments forming aggregates called biofilms. Microorganisms adhere to surfaces, forming colonies and changing their phenotype to produce extracellular polymeric matrix (EPS). This matrix shelters them from antibiotics, disinfectants, flows and external aggressions. Biofilms may be considered biological materials, whose properties are governed by environmental factors affecting cellular behavior. Recent attempts to engineer devices out of biofilms successfully produced electrooptical devices [2]. The advancement of synthetic biology is paving the way for the use of biofilms as bioindicators or biosensors in the environment [10]. There are efforts to use biofilms emitting optic signals as microsensors in microdevices. Bacteria can be genetically engineered to change their color in response to variations in the environment. Properly modified, bacteria growing in the devices could give local information of the temperature or other variables, without perturbing the internal flow, since the typical size of bacteria is of the order of microns. To indicate the magnitude of variables on the surfaces they attach to, biofilms should be homogeneous and thin. Pattern formation may largely disrupt the environment they grow in. To be able to exploit bacteria in a controlled way, we must understand the influence of external factors on their collective dynamics.

Biofilms are a mixture of living cells embedded in an exopolysaccharide matrix which contains different kinds of metabolic by-products, that can be generically considered as 'biomass'. In fact, the formation of biofilms in flows may be included in a more general group of physical processes where adhesion mechanisms drive agglomeration of matter to create different geometries. The mechanical behavior of the biomass (EPS, cells, debris) and its interaction with the flow seem to be relevant, allowing for growth of structures that do not align with the streamlines of the flow, but may cross the mainstream or wrap around tubes forming helices instead [11, 12].

In this paper, we propose a computational framework to study the growth of biological aggregates in flows triggered by adhesion of particles, much faster than growth due to nutrient consumption. The biofilm is considered a biomaterial with known average cohesive properties formed by a soft sticky matrix of EPS, debris, and other substances secreted by the cells included in it or floating around. We formulate stochastic rules for biomass adhesion, erosion and motion informed by the continuous flow fields around the expanding aggregate, that are approximated by a finite difference discretization strategy using a fixed mesh to reduce the computational cost. The resulting model is tested studying biofilm streamer formation in laminar corner flows.

The paper is organized as follows. In Sect. 2, we describe the general framework and collect the rules for biomass behavior. Section 3 illustrates the numerical results and discusses the insight gained on the dynamics of the aggregates.

2 Hybrid Description of Biofilms in Microflows

Hybrid models combine continuous descriptions of some relevant fields, such as concentrations, flow fields or EPS matrix production, with discrete descriptions of the cells [1, 5, 6]. The situation we examine here fits better as interaction of the surrounding fluid with a elastic biofilm structure whose growth is mediated by adhesion processes. From a computational point of view, biomass is considered as a mixture of bacteria and organic matter allocated on a grid which may behave in different ways in response to external conditions with a certain probability.

Let us denote by Ω_f the region occupied by fluid and by Ω_b the region occupied by biofilm. The whole computational region is divided in a grid of tiles. Each tile may be filled with either substratum, fluid, or biomass, as illustrated in Fig. 1. Since we have in mind applications to microflows, we choose the size of each tile to be of the order of the average size of one bacterium, about 1-2 μm.

The fluid surrounding the biofilm is governed by the incompressible Navier-Stokes equations:

$$\begin{aligned} \rho \mathbf{u}_t - \mu \Delta \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p = 0, \quad & \mathbf{x} \in \Omega_f, t > 0 \\ \operatorname{div} \mathbf{u} = 0, \quad & \mathbf{x} \in \Omega_f, t > 0 \end{aligned} \tag{1}$$

where $\mathbf{u}(\mathbf{x}, t)$ is the velocity and $p(\mathbf{x}, t)$ the pressure. ρ and μ stand for the density and viscosity of the fluid. The non-slip condition on the velocity holds at the biofilm/fluid interface Γ. A low cost prediction of the evolution of the velocity and pressure fields is provided by second order slight artificial compressibility schemes [3]. Approximated velocities and pressures can be improved using second order implicit gauge schemes [16], if necessary, at a higher cost.

Flow effects are felt by a biofilm on much shorter time scales (s) than growth effects (h) [4]. Biomass attaches, detaches and moves according to the flow fields at each location. Floating bacteria are carried by the fluid. The flow geometry selects preferential adhesion sites on the walls where biofilm seeds may be nucleated [12]. Biofilm nucleation may be successful or not depending on the surface nature and the bacterial strain. The flow also determines the strength of the biofilm [9, 15]. Once a biofilm seed is formed, biomass accumulation is a balance between biomass increase due to adhesion or cellular processes, and loss of bacteria due to erosion [14]. We describe below basic stochastic rules for adhesion, erosion and motion processes, having in mind the model case of bacterial streamers in laminar corner microflows, that will serve as a test later. We focus on fast processes. Growth due to nutrient consumption is neglected here.

Two main adhesion processes are taken into account:

- Adhesion of floating cells to walls. In laminar regimes, nucleation of biofilm seeds on the walls is often driven by the geometry. Corners or narrowings may produce secondary flows that drive cells and particles to the walls. Continuous adhesion of bacteria at preferential adhesion sites is taken care of by attaching N_s cells at each step. They distribute on the seed, inside a limited region where the secondary flow is expected to be relevant.

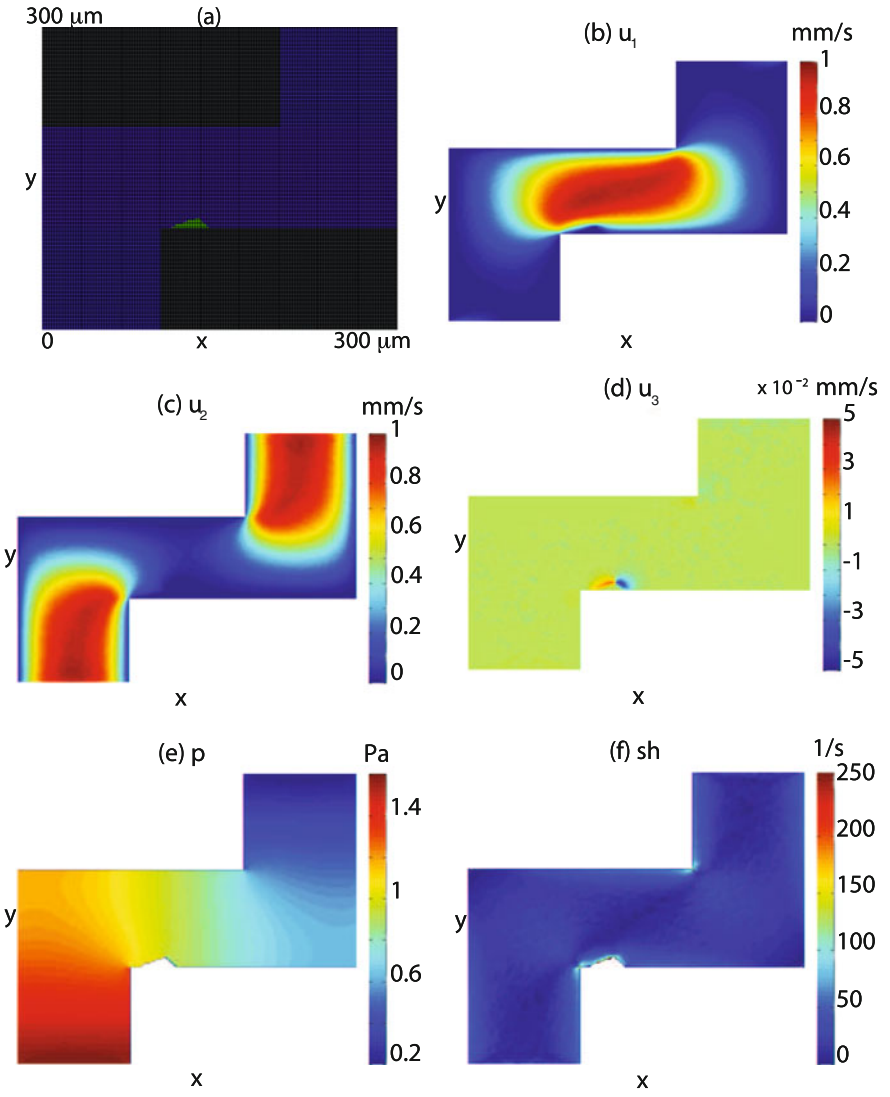


Fig. 1 Initial status of a central slice $z = z_0$ of the tubes: (**a**) Computational grid with biofilm seed (*green*), fluid (*blue*) and substratum (*black*). (**b**), (**c**), (**d**) Velocity components around the initial biofilm seed. (**e**) Pressure field. (**f**) Shear rate

- Once a biofilm seed sticks out from the wall, bacteria and particles swimming with the flow may hit it, and stick to it at a certain rate. Additional N_b biomass blocks are distributed between the tiles located at the biofilm/fluid interface.

N_s and N_b depend on the density of biomass floating in the fluid. N_s is affected by the likeliness of the specific bacterial strain selected to adhere to the walls.

Biomass tiles $\mathscr{C}$ located on the surface of the biofilm detach due to shear forces exerted by the flow [14]. A probability for biomass detachment is proposed in [8]:

$$P_e(\mathscr{C}) = \frac{1}{1 + \frac{\gamma}{\tau(\mathscr{C})}} = \frac{\tau(\mathscr{C})}{\tau(\mathscr{C}) + \gamma}. \tag{2}$$

γ is a measure of the biofilm cohesion. We assume it to be known and constant. $\tau(\mathscr{C})$ measures the shear force felt by cell $\mathscr{C}$. Here, we use the magnitude of the shear force due to the flow at the cell location $\tau_f(\mathscr{C})$, modified by a geometrical factor $f(\mathscr{C})$ that accounts for the local sheltering role of neighboring cells, see [6]. In our numerical experiments, $\tau_f(\mathscr{C})$ is usually set equal to the shear rate at location $\mathscr{C}$ multiplied by the fluid viscosity μ. The shear rate is defined as the spatial rate of change in the fluid velocity field [7]. As for the geometrical factor, it varies according to the main component of the flow, see [6]. In practice, we check erosion in the three directions. At each step and for each biomass tile $\mathscr{C}$ on the biofilm boundary, we detach biomass with probability $P_e(\mathscr{C})$. Erosion due to the flow may occur as detachment of single blocks or of whole clusters of biomass with a thinning connection to the rest of the biofilm.

Shear forces exerted by the flow on the biofilm surface detach biomass. Normal forces on biofilm surfaces may move them. The motion of a biofilm block may be seen as the result of the collective motion of small fragments of the aggregate.

The probability for biomass motion in the x directions is defined as:

$$P_x(\mathscr{C}) = \frac{1}{1 + \frac{\gamma}{|F_x(\mathscr{C})|}} = \frac{|F_x(\mathscr{C})|}{|F_x(\mathscr{C})| + \gamma}. \tag{3}$$

Similar expressions are used in the y and z directions. γ is again a measure of the biofilm cohesion. F_x is the force exerted by the flow in the x direction (on cell walls normal to the x direction) weighted with a geometrical factor accounting for neighbor protection similar to the one used in (2) [6]. F_y and F_z are its counterparts in the y and z direction. The forces are calculated using the values of the fluid stress tensor σ at the cell location: $\sigma \cdot \mathbf{n}$ for the chosen normal vector $\mathbf{n}$.

At each step and for each occupied tile on the biofilm boundary, the biomass moves in the x direction with probability $P_x(\mathscr{C})$ pushing its neighbors in that direction too. Motion is in the positive or negative sense depending on the sign of F_x. Similar rules are applied in the y and z directions.

3 Numerical Results

We will fix as a model case of study the growth of streamers in corner microflows, that is well documented experimentally [12]. The computational region is described in Fig. 1a. A pressure driven flow circulates through the ducts with maximum

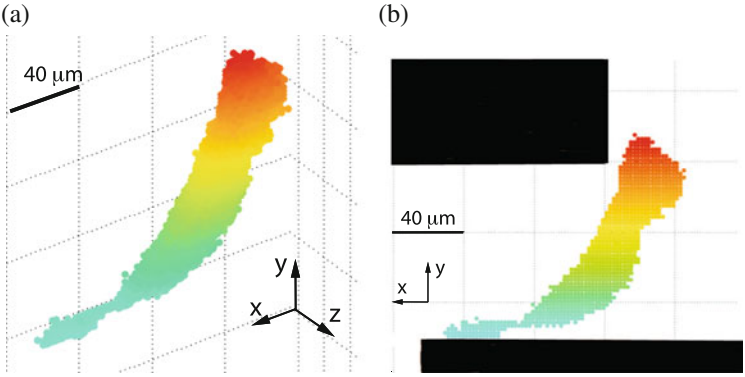


Fig. 2 Streamer grown for $\gamma = 15\,\text{Pa}$ at step $12,600$ of the adhesion-erosion-motion process. $N_s = 1$ around the initial seed and $N_b = 4$ along the biofilm body. The biofilm is merging with another seed growing at the opposite corner, which has been ignored in the plot: (**a**) front view, (**b**) side view

velocities of about 1 mm/s. The structure of the flow is represented in Fig. 1b–f. The density of the liquid is $10^3\,\text{kg/m}^3$ and its viscosity $\mu = 10^{-3}\,\text{Pa s}$. The bacterial size, and the tile size thereof, is taken to be $2\,\mu\text{m}$. The dimensions of the central straight fragment are $N \times M \times L\ \mu\text{m}$. Streamers grow mostly in the $N/3 \times M \times L\ \mu\text{m}$ region between corners. In real experiments, usual values for N, M and L are 600, 200 and 100. In the numerical tests selected here, we have divided those sizes by 2 to reduce the computational cost.

An initial biofilm seed is placed on the left corner at the bottom, see Fig. 1a. According to [12], the presence of secondary vortices in that area favors adhesion of particles to the wall, becoming a preferential adhesion site. Biomass will be attached to that seed, eroded and moved according to stochastic rules described above.

Numerical tests of biofilm growth are performed using this geometry, see Fig. 2. γ is a measure of the biofilm cohesion estimated from the biofilm Young modulus. Reference [12] gives values in the range 70–140 Pa. To reduce the computational cost, we adjust it so that our biofilms involve a small number of tiles. Images in Ref. [12] yield estimates for the adhesion time τ of 1 block of biomass per second. Each step of the adhesion-erosion-motion process occurs in a time scale τ.

Provided enough biomass attaches to the seed (to avoid streamer detachment) and to the biofilm body (to resist increasing erosion while crossing the current), the aggregate grows into the current, elongates with it, bends when it reaches the curve, approaches the opposite corner, and eventually merges with the additional biofilm seed that should be growing there. The observed effective growth rate is the balance between the biomass that attaches and detaches at each step, and varies during the spread process. It is usually larger before the thread tries to cross the main stream and decreases as it tries to reach the opposite corner while changing its shape.

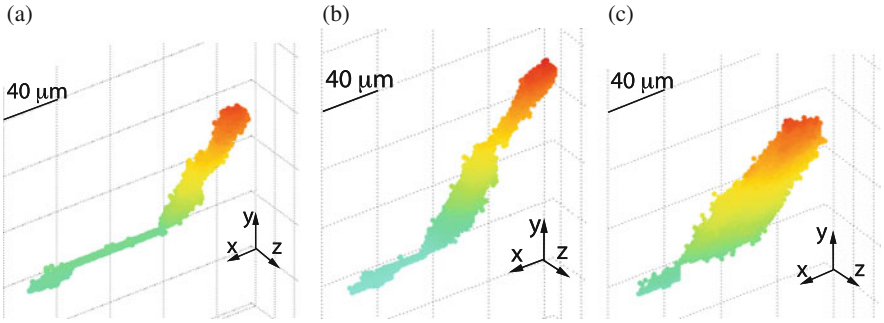


Fig. 3 Reducing the number of attached biomass blocks, streamers detach without reaching the opposite corner. (**a**) Decreasing N_b to 2, the streamer elongates, bends, detaches and regrows. The image corresponds to step 42, 600, just before the fourth detachment, with 1373 blocks. (**b**) Decreasing N_b to 3, the streamer becomes too thin and the top part encounters resistance to join the corner. It finally breaks off at step 15, 600, with 2151 blocks. (**c**) Decreasing N_s to 0.5 (one block attached each two steps), the connection of the streamer to the seed breaks off after step 9700 with 4792 blocks. Other parameter values as in Fig. 2. Distance between grid lines is always 40 μm

The aggregate grows into the region of minimum shear rate, that joins the two corners. Once formed, pressure variations move the filament downstream, curving it in a similar way to the experimentally observed threads, and leaving a thin joint with the seed. It reaches the opposite corner from behind, as observed in experimental photographs.

The number of biomass blocks to be attached depends on the selected biofilm cohesion. Too large values of N_b produce expanding balls. Too small adhesion rates to the biofilm N_b produce an elongated thread close to the wall, that eventually feels the corner flow and starts to gain biomass on the top, but may not receive enough biomass to resist the increased erosion and detaches, see Fig. 3a, b. For small values of N_s the connection between the streamer and the seed breaks off, see Fig. 3c. Too large adhesion rates to the seed N_s favor expansion parallel to the bottom substratum. If N_b is not large enough for the selected cohesion, the biofilm reaches the rightmost wall as shown in Fig. 4a. Increasing N_b, the biofilm may cross to the opposite corner sustained by a wider basis. If the initial adhesion rates are large enough for the considered cohesion, a sort of fan expands into the main stream. The fan becomes narrower as we reduce the adhesion rates.

Depending on the ratio N_b/N_s for the selected γ, we see narrower or wider streamers. If we increase the cohesion parameter γ, we must reduce the computational adhesion rates N_b and N_s to see similar behaviors. The failed streamer in Fig. 4a reaches successfully the opposite corner sustained by a wider basis when we slightly increase γ in Fig. 4b, c. If the biofilm cohesion is too small, the biofilm seed is eroded and eventually washed out. No thread is formed.

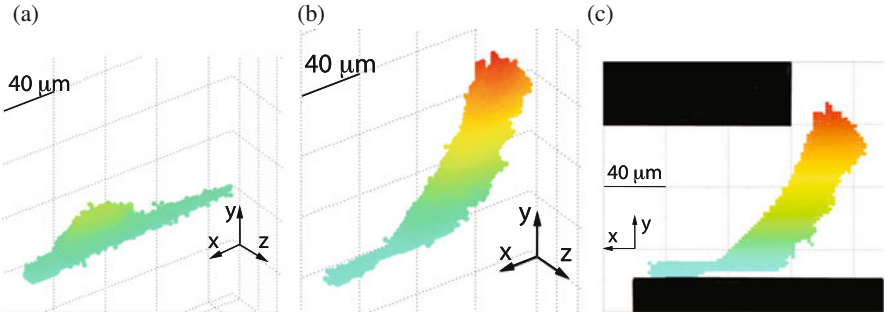


Fig. 4 (**a**) Increasing N_s to 2 the streamer remains parallel to the substratum until it reaches the wall at step 3200 with 3242 biomass blocks, for $\gamma = 15\,\text{Pa}$ and $N_b = 3$. Increasing γ to 20 Pa, the thread widens and crosses the current. (**b**) and (**c**) show the front and lateral views at step 15, 000, with 4702 blocks

These tests provide insight on the way these structures are formed. Threads experimentally observed [12], however, look more like thin jets and may require a different description. Streamers joining opposite corners appear to be attractor shapes that may be formed under different dynamics.

Acknowledgements D.R. Espeso and A. Carpio were supported by the Autonomous Region of Madrid and the Spanish MICINN through grants S2009/ENE-1597, and FIS2011-28838-C02-02. B. Einarsson was supported by a grant of the NILS program and project FIS2008-04921-C02-01.

References

1. Alpkvist, E., Picioreanu, C., Loosdrecht, M.C.M., Heyden, A.: Three-dimensional biofilm model with individual cells and continuum EPS matrix. Biotechnol. Bioeng. **94**, 961–979 (2006)
2. Castellon, E., Chavarria, M., de Lorenzo, V., Zayat, M., Levy, D.: An electro-optical device from a biofilm structure created by bacterial activity. Adv. Mater. **22**, 4846–4850 (2010)
3. Chorin, A.J.: A numerical method for solving incompressible viscous flow problems. J. Comput. Phys. **2**, 12–26 (1967)
4. Drescher, K., Shen, Y., Bassler, B.L., Stone, H.A.: Biofilm streamers cause catastrophic disruption of flow with consequences for environmental and medical systems. Proc. Natl. Acad. Sci. **110**, 4345–4350 (2013)
5. Eberhard, J.P., Efendiev, Y., Ewing, E., Cunningham, A.: Coupled cellular models for biofilm growth and hydrodynamic flow in a pipe. Int. J. Mult. Comput. Eng. **3**, 499–516 (2005)
6. Einarsson, B., Rodriguez, D., Carpio, A.: Biofilm growth on rugose surfaces. Phys. Rev. E **86**, 061914 (16 pp.) (2012)
7. Finlayson, B.A.: Introduction to Chemical Engineering Computing. Wiley, New York (2012)
8. Hermanovic, S.W.: A simple 2d biofilm model yields a variety of morphological features. Math. Biosci. **169**, 1–14 (2001)
9. Lecuyer, S., Rusconi, R., Shen, Y., Forsyth, A., Vlamakis, H., Kolter, R., Stone, H.A.: Shear stress increases the residence time of adhesion of *Pseudomonas aeruginosa*. Biophys. J. **100**, 341–350 (2011)

10. Morina, S., Pesceb, S., Tlilib, A., Costea, M., Montuelle, B.: Recovery potential of periphytic communities in a river impacted by a vineyard watershed. Ecol. Indic. **10**, 419–426 (2010)
11. Rodriguez, E.D.: Modeling and simulation of bacterial biofilms. Ph.D. Thesis, Universidad Carlos III de Madrid (2013)
12. Rusconi, R., Lecuyer, S., Autrusson, N., Guglielmini, L., Stone, H.A.: Secondary flow as a mechanism for the formation of biofilm streamers. Biophys. J. **100**, 1392–1399 (2011)
13. Schachter, B.: Slimy business-the biotechnology of biofilms. Nat. Biotechnol. **21**, 361–365 (2003)
14. Stoodley, P., Dodds, I., Boyle, J.D., Lappin-Scott, H.M.: Influence of hydrodynamics and nutrients on biofilm structure. J. Appl. Microbiol. Symp. Suppl. **85**, 19S–28S (1999)
15. Stoodley, P., Cargo, R., Rupp, C.J., Wilson, S., Klapper, I.: Biofilm material properties as related to shear-induced deformation and detachment phenomena. J. Ind. Microbiol. Biotechnol. **29**, 361–367 (2002)
16. Weinan, E., Lui, J.G.: Gauge method for viscous incompressible flows. Commun. Math. Sci. **1**, 317–332 (2003)

MS 16
MINISYMPOSIUM: METHODS FOR ADVANCED MULTI-OBJECTIVE OPTIMIZATION FOR eDFY OF COMPLEX NANO-SCALE CIRCUITS

Organizers

Salvatore Rinaudo[1] and Giuliana Gangemi[2]

Speakers

Helmut Graeb[3]
How Can We Include Pareto Front Computation, Discrete Parameter Values and Aging into Analog Circuit Sizing?

Carmelo Vicari[4], Mauro Olivieri[5], Zia Abbas[6] and M. Ali Khozoei[7]
Statistical Variation Aware ANN and SVM Model Generation for Digital Standard Cells

Giuliana Gangemi[2], Carmelo Vicari[4], Angelo Ciccazzo[8], Carmelo Vicari[4]
The MAnON Project: Methods for Advanced Multi-Objective Optimization for eDFY of Complex Nano-scale Circuits

[1]Salvatore Rinaudo, STMicroelectronics S.r.l, Catania, Italy.
[2]Giuliana Gangemi, STMicroelectronics S.r.l, Catania, Italy.
[3]Helmut Graeb, Technische Universitaet Muenchen, Muenchen, Germany.
[4]Carmelo Vicari, STMicroelectronics S.r.l, Catania, Italy.
[5]Mauro Olivieri, Sapienza University, Roma, Italy.
[6]Zia Abbas, Sapienza University, Roma, Italy.
[7]Mohammed Ali Khozoei, ITWM Fraunhofer Institute, Kaiserslautern, Germany.
[8]Angelo Ciccazzo, STMicroelectronics S.r.l, Catania, Italy.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_55

Mohammed Ali Khozoei[7], Matthias Hauser[9], Angelo Ciccazzo[8]
Waveform Modelling in Order to Speed Up Transient SPICE Simulations

Vittorio Latorre[10], Gianni Di Pillo[11] and Angelo Ciccazzo[8]
Yield Optimization in Electronic Circuits Design

Keywords

Behavioral models
Electrical design for yield
Multi objective optimization
Neural network
Process variation
Response surface models
Statistical analysis
Support vector machine

Short Description

Cost control, production efficiency, cycle time and yield are critical quality benchmarks for nano-electronics productions. An increasingly important downside of nano-CMOS technology scaling is the fact that the scaling of feature sizes cannot be accompanied by a suitable scaling of geometric tolerances. In addition, when getting into deep miniaturized dimensions, phenomena like edges or surfaces roughness, or the fluctuation of the number of doping atoms within the channels are becoming increasingly significant. As a result, the figures of merit of a circuit, such as performance and power, have become extremely sensitive to uncontrollable statistical process variations (PV).

To ensure stable manufacturability and secure high manufacturing yield, it is mandatory to manage complete design flows and to link traditional methods for design with Technology CAD models. In this context, multi-objective optimization algorithms and statistical analysis are essential on device and behavioural levels to secure high yielding by modelling the impact of inevitable process variations and doping fluctuations on IC performances. Statistical circuit modelling is a viable

[9]Matthias Hauser, ITWM Fraunhofer Institute, Kaiserslautern, Germany.
[10]Vittorio Latorre, Sapienza University, Roma, Italy.
[11]Gianni Di Pillo, Sapienza University, Roma, Italy.

solution to nano-electronics production quality, on which the European Community is already investing.

The CAD and Design Services group, part of the IPG R&D in STMicroelectronics, has created a consortium in order to develop, test and implement innovative "Methods for Advanced Multi-Objective Optimization for eDFY of complex Nanoscale Circuits": the MAnON Project.

The scope of the research activity has been to create "Process Variation"-aware and "Process Variation"-robust circuit design techniques, tools and models in the frame of the analogue and mixed-signal circuit industrial design.

The project has also been recognized to be relevant by the European Union which under the Marie Curie Action for the Industry-Academia Partnership, call 2009, has selected the MANON project and is co-funding the researchers directly involved in it, under the grant agreement FP7-MCA-IAPP 2009-251380. One hundred person months effort is directly funded by the European Union between Experienced and More Experienced Researchers.

How to Include Pareto Front Computation, Discrete Parameter Values and Aging into Analog Circuit Sizing

Helmut Graeb

Abstract Analog circuit sizing has strongly focused on the optimization of nominal performance and of the yield in the past. Recently, more topics in analog sizing have come up. These are Pareto optimization, optimization with discrete parameter values and consideration of aging effects in addition to manufacturing and operating tolerances. This contribution will illustrate these tasks and give problem formulations and solution approaches.

Keywords Analog circuit sizing • Multi objective optimization

1 Introduction

When talking about analog circuits, it should be noted that the described methods do not refer to analog circuits only, but to any type of circuit that is described with continuous signals and modeled with differential equations or netlists of compact circuit elements. Figure 1 gives examples of such circuits, as for instance, operational amplifiers, phase-locked loops, RF circuits, receiver frontends, and even digital gates or MEMS elements.

1.1 Parameter, Performances, Simulation

The core method to analyze such analog circuits is a numerical, SPICE-like simulation. We partition the descriptive variables of an analog circuit into *simulator input variables*, which we call

$$\textit{parameters} \quad \mathbf{x} \in \mathbb{R}^{n_x} \tag{1}$$

H. Graeb (✉)
Institute of Electronic Design Automation, Technische Universitaet Muenchen, München, Germany
e-mail: graeb@tum.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_56

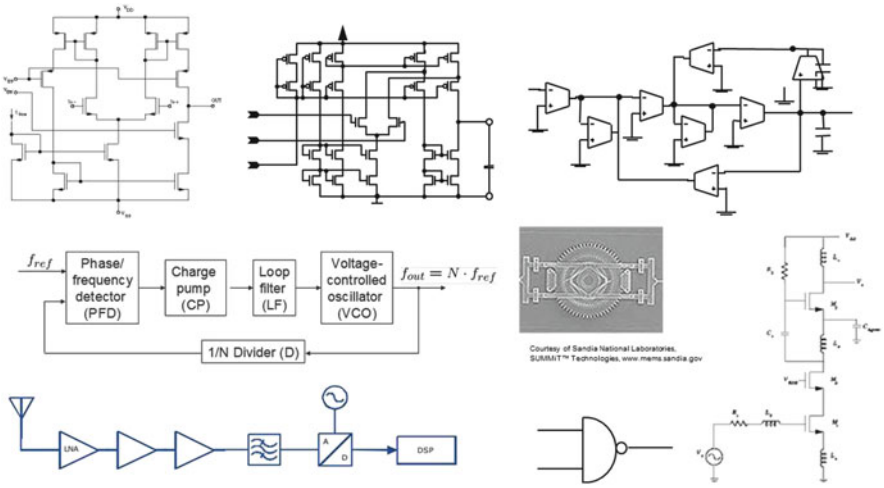


Fig. 1 Examples for circuits and systems that are designed on the abstraction level of analog circuits

and *simulator output variables*, which we call

$$\textit{performances} \quad \mathbf{f} \in \mathbb{R}^{n_f} \tag{2}$$

Hence, simulation maps parameters onto performances

$$\mathbf{x} \mapsto \mathbf{f} \tag{3}$$

Performances are for instance gain, slew rate, delay, power. Parameters are partitioned into three types,

- *design parameters* $\mathbf{x}_d \in \mathbb{R}^{n_{xd}}$, as for instance transistor geometries,
- *statistical parameters* $\mathbf{x}_s \in \mathbb{R}^{n_{xs}}$, as for instance threshold voltages, which model the statistical manufacturing variations with probability density functions, and
- *range parameters* $\mathbf{x}_r \in \mathbb{R}^{n_{xr}}$, as for instance temperature, which model the operating fluctuations with given intervals of values for which the circuit is specified to work properly.

Circuit simulation of analog circuits has two characteristics:

- It allows the *abstraction* from the physical level of description to a mathematical level.
- It is extremely *expensive* in terms of CPU time.

Hence application-specific optimization methods for circuit optimization need to be developed that specifically reduce the number of function evaluations during the optimization process.

1.2 Circuit Sizing, Constraints, Design Centering

Circuit sizing consists in optimizing the performance values (optimization objectives) by tuning the design parameters (optimization variables) while keeping constraints satisfied:

$$\min \mathbf{f}(\mathbf{x}_d) \text{ subject to } \mathbf{c}(\mathbf{x}_d) \geq \mathbf{0} \tag{4}$$

Here, we have assumed without loss of generality that performance should be minimal and constraints should be non-negative. Constraints are capturing basic design knowledge in form of requirements on geometries and voltages of transistors and specific transistor groups [3–5, 7, 10]. They are also representing minimum requirements on the performance values of a circuit. Sizing constraints are crucial for successful analog sizing.

Circuit sizing is a multicriteria optimization (MCO) problem, and it is a nonlinear optimization problem. One example of a practicable scalarization of the MCO problem is a least-squares approach, which aims at given target values, which are eventually updated during the optimization process:

$$\min \sum_{i=1}^{n_f} \left(f_i(\mathbf{x}_d) - f_{target,i}\right)^2 \text{ s.t. } \mathbf{c}(\mathbf{x}_d) \geq \mathbf{0} \tag{5}$$

Manufacturing and operation tolerances can be included by replacing the performances in Eqs. (4), (5) with so-called worst-case distances [1, 2, 6], which provide an x-sigma robustness and yield indicator for each performance and for the whole circuit. Analog sizing in this case becomes design centering, or, yield optimization.

Figure 2 shows the situation of an operational amplifier before and after optimization. We can see that the worst-case distances have been improved to at least 4-Sigma, which is the overall robustness of the circuit.

Several new requirements have emerged in analog sizing. Three of them are Pareto optimization, discrete parameter values, and aging.

Performance specification		Initial		Optimized	
Gain[dB]	≥ 65	76	2.5(„σ“)	76	4.2(„σ“)
Transit frequency [MHz]	≥ 30	67	7.7(„σ“)	58	4.5(„σ“)
Phase margin [°]	≥ 60	68	1.8(„σ“)	71	3.9(„σ“)
Slew rate [V/µs]	≥ 32	67	6.3(„σ“)	58	3.9(„σ“)
DC power [mW]	≤ 3,5	2.6	1.1(„σ“)	2.3	4.2(„σ“)
Overall yield		82.9%		99.9%	

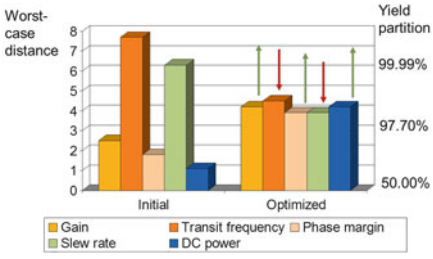


Fig. 2 Design centering of an operation amplifier. Performance, worst-case distance and yield values before and after optimization

2 Pareto Optimization

Circuit sizing leads to exactly one optimal trade-off between competing performances. Optimal trade-off means that one objective cannot be improved without degrading another one. However, often a whole set of possible trade-offs is required. This is illustrated in Fig. 3 (left). Each polytope corner represents one specific sizing run to one specific Pareto point. Three performances are shown. We can see that and how the folded-cascode structure of an operational amplifier yields better slew rate and that and how the Miller structure results in better gain values. Performance space exploration by means of Pareto optimization thus can contribute to selecting suitable circuit structures. Another application example is hierarchical sizing, where bottom-up Pareto optimization provides constraints for higher-level design parameters. This prevents top-level system optimization from producing unrealistic requirements on low-level implementations.

We have developed a method for analog Pareto optimization [8, 9, 11–15, 30–33] that has the following features:

- Pareto front is built successively, first all individual minima, then all two-dimensional fronts, then all three-dimensional fronts, and so on (Pareto fronts of dimension $n-1$ form borders of Pareto front of dimension n);
- deterministic optimization approach, which combines a goal attainment formulation with a minmax formulation, concurrent search threads exchange intermediate solutions, specific SQP solution;
- yield optimization/design centering is included by replacing performance values with worst-case performance values that refer to a required yield, increased evaluation effort is counteracted by a specific "lazy" worst-case analysis approach.

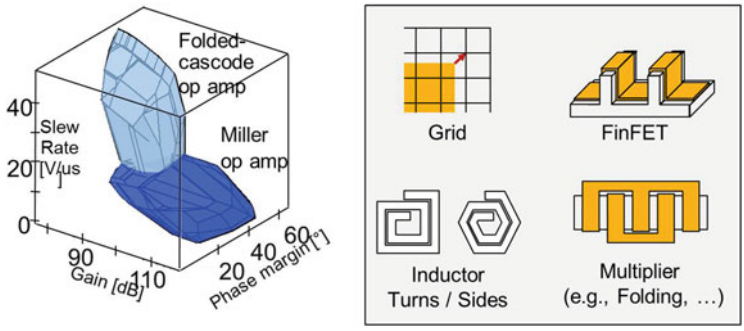


Fig. 3 *Left*: Comparison of performance spaces of two types of operation amplifiers. *Right*: examples of discrete parameters

3 Discrete Parameter Values

Discrete parameter values are frequently occurring in analog design. Figure 3 (right) illustrates four examples. For manufacturing reasons, the transistor shapes must lie on a grid. That means that the values have to be from a discrete set of values. New technologies like FinFETS just like the usual layout of transistors in terms of multipliers are discrete in that the sizing refers to a discrete number of fins or multipliers, not to a continuous space of values. The same holds for integrated inductors, where the sizing refers to a discrete number of turns or sides.

In fact, any analog sizing parameter is discrete. It is known that continuous optimization with subsequent rounding does not provide the optimal solution. We have developed a method for analog optimization with discrete parameters [23–29] with the following features:

- smooth objective function aiming at satisfying specification bounds;
- solution algorithm combining feasible SQP and Branch-and-Bound approach, reducing simulation cost by means of SQP model;
- two approaches depending on whether simulation can be done continuously or only discretely;
- yield optimization/design centering is included by replacing performance values with worst-case performance values that refer to a required yield, increased evaluation effort is counteracted by a specific "lazy" worst-case analysis approach.

4 Aging

Aging is becoming a critical issue in the ultradeep submicron era, even for analog circuits. We have developed a concept for aging that is based on the concept of lifetime yield [Fig. 4 (left)], where we analyze the percentage of circuits that work not only after production (test, fresh yield) but also for a certain age. Our solution approach [16–22] has the following features:

- fresh yield optimization for given area constraint;
- aging analysis during optimization only for sizing constraints (DC simulation);
- trade-off curve lifetime yield vs. required area [Fig. 4 (right)].

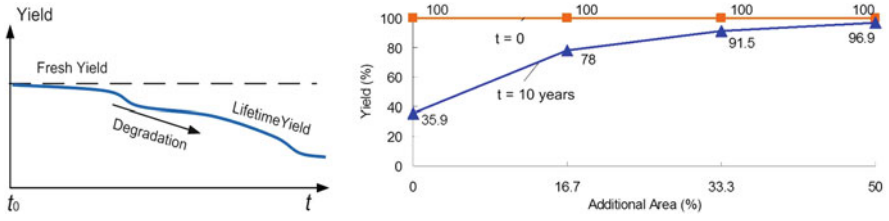


Fig. 4 *Left*: Yield degrades over lifetime. *Right*: Area vs. 10-years yield

References

1. Antreich, K., Graeb, H.: Circuit optimization driven by worst-case distances. In: IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pp. 166–169 (1991)
2. Antreich, K., Graeb, H., Wieser, C.: Circuit analysis and optimization driven by worst-case distances. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **13**(1), 57–71 (1994)
3. Eick, M., Strasser, M., Lu, K., Schlichtmann, U., Graeb, H.: Comprehensive generation of hierarchical placement rules for analog integrated circuits. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **30**(2), 180–193 (2011)
4. Eick, M., Graeb, H.: MARS: matching-driven analog sizing. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **31**(12), 1145–1158 (2012)
5. Eick, M., Graeb, H.: Towards automatic structural analysis of mixed-signal circuits. In: Fakhfakh, M., Tlelo-Cuautle, E., Castro-Lopez, R. (eds.) Analog/RF and Mixed-Signal Circuit Systematic Design, chap. 1, pp. 3–25. Springer, Berlin (2013)
6. Graeb, H.: Analog Design Centering and Sizing. Springer, Berlin (2007)
7. Graeb, H., Zizala, S., Eckmueller, J., Antreich, K.: The sizing rules method for analog integrated circuit design. In: IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pp. 343–349 (2001)
8. Graeb, H., Mueller, D., Schlichtmann, U.: Pareto Optimization of Analog Circuits considering Variability. In: European Conference on Circuit Theory and Design (ECCTD), pp. 28–31 (2007)
9. Graeb, H., Mueller-Gritschneder, D., Schlichtmann, U.: Pareto optimization of analog circuits considering variability. Int. J. Circuit Theory Appl. **37**(2), 283–299 (2009)
10. Massier, T., Graeb, H., Schlichtmann, U.: The sizing rules method for CMOS and bipolar analog integrated circuit synthesis. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **27**(12), 2209–2222 (2008)
11. Mueller, D., Stehr, G., Graeb, H., Schlichtmann, U.: Deterministic approaches to analog performance space exploration (PSE). In: ACM/IEEE Design Automation Conference (DAC), pp. 869–874 (2005)
12. Mueller, D., Stehr, G., Graeb, H., Schlichtmann, U.: Fast evaluation of analog circuit structures by polytopal approximations. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1479–1482 (2006)
13. Mueller, D., Graeb, H., Schlichtmann, U.: Trade-off design of analog circuits using goal attainment and wave front sequential quadratic programming. In: Design, Automation and Test in Europe (DATE), pp. 75–80 (2007)
14. Mueller-Gritschneder, D., Graeb, H.: Computation of yield-optimized Pareto fronts for analog integrated circuit specifications. In: Design, Automation and Test in Europe (DATE) (2010)
15. Mueller-Gritschneder, D., Graeb, H., Schlichtmann, U.: A successive approach to compute the bounded Pareto front of practical multi-objective optimization problems. SIAM J Optim. **20**(2), 915–934 (2009)
16. Pan, X., Graeb, H.: Degradation-aware analog design flow for lifetime yield analysis and optimization. In: IEEE International Conference on Electronics, Circuits and Systems (ICECS) (2009)
17. Pan, X., Graeb, H.: Lifetime yield optimization: towards a robust analog design for reliability. In: Design, Automation and Test in Europe (DATE) University Booth (2010)
18. Pan, X., Graeb, H.: Reliability analysis of analog circuits by lifetime yield prediction using worst-case distance degradation rate. In: IEEE International Symposium on Quality Electronic Design (ISQED) (2010)
19. Pan, X., Graeb, H.: Reliability analysis of analog circuits using quadratic lifetime worst-case distance prediction. In: IEEE Custom Integrated Circuits Conference (CICC) (2010)
20. Pan, X., Graeb, H.: Lifetime Yield Optimization of Analog Circuits Considering Process Variations and Parameter Degradations, chap. 6, pp. 131–146. InTech (2011)

21. Pan, X., Graeb, H.: Reliability optimization of analog circuits with aged sizing rules and area trade-off. In: edaWorkshop, pp. 33–38. VDE Verlag GmbH (2011)
22. Pan, X., Graeb, H.: Reliability optimization of analog integrated circuits considering the trade-off between lifetime and area. Microelectron. Reliab. **52**(8), 1559–1564 (2012)
23. Pehl, M., Graeb, H.: RaGAzi: A random and gradient-based approach to analog sizing for mixed discrete and continuous parameters. In: International Symposium on Integrated Circuits (ISIC) (2009)
24. Pehl, M., Graeb, H.: Dimensionierung Analoger Schaltungen mit diskreten Parametern unter Verwendung eines Zufalls- und Gradientenbasierten Ansatzes. In: ITG/GMM-Fachtagung Entwurf von analogen Schaltungen mit CAE-Methoden (ANALOG) (2010)
25. Pehl, M., Graeb, H.: An SQP and branch-and-bound based approach for discrete sizing of analog circuits, chap. 13, pp. 297–316. InTech (2011)
26. Pehl, M., Graeb, H.: Tolerance design of analog circuits using a branch-and-bound based approach. J. Circuits Syst. Comput. **21**(8), 1240022, 17p. (2012)
27. Pehl, M., Massier, T., Graeb, H., Schlichtmann, U.: A random and pseudo-gradient approach for analog circuit sizing with non-uniformly discretized parameters . In: IEEE International Conference on Computer Design (ICCD), pp. 188–193 (2008)
28. Pehl, M., Zwerger, M., Graeb, H.: Sizing analog circuits using an SQP and branch and bound based approach. In: IEEE International Conference on Electronics, Circuits and Systems (ICECS) (2010)
29. Pehl, M., Zwerger, M., Graeb, H.: Variability-aware automated sizing of analog circuits considering discrete design parameters. In: International Symposium on Integrated Circuits (ISIC) (2011)
30. Zou, J., Mueller, D., Graeb, H., Schlichtmann, U., Hennig, E., Sommer, R.: Fast automatic sizing of a charge pump phase-locked loop based on behavioral models. In: IEEE International Behavioral Modeling and Simulation Conference, pp. 100–105 (2005)
31. Zou, J., Mueller, D., Graeb, H., Schlichtmann, U.: A CPPLL hierarchical optimization methodology considering jitter, power and locking time. In: ACM/IEEE Design Automation Conference (DAC), pp. 19–24 (2006)
32. Zou, J., Mueller, D., Graeb, H., Schlichtmann, U.: Optimization of SC $\Sigma\Delta$ modulators based on worst-case-aware Pareto-optimal fronts. In: IEEE Custom Integrated Circuits Conference (CICC), pp. 607–610 (2007)
33. Zou, J., Mueller, D., Graeb, H., Schlichtmann, U.: Pareto-front computation and automatic sizing of CPPLLs. In: IEEE International Symposium on Quality Electronic Design, pp. 481–486 (2007)

Statistical Variation Aware ANN and SVM Model Generation for Digital Standard Cells

C. Vicari, M. Olivieri, Z. Abbas, and M. Ali Khozoei

Abstract Progressive CMOS technology scaling leads to high increase in statistical variations, whose impact on circuit performances must be taken into account already in the design phase. Reliable surrogate models can replace expensive circuit simulations to statistically characterize the figures of merit of a circuit with a reduced computational effort. We implemented a software framework which allows the automatic generation of surrogate models based on machine learning techniques such as Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs). These methodologies have been used to generate statistical variation aware models for leakages and propagation delays of a set of digital standard cells.

Keywords Artificial neural networks • Digital standard cells • Support vector machines

1 Introduction

Model generation for the circuit behavior is used to speed up analyses or optimizations by performing them on models instead of running circuit simulations. However, this speed-up is normally obtained at the penalty of less accurate analysis or optimization results. This error in accuracy depends on the model error. Generation of reliable models with reduced effort, ensuring model accuracy as high as possible is one of the challenging tasks especially with the enormous increase in statistical variations due to technology scaling [5].

C. Vicari (✉)
STMicroelectronics, Stradale Primosole 50, 95121 Catania, Italy
e-mail: carmelo.vicari@st.com

M. Olivieri • Z. Abbas
University La Sapienza, Via Eudossiana 18, 00184 Roma, Italy
e-mail: olivieri@diet.uniroma1.it; abbas@die.uniroma1.it

M.A. Khozoei
Fraunhofer ITWM, Fraunhofer-Platz 1, 67663 Kaiserlautern, Germany
e-mail: khozoei@itwm.fraunhofer.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_57

Transistors in circuits behave and interact in a complex manner and make circuit performances such as propagation delays and power dissipation more and more sensitive to these process variations [2, 8]. Furthermore, for the generation of a reliable model, the impact of both the local (mismatch) and global process variations must be considered thoughtfully. Global variations (e.g. oxide thickness) are chip-to-chip, wafer-to-wafer or batch-to-batch variations, while local variations (e.g. threshold voltages) may affect every device in a chip individually [1, 5].

In general, it is difficult to predict the relationship between process variations and circuit performances whose statistical distributions can be estimated via computationally expensive Monte Carlo simulations. The huge number of circuit simulations needed to perform a Monte Carlo analysis can be reduced by creating surrogate models which approximate the performances of a circuit as function of the statistical parameters. In these cases, Machine Learning techniques, such as *Artificial Neural Networks* (ANNs) [3] and *Support Vector Machines* (SVMs) [9], are promising approaches for building surrogate models with a reduced computational effort. The generation of such kind of models is not trivial and requires several tasks to obtain satisfactory results.

We have implemented a software framework which simplifies and automatizes the generation of surrogate models based on ANNs and SVMs for integrated circuits. We have used these methodologies to generate statistical variation aware models for leakages and propagation delays of a set of digital standard cells.

2 Model Generation

Figure 1 shows the architecture of the software framework used for the generation of surrogate models. It is based on WiCkeD^{TM}[7], a widely used EDA software tool for circuit analysis, modeling and optimization.

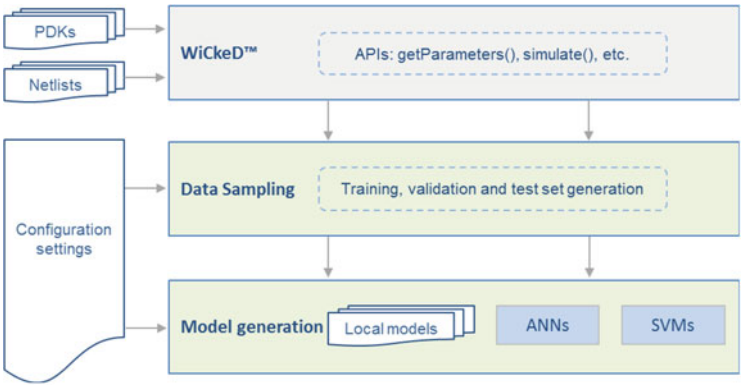


Fig. 1 The architecture of framework used to generate the models

The model generation flows starts from the left upper corner of Fig. 1. Here WiCkeD™ receives two primary inputs: (1) a set of transistor level netlists of the circuits to be modeled and (2) fabrication process information usually contained in a process design kit (PDK). Each netlist represents a specific analysis to be performed in order to measure the circuit performances of interest.

WiCkeD™ allows the selection of global and local process parameters and provides a set of application programming interfaces (APIs) to the other modules of the framework to execute different actions: running circuit simulations, getting parameter and performance information, executing sensitivity analyses to remove the process parameters which have a negligible influence on the performances of interest. It constitutes a common access to circuit information independently of the PDKs and simulators used.

The **Data sampling** module is a collection of methods whose main goal is to select the points which are representative of the process parameter space and obtain, for each point, the corresponding values of circuit performances from computer simulations. Different algorithms can be chosen for sampling the parameter space: random, regular grid and Latin Hypercube Sampling (*LHS*) [10] methods. Sample points and the corresponding simulation results are put together in unique objects that are used in the model generation phase as training, validation and test data sets.

The **Model generation** module implements a set of algorithms for the actual generation of surrogate models. A model can be seen as a black box which approximates the relationship between a set of input parameters (which in general are process parameters, but can be also design or operating parameters) and a set of outputs represented by circuit performances.

For each performance to be modeled a set of SVMs or ANNs are trained with the same (or eventually different) training set and the generalization error is calculated by using different error measures (i.e., mean relative error, max relative error, max absolute error and others). The model showing the lowest error is automatically set as favorite approximating function for the given performance.

To improve the accuracy, the framework allows the generation of a set of local models, each valid in a sub-region of the parameter space. When the model is used to predict the value of a performance corresponding to a point in the parameter space that falls in a sub-region where a local model exists, the framework automatically uses the local model for calculating the value of the performance.

Input and output parameters can be transformed using appropriate mathematical functions to simplify the learning process and a set of model parameters can be fine tuned to avoid phenomena such as overfitting or underfitting. Transformation functions, error types and thresholds, model parameters and other settings can be specified via a configuration module. The type of SVMs and ANNs used are briefly described in the following sections.

2.1 Type of Support Vector Machines Used

SVMs are mostly commonly used for binary classifications but there is one branch of SVM, *SVM regression* or *SVR*, which is able to fit a continuous function. We use ϵ-*SVR* [11], a support vector formulation for SVM regression. In a regression problem an SVM is trained with a set of points, $(\mathbf{x}^i, y^i), i = 1, \ldots, l$ being l the number of training samples, $\mathbf{x}^i \in \mathbb{R}^N$ the feature vector and $y^i \in \mathbb{R}$ the target output. The approximate function $y = f(\mathbf{x})$ is computed as follows:

$$y = f(\mathbf{x}) = \sum_{i=1}^{l} (\hat{\lambda}_i^* + \lambda_i^*) k(\mathbf{x}, \mathbf{x}^i) + b^*$$

where $(\hat{\lambda}_i^*, \lambda_i^*)$ is the solution of the following quadratic optimization problem:

$$\min_{\lambda, \hat{\lambda}} \Gamma(\lambda, \hat{\lambda}) = \min_{\lambda, \hat{\lambda}} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\hat{\lambda}_i - \lambda_i)(\hat{\lambda}_j - \lambda_j) k(\mathbf{x}^i, \mathbf{x}^j) - \sum_{i=1}^{l} (\hat{\lambda}_i - \lambda_i)\, y^i + \epsilon \sum_{i=1}^{l} (\hat{\lambda}_i + \lambda_i)$$

$$\sum_{i=1}^{l} (\hat{\lambda}_i - \lambda_i) = 0, \quad 0 \le \lambda \le C, \quad 0 \le \hat{\lambda} \le C, \quad i = 1, \ldots, l$$

where $\epsilon > 0$ and $C > 0$ are given parameters and $k(x, z) = e^{(-\gamma * |x - z|^2)}$ is a radial basis kernel function. In our work we have set $\epsilon = 0.0001$ and we have determined the values of C and γ via a fivefold cross validation technique. For the training and evaluation of SVM models we use the library *LIBSVM* [4].

2.2 Type of Artificial Neural Networks Used

We create ANNs models by using feed-forward neural networks whose structure is depicted in Fig. 2. In this kind of networks the input signal moves in only one direction and the output of the network, given by the output of the last layer neurons,

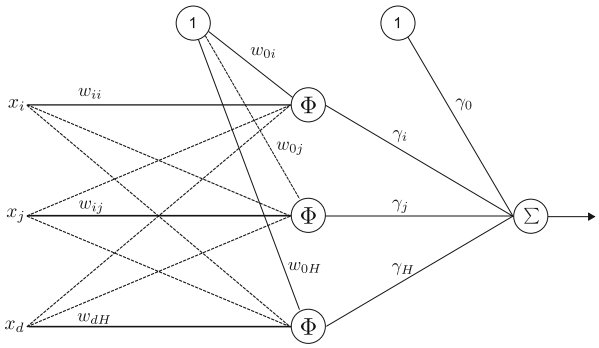


Fig. 2 The structure of the Artificial Neural Network used

is computed as showed in (1)

$$y = f(\mathbf{x}, \theta) = \gamma_0 + \sum_{h=1}^{H} \gamma_h \Phi(\mathbf{x}\, \mathbf{w}_h) \tag{1}$$

where $\theta = (\mathbf{w}_1, \ldots, \mathbf{w}_H, \gamma_0, \ldots, \gamma_H)$ is the vector of network weights, with $\mathbf{w}_h = (w_{oh}, \ldots, w_{dh})^T$ for $h = 1, \ldots, H$ and H denotes the number of neurons in the hidden layer and is a number proportional to the number of training samples and inputs of the model. $\mathbf{x} = (1, x_1, \ldots, x_d)$ is the vector of inputs and $\Phi(x) = 1 - \frac{2}{e^{2x}+1}$ is the activation function of each hidden unit and vector. Given a training set, the weights of the networks are estimated via the *Levenberg-Marquardt* back propagation method [6].

3 Digital Standard Cells Modeling

We have used the framework to statistically characterize a set of digital standard cells. We have estimated *propagation delays* and pattern dependent *leakage powers* of a NOT, and 2-inputs NOR and NAND cells as a function of the statistical variations. Each cell includes nine global process parameters and five local parameters per single transistor. All the models have been created for a fixed size of the digital cells and at nominal values of the operating parameters. The fabrication process information has been accessed from a 40 nm low power (LP) standard threshold voltage (SVT) CMOS PDK. ANNs and SVMs models have been compared to Response Surface Models (RSM), a widely used methodology for surrogate model generation.

Training data have been generated by using the Latin Hypercube Sampling algorithm. The total number of process parameters taken into consideration are 19 for the NOT gate and 29 for NOR and NAND cells. Only for leakage models we have performed a parameter screening by removing all the process parameters having an influence less than 1 %. In addition a **log** transformation of leakages has produced a considerable improvement of models accuracy.

In the following we report the modeling results for each standard cell. We denote with N_{train} the total number of training samples used to create a given model and with μ *error* and σ *error* the relative errors on the mean and standard deviation of a test set of 1000 samples.

The terms $leakage_0$ and $leakage_1$ denote the leakage power of the NOT gate when its input is 0 and 1 respectively, while t_{PHL} and t_{PLH} denote the high-to-low and low-to-high propagation delays, as shown in Table 1.

Similarly $leakage_{00}$, $leakage_{01}$, $leakage_{10}$ and $leakage_{11}$ corresponds to input patterns 00, 01, 10 and 11 in respective 2-input cells, while t_{PHL_A}, t_{PHL_B}, t_{PLH_A}, t_{PLH_B} represent the high-to-low and low-to-high propagation delays for inputs A and B. Tables 2 and 3 show the results of NOR and NAND cells respectively.

It is obvious from reported results, that ANN and SVM models are able to predict with sufficient accuracy the given performance figures of digital cells especially leakages, where RSM models are not accurate enough.

4 Conclusions

We have presented a framework for the generation of statistical aware models based on machine learning techniques such as ANNs and SVMs. The implemented methodologies have been used to statistically characterize a set of digital standard cells from a reduced number of circuit simulations. The obtained models show a good accuracy for different performances such as propagation delays and leakage powers.

Table 1 NOT cell

	RSM			ANN			SVM		
Performance	N_{train}	μ *error* (%)	σ *error* (%)	N_{train}	μ *error* (%)	σ *error* (%)	N_{train}	μ *error* (%)	σ *error* (%)
$leakage_0$	238	13.49	65.36	138	0.04	0.99	138	0.27	1.30
$leakage_1$	238	15.46	50.86	138	0.06	1.06	138	0.20	0.12
t_{PHL}	100	3.16	7.87	50	0.10	0.53	50	0.05	0.53
t_{PLH}	100	2.83	1.04	50	0.12	1.70	50	0.02	0.15

Number of training samples (N_{train}), μ *error* (%) and σ *error* (%) for each performance and modeling methodology. Error measures are computed on a common test set of 1000 samples

Table 2 NOR2 cell

	RSM			ANN			SVM		
Performance	N_{train}	μ *error* (%)	σ *error* (%)	N_{train}	μ *error* (%)	σ *error* (%)	N_{train}	μ *error* (%)	σ *error* (%)
$leakage_{00}$	258	8.73	46.19	158	2.13	13.96	158	0.51	2.56
$leakage_{01}$	258	1.88	24.91	158	0.14	0.72	158	0.44	1.44
$leakage_{10}$	258	3.16	30.47	158	0.50	3.19	158	0.12	1.82
$leakage_{11}$	258	0.36	14.82	158	0.07	1.49	158	0.16	0.04
t_{PHL_A}	150	0.07	6.91	100	0.03	1.45	100	0.001	1.08
t_{PHL_B}	150	0.30	1.78	100	0.05	2.76	100	0.01	1.29
t_{PLH_A}	150	0.61	2.01	100	0.06	1.07	100	0.01	0.61
t_{PLH_B}	150	0.54	1.13	100	0.12	1.29	100	0.06	0.75

Number of training samples (N_{train}), μ *error* (%) and σ *error* (%) for each performance and modeling methodology. Error measures are computed on a common test set of 1000 samples

Table 3 NAND2 cell

	RSM			ANN			SVM		
Performance	N_{train}	μ *error* (%)	σ *error* (%)	N_{train}	μ *error* (%)	σ *error* (%)	N_{train}	μ *error* (%)	σ *error* (%)
$leakage_{00}$	258	5.85	18.82	158	0.83	9.95	158	0.16	4.54
$leakage_{01}$	258	0.76	49.10	158	1.24	12.95	158	0.41	4.33
$leakage_{10}$	258	3.62	48.30	158	0.59	6.88	158	0.16	2.87
$leakage_{11}$	258	9.17	38.46	158	0.19	4.46	158	0.12	1.19
t_{PHL_A}	150	0.37	18.69	100	0.11	3.05	100	0.01	0.95
t_{PHL_B}	150	7.78	4.65	100	0.04	1.92	100	0.0048	0.82
t_{PLH_A}	150	2.09	2.77	100	0.10	0.86	100	0.008	1.22
t_{PLH_B}	150	1.87	2.34	100	0.07	0.0008	100	0.08	0.51

Number of training samples (N_{train}), μ *error* (%) and σ *error* (%) for each performance and modeling methodology. Error measures are computed on a common test set of 1000 samples

Acknowledgements This work has been carried out within the project *MAnON—Methods for Advanced Multi-Objective Optimization for eDFY of complex Nanoscale Circuits*, supported by the European Commission under grant no. 251380.

References

1. Abbas, Z., Olivieri, M., Yakupov, M., Ripp, A.: Design centering/yield optimization of power aware band pass filter based on CMOS current controlled current conveyor (CCCII+). Microelectron. J. **44**(4), 321–331 (2013). doi:http://dx.doi.org/10.1016/j.mejo.2012.11.004
2. Abbas, Z., Mastrandrea, A., Olivieri, M.: A voltage-based leakage current calculation scheme and its application to nanoscale MOSFET and FinFET standard-cell designs. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **22**(12), 2549–2560 (2014). doi: 10.1109/TVLSI.2013.2294550
3. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1995)
4. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
5. Graeb, H.E.: Analog Design Centering and Sizing, 1st edn. Springer Publishing Company, Heidelberg, Incorporated (2007)
6. Hagan, M., Menhaj, M.B.: Training feedforward networks with the Marquardt algorithm. IEEE Trans. Neural Netw. **5**(6), 989–993 (1994). doi: 10.1109/72.329697
7. Muneda (2014). http://www.muneda.com
8. Olivieri, M., Mastrandrea, A.: Logic drivers: a propagation delay modeling paradigm for statistical simulation of standard cell designs. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **22**(6), 1429–1440 (2014). doi: 10.1109/TVLSI.2013.2269838
9. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
10. Stein, M.: Large sample properties of simulations using latin hypercube sampling. Technometrics **29**(2), 143–151 (1987). http://www.jstor.org/stable/1269769
11. Vapnik, V.N.: Statistical Learning Theory. Wiley, Chichester (1998)

The MAnON Project

Methods for Advanced Multi-Objective Optimization for eDFY of Complex Nano-scale Circuits

Giuliana Gangemi, Carmelo Vicari, Angelo Ciccazzo, and Salvatore Rinaudo

Abstract The nano-CMOS technology scaling makes the figures of merit of a circuit, such as performance and power, extremely sensitive to uncontrollable statistical process variation (PV). In this context, multi-objective optimization algorithms and statistical analysis are essential to ensure stable manufacturing and secure high foundry yields. The CAD and Design Services group, part of the IPG R&D in STMicroelectronics, has created a consortium in order to develop, test and implement *"Methods for Advanced Multi-objective Optimization for eDFY of Complex Nano-scale Circuits"*: **the MAnON Project**. The contribution presents the industrial and scientific project challenges, the research results, and consequent methodology enhancements and their implementation into a software prototype in order to be usable inside a nanoelectronics industrial design environment.

Keywords Multiobjective optimization • Nanoscale circuits

1 Introduction: The MAnON Project

The project MAnON is the joint venture between academies, University La Sapienza in Rome, the Fraunhofer ITWM (Fraunhofer Institute for Industrial Mathematics) in Kaiserslautern, one leading edge EDA (Electronic Design Automation) Software House, which is MUNEDA GmbH, and the semiconductor company STMicroelectronics industry so to create a Transfer of Knowledge between the organizations in order to pass the mathematical knowledge on multi-objective optimization, symbolic techniques and numerical statistical simulation on one side, the industrial design experience, real test cases availability and EDA software modeling skills on the other.

G. Gangemi (✉) • C. Vicari • A. Ciccazzo • S. Rinaudo
STMicroelectronics S.r.l., Catania, Italy
e-mail: giuliana.gangemi@st.com; carmelo.vicari@st.com; angelo.ciccazzo@st.com; salvatore.rinaudo@st.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_58

The aim of the design activity is to optimize the performances of a circuit, such as gains, leakages and propagation delays, under all operating conditions and taking into account the statistical process variations (eDFY). Such an optimization is done simulating circuit models which are also PV aware. The creation of PV-aware models with traditional Monte Carlo techniques is very costly and time consuming, for this reason different methods are used to create such PV aware models.

The Consortium has been built in order to have the necessary multidisciplinary composure of knowledge and expertise. The activities have been carried as standard industrial EDA development project.

We have divided the entire research activities in four phases where theoretical activities of research have been followed by concrete and applied activities making the researchers exercise the acquired knowledge and skills (**learning by doing**). **Phase I** has included a preliminary study dedicated to the know-how acquisition, knowledge exchange and training activities of the researchers, these leading to the *specification of the methodological developments*. **Phase II** has started the *development of the new specific methods* and specified their implementation in a prototype tool and demonstrator, prerequisite of the **Phase III** where a *prototype/demonstrator* has been realized including provision of the test benches. To conclude, in **Phase IV**, the *verification and validation* of the prototype tools and methods is executed. At the time the present paper is being written the project has completed the development of the prototype and has begun its validation and benchmarking.

1.1 Methods, Tool and Test Specification

1.1.1 Simulation Challenge

The challenge of the project is to develop behavioral models for the selected test cases reducing the number and the duration of the simulations necessary to develop PV aware models (criteria number 2 and 3 in the following "Success Criteria" table) without having to reduce the number of variables to handle (criterion number 1) or neglecting the accuracy (criterion number 5), and maintaining the complexity of the process as much simple as possible (criterion number 4). The Measure of success set to evaluate the methodology enhancements are summarized in Table 1.

1.2 Methodology Enhancement: Research Development

Academic and industrial state of the art optimization and methods to generate circuits' behavioral models have been studied and analyzed on selected industrial test cases. The research activities started from the analysis of what is currently the state of art in the industrial design flow for the generation of e-DFY models and from its limitations. The cases presented, circuit and PDK do not allow the

Table 1 MAnON success criteria table

Success criteria table		
NUM	Field of application	Target (%)
1	Number of variables managed	>20
2	Time requirement	<20
3	HW requirement	<20
4	Number or steps of complexity	<10
5	Accuracy	>10

generation of accurate e-DFY models with the EDA tool Wicked which is based on the Response Surface Methodology (RSM) with an adequate accuracy to meet nowadays requirements of the design and optimization phase. For this reason, three new different methodologies have been investigated in this project and the different pros and cons are reported below. The methodologies taken into account are:

1. A combination of *Support Vector Machine* (*SVM*) surrogate models and a *Derivative-free mixed-integer black-box optimization* algorithm to be used for faster circuits yield estimation.
2. The usage of *Symbolic Model Order Reduction* (SMOR) techniques and *Neural Networks* (NN) for reducing the complexity of the system of differential equations describing the behavior of an integrated circuit, thus reducing drastically the simulation time.
3. Enhance the RSM models accuracy using RBF (radial basis functions) with automatic width variation, per function and/or input parameter. In addition, research into *table-based models* enabling:

   - Run-time check of coverage.
   - Incremental model updates.
   - Global *Process Variation* (PV) using interpolated sensitivity matrix.
   - Mismatch using interpolated covariance matrix and new performance distributions.

   A summary of the new methodologies is shown in Table 2.

We developed the enhancements able to manage a greater number of **design variables** (*MOS widths and lengths etc. etc.*), **environment variables** (*bias Voltage, temperature, etc. etc.)* and **process variables** (*statistically described through Gaussian or uniform distribution*), and better performing in terms of time requirements and maybe in term of needed hardware resources in order to complete the extraction of the behavioral models. The results of these new methodologies have been implemented in a SW prototype and demonstrator and is currently being validated on selected industrial test cases.

Table 2 Methodology features at a glance

	UNIRM	ITWM	MUNEDA
Method	• SVM (Support Vector Machine) • Derivative-free mixed-integer black-box optimization algorithms	• SMOR (Symbolic Model order reduction techniques) • NN (Neural Network)	• RBF (radial basis functions) with automatic width variation, per function and/or input parameter
Benefit	• Yield Analysis Optimization • Timing enhancements	• Timing • HW costs • Extended analysis by symbolic handling	• Improved fitting
Longer term	• Research to extend to the standard cell digital flow by including yield optimization criteria	• Include dynamics to data driven models	• Research into table-based models enabling: • Run-time check of coverage • Incremental model updates • Global PV using interpolated sensitivity matrix • Mismatch using interpolated covariance matrix and new performance distributions

With the results obtained on the selected test cases we can state that:

1. the new methodologies are able to manage a greater number of design variables (MOS widths and lengths etc. etc.), environment variables *(bias Voltage, temperature, etc. etc.)* and process variables *(statistically described through Gaussian or uniform distribution).*
2. the new methodologies are better performing in terms of time requirements and in terms of necessary hardware resources, in order to complete the behavior models extraction.

1.2.1 Foreword: Correct Interpretation of the Present Results

For sake of readability we have used the success criteria listed in Table 1 as a guideline but the information in the current phase of the project has to be interpreted. The table was made to set up front, at the beginning of the project, some objectives and tangible measures of success that could also drive the research activities.

In the present paragraph these criteria are used to give a rough measure of the goodness of the innovative methodologies and to justify the implementation in the

prototype but *its final purpose will be to measure the exercise of the enhanced methodologies* ***on a large number of tests****, activity that will only be possible at the end of the fourth phase of the project.* The main result obtained, up till now, is that we demonstrated in the documented tests that the NN and SVM methodologies are as much as accurate as the reference RSM (currently available) and that, case by case, topology by topology, circuit by circuit each of them offers some advantages either in time or number of variable that can be taken into account with respect to the generation time. Said the accuracy is maintained, all these trade-offs taken into account are enough to justify the implementation of the methodologies in a prototype, WiCkeD based, that will give the end user, the designer, the combined and increased advantage to have the possibility to choose the method to generate the models, according to the circuit or the case study need, altogether reusing the training set computation time for both methodologies (main gain of time).

1.2.2 Support Vector Machines

The results indicate that the proposed method is able to find a solution for the yield optimization problem comparable with the solution obtained by the benchmark WiCkeD, **with a reduction in computational effort of the order of 25 % and fulfilling the success criteria number 2, 3, 4 while we partially enhanced the number of variables managed, to some extent achieving the success criteria number 1.** *Overall it represents a reasonable success which justifies the implementation in a prototype.*

1.2.3 Neural Network

The results of this methodology have been slightly different depending on the test circuit on which it has been applied. Depending on the shape of the function which relates the independent variables (design, operating and process parameters) to the circuit performances, the Neural Networks have performed in some cases better and in other cases worse than the other modeling technologies. In general we have noticed that it has shown **enhancements with respect to the success criteria 2, 3 reducing the simulation time and with respect to the criteria number 4 maintaining the accuracy with respect the reference methodology.** *Overall it represents a reasonable success which justifies the implementation in a prototype.*

1.2.4 Response Surface Model (Behavioral Models Enhancements)

The RSM methodology seems to have reached its limit in term of the MAnON success criteria, but remains the industrial reference and it has been enhanced by using RBF (radial basis functions) with automatic width variation, per function and/or input parameter.

1.2.5 Conclusions: Methodologies Comparison Table

Although our experiments have shown incremental improvements with respect the reference benchmark, we believe that approaches based solely on a modified Response Surface Model implementation, or replacing it with Support Vector Machines (SVM) and Artificial Neural Network (ANN), are not sufficient to guarantee acceptable results over a large set of industrial test cases. In other words, improvements on behavioral models algorithms are still useful, but pursuing this single "attack front" has little chance of practical success. This is mainly due to the high-dimensionality of the space of input parameters. For this reason, we have proposed another solution focusing on workflow mechanisms to enable using behavioral models optimized for speed and flexibility as local interpolators. Local behavioral models can be created by using the most performing technique, chosen between RSM, SVM and ANN, for each specific test case.

Table 3 summarizes, with the clauses explained in the foreword the pros and cons of the different methods that will be made available to the designers within the MAnON prototype and why we consider valuable to implement such methods in the prototype. **As said above at the end of the project the real enhancement to the design flow will be provided by the tool itself that shall offer the designer the possibility to choose the best solution offered by each method according to the circuit and the question to be solved, reducing the training time.**

Table 3 Methodologies success comparison; refer to the foreword for interpretation

Success criteria table				
NUM	Field of application	Target (%)	NN	SVM
1	Number of variables managed	>20		ν[a]
2	Time requirement	<20	ν	ν
3	HW requirement	<20	ν	ν
4	Number of steps of complexity	<10	ν	ν
5	Accuracy	>10	[b]	[b]

[a]Partially

[b]Maintained the accuracy of the state of the art methodology used as benchmark

1.2.6 Additional Methodological Investigation and Test Cases

During the project activities it has been judged valuable by the consortium to add some investigations to what originally planned. On one side some effort has been dedicated to the possibility of using Neural Networks for modeling a transient analysis and on the other side the consortium has looked at extending the MAnON eDFY methodologies to the digital domain and build a CMOS library as additional test cases.

Using Neural Networks for Modelling a Transient Analysis
A novel approach based on the combined usage of Neural Networks and Bézier curves to approximate circuit waveforms obtained from transient analyses of analogue circuits has been investigated. The proposed methodology allows the generation of behavioral models that can be used to speed-up the large number of transient simulations needed for circuit analyses and optimization. The outputs of such models are not just numerical values of measured performances (spike, settling time, Vout stabilized) but time dependent functions. The results are very promising and will set the basis of future research and development activities.

Development of Additional Test Cases
The design of a CMOS standard cell library has been pursued as alternative test bench for optimization and modelling. The design of the library from scratch has been chosen to have full control over transistor sizing and technology parameter variations, for optimization and modelling experiments. The target of this activity was to set up a circuit test case with strong non-linearity for optimization and modelling in the context Design for Yield. The circuits have been characterized in 45 nm technology. Experiments on the optimization of digital standard cells (flip-flop) taken from the library have been conducted leading to partially successful results. The application of surrogate model development techniques to the case of leakage power and propagation delay in the digital standard cell library is currently in progress.

2 Conclusions

The project is in good shape to achieve the planned final objective and has also achieved further results:

- The research activities have led to the implementation of the SW prototype that will offer the end user, the designer, the combined and increased advantage to have the possibility to choose the method to generate surrogate models, according the circuit or the case study need, altogether reusing the training set computation time for both methodologies (main gain of time).

- Additional research work has *demonstrated the Neural Network methodology can be used to create a circuit model within the time domain still reducing the simulation time. The results are very promising and will set the basis of future research and development activities.*
- We have extended the MAnON eDFY methodologies to the digital domain and built a CMOS library as additional test cases with strong non-linearity for optimization and modelling in the context of Design for Yield. The circuits have been characterized in 45 nm technology and experiments on the optimization and modeling of digital standard cells (flip-flop) taken from the library have been conducted leading to partially successful results.

Waveform Modelling in Order to Speed Up Transient SPICE Simulations

Mohammed Ali Khozoei, Matthias Hauser, and Angelo Ciccazzo

Abstract The production of semiconductor integrated circuits is very complex and expensive. Therefore, it is essential to verify the designed circuits before they are fabricated. Due to the process variations, nanoscale circuits have to be simulated many times during the design flow. This kind of analysis can be very expensive because of their complexity and the high number of simulations. For this reason the semiconductor industry is deeply interested in using less complex but accurate models to speed up time consuming SPICE simulations.

This contribution presents a method that creates a compact model, which replaces a semiconductor integrated circuit or sub circuit to significantly reduce the transient simulation time.

Keywords Nanoscale circuits • Semiconductor integrated circuits simulations

1 Introduction

The simulation of semiconductor integrated circuits can be very challenging [5, 7]. Depending on development status of the circuit in the design flow, there exist different targets. Without loss of generality, we consider here the following two cases:

1. In the *design centering* [2] step, analogue circuits are simulated many times during their design flow, e.g. to find the design parameter values that optimize the production yield. In such situations the values of several circuit parameters are adapted and the resulting waveforms of some circuit outputs are verified. Here a speed up of the simulation can be realized by replacing the analogue circuit by approximations of these waveforms.

M.A. Khozoei • M. Hauser
Fraunhofer ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
e-mail: mohammed.ali.khozoei@itwm.fraunhofer.de; matthias.hauser@itwm.fraunhofer.de

A. Ciccazzo (✉)
STMicroelectronics Srl, Stradale Primosole 50, 95121 Catania, Italy
e-mail: angelo.ciccazzo@st.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_59

2. In case that the circuit contains a set of sub circuits, the complexity of each sub circuit has a lasting effect on the simulation time. Replacing some of the sub circuits with simplified models that approximate the sub circuit's output signals also leads here to a significant speed up.

In this paper we will present a method to create simplified models of non-linear dynamic systems with the use of artificial neural networks and Bézier curves. These models are able to approximate the time domain behaviour of the non-linear dynamic systems and reduce the effort for their simulations.

First of all a short introduction in Bézier curves is given. Then the field of application of the method is defined. And finally the method itself and its application to a modern analogue circuit is presented.

2 Quadratic Bézier Curves

A quadratic Bézier curve [6] is a parametric function defined by three **control points $\mathbf{P_0}$, $\mathbf{P_1}$** and $\mathbf{P_2}$:

$$B(t) = (1-t)^2\,\mathbf{P_0} + 2t(1-t)\,\mathbf{P_1} + t^2\,\mathbf{P_2},\ t \in [0,\ 1] \tag{1}$$

In the case $B(t)$ is given, the points $\mathbf{P_0} = B(0)$ and $\mathbf{P_2} = B(1)$ are the endpoints of the curve while $\mathbf{P_1}$ is unknown. Considering its derivative

$$B'(t) = 2(1-t)(\mathbf{P_1} - \mathbf{P_0}) + 2t(\mathbf{P_2} - \mathbf{P_1}),\ t \in [0,\ 1] \tag{2}$$

we can conclude that $\mathbf{P_1}$ is the intersection of the tangents at the endpoints $\mathbf{P_0}$ and $\mathbf{P_2}$. Hence, any quadratic Bézier curve is uniquely defined if all the three control points or if the endpoints and their corresponding tangents are given (Fig. 1).

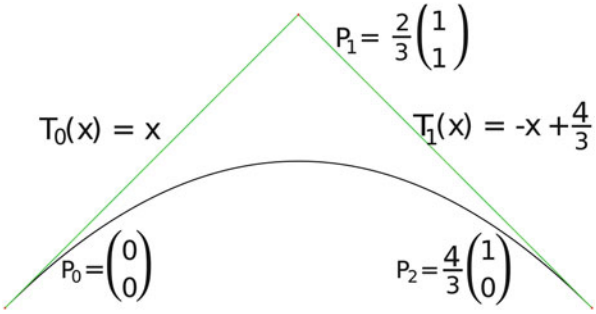


Fig. 1 The Bézier curve that represents the graph of the function $f(x) = x - \frac{3}{4}x^2,\ x \in [0,\ \frac{4}{3}]$ is defined by the control points $\mathbf{P_0}$, $\mathbf{P_1}$ and $\mathbf{P_2}$. $\mathbf{P_1}$ is the intersection of the tangents $T_0(x)$ and $T_1(x)$

3 Modelling

Given an analogue circuit, its transient output signals depend on several circuit parameters, like supply voltages or resistor values. We call these parameters **input variables**. The goal is now to find a simplified model for the functions that map the input variables to the corresponding transient output signals. Therefore we first define these functions and their approximations.

Definition 1 Let $\tilde{n}$ be the number of input variables and $X \subset \mathbb{R}^{\tilde{n}}$ be the set of their values. Then $f(t,\ X^k)$ is the continuous function of the **transient output signal** of the circuit for a given set of the input variables $X^k \in X$ in the time range $[T_0,\ T_1]$.

Definition 2 For a function $f(t,\ X^k)$ and an approximation error $\varepsilon_0 > 0$, S_k^f is the set of $\hat{n} \in \mathbb{N}$ quadratic Bézier curves $B_i(\tau)$ given by the control points $\{\mathbf{P_0^i},\ \mathbf{P_1^i},\ \mathbf{P_2^i}\}$:

$$S_k^f := \{B_i(\tau) \mid \tau \in [0,\ 1],\ i \in \{1,\ \ldots,\ \hat{n}\},\ B_i(0) = f(t_{i-1},\ X^k),$$
$$B_i(1) = f(t_i,\ X^k),\ t_i \in [T_0,\ T_1],\ t_0 = T_0,\ t_{\hat{n}} = T_1\}$$
$$\text{for all } X^k \in X.$$

such that for the parametrization functions $\varphi_i : [t_{i-1},\ t_i] \rightarrow [0,\ 1]$ hold that

$$|f(t,\ X^k) - G(B_i(\varphi_i(t)))| < \varepsilon_0,\ B_i(\tau) \in S_k^f,\ t \in [t_{i-1},\ t_i]$$
$$G : \mathbb{R}^2 \rightarrow \mathbb{R}$$
$$G([x,\ y]) = y$$

Furthermore C_k^f is the set of control points of S_k^f:

$$C_k^f = \{\{\mathbf{P_0^i},\ \mathbf{P_1^i},\ \mathbf{P_2^i}\} \mid \text{control points of } B_i(\tau) \in S_k^f\}$$

Note that S_k^f approximates a function $f(t,\ X^k)$ piecewise by Bézier curves. The output signal of an analogue circuit can be approximated if it can be described by a function $f(t,\ X^k)$ so that the set S_k^f exists.

Definition 3 Let $F_{\tilde{X}} = \{f(t, X^k) | X^k \in \tilde{X} \subset X\}$ be given. We call $F_{\tilde{X}}$ a **qualified** set if there exists an $0 < \varepsilon \in \mathbb{R}$ and a function $g : C \times X \rightarrow C$ such that for any functions $f(t, X^k)$, $f(t, X^l) \in F_{\tilde{X}}$ the following conditions are valid:

1. $|S_k^f| = |S_l^f|$
2. $\| g(C_k^f, X^l) - C_l^f \| \leq \varepsilon$

with $C = \bigcup C_k^f$ for all $X^k \in X$. Furthermore, we call the functions $f(t, X^k)$ and $f(t, X^l),\ t \in [T_0, T_1]$ **similar**.

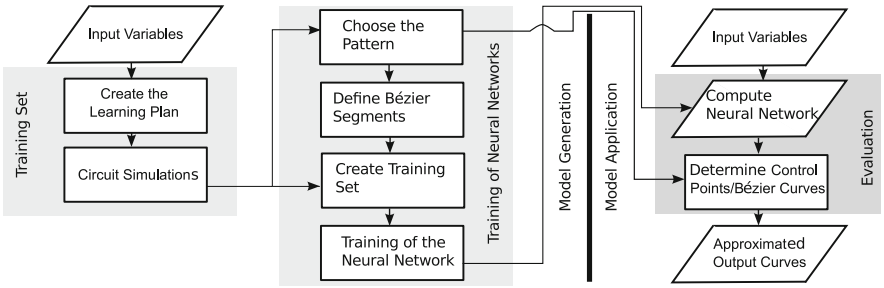


Fig. 2 Diagram of the model generation and its application for transient simulation

3.1 Modelling Assumptions

For modelling the transient behaviour of an output signal of an analogue circuit, we assume that the input variables are known and the corresponding output function $f(t,\ X^k)$ can be approximated by a set S_k^f, as defined in Definition 2.

Furthermore we assume that the set $\{f(t,\ X^k) \mid \forall\ X^k\ \in\ X\}$ can be split into n qualified sets so that for each subset a separate model can be created and combined afterwards.[1] Without loss of generality, we assume in the next sections that the waveforms of a circuit's output are similar for all $X^k \in X$.

3.2 Generation of the Neural Network

Figure 2 shows an overview of the algorithm of the model generation presented in this section and its application.

The goal of the developed algorithm is the generation of a set of Bézier curves S_k^f for any given input set $X^k \in X$ that approximates $f(t,\ X^k)$ sufficiently accurate. The idea is to use a neural network [3, 8, 9] that generates the set of control points C_k^f, which defines S_k^f.

To train the neural network, a training set is needed. It should contain a learning plan and the corresponding neural network's output values, which will be evaluated by the neural network for any valid inputs.

A **learning plan** L contains informative samples $L = \{X^k |\ k \in \{1, \ldots, n_T\}\} \subset X$ from the input space X. There are different methods like *latin hypercube* [4] to generate a meaningful sampling of the input space X. The choice of such a sampling method strongly depends on the demands of the machine learning method [10]. To evaluate the learning plan, the analogue circuit has to be simulated n_T times to determine the set of the **transient curves** $F_L = \{f(t,\ X^k) \mid X^k \in L,\ t \in [T_0, T_1]\}$.

[1] This is used in the example presented in Sect. 4.

3.2.1 Definition of the Training Set

We already assumed that the waveforms of the circuit's output are similar for all inputs $X^k \in X$. Now, we need to define the neural network so that the segments of S_k^f for all $X^k \in X$ can be generated from its outputs. First of all, we need to know how many Bèzier segments are contained in S_k^f and where the endpoints of these segments are located. Therefore, one of the output functions $f(t,\ X^p) \in F_L$, that has the typical shape for the functions in F_L, is chosen as a **pattern**. This function $f(t,\ X^p)$ is then approximated by a set of Bézier curves with their control points C_p^f for a predefined error ε_0^p. The fitting itself is done by a trial and error method. The **characteristic** of the set S_p^f defines the number of Bézier segments $\hat{n}$, and other degrees of freedom, e.g. the locations and tangents of the control points.

If the neural network is trained so that its outputs estimate the coordinates of the control points, the neural network has up to $4\hat{n} + 2$ outputs. Since the control points $\mathbf{P_1^i}$ are not on the Bézier curves, their estimation and verification is difficult.

Instead of this we use the tangents of the endpoints to evaluate the control points $\mathbf{P_1^i}$, as described in Sect. 2. Therefore, we determine the endpoints of the segments such that the gradients of their tangents have the minimal variation over all functions in F_L. In this case the neural network has up to $3\hat{n} + 3$ outputs. Especially, note that the gradients of the tangents will not change if the corresponding Bézier segment is only shifted for all $X^k \in L$. Finally the training set will be created by saving the samples L, the endpoints of Bézier segments and the gradient of their tangents for all $S_k^f, X^k \in L$. In some cases the effort of training the network can be reduced if the training set contains an output parameter that is constant over all $X^k \in L$. In these cases, the corresponding value can be imported from the pattern during the waveform approximation for any input $X^l \in X$. With this, the neural network's training set is defined.

3.2.2 Training of the Neural Network and Further Improvements

The neural network is trained by the described training set. Please note that the pattern has a significant influence on the accuracy of the model. If the created model is not sufficiently accurate, choosing a new pattern may increase the model performance considerably. Fortunately this can be done without any new simulations of the original circuit and depending on the output function's complexity the training of the new model is run in a short time.

Furthermore depending on the considered waveforms, it is sufficient to use linear Bèzier curves to model certain segments. This reduces the degrees of freedom and thus the number of the outputs of the neural network.

Please note that the model for the endpoint $\mathbf{P_2^i}$ of the segment i can be enhanced, if we use $\mathbf{P_0^i}$ as additional input of the neural network. Adding more already known points as inputs is also possible, but then more training samples and consequently more simulations of the circuit are needed to train this extended neural network.

In general, other characteristics of the original output curves can be used for increasing the quality of the approximating Bèzier curves. This means, if there is a function h mapping the characteristics D_k^f to the control points C_k^f, it can be used to recover S_k^f by its control points $C_k^f = h(D_k^f)$, where D_k^f is the result of the neural network for a given input set X^k.

4 Application of the Method on an Industrial Analogue Circuit

In this section, we present the application of the described method in the industrial environment. Therefore, *STMicroelectronics Catania* provides a voltage reference device, which has four different output voltage levels (0, 4, 4.5 and 5 V). The output voltage can be switched between the levels after a settling time. Except for 0 V, switching the output voltage to other levels causes an overshoot or undershoot, which has to be approximated by the model. Therefore the output levels 4 and 4.5 V are approximated by two models each and the output level 5 V can be approximated by one model. To generate these five models, the program *Design* [8] provided by *Fraunhofer ITWM* [1] is used to train the neural networks. The training plan contains 200 samples of the parameters' space given in the Table 1. For the simulation within the *Spectre Circuit Simulator*, the device is replaced by a *Verilog-A* code, which computes the respective Bèzier segments. Table 2 shows the number of the segments and neural network outputs for each model.

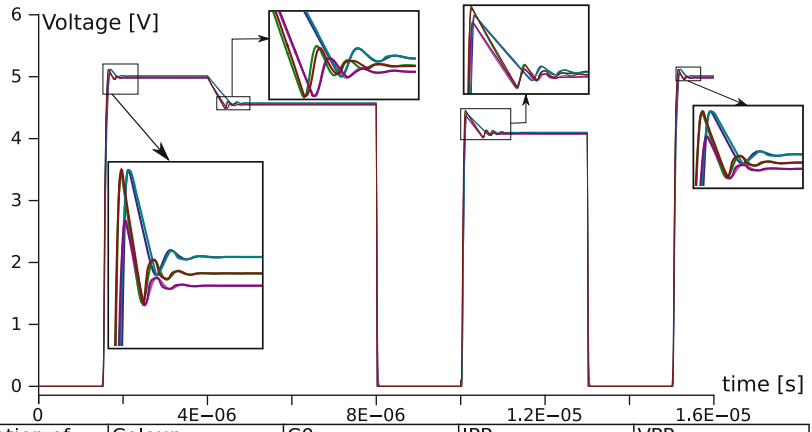
Table 1 Parameters ranges for the training of the neural network

	Load capacity C_0	Biasing current I_{pp}	Biasing voltage V_{pp}
Lower boundary	10 pF	1.75 μA	6.8 V
Upper boundary	40 pF	2.25 μA	7.2 V

Table 2 Models' information

Level	Segments	Endpoints (neural network)	Slopes (neural network)
5 V overshoot	15	16 (32[a], 26[b])	16 (16[a], 4[b])
4.5 V undershoot	16	17 (34[a], 27[b])	17 (17[a], 7[b])
4 V overshoot	19	20 (40[a], 34[b])	20 (20[a], 5[b])

[a]The maximum number of the required outputs of the neural network
[b]The actual number of neural network's outputs. The pattern will be used to define the values that are not given by the neural network



Simulation of	Colour	C0	IPP	VPP
analogue circuit		1.8E−11	1.80502E−06	7.19
analogue circuit		1.25E−11	2.245E−06	6.85
analogue circuit		1.005E−11	2E−06	6.99242
Model		1.8E−11	1.80502E−06	7.19
Model		1.25E−11	2.245E−06	6.85
Model		1.005E−11	2E−06	6.99242

Fig. 3 Using Bézier segments to approximate the output waveforms of a voltage reference device

To compute the waveform in a given time range, De Casteljau's algorithm[2] is used. Figure 3 compares the generated output waveforms of the models with the transient simulation results of the analogue circuit.

Acknowledgements The present research work was supported by *Methods for Advanced Multi-Objective Optimization for eDFY of complex Nanoscale Circuits—ManOn* Marie Curie Fellowships for the Transfer of Knowledge (ToK). ST Microelectronics Catania—Italy, and Fraunhofer Institute for Industrial Mathematics (ITWM) in Kaiserslautern—Germany.

References

1. Fraunhofer ITWM: Fraunhofer Institute for Industrial Mathematics. Website. www.itwm.fraunhofer.de (2014)
2. Graeb, H.E.: Analog Design Centering and Sizing. Springer, Dordrecht (2007)
3. Haykin, S.: Neural Networks: a Comprehensive Foundation, 2nd edn. Prentice Hall, Upper Saddle River (1998)
4. Loh, W.L.: On latin hypercube sampling. Ann. Stat. **24**(5), 2058–2080 (1996)

[2]The computation of a transient output curve with 80 points in time needs about 4 ms. Fortunately this time just depends on the number of sample points and not on the complexity of the underlying circuit.

5. Lorenz, J.K., Bar, E., Clees, T., Evanschitzky, P., Jancke, R., Kampen, C., Paschen, U., Salzig, C.P., Selberherr, S.: Hierarchical simulation of process variations and their impact on circuits and systems: results. IEEE Trans. Electron Devices **58**(8), 2227–2234 (2011)
6. Prautzsch, H., Boehm, W., Paluszny, M.: Bézier and B-spline Techniques. Springer, Berlin [u.a.] (2002)
7. Rappitsch, G., Seebacher, E., Kocher, M., Stadlober, E.: Spice modeling of process variation using location depth corner models. IEEE Trans. Semicond. Manuf. **17**(2), 201–213 (2004)
8. Sarishvili, A.: Design - the customized predictive analytics tool. Website. www.itwm.fraunhofer.de/design-predictive-analytics (2014)
9. Sarishvili, A., Andersson, C., Franke, J., Kroisandt, G.: On the consistency of the blocked neural network estimator in time series analysis. Neural Comput. **18**(10), 2568–2581 (2006)
10. Vapnik, V.N.: Statistical Learning Theory. Wiley, Chichester (1998)

Yield Optimization in Electronic Circuits Design

Angelo Ciccazzo, Gianni Di Pillo, and Vittorio Latorre

Abstract In this work we propose an approach that combines a Support Vector learning Machine with a Derivative-Free black box optimization algorithm in order to maximize the yield in the production of electronic circuits. This approach is tested on a circuit provided by ST-Microelectronics, to be employed in consumer electronics. The results of the approach are compared with the results of WiCkeD, a commercial software largely used for integrated circuits analysis.

Keywords Derivative-free optimization • Electronic circuits • Support vector machines

1 Introduction

In recent years there is a great increase in the complexity of digital integrated circuits, with the number of components in a single circuit doubling every year. Therefore the issues coming from scaling the devices dimensions to nanometer size become more and more difficult to handle. Some major challenges in the manufacturing process come from parametric variations such as intra and inter die variation of channel length, oxide thickness and doping concentration. These variations have a deep impact on the circuit performances which must satisfy given specifications.

A circuit model must be tested in several ways during the design phase so that there is an high probability that the final manufactured circuits satisfy the specifications. These tests are generally performed with time-consuming circuit simulations. Therefore there is a great interest in developing methods capable of performing reliable analysis with the use of less simulations as possible.

A. Ciccazzo
ST-Microeletronics, Stradale Primosole 50, 95121 Catania, Italy
e-mail: angelo.ciccazzo@st.com

G. Di Pillo • V. Latorre (✉)
Department of Computer, Controls and Management Engineering A. Ruberti, "Sapienza" University of Rome, Via Ariosto 25, 00185 Rome, Italy
e-mail: dipillo@dis.uniroma1.it; latorre@dis.uniroma1.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_60

In this paper we use a particular kind of learning machine, the Support Vector Machines (SVM) as surrogate model for the circuit simulations [3] and an efficient derivative free (DF) mix-integer optimization algorithm [6] in order to perform this analysis. Surrogate models have been used to approximate circuit performance evaluations [1, 4, 5], and in this work the SVM is used to perform a Montecarlo (MC) analysis that can be used, for a specific design choice, to calculate the yield, that is the probability that a circuit satisfies its specifications. Then the DF algorithm is used in order to maximize the yield. A similar approach has been used in [2] where instead of a MC analysis a robust design using a different derivative free algorithm has been implemented. Our results are compared with the results obtained by using the commercial software WiCkeD,[1] widely popular in the electronic industrial sector. The comparison is performed using as a test case an actual DC-DC converter circuit designed and produced by ST-Microelectronics.

2 Problem Definition

Circuit variables can be divided into three different classes:

- Design Variables x_d: these variables represent the geometrical dimensions of the components in the circuits (e.g. channel widths and lengths);
- Operating Variables x_o: these variables model operating conditions (e.g. supply voltage and temperature);
- Statistical Variables x_p: these variables are usually subject to uncertainty due to fluctuations in the manufacturing process and are generally modeled by Gaussian or Uniform Distributions (e.g. oxide thickness, threshold voltage and channel length reduction).

The variables x_d, x_o, x_p are vector variables of suitable dimensions.

In the manufacturing processes the electronic devices are produced in series, but each device is not the same in terms of electrical performances. This brings to the uncertainty in the circuit manufacturing process modeled by the Statistical Variables x_p. Another source of uncertainty are the conditions in which the circuit should operate after the production. These conditions as the working temperature and the supply voltage, are considered in the Operating Variables x_o. During the yield optimization process the Operating Variables x_o are set to their worst cases, that is the most unfavorable values of these variables for the circuit performances.

The Statistical and Operating Variables are not under the control of the circuit designer. The only variables that can be manipulated in order to improve the circuit performances are the Design Variables x_d. Therefore these variables are manipulated during the optimization process. The Design Variables have some lower and upper

[1] MunEDA inc, WiCkeD, a Tool Suite for Nominal and Statistical Custom IC Design.

bounds that they cannot exceed, and that are modeled with box constraints in the optimization process:

$$l_{x_d}^k \leq x_d^k \leq u_{x_d}^k \quad \text{for } k = 1, \ldots, n,$$

where n is the number of design variables.

The outputs of the circuit design process are the Performance Features $p_i(x_d, x_o, x_p)$ for $i = 1, \ldots, m$, where m is the number of performances. The Performance Features are quantities that represent the behavior of the circuits such as: delay, gain, phase, margin, slew rate, etc. The Performance Features must satisfy given constraints for a circuit to be in full working order. Usually these constraints are upper and lower bounds on the values of the performances:

$$l_i \leq p_i(x_d, x_o, x_p) \leq u_i \quad \text{for } i = 1, \ldots, m. \tag{1}$$

The problem we aim to solve is the Performance Centering Problem for Yield Optimization. For the Performance Features, we aim to manipulate the Design Variables so that there is an high probability that the produced circuits are in full working order with the Operating Variables at their worst cases and despite the variations of the Statistical Variables.

3 Proposed Methodology

We assume that before the optimization process the values of the Operating Variables x_o are given at the worst cases. From Eq. (1) it is possible to notice that there are $2m$ constraints on the Performances. The worst cases conditions for the Operating Variables are associated with the constraints on the Performances, that is a worst case for every constraint on the performances. Therefore there are $2m$ worst case conditions in the yield optimization problem. It is assumed that if the Performance Specifications are satisfied at the worst cases, they are satisfied for any other values of the Operating Variables.

As we said in the introduction, the goal of the Yield Optimization is to maximize the yield (i.e. the percentage of manufactured circuits that satisfy the Performance Specifications when varying Statistical Variables).

Given the Design Choice $\tilde{x}_d$ and the values of the Operating Variables at the worst cases $\tilde{x}_o$, let:

$$A_p = \{x_p \mid l \leq p(\tilde{x}_d, \tilde{x}_o, x_p) \leq u\};$$

then the yield Y can be formally defined as:

$$Y = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \delta(x_p) \cdot pdf(x_p) \cdot dx_p = E\{\delta(x_p)\}$$

where

$$\delta(x_p) = \begin{cases} 1, & x_p \in A_p \\ 0, & otherwise, \end{cases}$$

and $pdf(x_p)$ denotes the probability density function of the Statistical Variables.

An estimator for the expectation value is:

$$\hat{Y} = \hat{E}\{\delta(x_p)\} = \frac{1}{n_s} \sum_{\mu=1}^{n_s} \delta(x_p^{(\mu)}) = \frac{n_{ok}}{n_s},$$

with $x_p^{(\mu)}$, $\mu = 1, \ldots, n_s$ normally distributed sample elements. Therefore, the estimator is given by the number of the sample elements which satisfy the specifications n_{ok} divided by the total number of elements in the sample n_s.

One approach to maximize the yield consists in randomly generating a certain number of circuits simulations through a MC analysis varying the values of the statistical variables and minimizing the violation of the specifications in all the considered operating cases. In particular $l = 1 + 2m$ cases are considered, that is the worst cases conditions plus the typical condition (temperature 27 °C, supply voltage 2.3 V). For each of these l working conditions a Montecarlo analysis is generated. Therefore the problem that is solved is:

$$\begin{aligned} &\min_{x_d} \sum_{\mu=1}^{n_s} \sum_{j=1}^{l} \sum_{i=1}^{m} \left\{ \max\{0, l_i - p_i(x_d, \bar{x}_{o,j}, \bar{x}_p^{(\mu)})\} + \max\{0, p_i(x_d, \bar{x}_{o,j}, \bar{x}_p^{(\mu)}) - u_i\} \right\} \\ &s.t. \;\; l_{x_d}^k \le x_d^k \le u_{x_d}^k \quad \text{for } k = 1, \ldots, n. \end{aligned} \tag{2}$$

The main drawback of this approach is that a large number of time consuming simulations must be performed in order to realize a reliable MC Analysis on the Statistical Variables.

Our approach consists in using the SVM as surrogate model for the simulations in the MC Analysis. In detail, every time the objective function has to be evaluated, the simulator generates a set of input-output samples used as training set for the SVM. Once the SVM has been generated, it is used as surrogate model to generate a Montecarlo with a large number of points that approximates the yield.

The generation of the surrogate model via SVM is explained in detail in [3]. Furthermore it can be easily noticed that the objective function in (2) is non-smooth, therefore it is approximated by a zero-norm type of objective function and the

following DF problem is solved:

$$\begin{aligned} \min_{x_d} & \sum_{\mu=1}^{n_s} \sum_{j=1}^{l} \sum_{i=1}^{m} \Big\{ \log\left(\max\{0, l_i - p_i(x_d, \bar{x}_{o,j}, \bar{x}_p^{(\mu)})\} + \epsilon\right) + \\ & \log\left(\max\{0, p_i(x_d, \bar{x}_{o,j}, \bar{x}_p^{(\mu)}) - u_i\} + \epsilon\right) \Big\} \\ s.t. \; & l_{x_d}^k \le x_d^k \le u_{x_d}^k \quad \text{for } k = 1, \ldots, n, \end{aligned} \tag{3}$$

where ϵ is a positive parameter close to zero.

4 Results

We present the results obtained using as a test case the DC-DC converter for AMOLED display panels shown in Fig. 1. In this circuit it is relevant the delay between the time the signal to turn on/off the circuit is sent and the time the circuit is actually turned on/off. The longer the difference between these delays is, the more the power losses increase. Hence, this circuit has three performances, Delay 1 (on), Delay 2 (off) and Delay S (symmetry). The performance we are most interested in is the Delay S that represents a measure of the overall efficiency of the circuit. The problem has eight Design Variables, nine Statistical Variables and two Operating Variables, the supply voltage and the temperature. The worst cases for Delay 1 and Delay 2 coincide both at the upper bound and the lower bound. This means that there are a total of $l = 5$ values for the Operating Variables, one at typical condition and four for the worst case conditions.

The tests have been performed at ST-Microelectronics headquarter in Catania on a computational grid with more than 800 processors. This brings a further challenge in measuring the goodness of the results. As a matter of facts the processors used for the simulation are chosen according to the load on the grid and the different processors cannot be expected to have the same performances. Consequently using the time needed for a run to be completed to measure the speed of the computation is not a viable option. Therefore we will use the total number of simulations needed in a run to measure the speed of the computation, as the simulations are considered the principal computational load of the procedure.

Once the optimization procedure is completed, we test the obtained optimal values for the Design Variables with four 250 points Montecarlo generated through simulations in every worst conditions, for a total of 1000 points. In Table 1 we report the yield for the final Montecarlo analysis obtained in the worst cases for the three performances, using WiCkeD and the proposed method, together with the values of the operational parameters in the worst cases. The total yield is calculated as the percentage of circuits that satisfy all the three performances at the same time. From these results we observe that, as concerns the yield, the two methods are able to get comparable results. WiCkeD is substantially superior for what concerns the Delay 1

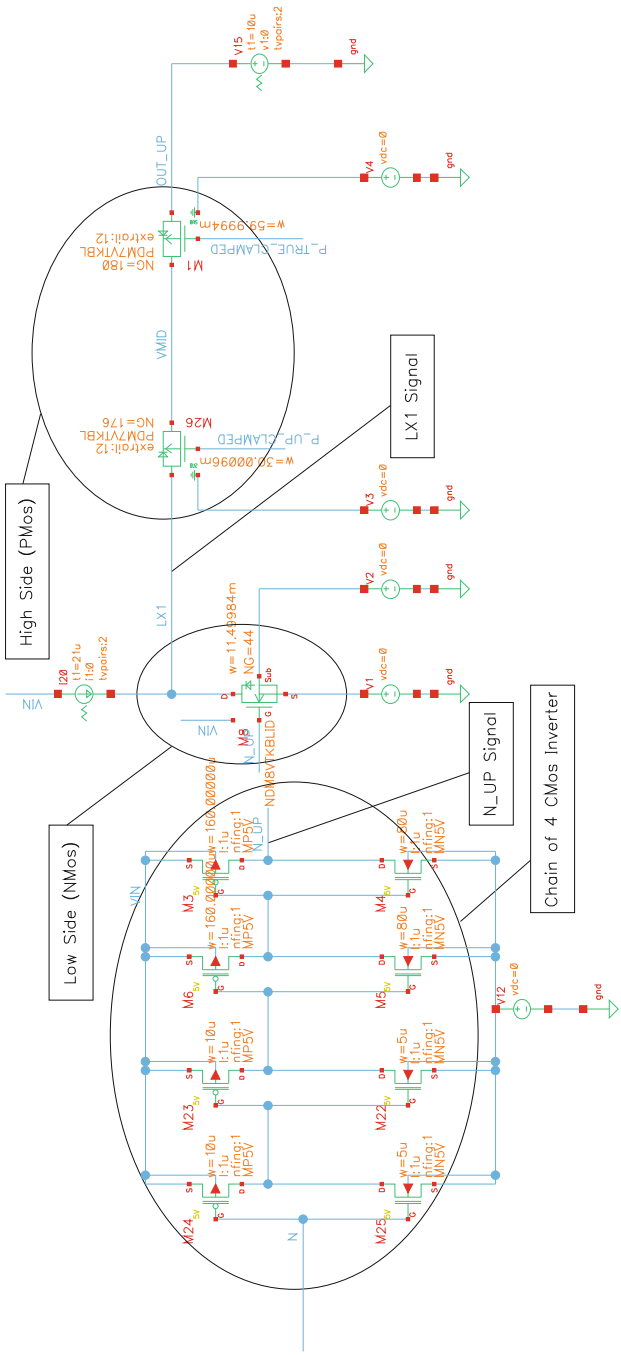


Fig. 1 The DC-DC converter circuit

Table 1 Comparisons of the results between the proposed method and Wicked

Specification	Temp.	Voltage	Yield wicked (%)	Yield SVM-DF (%)
Delay 1 lower	120	2.3	100	100
Delay 1 upper	120	4.8	94.30	89.50
Delay 2 lower	120	2.3	99.20	100
Delay 2 upper	120	4.8	96.20	94.90
Delay S lower	−40	2.3	100	100.00
Delay S upper	−40	4.8	96.60	100.00
Total yield			91.40	89.30
Total simulations			11,000	8600

Upper and a has slightly better performance for the Delay 2 Upper, while SVM-DF fares better for the Delay 2 Lower and the Delay Symmetry Upper. The difference in performance between the two methods in the Delay 1 Upper case influences the total yield making WiCkeD superior. On the other side, the performance we are most interested in is the Delay Symmetry. Therefore the solution obtained by the SVM-DF is better from a design point of view because it reaches the 100 % of Delay Symmetry.

From the point of view of computational burden, the proposed method generates 100 circuit simulations every iteration in order to train the SVMs and the stopping condition for the optimization algorithm occurs after 86 function evaluations, for a total of 8600 simulations. On the other hand WiCkeD takes 11,000 simulation to find the results in Table 1. All simulations are done using the ELDO simulation software [7].

In conclusion these preliminary results indicate that the proposed method is able to find a solution for the yield optimization problem comparable or even slightly superior with the solution obtained by the benchmark WicKeD, with only the 75 % of computational cost.

References

1. Boolchandani, D., Garg, L., Khandelwal, S., Sahula, V.: Variability aware yield optimal sizing of analog circuits using SVM-genetic approach. In: 2010 XIth International Workshop on Symbolic and Numerical Methods, Modeling and Applications to Circuit Design (SM2ACD), pp. 1–6 (2010)
2. Ciccazzo, A., Latorre, V., Liuzzi, G., Lucidi, S., Rinaldi, F.: Derivative-free robust optimization for circuit design. J. Optim. Theory Appl. (2013) doi:10.1007/s10957-013-0441-2
3. Ciccazzo, A., Di Pillo, G., Latorre, V.: Support vector machines for surrogate modeling of electronic circuits. Neural Comput. Appl. **24**, 69–76 (2014)
4. Ilumoka, A.A.: A modular neural network approach to microelectronic circuit yield optimization. Microelectron. Reliab. **38**, 571–580 (1998)
5. Jing, M., Hao, Y., Zhang, J.F., Ma, P.J.: Efficient parametric yield optimization of VLSI circuit by uniform design sampling method. Microelectron. Reliab. **45**, 155–162 (2005)

6. Liuzzi, G., Lucidi, S., Rinaldi, F.: Derivative-free methods for bound constrained mixed-integer optimization. Comput. Optim. Appl. **53**, 505–526 (2012)
7. Mentor Graphics Corporation (2014). Website: http://www.mentor.com/products/ic_nanometer_design/analog-mixed-signal-verification/eldo/

MS 17
MINISYMPOSIUM: MODELING AND OPTIMIZATION OF INTERACTING PARTICLE SYSTEMS

Organizers

Stephan Martin[1] and René Pinnau[2]

Speakers

Maria Bruna[3] and Jon Chapman[4]
From Discrete to Continuum in Stochastic Models of Diffusion of Finite-Size Particles

Concetta Drago[5] and Vittorio Romano[6]
Optimal Control for Semiconductor Diodes Design Based on the MEP Energy-Transport Model

Jochen Kall[7]
A Different Perspective on Riemann Problems

[1]Imperial College, London, England.
[2]TU Kaiserslautern, Kaiserslautern, Germany.
[3]OCIAM, Oxford, England.
[4]OCIAM, Oxford, England.
[5]Univ. Catania, Catania, Italy.
[6]Univ. Catania, Catania, Italy.
[7] TU Kaiserslautern, Kaiserslautern, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_61

Stephan Martin[8]
Interacting Particle Systems for Swarming and Their Kinetic PDEs

Claudia Totzeck[9]
Optimal Control of Interacting Particle Systems with External Agents

Marie-Therese Wolfram[10]
Mean Field Game Approaches in Pedestrian Dynamics

Keywords

Biology
Interacting particle systems
Modeling
Optimization
Semiconductors

Short Description

The aim of this mini symposium is to present recent advances in the modeling, simulation and optimization of interacting classical and quantum particle systems with applications in biology and semiconductor device modeling.

[8]Imperial College, London, England.
[9]TU Kaiserslautern, Kaiserslautern, Germany.
[10]RICAM, Linz, Austria.

Numerical Sensitivity Analysis for an Optimal Control Approach in Semiconductor Design Based on the MEP Energy Transport Model

Concetta R. Drago and Vittorio Romano

Abstract An optimal control approach based on the adjoint method for the design of a semiconductor device is considered. A consistent energy transport model, free of any fitting parameters, formulated on the basis of the maximum entropy principle (MEP) is used as mathematical model. The robustness of the optimal control approach is verified by a numerical sensitivity analysis, performed by introducing a Gaussian noise in the reference doping profile.

Keywords Maximum entropy principle • Optimal control • Semiconductor design • Sensitivity analysis

1 Introduction

One of the main objectives in optimal semiconductor design is to get an improvement in the current flow at a specified Ohmic contact of the device, for a fixed applied voltage, via a modification in the doping density profile. Recently the problem of optimal control in semiconductor design was tackled by using the adjoint calculus, see [1–4], where an optimal control approach have been presented for the drift diffusion model. In [5–8] the same optimal control approach was extended to the classical Stratton energy transport model. Energy-transport models [9] take into account also the thermal effects of the electron flow inside the semiconductor devices. Usually they are based on phenomenological constitutive equations for the particle flux and the energy-flux, depending on a set of parameters which are fitted to homogeneous bulk material Monte Carlo simulations. In [10–12] a model free of any fitting parameters has been developed for the electron transport in silicon, where the parameters appearing in the constitutive laws are directly related to the collision operators of the semiclassical Boltzmann transport equation for electrons in semiconductors. Such an approach, based on the maximum entropy principle

C.R. Drago • V. Romano (✉)
Dipartimento di Matematica e Informatica, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy
e-mail: concetta.drago@gmail.com; romano@dmi.unict.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_62

(hereafter MEP), takes into account all the relevant scattering mechanisms in silicon, i.e. scattering of electrons with acoustic and non-polar phonons and with impurities.

The extension to MEP energy transport model of the optimal approach was investigated in [13] (for an approach based on evolutionary algorithms see [14]). In this paper a sensitivity numerical analysis is performed by introducing a Gaussian noise in the reference doping profile. The results show the robustness of the method.

2 MEP Energy Transport Model

The energy transport model, obtained for semiconductor in [12] starting from the hydrodynamical model based on the maximum entropy principle [10, 11], is given in the stationary case by the following set of balance equations for the electron density n and energy W, coupled with the Poisson equation for the electric potential ϕ:

$$\operatorname{div}(n\mathbf{V}) = \mathbf{0}, \qquad \operatorname{div}(n\mathbf{S}) = nq\mathbf{V}\cdot\nabla\phi + nC_W \tag{1}$$

$$\operatorname{div}(\varepsilon\nabla\phi) = q(n - C). \tag{2}$$

C is the doping profile, which depends only on the position x, q is the (positive) elementary charge, ϵ is the dielectric constant. The system is closed with the following constitutive relations for the velocity $\mathbf{V}$ and the energy-flux $\mathbf{S}$ [15, 16]

$$\mathbf{J} = n\mathbf{V} = \frac{c_{22}}{D}\nabla(nU) - \frac{c_{12}}{D}\nabla(nF) - q\lambda^W n\nabla\phi\left[\frac{c_{22}}{D}U - \frac{c_{12}}{D}F\right], \tag{3}$$

$$\mathbf{H} = n\mathbf{S} = \frac{c_{11}}{D}\nabla(nF) - \frac{c_{21}}{D}\nabla(nU) - q\lambda^W n\nabla\phi\left[\frac{c_{11}}{D}F - \frac{c_{21}}{D}U\right] \tag{4}$$

where $\mathbf{J}$ is the electron current density and $\mathbf{H}$ the energy-flux density. Here $D = c_{11}c_{22} - c_{12}c_{21}$. All the coefficients c_{ij} and the functions U, F depend on the energy W. We define $\alpha(W) = \frac{c_{22}}{D}U - \frac{c_{12}}{D}F$ and $\beta(W) = \frac{c_{11}}{D}F - \frac{c_{21}}{D}U$. The energy production term has a relaxation form $C_W = -\frac{W-W_0}{\tau_W}$, where τ_W is the energy relaxation time, which depends also on W, and $W_0 = \frac{3}{2}k_BT_L$ is the energy at equilibrium. T_L is the lattice temperature, here assumed to be constant. The expression of U, F, C_W, c_{ij} have been obtained in [10, 11] both for the parabolic band and Kane's dispersion relation. Finally λ^W is the Lagrangian multiplier associated to the energy W. For more details regarding the model we refer to [10–12, 17].

System (1) and (2) has to be supplemented with suitable boundary conditions. We assume that the boundary $\partial\Omega$ of the domain Ω splits into two disjoint parts Γ_D and Γ_N, where Γ_D represents the Ohmic contacts of the device and Γ_N represents the insulating parts of the boundary. Let $\boldsymbol{\nu}$ denote the unit outward normal vector

along the boundary. We will consider the following mixed boundary conditions:

$$n = n_D, \quad \nabla W \cdot \boldsymbol{\nu} = 0, \quad \phi = \phi_D \qquad \text{on } \Gamma_D, \tag{5a}$$

$$\nabla n \cdot \boldsymbol{\nu} = \nabla W \cdot \boldsymbol{\nu} = \nabla \phi \cdot \boldsymbol{\nu} = 0 \qquad \text{on } \Gamma_N. \tag{5b}$$

Here n_D and ϕ_D are the $H^1(\Omega)$-extensions of fixed functions defined on Γ_D.

3 Design Problem and Analytical Setting

The design objective we investigate consists in adjusting the current $I = \int_{\Gamma_0} \mathbf{J} \cdot d\boldsymbol{\nu}$ at some given Ohmic contact $\Gamma_0 \subset \Gamma_D$ via a change in the reference doping profile $\bar{C}$. At the contact Γ_0 the desired current I_g is prescribed and deviations of the doping profile from $\bar{C}$ are allowed in order to achieve this current flow. In other words the minimization of cost functionals of the form

$$F_{\gamma_1,\gamma_2}(n, W, \phi, C) = \frac{\gamma_1}{2}\left[\int_{\Gamma_0} \mathbf{J} \cdot d\boldsymbol{\nu} - I_g\right]^2 + \frac{\gamma_2}{2}\int_{\Omega} |\nabla(C - \bar{C})|^2 dx, \tag{6}$$

is considered, where γ_1 and γ_2 are nonnegative balance parameters. C enters as a source term in the MEP energy transport model, which can be interpreted as a constraint to the minimization problem, determining the current I by the state variables (n, W, ϕ). In the following we set $I_g = 1.5 \cdot \bar{I}$.

We can rewrite (1) and (2) as $f(n, W, \phi, C) = 0$, where the nonlinear mapping $f : X \times \mathscr{C} \to Z^*$ is defined by

$$f(n, W, \phi, C) \stackrel{\text{def}}{=} \begin{pmatrix} div\, \mathbf{J} \\ div\, \mathbf{H} - q\mathbf{J} \cdot \nabla\phi - nC_W \\ \operatorname{div}(\epsilon\nabla\phi) - qn + qC \end{pmatrix}. \tag{7}$$

Altogether, this yields the constrained minimization problem

$$\min_{X \times \mathscr{C}} F_{\gamma_1,\gamma_2}(n, W, \phi, C) \text{ subject to } f(n, W, \phi, C) = 0 \text{ and the boundary condition (5).}$$

Here $X = (n, W, \phi) = y_D + X_0$ is the state space with $y_D \stackrel{\text{def}}{=} (n_D, 0, \phi_D)$ denoting the boundary data, and $X_0 = H_0^1(\Omega \cup \Gamma_N) \times H^1_{0,\partial\Omega}(\Omega \cup \partial\Omega) \times \big(H_0^1(\Omega \cup \Gamma_N) \cap L^\infty(\Omega)\big)$, equipped with the norm $\|y\|_{X_0} \stackrel{\text{def}}{=} \|n\|_{H^1(\Omega)} + \|W\|_{H^1(\Omega)} + \|\phi\|_{H^1(\Omega)} + \|\phi\|_{L^\infty(\Omega)}$. $Z \stackrel{\text{def}}{=} [H^1(\Omega)]^3$ is the set of the co-states while $\mathscr{C}$ is the set of the admissible controls given by $\mathscr{C} = \{C \in H^1(\Omega) : C = \bar{C} \text{ on } \Gamma_D\}$.

4 First-Order Optimality System

The first-order optimality system is derived by using the Lagrangian $\mathscr{L} : X \times \mathscr{C} \times Z \to \mathbb{R}$ associated to the minimization problem

$$\mathscr{L}(n, W, \phi, C; \mu_1, \mu_2, \mu_3) \stackrel{\text{def}}{=} F_{\gamma_1,\gamma_2}(n, W, \phi, C) + \int_\Omega \mu_1 \,\text{div}\mathbf{J}\, dx + \int_\Omega \mu_2 \,\text{div}\mathbf{H}\, dx$$
$$-q\int_\Omega \mu_2\, \mathbf{J}\cdot\nabla\phi\, dx - \int_\Omega \mu_2\, nC_W\, dx + \int_\Omega \mu_3 \,\text{div}(\epsilon\nabla\phi)\, dx - q\int_\Omega \mu_3\,(n - C)\, dx$$

where $\mu = (\mu_1, \mu_2, \mu_3) \in Z$ are the lagrangian multipliers arising from the constraints (1) and (2) . Thus the first order optimality system is given by

$$\nabla_{(y,C,\mu)}\mathscr{L}(y, C, \mu) = 0. \tag{8}$$

The variations of $\mathscr{L}$ with respect to μ yield the state equations $f(y, C) = 0$ itself, while the variation with respect to the state variable $y = (n, W, \phi)$ leads to the co-state equations (see [13] for the details) subject to the boundary conditions:

$$\mu_1 = \gamma_1\left[-\int_{\Gamma_0}\mathbf{J}\cdot\boldsymbol{\nu} + I_g\right] \text{ on } \Gamma_0; \quad \mu_1 = 0 \quad \text{on } \Gamma_D\setminus\Gamma_0; \quad \nabla\mu_1\cdot\boldsymbol{\nu} = 0 \quad \text{on } \Gamma_N \tag{9}$$

$$\mu_2 = \mu_3 = 0 \quad \text{on } \Gamma_D \quad \text{and} \quad \nabla\mu_2\cdot\boldsymbol{\nu} = \nabla\mu_3\cdot\boldsymbol{\nu} = 0 \quad \text{on } \Gamma_N \tag{10}$$

Moreover taking variation of $\mathscr{L}$ w.r.t the control $C \in \mathscr{C}$ leads to the optimality condition given by:

$$\gamma_2\triangle(C - \bar{C}) = q\mu_3 \tag{11a}$$

along with the boundary conditions

$$C = \bar{C} \text{ on } \Gamma_D, \quad \nabla C\cdot\boldsymbol{\nu} = \nabla\bar{C}\cdot\boldsymbol{\nu} \text{ on } \Gamma_N. \tag{11b}$$

Thus the first-order necessary optimality condition (8) consists of the state equations (1) and (2), the adjoint system and the optimality conditions (11).

5 Numerical Method

For the solution of (3), we will adopt the following steepest descent gradient algorithm: (1) choose $C_0 \in \mathscr{C}$; (2) for $k = 1, 2, \ldots$ compute $C_k = C_{k-1} - \theta_k\hat{F}'(C_{k-1})$. $\hat{F}(C) \stackrel{\text{def}}{=} F(y(C), C)$ denotes the reduced cost functional, where $y = (n, W, \phi)$ stands for the state variables and $\hat{F}'(C)$ is the Riesz representative of its

first variation. The evaluation of $\hat{F}'(C)$ requires the solution of the nonlinear state system (1) and (2) for the state variables (n, W, ϕ), as well as a solution of the linear adjoint system for μ, and finally solve a linear Poisson problem (11) to get the correct Riesz representative.

In the following simulations we have used a constant value for the parameter θ_k, which has revealed promising for the overall performance of the algorithm.

In [13] numerical results are shown for the one-dimensional n^+-n-n^+ diodes. The semiconductor domain is given by the interval $\Omega = (0, L)$, with $L > 0$. As reference doping profile a step–wise function is taken

$$\bar{C}(x) = \begin{cases} N_D^+ \ x \in [0, (L - L_c)/2] \\ N_D \ \ x \in](L - L_c)/2, (L + L_c)/2[\\ N_D^+ \ x \in [(L + L_c)/2, L] \end{cases}$$

where L_c is the channel length and $N_D^+ = 10^{18}$ cm^{-3}, $N_D = 10^{16}$ cm^{-3} are the high and low doping concentrations, respectively. The length of the n^+-regions is 0.1 μm. For the channel length the following cases were considered: 0.4 and 0.2 μm.

As boundary conditions we imposed that the density equals the doping and set the voltage equal to zero and to the bias voltage on the right and left boundaries, respectively. Concerning the energy we impose homogeneous Neumann condition at the edges of the device [16]. We set the parameters $\gamma_1 = 1$ V ps and $\gamma_2 = 10^{-20}$ eV μm^5/ ps and take a constant step-size $\theta_k = 10^{-3}$. The iteration stops when the difference of two consecutive iterates over the last iterate is below some specified threshold we have taken equal to 0.001.

The state system has been discretized by a variant of the well-known exponentially fitted Scharfetter–Gummel scheme [16]. The computation has been performed on a uniform grid of 120 and 96 intervals, respectively for the 0.4 μm and the 0.2 μm channel length case. The numerical values of the physical parameters used in the numerical simulations are those in [16].

6 A Sensitivity Analysis

In order to assess the robustness of the proposed optimization method even in the case where there may be a loss of regularity, we introduce a Gaussian noise in the reference doping profile with zero mean and standard deviation equals to 0.1.

In Figs. 1 and 2 we report the results obtained for the 0.4 and 0.2 μm channel length cases. Although the irregularity in the doping profile, the numerical approach has proved very efficient and stable. Moreover also a diffusive effect has been included, by replacing the stepwise function used in the previous section for $C(x)$ with the following regularization

$$C(x) = C_0 - d_0 \left(\tanh \frac{x - x_1}{s} - \tanh \frac{x - x_2}{s} \right),$$

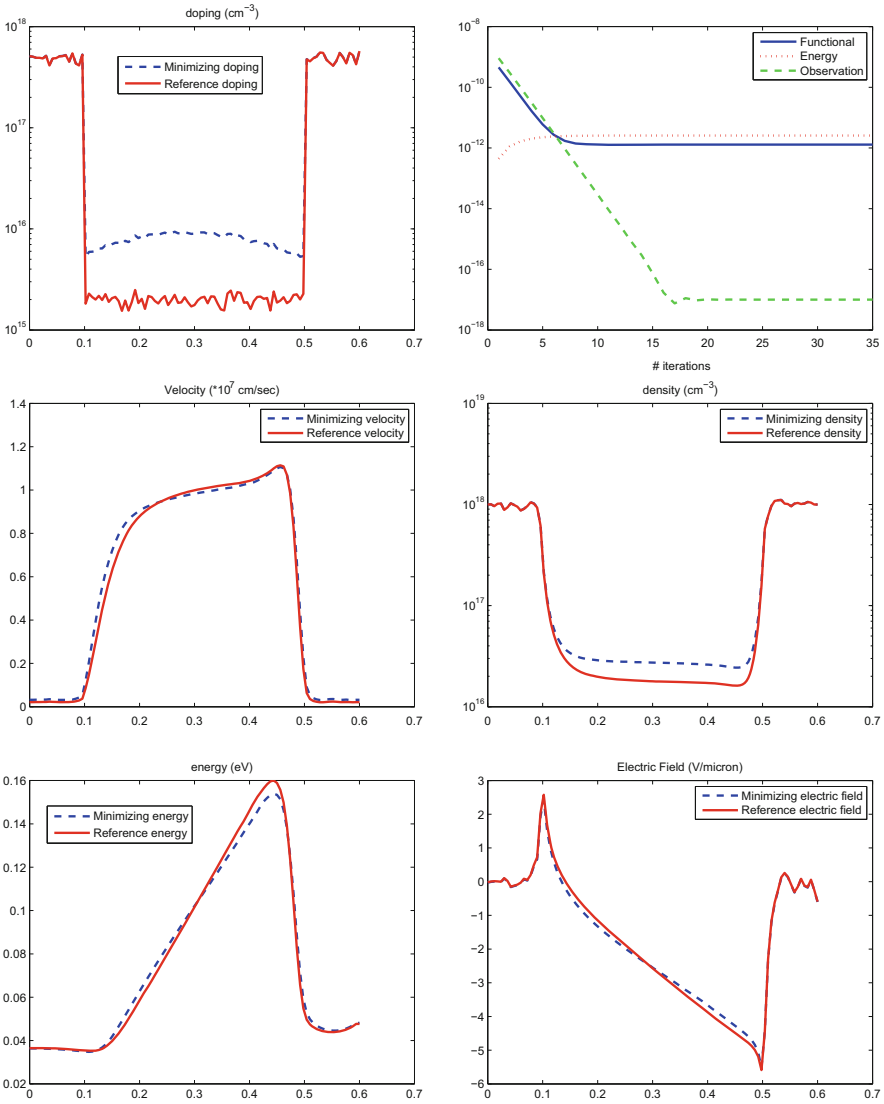


Fig. 1 Optimized doping profile, evolution of the cost functional, electron mean velocity, electron density, energy and electric field for a biasing voltage of 1 V and a channel of 0.4 μ

where $C_0 = C(0)$, $d_0 = C_0\left(1 - \frac{N_D}{N_D^+}\right)/2$, $x_1 = 0.1\,\mu\text{m}$, $x_2 = x_1 + L_c$. The parameter s gives a measure of the diffusion of the dopants into the crystal and produces a change of the channel length, influencing the values of the current.

The value s has been randomly generated according to a Gaussian distribution in samples of 20 elements. In Table 1 we report the results obtained for a channel

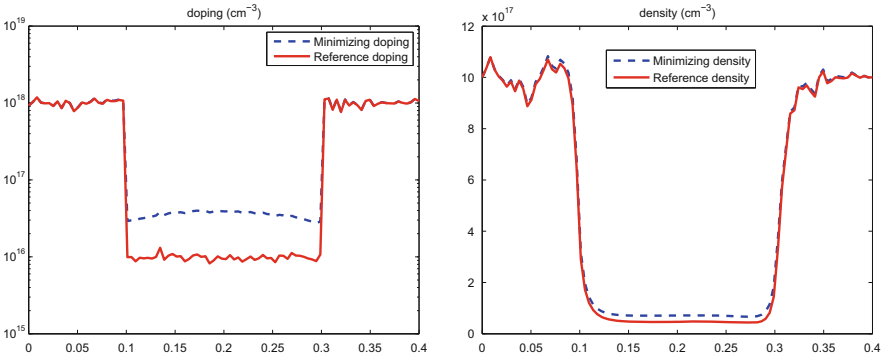


Fig. 2 Optimized doping profile and electron density. The qualitative behavior of the other variables are as in Fig. 1

Table 1 Current mean values and standard deviations for Gaussian noise introduced into the reference doping profile

Noise mean	Noise standard deviation	Current mean value	Current standard deviation
0.001	0.001	1,268,949.2178	2867.4900
0.001	0.01	1,345,523.8898	78,239.9692
0.001	0.1	15,245,077,447.3599	123,470.9579

length of 0.25 μm, by varying the mean and the standard deviation. As expected the larger is the standard deviation, the larger is the spread of the current.

References

1. Burger, M., Hinze, M., Pinnau, R.: Optimization models for semiconductor dopant profiling. Transport Phenomena and Kinetic Theory (MSSETS), pp. 91–115. Birkhäuser, Boston (2007)
2. Burger, M., Pinnau, R.: Fast optimal design for semiconductor devices. SIAM J. Appl. Math. **64**(1), 108–126 (2003)
3. Hinze, M., Pinnau, R.: An optimal control approach to semiconductor design. Math. Models Methods Appl. Sci. **12**(1), 89–107 (2002)
4. Hinze, M., Pinnau, R.: Mathematical tools in optimal semiconductor design. Bull. Inst. Math. Acad. Sin. (New Series) **2**(2), 569–586 (2007)
5. Drago, C.R., Anile, A.M.: An optimal control approach for an energy transport model in semiconductor design. In: Scientific Computing in Electrical Engineering, SCEE'04. Mathematics in Industry, vol. 9, pp. 323–331. Springer, Berlin (2006)
6. Drago, C.R.: On fast optimal control for energy-transport-based semiconductor design. In: Scientific Computing in Electrical Engineering, SCEE'06. Mathematics in Industry, vol. 11, pp. 347–355. Springer, Berlin (2007)
7. Drago, C.R., Pinnau, R.: Optimal dopant profiling based on energy-transport semiconductor models. Math. Models Methods Appl. Sci. **18**(2), 195–214 (2008)
8. Drago, C.R., Marheineke, N., Pinnau, R.: Semiconductor device optimization in the presence of thermal effects. Z. Angew. Math. Mech. **93**(9), 700-705 (2013)

9. Jüngel, A.: Quasi-Hydrodynamic Semiconductor Equations. Birkhauser, Boston (2001)
10. Anile, A.M., Romano, V.: Non-parabolic band transport in semiconductors: closure of the moment equations. Contin. Mech. Thermodyn. Phys. **11**, 307–325 (1999)
11. Romano, V.: Non-parabolic band transpor in semiconductors: closure of the production terms in the moment equations. Contin. Mech. Thermodyn. **12**, 31–51 (2000)
12. Romano, V.: Non-parabolic band hydrodynamical model for silicon semiconductors and simulation of electron devices. Math. Methods Appl. Sci. **24**, 439–471 (2001)
13. Drago, C.R., Romano, V.: An optimal control for semiconductor diodes design based on the MEP energy-transport model. Preprint, University of Catania
14. Stracquadanio, G., Romano, V., Nicosia, G., Semiconductor device design using the BIMADS algorithm. J. Comput. Phys. **242**, 304–320 (2013)
15. Romano, V.: 2D simulation of a silicon MESFET with a non parabolic hydrodynamical model based on the maximum entropy principle. J. Comput. Phys. **176**, 70–92 (2002)
16. Romano, V.: 2D numerical simulation of the MEP energy-transport model with a finite difference scheme. J. Comput. Phys. **221**, 439–468 (2007)
17. Anile, A.M., Mascali, G., Romano, V.: Recent developments in hydrodynamical modeling of semiconductors. In: Mathematical Problems in Semiconductor Physics. Lecture Notes in Mathematics, vol. 1832, pp. 1–54. Springer, Berlin (2003)

MS 18
MINISYMPOSIUM: MULTIPHYSICS SIMULATION IN ELECTRICAL ENGINEERING

Organizers

Michael Günther[1]

Speakers

Johanna Kerler[2] and Tatjana Stykel[2]
Nonlinear Model Reduction for Simulation of Coupled Systems

Lennart Jansen[3]
Exponential Integration Methods for Water Transportation Networks

Christoph Hachtel[4], Michael Günther[4] and Andreas Bartel[4]
Model Order Reduction for Multirate ODE-Solvers in Multiphysics Applications

[1]Michael Günther, Bergische Universität Wuppertal, Germany.

[2]Ljohanna Kerler and Tatjana Stykel, Universität Augsburg, Germany.

[3]Lennart Jansen, Humboldt Universität zu Berlin, Germany.

[4]Christoph Hachtel, Michael Günther and Andreas Bartel, Bergische Universität Wuppertal, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_63

Valentina Koliskina[5], Andrei Kolyshkin[5], Olev Martens[5], Rauno Gordon[5], Raul Land[5] and Andrei Pokatilov[5]
Eddy Current Model for Non-destructive Testing of Electrically Conducting Materials with Cylindrical Symmetry

Keywords

Change in impedance
Eddy current testing
Model order reduction
Multiphysics application
Multirate integration schemes
TREE method

Short Description

Coupled problems usually arise in electrical engineering problems, if multiphysical effects have to be addressed. These problems consist of systems of ordinary differential equations, differential-algebraic equations and partial differential equations, which are linked via source terms and boundary conditions. To be feasible, the simulation has to be tailored to the coupled structure of the problem, exploiting the different activity levels in the different sub systems. One idea, discussed in the contribution by Hachtel et al., is to combine the ideas of model order reduction with multirate time integration. This work is part of structured research program *KoSMos: Model reduction based simulation of coupled PDAE systems* funded by the German ministry on research and technology.

The second contribution by Koliskina et al. deals with the simulation of eddy current problems for quality testing of conducting materials, which demand for modelling both electromagnetic and electrical effects. Here the authors propose a semi-analytical approach.

[5]Valentina Koloskina, Andrei Kolyshkin, Olev Martens, Rauno Gordon, Raul Land and Andrei Pokatilov, Riga Technical University, Latvia.

Eddy Current Model for Nondestructive Testing of Electrically Conducting Materials with Cylindrical Symmetry

Valentina Koliskina, Andrei Kolyshkin, Olev Märtens, Rauno Gordon, Raul Land, and Andrei Pokatilov

Abstract Eddy current method is widely used in practice for quality testing of conducting materials (examples include determination of electrical conductivity, thickness of metal coatings, identification of flaws in a conducting medium). In the present paper a semi-analytical method for solution of direct eddy current problems for the case of a conducting medium of finite size is considered. The method is applied to several eddy current problems with cylindrical symmetry. The following problem is analyzed in detail. Consider a coil with alternating current located above a conducting medium in the form of a circular cylinder (such a model can be used for design of coin validators which are based on the estimation of electrical conductivity of a coin). We assume that the electromagnetic field is exactly zero at a sufficiently large distance from the coil (the distance can be chosen on the basis of the required accuracy of the solution). The solution is constructed using the method of separation of variables which includes two steps where numerical calculations are necessary: (a) computation of complex eigenvalues without good initial guess for the roots and (b) solution of a system of linear algebraic equations. Computations of the change in impedance of the coil for different frequencies with the semi-analytical method are in good agreement with experimental data and results of numerical simulation with finite element method. Solution of other problems with cylindrical symmetry is also discussed (a flaw in the form of a circular cylinder in a conducting half-space or a plate). Such models can be used for the analysis of quality of spot welding (in case of a volumetric flaw) and estimation of the effect of corrosion (for surface flaws).

Keywords Eddy current model • Electrically conducting materials

V. Koliskina • A. Kolyshkin (✉)
Riga Technical University, 1 Kalku Street, Riga, LV 1658, Latvia
e-mail: v.koliskina@gmail.com; andrejs.koliskins@rtu.lv

O. Märtens • R. Gordon • R. Land
Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
e-mail: olev.martens@ttu.ee; rauno.gordon@ttu.ee; raul.land@ttu.ee

A. Pokatilov
Metrosert, Teaduspargi 8, 12618 Tallinn, Estonia
e-mail: andrei.pokatilov@metrosert.ee

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_64

1 Introduction

Mathematical models of eddy current testing problems are usually based on the assumption that a conducting medium is infinite in one or two spatial dimensions [1, 2, 6]. Analytical solutions of the corresponding systems of equations for the vector potential can be obtained in such cases by the method of integral transforms (for example, Fourier or Hankel integral transforms).

Recently a quasi-analytical approach for the solution of eddy current testing problems is suggested in [3]. The main idea of the method is that the vector potential is assumed to be exactly zero at a sufficiently large distance $r = b$ from eddy current coil. Recommendations on the selection of the value of b are given in [3].

The main advantage of the proposed method (called the TREE method by the authors in [3]) in comparison with analytical methods used for infinite domains is that with the TREE method one can construct quasi-analytical solutions for the cases where a conducting medium has a finite size. Such models are important in applications since with the TREE method it is possible to model the presence of inhomogeneities (or flaws) in the conducting medium.

In the present paper we construct quasi-analytical solutions for the case where a cylindrical coil with alternating current is located above a conducting medium which is either in the form of a circular cylinder of finite height and radius or contains a flaw of cylindrical shape whose axis coincides with the axis of the coil. The solution of the first problem is discussed in detail. Such a model can be used for design of coin validators which are based on the estimation of electrical conductivity of a coin.

2 Problem Solution

Consider a coil with alternating current of frequency f ($\omega = 2\pi f$) located above a conducting cylinder of radius c and height d. The axis of the coil coincides with the axis of the cylinder. The inner and outer radii of the coil are r_1 and r_2, respectively, z_1 is the lift-off and $z_2 - z_1$ is the height of the coil. The number of turns in the coil is denoted by N. The starting point (as usual in such type of problems) is the solution for a filamentary coil of radius r_0 located at a distance h from the cylinder.

We introduce a system of cylindrical polar coordinates (r, φ, z) centered at the axis of the coil (the plane $z = 0$ coincides with the top surface of the cylinder). Due to azimuthal symmetry the vector potential has only one non-zero component in the φ-direction. Let us denote by R_0, R_1 and R_2 the regions in space where $z > 0$, $-d < z < 0$ and $z < -d$, respectively. The amplitudes of the vector potential in regions R_0, R_1 and R_2 are denoted by A_0, A_1 and A_2, respectively. Since region R_1 is not homogeneous in the radial direction we use the notations A_1^{con} and A_1^{air} in regions $0 < r < c$ and $c < r < b$, respectively. The system of equations for the amplitudes

of the vector potential in regions R_0, R_1 and R_2 has the form

$$\frac{\partial^2 A_0}{\partial r^2} + \frac{1}{r}\frac{\partial A_0}{\partial r} - \frac{A_0}{r^2} + \frac{\partial^2 A_0}{\partial z^2} = -\mu_0 I\delta(r - r_0)\delta(z - h), \tag{1}$$

$$\frac{\partial^2 A_1}{\partial r^2} + \frac{1}{r}\frac{\partial A_1}{\partial r} - \frac{A_1}{r^2} - j\omega\sigma(r)\mu_0 A_1 + \frac{\partial^2 A_1}{\partial z^2} = 0, \tag{2}$$

$$\frac{\partial^2 A_2}{\partial r^2} + \frac{1}{r}\frac{\partial A_2}{\partial r} - \frac{A_2}{r^2} + \frac{\partial^2 A_2}{\partial z^2} = 0, \tag{3}$$

where $j = \sqrt{-1}$, $\sigma(r) = 0$ if $c < r < b$ and $\sigma(r) = \sigma$ if $0 < r < c$. The current in the coil is assumed to be of the form

$$\boldsymbol{I}^e = Ie^{j\omega t}\boldsymbol{e}_\varphi. \tag{4}$$

The functions A_0, A_1 and A_2 are equal to zero at $r = b$:

$$A_i \mid_{r=b} = 0, \quad i = 0, 2, \quad A_1^{air} \mid_{r=b} = 0. \tag{5}$$

Interface conditions at $r = c$ have the form

$$A_1^{con} \mid_{r=c} = A_1^{air} \mid_{r=c}, \quad \frac{\partial A_1^{con}}{\partial r} \mid_{r=c} = \frac{\partial A_1^{air}}{\partial r} \mid_{r=c}. \tag{6}$$

The conditions at $z = 0$ and $z = -d$ are

$$A_0 \mid_{z=0} = A_1 \mid_{z=0}, \quad \frac{\partial A_0}{\partial z} \mid_{z=0} = \frac{\partial A_1}{\partial z} \mid_{z=0}, \tag{7}$$

$$A_1 \mid_{z=-d} = A_2 \mid_{z=-d}, \quad \frac{\partial A_1}{\partial z} \mid_{z=-d} = \frac{\partial A_2}{\partial z} \mid_{z=-d}. \tag{8}$$

Using the method of separation of variables and superposition principle, the solution in region R_0 is obtained as follows

$$A_0(r, z) = \sum_{i=1}^{\infty} D_{1i} e^{-\lambda_i z} J_1(\lambda_i r) + \frac{\mu_0 I r_0}{b^2} \sum_{i=1}^{\infty} \frac{J_1(\lambda_i r_0)}{\lambda_i J_0^2(\lambda_i b)} e^{-\lambda_i |z-h|} J_1(\lambda_i r), \tag{9}$$

where $\lambda_i = \alpha_i / b$ and α_i, $i = 1, 2, \ldots$ are the roots of the equation $J_1(\alpha) = 0$.

The solution in region R_1 can be written in the form

$$A_1^{con}(r,z) = \sum_{i=1}^{\infty}[D_{2i}e^{p_i z} + D_{3i}e^{-p_i z}J_1(q_i r), \tag{10}$$

$$A_1^{air}(r,z) = \sum_{i=1}^{\infty}[(D_{4i}J_1(p_i r) + D_{5i}Y_1(p_i r))e^{p_i z} + (D_{6i}J_1(p_i r) + D_{7i}Y_1(p_i r))e^{-p_i z}, \tag{11}$$

where p_i are unknown eigenvalues and $p_i = \sqrt{q_i^2 + j\omega\sigma\mu_0}$. The solution in region R_2 is

$$A_2(r,z) = \sum_{i=1}^{\infty} D_{8i}e^{\lambda_i z}J_1(\lambda_i r). \tag{12}$$

It can be shown that eigenvalues p_i are the complex roots of the equation

$$p_i J_1(q_i c)T'(p_i c) = q_i J_1'(q_i c)T(p_i c), \tag{13}$$

where $T(p_i r) = J_1(p_i r)Y_1(p_i b) - J_1(p_i b)Y_1(p_i r)$. The unknown coefficients in (9)–(12) are determined from boundary and interface conditions (5)–(8). The resulting formulas for the coefficients are bulky and are not presented here. The details of the derivation for similar problems can be found elsewhere [4, 5]. It can be shown that the induced vector potential in air due to the presence of the conducting cylinder is given by the formula

$$A_0^{ind}(r,z,r_0,h) = \sum_{i=1}^{n} D_{1i}e^{-\lambda_i z}J_1(\lambda_i r), \tag{14}$$

where the series are truncated at $i = n$.

The induced vector potential in air due to currents in the whole coil is given by

$$A_{0coil}^{ind}(r,z) = \int_{r_1}^{r_2}\int_{z_1}^{z_2} A_0^{ind}(r,z,r_0,h)dr_0 dh. \tag{15}$$

Using (15) and the following formula for the computation of the change in impedance of the coil over the volume, V, of the coil

$$Z^{ind} = \frac{i\omega}{I^2}\iiint_V \boldsymbol{A}_{0coil}^{ind}\cdot\boldsymbol{I}\,dV$$

we obtain

$$Z^{ind} = \frac{2j\omega\pi\mu_0 N^2}{(r_2 - r_1)^2 (z_2 - z_1)^2} \sum_{m=1}^{n} \frac{(e^{-\lambda_m z_2} - e^{-\lambda_m z_1})}{\lambda_m^3} \int_{\lambda_m r_1}^{\lambda_m r_2} \xi J_1(\xi)\, d\xi$$
$$\times \sum_{i=1}^{n} Y_{mi} \frac{(e^{-\lambda_i z_2} - e^{-\lambda_i z_1})}{\lambda_i^3} \int_{\lambda_i r_1}^{\lambda_i r_2} \eta J_1(\eta)\, d\eta. \tag{16}$$

The elements of the matrix Y are bulky and are not shown here for brevity.

3 Numerical Results and Discussion

Formula (16) is used to compute the change in impedance of the coil. Calculations are performed with "Mathematica". The following values of the parameters are used for calculations: $\sigma = 9.6\,\text{Ms/m}$, $b = 60\,\text{mm}$, $d = 1.93\,\text{mm}$, $r_2 = 6\,\text{mm}$, $r_1 = 3\,\text{mm}$, $z_2 = 0.39\,\text{mm}$, $z_1 = 0.06\,\text{mm}$, $N = 100$. The results of calculations are shown in Fig. 1. The calculated points (from top to bottom) correspond to the following values of the frequency f: 1125, 1598, 2270, 3224 and 4579 Hz. The smaller points of the graph represent theoretical calculations while larger points correspond to numerical simulations by finite element method. The upper limit of the summation index in (16) is fixed at $n = 68$. Comparison of the computational results obtained for other values of n showed that the chosen value is quite satisfactory in terms of computational accuracy.

Several computational steps are necessary in order to calculate the change in impedance. First, the set of eigenvalues λ_i has to be calculated. This can easily be done in "Mathematica" using the built-in routine BesselZeros. Second, a set of complex eigenvalues p_i should be calculated. The computational procedure is based

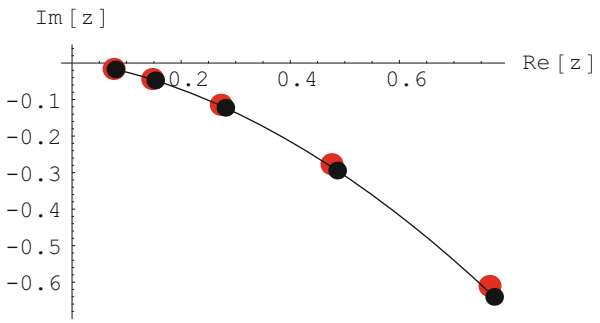


Fig. 1 Comparison of theoretical calculations with numerical simulations

on the method described in [7, 8]. Third, several systems of linear equations have to be solved numerically in order to determine expansion coefficients. Details of the numerical aspects of the procedure are given in [9].

In addition, the problem is solved by means of a finite element method. The details of numerical simulation can be found in [10]. As can be seen from Fig. 1, good agreement is found between theory (semi-analytical method presented in the paper) and finite element method.

The problem is also analyzed experimentally. Experiments are performed at Tallinn University of Technology as a part of the work on the SAFEMETAL project "Increasing EU citizen security by utilizing innovative intelligent signal processing systems for euro-coin validation and metal quality testing" in the framework of FP7 program for the period from 2010 till 2012. Experimental results related to the project are published in [11].

The approach presented in the paper can be used for other axisymmetric problems in eddy current testing. In particular, quasi-analytical solutions for the case of cylindrical flaws (either in the form of a cylindrical inclusion in half-space or surface flaws in a half-space or plate) are constructed in [4, 5, 9].

Acknowledgements This work was partially supported by the grant 623/2014 of the Latvian Council of Science.

References

1. Tegopoulos, J.A., Kriezis, E.E.: Eddy Currents in Linear Conducting Media. Elsevier, Amsterdam (1985)
2. Antimirov, M.Ya., Kolyshkin, A.A., Vaillancourt, R.: Mathematical Models for Eddy Current Testing. CRM, Montréal (1997)
3. Theodoulidis, T.P., Kriezis, E.E.: Eddy Current Canonical Problems (with Applications to Nondestructive Evaluation). Tech Science Press, New York (2006)
4. Koliskina, V.: Calculation of the change in impedance of a coil located above a conducting medium with a flaw. In: Proceedings of the International Conference on Application of Contemporary Non-destructive Testing in Engineering, Portoroz, September 4–6, pp. 327–332. University of Ljubljana, Ljubljana (2013)
5. Kolyshkin, A.: Solution of direct eddy current problems with cylindrical symmetry. In: Proceedings of the International Conference on Application of Contemporary Non-destructive Testing in Engineering, Portoroz, September 4–6, pp. 347–351. University of Ljubljana, Ljubljana (2013)
6. Dodd, C.V., Deeds, W.E.: Analytical solutions to eddy-current probe-coil problems. J. Appl. Phys. **39**, 2829–2838 (1968)
7. Delves, L.M., Lynness, J.N.: A numerical method for locating zeros of an analytic function. Math. Comput. **21**, 543–560 (1967)
8. Lynness, J.N. : Numerical algorithms based on the theory of complex variables. In: Proceedings of the ACM, pp. 125–133 (1967)
9. Koliskina, V.: Analytical and quasi-analytical solutions of direct problems in eddy current testing. Ph.D. thesis, Riga Technical University (2013)

10. Gordon, R., Märtens, O., Kolyshkin, A.: Comparison of Eddy-Current Theory and Finite Element Method for Metal Evaluation. Lecture Notes on Impedance Spectroscopy: Measurement, Modeling and Applications, vol. 3 pp. 41–65. CRC Press/Balkema, Leiden, Boca Raton (2012)
11. Gordon, R., Märtens, O., Land, R., Min, M., Rist, M., Gavrijaseva, A., Pokatilov, A., Kolyshkin, A.: Eddy-Current Validation of Euro Coins. Lecture Notes on Impedance Spectroscopy: Measurement, Modeling and Applications, vol. 3, pp. 47–63. CRC Press/Balkema, Leiden, Boca Raton (2012)

Model Order Reduction for Multirate ODE-Solvers in a Multiphysics Application

Christoph Hachtel, Michael Günther, and Andreas Bartel

Abstract Given a multiphysics problem with components of different dynamical behaviour reduction-multirate methods start with a model order reduction of the slow part system and apply than a multirate ODE-integration to the whole system. This approach lets us profit as much as possible from properties of the given system related to computational efficiency. In this paper we present the motivation and the idea behind this reduction-multirate approach.

Keywords Model order reduction • Multirate ODE-solvers

1 Introduction

In general, the modeling of a multiphysical setting leads to a coupled system with largely differing dynamical behaviour. Possibly after a semi-discretization of the spatial variables, these models are often given by coupled systems of ODEs. Now, the existence of stiff parts suggests which type of time domain method should be applied. Furthermore the most active part, i.e., the part with the highest frequencies, determines the step size to be used.

Multirate ODE-solvers allow us to use different step sizes for each subsystems. The use of inherent step sizes for the subsystems with different dynamical behaviour gives us potential to enhance the numerical efficiency (the performance concerning computation time). The crucial part of a multirate solver is the coupling of the different scales, i.e., the computation of the coupling variables. We follow the idea of *compound-step* methods, which was first presented in [7]. Often the physics of the underlying systems justifies the usage of a certain coupling type. Although one saves computation time due to larger step sizes for latent components (macrostep), usually a large and stiff system remains to be solved in each macrostep.

C. Hachtel (✉) • M. Günther • A. Bartel
Lehrstuhl für Angewandte Mathematik und Numerische Analysis, Fachbereich C Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany
e-mail: hachtel@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de; bartel@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_65

In the last years, model order reduction has been developed to a reliable technique to solve high dimensional systems of differential equations efficiently [1]. Until now there has been no work on combining model order reduction with multirate ODE-solvers. We present related ideas and concepts of the reduction-multirate methods and a suitable multiphysics example.

2 Mixed Multirate Methods

Multirate integration schemes are interesting for systems of differential equations with parts of very different dynamic behaviour. As in most of the previous works about multirate methods we consider a system with very fast dynamic changes, the so-called active part, and a considerable slower part, the so-called latent behaviour, both parts depend on each other. In an ODE-framework that reads

$$\dot{\mathbf{y}}_A = \mathbf{f}_A\left(\mathbf{y}_A, \boxed{\mathbf{y}_L}\right) \quad \mathbf{y}_A(t_0) = \mathbf{y}_{A,0} \tag{1}$$

$$\dot{\mathbf{y}}_L = \mathbf{f}_L\left(\boxed{\mathbf{y}_A}, \mathbf{y}_L\right) \quad \mathbf{y}_L(t_0) = \mathbf{y}_{L,0}. \tag{2}$$

The coupling is illustrated by the boxes around the coupling terms. There are several approaches how to get such a partition. While [5] or [8] deal with a given monolithic system and partition it dynamically we are following the setting of a given partition like in [7] or [2]. This is justified since we are considering multiphysics problems and usually the underlying physical behaviour defines a certain dynamical behaviour. The idea of a mixed multirate integration scheme is given in [2] and is based on the idea of multirate compound step Runge-Kutta methods first presented in [7]. In the latter the coupling is realized by integrating the latent and the active component coupled together but with different stepsizes: The latent component with a large macrostep H, the active one with a small microstep of size $h = H/m$. The remaining $m - 1$ microsteps are computed with interpolating the latent component. Günther and Rentrop presented ROW-methods for multirate integration schemes in [6] and Bartel and Günther developed W-methods for compound-step multirate integrators in [3]. Here compound-step and remaining micro-steps are computed with the same integration scheme. In mixed multirate methods different schemes can be used for compound and micro-steps. That can be reasonable if the dimension of the active part is small compared to the whole system and a high accuracy is desired so a method of higher order can be applied to the remaining microsteps. This is exactly the case for the here presented topic so we follow [2] and apply a 2(3)-ROW-scheme for the compound step and a 3(4)-ROW-scheme for the remaining micro-steps. A set of coefficients can be also found in [2].

3 Model Order Reduction with Balanced Truncation

In multirate context we deal very often with a high-dimensional slow part of the ODE-system and only few active components. The question is whether we can gain efficiency not only by adapting stepsize but also to consider the dimension of the slow part: The idea is to apply a model order reduction before integrating the system. We assume a linear ODE-system for the slow part system so we can use the methods of linear model order reduction. We now present briefly the method of balanced truncation as it can be found in [1]. For a given linear time invariant (LTI) system

$$\dot{\mathbf{x}} = \mathbf{A} \cdot \mathbf{x} + \mathbf{B} \cdot \mathbf{u}(t) \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbb{R}^n \tag{3}$$

$$\mathbf{y}(t) = \mathbf{C} \cdot \mathbf{x} \tag{4}$$

model order reduction computes rectangular biorthogonal projection matrices $\mathbf{V}_r$, $\mathbf{W}_r$ so that the dimension r of reduced system matrices $\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$, $\mathbf{W}_r^T\mathbf{B}$, $\mathbf{C}\mathbf{V}_r$ is relevantly smaller than of the original system ($r \ll n$). While the output of the reduced system $\mathbf{y}_r(t)$ shall approximate the original output as good as possible. The idea of balanced truncation is now to keep all important states and truncate all states who need a lot of energy to be reached and to be observed. Truncating states that are difficult to reach and to observe become equivalent if the system is balanced. One gets such a balanced system by solving Lyapunov-Equations and construct a suitable transformation matrix. Balanced truncation has many advantages over other MOR methods, e.g. (1) the input and the output matrix are considered in the computation of projection matrices and (2) an efficient error estimate is available:

$$\|\mathbf{H} - \mathbf{H}_r\|_{\mathscr{H}_\infty} \leq 2 \sum_{i=r+1}^{n} \sigma_i \tag{5}$$

while $\mathbf{H}$ denotes the transfer function, σ the eigenvalues of the Gramian matrices of the balanced system and r is the dimension of the truncated system. Due to the fact that for this method the Lyapunov-Equations have to be solved, the method is less efficient for high dimensional problems.

4 A Multiphysics Application: An Electric-Thermal Problem

Benefits from multirate ODE-solvers can only be expected if applied to a system with differing dynamic behaviour. Here we consider an electric circuit in where the thermal behaviour of a resistor is included. This results in a coupled system of the network equations and the heat equation. While voltages change very fast, heating or cooling of devices is a much slower process. So this example suits for using multirate integration methods. Before applying the time integration, a semi-

Fig. 1 Circuit diagram

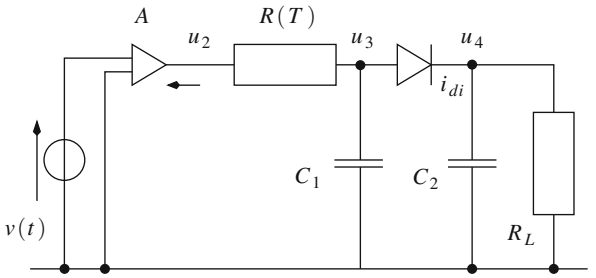


Table 1 Parameters of the electric circuit

Decide	Parameter	Decide	Capacity
Amplification	$A = 300$	Capacity 1	$C_1 = 1\,\mathrm{F}$
Load resistance	$R_L = 0.3\,\mathrm{k\Omega}$	Capacity 2	$C_2 = 100\,\mu\mathrm{F}$

Pulsed voltage source $v(t) =$

$$\begin{cases} 0.5\sin(\pi t/(2.5\cdot 10^{-5}\,\mathrm{s}))\ [\mathrm{mV}] & \text{if } t < 2.5\cdot 10^{-5}\,\mathrm{s} \\ 0\ [\mathrm{V}] & \text{otherwise} \end{cases}$$

discretisation of space is performed for the heat equation. High accuracy demands as well as fine structures may lead to a large scale system. Therefore a model order reduction of the slow components will improve efficiency. The presented model is adapted from [4] with some modifications.

Circuit Modeling The electric part is represented by the circuit diagram in Fig. 1. It is obvious that the corresponding nodal equations describe a relative stiff system of differential equations. So the mixed multirate ROW-method presented in Sect. 2 is not a *natural choice*. To be able to apply this method to this circuit we use some *unphysical* parameters amongst others for the capacitances. Table 1 shows all relevant parameters. The ODE model reads

$$C_1\dot{u}_3 = (u_2 - u_3)/R(T) - i_{\mathrm{di}}(u_3 - u_4, T_{\mathrm{di}}) \tag{6}$$

$$C_2\dot{u}_4 = i_{\mathrm{di}}(u_3 - u_4, T_{\mathrm{di}}) - u_4/R_L \tag{7}$$

with the node voltages u_3, u_4 and $u_2 = Av(t)$, the resistors's temperature T and the diode's temperature T_{di}. Between node two and three we consider a copper wire of length l and model it as a 1-D thermal dependent resistor. Let $a(x) = a_0 \cdot 1/(1 + (2/l)^2(l-x)x)$ denote the cross section of the wire while x represents the spatial coordinate; so at half of the length of the wire the cross section is half of the cross section at the ends. So we expect higher temperatures in the middle of the resistor. We assume a local resistance of the following type:

$$\rho(T) = r_0(1 + \alpha(T - T_{\mathrm{meas}})) \tag{8}$$

with thermal coefficient α and specific resistance r_0 at temperature T_{meas}. We get the total resistance $R(T)$ by integrating the local resistance over the length of the wire l

with respect to the cross section

$$R(T) = \int_0^l \frac{\rho(s, T(t,s))}{a(s)} ds = \int_0^l \tilde{\rho}(s, T(t,s)) ds. \tag{9}$$

The diode is also temperature dependent and has a strong nonlinear behaviour, for the characteristic curve and more details see [4].

Thermal Modeling and Coupling The starting point of the thermal model is the 1-D heat equation for diffusive heat transport, which we use for the copper wire (resistor):

$$M'_W \dot{T} = \frac{\partial}{\partial x}\left(\Lambda(x)\frac{\partial T}{\partial x}\right) + \text{sources} \tag{10}$$

with thermal mass of the wire M'_W and local 1-D conductivity $\Lambda(x) = \lambda(x) \cdot a(x)$. The sources term is comprised of two effects: (a) Local self heating due to the electric current. In fact, the dissipated power $P_W = u_R^2/R$ of the resistor results in heating the wire; (b) Cooling to the ambient temperature T_{env}, which is given by Newton's cooling $C = -\gamma S'(T - T_{\text{env}})$ with surface S'. For further details see [4].

To be able to apply the multirate ODE-integration scheme presented in Sect. 2, we discretise space in the parabolic PDE (10) first (method of lines). We equip the wire with an equidistant grid $I_h : X_i = i \cdot h,\ i = 0, \ldots, N$ with $X_N = N \cdot h = l$ and use a finite volume approach. For that we sub-divide the wire in cells of length h in the inner and $h/2$ at the boundaries. A schematic representation is given in Fig. 2. The heat conduction over one single cell can be described by means of its inflow minus outflow. So we get the approximation

$$M'_{W,i}\dot{T}_i = \Lambda \frac{T_{i+1} - 2T_i + T_{i-1}}{h^2} + P'_{W,i} - \gamma S'_{W,i}(T_i - T_{env}) \tag{11}$$

for the inner cells while i denotes the property of the i-th cell, $i = 1, \ldots, N-1$. For the boundary cells we have

$$M'_{W,0}\dot{T}_0 = \Lambda(T_1 - T_0)/h + P'_{W,0} - \gamma S'_{W,0}(T_0 - T_{env}) \tag{12}$$

$$M'_{W,N}\dot{T}_N = \Lambda(T_{N-1} - T_N)/h + P'_{W,N} - \gamma S'_{W,N}(T_N - T_{env}). \tag{13}$$

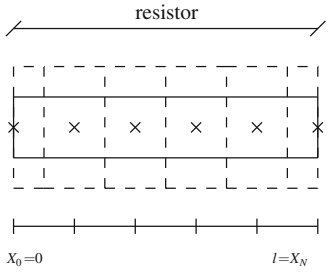


Fig. 2 Finite volume discretised resistor

The diode is temperature dependent but without own thermal mass. So we just set the temperature at the end of the copper wire to be the temperature of the diode.

The coupling terms have been given indirectly in the above models.

(1) *Circuit to thermal:* Joule's law gives the dissipated power at the resistor. By adding an additional differential equation to the circuit equations,

$$\dot{e} = u_r \cdot i_r = (u_2 - u_3)^2 / R(T) \tag{14}$$

total energy e is computed in each time step. And $P_W = e/\Delta t$ gives us the required power for some time step Δt.

(2) *Thermal to circuit:* Since the resistance $R(T)$ depends on the temperature profile T we need the temperature distribution in the resistor to compute it, for a given distribution we use Eq. (9) to compute the total resistance. In addition the diode's current depends on the wire temperature of the last cell.

Numerical Results To the coupled thermal-electric problem we apply the mixed multirate ODE-integration scheme from Sect. 2. The active part is given by the circuit equations (6)–(7) and (14) while the latent part is given by the semi-discretised heat equation (11)–(13). The coupling is realised in the compound step of the multirate scheme and in the off-diagonal Jacobian matrices. Core of this work is the combination of the multirate integration with a model order reduction for the latent part given for example in Sect. 3. In a early stage of research we restrict the setting to *linear* MOR. A thermal model is a priori non-linear, due to a relative short simulation time (0.1 s) and an even shorter timespan where voltage is applied in the circuit and several small assumptions like in Eq. (8). We can linearise the thermal behaviour of the resistor so that we get a linear system of the form

$$\dot{T} = \mathbf{A} \cdot T + \mathbf{B} \cdot [T_{env},\ P_W] \tag{15}$$

$$[R(T),\ T_N] = \mathbf{C} \cdot T. \tag{16}$$

To get an impression of the electric behaviour of the system Fig. 3 shows the resulting voltage u_3 at node three. First we are investigating the influence of the linearisation of the thermal components. Figure 4 shows the temperature of the central cell in the resistor and of the diode in a linear and a nonlinear model. As we see the influence of the linearisation is negligible.

For this results the resistor was discretised in $n = 10$ cells. If we consider a more detailed discretisation the dimension of the system and the computation time would increase. For this case a model order reduction promises more efficiency in computation time. Due to that we will focus on model order reduction performed in a pre-processing step and the influences on the simulation results in the future research.

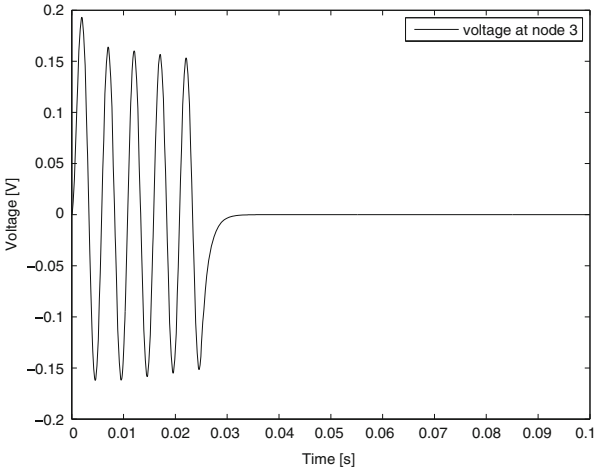


Fig. 3 Voltage at node 3

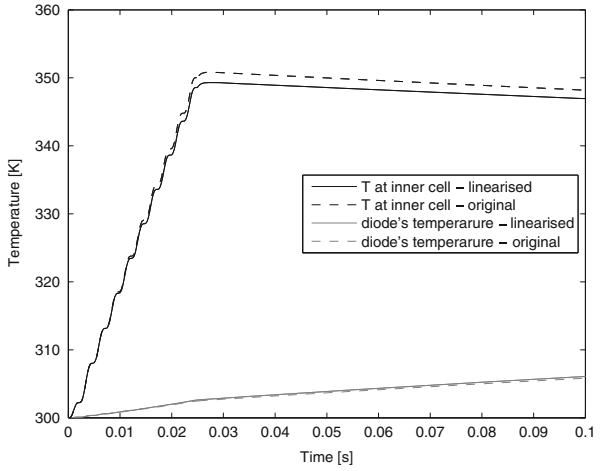


Fig. 4 Resistor's and diode's temperature

5 Conclusions and Outlook

Combining a model order reduction with a multirate integration scheme can increase the efficiency of the broad algorithm. Since the multirate problem already provides a partitioning relating to the dynamics of the system we already know that there is no fast change in the latent component so that the error due to the model order reduction might be controllable. Especially for linear latent components the wide theory of linear model order reduction gives efficient and reliable error information.

The influence of the model order reduction on the multirate integrator is not yet considered so it is desirable to have all-in-one multirate-MOR error bounds. After all in real-world multiphysics problems a latent linear component is not available so the question is whether a linearisation following linear MOR or a nonlinear MOR gives better results.

Acknowledgements This work was supported by the Research Network KoSMos: *Model Reduction Based Simulation of coupled PDAE Systems* funded by the German Federal Ministry of Education and Science (BMBF), grant no. 05M13PXA. Responsibility for the contents of this publication rests with the authors.

References

1. Antoulas, A.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia, PA (2005)
2. Bartel, A.: Multirate row methods of mixed type for circuit simulation. In: van Rienen, U., Günther, M., Hecht D. (eds.) Scientific Computing in Electrical Engineering. Lecture Notes in Computational Science and Engineering. Springer, Berlin (2001)
3. Bartel, A., Günther, M.: A multirate W-method for electrical networks in state-space formulation. J. Comput. Appl. Math. **147**, 411–425 (2002). Elsevier B.V., Amsterdam
4. Bartel, A., Günther, M., Schulz, M.: Modeling and discretization of a thermal-electric test circuit. In: Antreich, K., Bulirsch, R., Gilg, A., Rentrop, P. (eds.) Modeling, Simulation and Optimization of Integrated Circuits. International Series of Numerical Mathematics, vol. 146, pp. 187–201. Springer, Berlin (2003)
5. Engstler, C., Lubich, C.: Multirate extrapolation methods for differential equations with different time scales. Computing **58**, 173–185 (1997)
6. Günther, M., Rentrop, P.: Multirate ROW-methods and latency of electric circuits. Appl. Numer. Math. **13**, 83–102 (1993)
7. Kværnø, A., Rentrop, P.: Low order multirate Runge-Kutta methods in electric circuit simulation (1999). Preprint no. 99/1, IWRMM, University of Karlsruhe
8. Savcenco, V.: Multirate numerical integration for ordinary differential equations. Ph.D. thesis, Universiteit van Amsterdam (2007)

MS 19
MINISYMPOSIUM: MULTIPHYSICS SIMULATIONS WITH INDUSTRIAL APPLICATIONS

Organizers

Stefano Micheletti[1] and Simona Perotto[2]

Speakers

Suzanne Shontz[3] and Corina Drapaca[4]
Optimal Shunt Placement for Hydrocephalus Treatment via Patient-Specific Multiphysics Simulations

Simone Pezzuto[5], Johan Hake[6], Samuel Wall[7] and Joakim Sundnes[8]
Numerical and Constitutive Aspects of the Cardiac Electromechanical Coupling

Matteo Pischiutta[9] and Gianni Arioli[10]
A Reduced Nonlinear Model for the Simulation of Two Phase Flow in a Horizontal Pipe

[1]Stefano Micheletti, Politecnico di Milano, Italy.
[2]Simona Perotto, Politecnico di Milano, Italy.
[3]Suzanne Shontz, Mississippi State University, USA.
[4]Corina Drapaca, Pennsylvania State University, USA.
[5]Simone Pezzuto, Simula Research Laboratory, Norway.
[6]Johan Hake, Simula Research Laboratory, Norway.
[7]Samuel Wall, Simula Research Laboratory, Norway.
[8]Joakim Sundnes, Simula Research Laboratory, Norway.
[9]Matteo Pischiutta, Politecnico di Milano, Italy.
[10]Gianni Arioli, Politecnico di Milano, Italy.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_66

Marco Panzeri[11], Monica Riva[12], Alberto Guadagnini[13] and Shlomo Neuman[14]
Field Application of Groundwater Flow Data Assimilation Through Moment Equation Based Ensemble Kalman Filter

Thierry Coupez[15], Elie Hachem[16], Luisa Silva[17] and Hugues Digonnet[18]
Multiphase Flow Computation at High Reynolds with Anisotropic Adaptive Meshing and Stabilized FE Solver

Melania Carfagna[19], Filomena Iorizzo[20] and Alfio Grillo[21]
Mathematical Characterization of a Heat Pipe by Means of the Non-isothermal Cahn-Hilliard Model

Patrick Farrell[22] and Simon Funke[23]
The Optimisation of Tidal Turbine Arrays

Timo Van Opstal[24] and Harald Van Brummelen[25]
Finite-Element/Boundary-Element Coupling for Inflatables

Dimitar Iliev[26], Oleg Iliev[27], Ralf Kirsch[28], Zahra Lakdawala[29], Andro Mikelic[30] and Galina Printsypar[31]
Multiphysics and Multiscale Models for the Simulation of Filtration Processes

[11]Marco Panzeri, Politecnico di Milano, Italy.
[12]Monica Riva, Politecnico di Milano, Italy.
[13]Alberto Guadagnini, Politecnico di Milano, Italy.
[14]Shlomo Neuman, University of Arizona, USA.
[15]Thierry Coupez, Ecole Centrale de Nantes, France.
[16]Elie Hachem, Ecole des Mines de Paris, France.
[17]Luisa Silva, Ecole des Mines de Paris, France.
[18]Hugues Digonnet, Ecole des Mines de Paris, France.
[19]Melania Carfagna, DISMA “G.L. Lagrange" Politecnico di Torino, Italy.
[20]Filomena Iorizzo, ARGOTEC S.R.L., Torino, Italy.
[21]Alfio Grillo, DISMA “G.L. Lagrange” Politecnico di Torino, Italy.
[22]Patrick Farrell, University of Oxford, UK.
[23]Simon Funke, Simula Research Laboratory, Norway.
[24]Timo Van Opstal, Norwegian University of Science and Technology, Norway.
[25]Harald Van Brummelen, TUE, The Netherlands.
[26]Dimitar Iliev, Fraunhofer Institute for Industrial Mathematics, Germany.
[27]Oleg Iliev, Fraunhofer Institute for Industrial Mathematics, Germany.
[28]Ralf Kirsch, Fraunhofer Institute for Industrial Mathematics, Germany.
[29]Zahra Lakdawala, Fraunhofer Institute for Industrial Mathematics, Germany.
[30]Andro Mikelic, University “Claude Bernard" Lyon 1, France.
[31]Galina Printsypar, King Abdullah University of Science and Technology, Saudi Arabia.

Marianna Signorini[32], Stefano Micheletti[1] and Simona Perotto[2]
An Optimal Control Approach to Full Waveform Inversion (FWI)

Andreas Steinbrecher[33]
Numerical Treatment of High Index Dynamical Systems

Alfio Grillo[21], Raphael Prohl[34], Gabriel Wittum[35] and Salvatore Federico[36]
A Model of Structural Reorganisation in Statistically Oriented Fibre-Reinforced Biological Materials

Keywords

Fluid-structure interaction
Micro-electro-mechanical systems
Multiphase flow
Multiphysics simulations
Structural-pore pressure coupling

Short Description

Many problems in engineering and applied sciences are characterized by a coupling among several physical fields. This involves a strong interaction of different disciplines. Remarkable instances are provided by fluid-structure interaction problems, structural-pore pressure coupling or Micro-Electro-Mechanical Systems (MEMS). The mathematical modeling and numerical simulation of multiphysics problems is still an ambitious challenge for the scientific computing.

Goal of this mini-symposium has been to provide some examples of multiphysics modeling with a particular emphasis on industrial applications.

[32]Marianna Signorini, Politecnico di Milano, Italy.
[33]Andreas Steinbrecher, Tu Berlin, Germany.
[34]Raphael Prohl, University of Frankfurt, Germany.
[35]Gabriel Wittum, University of Frankfurt, Germany.
[36]Salvatore Federico, University of Calgary, Canada.

A Reduced Nonlinear Model for the Simulation of Two Phase Flow in a Horizontal Pipe

Matteo Pischiutta, Gianni Arioli, and Alberto Di Lullo

Abstract In the last 10 years many 3D numerical schemes have been developed for the study the flow of a mixture of liquid and gas in a pipeline (Frank, Numerical simulation of slug flow regime for an air-water two-phase flow in horizontal pipes. In: The 11th international topical meeting on nuclear reactor thermal-hydraulics (NURETH-11), Avignon, 2005; Vallée et al., Nucl Eng Des 238(3):637–646, 2008; Höhne, Experiments and numerical simulations of horizontal two-phase flow regimes. In: Proceeding of the seventh international conference on CFD in the minerals and process industries, Melbourne, 2009; Bartosiewicz et al., Nucl Eng Des 240(9):2375–2381, 2010) but although they offer a very good accuracy, they are rarely fit for modelling a long pipe, due to the high computational costs. Then one is usually led to consider 1D models, see e.g. the works of Issa and his group (Issa and Kempf, Int J Multiphase Flow 29(1):69–95, 2003). Such models offer much faster simulations than 3D schemes, on the other hand they almost completely miss the dynamics in the transversal direction. Here we present a model able of representing the full 3D dynamics, but with the computational cost typical of 1D simulation. The main feature of our model consists in describing the dynamical variables in the direction transversal to the pipe by means of a family of functions depending on a set of parameters. The model is then solved by a standard finite volume scheme.

Keywords Reduced nonlinear models • Reynolds-averaged Navier-Stokes equations • Two phase flow

1 Introduction

A common starting point for the simulation of the flow of a two phase fluid in a pipeline is provided by the Reynolds-averaged Navier-Stokes equations (RANS)

M. Pischiutta (✉) • G. Arioli
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy
e-mail: matteo.pischiutta@polimi.it; gianni.arioli@polimi.it

A. Di Lullo
ENI S.p.A, Exploration and Production Division, Milan, Italy
e-mail: alberto.dilullo@eni.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_67

[4, 8], which for the k-th phase read:

$$\frac{\partial}{\partial t}(\rho_k \alpha_k) + \nabla \cdot (\rho_k \alpha_k \mathbf{U}_k) = 0\,, \tag{1}$$

$$\frac{\partial}{\partial t}(\rho_k \alpha_k \mathbf{U}_k) + \nabla \cdot (\rho_k \alpha_k \mathbf{U}_k \otimes \mathbf{U}_k) + \nabla \cdot (\alpha_k \mathbf{R}_k^{eff}) = -\alpha_k \nabla p_k + \rho_k \alpha_k \mathbf{g} + \mathbf{M}_k\,, \tag{2}$$

where k denotes the phase ($k = g, l$), α_k is the fraction of volume occupied by the phase such that $\alpha_g + \alpha_l = 1$, $\mathbf{U}_k$ is the velocity, p_k is the pressure, $\mathbf{R}_k^{eff} = -\rho_k \nu_k^{eff} \left(\nabla \mathbf{U}_k + (\nabla \mathbf{U}_k)^T\right)$ is the strain tensor, which takes into account both viscous and turbulent effects, $\nu_k^{eff} = \nu_k + \nu_{t,k}$ is the effective viscosity, and $\mathbf{M}_k$ describes the exchange of momentum between the phases. Respective 3D numerical schemes have been developed in [2, 3, 5, 9]. Although they offer a very good accuracy, they can rarely be used for long pipe simulations because of the high computational costs. Our main purpose is the development of an efficient numerical scheme that can simulate a full 3D flow fast and accurately.

2 Equation Reduction

We partition the length of the pipeline $x = [0, L]$ in N uniform cells and we set x_i, $i = 1, \ldots, N$ to be the centres of the cells. Consider a cell $\mathscr{V}$, that is a cylinder bounded by the disks $\mathscr{A}_{in}, \mathscr{A}_{out}$ and the wall $\mathscr{W}$. Assuming constant densities, integrating Eqs. (1) and (2) in $\mathscr{V}$ and using Gauss' theorem we get:

$$\frac{\partial}{\partial t}\int_{\mathscr{V}} \rho_k \alpha_k + \int_{\partial \mathscr{V}} \rho_k \alpha_k \mathbf{U_k} \cdot \mathbf{n} = 0, \tag{3}$$

$$\frac{\partial}{\partial t}\int_{\mathscr{V}} \rho_k \alpha_k \mathbf{U}_k + \int_{\partial \mathscr{V}} \rho_k \alpha_k (\mathbf{U}_k \otimes \mathbf{U}_k) \cdot \mathbf{n} + \int_{\partial \mathscr{V}} (\alpha_k \mathbf{R}_k^{eff}) \cdot \mathbf{n} = -\int_{\mathscr{V}} \alpha_k \nabla p_k + \int_{\mathscr{V}} \rho_k \alpha_k \mathbf{g} + \int_{\mathscr{V}} \mathbf{M}_k\,, \tag{4}$$

where $\partial \mathscr{V} = \mathscr{A}_{in} \cup \mathscr{A}_{out} \cup \mathscr{W}$ e $\mathbf{n}$ is the normal vector to the surface. We assume that the transversal components of the velocity are zero, $\mathbf{U}_k = (u_k, 0, 0)^T$, where $u_k = u_k(x, y, z)$, and constant densities. Then $\mathbf{M}_k = (M_k, 0, 0)$. We enforce the no-slip boundary condition on the wall, $u_k|_{\mathscr{W}} = 0$, so that the conservation equations become:

$$\frac{\partial}{\partial t}\int_{\mathscr{V}} \alpha_k + \left[\int_{\mathscr{A}} \alpha_k u_k\right]_{in}^{out} = 0, \tag{5}$$

$$\frac{\partial}{\partial t}\int_{\mathscr{V}} \alpha_k u_k + \left[\int_{\mathscr{A}} \alpha_k u_k^2\right]_{in}^{out} - \left[\int_{\mathscr{A}} 2\alpha_k \nu_k^{eff} \frac{\partial u_k}{\partial x}\right]_{in}^{out} = -\int_{\mathscr{V}} \frac{\alpha_k}{\rho_k} \frac{\partial p_k}{\partial x} + \int_{\mathscr{W}} \alpha_k \nu_k^{eff} \frac{\partial u_k}{\partial \mathbf{n}} + \int_{\mathscr{V}} \frac{M_k}{\rho_k}\,, \tag{6}$$

where $\left[\int_{\mathscr{A}} \psi\right]_{in}^{out} = \int_{\mathscr{A}_{out}} \psi - \int_{\mathscr{A}_{in}} \psi$. Our model consists in describing the transverse profile of the unknown variables α_k and u_k at each point in space through given functions depending on some parameters which in turn vary in the longitudinal direction. Our choice for $\alpha_g(x, y, z)$ is as follows:

$$\alpha_g(x, y, z) = \begin{cases} 0 & y \le \beta(x), \\ \dfrac{y - \beta(x)}{\delta(x)} & \beta(x) < y < \beta(x) + \delta(x), \\ 1 & y \ge \beta(x) + \delta(x), \end{cases} \tag{7}$$

where $\beta(x)$ is related to the level of the liquid and $\delta(x)$ is the width of the interface. Concerning the velocity, as a first simple attempt to model the flow, we choose a profile parabolic in z:

$$u_k(x, y, z) = U_k(x, y) \frac{R^2 - y^2 - z^2}{R^2 - y^2}, \tag{8}$$

where R is the (constant) radius of the pipe. We also choose a linear dependence on y for the liquid phase linear on the plane $z = 0$:

$$U_l(x, y) = \frac{y + R}{\beta(x) + R} \gamma(x), \quad y \le \beta(x) + \delta(x), \tag{9}$$

and a quadratic dependence for the gas phase:

$$U_g(x, y) = \zeta(x)(y - R)(\beta(x) - y) + \gamma(x) \frac{y - R}{\beta(x) - R}, \quad y \ge \beta(x). \tag{10}$$

The graphs of the center line profiles of the variables are reported in Fig. 1.

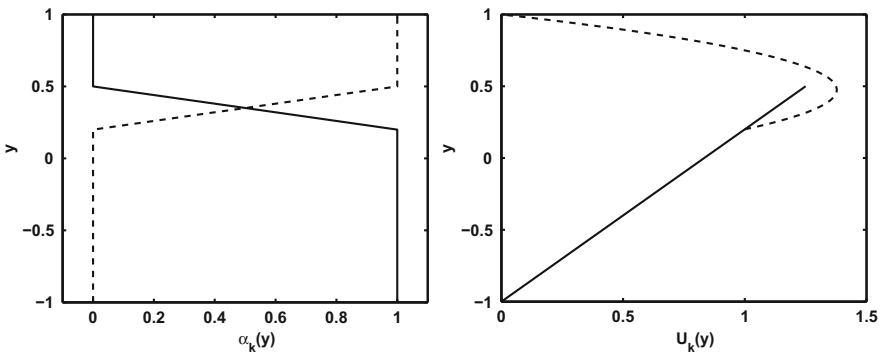


Fig. 1 *Left*: graph of α_k, *right*: graph of U_k. In *solid line*, the graphs relative to the liquid phase, in *broken line*, the graphs relative to the gas phase. The parameter are: $\beta = 0.2$, $\delta = 0.3$, $\gamma = 1$, $\zeta = 5$

We assume that the pressure of the gas is constant in the transversal directions, while the liquid is subject to the hydrostatic pressure:

$$p_g(x,y,z) = P(x)\,, \qquad p_l(x,y,z) = P(x) + \rho_l g(\beta + \delta/2 - y)\,, \tag{11}$$

which leads to a term proportional to the slope of the free surface $\partial(\beta + \delta/2)/\partial x$ in the momentum equation of the liquid phase. As a first simple approximation, we adopt a zero-equation model for the computation of the turbulent viscosity [1], that is $\nu_k = \frac{\bar{U}_k l_k}{Re_k}$, where $\bar{U}_k$ is the maximum phase velocity, l_k is a geometric length scale and $Re_k = 1000$ is a free parameter.

3 Numerical Integration

We assume that all the unknowns are constant in the cell and located in the centres of the cells. Hence, all integrals in (5) and (6) can be computed explicitly, obtaining functions of the parameters β, δ, γ e ζ. We set the following notation:

$$\mathbb{H}_l(\beta,\delta) := \int_{\mathscr{A}} \alpha_l\,, \qquad \mathbb{H}_g(\beta,\delta) := \int_{\mathscr{A}} \alpha_g\,, \tag{12}$$

$$\gamma\mathbb{F}_l(\beta,\delta) := \int_{\mathscr{A}} \alpha_l u_l\,, \qquad \zeta\mathbb{F}_{g,1}(\beta,\delta) + \gamma\mathbb{F}_{g,2}(\beta,\delta) := \int_{\mathscr{A}} \alpha_g u_g\,, \tag{13}$$

$$\mathbb{G}_l(\beta,\delta,\gamma) := \int_{\mathscr{A}} \alpha_l u_l^2\,, \qquad \mathbb{G}_g(\beta,\delta,\gamma,\zeta) := \int_{\mathscr{A}} \alpha_l u_g^2\,, \tag{14}$$

$$\mathbb{W}_l(\beta,\delta,\gamma) := \int_{\mathscr{W}} \alpha_l \frac{\partial u_l}{\partial \mathbf{n}}\,, \qquad \mathbb{W}_g(\beta,\delta,\gamma,\zeta) := \int_{\mathscr{W}} \alpha_g \frac{\partial u_g}{\partial \mathbf{n}}\,. \tag{15}$$

Time is discretized with explicit Euler method. Equation (5) reads (subscripts denote the cells, superscripts denote time):

$$\int_{\mathscr{V}} \alpha_i^{n+1} = \int_{\mathscr{V}} \alpha_i^n + \Delta t\Big[\int_{\mathscr{A}} \alpha_{i-1/2}^n u_{i-1/2}^n - \int_{\mathscr{A}} \alpha_{i+1/2}^n u_{i+1/2}^n\Big]. \tag{16}$$

The width of the interface should be dictated by some closure equation, $\delta_i^{n+1} = \mathbb{B}(\beta_i^n, \delta_i^n, \gamma_i^n, \zeta_i^n)$, but our current choice is to keep it constant. We found an optimal value at $\delta = 10^{-2}$ mm. We remark that the width of the interface affects the drag between the phases. We use the mass conservation equation of the liquid phase to update the free surface level β:

$$\mathbb{H}_l(\beta_i^{n+1}, \delta_i^{n+1}) = \mathbb{H}_l(\beta_i^n, \delta_i^n) + \frac{\Delta t}{\Delta x}\Big[\gamma_{i-1/2}^n \mathbb{F}_l(\beta_{i-1/2}^n, \delta_{i-1/2}^n) - \gamma_{i+1/2}^n \mathbb{F}_l(\beta_{i+1/2}^n, \delta_{i+1/2}^n)\Big], \tag{17}$$

that is a non-linear equation in β_i^{n+1} which is solved in every cell with standard Newton method. The conservation of the gas phase is used to obtain an equation for the pressure. To compute the interface values $\gamma_{i-1/2}^n$, $\beta_{i-1/2}^n$ and $\delta_{i-1/2}^n$ we have to interpolate. Since we are considering only positive velocities we chose the *upwind* interpolation (e.g.: $\beta_{i-1/2}^n = \beta_{i-1}^n$). We discretize the equation for the conservation of momentum for the liquid phase as follows:

$$\frac{\partial}{\partial t}\int_{\mathscr{V}} \alpha_l u_l \simeq \Delta x \frac{\partial}{\partial t}\left(\gamma_i \mathbb{F}_l(\beta_i, \delta_i)\right) \simeq \Delta x \frac{\Delta\left(\gamma_i \mathbb{F}_l(\beta_i, \delta_i)\right)}{\Delta t}, \tag{18}$$

$$\left[\int_{\mathscr{A}} \alpha_l u_l^2\right]_{in}^{out} \simeq \mathbb{G}_l(\beta_{i+1/2}, \delta_{i+1/2}, \gamma_{i+1/2}) - \mathbb{G}_l(\beta_{i-1/2}, \delta_{i-1/2}, \gamma_{i-1/2}), \tag{19}$$

$$-\int_{\mathscr{V}} \frac{\alpha_l}{\rho_l}\frac{\partial p_l}{\partial x} = -\int_{\mathscr{V}} \frac{\alpha_l}{\rho_l}\frac{\partial P}{\partial x} - \int_{\mathscr{V}} \alpha_l g \frac{\partial(\beta+\delta/2)}{\partial x} \simeq -\Delta x \mathbb{H}_l(\beta_i, \delta_i)\left(\frac{1}{\rho_l}\frac{\partial P}{\partial x}\bigg|_i + g\frac{\partial(\beta+\delta/2)}{\partial x}\bigg|_i\right). \tag{20}$$

The gradient of the pressure will be discussed later, while the slope of the free surface is handled with a centred scheme. Since the viscosity is constant in the section of the tube, we have:

$$\int_{\mathscr{W}} \alpha_l \nu_l^{eff} \frac{\partial u_l}{\partial \mathbf{n}} \simeq \nu_{l,i}^{eff}\, \mathbb{W}_l(\beta_i, \delta_i, \gamma_i). \tag{21}$$

The drag force between the phases is computed with [4, 8]:

$$M_l = \frac{3}{4}\alpha_l \alpha_g \left(\alpha_l \frac{C_{D,l}\,\rho_g}{d_l} + \alpha_g \frac{C_{D,g}\,\rho_l}{d_g}\right)|u_r|\, u_r, \tag{22}$$

where d_l and d_g are the typical diameters of the drops/bubbles of the phases, $u_r = u_g - u_l$ is the relative velocity and $C_{D,l}$, $C_{D,g}$ are coefficients computed with Schiller-Naumann formula. Since $\alpha_l\alpha_g \neq 0$ only at the interface $\beta < y < \beta + \delta$, we have:

$$\int_{\mathscr{V}} \frac{M_l}{\rho_l} \simeq \Delta x \int_{\beta}^{\beta+\delta}\left(\int_{-\sqrt{R^2-y^2}}^{-\sqrt{R^2-y^2}} \frac{M_l}{\rho_l} dz\right) dy = \frac{\Delta x}{\rho_l}\mathbb{M}_l(\beta_i, \delta_i, \gamma_i, \zeta_i). \tag{23}$$

The integrals are computed with a quadrature scheme. We compute the diffusion term using the average velocities:

$$\bar{u}_l = \frac{\int_{\mathscr{A}} \alpha_l u_l}{\int_{\mathscr{A}} \alpha_l} = \gamma \frac{\mathbb{F}_l(\beta, \delta)}{\mathbb{H}_l(\beta, \delta)}$$

approximating $\partial u_l/\partial x$ with $\partial \bar{u}_l/\partial x$ (i.e. a term that does not depend on y and z). Then we define:

$$\left[\int_{\mathscr{A}} 2\alpha_l \nu_l^{eff} \frac{\partial \bar{u}_l}{\partial x}\right]_{in}^{out} \simeq \left[2\frac{\partial \bar{u}_l}{\partial x}\nu_l^{eff}\int_{\mathscr{A}}\alpha_l\right]_{in}^{out} := \left[\mathbb{D}_l(\beta_{i+1/2},\delta_{i+1/2},\gamma_{i+1/2}) - \mathbb{D}_l(\beta_{i-1/2},\delta_{i-1/2},\gamma_{i-1/2}))\right]. \tag{24}$$

Finally, the momentum equation for the liquid phase reads:

$$\begin{aligned}\Delta x\frac{\Delta\left(\gamma_i\mathbb{F}_l(\beta_i,\delta_i)\right)}{\Delta t} &+ \left[\mathbb{G}_l(\beta_{i+1/2},\delta_{i+1/2},\gamma_{i+1/2}) - \mathbb{G}_l(\beta_{i-1/2},\delta_{i-1/2},\gamma_{i-1/2}))\right] \\ &- \left[\mathbb{D}_l(\beta_{i+1/2},\delta_{i+1/2},\gamma_{i+1/2}) - \mathbb{D}_l(\beta_{i-1/2},\delta_{i-1/2},\gamma_{i-1/2}))\right] = \\ &- \Delta x\mathbb{H}_l(\beta_i,\delta_i)\left(\frac{1}{\rho_l}\frac{\partial P}{\partial x}\bigg|_i + g\frac{\partial(\beta+\delta/2)}{\partial x}\bigg|_i\right) \\ &+ \nu_{l,i}^{eff}\,\mathbb{W}_l(\beta_i,\delta_i,\gamma_i) + \frac{\Delta x}{\rho_l}\mathbb{M}_l(\beta_i,\delta_i,\gamma_i,\zeta_i),\end{aligned} \tag{25}$$

which is a linear equation for the unknown γ_i^{n+1}. The gas phase is handled similarly, leading to a linear equation for the unknown ζ_i^{n+1}.

The equations are integrated according to the following scheme:

1. **Mass conservation**: the free surface is updated solving (17) for each cell.
2. **Velocity predictor**: the auxiliary values γ^* and ζ^* are computed with an explicit Euler iteration of the momentum equations.
3. **Pressure correction**: the updated values γ^{n+1} and ζ^{n+1} must satisfy the continuity equation $\left[\int_{\mathscr{A}}(\alpha_l^{n+1}u_l^{n+1} + \alpha_g^{n+1}u_g^{n+1})\right]_{in}^{out} = 0$, which reads:

$$\left[\gamma^{n+1}\mathbb{F}_l(\beta^{n+1},\delta^{n+1}) + \zeta^{n+1}\mathbb{F}_{g,1}(\beta^{n+1},\delta^{n+1}) + \gamma^{n+1}\mathbb{F}_{g,2}(\beta^{n+1},\delta^{n+1}))\right]_{in}^{out} = 0. \tag{26}$$

   A manipulation of Eq.(26) leads to:

$$\left[\gamma_i^*\mathbb{F}_l(\beta_i^{n+1},\delta_i^{n+1}) + \zeta_i^*\mathbb{F}_{g,1}(\beta_i^{n+1},\delta_i^{n+1}) + \gamma_i^*\mathbb{F}_{g,2}(\beta_i^{n+1},\delta_i^{n+1})\right]_{in}^{out} = \Delta t\left[\left(\frac{\mathbb{H}_l(\beta_i^{n+1},\delta_i^{n+1})}{\rho_l} + \frac{\mathbb{H}_g(\beta_i^{n+1},\delta_i^{n+1})}{\rho_g}\right)\frac{\partial P}{\partial x}\bigg|_i^*\right]_{in}^{out} \tag{27}$$

   which is the equation for the pressure correction: $\frac{\partial P}{\partial x}\big|_i^* = \frac{\partial P}{\partial x}\big|_i^{n+1} - \frac{\partial P}{\partial x}\big|_i^n$.
4. **Velocity correction**: once $\frac{\partial P}{\partial x}\big|_i^*$ is known, we can compute γ^{n+1} e ζ^{n+1}.

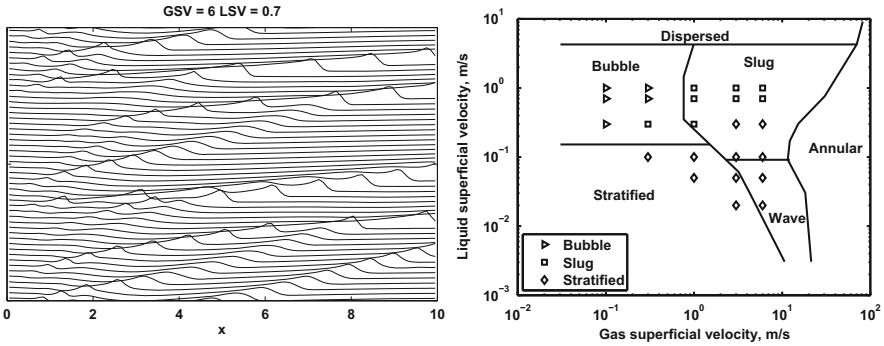


Fig. 2 Snapshots (*left*) and flow map (*right*), where the solid boundaries of the different flow regimes are provided by Mandhane et al. [7], and the mark are the results of our model

4 Results

Figure 2 represents some results of our simulations. The picture on the left displays some snapshots of the free surface in slug flow at different instants in time; the curves are represented each 0.2 s and are shifted relative to each other with respect to time. The liquid superficial velocity is $LSV = 0.7$ m/s and the gas superficial velocity is $GSV = 6$ m/s, in a pipeline with 4 cm of diameter and 10 m long. We can clearly observe that our model can reproduce the continuous generation of slug in the pipe.

The picture on the right compares our results for different regimes to the benchmark given by the flow map provided in [7]. We observe that our model is able to reproduce a good transitions between the elongated bubble and the slug regimes, whereas the exact transition between the stratified/wave flow and the slug flow is more difficult to be captured. Moreover, we cannot distinguish the stratified from the wave regime. Nevertheless, the comparison with the experimental data show that our model is able to predict the flow map with an acceptable accuracy.

5 Conclusions

These are the main features of the scheme that we introduced: The numerical algorithm is very fast, it is comparable with 1d schemes, cf. [6]. The results are in quantitative and qualitative accord with experimental data. The model is very flexible. The transversal profiles can be changed, while keeping the basic structure of the model and its numerical implementation. The numerical algorithm is stable, it does not pose any problem in the slug regions. The model has been implemented for horizontal pipelines. It could be easily adapted to moderate slopes, but it requires serious modifications to handle vertical pipelines.

In conclusion, we proved the feasibility of a reduction scheme for the two phase fluid dynamics in a pipeline, consisting in choosing a specific shape for the dynamic variables in the transversal directions depending on parameters.

Acknowledgements The financial support from ENI S.p.A. is gratefully acknowledged.

References

1. Ansys, CFX: Ansys CFX-solver theory guide (2010)
2. Bartosiewicz, Y., Seynhaeve, J.M., Vallée, C., Höhne, T., Laviéville, J.M.: Modeling free surface flows relevant to a PTS scenario: comparison between experimental data and three RANS based CFD-codes. Comments on the CFD-experiment integration and best practice guideline. Nucl. Eng. Des. **240**(9), 2375–2381 (2010)
3. Frank, T.: Numerical simulation of slug flow regime for an air-water two-phase flow in horizontal pipes. In: The 11th International Topical Meeting on Nuclear Reactor Thermal-Hydraulics (NURETH-11), Avignon (2005)
4. Hill, D.P.: The computer simulation of dispersed two-phase flow. Ph.D. thesis, University of London (1998)
5. Höhne, T.: Experiments and numerical simulations of horizontal two-phase flow regimes. In: Proceeding of the Seventh International Conference on CFD in the Minerals and Process Industries, Melbourne (2009)
6. Issa, R., Kempf, M.: Simulation of slug flow in horizontal and nearly horizontal pipes with the two-fluid model. Int. J. Multiphase Flow **29**(1), 69–95 (2003)
7. Mandhane, J., Gregory, G., Aziz, K.: A flow pattern map for gas-liquid flow in horizontal pipes. Int. J. Multiphase Flow **1**(4), 537–553 (1974)
8. Rusche, H.: Computational fluid dynamics of dispersed two-phase flows at high phase fractions. Ph.D. thesis, Imperial College (2002)
9. Vallée, C., Höhne, T., Prasser, H.M., Sühnel, T.: Experimental investigation and CFD simulation of horizontal stratified two-phase flow phenomena. Nucl. Eng. Des. **238**(3), 637–646 (2008)

Mathematical Characterisation of a Heat Pipe by Means of the Non-isothermal Cahn-Hilliard Model

Melania Carfagna, Filomena Iorizzo, and Alfio Grillo

Abstract The aim of this contribution is to provide a thorough description of a heat pipe. This is a particular type of heat exchanger used in a variety of industrial applications, such as the cooling of electrical devices and solar cells, the temperature equalisation in spacecrafts, or the reduction of local heat gains in reactors and air-conditioning systems. Usually, lumped parameter models are used to study the behaviour of heat pipes and the thermal ranges in which they work optimally. In the following analysis, a quite comprehensive thermo-fluid dynamic model of the liquid/vapour pair operating in a heat pipe is developed. The model, which accounts for several phenomena taking place in this kind of devices, has the purpose of predicting the optimal thermal range of a given heat pipe, and preventing the occurrence of off-design conditions. The present investigation is done by considering a heat pipe working in zero-gravity conditions, to be used for Aerospace applications.

Keywords Heat pipe • Non-isothermal Cahn-Hilliard model

1 Problem Statement: Mathematical and Numerical Issues

A heat pipe is a thermal device used to transport and drain heat from a hot zone to a cold one. It can be differently structured. The one considered here is a metallic tube, filled with a small amount of fluid, and then welded at the ends. One of its two ends is placed in contact with a heat source, while the other one is refrigerated. The thermodynamic conditions of the fluid inside the pipe are such that the liquid and vapour phases of the fluid coexist. The liquid phase evaporates in the hot zone of the tube, while the vapour condenses in the cold one. The processes of evaporation

M. Carfagna (✉) • A. Grillo
DISMA "G.L.Lagrange", C.so Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: melania.carfagna@polito.it; alfio.grillo@polito.it

F. Iorizzo
Argotec s.r.l., via Cervino 52, 10155 Torino, Italy
e-mail: filomena.iorizzo@argotec.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_68

and condensation form a cycle, which takes place at saturation temperature. Usually, this temperature is lower than the one at the atmospheric pressure. More precisely, the thermodynamic cycle experienced by the fluid in the heat pipe consists of the following stages:

- Evaporation of the liquid at the hot end;
- Flux of the produced vapour from the evaporator to the condenser by means of a pressure gradient induced by the thermal difference between the hot and the cold end of the pipe;
- Condensation of vapour at the cold end;
- Reflux of the condensed liquid at the evaporator zone due to a capillary structure that covers the internal wall of the tube.

When the heat pipe works in nominal conditions, the latent heat H_{lv}, which is supplied during evaporation and subtracted during condensation, balances the heat sources and sinks applied from the outside. In principle, this thermal balance maintains the whole heat pipe at almost the same temperature. Nevertheless, in order to control and predict the occurrence of off-design conditions, a non-isothermal model of the heat pipe is required [2].

The Cahn-Hilliard model is widely used to model phenomena such as capillary waves and moving contact lines [1] between two fluids as well as near-critical point phenomena, such as the spinodal decomposition and phase transitions. It describes the interface between two fluid phases as a thin transition layer in which the two phases coexist and form, thus, a mixture. To model the coupling that occurs at the interface between these two weakly miscible [7] (and often weakly compressible) fluids in contact with each other, it is postulated that the system possesses a Helmholtz free energy density of the type

$$F(\varphi, \nabla\varphi) = \tfrac{1}{2}\lambda|\nabla\varphi|^2 + U(\varphi), \quad U(\varphi) = \tfrac{1}{4}\lambda\varepsilon^{-2}(1-\varphi^2)^2. \tag{1}$$

Here, the term $\frac{1}{2}\lambda|\nabla\varphi|^2$ describes a weak interaction between the two phases (this interaction is often referred to as "non-local" in the literature), whereas $U(\varphi)$ is a globally non-convex, double-well potential vanishing at $\varphi = \pm 1$. The field φ is sometimes referred to as "colour function". In this work, it is defined as an affine function of the mass fraction c of one of the two phases, i.e. $\varphi = 2c - 1$. The parameters λ and ε are quantities related, respectively, to the chemical potential of the system at the interface, i.e., the surface tension force σ_{t}, and the thickness of the layer.

The thermodynamic consistency is fulfilled by constitutive relations that ensure the non-negativeness of the entropy production of the system. In particular, this requirement yields a characterisation of the total stress tensor $\boldsymbol{\sigma}$, entropy flux $\boldsymbol{q}_{\mathrm{s}}$, and diffusive mass flux vector $\boldsymbol{J}_{\mathrm{d}}$:

$$\boldsymbol{\sigma} = \boldsymbol{\tau} - p\boldsymbol{I} - \boldsymbol{K}, \tag{2}$$

$$T\boldsymbol{q}_{\mathrm{s}} = \boldsymbol{q} - \dot{\varphi}\left[\nabla.\left(\frac{\partial F}{\partial \nabla \varphi}\right)\right], \tag{3}$$

$$\boldsymbol{J}_{\mathrm{d}} = -\varrho\mu\nabla\theta. \tag{4}$$

In (2)–(4), $\boldsymbol{\tau}$ represents the viscous part of the overall stress $\boldsymbol{\sigma}$, p is pressure, $\boldsymbol{K}$ is known as Korteweg stress tensor, $\boldsymbol{q}$ is the heat flux vector, μ is the motility of the liquid/vapour pair, ϱ is the mass density of the mixture, and θ is the chemical potential, which reads

$$\varrho\theta = 2\frac{\lambda}{\varepsilon^2}\left(-\varepsilon^2\Delta\varphi + (\varphi^2-1)\varphi\right) - 2\frac{\partial\varrho}{\partial\varphi}\frac{1}{\varrho}(p+F). \tag{5}$$

The superimposed dot denotes the substantial derivative, i.e. $\dot{\varphi} = \partial_t\varphi + \nabla\varphi.\boldsymbol{u}$, with $\boldsymbol{u}$ being the velocity of the two-fluid system. For a quasi-immiscible pair, as the one considered in this discussion, the Korteweg stress tensor $\boldsymbol{K}$ is usually related to the surface tension force [5]. By substituting (5) into (4), and considering the balance laws of mass, linear momentum and energy, the following system of partial differential equations is obtained

$$\nabla\cdot\boldsymbol{u} = \frac{\varrho_{\mathrm{l}}-\varrho_{\mathrm{v}}}{\varrho_{\mathrm{l}}\varrho_{\mathrm{v}}}\varrho\left(\nabla\cdot\left[-2\mu\nabla\left(\frac{\lambda}{\varepsilon^2}\psi\right)\right] + \frac{\Gamma_{\mathrm{v}}}{\varrho}\right), \tag{6}$$

$$\dot{\varphi} = -2\nabla\cdot\left[-2\mu\nabla\left(\frac{\lambda}{\varepsilon^2}\psi\right)\right] - 2\frac{\Gamma_{\mathrm{v}}}{\varrho}, \tag{7}$$

$$\psi = -\varepsilon^2\Delta\varphi + (\varphi^2-1)\varphi, \tag{8}$$

$$\varrho\dot{\boldsymbol{u}} = \nabla\cdot(\boldsymbol{\tau} - p\mathbf{I}) - \boldsymbol{F}_{\mathrm{st}}, \tag{9}$$

$$\varrho C_{\mathrm{p}}\dot{T} = \nabla\cdot(k\nabla T) + \boldsymbol{\tau} : \boldsymbol{d} + \dot{\varphi}\lambda\Delta\varphi + H_{\mathrm{lv}}\Gamma_{\mathrm{v}}. \tag{10}$$

Here, ϱ_{l} and ϱ_{v} denote, respectively, the mass densities of the liquid and vapour, Γ_{v} is the evaporation/condensation mass flow rate, which has been modeled by means of the *Knudsen* relation for the liquid/vapour phase change:

$$\Gamma_{\mathrm{v}} = C\sqrt{\frac{M}{2\pi R}}\left(\frac{p_{\mathrm{sat}}(T_{\mathrm{l}})}{\sqrt{T_{\mathrm{l}}}} - \frac{p_{\mathrm{v}}}{\sqrt{T_{\mathrm{v}}}}\right), \tag{11}$$

with M being the molar mass of the liquid and R the Universal Gas Constant. The subscripts "l" and "v" indicate, respectively, whether a given physical quantity is evaluated in the liquid or in the vapour phase. The term C is an accommodation coefficient to be adapted to the considered fluid [3, 6]. The saturation pressure p_{sat} appearing in Eq. (11), also known as vapour tension of the substance, is evaluated

punctually in the interior of the heat pipe by means of the *Clausius-Clapeyron* formula[6]

$$p_{\mathrm{sat}}(T) = p_0 \exp\left[\frac{MH_{\mathrm{lv}}}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right]. \tag{12}$$

The term $\boldsymbol{F}_{\mathrm{st}} = \nabla.\boldsymbol{K}$, under appropriate assumptions, can be written as

$$\boldsymbol{F}_{\mathrm{st}} = -\frac{\lambda}{\varepsilon^2}\psi\nabla\varphi.$$

The parameters, which refer to the two-fluid system as a mixture, i.e., the specific heat at constant pressure C_{p}, the thermal conductivity k, $\boldsymbol{u}$ and ϱ, are defined as usually done for a two-constituent mixture in the framework of Mixture Theory. The term $\boldsymbol{d} = \mathrm{sym}(\nabla\boldsymbol{u})$ is the symmetric part of the velocity gradient.

Equations (6)–(10) are considered hereafter to model the two-fluid system in the heat exchanger. Equation (6) represents the mass balance law of the liquid/vapour system as a whole. It states that the velocity flux of the system's centre of mass is not solenoidal. This differs from many other models (cf., e.g., [4, 8]), which rely on the constraint that the velocity field is divergence-free. Equation (7) represents the modified concentration balance law of one of the two phases. Equation (8) is an auxiliary equation, introduced to eliminate the fourth order derivative of the colour function φ, which would otherwise arise in the first term on the right-hand-side of Eq. (7). Finally, Eqs. (9) and (10) are, respectively, the local forms of the balance laws of linear momentum and energy of the system as a whole.

The weak form of (6)–(10) has been implemented in a commercial Finite Element Software, by modifying an already existing model, in which the non-standard terms, i.e., the ones linked to the presence of a thermal field, the phase change and the weak miscibility, are not included. For numerical purposes, in (6)–(10) the contributions of compressibility have been neglected. The parameters ε, λ and μ in Eq. (1) are defined as in [8]. The parameters ϱ_{l}, ϱ_{v}, H_{lv}, σ_{t}, C_{p}, and k, which define the nature of the chosen working fluid, are here defined as polynomial functions of the temperature. Depending on which requirements the heat exchanger has to satisfy, any working fluid can be chosen (e.g., liquid metals, water, ammonia, acetone, alcohol, nitrogen and helium) and, in principle, it can be characterized by the same set of parameters. In this contribution, the chosen working fluid is covered by a non-disclosure agreement. Finally, the geometric setting employed in the performed numerical simulations is schematically shown in Fig. 1a.

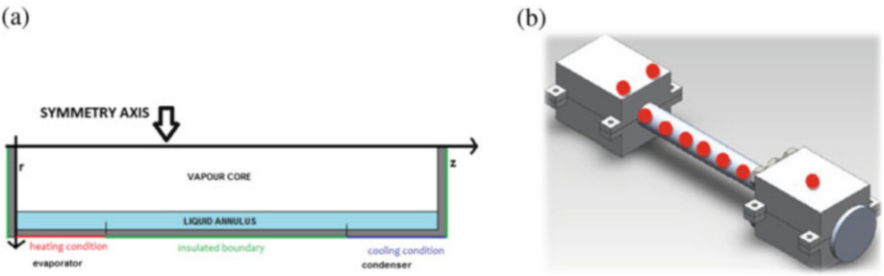


Fig. 1 (**a**) Axially symmetric geometry of the system considered in the numerical simulations. (**b**) The experimental apparatus. The *red dots* correspond to the positions of the thermocouples (Color figure online)

2 Validation and Results

The numerical model was validated by experimental data provided by Argotec s.r.l. that designed the heat pipe and the experimental apparatus and performed the test campaign. The pipe was heated at one of its two ends (where an imposed heat flux is prescribed) by means of resistors, and cooled down at the other end, which was inserted in a cooling jacket. The latter one was connected with a thermal bath via a bundle of tubes in which a refrigerating fluid was conveyed. The remaining part of the pipe was thermally insulated and, thus, referred to as "adiabatic zone".

The output data were obtained by positioning thermocouples (with a precision of 1 K) along the pipe, as shown in Fig. 1b. The temperature at each thermocouple position was measured at different instants of time until the system reached stationary working conditions. The system described above, and the experiments conduced on it, were simulated for different values of the thermal power supplied to the evaporation zone. The results obtained by the model presented in this contribution were validated by comparison with the experimental outputs referred to the system's stationary working conditions. The experimental and simulated outcomes are reported in Fig. 2, and refer to the values of temperature on the outer wall of the heat pipe, evaluated with respect to a reference temperature.

In the adiabatic zone, the relative percentage of error between the measured temperature and the simulated one is below 1 % for all the performed simulations.

After validating the model, it has to be assessed whether or not the heat pipe works in nominal conditions. This investigation is carried out by analysing the temperature, pressure and velocity fields as well as the chemical potential in different zones of the pipe.

The first result presented here (see Fig. 3a) refers to the evaluation of the discrepancies between the saturation state $(p_{\mathrm{sat}}, T_{\mathrm{sat}})$ and the output state of the system (p, T). Hereafter, $(p_{\mathrm{l}}, T_{\mathrm{l}})$ and $(p_{\mathrm{v}}, T_{\mathrm{v}})$ represent the output states taken at the bottom of the liquid film and on the symmetry axis of the numerical representation of the heat pipe, respectively. If the pair $(p_{\mathrm{v}}, T_{\mathrm{v}})$, especially in those

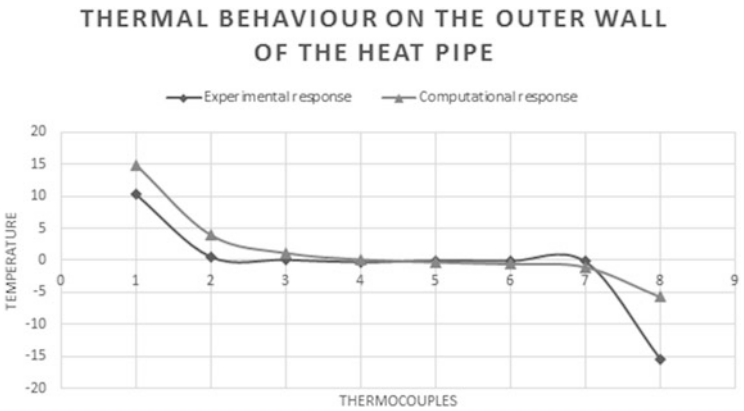


Fig. 2 Validation of the model: numerical and experimental thermal outputs on the outer wall of the pipe. *Diamonds* and *triangles* correspond to the positions of the thermocouples

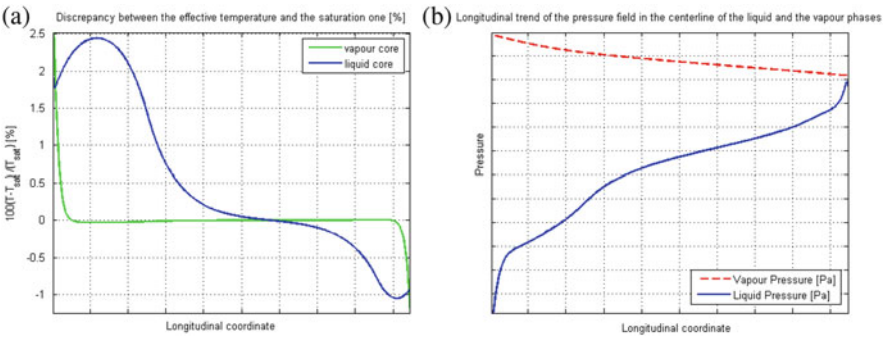


Fig. 3 (**a**) Relative discrepancy (%) between the computed saturation temperature and T_v and T_l, respectively. (**b**) Longitudinal trend of the pressure field in the vapour and in the liquid core

points corresponding to the adiabatic zone, exhibits strong deviations from the saturation state (p_{sat}, T_{sat}), then the heat pipe could fail to work on-design. Indeed, this undesired occurrence implies that H_{lv} does not match the heat that is supplied and subtracted from the outside, leading to a overheating of the device. Therefore, a good estimate of the departure of the thermodynamical state of the system from the saturation one is useful to evaluate *a posteriori* one of the possible operating limits, which is a critical supplied power, above which the heat pipe fails and, consequently, to forecast this failure. For this purpose, T_{sat} is determined with the aid of the Clausius-Clapeyron formula (12), and then compared with the output temperature T_v.

In the present case study, the absolute difference $|T_v - T_{sat}(p)|$ remains less than 1 K almost everywhere in those numerical tests that reproduce a successful experiment. For instance, in the case reported in Fig. 3a, the relative difference, expressed as a percentage, is very small and presents a peak in the two ends of

the pipe ($|T - T_{\text{sat}}|_{\max} \approx 7$ K in the evaporation zone, and $|T - T_{\text{sat}}|_{\max} \approx 3$ K in the condenser).

Another failure of the heat pipe is represented by a loss in capillary pressure of the liquid film. This occurrence can be produced by the high viscosity of the chosen fluid, or a too small capillary pressure, which instead should guarantee the liquid reflux to the evaporator. The velocity field of the liquid phase must be counter-current to the vapour flux in this device, since the former one has to reach the evaporator, whereas the latter one must flow towards the condenser. In Fig. 3b, the counter-current nature of the velocity in each phase can be deduced from the longitudinal trend of the pressure. In fact, the pressure drop is positive in the vapour and negative in the liquid. Moreover, the pressure jump $|p_{\text{l}} - p_{\text{v}}|$ rising at the evaporator should be as high as possible. This difference is a measure of the capillary pump that acts on the liquid, thereby allowing for its reflux. In the model presented in this work, however, the wick structure is absent. Nevertheless, the pressure difference $|p_{\text{l}} - p_{\text{v}}|$ gives an estimation of the capillary pressure that is sufficient to ensure the reflux of the liquid to the evaporator. This capillary action is partially ascribable to the contribution of the Korteweg stress to the overall stress tensor defined in (2).

As an outcome of the analysis of the results illustrated in Fig. 3a and b, the considered experimental protocol and the performed simulations correspond to the on-design working condition of the device.

3 Conclusions

The purpose of this work is to support a conscious industrial design of the optimal heat pipe by providing a mathematical model conceived to be rigorous and efficient, but also manageable and implementable on in-house machines.

A two-phase computational fluid dynamic model for estimating the liquid and vapour thermodynamic state inside a heat pipe is developed. For this purpose, the dissipative aspects of the studied problem have been taken into account, and the model has been elaborated in such a way that the Helmholtz free energy density of the two-fluid system depends both on the phase field and on the first gradient of this order parameter, as is the case in the Cahn-Hilliard theory. As a consequence, the Cahn-Hilliard model for phase transitions is generalised to the non-isothermal case. As obtained from the comparison of the numerical outcomes with the experimental data, provided by Argotec s.r.l., the developed numerical model describes effectively some features of the complicated behaviour of a heat pipe.

After validating the model, particular attention is paid to checking the temperature and pressure fields of the liquid and the vapour phases. All the obtained results can be also used for evaluating a posteriori the operating limits of the heat pipe. Indeed, it seems important to comprehend with an adequate amount of confidence in which way off-design conditions may arise. A useful application of the present

model could be put into practice, to refine the design and the industrial production of this particular heat exchanger.

Further characterisations and particularisations of the present mathematical model are the object of current investigations, which aims at suiting the present model to the demands of the company, which busily cooperates in the model development by providing new ideas, new test cases, and new related experimental data.

References

1. Anderson, D.M., McFadden, G.B.: Diffuse-interface methods in fluid mechanics. Annu. Rev. Fluid Mech. **30**, 139–165 (1998)
2. Chi, S.W.: Heat Pipe Theory and Practice. Hemisphere Publishing Corporation, Washington (1976)
3. Hall, M.L., Doster, J.M.: A sensitivity study of the effects of evaporation/condensation accommodation coefficients on transient heat pipe modeling. Int J. Heat Mass Transf. **33**(3), 465–481 (1990)
4. Jamet, D.: Diffuse interface models in fluid mechanics (2014). http://pmc.polytechnique.fr/mp/GDR/docu/Jamet.pdf
5. Kim, J.: A continuous surface tension force formulation for diffuse-interface models. J. Comput. Phys. **204**, 784–804 (2005)
6. Lips, S., Bonjour, J., Lefèvre, F.: Investigation of evaporation and condensation process specific to grooved flat heat pipes. Front. Heat Pipes (FHP) **1**, 023001 (2010)
7. Lowengrub, J., Truskinovsky, L.: Quasi-incompressible Cahn-Hilliard fluids and topological transitions. Proc. R. Soc. Lond. **454**, 2617–2654 (1998)
8. Yue, P., Feng, J.J., C. Liu, C., Shen, J.: A diffuse-interface method for simulating two-phase flows of complex fluids. J. Fluid Mech. **515**, 293–317 (2004)

MS 20
MINISYMPOSIUM: NATURE'S NATURAL ORDER: FROM INDIVIDUAL TO COLLECTIVE BEHAVIOUR AND SELF-ORGANIZATION

Organizers

Hermes Gadelha[1] and Philip Maini[2]

Speakers

George Chamoun[3], Mazen Saad[4], Raafat Talhouk[5]
Mathematical and Numerical Analysis of a Coupled Anisotropic Chemitaxis-Fluid Model

Hermes Gadelha[1]
Mechanoregulation of Molecular Motors in Flagella

Marco Polin[6], Douglas Brumley[7], Kirsty Y. Wan[8], Raymond E. Goldstein[9]
Flagellar Synchronization Through Direct Hydrodynamic Interactions

[1]Hermes Gadelha, University of Oxford, United Kingdom.

[2]Philip Maini, University of Oxford, United Kingdom.

[3]Georges Chamoun, Laboratoire de Mathematiques Jean Leray, Nantes, France.

[4]Mazen Saad, Laboratoire de Mathematiques Jean Leray, Nantes, France.

[5]Raafat Talhouk, LAboratorie de MAthematiques, Hadath, Beyrouth, Liban.

[6]Marco Polin, University of Warwick, Coventry, United Kingdom; University of Cambridge, Cambridge, United Kingdom.

[7]Douglas Brumley, University of Cambridge, United Kingdom; Massachusetts Institute of Technology, United States.

[8]Kirsty Y. Wan, University of Cambridge, Cambridge, United Kingdom.

[9]Raymond E. Goldstein, University of Cambridge, Cambridge, United Kingdom.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_69

Ulrich Dobramysl[10], Radek Erban[11]
Stochastic Multi-Scale Modeling of Filopodial Growth

Keywords

Collective behaviour
Mathematical biology
Self-organization

Short Description

Ordered randomness is ubiquitous in nature. Every single biological system, from molecular motor dynamics in intracellular processes, multicellular organisms to human social behaviour, from micro- to macro-scale, is susceptible to arbitrary events, which did not have to happen in a certain way, but which did, and completely defined the final system. Interestingly, despite the inherent complexity of many interacting individual agents, order and structure may arise naturally, producing robust global behaviours with emergent properties that qualitatively differ from those of its individual units. In this mini-symposium, we will explore how mathematical modelling can be utilized to predict and interpret the different facets of self-organization and collective behaviour in biology, establishing what is known and identifying further challenges. The proposed mini-symposium will additionally highlight that this is a fertile and challenging area of inter-disciplinary research for applied mathematicians, while demonstrating the importance of future observational and theoretical studies in understanding the underlying mechanisms of self-organization and collective behaviour.

[10]Ulrich Dobramysl, University of Cambridge, United Kingdom.
[11]Radek Erban, University of Oxford, United Kingdom.

Convergence Analysis and Numerical Simulations of Anisotropic Keller-Segel-Fluid Models

Georges Chamoun, Mazen Saad, and Raafat Talhouk

Abstract In order to study the dynamics of anisotropic chemotaxis-fluid models, a detailed numerical analysis is established in this paper. To discretize this type of models, a monotone combined scheme is proposed as a compromise between the nonconforming finite elements, enabling in particular the use of general meshes and the discretization of anisotropic diffusion tensors, and between the finite volumes enabling to avoid spurious oscillations in the convection-dominated regime. Moreover, this monotone scheme ensures the discrete maximum principle and therefore the confinement of the density of cells and the positivity of the chemical concentration. Finally, a test is given to illustrate the numerical study.

Keywords Anisotropic chemotaxis-fluids • Anisotropic Keller-Segel-fluid models

1 Introduction

Chemotaxis, movement toward or away from chemicals, is a universal attribute of motile cells and organisms. In a large variety of ecological systems, cells often live in a viscous fluid so that cells and chemical substrates are also transported with the

G. Chamoun (✉)
Ecole Centrale de Nantes, Laboratoire de Mathématiques Jean Leray, UMR CNRS 6629, 1 rue de la Noé, 44321 Nantes, France

Université Libanaise, EDST et Faculté des sciences, Laboratoire de Mathématiques, Hadath, Lebanon
e-mail: georges.chamoun@ec-nantes.fr

M. Saad
Ecole Centrale de Nantes, Laboratoire de Mathématiques Jean Leray, UMR CNRS 6629, 1 rue de la Noé, 44321 Nantes, France
e-mail: mazen.saad@ec-nantes.fr

R. Talhouk
Université Libanaise, EDST et Faculté des sciences, Laboratoire de Mathématiques, Hadath, Lebanon
e-mail: rtalhouk@ul.edu.lb

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_70

fluid, and meanwhile the motion of the fluid is under the influence of gravitational forcing generated by aggregation of cells. For example, E. coli cells often swim towards amino acids and sugars. Nevertheless, the mathematical modelling of cell movement which goes back to Patlak [12], (see [10]) do neglect the surrounding fluid and would fail to predict the dynamics of cells influenced by the fluid. Thus, it is interesting and important in biology to study some phenomenon of chemotaxis on the basis of the coupled cell-fluid model. This paper is devoted to the numerical analysis of the following system of Keller-Segel equations coupled to Stokes equations,

$$\begin{cases} \partial_t N - \nabla \cdot (S(x)a(N)\nabla N) + \nabla \cdot (S(x)\chi(N)\nabla C) + u \cdot \nabla N = 0, \\ \partial_t C - \nabla \cdot (M(x)\nabla C) + u \cdot \nabla C = \alpha N - \beta C, \\ \partial_t u - \nu \Delta u + \nabla P = -N\nabla\phi, \\ \nabla \cdot u = 0, \end{cases} \tag{1}$$

where Ω is an open bounded domain in $\mathbb{R}^d$ $(d = 2, 3)$ with smooth boundary $\partial\Omega$. The system is supplemented by the following boundary conditions on $\partial\Omega \times (0, T)$,

$$S(x)a(N)\nabla N \cdot \eta = 0,\ M(x)\nabla C \cdot \eta = 0,\ u = 0, \tag{2}$$

where η is the exterior unit normal to $\partial\Omega$. The initial conditions on Ω are given by,

$$N(x, 0) = N_0(x),\ C(x, 0) = C_0(x),\ u(x, 0) = u_0(x). \tag{3}$$

The fluid flow is governed by the incompressible Stokes equations with velocity field u, pressure P and viscosity ν. Both, the chemical concentration and the density of cells are denoted by C and N, respectively. Anisotropic and heterogeneous tensors are denoted by $S(x)$ and $M(x)$. The function $\chi(N)$ is usually written in the form $\chi(N) = N\tilde{h}(N)$ where $\tilde{h}$ is commonly referred to as the chemotactic sensitivity function. The cross diffusion term namely : $\nabla \cdot (S(x)\chi(N)\nabla C)$ can be view as a convective term of the flux $\chi(N)$ according to the sign of gradient of ∇C. Furthermore, if $\chi(N)$ is positive then we observe the chemoattractant mechanism and in the other case we observe the chemorepellent mechanism. Moreover, the density-dependent diffusion coefficient is denoted by $a(N)$.

The function $h(N, C) = \alpha N - \beta C$ describes the rates of production and degradation of the chemical signal (chemoattractant). It can be seen in the model (1) that the fluid is coupled to the chemotaxis equations through both the transport of cells and chemical substrates $u \cdot \nabla N$, $u \cdot \nabla C$ and the external gravitational force $g = -N\nabla\phi$ exerted by a cell onto the fluid along the upwards unit vector z. The question of global existence of weak solutions of the degenerate model (1) has been answered in [2, 4] and hence it is well-posed. Interested by experiments

able to predict the dynamics of cells influenced by a viscous fluid through transport and gravitational force and motivated by results described in [3, 5] which explain the dynamics of anisotropic chemotaxis models in a fluid at rest ($u = 0$), we extend in this paper the numerical analysis of Chamoun et al. [5] to the model (1). To our knowledge, there are only a few numerical results given for related systems (see [6, 11]). For example, the finite element method has been used to illustrate the behavior of the elliptic-parabolic Keller-Segel-Stokes system with numerical examples in [11].

2 Setting of the Problem

We assume first that $\chi(0) = 0$ and the chemotactical sensitivity $\chi(N)$ vanishes when $N \geq 1$. This threshold condition has a clear biological interpretation called the volume-filling effect (see [9]). The main assumptions are:

$$\chi : [0,1] \longmapsto \mathbb{R} \text{ is continuous and } \chi(0) = \chi(1) = 0\,, \tag{4}$$

$$\begin{aligned} a, f : [0,1] \longmapsto \mathbb{R}^+ \text{ are continuous, } a(0) = a(1) \\ = f(0) = 0 \text{ and } a(s) > 0 \text{ for } 0 < s < 1\,. \end{aligned} \tag{5}$$

Next, we require

$$\nabla\phi \in (L^\infty(\Omega))^d \text{ and } \phi \text{ is independent of time.} \tag{6}$$

The permeabilities S, M: $\Omega \longrightarrow \mathscr{M}_d(\mathbb{R})$ where $\mathscr{M}_d(\mathbb{R})$ is the set of symmetric matrices $d \times d$, verify:

$$S_{i,j} \in L^\infty(\Omega),\ M_{i,j} \in L^\infty(\Omega),\ \forall i,j \in \{1,..,d\}\,, \tag{7}$$

and there exist $c_S \in \mathbb{R}_+^*$ and $c_M \in \mathbb{R}_+^*$ such that a.e $x \in \Omega$, $\forall \xi \in \mathbb{R}^d$,

$$S(x)\xi \cdot \xi \geq c_S|\xi|^2,\ M(x)\xi \cdot \xi \geq c_M|\xi|^2\,. \tag{8}$$

Finally, we introduce basic spaces associated to the Stokes equation,

$$\wp = \{u \in \mathscr{D}(\Omega), \nabla \cdot u = 0\},\ V = \bar{\wp}^{H_0^1(\Omega)} \text{ and } H = \bar{\wp}^{L^2(\Omega)}\,, \tag{9}$$

where V and H are the closure of $\wp$ in $H_0^1(\Omega)$ and $L^2(\Omega)$ respectively.

3 Combined Finite Volume-Nonconforming Finite Element Scheme

This section is devoted to the formulation of a combined scheme for the anisotropic chemotaxis-fluid model (1) which has been recently proposed and studied for parabolic equations in [8] and for anisotropic Keller-Segel models in [5].

3.1 Space and Time Discretization of Ω

We consider a family $\mathscr{T}_h$ of meshes of the domain Ω, consisting of disjoint closed simplices. The size of the mesh $\mathscr{T}_h$ is defined by $h{:=}\max_{K\in\mathscr{T}_h} diam(K)$. We also make the following shape regularity assumption: $\exists k_{\mathscr{T}} > 0$ such that $\min_{K\in\mathscr{T}_h} \frac{|K|}{(diam(K))^d} \geq k_{\mathscr{T}}$, $\forall h > 0$. We also use a dual partition $\mathscr{D}_h$ of disjoint closed simplices called control volumes of Ω such that $\bar{\Omega} = \cup_{D\in\mathscr{D}_h}\bar{D}$. There is one dual element D associated with each side $\sigma_D = \sigma_{K,L} \in \mathscr{E}_h$. We construct it by connecting the barycenters of every $K \in \mathscr{T}_h$ that contains σ_D through the vertices of σ_D. The point P_D is referred to as the barycenter of the side σ_D. For all $D \in \mathscr{D}_h$, denote by $|D|$ the measure of D, by $\mathscr{N}(D)$ the set of neighbors of the volume D, by $\sigma_{D,E}$ the interface between a dual volume D and E and by $\eta_{D,E}$ the unit normal vector to $\sigma_{D,E}$ outward to D (see Fig. 1).

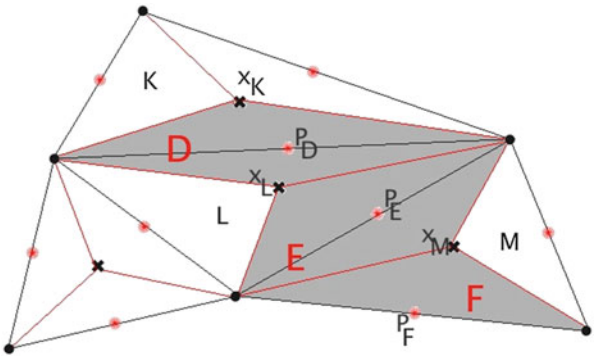


Fig. 1 Dual volumes associated with edges of the primal mesh

Next, we define the following finite-dimensional spaces:

$$X_h := \{\varphi_h \in L^2(\Omega);\ \varphi_h|_K \text{ is linear } \forall K \in \mathscr{T}_h,\ \varphi_h \text{ is continuous at the points } P_D,\ D \in \mathscr{D}_h^{int}\}, \tag{10}$$

$$X_h^0 := \{\varphi_h \in X_h;\ \varphi_h(P_D) = 0,\ \forall D \in \mathscr{D}_h^{ext}\}.$$

The basis of X_h is spanned by the shape functions φ_D, $D \in \mathscr{D}_h$, such that $\varphi_D(P_E) = \delta_{DE}$, $E \in \mathscr{D}_h$, δ being the Kronecker delta. We equip X_h^0 with the scalar product $((N_h, V_h))_h = \sum_{K\in\mathscr{T}_h} \int_K \nabla N_h \cdot \nabla V_h dx$ and the seminorm $||N_h||_{X_h}^2 := \sum_{K\in\mathscr{T}_h} \int_K |\nabla N_h|^2 dx$ which becomes a norm on X_h^0.

Let us consider a constant time step $\Delta t \in [0, T]$. A discretization of $[0, T]$ is given by $\tilde{N} \in \mathbb{N}^*$ such that $t_n = n\Delta t$, for $n \in \{0, \ldots., \tilde{N} + 1\}$. The discrete unknowns are denoted by $\left\{w_D^n,\ D \in \mathscr{D}_h,\ n \in \{0, \ldots, , \tilde{N} + 1\}\right\}$ where $w = N,\ C$ or u.

3.2 Combined Scheme for the System (1)

Due to the incompressibility condition $\nabla \cdot u = 0$, it was shown in [7, 13] that it is not possible to approximate the space V defined in (9) by the most simple finite elements where the results, even for the Stokes equations, are less general and vary according to the dimension since no basis of the approximate space V_h is available. For this reason, the approximation by means of nonconforming finite element methods chosen in this subsection is certainly very useful for Stokes problems.

Let us denote the approximation of the flux $S(x)\nabla C \cdot \eta_{D,E}$ (resp. $u \cdot \eta_{D,E}$) on the interface $\sigma_{D,E}$ by $\delta C_{D,E}$ (resp. $u_{D,E}$). Then, we approximate the numerical flux $S(x)\chi(N)\nabla C \cdot \eta_{D,E}$ by means of the values N_D, N_E and $\delta C_{D,E}$ through a numerical flux function $G(N_D, N_E, \delta C_{D,E})$ satisfying classical properties given in [5]. Similarly, for $\chi(N) = N$ and $S(x) = Id$, we approximate the flux $Nu \cdot \eta_{D,E}$ as an upwind convection function $G_1(N_D, N_E, u_{D,E}) = u_{D,E}^+ N_D - u_{D,E}^- N_E$, where $u_{D,E}^+$ and $u_{D,E}^-$ denote the positive and negative parts of $u_{D,E}$. Moreover, for all $N_h = \sum_{D\in\mathscr{D}_h} N_D \varphi_D \in X_h$, we define a discrete function of $A(N_h)$ as $A_h(N_h) = \sum_{D\in\mathscr{D}_h} A(N_D)\varphi_D$.

Finally, a combined finite volume-nonconforming finite element scheme for the discretization of the model (1) is given by the following iterative algorithm: Suppose that the solution $(\tilde{N}_h^n, \tilde{C}_h^n, u_h^n, p_h^n)$ at time t_n is known, then we compute the solution $(\tilde{N}_h^{n+1}, \tilde{C}_h^{n+1}, u_h^{n+1}, p_h^{n+1})$ at time t_{n+1} through two steps.

- **First step:** (Computation of u_h^{n+1} and p_h^{n+1})
  Let V_h be a subspace of the preceding space X_h such that

$$V_h = \{u_h \in X_h,\ div_h(u_h) = 0\} \text{ where the discrete divergence,} \tag{11}$$

$$div_h(u_h) = \sum_{K \in \mathcal{T}_h} \eta_K 1_K;\ \eta_K = \frac{1}{|K|} \int_K \nabla \cdot u_h \, dx\,.$$

The practical computation of u_h^{n+1} in V_h is not easy (see [13]). For that, we interpret our problem as a variational problem in X_h with linear constraints. We define the space of piecewise constant functions

$$Y_h = \{p_h = \sum_{K \in \mathcal{T}_h} \eta_K 1_K \, dx;\ 1_K \text{ is the characteristic function of } K\}$$

and we compute practical solutions $u_h^{n+1} \in X_h$ and $p_h^{n+1} \in Y_h$ using the classical Uzawa's algorithm, as the limits of two sequences of elements

$$u_h^{n+1,r} \in X_h \text{ and } p_h^{n+1,r} \in Y_h,\ r = 0, 1, .., +\infty\,.$$

We start the algorithm with an arbitrary element $p_h^{n+1,0}$. When $u_h^{n+1,r}$ is known, we define $u_h^{n+1,r+1}$ and $p_h^{n+1,r+1}$ by

$$\frac{1}{\Delta t}(u_h^{n+1,r+1} - u_h^n, v_h) + \nu((u_h^{n+1,r+1}, v_h))_h - (p_h^{n,r+1}, div_h v_h) = (g^n, v_h),\ \forall v_h \in X_h\,, \tag{12}$$

where $g^n = -\tilde{N}_h^n \nabla\phi \in L^2(\Omega)$.

$$(p_h^{n+1,r+1} - p_h^{n,r}, q_h) + \rho\big(div_h(u_h^{n+1,r+1}), q_h\big) = 0,\ \forall q_h \in Y_h\,. \tag{13}$$

The existence and the uniqueness of the solution $u_h^{n+1,r+1}$ follow from the projection theorem. Regarding the convergence of this algorithm, we have the following Proposition proved in [13, Chap. VII, Proposition 6.7],

Proposition 1 *If* $0 < \rho < \frac{2\nu}{d}$ *then, as* $r \to +\infty$, $u_h^{n+1,r+1}$ *converges to* u_h^{n+1} *in* X_h *and* $p_h^{n+1,r+1}$ *converges to* p_h^{n+1} *in* $Y_h/\mathbb{R}$.

- **Second step:** Given u_h^{n+1} from the first step, we compute $\tilde{N}_h^{n+1}$ and $\tilde{C}_h^{n+1}$ by:

$$|D|\frac{N_D^{n+1}-N_D^n}{\Delta t}-\sum_{E\in\mathscr{D}_h}\mathscr{S}_{D,E}A(N_E^{n+1})+\sum_{E\in\mathscr{N}(D)}G(N_D^{n+1},N_E^{n+1};\delta C_{D,E}^{n+1}) \tag{14}$$

$$+\sum_{E\in\mathscr{N}(D)}G_1(N_D^{n+1},N_E^{n+1};u_{D,E}^{n+1})=0\,,$$

$$|D|\frac{C_D^{n+1}-C_D^n}{\Delta t}-\sum_{E\in\mathscr{D}_h}\mathscr{M}_{D,E}C_E^{n+1}+\sum_{E\in\mathscr{N}(D)}G_1(C_D^{n+1},C_E^{n+1};u_{D,E}^{n+1})=h(N_D^n,C_D^{n+1})\,. \tag{15}$$

The diffusion matrix $\mathscr{S}$ (resp. $\mathscr{M}$) of elements $\mathscr{S}_{D,E}$ (resp. $\mathscr{M}_{D,E}$) for $D,E\in\mathscr{D}_h$ is the stiffness matrix of the nonconforming finite element method. So that,

$$\mathscr{S}_{D,E}=-\sum_{K\in\mathscr{T}_h}(\mathscr{S}(x)\nabla\varphi_E,\nabla\varphi_D)_{0,K}\text{ and }\mathscr{M}_{D,E}=-\sum_{K\in\mathscr{T}_h}(M(x)\nabla\varphi_E,\nabla\varphi_D)_{0,K}\,.$$

$$\text{Otherwise, }\delta C_{D,E}^{n+1}=S_{D,E}\big(C_E^{n+1}-C_D^{n+1}\big)\text{ and }u_{D,E}^{n+1}=\int_{\sigma_{D,E}}u_h^{n+1}\cdot\eta_{D,E}\,d\gamma\,.$$

Now, we state a convergence result of the combined scheme under the assumption that all transmissibilities coefficients are positive:

$$\mathscr{S}_{D,E}\geq 0\text{ and }\mathscr{M}_{D,E}\geq 0,\ \forall D\in\mathscr{D}_h,\ E\in\mathscr{N}(D)\,. \tag{16}$$

Theorem 1 (Convergence of the Combined Scheme) *Assume* (4)...(9). *Consider* $0\leq N_0\leq 1$, $C_0\geq 0$, $u_0\in L^\infty(\Omega)$ *and* $\nabla\cdot u_0=0$. *Under the assumption* (16), *one has:*

1) There exists a solution $(\tilde{N}_{h,\Delta t},\tilde{C}_{h,\Delta t})$ *of the discrete system* (14) *and* (15).
2) Any sequence $(h_m)_m$ *decreasing to zero possesses a subsequence such that* $(N_{h_m},C_{h_m},u_{h_m})$ *converges a.e. on* Q_T *to a solution* (N,C,u) *of the system* (1).

Remark 1 If assumption (16) is not satisfied, one can use a nonlinear technique inspired from [1] to correct the diffusive flux blocking the discrete maximum principle and to maintain the monotonicity and the convergence of the corrected numerical scheme. One can see [5, Sect. 4] for more details.

4 Numerical Experiment

The driven cavity flow is one of the most studied fluid problems in computational fluid dynamics field. The fluid in this problem is contained in a square domain with Dirichlet boundary conditions on all sides, with three stationary sides and one upper moving side (with velocity $(1, 0)$ tangent to the side). In this section, we present a numerical test to show the dynamics of solutions of the system (1) in a driven cavity flow discretized by the combined method along the algorithm detailed in the Sect. 3.2 with $a(N) = N(1-N)$, $\chi(N) = cN(1-N)^2$ and $c = 0.1$. First, we choose $dt = 0.0005$, $\alpha = 0.01$, $\beta = 0.05$, $D = 0.001$, $d = 2 \times 10^{-4}$ (diffusion coefficient of C), $\nu = 5 \times 10^{-3}$, $c_2 = 0$, $c_1 = 1$ and $\nabla\phi = (0, 1)$. Next, we consider the tensors: $S = [8, -7; -7, 20]$ and $M = Id$. Simulations of this test are done on the mesh given in Fig. 2a and initial conditions are defined by regions in Fig. 2b. We see in Fig. 3 the anisotropic chemotaxis attitude of cells transported at the same time by the fluid. We can also remark that the cells split into several parts due to the velocity of the fluid which accelerates a part of cells towards the chemoattractant.

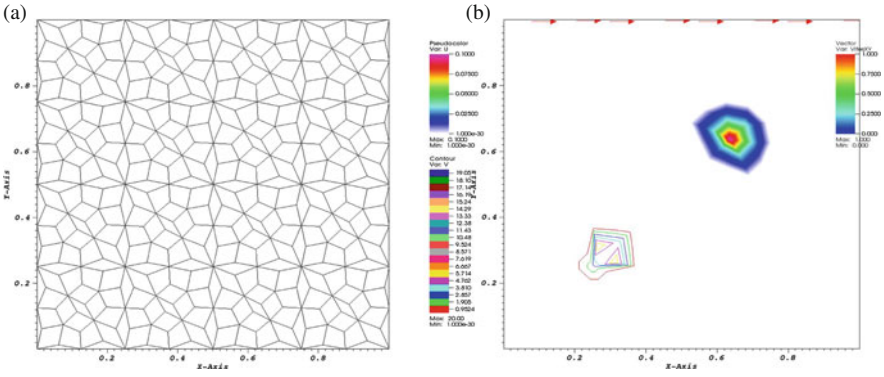


Fig. 2 The initial density is defined by $N_0(x, y) = 0.1$ in the square $(x, y) \in \{0.6, 0.7\} \times \{0.6, 0.7\}$ and 0 otherwise. The initial concentration of chemo-attractant is defined by $C_0(x, y) = 20$ in the square $(x, y) \in \{0.25, 0.35\} \times \{0.25, 0.35\}$ and 0 otherwise. A constant speed (1,0) is imposed on the upper wall of the space domain and the pressure is neglected in the whole domain. (**a**) Dual mesh of the space domain (352 diamonds). (**b**) Initial conditions

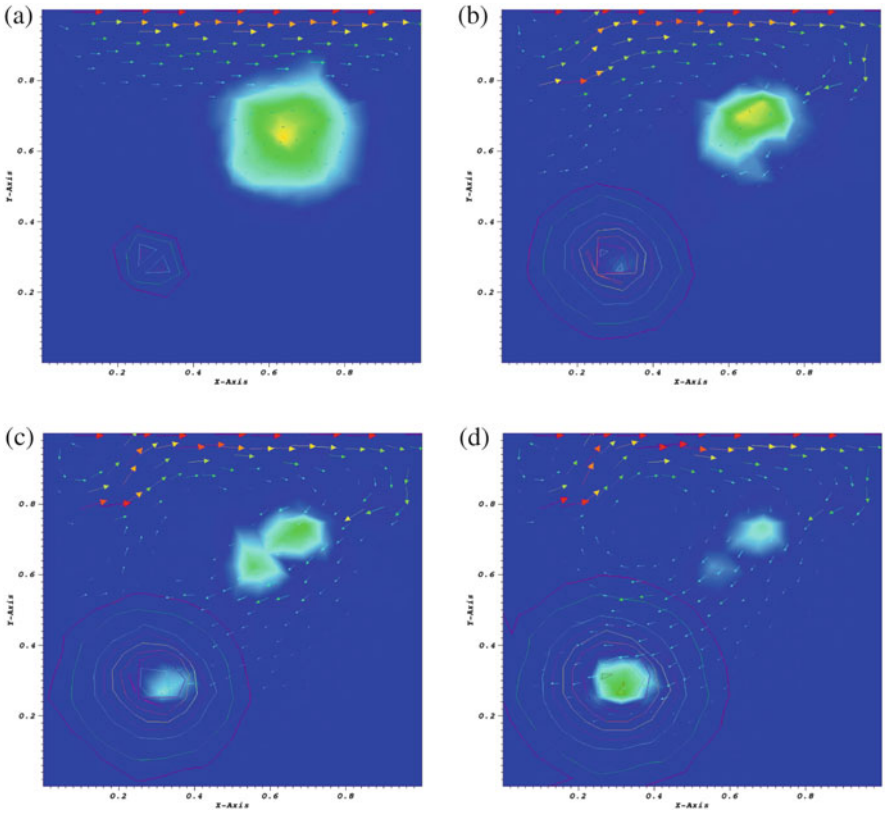


Fig. 3 Evolution in time of the cell density via a chemo-attractant in a fluid. (**a**) $0 \leq N(t = 4) \leq 0.01$, $0 \leq C(t = 4) \leq 8.35$. (**b**) $0 \leq N(t = 25) \leq 0.028$, $0 \leq C(t = 25) \leq 0.46$. (**c**) $0 \leq N(t = 40) \leq 0.03852$, $0 \leq C(t = 40) \leq 0.1852$. (**d**) $0 \leq N(t = 50) \leq 0.04591$, $0 \leq C(t = 50) \leq 0.06134$

Acknowledgements The authors would like to thank the National Council for Scientific Research (Lebanon), the Ecole Centrale de Nantes and the Lebanese University for their support to this work.

References

1. Cancès, C., Cathala, M., Le Poitier, C.: Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations. Numer. Meth. **125**, 387–417 (2013)
2. Cavalli, F., et al.: 3D simulations of early blood vessel formation. Comput. Phys. **225**, 2283–2300 (2007)
3. Chamoun, G., Saad, M., Talhouk, R.: Mathematical and numerical analysis of a modified Keller-Segel model with general diffusive tensors. J. Biomath. **2**, 1312273 (2013). http://dx.doi.org/10.11145/j.biomath.2013.12.07

4. Chamoun, G., Saad, M., Talhouk, R.: A coupled anisotropic chemotaxis-fluid model: the case of two-sidedly degenerate diffusion. Comput. Math. Appl. **68**, 1052–1070 (2013). http://dx.doi.org/10.1016/j.camwa.2014.04.010
5. Chamoun, G., Saad, M., Talhouk, R.: Monotone combined edge finite volume-non conforming finite element for anisotropic Keller-Segel model. NMPDE **30**(3), 1030–1065 (2014)
6. Chertok, A., Fellner, K., Kurganov, A., Lorz, A., Markowich, P.A.: Sinking, merging and stationary plumes in a coupled chemotaxis-fluid model: a high-resolution numerical approach. Fluid Mech. **694**, 155–190 (2012)
7. Crouzeix, M., Raviart, P.-A.: Conforming and nonconforming finite element methods for solving stationary Stokes equations I. Revue française d'automatique, informatique, recherche opérationnelle **7**(3), 33–75 (1973)
8. Eymard, R., Hilhorst, D., Vohralík, M.: A combined finite volume-nonconforming/mixed hybrid finite element scheme for degenerate parabolic problems. Numer. Math. **105**, 73–131 (2006)
9. Hillen, T., Painter, K.: Volume filling effect and quorum-sensing in models for chemosensitive movement. Can. Appl. Math. **10**, 501–543 (2002)
10. Keller, E.F., Segel, L.A.: Initiation of slime mold aggregation viewed as an instability. J. Theor. Biol. **26**, 399–415 (1970)
11. Lorz, A.: Coupled Keller-Segel Stokes model: global existence for small initial data and blow-up delay. Commun. Math. Sci. **10**, 555–574 (2012)
12. Patlak, C.S.: Random walk with persistence and external bias. Bull. Math. Biophys. **15**(3), 311–338 (1953). http://dx.doi.org/10.1007/BF02476407
13. Temam, R.: Navier-Stokes Equations. AMS Chelsea edition, Providence, RI (2000)

MS 21
MINISYMPOSIUM: MATHEMATICAL AND NUMERICAL MODELLING OF THE CARDIOVASCULAR SYSTEM

Organizers

Piero Colli Franzone[1], Luca F. Pavarino[2] and Simone Scacchi[3]

Speakers

Toni Lassila[4]
Computational Simulation of Heart Function with an Orthotropic Active Strain Model of Electromechanics

Martin Weiser[5]
Spectral Deferred Correction Methods for Adaptive Electro-Mechanical Coupling in Cardiac Simulation

Dorian Krause[6]
A Lightweight Approach to Parallel Adaptivity: Design, Implementation and Application in Electrophysiology

[1]Piero Colli Franzone, Universitá degli Studi di Pavia, Pavia, Italy.
[2]Luca F. Pavarino, Universitá degli Studi di Milano, Milano, Italy.
[3]Simone Scacchi, Universitá degli Studi di Milano, Milano, Italy.
[4]Toni Lassila, EPFL, Lausanne, Switzerland.
[5]Martin Weiser, ZIB, Berlin, Germany.
[6]Rolf Krause, Universitá della Svizzera Italiana, Lugano, Switzerland.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_71

Daniele Boffi[7]
Advances in the Mathematical Theory of the Finite Element Immersed Boundary Method

Joakim Sundnes[8]
Computational Models of Electro-Mechanical Interactions in the Heart

Paola Causin[9]
Impact of Blood Flow in Ocular Pathologies: Can Mathematical and Numerical Modeling Help Preventing Blindness?

Christian Vergara[10]
Inexact Schemes for the Fluid-Structure Interaction with Application to the Ascending Aorta

Naima Aissa[11]
Global Existence of Weak Solutions to an Angiogenesis Model

Keywords

Angiogenesis model
Cardiac simulation
Cardiovascular system
Computational cardiology
Electrophysiology

Short Description

The numerical simulation of the cardiac fluidomechanical and electrical activity is a very challenging task. Different multiscale and nonlinear effects such as the propagation of the electrical activation front, the chemical reactions within the ion channels, the orthotropic fiber architecture, the constitutive laws and active tension of the cardiac tissue have to be properly taken into account. Furthermore, the numerical methods required to simulate efficiently such complex models needs to be carefully devised and adapted to properly couple/decouple the different submodels.

[7]Daniele Boffi, Universitá degli Studi di Pavia, Pavia, Italy.
[8]Joakim Sundnes, SIMULA, Oslo, Norway.
[9]Paola Causin, Universitá degli Studi di Pavia, Pavia, Italy.
[10]Christian Vergara, MOX, Politecnico di Milano, Milano, Italy.
[11]Naima Aissa, Ecole Politechnique, Paris, France.

This minisymposium aims at bringing together researchers in computational cardiology, focusing on the latest developments and on the new research pathways and applications.

ECMI motivation/relevance: In silico studies of the cardiovascular system are very relevant to ECMI since they are a crucial part of the ongoing efforts to bridge advanced research and clinical applications, as e.g. in the Virtual Physiological Human (VPH) project. The main goal of these studies is to develop, test and implement integrative biomedical science and technology-facilitated applications, as well as to improve current simulation techniques.

On a Spatial Epidemic Propagation Model

István Faragó and Róbert Horváth

Abstract Most of the models of epidemic propagations do not take into the account the spatial distribution of the individuals. They give only the temporal change of the number of the infected, susceptible and recovered patients. In our presentation we present a spatial epidemic propagation model and give some of its qualitative properties both in the continuous and the finite difference numerical case: boundedness, nonnegativity preservation, the condition of forming epidemic waves. Some of the results are demonstrated on numerical tests.

Keywords Biomedical science • Spatial epidemic propagation model

1 Introduction

Most of the living populations have to cope with different diseases. Some of these diseases are communicable and can decrease the size of the population dramatically. This is why people are eager to understand the mechanism of epidemics and try to prevent their outbreak and propagation by efficient and affordable means (e.g. hygiene, vaccination).

One of the tools of the investigation of epidemics may be the construction of mathematical models and the analysis of the solutions of these models [1–3]. In 1927, Kermack and McKendrick [4] created an epidemic model (also known as SIR

I. Faragó (✉)
Department of Applied Analysis and Computational Mathematics, Eötvös Loránd University, Pázmány Péter stny. 1/C, 1117 Budapest, Hungary

MTA-ELTE NumNet Research Group, Budapest, Hungary
e-mail: faragois@cs.elte.hu

R. Horváth
Department of Analysis, Budapest University of Technology and Economics, Egry J. u. 1., 1111 Budapest, Hungary

MTA-ELTE NumNet Research Group, Budapest, Hungary
e-mail: rhorvath@math.bme.hu

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_72

model) in the form of a system of ordinary differential equations

$$\begin{aligned} S' &= -aSI, \\ I' &= aSI - bI, \\ R' &= bI, \end{aligned} \tag{1}$$

where $I = I(t)$, $S = S(t)$ and $R = R(t)$ denote the number of infective, susceptible and removed (by immunity or death) individuals as a function of time t, respectively. The contact rate a and recovery coefficient b are positive known numbers. This model has been improved several times taking into the account also births, deaths, latent periods, reinfections, incubations etc. [1, 2]. These models assume that the population is homogeneous, that is do not handle the different spatial positions of the individuals. There are several methods to bring also spatial dependence into the picture. For example, it is possible to investigate subpopulations inside the original population that are connected somehow into a network. Other possibility is to allow the motion of the individuals in the population [3]. We will consider a third model. We assume that the speed of the motion of the individuals can be neglected compared to the speed of the disease and the infection is localized in that sense that a member of the population can infect only members in its well defined neighbourhood. This property is brought into the model by integral coefficients.

Based on the above considerations, we arrive at a modified SIR model (see e.g. [3]) in the form of a system of partial differential equations equipped with suitable initial and boundary conditions

$$\begin{aligned} S'_t(x,t) &= -\left(\int_{N(x)} W(|x'-x|)I(x',t)\,\mathrm{d}x'\right) S(x,t), \\ I'_t(x,t) &= \left(\int_{N(x)} W(|x'-x|)I(x',t)\,\mathrm{d}x'\right) S(x,t) - bI(x,t), \\ R'_t(x,t) &= bI(x,t), \end{aligned} \tag{2}$$

where now $S = S(x,t)$, $I = I(x,t)$ and $R = R(x,t)$ depend also on the spatial position and give the densities of the corresponding parts of the population. The nonnegative weighting function W is supposed to depend only on the distance of the points x' and x, and $N(x)$ denotes a prescribed neighbourhood of the point x.

The model (2) can be simplified further. Let us suppose that the spatial dimension of the problem is one, and that $N(x) = [x-\delta, x+\delta]$ is a symmetric interval around any fixed point x. Let us approximate I with its second order spatial Taylor series. In this way we arrive at the system [3]

$$\begin{aligned} S'_t &= -S\left(\theta I + \phi I''_{xx}\right), \\ I'_t &= S\left(\theta I + \phi I''_{xx}\right) - bI, \\ R'_t &= bI, \end{aligned} \tag{3}$$

where

$$\theta = \int_{-\delta}^{\delta} W(|u|)\,\mathrm{d}u, \quad \phi = \frac{1}{2}\int_{-\delta}^{\delta} u^2 W(|u|)\,\mathrm{d}u \tag{4}$$

are positive constants that can be computed from the model (namely from $N(x)$ and W) directly.

2 Properties of the Simplified Model

It is a natural requirement for the mathematical and numerical models of any real life phenomenon that the solutions of the models must possess some basic qualitative properties of the original process. In the present case such qualitative properties are as follows. We formulate them simultaneously with the properties of the mathematical model (3).

[P1] The size of the population at a given spatial position cannot change in time. This means that $S + I + R$ must be constant at any given spatial position.

[P2] The number of the susceptibles cannot grow and the number of the recovered cannot decrease. That is S is a nonincreasing and R is a nondecreasing function of time at any fixed spatial point.

[P3] The number of the susceptible, infective and recovered members must be nonnegative. S, I and R must be always nonnegative if $S > 0$, $I \geq 0$ and $R \equiv 0$ are satisfied at the initial time instant.

Remark 1 Let us notice that property [P2] does not follow from the model directly. If S is positive and $\theta I + \phi I''_{xx}$ is negative in the initial state at a certain point, then the time derivative of S would be positive, according to the first equation in (3). This is qualitatively incorrect.

The validity of the qualitative properties [P1]–[P3] can be guaranteed by the following theorem.

Theorem 1 *If the condition*

$$\theta I + \phi I''_{xx} \geq 0 \tag{5}$$

is satisfied then properties [P2] and [P3] are true for the solution of problem (3)*. Property [P1] is true without any restrictions.*

Proof Property [P1] follows after the addition of the three equations in (3). Let us turn to the proof of properties [P2] and [P3]. Let us divide the first equation in (3) by S, and integrate from 0 to $t^\star$ with respect to time t (the spatial position is kept

fixed). We obtain

$$\log(S(x,t^\star)) - \log(S(x,0)) = -\int_0^{t^\star} (\theta I(x,t) + \phi I''_{xx}(x,t))\,\mathrm{d}t,$$

and reformulating it we have

$$S(x,t^\star) = S(x,0)\exp\left(\int_0^{t^\star} -(\theta I(x,t) + \phi I''_{xx}(x,t))\,\mathrm{d}t\right).$$

Based on (5), the monotone decrease in time and the nonnegativity of S can be seen.

From the second equation of (3) we obtain that, because $S(\theta I + \phi I''_{xx}) \geq 0$, the solution function $I(x,t)$ satisfies the inequality $I(x,t) \geq I(x,0)\exp(-bt)$. This means that I is also nonnegative.

Because I is nonnegative, R is monotone increasing in time, thus it is nonnegative. This follows from the third equation in (3).

We also see that I goes to zero as t tends to infinity, because if $I(x,t) \geq I_0 > 0$ satisfied for some appropriate value I_0 then R would go to infinity (third equation in (3)) but this would contradict to property [P1]. This completes the proof.

Remark 2 Condition (5) depends on the values of the solution I in the whole solution domain of system (3). We cannot give requirements for the initial and boundary conditions that would guarantee the validity of the condition a priori. Despite of this, we found that in the numerical examples (see later) the fulfilment of the condition at the initial state was enough to its validity at later time instants.

Remark 3 Condition (5) guarantees the monotone decrease of the number of the susceptibles [first equation in (3)]. Theorem 1 can be formulated alternatively as follows: If the number of the susceptibles is decreasing in time then the other qualitative properties also hold.

Now we turn to the question whether system (3) can possess travelling wave solutions. This would make it able to mimic epidemics. Thus, following [3], we are looking for solutions in the form

$$S(x,t) = \tilde{S}(x-ct),\quad I(x,t) = \tilde{I}(x-ct),\quad R(x,t) = \tilde{R}(x-ct), \tag{6}$$

where c is a constant that denotes the wave speed, and the univariate functions $\tilde{I}$ and $\tilde{S}$ have the properties

$$\lim_{\xi\to\pm\infty} \tilde{I}(\xi) = 0,\quad \lim_{\xi\to\pm\infty} \tilde{I}'(\xi) = 0,\quad \lim_{\xi\to\infty} \tilde{S}(\xi) = \tilde{S}^\infty > 0. \tag{7}$$

These expressions formulate the fact that there are no infected members at the time instants $t = \pm\infty$ and the density of the susceptible members is a positive constant at the time instant $t = -\infty$.

Inserting functions (6) into the system (3) and integrating the equations from ξ to ∞ and taking into the consideration the assumptions (7), after some simple manipulations we obtain the system of ordinary differential equations

$$\begin{aligned}\tilde{I}' &= \frac{c}{\phi}\log(\tilde{S}/\tilde{S}^{\infty}) - \frac{\theta c}{b\phi}(\tilde{I}+\tilde{S}-\tilde{S}^{\infty}),\\ \tilde{S}' &= \frac{b}{c}\tilde{I} - \frac{c}{\phi}\log(\tilde{S}/\tilde{S}^{\infty}) + \frac{\theta c}{b\phi}(\tilde{I}+\tilde{S}-\tilde{S}^{\infty}),\\ \tilde{R}' &= -\frac{b}{c}\tilde{I}.\end{aligned} \tag{8}$$

Let us consider the first equation in (8) at $\xi = -\infty$. Based on the assumptions (7), the equality can be true if and only if

$$\frac{\tilde{S}^{-\infty}}{\tilde{S}^{\infty}} = \exp\left(\frac{\tilde{S}^{\infty}\theta}{b}\left(\frac{\tilde{S}^{-\infty}}{\tilde{S}^{\infty}} - 1\right)\right),$$

where $\tilde{S}^{-\infty} = \lim_{\xi\to-\infty}\tilde{S}(\xi)$. The equality is trivially true if $\tilde{S}^{-\infty} = \tilde{S}^{\infty}$. If $\tilde{S}^{\infty}\theta/b \le 1$ then this is the only solution. In this case, however, the number of the susceptible members does not change, that is no epidemic occurs. Thus the condition

$$\tilde{S}^{\infty} > b/\theta \tag{9}$$

(the initial density of the susceptible members must be sufficiently large) is a necessary condition for the propagation of the disease. In this case $\tilde{S}^{-\infty} < b/\theta$, that is the epidemic wave does not leave enough susceptible members back to be able to sustain a new wave.

It is possible to gain a lower bound for the speed of the epidemic. One of the critical points of the first two equations in (8) is $(\tilde{S},\tilde{I}) = (\tilde{S}^{\infty},0)$. Linearising the system at this critical point, the eigenvalues of the coefficient matrix are

$$\lambda_{1,2} = \frac{-c \pm \sqrt{c^2 + 4\tilde{S}^{\infty}b\phi - 4(\tilde{S}^{\infty})^2\phi\theta}}{2\tilde{S}^{\infty}\phi} = \frac{-c \pm \sqrt{c^2 - 4\tilde{S}^{\infty}\phi(b-\tilde{S}^{\infty}\theta)}}{2\tilde{S}^{\infty}\phi}.$$

If (9) is satisfied then the critical point is either a stable spiral point or a stable node. Because around a spiral point $\tilde{S}$ would take greater values than $\tilde{S}^{\infty}$ (more susceptible than at the starting instant) the critical point must be a stable node. That is only epidemic waves with speed

$$c \ge 2\sqrt{\tilde{S}^{\infty}\phi(\tilde{S}^{\infty}\theta - b)}$$

can exist.

3 Numerical Solution of the Simplified Model and Its Qualitative Properties

We solve (3) numerically by the finite difference method on a finite spatial interval $[0, L]$. At the two ends of the interval homogeneous Neumann boundary conditions are applied.

We define a uniform spatial grid $\omega_h = \{x_k \in [0, L] \,|\, x_k = kh,\ k = 0, \ldots, N,\ h = L/N\}$ and a time step $\tau > 0$. The functions S, I and R are approximated respectively by the grid functions s^n, i^n and r^n at the nth time level $t = n\tau$. For $n = 0$, the grid functions are known from certain initial conditions.

Let us consider the discretization scheme

$$\begin{aligned}
\frac{s_k^{n+1} - s_k^n}{\tau} &= -s_k^n \left(\theta i_k^n + \phi \frac{i_{k-1}^n - 2i_k^n + i_{k+1}^n}{h^2} \right), \\
\frac{i_k^{n+1} - i_k^n}{\tau} &= s_k^n \left(\theta i_k^n + \phi \frac{i_{k-1}^n - 2i_k^n + i_{k+1}^n}{h^2} \right) - b i_k^n, \\
\frac{r_k^{n+1} - r_k^n}{\tau} &= b i_k^n,
\end{aligned} \tag{10}$$

for the indices $k = 0, \ldots, N$, where we define the values with the spatial indices -1 and $N + 1$ to be the values with indices 1 and $N - 1$, respectively (homogeneous Neumann boundary).

The discrete versions of the qualitative properties [P1]–[P3] can be easily formulated for the numerical solution simply changing the functions S, I and R to the mesh functions s^n, i^n and r^n.

Theorem 2 *The finite difference scheme* (10) *satisfies the discrete version of [P1] and if the relation*

$$0 \le \theta i_k^n + \phi \frac{i_{k-1}^n - 2i_k^n + i_{k+1}^n}{h^2} \tag{11}$$

is true for all possible indices k and n, and the time step satisfies the condition

$$\tau \le \min \left\{ \frac{1}{M(\theta + 4\phi/h^2)}, \frac{1}{b} \right\}, \tag{12}$$

where $M = \max_{x \in [0,L]} \{S(x, 0) + I(x, 0) + R(x, 0)\}$*, then the discrete versions of the properties [P2] and [P3] are also satisfied.*

Proof The proof can be carried out analogously to the proof of Theorem 1. The first part of the statement follows again after adding the three equations in (10).

Let us introduce the notation

$$j_k^n = \left(\theta i_k^n + \phi \frac{i_{k-1}^n - 2i_k^n + i_{k+1}^n}{h^2}\right) \geq 0,$$

and define C_0 as the maximum of the values j_k^n. Then, from the first equation of (10), we have $s_k^{n+1} = (1 - \tau j_k^n) s_k^n$. If $\tau \leq 1/C_0$ then s_k^n is monotonically decreasing in n and remains nonnegative.

From the second equation in (10), based on the nonnegativity of the first term on the right hand side, we obtain that $i_k^n \geq (1 - \tau b)^n i_k^0 \geq 0$. This implies that r_k^n is also nonnegative, and that all grid functions are bounded by M from above. This gives that $C_0 = M(\theta + 4\phi/h^2)$ is a good choice. This completes the proof.

Remark 4 Based on the proof of the previous theorem, we can state that the numerical scheme is stable in maximum norm provided that the condition of Theorem 2 is satisfied.

We turn to the numerical verifications of the results of the previous section. We have seen that epidemic waves can occur for special parameter choices. We are going to demonstrate this effect on some numerical test examples. We are also going to check the validity of the properties [P1]–[P3] for the numerical solution.

We set $L = 10$, $\delta = 7/100$ and $b = 5/100$. The weighting function is defined to be $W(|u|) = 1 - |u|/\delta$ for $|u| \in [0, \delta]$ and zero otherwise (Fig. 1). With this choice, formulas in (4) give $\theta = \delta$ and $\phi = \delta^3/12$. The spatial step size is set to $h = 1/30$ and the time step is chosen as $\tau = 1$ according to the upper bound (12) $\tau \leq 5.7837$ ($M = 1$, see the initial conditions later).

The first initial condition can be seen on the left panel of Fig. 2. In the middle of the interval $[0, 10]$ 80 % of the individuals are infected, the others are susceptible. Because $\tilde{S}^\infty = 1 > b/\theta = 0.7142$, the necessary condition of the birth of an epidemic wave is satisfied. There are enough susceptibles to sustain the wave. Albeit, the condition is only necessary, the numerical test shows that now it is also sufficient. The density functions are plot at the time instant $t = 1500$ (right panel of Fig. 2). The left hill on the dashed curve (infected part) moves as a wave to the left and the right one to the right. After the epidemic wave passed the density of

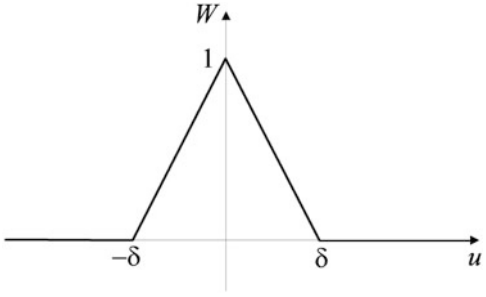


Fig. 1 Graph of the weighting function W

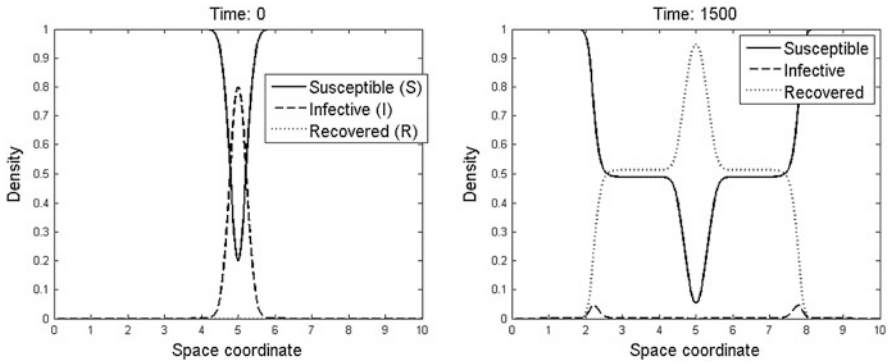


Fig. 2 Epidemic case: the initial conditions (*left panel*) and the states at the time level $t = 1500$ (*right panel*). The left hill on the *dashed curve* (infected part) moves to the left and the right one to the right

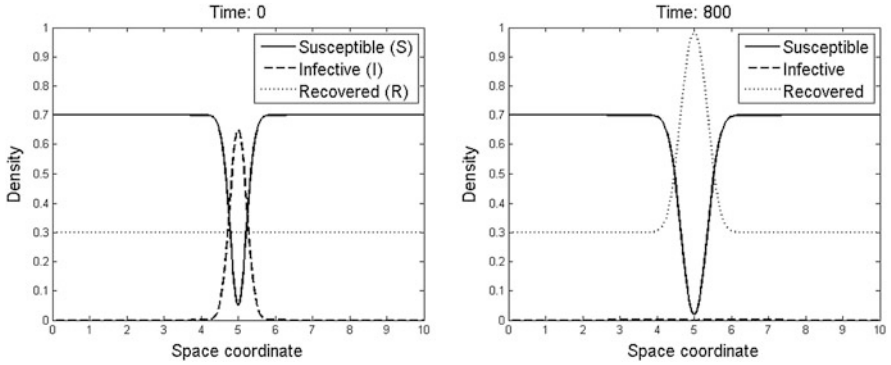


Fig. 3 Non-epidemic case: the initial conditions (*left panel*) and the states at the time level $t = 800$ (*right panel*)

susceptibles drops down to around 0.5 which is not enough to generate a new wave later.

The value $b/\theta = 0.7142$ shows that if more than 28.58 % of the population is immunized before the disease starts (left panel of Fig. 3) then an epidemic wave is not able to develop. The disease localized only around its starting position. The capture of the densities at $t = 800$ is seen on the right panel of Fig. 3. In this way, we are able to obtain an immunization strategy. We can give that how many individuals must be immunized before the epidemic wave reaches a given region and we can stop the propagation of the disease.

Regarding the fulfilment of the qualitative properties, we can state that all the properties [P1]–[P3] were satisfied in the numerical tests. Executing several models with different initial conditions and parameters, we surmise that it is enough to guarantee (11) only for the initial state ($n = 0$) provided that the time step is sufficiently small. In this way the qualitative properties may be guaranteed a priori.

Acknowledgements Both authors were supported by the Hungarian Research Fund OTKA under grant no. K112157.

References

1. Anderson, R.M.: Population Dynamics of Infectious Diseases: Theory and Applications. Chapman and Hall, London (1982)
2. Brauer, F., Castillo-Chávez, C.: Mathematical Models in Population Biology and Epidemiology. Springer, New York (2001)
3. Jones, D.S., Sleeman, B.D.: Differential Equations and Mathematical Biology. Chapman & Hall/CRC Mathematical Biology and Medicine Series. CRC Press, Boca Raton (2011)
4. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proc. R. Soc. A Math. Phys. Eng. Sci. **115**(772), 700–721 (1927)

MS 22
MINISYMPOSIUM: NEW PROGRESS ON NUMERICAL MODELING OF GEOPHYSICAL FLOWS FOR ENVIRONMENT, NATURAL HAZARDS, AND RISK EVALUATION

Organizers

Manuel J. Castro Díaz[1], Carlos Parés Madroñal[2] and Giovanni Russo[3]

Speakers

José M. González Vida[4], Manuel J. Castro[1], Jorge Macías, Marc de La Asunción, Sergio Ortega and Carlos Sánchez-Linares
2D GPU-Based HySEA Model for Tsunami Simulation. Some Practical Examples

Edie Miglio[5]
Multidimensional Coupling for Shallow Water Flows

Manuel J. Castro Diaz, Yuanzhen Cheng, Alina Chertock[6], Alexander Kurganov and Tomas Morales de Luna
Path-Conservative Central-Upwind Schemes for Nonconservative Hyperbolic Systems

[1]Manuel J. Castro Diaz, Departamento de Análisis Matemático, Universidad de Málaga, Málaga, Spain. email: castro@anamat.cie.uma.es.

[2]Carlos Carlos Parés Madroñal, Departamento Análisis Matemático, Universidad de Málaga, Málaga, Spain. email: pares@uma.es.

[3]Giovanni Russo, Dipartimento di Matematica ed Informatica, Università di Catania. email: russo@dmi.unict.it.

[4]José Manuel González Vida, Universidad de Málaga, Málaga, Spain.

[5]Edie Miglio, Politecnico di Milano, Milano, Italy.

[6]Alina Chertock, North Carolina State University, North Carolina, USA.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_73

Alessandro Valiani[7] and Valerio Caleffi
Some Results on the Numerical Treatment of Movable Bed Shallow Water Equations

Christophe Berthon[8]
Well-Balanced Schemes for Shallow-Water Models with Strongly Non-linear Source Terms

Alina Chertock, Michael Dudzinski, Alexander Kurganov[9] and Maria Lukacova-Medvidova
Well-Balanced Schemes for the Shallow Water Equations with Coriolis Force

Enrique Fernandez-Nieto[10], José María Gallardo Molina and Paul Vigneaux
Modelling and Numerical Approach of Viscoplastic Avalanches

Marie-Odile Bristeau, Raouf Hamouda and Jacques Sainte-Marie[11]
Simulations of 3D Navier-Stokes Equations with Free Surface for Hydrodynamics-Biology Coupling and Marine Energies

Tomás Morales de Luna[12], Enrique D. Fernández Nieto and Manuel J. Castro Díaz
A Multilayer Approach for the Simulation of Suspended Sediment Transport

Gianni Pagnini[13] and Andrea Mentrelli
The Randomized Level Set Method and an Associated Reaction-Diffusion Equation to Model Wildland Fire Propagation

Keywords

Geophysical flow
Mathematical models for the environment
Natural hazard
Risk management

Short Description

It is undisputed that the numerical simulation of geophysical processes plays nowadays a central role for understanding the interactions of complex natural

[7]Alessandro Valiani, Università di Ferrara, Italy.
[8]Christophe Berthon, Université de Nantes, France.
[9]Alexander Kurganov, Tulane University, Louisiana, USA.
[10]Enrique D. Fernandez Nieto, Universidad de Sevilla, Spain.
[11]Jaques Sainte-Marie, UPCM- Paris 6, Paris, France.
[12]Tomás Morales de Luna Univiversidad de Cordoba, Cordoba, Spain.
[13]Gianni Pagnini, Basque Center for Applied Mathematics, Bilbao, Spain.

phenomena and for making predictions when systems are perturbed. Moreover, in many cases, numerical models play an important role in the design of early warning system for natural disasters like tsunami or storm/hurricane alert systems.

The main goal of the mini-symposium will be the discussion and presentation of state-of-the-art computational and numerical methods for the next generation of geophysical flow models for environment, natural hazards, and risk evaluation with a focus on finite volume discretizations and HPC techniques for faster than real time simulations.

Motivation/Relevance to ECMI This minisymposium fits in one of the specific topics of the Congress: 'Mathematical methods in environment'. Mathematical models for environment, natural hazards, and risk evaluations are useful tools for public and private companies in different fields such as civil protection, hydraulic engineering, assurances, etc. Therefore, the thematic of this minisymposium also fits in one of the main goals of ECMI: to promote the use of mathematical models in activities of social or economic importance.

The Randomized Level Set Method and an Associated Reaction-Diffusion Equation to Model Wildland Fire Propagation

Gianni Pagnini and Andrea Mentrelli

Abstract Front propagation can be studied by two alternative approaches: the level set method and the reaction-diffusion equation. When a front propagates in a random environment it gets a random character and these two approaches can indeed be considered complementary and reconciled. In fact, if the level set contour is randomized accordingly to the probability density function of the front particle displacement, the resulting averaged process emerges to be governed by an evolution equation of the reaction-diffusion type. This approach turns out to be useful to simulate random effects in wildland fire propagation as those due to turbulent heat convection and fire spotting phenomena.

Keywords Geophysical flow • Randomized level set method • Reaction-diffusion equations • Wildland fire propagation models

1 Introduction

Wildland fire propagation is a complex multi-scale, as well as a multi-physics and multi-discipline process, strongly influenced by the atmospheric wind. Wildland fire is fed by the fuel on the ground and displaced, beside meteorological and orographical factors, also by the hot air that pre-heats the fuel and aids the fire propagation. Heat transfer is turbulent due to the Atmospheric Boundary Layer and

G. Pagnini (✉)
BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo 14, 48009 Bilbao, Spain

Ikerbasque, Basque Foundation for Science, Calle de María Díaz de Haro 3, 48013 Bilbao, Basque Country, Spain
e-mail: gpagnini@bcamath.org

A. Mentrelli
Department of Mathematics, Alma Mater Research Center on Applied Mathematics $(AM)^2$, University of Bologna, via Saragozza 8, 40123 Bologna, Italy

BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo 14, 48009 Bilbao, Basque Country, Spain
e-mail: andrea.mentrelli@unibo.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_74

the fire-induced flow. Moreover, fire generates firebrands that when land on the ground are further sources of fire. Both turbulence and jump-length of firebrands are random processes that affect the fireline propagation.

Fire propagation has been mainly modelled in the literature by using reaction-diffusion type equations, see e.g. [1, 4], and the level set method, see e.g. [3, 5]. Here, an approach based on the level set method is proposed to model the global random effects on fire front propagation due to turbulence and fire spotting. Actually, a reaction-diffusion equation associated to the level-set method is derived.

2 Model Formulation

Let $\Gamma(t)$ be the fire line contour then, in a two dimensional domain, it can be represented as an isoline of an auxiliary function $\gamma(\boldsymbol{x}, t)$, i.e. $\Gamma(t) = \{\boldsymbol{x}, t : \gamma(\boldsymbol{x}, t) = \gamma_0 = constant\}$. The evolution equation of the isoline γ_0 is given by

$$\frac{D\gamma}{Dt} = \frac{\partial \gamma}{\partial t} + \frac{d\boldsymbol{x}}{dt} \cdot \nabla \gamma = \frac{D\gamma_0}{Dt} = 0 \,. \tag{1}$$

Let the motion of the surface points be directed towards the normal direction then

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{V}(\boldsymbol{x}, t) = \mathscr{V}(\boldsymbol{x}, t)\,\widehat{\boldsymbol{n}} \,, \quad \widehat{\boldsymbol{n}} = -\frac{\nabla \gamma}{||\nabla \gamma||} \,, \tag{2}$$

and (1) becomes

$$\frac{\partial \gamma}{\partial t} = \mathscr{V}(\boldsymbol{x}, t)\, ||\nabla \gamma|| \,, \tag{3}$$

which is the *ordinary* level set equation. Let $\varphi(\gamma(\boldsymbol{x}, t))$ be an indicator function such that

$$\varphi(\gamma(\boldsymbol{x}, t)) = \begin{cases} 1 \,, \ \gamma(\boldsymbol{x}, t) > \gamma_0 \,, \ \boldsymbol{x} \in \Omega(t) \,, \ \text{burned area} \,, \\ \\ 0 \,, \ \gamma(\boldsymbol{x}, t) \leq \gamma_0 \,, \ \boldsymbol{x} \notin \Omega(t) \,, \ \text{unburned area} \,. \end{cases} \tag{4}$$

The boundary of $\Omega(t)$ is $\Gamma(t)$, that is the front line contour of the wildland fire. Quantity $\mathscr{V}(\boldsymbol{x}, t)$ is identified with the so-called Rate Of Spread (ROS) [3, 5]. Several determinations of the ROS have been proposed, see e.g. [2, 3, 9]. The present formulation holds for any determination of the ROS.

Let the burning fireline be embodied by a large number of *active* flame holders. Let the motion of each *active* flame holder belonging to the fireline be random due to turbulence and fire spotting effects. For any realization indexed by ω, the random trajectory of each *active* flame holder is stated to be $\boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_{ROS}(t,\overline{\boldsymbol{x}}_0) + \chi^{\omega} + \xi^{\omega}$, where χ and ξ are two random noises that reproduce the randomness of turbulence and fire spotting. The deterministic component $\overline{\boldsymbol{x}}_{ROS}$ corresponds to the motion obtained by literature determination of the ROS [2, 3, 9]. The trajectory of a single *active* flame holder is marked out by the one-particle density function $f^{\omega}(\boldsymbol{x};t) = \delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}_0))$, where $\delta(\boldsymbol{x})$ is the Dirac-delta function. The random trajectory $\boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}_0)$ has the same fixed initial condition $\boldsymbol{X}^{\omega}(0,\overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_{ROS}(0,\overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_0$ in all realizations. Let $\gamma(\overline{\boldsymbol{x}}_0,0)$ be the initial fixed fireline contour, the evolution in time of the fireline according to the ω-realization of the trajectories of the *active* flame holders follows to be

$$\gamma^{\omega}(\boldsymbol{x}(t)) = \int_{\Gamma_0} \gamma(\overline{\boldsymbol{x}}_0,0)\,\delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}_0))\,d\overline{\boldsymbol{x}}_0\,, \tag{5}$$

where $\Gamma_0 = \{\boldsymbol{x} : \gamma(\overline{\boldsymbol{x}},0) = \gamma_0\}$.

Denoting by $\langle\cdot\rangle$ the ensemble average, the average trajectory $\langle \boldsymbol{X}^{\omega}(t;\overline{\boldsymbol{x}}_0)\rangle = \overline{\boldsymbol{x}}(t,\overline{\boldsymbol{x}}_0)$ is driven by the deterministic velocity field $d\overline{\boldsymbol{x}}/dt = \boldsymbol{V}(\overline{\boldsymbol{x}},t)$. Then, trajectory $\overline{\boldsymbol{x}}(t,\overline{\boldsymbol{x}}_0)$ emerges to be time-reversible and the Jacobian of the transformation follows to be $J = d\overline{\boldsymbol{x}}_0/d\overline{\boldsymbol{x}} \neq 0$. To study the potentialities of the proposed approach, the working hypothesis $J = 1$ is made. Finally, by time inversion and ensemble averaging, from (5) the effective fire front contour emerges to be in terms of the indicator function $\varphi(\boldsymbol{x},t)$ as follows

$$\begin{aligned}\langle\varphi^{\omega}(\boldsymbol{x}(t))\rangle &= \langle \int_{R^2} \varphi(\overline{\boldsymbol{x}},t)\,\delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}))\,d\overline{\boldsymbol{x}}\rangle = \int_{R^2} \varphi(\overline{\boldsymbol{x}},t)\,\langle\delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}))\rangle\,d\overline{\boldsymbol{x}} \\ &= \int_{R^2} \varphi(\overline{\boldsymbol{x}},t) f(\boldsymbol{x};t|\overline{\boldsymbol{x}})\,d\overline{\boldsymbol{x}} = \varphi_e(\boldsymbol{x},t)\,,\end{aligned} \tag{6}$$

where $f(\boldsymbol{x};t|\overline{\boldsymbol{x}}) = \langle\delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}))\rangle$ is the probability density function (PDF) of the distribution of the particles of the fireline contour around the average front location $\overline{\boldsymbol{x}}$ and the definition of $\varphi(\overline{\boldsymbol{x}},t)$ stated in (4) has been used.

Field variable $\varphi_e(\boldsymbol{x},t)$ is computed from formula (6) where indicator function $\varphi(\overline{\boldsymbol{x}},t)$ follows from solving the level set equation driven by an average front velocity.

By applying the Reynolds transport theorem to (6), the evolution equation of the effective fire front $\varphi_e(\boldsymbol{x},t)$ is [6]

$$\frac{\partial\varphi_e}{\partial t} = \int_{\Omega(t)} \frac{\partial f}{\partial t}\,d\overline{\boldsymbol{x}} + \int_{\Omega(t)} \nabla_{\overline{\boldsymbol{x}}}\cdot[\boldsymbol{V}(\overline{\boldsymbol{x}},t) f(\boldsymbol{x};t|\overline{\boldsymbol{x}})]\,d\overline{\boldsymbol{x}}\,, \tag{7}$$

that is the reaction-diffusion equation associated to the level set equation (3).

Since the effective fireline contour $\varphi_e(\boldsymbol{x}, t)$ is a smooth function continuously ranging from 0 to 1, a criterion to mark burned points have to be stated. Here points $\boldsymbol{x}$ such that $\varphi_e(\boldsymbol{x}, t) > 0.5$ are marked as burned and the effective burned area emerges to be $\Omega_e(t) = \{\boldsymbol{x}, t : \varphi_e(\boldsymbol{x}, t) > 0.5\}$. However, beside this criterion, a further criterion associated to an ignition delay due to the pre-heating action of the hot air or to the landing of firebrands is introduced. Hence, in the proposed modelling approach, an unburned point $\boldsymbol{x}$ will be marked as burned when one of these two criteria is met.

This ignition delay, due to a certain *heating-before-burning mechanism*, can be depicted as an accumulation in time of heat [7], i.e.

$$\psi(\boldsymbol{x}, t) = \int_0^t \varphi_e(\boldsymbol{x}, \eta) \, \frac{d\eta}{\tau} \,, \tag{8}$$

where $\psi(\boldsymbol{x}, 0) = 0$ corresponds to the unburned initial condition and τ is a characteristic ignition delay. Since the fuel can burn because of two pathways, i.e. hot-air heating and firebrand landing, the resistance analogy suggests that τ can be approximatively computed as resistances acting in parallel, i.e.

$$\frac{1}{\tau} = \frac{1}{\tau_h} + \frac{1}{\tau_f} = \frac{\tau_f + \tau_h}{\tau_h \tau_f} \,, \tag{9}$$

where τ_h and τ_f are the ignition delays due to hot air and firebrands, respectively.

The amount of heat is proportional to the increasing of the fuel temperature $T(\boldsymbol{x}, t)$, then

$$\psi(\boldsymbol{x}, t) \propto \frac{T(\boldsymbol{x}, t) - T(\boldsymbol{x}, 0)}{T_{ign} - T(\boldsymbol{x}, 0)} \,, \quad T(\boldsymbol{x}, t) \leq T_{ign} \,, \tag{10}$$

where T_{ign} is the ignition temperature. Finally, when $\psi(\boldsymbol{x}, t) = 1$ the ignition temperature is assumed to be reached, so that a new ignition occurs in $(\boldsymbol{x}, t)$ and, with reference to (6), the modelled fire goes on by setting $\varphi(\boldsymbol{x}, t) = 1$.

3 Discussion and Conclusions

The present analysis constitutes a proof-of-concept and it needs to be subjected to a future validation. Hence, numerical results showed in Figs. 1, 2, 3, 4, 5 are understood as explorative exercises to investigate the potentialities of the approach. From comparison of the level set method against the proposed model when only

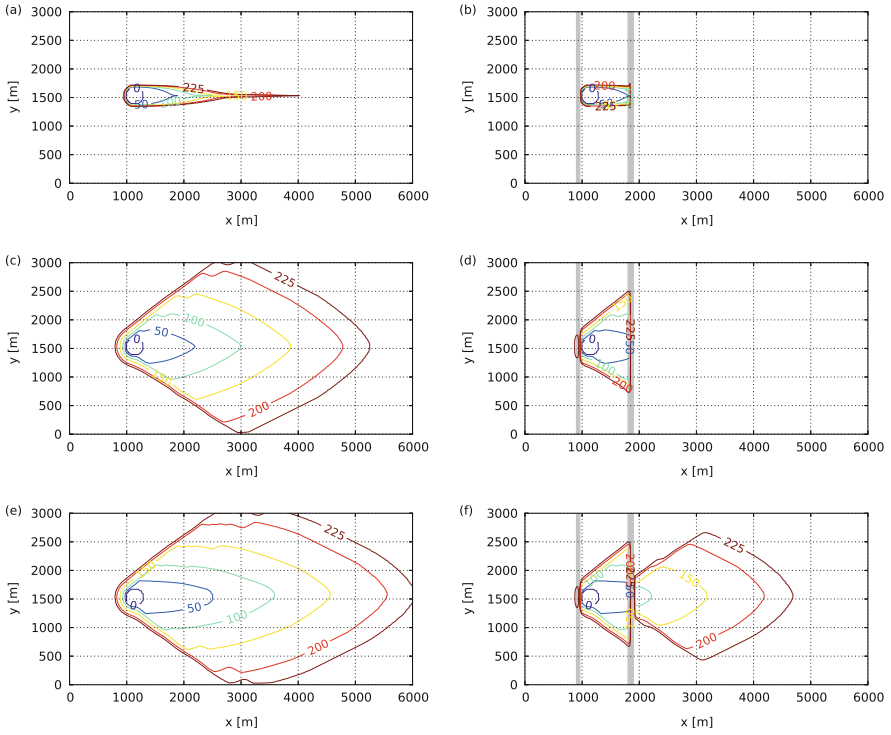


Fig. 1 Time evolution of the firefront in absence (*on the left*, **a**) and presence (*on the right*, **b**) of two fire-break zones (*grey stripes*). The results are obtained by adopting the level set method (*top row*), by the present modelling approach when only turbulence is taken into account (*middle row*), and when both turbulence and fire spotting are considered (*bottom row*). The labels on the contour lines represent the propagation time (expressed in minutes). Following [8], turbulence has been parameterized with a Gaussian PDF and fire spotting with a stationary log-normal distribution for jump-length of embers with mean stated equal to $\mu = 1.32\, I_f^{0.26}\, U_t^{0.11} - 0.02$ and standard deviation $s = 4.95\, I_f^{-0.01}\, U_t^{-0.02} - 3.48$, where U_t is the modulus of the mean wind, assumed constant both in value ($6.70\,\mathrm{m\,s^{-1}}$) and direction (x-axis), and $I_f = I + I_t$ where $I = 10{,}000\,\mathrm{kW\,m^{-1}}$ is the fire intensity and $I_t = 0.015\,\mathrm{kW\,m^{-1}}$ is the tree torching intensity. Other simulation parameters are: $V_{ROS} = I/(Hw_0)$ where $H = 22{,}000\,\mathrm{kJ\,kg^{-1}}$ is the fuel low heat of combustion and $w_0 = 2.243\,\mathrm{kg\,m^{-2}}$ is the oven-dry mass of fuel, $\mathscr{D} = 0.04\,\mathrm{m^2\,s^{-1}}$, $\tau_h = 600\,\mathrm{s}$, $\tau_f = 60\,\mathrm{s}$ and the width of fire-breaks is 60 m in the windward sector and 90 m in the leeward sector

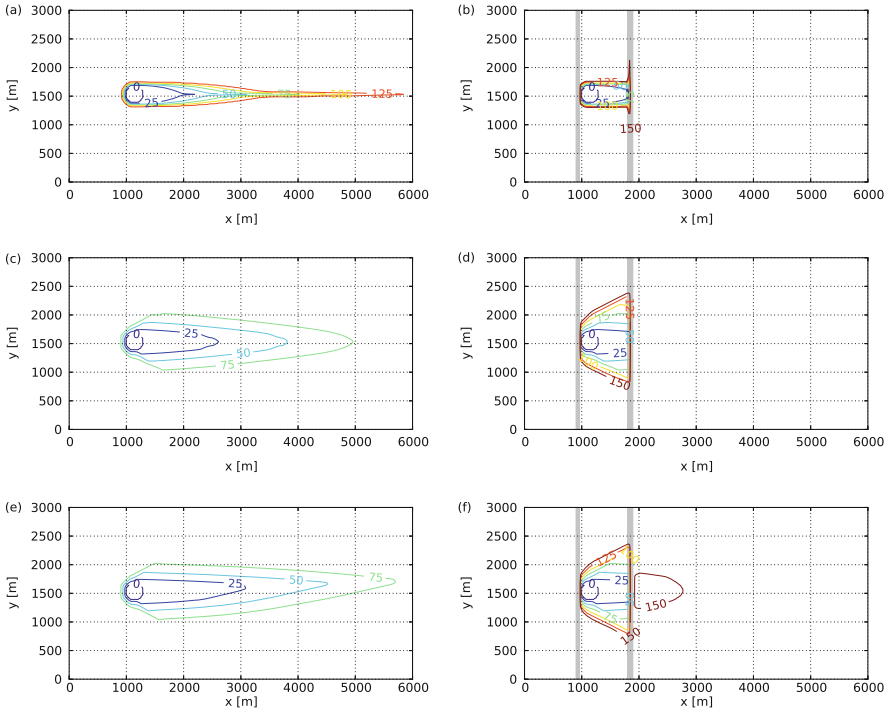


Fig. 2 The same as in Fig. 1 but when $U_t = 6.70\,\mathrm{m\,s^{-1}}$ and $I = 30{,}000\,\mathrm{kW\,m^{-1}}$

turbulence and when both turbulence and fire spotting are taken into account, it emerges the suitability of the proposed approach to simulate a fire that overcomes a fire-break zone, in contrast to the level set method. Moreover, it emerges also that the inclusion of turbulence allows for simulating fire flank and backing fire

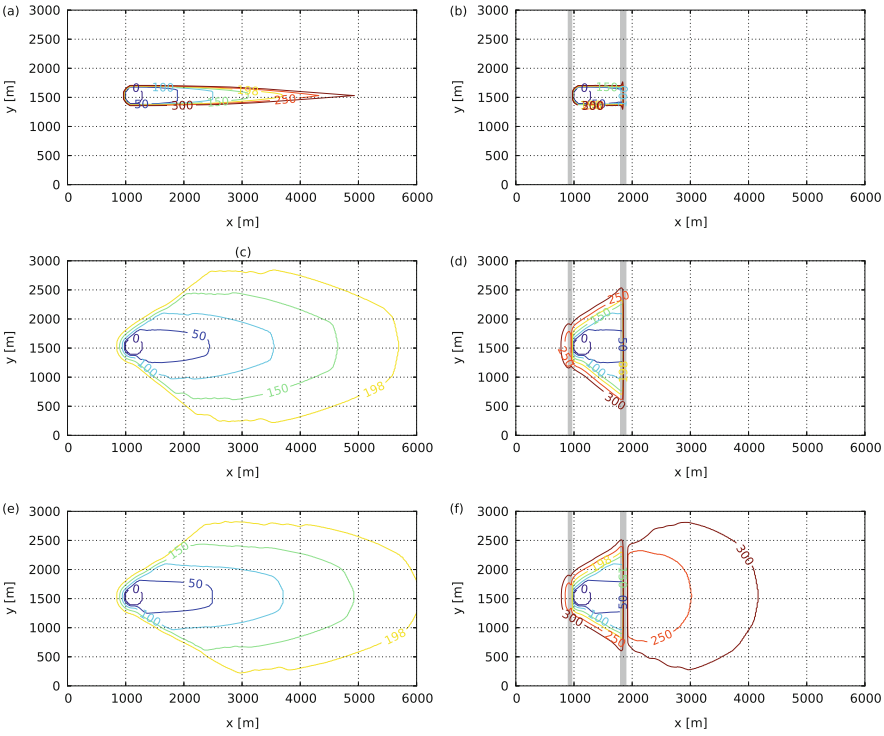


Fig. 3 The same as in Fig. 1 but when $U_t = 17.88\,\mathrm{m\,s^{-1}}$ and $I = 10{,}000\,\mathrm{kW\,m^{-1}}$

and the inclusion of hot air preheating and ember landing enhances the frontline propagation. This richness of model behaviours supports the proposed formulation as a promising approach to simulate the complex phenomenology of real wildland fire propagation.

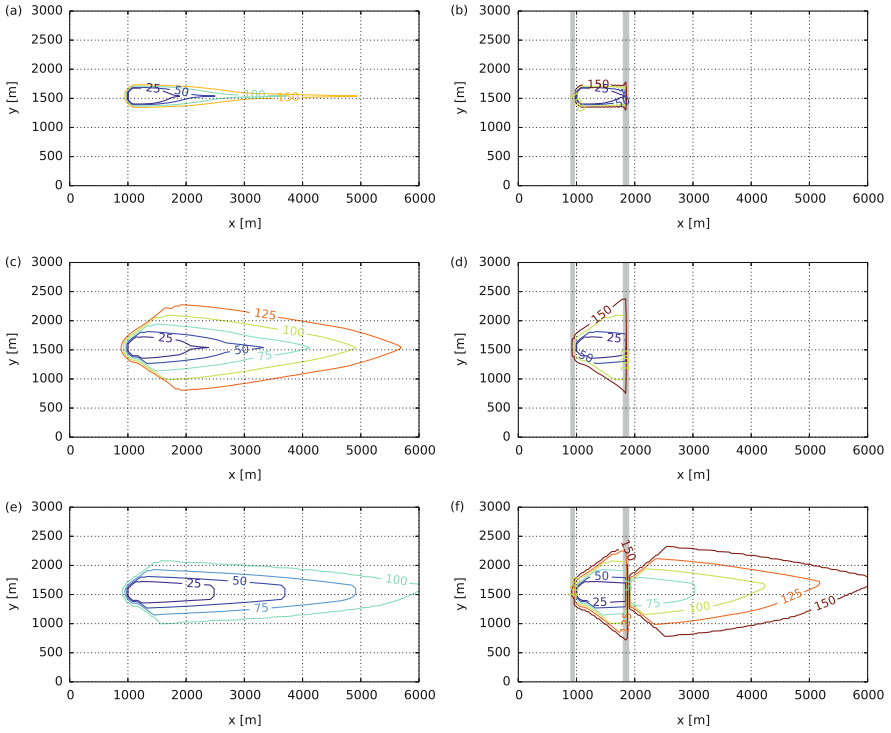


Fig. 4 The same as in Fig. 1 but when $U_t = 17.88\,\mathrm{m\,s^{-1}}$ and $I = 20{,}000\,\mathrm{kW\,m^{-1}}$

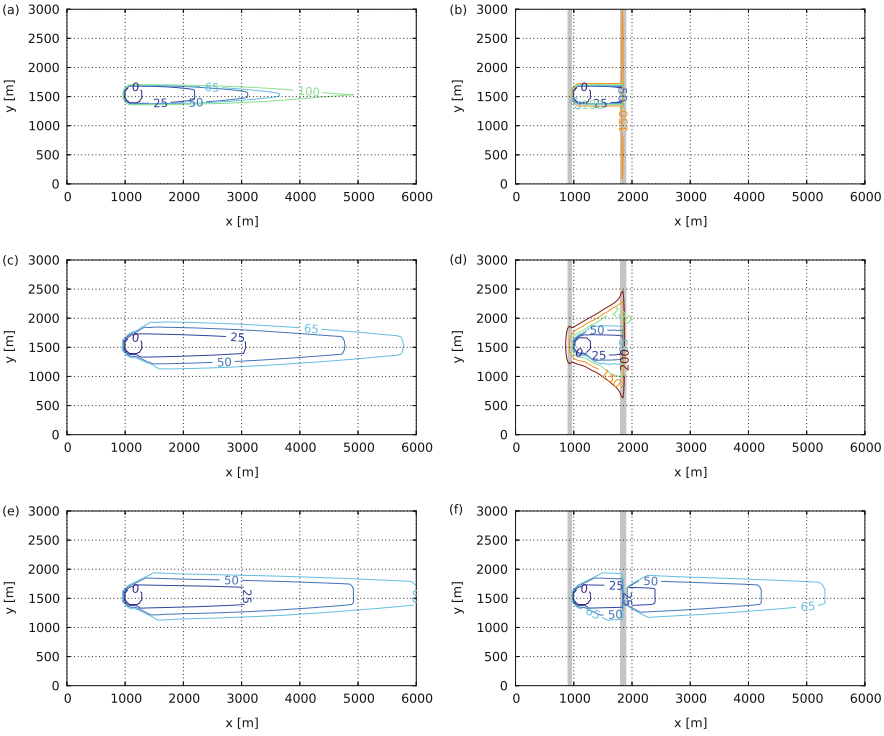


Fig. 5 The same as in Fig. 1 but when $U_t = 17.88\,\mathrm{m\,s^{-1}}$ and $I = 30{,}000\,\mathrm{kW\,m^{-1}}$

Acknowledgements This research is supported by GNFM/INdAM Young Researchers Project 2013, by Bizkaia Talent and European Commission through COFUND programme under Grant AYD-000-226, and also by the Basque Government through the BERC 2014–2017 program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa accreditation SEV-2013-0323.

References

1. Asensio, M.I., Ferragut, L.: On a wildland fire model with radiation. Int. J. Numer. Methods Eng. **54**, 137–157 (2002)
2. Balbi, J.H., Morandini, F., Silvani, X., Filippi, J.B., Rinieri, F.: A physical model for wildland fires. Combust. Flame **156**, 2217–2230 (2009)
3. Mallet, V., Keyes, D.E., Fendell, F.E.: Modeling wildland fire propagation with level set methods. Comput. Math. Appl. **57**, 1089–1101 (2009)
4. Mandel, J., Bennethum, L.S., Beezley, J.D., Coen, J.L., Douglas, C.C., Kim, M., Vodacek, A.: A wildland fire model with data assimilation. Math. Comput. Simulat. **79**, 584–606 (2008)
5. Mandel, J., Beezley, J.D., Kochanski, A.K.: Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011. Geosci. Model. Dev. **4**, 591–610 (2011)

6. Pagnini, G., Bonomi, E.: Lagrangian formulation of turbulent premixed combustion. Phys. Rev. Lett. **107**, 044503 (2011)
7. Pagnini, G., Massidda, L.: The randomized level-set method to model turbulence effects in wildland fire propagation. In: Spano, D., Bacciu, V., Salis, M., Sirca, C. (eds.) Modelling Fire Behaviour and Risk. Proceedings of the International Conference on Fire Behaviour and Risk. ICFBR 2011, Alghero, 4–6 Oct 2011, pp. 126–131 (2012). ISBN 978-88-904409-7-7
8. Pagnini, G., Mentrelli, A.: Modelling wildland fire propagation by tracking random fronts. Nat. Hazards Earth Syst. Sci. **14**, 2249–2263 (2014)
9. Rothermel, R.C.: A mathematical model for predicting fire spread in wildland fires. Technical Report. Research Paper INT-115, USDA Forest Service, Intermountain Forest and Range Experiment Station, Ogden, Utah 84401 (1972). Available at: http://www.treesearch.fs.fed.us/pubs/32533

MS 23
MINISYMPOSIUM: NON-HYDROSTATIC WAVE PROPAGATION WITH DEPTH AVERAGED EQUATIONS: MODELS AND METHODS

Organizers

Anargiros I. Delis[1] and Mario Ricchiuto[2]

Speakers

A.I. Delis[1] and M. Kazolea[3]
Advanced Numerical Simulation of Near-Shore Processes by Extended Boussinesq-Type Models on Unstructured Meshes

Claes Eskilsson[4] and Allan P. Engsig-Karup[5]
On Devising Boussinesq-Type Equations with Bounded Eigen-Spectra: Two Horizontal Dimensions

[1]Anargiros I. Delis, School of Production Engineering & Management, Technical University of Crete, Chania, Crete, Greece.

[2]Mario Ricchiuto, Team CARDAMOM, Inria Bordeaux Sud-Ouest, France.

[3]Maria Kazolea, Team CARDAMOM, Inria Bordeaux Sud-Ouest, France.

[4]Claes Eskilsson, Shipping and Marine Technology, Chalmers University of Technology, Goteborg, Sweden.

[5]Allan P. Engsig-Karup, DTU Compute, Danish Technical University, Lyngby, Denmark.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_75

Andrea Filippini[6], Stevan Bellec[7], Mathieu Colin[8] and Mario Ricchiuto[2]
On Nonlinear Shoaling Properties of Enhanced Boussinesq Models

Keywords

Boussinesq-type models
Near-shore hydrodynamics
Wave propagation

Short Description

Near shore hydrodynamics involves very complex processes involving the transformation and dissipation of ocean waves, as well as their impact on the shore. This is a domain with enormous impact in the field of near-shore engineering (design of harbours, coastal defence structures, etc.), offshore engineering (platforms design, pipelines, etc.), naval engineering (design of vessels with optimized properties), and environmental management (morphodynamic evolution, pollutant transport, etc.).

Due to the multi-scale nature of this phenomena, the coastal engineering community has turned long ago to approximate, asymptotic, depth averaged models, often referred to Boussinesq-type equations. These partial differential systems are obtained from the incompressible Euler equations under some assumptions on wave length and wave amplitude. The resulting partial differential equations are very complex due to the appearance of nonlinear high order differential terms modelling wave dispersion, wave shoaling an steepening. The derivation of improved variants of these models, having properties closer to those of the Euler system, is a challenge in itself. The numerical discretization of the resulting equations is another challenge.

This symposium aims at providing an overview of these aspects. The topics covered in the talks range from the choice of the form of the PDEs, to their dispersion optimisation, to their discretization with appropriate very high resolution numerical methods. Additional topics include the treatment of wave breaking and of moving shorelines, which are phenomena of paramount importance for the hydrodynamics in the near shore region.

[6]Andrea Filippini, Team CARDAMOM, Inria Bordeaux Sud-Ouest, France.
[7]Stevan Bellec, Institut de Mathématiques de Bordeaux and, Team CARDAMOM, Inria Bordeaux Sud-Ouest, France.
[8]Matieu Colin, Institut de Mathématiques de Bordeaux and, Team CARDAMOM, Inria Bordeaux Sud-Ouest, France.

Advanced Numerical Simulation of Near-Shore Processes by Extended Boussinesq-Type Models on Unstructured Meshes

A.I. Delis and M. Kazolea

Abstract A numerical code that employs a higher-order finite volume scheme on unstructured meshes for approximating enhanced Boussinesq-type equations is presented. The objective of this study is to further investigate wave propagation over complex bathymetries using the developed code and to present an approach for the parallelization of the resulted code, along with preliminary numerical results.

Keywords Boussinesq-type models • Near-shore hydrodynamics • Unstructured meshes • Wave propagation

1 Introduction

Accurate simulations of water wave's propagation is of fundamental importance to marine and coastal engineering. Over the last decades, Boussinesq-type equations (BTEs) have been widely used to describe wave transformations in coastal regions, [2]. The success of the BTEs is mainly due to the optimal blend of physical adequacy, in representing all main physical phenomena, and to their relative computational ease. However, the accurate and efficient numerical approximation of BTEs is still in the focus of on-going research especially in terms of higher-order numerics and the adaptive mathematical/numerical description of the flow.

Recently the Finite Volume (FV) method has become the numerical approach of choice for BTMs. This can be attributed to the significantly less computational effort required, compered to other methods (such as the Finite Element). For both 1D and 2D computations, classical FV schemes have been modified to solve enhanced BTEs using the Finite Difference (FD) approach for the discretization of the dispersive

A.I. Delis (✉)
School of Production Engineering & Management, Technical University of Crete, 73100 Chania, Greece
e-mail: adelis@science.tuc.gr

M. Kazolea
INRIA Bordeaux Sud-Ouest, 33405 Talence cedex, France
e-mail: maria.kazolea@inria.fr

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_76

terms, [4, 11, 12, 15], on structured meshes. Very recently, and for the first time, a higher-order FV approach on unstructured meshes was introduced in [5]. Along these lines, the TUCWave model, [5, 6], is briefly presented here. The model employs a novel well-balanced FV scheme for approximating the BTEs of Nwogu [8]. The objective of the present study is to further investigate wave propagation over complex three-dimensional bathymetries. Furthermore, a first approach for the parallel realization of the TUCWave code is also given along with preliminary results.

The BTEs of Nwogu describe accurately weakly non-linear/weakly dispersive waves, i.e. with Stokes number $S = \epsilon/\mu^2 = O(1)$, where $\epsilon := A/h$ with A the wave's amplitude and h the still water level, and $\mu^2 := h^2/L^2$ the water depth to wave length (L) ratio. Following [5], the conservative-like form of the equations reads as:

$$\partial_t \mathbf{U} + \nabla \cdot \mathscr{F}(\mathbf{U}^\star) = \mathbf{S} \quad \text{on} \quad \Omega \times [0,t] \subset \mathbb{R}^2 \times \mathbb{R}^+, \tag{1}$$

where $\mathbf{U}^\star = [H, Hu, Hv]^{\mathrm{T}}$ are the physically conservative variables, $\mathbf{U}$ is the vector of the actual solution variables, with $H = h + \eta$ being the total water depth and

$$\mathbf{U} = \begin{bmatrix} H \\ P_1 \\ P_2 \end{bmatrix}, \quad \mathscr{F} = [\mathbf{F}, \mathbf{G}] = \begin{bmatrix} Hu & Hv \\ Hu^2 + \frac{1}{2}gH^2 & Huv \\ Huv & Hv^2 + \frac{1}{2}gH^2 \end{bmatrix},$$

where, solving at an optimized distance $z_a = -0.531h$ from the still water level,

$$\mathbf{P} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = H\left[\mathscr{C} + \mathbf{u}\right], \text{ with } \mathscr{C} = \frac{z_a^2}{2}\nabla(\nabla \cdot \mathbf{u}) + z_a \nabla(\nabla \cdot h\mathbf{u}). \tag{2}$$

The source term vector, $\mathbf{S} = \mathbf{S_b} + \mathbf{S_f} + \mathbf{S_d}$, includes the bed topography's (b) slope $\mathbf{S_b}$, the bed friction effects $\mathbf{S_f}$, given in this work in terms of the Manning coefficient n_m, and the dispersive terms $\mathbf{S_d}$. These terms read as

$$\mathbf{S_b} = [0, \ -gH(\nabla \cdot b)]^{\mathrm{T}}, \ \mathbf{S_d} = [-\psi_c, \ \ \mathbf{u}\psi_c + \boldsymbol{\psi}_\mathbf{M}]^{\mathrm{T}}, \ \mathbf{S_f} = \left[0, \ -gn_m^2 \mathbf{u}||\mathbf{u}||h^{-1/3}\right]^{\mathrm{T}}$$

with

$$\psi_c = \nabla \cdot \left[\left(\frac{z_a^2}{2} - \frac{h^2}{6}\right) h\nabla(\nabla \cdot \mathbf{u}) + \left(z_a + \frac{h}{2}\right) h\nabla(\nabla \cdot h\mathbf{u})\right], \quad \boldsymbol{\psi}_\mathbf{M} = \partial_t H \mathscr{C}. \tag{3}$$

Equations (1) have flux terms identical as those in the Nonlinear Shallow Water equations (NSWE) and $\mathbf{P}$ contains all time derivatives in the momentum equations, including part of the dispersion terms. Vector $\mathbf{S_d}$ contains only spatial derivatives since $\partial_t H$ is explicitly defined by the mass equation. The NSWE are recovered when the dispersive terms in $\mathbf{P}$ and $\mathbf{S_d}$ are vanishing.

2 The Numerical Scheme and Parallelization Strategy

To numerically solve equations (1) we use a Godunov-type FV scheme [5, 6]. The FV approach is of the node-centered median-dual type where the control volumes are elements dual to the primal triangular mesh. The FV integration of (1) over each computational cell, C_P, leads to the semi-discrete form of the scheme as:

$$\frac{\partial \mathbf{U}_P}{\partial t} = -\frac{1}{|C_P|}\sum_{Q\in K_P} \Phi_{PQ} - \frac{1}{|C_P|}\Phi_{P,\Gamma} + \frac{1}{|C_P|}\iint_{C_P} \mathbf{S} d\Omega, \quad P = 1, \ldots, N \tag{4}$$

where $\mathbf{U}_P$ is the volume-averaged value of $\mathbf{U}$ at a given time, K_P is the set of the neighboring nodes to P, Γ is the boundary of the computational domain Ω and Φ_{PQ}, $\Phi_{P,\Gamma}$ are the numerical flux vectors across each internal and boundary face, respectively. The numerical fluxes are evaluated solving a Riemann problem at cell interfaces using the approximate Riemann solver of Roe [10]. To reach higher-order spatial accuracy an extension of the MUSCL methodology of Van Leer [16] is used. Each component of the physical variables and bed topography, b, is extrapolated using solution gradients obtained using a combination of centered and upwind gradients. In this way a third-order well-balanced scheme is obtained for the advection part of the models while consistent FV approximations are implemented for the gradient and divergence operators in the dispersive terms [5]. Details on the numerical model for wet/dry front treatment, boundary conditions and discretization of the dispersive terms, can be found in [5, 6].

For the time discretization an optimal third-order explicit Strong Stability Preserving Runge-Kutta (RK) scheme is utilized. In the RK scheme the velocity field $\mathbf{u} = [u,\ v]^{\mathrm{T}}$ must be recovered from the computed values of the new solution variable $\mathbf{P} = [P_1\ P_2]$ at each node. Discretizing $\mathbf{P}$ in (2), a linear system $\mathbf{AV=C}$ occurs with $\mathbf{V}=[\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_N]^{\mathrm{T}}$ and $\mathbf{C} = [\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N]^{\mathrm{T}}$. Matrix $\mathbf{A} \in \mathbf{R}^{2N\times 2N}$ is sparse, structurally symmetric and mesh depended. The properties $\mathbf{A}$ depend on the physical conditions of the problem to be solved and is stored in a compressed sparse row (CSR) format. The system is solved using the Bi-Conjugate Gradient Stabilized method(BiCGStab). The ILUT pre-conditioner from SPARSKIT package is implemented and the reverse Cuthill–McKee (RCM) algorithm is also employed to reorder the matrix elements as to minimize its bandwidth. Convergence to the solution was obtained in one or two iterations for the test problems considered.

A new wave breaking technique is also incorporated in TUCWave code. It is based on a hybrid BT/NSWE approach [6, 12, 15]. Once a wave breaking interface occurs, BTEs are turned into NSWE by switching off the dispersive terms. In this way, the wave breaking interface is treated as a bore by the NSWE and the shock-capturing FV scheme. Using a new set of physical criteria we first estimate the location of breaking waves and then the NSWE are solved in the breaking regions and BTEs elsewhere. We briefly describe the new methodology below: (1) *Computation of wave breaking criteria for each computational cell*: two phase resolving criteria are used, [6]: (a) the surface variation: $|\partial_t \eta| \geq \gamma\sqrt{gh}$, with

$\gamma \in [0.3,\ 0.65]$ and (b) the local slope angle: $||\nabla\eta||_2 \geq \tan(\phi_c)$, where $\phi_c \approx 30°$ is the critical front face angle at the initiation of breaking. (2) *If at least one of the criteria is satisfied*: we flag the relative nodes as breaking ones. (3) *Distinguish different breaking waves*: creating a dynamical list that contains the breaking nodes of such a wave and different breaking waves are treated individually. The wave front of each breaking wave is then handled as a bore by the NSWE dissipating energy. (4) *Switch back to BTEs* for non-breaking undular bores as characterized based on their Froude, *Fr*, number, i.e. $Fr \leq 1.3$ [6, 14]. (5) For each breaking wave an *extension of the computational region* governed by the NSWE is performed, as to avoid non-physical effects that may appear at the interface between a zone governed by the BTEs and a zone governed by NSWE [6].

For the parallelization approach an explicit partition of the global solution domain (Ω) into overlapping subdomains ($\Omega_s,\ s = 1, 2, \ldots, Pr$), each being attributed to a single processor, is performed. The amount of overlap between the subdomains is always one layer of shared computational cells. For the partitioning we use an unstructured strategy based on the METIS software package and the subdomains have approximately the same number of nodes. The implementation of the resulting parallel algorithm employs the commonly used Message Passing Interface (MPI) library. Each subdomain solver is responsible for finding the solution of (1) in Ω_s. The main differences between the single and the parallel approach is the area of the discretization and the solution process for the recovery of the velocity field in each subdomain. More precisely, the result of the parallel discretization is that the global sparse matrix $\mathbf{A}$ is distributed as a set of subdomain matrices $\mathbf{A_s}$. Each processor s constructs its own matrix $\mathbf{A_s}$ in the preprocessing stage. As such, first its structure is stored in the CSR format and reordered using the RCM reordering technique. Then the pre-conditioner (ILUT) of the reordered matrix $\mathbf{A_s}$ is computed at this pre-processing stage and subsequently utilized to solve the linear system at each time step. Like the global matrix $\mathbf{A}$, the properties of each matrix $\mathbf{A_s}$ vary depending on the physical situation of the problem solved, the type of the grid used and additionally the (sub)domain's, (Ω_s), structure. To solve the sparse linear system among the subdomains, we use an additive Schwarz iteration technique [3]. More precisely and during each time step, each linear system $\mathbf{A_s V_s^k = P_s^{k-1}}$ is solved for k=1,2,... times until global convergence of the solution is achieved. The global solution $\mathbf{V_g^k}$ exists only logically and it is composed by the intermediate subdomain solutions $\mathbf{V_s^k}$. Appropriate local and global-type monitors are used [3, 7].

3 Numerical Tests and Results

3.1 2D Solitary Wave Propagation in a Channel

In an $h = 10\,\text{m}$ deep channel with $(x, y) \in [-100\,\text{m}, 2400\,\text{m}] \times [-5, 5\,\text{m}]$ an $A = 2\,\text{m}$ high solitary wave, i.e. $\epsilon = 0.2$, is initially positioned at $x = 200\,\text{m}$ and the asymptotically analytical solution for η and $\mathbf{u}$ can be found in [17]. A triangular

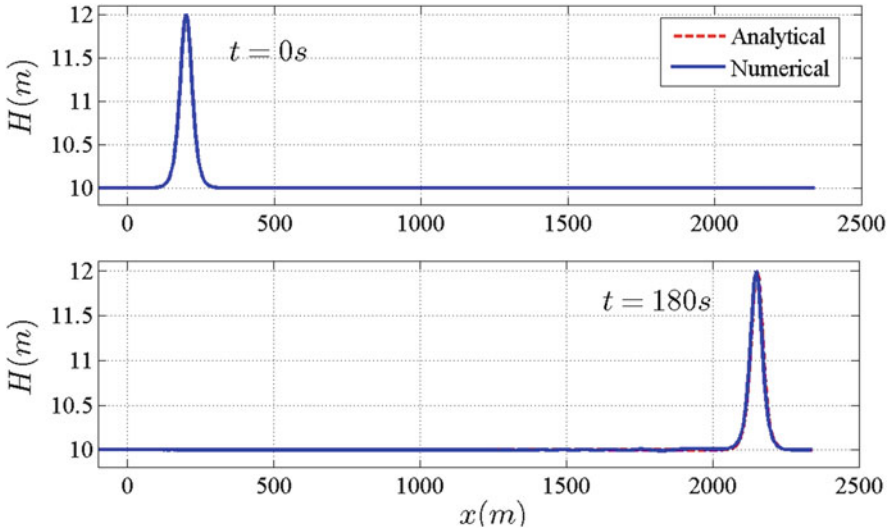


Fig. 1 Solitary wave profiles along a channel of constant depth

mesh consisting of equilateral triangles, with edge size $h_N = 0.75$ m, was used, leading to a mesh of $N = 53{,}304$ nodes. The CFL number used was set equal to 0.6. Figure 1 shows the initial solitary wave and the computed waveform along the channel for $y = 0$ at $t = 180$ s. The computed permanent waveform maintains its symmetry and phase speed which are very close to the (asymptotic) analytical solution.

3.2 Wave Propagation Over an Elliptic Shoal

Berkhoff et al. [1] carried out an experiment to study the refraction and diffraction of 2D monochromatic waves over a complex bathymetry, which has been used for validation of BTMs, e.g. [9, 17]. The topography consists of an elliptic shoal, over an inclined slope of $1/50$. The depth is $h = 0.45$ m at the wave maker. The incoming waves have period $T = 1$ s and $A = 0.0232$ m and $S = 1.13$. An internal source function [17] was used to generate the waves. Surface elevation was measured at sections $x_m = 0$ and $y_m = [1.0,\ 3.0,\ 5.0,\ 7.0,\ 9.0]$ and the mean wave height is computed using the zero-up crossing technique. The grid consists of triangles with edge size of $h_N = 0.1$ m and has been refined in the region of the shoal with $h_N = 0.05$ m and a CFL value of 0.3 was used. The simulation period is 50 s and the ten last waves are employed to estimate the wave height. Results are reported in Fig. 2. The agreement between the numerical results and the experimental data is quite satisfactory and comparable to others found in the literature e.g. [9, 17]. Wave's focusing behind the shoal, due to refraction, is well reproduced. In section 5

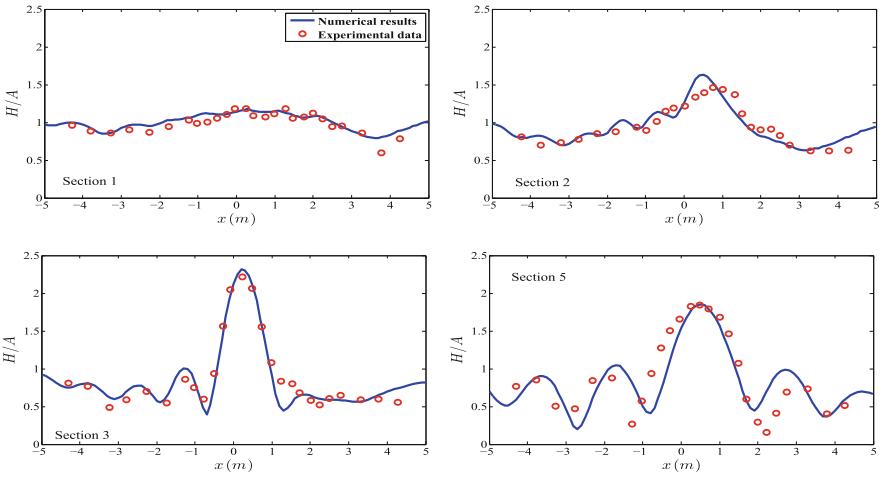


Fig. 2 Comparison of the computed and experimental wave heights in sections 1–3 and 5

the maximum amplification factor is well predicted compared to other results in the literature, e.g. in [17], where it is usually underestimated.

3.3 Solitary Wave Propagation Over a Three-Dimensional Reef

In [13] a laboratory experiment to study phenomena such as wave shoaling, refraction and breaking was presented. Bathymetry information, positions of wave gauges (WGs) that measure the free surface elevation and position of Acoustic Doppler Velocimeters (ADVs) can be found in [11, 13]. An unstructured mesh refined along the shelf with $N = 87,961$. WGs $1-7$ are located along $y = 0\,\text{m}$, gauges $8-13$ along $y = 5\,\text{m}$ and gauges $14-17$ along $x = 25\,\text{m}$. A solitary wave of $0.39\,\text{m}$ in height is placed along $x = 5\,\text{m}$. Figure 3 shows series of snapshots of the wave propagating over the shallow shelf, creating a strongly plunging breaker. Initially, the wave propagates unchanged since the topography is flat. As the wave approaches the shelf apex, breaking begins along the center-line. The bore front propagates onshore while the wave along the sides shoals. Up to time $t = 8\,\text{s}$ a plunging wave has been developed along, the entire length of the reef edge, and a new bore has been developed at the apex of the shelf, which propagates over and away from the sill. At $t = 16\,\text{s}$ a third bore-front is visible further onshore and it is a portion of the first bore which has been reflected off the top of the planar beach generating an offshore flow. Five seconds later the third bore-front converges at the apex of the shelf as a refraction phenomenon, while the flow at the top continues to move forward. Figure 4 shows the computed and recorded elevation time series at WGs 1, 5, 10 and 13. Figure 5 shows the same for the wave gauges located at the

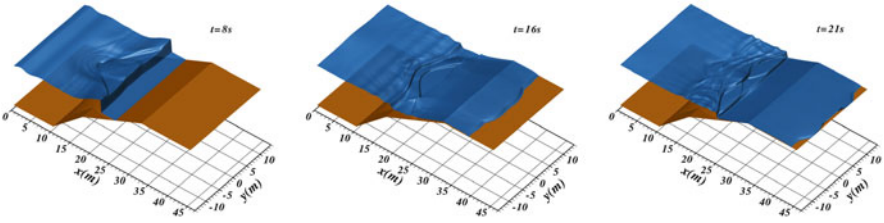


Fig. 3 Water surface for solitary wave propagation on a 3D reef at different times

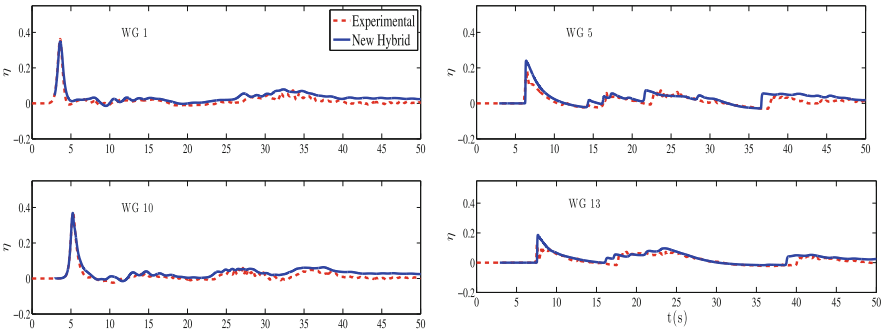


Fig. 4 Propagation on a 3D reef for WGs along the centerline (*top*) and along $y = 5$ m (*bottom*)

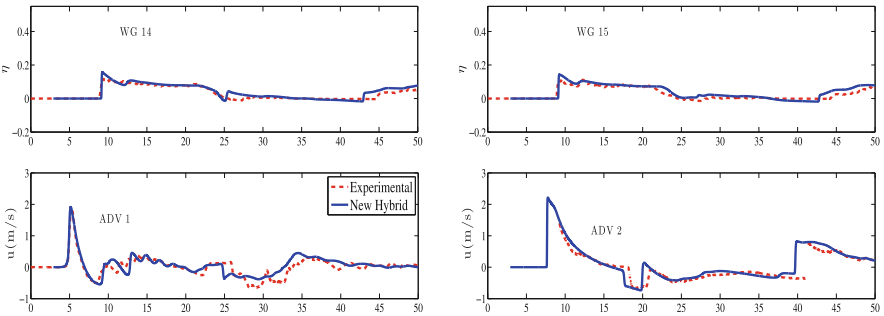


Fig. 5 Propagation on a 3D reef, WGs at the edge of the reef flat (*top*) and of the velocity (*bottom*)

edge of the reef flat and compares also recorder and computed velocity components in the x-direction for two WGs. The numerical results agree very well with the measurements at the presented wave gauges.

3.4 Regular Wave Propagation Over a Submerged Bar

A regular wave propagation over a submerged bar test is implemented using the parallel version of TUCWave code to investigate the frequency dispersion characteristics and nonlinear interactions. The dimensions of the computational

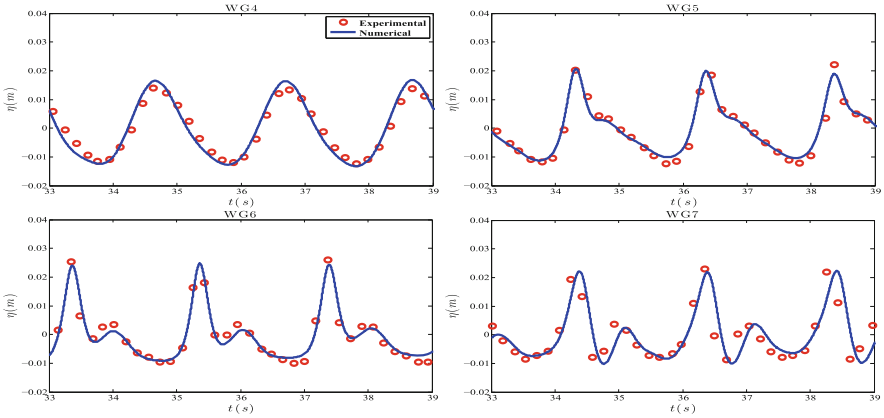


Fig. 6 Time series of surface elevation for periodic wave propagation over a submerged bar

domain were set to $(x, y) \in [-10, 30\,\text{m}] \times [0, 0.8\,\text{m}]$. A fine mesh of $N = 170,194$ nodes was used in $Pr = 10$ processors. Regular waves with $A = 0.01\,\text{m}$, and $T = 2.02\,\text{s}$ was generated at $x = 0\,\text{m}$. The free-surface elevations are recorded at WGs over and behind the bar as in the laboratory experiment, placed along the flume at $x = 10.5,\ 12.5,\ 13.5,\ 14.5\,\text{m}$. The time series of the free surface elevation at the wave gauges along the center-line are sown in Fig. 6. The generated waves propagate without changing their shape, until they reach the front slope where the waves shoal since nonlinear effects cause the waves to steepen. The wave amplitude grows and the surface profile becomes asymmetric. The back slope causes the waves to breakup into independent waves traveling at their own speed. The numerical results provide good agreement with the experimental data for WGs 4 and 5 and maintain relatively good agreement with the experimental data at WGs 6 and 7 over the crest and the lee-slope.

References

1. Berkhoff, J.C.E., Booy, N., Radder, A.C.: Verification of numerical wave propagation models for simple harmonic linear water waves. Coast. Eng. **6**, 255 (1982)
2. Brocchini, M.: A reasoned overview on Boussinesq-type models: the intreplay between physics, mathematics and numerics. Proc. R. Soc. **469**, 1–27 (2013)
3. Cai, X., Pederson, G.K., Langtagen, H.P.: A parallel multi-subdomain strategy for solving Boussinesq water wave equations. Adv. Water Resour. **28**, 215–233 (2005)
4. Kazolea, M., Delis, A.I.: A well-balanced shock-capturing hybrid finite volume-finite difference numerical scheme for extended 1D Boussinesq models. Appl. Numer. Math. **67**, 167–186 (2013)
5. Kazolea, M., Delis, A.I., Nikolos, I.K., Synolakis, C.E.: An unstructured finite volume numerical scheme for extended Boussinesq-type equations. Coast. Eng. **69**, 42–66 (2012)

6. Kazolea, M., Delis, A.I., Synolakis, C.E.: Numerical treatment of wave breaking on unstructured finite volume approximations for extended Boussinesq-type equations. J. Comput. Phys. **271**, 318–349 (2014)
7. Kazolea, M., Delis, A.I., Vavilis, P.S.: Parallel implementation of an unstructured finite volume solver for 2D BT equations. (in preparation)
8. Nwogu, O.: An alternative form of the boussinesq equations for nearshore wave propagation. J. Waterw. Port Coast. Ocean Eng. **119**, 618–638 (1994)
9. Ricchiuto, M., Filippini, A.: Upwind residual discretization of enhanced Boussinesq equations for wave propagation over complex bathymetries. J. Comput. Phys. **271**, 306–341 (2014)
10. Roe, P.L.: Approximate Riemann solvers, parameter vectors, and difference schemes. J. Comput. Phys. **43**, 357–372 (1981)
11. Roeber, V., Cheung, K.F.: Boussinesq-type model for energetic breaking waves in fringing reef environment. Coast. Eng. **70**, 1–20 (2012)
12. Shi, F., Kirby, J.T., Harris, J.C., Geiman, J.D., Grilli, S.T.: A high-order adaptive time-stepping TVD solver for Boussinesq modeling of breaking waves and coastal inundation. Ocean Model. **43–44**, 36–51 (2012)
13. Swigler, D.T.: Laboratory study of the three-dimensional turbulence and kinematic properties associated with a breaking solitary wave. Ph.D. thesis, Texas A&M University, College Station, Texas (2009)
14. Tissier, M., Bonneton, P., Marche, F., Chazel, F., Lannes, D.: A new approach to handle wave breaking in fully non-linear Boussinesq models. Coast. Eng. **67**, 54–66 (2012)
15. Tonelli, M., Petti, M.: Finite volume scheme for the solution of 2D extended Boussinesq equations in the surf zone. Ocean Eng. **37**, 567–582 (2010)
16. van Leer, B.: Towards the ultimate conservative difference scheme III. Upstream centered finite difference schemes for ideal compressible flow. J. Comput. Phys. **23**, 263–275 (1977)
17. Wei, G., Kirby, J.T., Sinha, A.: Generation of waves in Boussinesq models using a source function approach. Coast. Eng. **36**, 271 (1999)

On Devising Boussinesq-Type Equations with Bounded Eigenspectra: Two Horizontal Dimensions

Claes Eskilsson and Allan P. Engsig-Karup

Abstract Boussinesq-type equations are used to describe the propagation and transformation of free-surface waves in the nearshore region. The nonlinear and dispersive performance of the equations are determined by tunable parameters. Recently the authors presented conditions on the free parameters under which a Nwogu-type equations would yield bounded eigenspectra (Eskilsson and Engsig-Karup, J Comput Phys 271:261–280, 2014). This leads to a global conditional CFL time-step restriction which is shown to not be affected by the discretisation method and in this sense the CFL condition is *tamed* to impose a minimal constraint. In this paper we extend the previous study and provide numerical experiments which confirms the theoretical results also is valid in two horizontal dimensions.

Keywords Boussinesq-type equations • CFL condition • Free-surface waves

1 Introduction

In coastal engineering application the use of classical FEM for low-order spatial discretisation of wave equations has been used to describe complex geometries by using graded meshes. An immediate advantage of graded meshes is the ability to significantly decrease the total number of unknowns in the discretisation making it possible to improve the computational efficiency by spatially only resolving geometry or features of the solution where it is needed [11]. The use of flexible mesh discretisation methods is gaining more popularity since with this basis it becomes more straightforward to describe wave-structure interactions in complex settings (e.g. harbour and coastal areas), cf. recent review given in [1]. The downside for

C. Eskilsson (✉)
Department of Shipping and Marine Technology, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden
e-mail: claes.eskilsson@chalmers.se

A.P. Engsig-Karup
Department of Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
e-mail: apek@dtu.dk

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_77

most wave models is that the use of such graded meshes may impose more severe time-stepping restrictions due to global condition CFL stability constraints. From an algorithmic perspective remedies to this problem is found in local time-stepping algorithms, semi-implicit temporal integration, or the use of only low-order and less accurate simulation resulting from artificial dispersion and dissipation errors. High-order discretisation have the potential to address all of these downsides, however, is known to often have operator eigenspectra which grow fast and therefore CFL conditions often become to severe for practical use. Thus, state-of-the-art techniques such as the high-order spectral element/*hp* method (SEM) have still not caught much interest in coastal engineering applications [6] despite that it offer good and flexible opportunities for balancing work effort and accuracy.

In this work, we consider the extension [5] and device a Boussinesq-type model in two space dimensions, which is amendable to use high-order SEM and unstructured meshes. As in [2–4], we demonstrate that graded meshes does not pose a significantly challenge for an implicitly-implicit formulations for wave-structure problems since the global CFL condition is governed by bounded eigenspectra and in this sense can be tamed [13].

2 Boussinesq Equations

Consider the weakly nonlinear and dispersive Boussinesq-type equations due to Nwogu [10]. In the original form, the set of equations—with unbounded operator spectrum—can be stated in two horizontal dimensions as

$$\eta_t + \nabla \cdot (d\mathbf{u}) + \epsilon \nabla \cdot (\eta \mathbf{u}) + \mu^2 \Gamma_{20} = \mathcal{O}(\epsilon^2, \mu^2\epsilon, \mu^4), \tag{1a}$$

$$\mathbf{u}_t + \nabla \eta + \epsilon(\mathbf{u} \cdot \nabla)\mathbf{u} + \mu^2 \Delta_{20} = \mathcal{O}(\epsilon^2, \mu^2\epsilon, \mu^4), \tag{1b}$$

where the dispersive terms read

$$\Gamma_{20} = \nabla \cdot [a_1 d^3 \nabla(\nabla \cdot \tilde{\mathbf{u}} + a_2 d^2 \nabla(\nabla \cdot (d\mathbf{u}))], \tag{2a}$$

$$\Delta_{20} = b_1 d^2 \nabla(\nabla \cdot \mathbf{u}_t) + b_2 d \nabla(\nabla \cdot (d\tilde{\mathbf{u}}_t)). \tag{2b}$$

Here $\eta(\mathbf{x}, t)$ is the free surface elevation, $\mathbf{u}(\mathbf{x}, t)$ is the horizontal velocity at the reference level z_α and $d(\mathbf{x})$ is the still water depth. The constants (a_1, a_2, b_1, b_2) govern the dispersive properties of the equations.

Using the enhancement technique [9] additional free parameters are introduced to the equations. This approach exploits that the Boussinesq-type equations are derived by a truncation procedure, and as long as modification are on the size of the magnitude of the truncation errors, the approach is feasible. The resulting equations

read [9]:

$$\Gamma_{20} = \nabla \cdot \left[a_1 d^3 \nabla (\nabla \cdot \mathbf{u}) + (a_2 - \beta_1) d^2 \nabla (\nabla \cdot (d\mathbf{u})) - \beta_1 \left(d^2 \nabla \eta_t \right) \right], \tag{3a}$$

$$\Delta_{20} = (b_1 - \alpha_1) d^2 \nabla (\nabla \cdot \mathbf{u}_t) + b_2 d \nabla (\nabla \cdot (d\mathbf{u}_t)) - \alpha_1 d^2 \nabla \left(\nabla^2 \eta \right), \tag{3b}$$

where an optimum choice—with regard to linear dispersion characteristics—of the free parameters is $(b_1 + b_2, \alpha_1, \beta_1) = (-0.395, 0.011, 0.039)$. The equations exhibit an unbounded eigenspectrum and we will hence refer to this setting as unbounded Boussinesq equations (UBE). However, as shown in [5], under certain conditions on the free parameters this set of equations will exhibit a bounded eigenspectrum. This observation was numerically supported by simulations in one horizontal dimension. Using the setting $(b_1 + b_2, \alpha_1, \beta_1) = (-0.400, 0, 0.015)$ gives a bounded eigenspectrum while retaining good dispersion characteristics. We will refer to this setting as bounded Boussinesq equations (BBE).

3 Numerical Discretization

To develop a high-order spectral/*hp* element method [8] that works on unstructured meshes we use the Method of Lines, where a semi-discrete system of equations is formed by spatial discretisation using a Galerkin method.

We introduce approximate solutions, e.g., u_δ to u, where $u_\delta \in \chi$ with $\chi = \{u | u \in H^1(\Omega), u(\mathbf{x}) = g(\mathbf{x}), \mathbf{x} \in \partial\Omega\}$ is the trial space and $\mathcal{V} = \{v | v \in H^1(\Omega), v(\mathbf{x}) = 0, \mathbf{x} \in \partial\Omega\}$ is the set of all test functions.

We partition the domain $\Omega_h \subset \Omega$ to obtain a tessellation $\mathcal{T}_h$ which consists of N_{el} non-overlapping elements $\mathcal{T}_e$ such that $\cup_{e=1}^{N_{el}} \mathcal{T}_e = \mathcal{T}_h$. We approximate the solutions $(\eta, u, v, w, q) \in \mathcal{V} \subset C^0(\Omega)$ with piece-wise continuous P'th order polynomial approximations $(\eta_\delta, u_\delta, v_\delta, w_\delta, q_\delta) \in \mathcal{V}_\delta$ where the discrete space is given as $\mathcal{V}_\delta = \{v \in H^1(\mathcal{T}_h) : v|_{\mathcal{T}_e} \in \mathcal{P}^P(\mathcal{T}_e) \in \mathcal{T}_h\}$. Each element $\mathcal{T}_e$ is filled with N_P local nodes.

In two space dimensions, we can represent the approximate solutions in the form of a nodal expansion

$$\eta_\delta(\mathbf{x}, t) = \sum_{e=1}^{N_{el}} \sum_{n=1}^{N_P} \hat{\eta}_n^e(t) l_n(\chi_e^{-1}(\mathbf{x})), \quad (\mathbf{x}) \in \mathcal{T}_h, \quad t \geq 0 \tag{4}$$

in which $\hat{\eta}_n^e(t)$ are local expansion coefficients and $l_n(\mathbf{x})$ is the multivariate n'th Lagrange polynomial with cardinal property $l_n(\mathbf{x}_i) = \delta_{ni}$ defined from the set of unique vertices $(\mathbf{x}_i) = \chi_e(\mathbf{r}_i)$ defining the local nodes on element e. $\chi_e : \mathbf{r} \to \mathbf{x}$ is a local affine co-ordinate mapping from a standard element $\mathcal{T}_{st}$ to an elemental region $\mathcal{T}_e = \{x | x \in \chi_e(\mathbf{r})\}$. For a straight-sided reference triangle, we have $\mathcal{T}_{st} = \{\mathbf{r} = (r, s) | (r, s) \geq -1; r + s \geq 0\}$. On the standard triangular element one can use

the symmetric and optimized node distribution generated by an explicit warp and blend procedure described in [12]. In a similar way, for a straight-sided reference quadrilateral, we have $\mathcal{Q}_{st} = \{\mathbf{r} = (r,s) | -1 \leq (r,s) \leq 1\}$ and we use the nodes defined by a tensor product of Gauss-Lobatto-Legendre points in the two directions.

For brevity we present the method for the constant depth case. The weak formulation of (1) and (3) can be stated as

$$\iint_{\mathcal{T}_h} v_\delta [\left(1 - \mu^2 \beta_1 d^2 \nabla^2\right) \eta_t + \nabla \cdot ((d + \epsilon\eta)\,\mathbf{u}) + \mu^2 \left(\alpha + \frac{1}{3} - \beta_1\right) d^3 \nabla^2 w] d\mathbf{x} = 0, \tag{5a}$$

$$\iint_{\mathcal{T}_h} v_\delta [\mathbf{u}_t + \mu^2 (\alpha - \alpha_1)\, d^2 \nabla w_t + \nabla\eta + \epsilon(\mathbf{u}\cdot\nabla)\mathbf{u} - \mu^2 \nabla q] d\mathbf{x} = 0 \tag{5b}$$

where $\alpha = b_1 + b_2$ and having neglected all $\mathcal{O}(\epsilon^2, \mu^2\epsilon, \mu^4)$-terms. Further, we have introduced the auxiliary variables $w = \nabla \cdot \mathbf{u}$ and $q = \alpha_1 d^2 (\nabla^2 \eta)$ to resolve the third-order spatial derivatives. Applying the Divergence theorem gives

$$\iint_{\mathcal{T}_h} [\left(v_\delta + \mu^2 \beta_1 d^2 \nabla v_\delta \nabla\right) \eta_t - \nabla v_\delta \cdot ((d + \epsilon\eta)\,\mathbf{u}) - \mu^2 \left(\alpha + \frac{1}{3} - \beta_1\right) d^3 \nabla v_\delta \nabla w] d\mathbf{x} = 0, \tag{6a}$$

$$\iint_{\mathcal{T}_h} [v_\delta \mathbf{u}_t + \mu^2 (\alpha - \alpha_1)\, d^2 v_\delta \nabla w_t + v_\delta \nabla\eta + \epsilon v_\delta (\mathbf{u}\cdot\nabla)\mathbf{u} - \mu^2 v_\delta \nabla q] d\mathbf{x} = 0 \tag{6b}$$

and

$$\iint_{\mathcal{T}_h} v_\delta w\, d\mathbf{x} = \iint_{\mathcal{T}_h} [-(\nabla v_\delta \cdot \mathbf{u})]\, d\mathbf{x}, \tag{7a}$$

$$\iint_{\mathcal{T}_h} v_\delta q\, d\mathbf{x} = \iint_{\mathcal{T}_h} v_\delta \left[-\alpha_1 d^2 (\nabla v_\delta \cdot \nabla\eta)\right] d\mathbf{x}, \tag{7b}$$

where all arising boundary terms have been omitted as we in this study are only concerned with periodic and impermeable vertical wall boundaries. For periodic domains the boundary terms naturally cancels out. For the slip wall condition the

zero net flux conditions in the normal direction of the walls corresponds to

$$\mathbf{u} \cdot \mathbf{n} = 0, \quad \nabla \eta \cdot \mathbf{n} = 0 \quad \nabla w \cdot \mathbf{n} = 0, \quad \mathbf{x} \in \partial \Omega, \tag{8}$$

giving that all arising boundary terms are zero.

In the simulations presented we integrate in time using the explicit third-order Adams-Bashforth scheme. The resulting linear system is solved with GMRES with an ILU preconditioner. We note that it is well-known that equal-order simulations give rise to stability problems. However, rather than using different order for the free surface and the velocity variables we here apply a weak exponential filter [7] on interior bubble modes, if needed, in order to stabilize the solution.

4 Eigenvalue Analysis

Writing the equations in semi-discrete form $\mathbf{A}\partial_t \mathbf{U} = \mathbf{B}\mathbf{U}$ we are interested in the eigenvalues, λ_i, of the operator $\mathbf{A}^{-1}\mathbf{B}$, in order to understand the stability of the scheme. In Fig. 1 we present numerically obtained maximum eigenvalues as function of the polynomial order. The eigenvalues are purely imaginary and, as expected, the BBE equations remain bounded while the UBE equations grows as P^2. It is stressed that the magnitude of the bounded eigenvalues do not depend on the numerical discretization, but is a property of the governing equations. This is illustrated in Fig. 1 where the magnitude of the eigenvalue for the BBE is the same on both periodic structured quadrilaterals and unstructured triangles with wall boundaries. Please note that the computed eigenvalues for the BBE coincide with the analytic value given in [5].

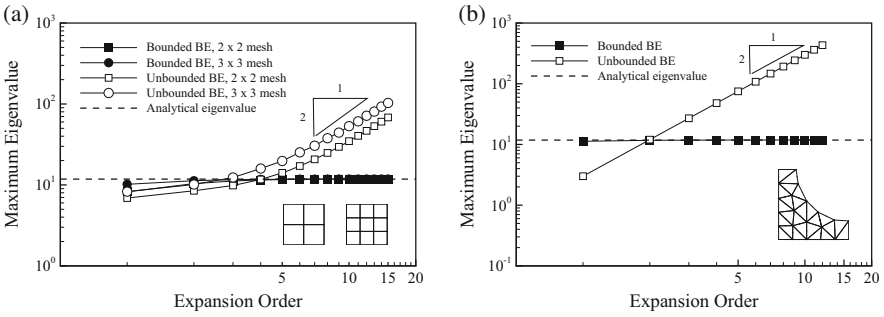


Fig. 1 Numerically observed eigenvalues. (**a**) Structured quadrilateral elements with periodic boundaries and (**b**) unstructured triangular elements with wall boundaries

5 Numerical Example

A frequently used test case for Boussinesq-type equations is the run-up of a solitary wave on a cylinder. This case highlights the typical situation where the geometry are represented with sufficient accuracy by generating an unstructured mesh with a significantly higher mesh density around structures and topological features.

The size of the wave tank is $[x, y] \in [-33, 25] \times [-19.2, 19.2]\,\mathrm{m}$ with a depth of 1 m. We utilize the symmetry of the set-up and simulate only the lower half of the domain. All boundaries are treated as wall/symmetry boundaries. The computational domain is decomposed into two different meshes with (a) 269 and (b) 116 triangular elements, see Fig. 2. Both meshes resolve the cylinder with approximately the same number of boundary edges.

The initial condition is given by Laitone's first order solitary wave solution, centered at $x = -17\,\mathrm{m}$, and we integrate the solution for 12.5 s. Figure 3 shows

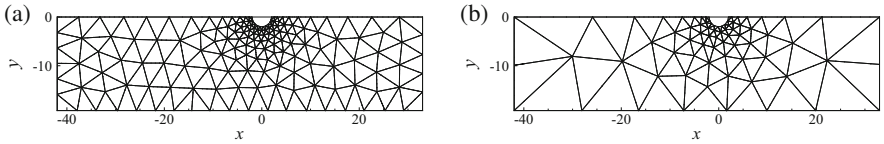


Fig. 2 Illustration of used meshes: (**a**) 269 elements (Mesh A) and (**b**) 116 elements (Mesh B)

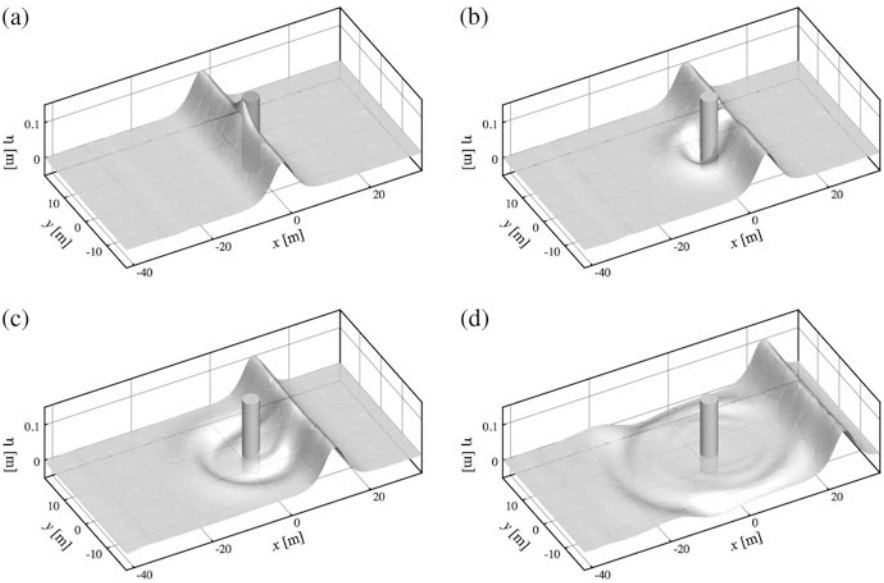


Fig. 3 Solitary wave impinging on a vertical cylinder. Simulation using mesh A with $P = 6$. At time: (**a**) 4.5 s, (**b**) 6.5 s, (**c**) 8.5 s and (**d**) 12.5 s

Table 1 Maximum allowed time step (in seconds) for a solitary wave impinging on a cylinder

		BBE	UBE	Rel. diff.
Mesh A	$P = 4$	6.25E−02	5.21E−02	1.20
	$P = 5$	6.25E−02	4.03E−02	1.55
	$P = 6$	6.25E−02	3.21E−02	1.95
Mesh B	$P = 6$	6.25E−02	3.68E−02	1.70
	$P = 7$	6.25E−02	3.05E−02	2.05
	$P = 8$	6.25E−02	2.50E−02	2.50

the simulated run-up and subsequent scattering of the solitary wave. For mesh A we need $P \geq 5$ and for mesh B we need $P \geq 7$ in order to get acceptable results.

In Table 1 we present the influence of the polynomial order on the maximum allowed time step. Again, a novel feature is that the maximum stable time step for BBE simulations are *not* dependent on the choice of h and p. Thus it is possible to have large differences in element size as illustrated by mesh B. As a result, from Table 1 it can be seen that for practical computations it is possible to obtain significant speed up by taking advantage of high-order elements compared to standard BE settings without compromising accuracy.

6 Conclusion

We have analysed and demonstrated that an implicitly-implicit formulation of a Nwogu-type equation in two spatial dimensions can be made accurate and efficient without compromising robustness by using unstructured high-order spectral element/*hp* methods. This work opens the road towards robust mesh-adaptive solutions considered in ongoing work.

References

1. Brocchini, M.: A reasoned overview on Boussinesq-type models: the interplay between physics, mathematics and numerics. Proc. R. Soc. Lond. A **469**(2160), 20130496 (2013)
2. Engsig-Karup, A.P.: Unstructured nodal DG-FEM solution of high-order Boussinesq-type equations. Ph.D. thesis, Department of Mechanical Engineering, Technical University of Denmark (2006)
3. Engsig-Karup, A., Hesthaven, J., Bingham, H., Madsen, P.: Nodal DG-FEM solutions of high-order Boussinesq-type equations. J. Eng. Math. **56**, 351–370 (2006)
4. Engsig-Karup, A., Hesthaven, J., Bingham, H., Warburton, T.: DG-FEM solution for nonlinear wave-structure interaction using Boussinesq-type equations. Coast. Eng. **55**, 197–208 (2008)
5. Eskilsson, C., Engsig-Karup, A.P.: On devising Boussinesq-type models with bounded eigenspectra: one horizontal dimension. J. Comput. Phys. **271**, 261–280 (2014)

6. Eskilsson, C., Engsig-Karup, A.P., Sherwin, S.J., Hesthaven, J.S., Bergdahl, L.: The next step in coastal numerical models: spectral/hp element methods? In: Proceedings of the WAVES2005 Conference, Madrid, 7–12 (2005)
7. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications. Springer, Berlin (2008)
8. Karniadakis, G., Sherwin, S.: Spectral/hp Element Methods for Computational Fluid Dynamics, 2nd edn. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford (2005)
9. Madsen, P., Schäffer, H.A.: Higher order Boussinesq-type equation for surface gravity waves—derivation and analysis. Proc. R. Soc. Lond. A **356**, 3123–3181 (1998)
10. Nwogu, O.: Alternative form of Boussinesq equations for nearshore wave propagation. J. Waterw. Port Coast. Ocean Eng. ASCE **6**(119), 618–638 (1993)
11. Sørensen, O.R., Schäffer, H.A., Sørensen, L.S.: Boussinesq-type modelling using an unstructured finite element technique. Coast. Eng. **50**, 181–198 (2004)
12. Warburton, T.: An explicit construction for interpolation nodes on the simplex. J. Eng. Math. **56**(3), 247–262 (2006)
13. Warburton, T., Hagstrom, T.: Taming the CFL number for discontinuous Galerkin methods on structured meshes. SIAM J. Numer. Anal. **46**(6), 3151–3180 (2008)

On Nonlinear Shoaling Properties of Enhanced Boussinesq Models

A.G. Filippini, S. Bellec, M. Colin, and M. Ricchiuto

Abstract In this paper, we investigate the nonlinear properties of Boussinesq models. In particular, we consider the wave shoaling obtained in physical regimes which go from linear to weakly nonlinear, to the wave breaking limit. For a given asymptotic accuracy in terms of dispersion and nonlinearity, we consider two families of models: the first depending on derivatives of the velocity, the second on derivatives of the volume flux. We show that, while linear dispersion and linear shoaling characteristics are strongly dependent on the type of dispersive terms introduced, when approaching the nonlinear regime the only influencing factor is whether the model is in amplitude-velocity of amplitude-flux form. We investigate these two alternative formulations of several known models, and propose a new model with a compact differential form, and the same linear characteristics of the model of Nwogu. The nonlinear shoaling properties of the models are investigated numerically showing that inside one given family, all the models have almost identical behaviour.

Keywords Boussinesq-type models • Wave shoaling

1 Introduction

This paper deals with Boussinesq-Type (BT) models for wave propagation and transformation. In the near shore region one has to deal with both nonlinear and dispersive effects which make the task of accurate modelling very difficult. Accounting for genuinely nonlinear effects is a research topics of high priority [4]. The simplest depth averaged model, the nonlinear shallow water equations system (NLSW), while capable of describing the energy dissipation in breaking regions [14], does not account for wave dispersion. In this work, we consider weakly

A.G. Filippini (✉) • S. Bellec • M. Ricchiuto
Inria, 200 av. de la Vieille Tour, 33405 Talence, France
e-mail: andrea.filippini@inria.fr; stevan.bellec@inria.fr; mario.ricchiuto@inria.fr

M. Colin
IPB, 351 cours de la libération, 33405 Talence Cedex, France
e-mail: mcolin@math.u-bordeaux.fr

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_78

nonlinear and dispersive BT models obtained by adding linear differential terms to the NLSW system. These terms account for non hydrostatic effects, they improve the linear frequency dispersion, however, they do not include any dissipative effects. These effects can be recovered by locally reverting to the NLSW equations in properly detected regions, or by explicitly adding eddy viscosity terms [4, 9, 14]. In this case, the BT equations are required to accurately predict wave shapes and amplitudes to allow the breaking process to be triggered at the right time and place. In particular, they should provide an accurate description of the shoaling process. Differently from linear dispersion and shoaling, which can be studied analytically [5], nonlinear shoaling can only be investigated numerically. There exist several types of BT models with different linear dispersion and shoaling properties. For a given linear dispersion relation, and within the same asymptotics in terms of the nonlinearity $\epsilon = a/d$ and dispersion $\sigma = d/\lambda$ parameters (a the wave amplitude, d the mean water level, λ the wavelength), one can find at least two different models. These two systems of PDEs differ in the fact that the dispersive terms contain either derivatives of the velocity, or of the volume flux; thus we refer to them as to models in *wave amplitude-velocity* or *wave amplitude-volume flux* form.

The aim of this paper is to assess the nonlinear shoaling properties of BT models which have the same linear properties, but different non-linear PDE structure. We consider in our analysis the models of Peregrine (P) [12], Beji-Nadaoka (BN) [3], Madsen-Sørensen (MS) [10] and Nwogu (N) [11]. For all of them, we manipulate the differential equations adding terms which, keeping the same asymptotic accuracy, allow to obtain a *wave amplitude-velocity* model, for those systems written in *wave amplitude-volume flux* form (e.g. the MS model), and vice versa for models in *wave amplitude-velocity* form (e.g. the P, BN, N models). We show that, when approaching the nonlinear regime, only the nonlinear structure of the PDE has an influence on the shoaling effects.

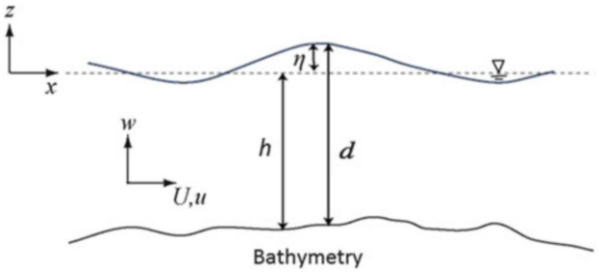


Fig. 1 Sketch of the free surface flow problem, main parameters description

2 Presentation of the Models

The most common BT model is perhaps the one of Peregrine [12]. Starting from the incompressible Euler equations (see Fig. 1), considering asymptotic expansions in terms of the nonlinearity and dispersion parameters, ϵ and σ, and depth averaging the resulting expressions, one can show that for $\epsilon = \mathcal{O}(\sigma^2)$

$$\begin{cases} \eta_t + [(h+\epsilon\eta)u]_x = 0 \\ u_t + \epsilon u u_x + \eta_x + \sigma^2 \left(\dfrac{h^2}{6} u_{xxt} - \dfrac{h}{2} [hu]_{xxt} \right) = \mathcal{O}(\epsilon\sigma^2, \sigma^4) \end{cases} \tag{1}$$

We refer to the Peregrine (P) model as the dimensional version of (1). Denoting the NLSW flux by $F^{SW} = uq + gd^2/2$, for the P model the volume flux $q = du$ verifies

$$q_t + F^{SW}_x - gdh_x + dP_t = 0\,, \quad P = P(u) = \frac{h^2}{6} u_{xx} - \frac{h}{2}[hu]_{xx}\,, \tag{2}$$

Assuming $h_t = 0$, and since in non-dimensional form $d = h + \epsilon\eta$, we have

$$(h+\epsilon\eta)\sigma^2 \left(\frac{h^2}{6} u_{xxt} - \frac{h}{2}(hu)_{xxt} \right) = \sigma^2 \left(\frac{h^3}{6} \left(\frac{q}{h} \right)_{xxt} - \frac{h^2}{2} q_{xxt} \right) + \mathcal{O}(\epsilon\sigma^2, \sigma^4)\,.$$

Thus, in the same asymptotics, we can replace (2) by

$$Q_t + F^{SW}_x - gdh_x = 0\,, \quad Q = Q(q) = q + \frac{h^3}{6} \left(\frac{q}{h} \right)_{xx} - \frac{h^2}{2} q_{xx} \tag{3}$$

leading to the model presented in Abbott (A) [1]. Even if the P and A models are identical in the linear limit, they are substantially different in the nonlinear case. In particular, these are not the same PDEs written in terms of different unknowns; they actually include different differential terms. The difference between these terms allows to express the dispersive operators in terms of q instead of in terms of u; in such sense the P and A model represent respectively the *amplitude-velocity* and the *amplitude-flux* form of the same linear dispersion relation.

As for the P and the A systems, two set of PDEs exist for a given couple linear dispersion relation-linear shoaling parameter. All dispersion enhanced BT models admit a *amplitude-velocity* form, and an *amplitude-flux* equivalent. Details on the derivation are given e.g. in [6] and in the second volume of [5], and are left out due to space limitations. Here we apply this theory to four linear dispersion relations, corresponding to the P model above and to the enhanced models of BN, MS and N. In total 8 models are considered. Three of them are new variants of existing models, so we speak of the *amplitude-flux* form of the BN model as the Beji-Nadaoka-Abbott (BNA) model, of the *amplitude-velocity* form of the MS model as the Madsen-Sørensen-Peregrine (MSP) model, and of the *amplitude-flux* form of the N model as

Nwogu-Abbott (NA) model. They can be generally recast as

$$\begin{cases} h_t + \mathbf{K}_x = 0 \\ \mathbf{Q}_t + F_x^{SW} - ghd_x + h\mathbf{P}_t + g\mathbf{R} + u\boldsymbol{\Psi} = 0 \end{cases} . \tag{4}$$

For the definition of the differential operators **K**, **Q**, **P**, **R** and $\boldsymbol{\Psi}$ for each specific model we refer to [6] and [3, 10, 11]. Here we observe that for models P, BN, MSP and N, written in *wave amplitude-velocity* form, $\mathbf{Q} = q$ and **P** is structurally similar to P(u) in (2); instead for models A, BNA, MS and NA, written in *wave amplitude-volume flux* form, **Q** is an elliptic operator structurally similar to Q(q) in (3), and $\mathbf{P} = 0$. Only for the BN and MS models $\mathbf{R} \neq 0$, in particular **R** has the structure of $P(\eta)$; likewise, $\boldsymbol{\Psi} \neq 0$ only for the N model, assuming the structure of P(u). Finally, $\mathbf{K} \neq q$ only for the N model, with $\mathbf{K} = q + dP(u)$ in the *wave amplitude-velocity* case, while $\mathbf{K} = Q(q)$ in the *wave amplitude-volume flux* case. For each model, we will consider the standard dispersion/shoaling optimised values of the constants. Note that, among the 8 BT models taken into account, two of them present the following very simple structure:

$$\begin{cases} h_t + \mathbf{K}_x = 0 \\ \mathbf{Q}_t + F_x^{SW} - ghd_x = 0 \end{cases} \tag{5}$$

The first is the A system defined by (3) and $\mathbf{K} = q$, the other one is the NA model taking in (5) $\mathbf{K} = q + A_1 h^2 q_{xx} + A_2 h^3 (q/h)_{xx}$ and $\mathbf{Q} = q + B_1 h^2 q_{xx} + B_2 h^3 (q/h)_{xx}$ [6]. This system is the only enhanced BT model we know of sharing a compact structure very close to that of the NLSW equations.

3 Numerical Experiments: Shoaling Tests

The numerical tests discussed hereafter have been repeated with two different discretizations and on several meshes, to ensure scheme and mesh independent results. The scheme used are the finite difference scheme proposed by Wei and Kirby, which discretizes the shallow water terms using fourth-order formulas and the dispersive terms to second order accuracy [15], and a P^1 continuous finite element method based on a standard Galerkin solution of the elliptic sub-problems defining **K**, **Q**, **P**, etc, plus a Galerkin projection for the first order PDEs (4) or (5). This procedure has been recently used in [13] for the MS equations, and shown analytically and numerically to have accuracy close to a fourth order finite difference scheme, and to that of the scheme of [15]. As in [13] high order implicit time integration is used to allow the choice of the time step based on physical arguments. In all the tests the two discretizations have given virtually indistinguishable results. Due to space limitations, we will not show this comparison, but only report the main findings.

3.1 *Linear Shoaling Test*

As discussed in Sect. 2, we consider BT equations which, in couples, reduce to the same four linearized systems. In particular the P together with the A model, the BN together with the BNA model, the MSP together with the MS model and the N together with the NA model degenerate to the same linear systems. We refer the interested reader to the references given in Sect. 2. These systems, thus, should manifest the same linear dispersion and linear shoaling behaviour.

To verify that our implementation correctly reproduces the linear shoaling, and in particular that indeed different models collapse onto one another for small amplitude waves, we perform an experiment proposed in [10] : a monochromatic wave of amplitude $a = 0.05\,\mathrm{m}$ and period $T = 4\,\mathrm{s}$, propagating over a depth $h_0 = 13\,\mathrm{m}$, and shoaling over a slope of 1:50 starting 50 m from the inlet. These data give a nonlinearity parameter $\epsilon \in [0.0038, 0.25]$, which is in the linear range. The generation of the periodic signals is performed adding a source to the mass equation (see [13] for details on this aspect, and [10] for further details on this test).

The results of the test are summarized on Fig. 2 where we have reported the distribution of the wave amplitude for all the models. It can be observed that the implemented schemes well reproduce the theoretical linear behaviors expected for all models. In fact Fig. 2 shows that the schemes of N and NA give nearly identical results, as well as the models MSP and MS and the models BN and BNA do.

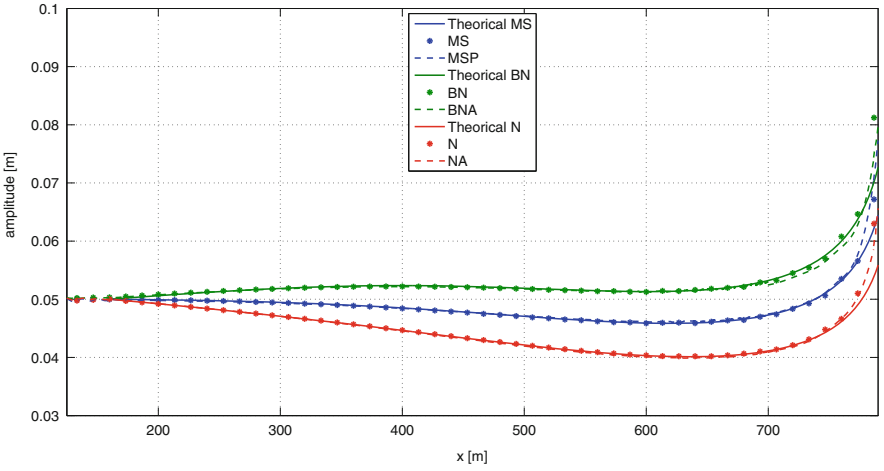


Fig. 2 Linear shoaling of a periodic wave from deep to shallow water: envelope of the maximum elevations computed by the several models and comparison w.r.t. the theoretical results

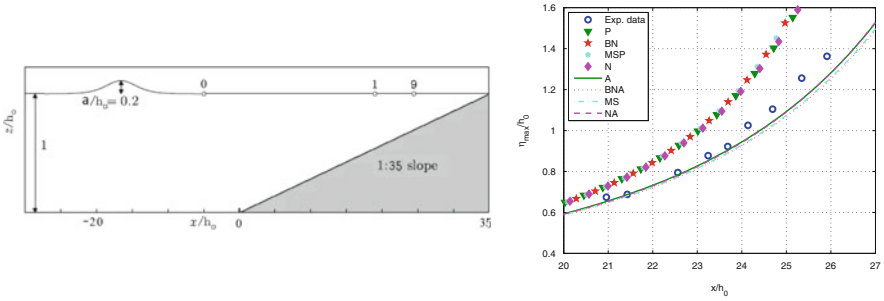


Fig. 3 Shoaling of a solitary wave. *Left*: computational set-up. *Right*: comparison between computed and measured maximum value of the relative wave height

3.2 Nonlinear Shoaling Test

We compare now the nonlinear properties of the models on the shoaling test of Grilli et al. in [7] : a solitary wave of amplitude $A/h_0 = 0.2$ m propagating on a depth $h_0 = 0.44$ m, and shoaling on a slope of 1:35. In this test $\epsilon = a_0/h \in [0.45, 1.7]$ which is in the nonlinear range $\epsilon \geq 1$. A sketch of the test is given on the left of Fig. 3, also showing the position of the gauges where wave height measurements are available.

We present the results on Figs. 3, 4, and 5, in terms of evolution of the wave maximum in space, and of temporal evolution of the wave height in the gauges. In both cases, the experimental data of [7] are reported as well. Looking at the figures we see clearly that only two main behaviors are observed. All the models derived in terms of *amplitude-velocity* (Fig. 5) quickly over-shoal. The evolution of the peak height is quite independent of the linear dispersion relation of the models which all give almost the same curves. The same is true for the models derived in terms of *amplitude-volume flux* (Fig. 4), which however give a shoaling height closer to the experimental ones.

In particular, it is reported in [7] than the breaking point is gauge 9. We can see from Fig. 4, and 5 that models obtained in terms of velocity all have already given higher waves already in gauge 5, while the models obtained in terms of volume flux under-shoal and never reach this height. This means that very early breaking will be likely to occur if one uses one family of models, while late or no breaking will be observed with the others. Note that here no breaking modeling is considered, so velocity based models keep on shoaling giving very tall and steep waves in the last gauges. The same figures show that the steep front of the waves are generally better described by models with dispersive terms written in terms of the velocity, while the tail of the wave is much better approximated by models in volume flux form.

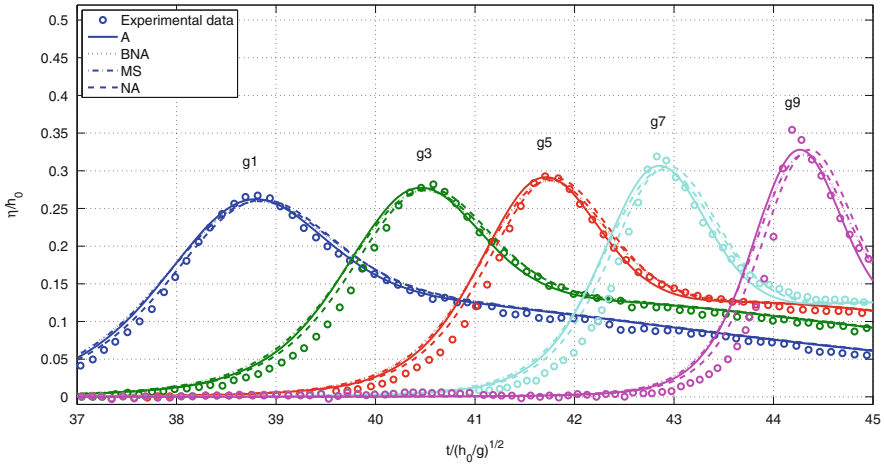


Fig. 4 Shoaling of a solitary wave up to breaking: comparison between computated and measured data for relative wave height at gauges 1, 3, 5, 7 and 9 and for the BT model of Abbott, Madsen and Sørensen, and Nwogu-Abbott

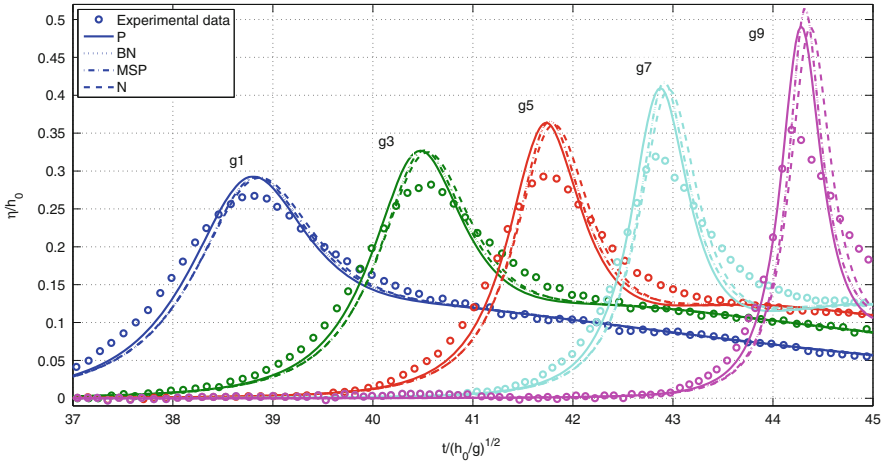


Fig. 5 Shoaling of a solitary wave up to breaking: comparison between computated and measured data for relative wave height at gauges 1, 3, 5, 7 and 9 and for the BT model of Peregrine, Nwogu, and Beji and Nadaoka

4 Conclusions and Perspectives

We have discussed the nonlinear behaviour of weakly nonlinear Boussinesq models. Under the same asymptotic accuracy, a linear dispersion relation/shoaling coefficient define at least two models written either in terms of *velocity* or *volume flux* derivatives. This has allowed in this paper to reformulate existing models, propose new ones, and studied their linear and nonlinear properties. The results show that in the nonlinear case only the *amplitude-velocity* form or *amplitude-volume flux* form counts: models with different linear shoaling and dispersion relations give practically the same results.

Future works will involve the study of the coupling of these models with wave breaking criteria to assess the impact of these properties on the behavior of breaking models, and the extension of this study to genuinely nonlinear equations and to the set of equations recently developed in [2] and [8] which couples the dispersive properties of BT models and the dissipative features of the NLSW equations.

Acknowledgements Work partially funded by the TANDEM contract, reference ANR-11-RSNR-499 0023-01 of the French Programme Investissements d'Avenir.

References

1. Abbott, M.B., Petersen, H.M., Skovgaard, O.: Computations of short waves in shallow water. In: Proceedings of 16th Conference on Coastal Engineering, Coastal Engineering, pp. 414–433 (1978)
2. Antuono, M., Liapidevskii, V., Brocchini, M.: Dispersive nonlinear shallow water equations. Stud. Appl. Math. **122**(1), 1–28 (2009)
3. Beji, S., Nadaoka, K.: A formal derivation and numerical modeling of the improved Boussinesq equations for varying depth. Ocean Eng. **23**, 691–704 (1996)
4. Brocchini, M.: A reasoned overview on Boussinesq-type models: the interplay between physics, mathematics and numerics. Proc. R. Soc. Lond. A **469**(2160), 20130496 (2014)
5. Dingemans, M.W.: Water Wave Propagation over Uneven Bottoms. Advanced Series Ocean Engineering. World Scientific, Singapore (1997)
6. Filippini, A.G., Bellec, S., Colin, M., Ricchiuto, M.: On the nonlinear behavior of Boussinesq type models: Amplitude-velocity vs wave amplitude-flux forms. Coast. Eng. **99**, 109–123 (2015)
7. Grilli, S.T., Subramanya, R., Svendsen, I.A., Veeramony, J.: Shoaling of solitary waves on plane beaches. J. Waterw. Port. C. ASCE **120**, 609–628 (1994)
8. Grosso, G., Antuono, M., Brocchini, M.: Dispersive nonlinear shallow water equations: some numerical results. J. Eng. Math. **67**(1–2), 71–84 (2010)
9. Kazolea, M., Delis, A.I., Synolakis, C.: Numerical treatment of wave breaking on unstructured finite volume approximations for extended Boussinesq type equations. J. Comput. Phys. **271**, 281–305 (2014)
10. Madsen, P.A., Sorensen, O.R.: A new form of the Boussinesq equations with improved dispersion characteristics. Part 2: a slowing varying bathymetry. Coast. Eng. **18**, 183–204 (1992)

11. Nwogu, O.: An alternative form of the Boussinesq equations for near-shore wave propagation. J. Waterw. Port. C. ASCE **119**, 618–638 (1994)
12. Peregrine, D.H.: Long waves on a beach. J. Fluid. Mech. **27**, 815–827 (1967)
13. Ricchiuto, M., Filippini, A.G.: Upwind residual discretization of enhanced Boussinesq equations for wave propagation over complex bathymetries. J. Comput. Phys. **271**, 306–341 (2014)
14. Tonelli, M., Petti, M.: Simulation of wave breaking over complex bathymetries by a Boussinesq model. J. Hydraulic Res. **49**, 473–486 (2011)
15. Wei, G., Kirby, J.T.: A time-dependent numerical code for extended Boussinesq equations. J. Waterw. Port. C. ASCE **120**, 251–261 (1995)

MS 24
MINISYMPOSIUM: NUMERICAL METHODS IN VOLCANO GEOPHYSICS

Organizers

Gilda Currenti[1] and Eugenio Sansosti[2]

Speakers

Sebastien Court[3], Olivier Bodart[4], Valerie Cayol[5] and Jonas Koko[6]
Fictitious Domain Methods for Fracture Models in Elasticity

Armando Coco[7], Gilda Currenti[1], Ciro Del Negro[8], Joachim Gottsmann[9] and Giovanni Russo[10]
Geophysical Changes in Hydrothermal-Volcanic Areas: A Finite-Difference Ghost-Point Method to Solve Thermo-Poroelastic Equations

[1]Gilda Currenti, Istituto Nazionale di Geofisica e Vulcanologia, Catania, Italy.
[2]Eugenio Sansosti, Istituto per il Rilevamento Elettromagnetico dell'Ambiente, CNR, Napoli, Italy.
[3]Sebastien Court, Universite Blaise Pascal, France.
[4]Olivier Bodart, Universite Blaise Pascal, France.
[5]Valerie Cayol, Universite Blaise Pascal, France.
[6]Jonas Koko, Universite Blaise Pascal, France.
[7]Armando Coco, University of Bristol, UK.
[8]Ciro Del Negro, Istituto Nazionale di Geofisica e Vulcanologia, Catania, Italy.
[9]Joachim Gottsmann, University of Bristol, UK.
[10]Giovanni Russo, Università di Catania, Italy.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_79

Giuseppe Nicosia[11], Piero Conca[12], Jole Costanza[13], Giovanni Carapezza[14], Gilda Currenti[1], Ciro Del Negro[8]
Multi-Objective Optimization, Sensitivity and Robustness Analyses for the Inverse Modeling of Ground Deformation and Gravity Changes of the 1981 Etna Eruption

Antonio Troiano[15], Maria Giulia Di Giuseppe[16], Alessandro Fedele[17], Renato Somma[18], Claudia Troise[19] and Giuseppe De Natale[20]
Numerical Simulation Applied to the Solfatara-Pisciarelli Shallow Hydrothermal System

Keywords

Finite differences
Fluid flow simulation
Volcano geophysics

Short Description

Volcanology is being evolved through quantitative approaches employed to investigate volcanoes, understand their dynamics, and forecast their hazards. The increasing use of remote sensing technologies has enhanced our ability to detect and track different volcanic processes accompanying the ascent of magma from source to surface, including hydrothermal activity, magma intrusion, conduit flow dynamics, pyroclastic flows and ash dispersal. The purpose of the minisymposium is to present the state-of-the-art in the modeling-based assessment of satellite remote sensing observations applied to open questions and problems in volcano geophysics.

[11]Giuseppe Nicosia, Università di Catania, Italy.
[12]Piero Conca, Università di Catania, Italy.
[13]Jole Costanza, Università di Catania, Italy.
[14]Giovanni Carapezza, Università di Catania, Italy.
[15]Antonio Troiano, Istituto Nazionale di Geofisica e Vulcanologia, Napoli, Italy.
[16]Maria Giulia Di Giuseppe, Istituto Nazionale di Geofisica e Vulcanologia, Napoli, Italy.
[17]Alessandro Fedele, Istituto Nazionale di Geofisica e Vulcanologia, Napoli, Italy.
[18]Renato Somma, Istituto Nazionale di Geofisica e Vulcanologia, Napoli, Italy.
[19]Claudia Troise, Istituto Nazionale di Geofisica e Vulcanologia, Napoli, Italy.
[20]Giuseppe De Natale, Istituto Nazionale di Geofisica e Vulcanologia, Napoli, Italy.

In proposing the ECMI Minisymposium “Numerical Methods in Volcano Geophysics”, our motivation is to bring geophysics and volcanology problems into focus and highlight the complexity in modelling the involved processes. Mathematical sciences play a central role in the effort to solve several challenges posed in volcano geophysics both at the modeling and computational level. The minisymposium will address themes with specialists presenting the theory and implementation of various numerical approaches applied to advance our knowledge of volcanic processes and quantitatively assess the volcanic hazards.

Fictitious Domain Methods for Fracture Models in Elasticity

Olivier Bodart, Valérie Cayol, Sébastien Court, and Jonas Koko

Abstract In this paper we are interested in a linear elasticity system modeling the presence of a crack inside a volcano. The traction force on this crack induces discontinuities of the displacement field. The computation of the latter is carried out with a finite element method for which the boundary of the crack is taken into account with a *fictitious domain* approach; It means that the mesh we consider does not fit to the crack. The interest of this approach lies in a framework where the position and the shape of the crack is lead to evolve, and in that case no re-meshing is required.

Keywords Fictitious domain approach • Fracture models

1 Introduction

Cracks (or fractures) play a major role in crustal deformations, whether acting in tensile mode or in shear mode. As a consequence, the simulation of displacements produced by inside cracks is an important issue for geological applications. However, when the position and the shape of a crack have to be updated from a step to the next, for instance when studying the propagation of a crack or when inverting surface deformation, such computations are expensive, making the modeling challenging.

This study presents a method that efficiently addresses the modeling of cracks, using a *fictitious domain* approach such that the cracks do not have to fit the mesh. The literature in this field of research is more and more abundant; Let us just cite the

O. Bodart • V. Cayol • S. Court (✉) • J. Koko
Université Blaise Pascal - Labex ClerVolc, Campus des Cézeaux 24 avenue des Landais, BP 80026, 63171 Aubière cedex, France
e-mail: sebastien.court@math.univ-bpclermont.fr

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_80

original article of the eXtended Finite Element Method [8] for taking into account cracks, or [6] in the context of fluid mechanics. The method we use is inspired from XFEM, but no enrichment of basis functions by singular functions is required near the crack region. The principles of our method are more comparable with the ones of [7] or [3], with the originality that we have to tackle here discontinuous fields of displacement. An other advantage of our method is the simplicity of the implementation.
The paper is divided as follows: In Sect. 2 we set the theoretical problem, and transform it into a variational problem. Next in Sect. 3 we explain the discretization we develop here, in particular the way the *fictitious domain* method is carried out. In the last Sect. 4 we provide numerical experiments, namely convergence tests and also physical experiments.

2 Setting of the Problem

Given a domain Ω of $\mathbb{R}^2$, and a crack $\Gamma_T \subset\subset \Omega$ represented by an injective curve, we consider a steady linear elasticity model governed by the following system:

$$\begin{cases} -\text{div}\, \sigma_L(\mathbf{u}) = \mathbf{f} & \text{in } \Omega, \\ \mathbf{u} = 0 & \text{on } \partial\Omega, \\ \sigma_L(\mathbf{u})\mathbf{n} = p\mathbf{n} & \text{on } \Gamma_T. \end{cases}$$

In this system the displacement of the solid is denoted by $\mathbf{u}$, some external forces (like the gravity) by $\mathbf{f}$, and $\sigma_L(\mathbf{u}) = 2\mu\varepsilon(\mathbf{u}) + \lambda(\text{div}\, \mathbf{u})\text{i}_{\mathbb{R}^2}$ denotes the Lamé stress tensor, with $\varepsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$. The coefficients μ and λ can be related to the Poisson coefficient ν and the Young modulus E by the formulas below:

$$\lambda = \frac{E\nu}{(1-2\nu)(1+\nu)}, \qquad \mu = \frac{E}{2(1+\nu)}.$$

The traction force of value $p > 0$ is applied on the both sides of the crack Γ_T, so we have to make precise the *outward* normal $\mathbf{n}$ on Γ_T. Moreover, for giving a sense to the both sides of the crack, we have to be able to determine whether a point of the domain lies on one side or the other of the crack. For that, the most convenient way we have found consists in uncoupling the problem by setting two unknowns displacements instead of a global one (Fig. 1).

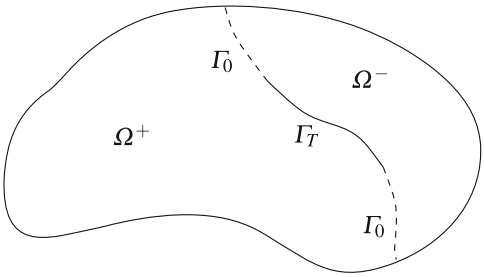


Fig. 1 Splitting of the domain according to the crack

2.1 Uncoupling the Problem from the Discontinuities of the Displacement Field

We extend the crack Γ_T to Γ, as represented below:

The global domain Ω is now split into two sub-domains Ω^+ and Ω^-. We have:

$$\Gamma = \Gamma_0 \cup \Gamma_T, \qquad \Omega = \Omega^+ \cup \Gamma \cup \Omega^-.$$

Let us now denote $\mathbf{u}^+ = \mathbf{u}_{|\Omega^+}$ and $\mathbf{u}^- = \mathbf{u}_{|\Omega^+}$. On the artificial boundary Γ_0—which is not connected—we have to ensure the continuity of the displacement, that is to say $\mathbf{u}^+ - \mathbf{u}^- = 0$. The system we are now interested in is the following[1]:

$$\begin{cases} -\text{div}\,\sigma_L(\mathbf{u}) = \mathbf{f} & \text{in } \Omega^+ \cup \Omega^-, \\ \mathbf{u} = 0 & \text{on } \partial\Omega, \\ (\sigma_L(\mathbf{u})\mathbf{n})^{\pm} = p\mathbf{n}^{\pm} & \text{on } \Gamma_T, \\ [\mathbf{u}] = 0 & \text{across } \Gamma_0 = \Gamma \setminus \Gamma_T, \\ [\sigma_L(\mathbf{u})]\,\mathbf{n}^+ = 0 & \text{across } \Gamma_0. \end{cases} \tag{1}$$

The notation $[\boldsymbol{\varphi}]$ refers to the jump of a function $\boldsymbol{\varphi}$ across Γ_0. Of course, the homogeneous Dirichlet condition on $\partial\Omega$ can be replaced by non-homogeneous data mixing Neumann conditions and Dirichlet conditions.

2.2 Continuous Formulation

Consider the following functional spaces:

$$\mathbf{V}^+ = \left\{\mathbf{v} \in \mathbf{H}^1(\Omega^+) \mid \mathbf{v} = 0 \text{ on } \partial\Omega \cap \partial\Omega^+\right\},$$

[1]The symbol $\pm$ represents the fact that we consider both formulations involving the symbols $+$ and $-$, in a sake of concision. The outward normal of domain $\Omega^{\pm}$ is denoted by $\mathbf{n}^{\pm}$.

$$\mathbf{V}^- = \left\{\mathbf{v} \in \mathbf{H}^1(\Omega^-) \mid \mathbf{v} = 0 \text{ on } \partial\Omega \cap \partial\Omega^-\right\},$$
$$\mathbf{W} = \mathbf{H}^{-1/2}(\Gamma_0) = \left(\mathbf{H}^{1/2}(\Gamma_0)\right)'.$$

We choose to impose the jump condition on Γ_0 by a multiplier $\boldsymbol{\lambda}$. A weak solution of system (1) can be seen as the stationary point in $\mathbf{V}^+ \times \mathbf{V}^- \times \mathbf{W}$ of the following Lagrangian:

$$\begin{aligned} L(\mathbf{u}^+, \mathbf{u}^-, \boldsymbol{\lambda}) = \frac{1}{2}\int_{\Omega+} \sigma_L(\mathbf{u}^+) : \varepsilon(\mathbf{u}^+)\mathrm{d}\Omega^+ + \frac{1}{2}\int_{\Omega-} \sigma_L(\mathbf{u}^-) : \varepsilon(\mathbf{u}^-)\mathrm{d}\Omega^- \\ - \int_{\Omega+} \mathbf{f}\cdot\mathbf{u}^+\mathrm{d}\Omega^+ - \int_{\Omega-} \mathbf{f}\cdot\mathbf{u}^-\mathrm{d}\Omega^- - \int_{\Gamma_T} \mathbf{u}^+\cdot p\mathbf{n}^+\mathrm{d}\Gamma_T \\ - \int_{\Gamma_T} \mathbf{u}^-\cdot p\mathbf{n}^-\mathrm{d}\Gamma_T \\ + \left\langle\boldsymbol{\lambda}; (\mathbf{u}^+ - \mathbf{u}^-)\right\rangle_{\mathbf{H}^{-1/2}(\Gamma_0);\mathbf{H}^{1/2}(\Gamma_0)}. \end{aligned}$$

In this expression $\sigma_L(\mathbf{u}) : \varepsilon(\mathbf{u}) = \mathrm{trace}\left(\sigma_L(\mathbf{u})\varepsilon(\mathbf{u})^T\right)$ denotes the classical inner product for matrices. Recall that the bilinear form bilinear form $(\mathbf{u}, \mathbf{v}) \mapsto \sigma_L(\mathbf{u}) : \varepsilon(\mathbf{v})$ is symmetric. Note that in the expression of L the jump condition $[\sigma_L(\mathbf{u})]\,\mathbf{n}^+ = 0$ across Γ_0 is no longer taken into account. Indeed, the first-order optimality conditions for L give for all test function $\mathbf{v} \in \mathbf{V}^\pm$ the following equality

$$\int_{\Omega^\pm} \sigma_L(\mathbf{u}^\pm) : \varepsilon(\mathbf{v})\mathrm{d}\Omega^\pm \pm \langle\boldsymbol{\lambda}; \mathbf{v}\rangle_{\mathbf{W},\mathbf{W}'} = \int_{\Omega^\pm} \mathbf{f}\cdot\mathbf{v}\mathrm{d}\Omega^\pm + \int_{\Gamma_T} \mathbf{v}\cdot p\mathbf{n}^\pm\mathrm{d}\Gamma_T,$$

On the other hand, taking the inner product by $\mathbf{v}$ of the first equation of (1) yields, after integration by parts

$$\int_{\Omega^\pm} \sigma_L(\mathbf{u}^\pm) : \varepsilon(\mathbf{v})\mathrm{d}\Omega^\pm - \left\langle\sigma_L(\mathbf{u}^\pm)\mathbf{n}^\pm; \mathbf{v}\right\rangle_{\mathbf{W},\mathbf{W}'} = \int_{\Omega^\pm} \mathbf{f}\cdot\mathbf{v}\mathrm{d}\Omega^\pm + \int_{\Gamma_T} \mathbf{v}\cdot p\mathbf{n}^\pm\mathrm{d}\Gamma_T.$$

Thus we can deduce $\boldsymbol{\lambda} = -\sigma_L(\mathbf{u}^+)\mathbf{n}^+$ and $\boldsymbol{\lambda} = \sigma_L(\mathbf{u}^-)\mathbf{n}^-$.

The variational problem derived from this functional L is then

$$\text{Find } (\mathbf{u}^+, \mathbf{u}^-, \boldsymbol{\lambda}) \text{ in } \mathbf{V}^+ \times \mathbf{V}^- \times \mathbf{W} \text{ such that} \tag{2}$$

$$\begin{cases} \mathscr{A}^\pm((\mathbf{u}^+, \mathbf{u}^-, \boldsymbol{\lambda}); \mathbf{v}) = \mathscr{L}^\pm(\mathbf{v}) & \forall \mathbf{v} \in \mathbf{V}^\pm, \\ \mathscr{B}((\mathbf{u}^+, \mathbf{u}^-, \boldsymbol{\lambda}); \boldsymbol{\mu}) = 0, & \forall \boldsymbol{\mu} \in \mathbf{W}, \end{cases} \tag{3}$$

where we set

$$\begin{aligned}
\mathscr{A}^{\pm}((\mathbf{u}^{+},\mathbf{u}^{-},\boldsymbol{\lambda});\mathbf{v}) &= \int_{\Omega^{\pm}} \sigma_L(\mathbf{u}^{\pm}) : \varepsilon(\mathbf{v})\mathrm{d}\Omega^{\pm} \pm \langle \boldsymbol{\lambda};\mathbf{v}\rangle_{\mathbf{H}^{-1/2}(\Gamma_0);\mathbf{H}^{1/2}(\Gamma_0)}\,,\\
\mathscr{L}^{\pm}(\mathbf{v}) &= \int_{\Omega^{\pm}} \mathbf{f}\cdot\mathbf{v}\mathrm{d}\Omega^{\pm} + \int_{\Gamma_T} \mathbf{v}\cdot p\mathbf{n}^{\pm}\mathrm{d}\Gamma_T, \qquad (4)\\
\mathscr{B}((\mathbf{u}^{+},\mathbf{u}^{-},\boldsymbol{\lambda});\boldsymbol{\mu}) &= \langle \boldsymbol{\mu};\mathbf{u}^{+}\rangle_{\mathbf{H}^{-1/2}(\Gamma_0);\mathbf{H}^{1/2}(\Gamma_0)} - \langle \boldsymbol{\mu};\mathbf{u}^{-}\rangle_{\mathbf{H}^{-1/2}(\Gamma_0);\mathbf{H}^{1/2}(\Gamma_0)}\,.
\end{aligned}$$

3 Discrete Formulation

The discrete formulation we develop in this paper is similar to the ones given in [7] and [3]: It is a *fictitious domain* method, in which the degrees of freedom for the multiplier on the boundary Γ are independent of the mesh. Let us first explain how we proceed for taking into account degrees of freedom which do not lie on the edges of the mesh originally.

3.1 Discretization

The fictitious domains for the unknowns are first considered on the whole domain Ω. Let us consider some discrete finite element spaces, $\tilde{\mathbf{V}}_h \subset \mathbf{H}^1(\Omega)$ and $\tilde{\mathbf{W}}_h \subset \mathbf{L}^2(\Omega)$. These spaces can be defined on the same structured mesh of Ω, that can be chosen Cartesian. We set for the displacement

$$\tilde{\mathbf{V}}_h = \left\{\mathbf{v}_h \in C(\overline{\Omega}) \mid \mathbf{v}_{h|\partial\Omega} = 0,\ \mathbf{v}_{h|T} \in P(T),\ \forall T \in \mathscr{T}_h\right\}, \qquad (5)$$

where $P(T)$ is a finite dimensional space of regular functions such that $P(T) \supseteq P_k(T)$ for some integer $k \geq 1$. See [4] for more details. The mesh parameter stands for $h = \max_{T\in\mathscr{T}_h} h_T$, where h_T is the diameter of the triangle T. The set $\tilde{\mathbf{W}}_h$ can be defined similarly, with the difference that the degree of the polynomial base functions chosen for $\tilde{\mathbf{W}}_h$ has to be lower than the one chosen for the displacement, in order to satisfy an inf-sup condition. We define

$$\mathbf{V}_h^{+} := \tilde{\mathbf{V}}_{h|\Omega^{+}}\,, \quad \mathbf{V}_h^{-} := \tilde{\mathbf{V}}_{h|\Omega^{-}}\,, \quad \mathbf{W}_h := \tilde{\mathbf{W}}_{h|\Gamma_0}\,,$$

which are natural discretization of $\mathbf{V}^{+}$, $\mathbf{V}^{-}$ and $\mathbf{H}^{-1/2}(\Gamma_0)$ respectively. This approach looks like the eXtended Finite Element Method [8], but here the standard basis functions near the boundary Γ are not enriched—by singular functions—but only multiplied by the Heaviside function ($H(\mathbf{x}) = 1$ for $\mathbf{x} \in \Omega^{\pm}$ and $H(\mathbf{x}) = 0$ for $\mathbf{x} \in \Omega \setminus \Omega^{\pm}$), and the products are substituted in the variational formulation of the problem. This kind of strategy is also adopted in [5] and [2] for instance.

3.2 Matrix Formulation

An approximation of problem (3) is given as follows

$$\text{Find } (\mathbf{u}_h^+, \mathbf{u}_h^-, \boldsymbol{\lambda}_h) \text{ in } \mathbf{V}_h^+ \times \mathbf{V}_h^- \times \mathbf{W}_h \text{ such that}$$

$$\begin{cases} a^+(\mathbf{u}_h^+, \mathbf{v}_h^+) + b(\lambda_h, \mathbf{v}_h^+) = \mathscr{L}^+(\mathbf{v}_h^+) & \forall \mathbf{v}_h^+ \in \mathbf{V}_h^+, \\ a^-(\mathbf{u}_h^-, \mathbf{v}_h^-) - b(\lambda_h, \mathbf{v}_h^-) = \mathscr{L}^-(\mathbf{v}_h^-) & \forall \mathbf{v}_h^- \in \mathbf{V}_h^-, \\ b^+(\boldsymbol{\mu}_h, \mathbf{u}_h^+) - b^-(\boldsymbol{\mu}_h, \mathbf{u}_h^-) = 0, & \forall \boldsymbol{\mu}_h \in \mathbf{W}_h, \end{cases}$$

where

$$a^+(\mathbf{u}_h^+, \mathbf{v}_h^+) = \int_{\Omega^+} \sigma_L(\mathbf{u}_h^+) : \varepsilon(\mathbf{v}_h^+) \mathrm{d}\Omega^+, \qquad a^-(\mathbf{u}_h^-, \mathbf{v}_h^-) = \int_{\Omega^-} \sigma_L(\mathbf{u}_h^-) : \varepsilon(\mathbf{v}_h^-) \mathrm{d}\Omega^-, \tag{6}$$

$$b^+(\boldsymbol{\mu}_h, \mathbf{u}_h^+) = \int_{\Gamma_0} \mu_h \cdot \mathbf{u}_h^+ \mathrm{d}\Gamma_0, \qquad b^-(\boldsymbol{\mu}_h, \mathbf{u}_h^-) = \int_{\Gamma_0} \mu_h \cdot \mathbf{u}_h^- \mathrm{d}\Gamma_0. \tag{7}$$

Note that the duality product between $\mathbf{H}^{-1/2}(\Gamma_0)$ and $\mathbf{H}^{1/2}(\Gamma_0)$ has been turned into the scalar product of $\mathbf{L}^2(\Gamma_0)$. We could avoid this by using a Laplace-Beltrami operator, but for a sake of simplicity (and under stronger regularity assumptions) we proceed like this. In matrix notation, this formulation gives

$$\left(\begin{array}{cc|c} A^+ & 0 & B^{+T} \\ 0 & A^- & -B^{-T} \\ \hline B^+ & -B^- & 0 \end{array}\right) \left(\begin{array}{c} \mathbf{U}^+ \\ \mathbf{U}^- \\ \hline \Lambda \end{array}\right) = \left(\begin{array}{c} \mathbf{F}^+ \\ \mathbf{F}^- \\ \hline 0 \end{array}\right),$$

where $\mathbf{U}^+$, $\mathbf{U}^-$ and Λ are the degrees of freedom of $\mathbf{u}_h^+$, $\mathbf{u}_h^-$ and $\boldsymbol{\lambda}_h$ respectively. The matrices A^+, A^-, B^+ and B^- are the discretization of (6)–(7). If we denote by $\{\boldsymbol{\varphi}_i^+\}$, $\{\boldsymbol{\varphi}_i^-\}$ and $\{\boldsymbol{\psi}_i\}$ the basis functions of the spaces $\mathbf{V}_h^+$, $\mathbf{V}_h^-$ and $\mathbf{W}_h$ respectively, we have:

$$\left(A^+\right)_{ij} = \int_{\Omega^+} \sigma_L(\boldsymbol{\varphi}_i^+) : \varepsilon(\boldsymbol{\varphi}_j^+) \mathrm{d}\Omega^+, \qquad (A^-)_{ij} = \int_{\Omega^-} \sigma_L(\boldsymbol{\varphi}_i^-) : \varepsilon(\boldsymbol{\varphi}_j^-) \mathrm{d}\Omega^-,$$

$$\left(B^+\right)_{ij} = \int_{\Gamma_0} \boldsymbol{\psi}_i \cdot \boldsymbol{\varphi}_j^+ \mathrm{d}\Gamma_0, \qquad (B^-)_{ij} = \int_{\Gamma_0} \boldsymbol{\psi}_i \cdot \boldsymbol{\varphi}_j^- \mathrm{d}\Gamma_0.$$

The vectors $\mathbf{F}^{\pm}$ are the discretization of (4).

4 Numerical Experiments

For illustrating the implementation of our method, we first make computations for artificial data corresponding to a given exact solution. Next, given a physical situation, we represent the computational domain Ω deformed by the computed displacement.

4.1 Convergence Rates

Given a square $\Omega = [0; 1] \times [0; 1]$, we consider for Γ a straight line splitting Ω into two parts. Since the main difficulty of our problem lies in the implementation of the jump condition, we only take a look at Γ_0, and so $\Gamma_T = \emptyset$. Tests with a non-trivial crack Γ_T are given in Sect. 4.2, or in [1] for instance. The imposed jump (D_1, D_2) on Γ_0 is chosen to be constant. We consider the following exact solution:

$$\mathbf{u}_{ex}(x, y) = \begin{pmatrix} (x + y)\cos(x) \\ (x - y)\sin(y) \end{pmatrix} \text{ if } y > 0.53,$$

$$\mathbf{u}_{ex}(x, y) = \begin{pmatrix} (x + y)\cos(x) - D_1 \\ (x - y)\sin(y) - D_2 \end{pmatrix} \text{ if } y \leq 0.53.$$

In the figures below we compute the relative errors on the displacement, between the exact solution above and the computed one, for different choices of finite elements, and we deduce an approximation of the order of convergence (Figs. 2, 3, 4, and 5).

4.2 Physical Tests

For a rectangle $[0; 100] \times [0; 50]$ with a mesh size $h = 1$, we consider a crack whose position and shape—a segment- can be guessed on the pictures below. In the first test the crack is vertical, in the second one it is inclined, and in the third test it is inclined and touching the surface. On the surface we consider a Neumann-type homogeneous condition: $\sigma_L(\mathbf{u})\mathbf{n} = 0$. The intensity $p > 0$ of the traction applied on the both sides of the crack is chosen to be constant. After computation of the displacement, we deform the initial rectangle according to the effects of the displacement, by amplifying the effects of the deformation (Figs. 6, 7, and 8).

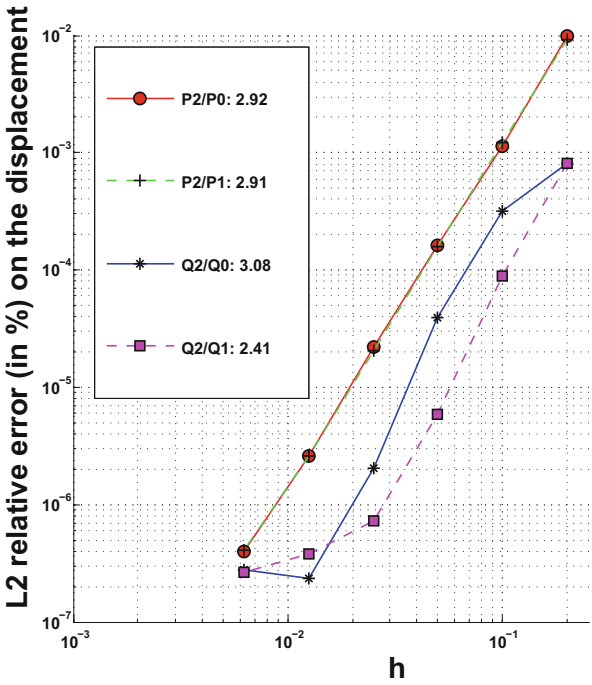


Fig. 2 $\mathbf{L}^2(\Omega)$-relative error (in %)

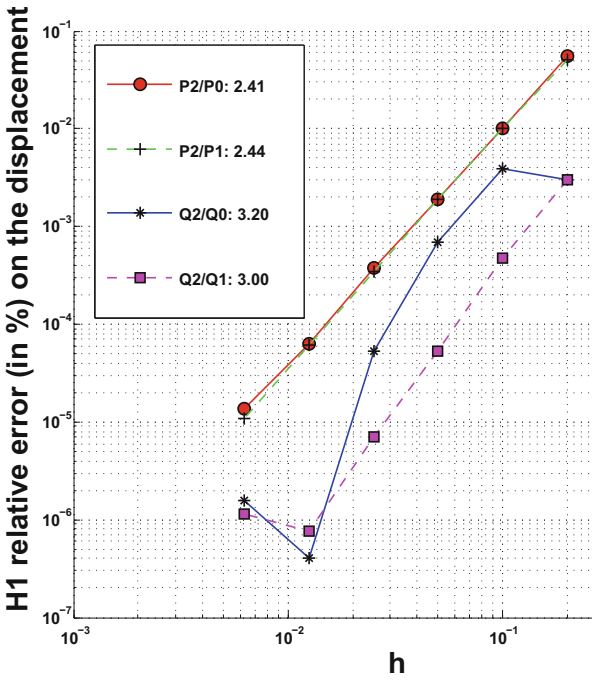


Fig. 3 $\mathbf{H}^1(\Omega)$-relative error (in %)

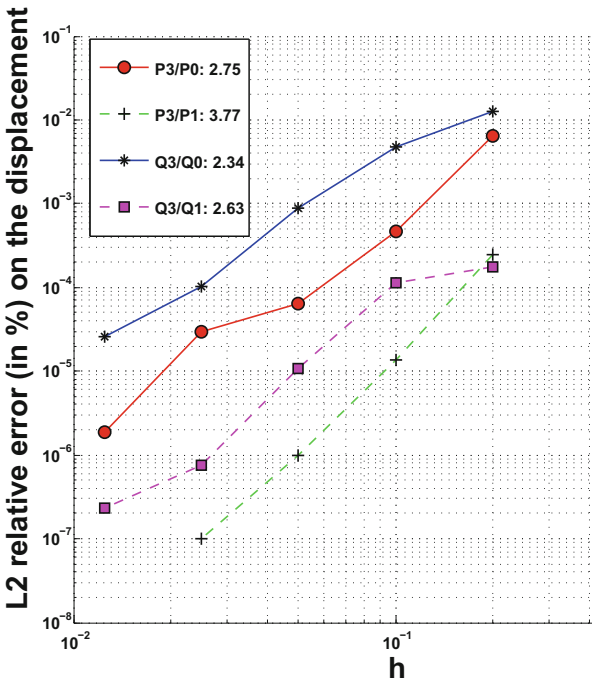


Fig. 4 $\mathbf{L}^2(\Omega)$-relative error (in %)

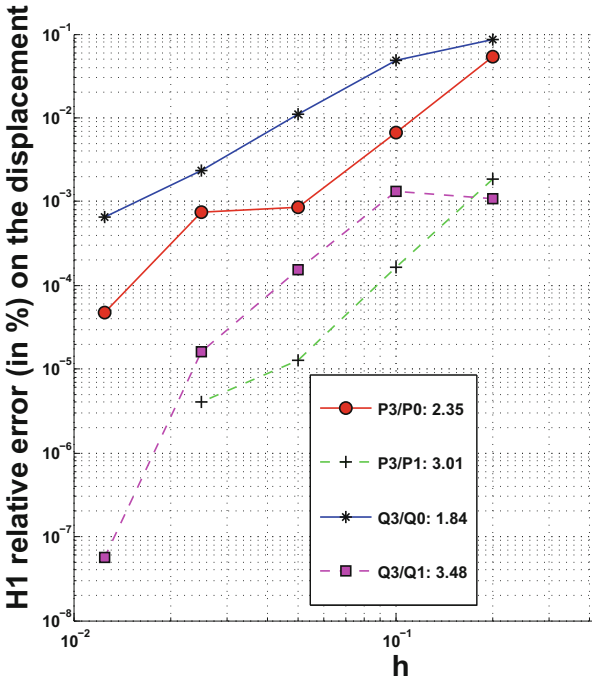


Fig. 5 $\mathbf{H}^1(\Omega)$-relative error (in %)

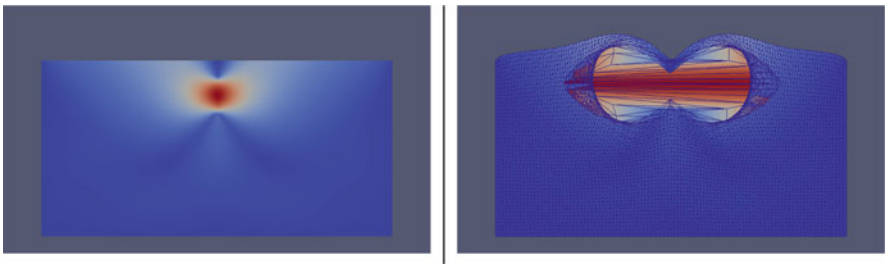


Fig. 6 Displacement due to a vertical inside crack applying traction forces

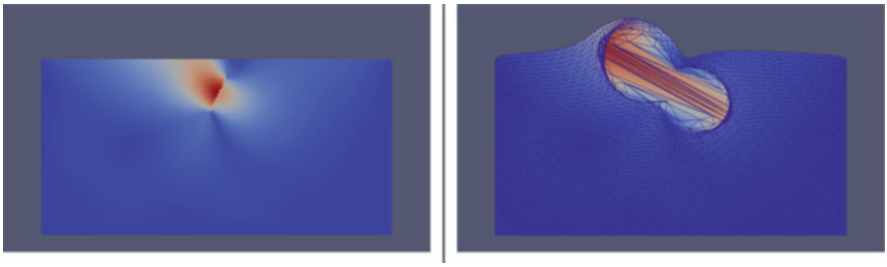


Fig. 7 Displacement due to an inclined inside crack applying traction forces

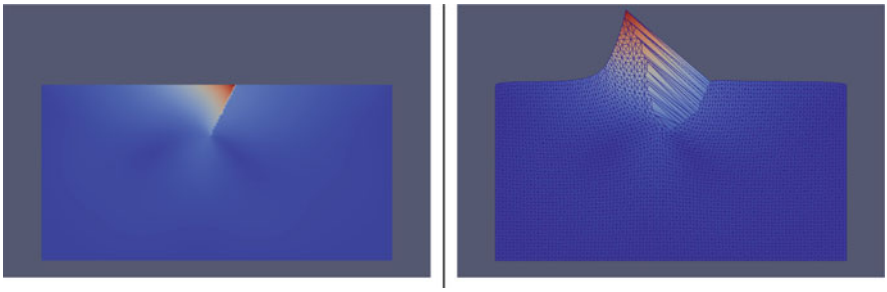


Fig. 8 Displacement due to an inclined crack intersecting the surface and applying traction forces

5 Conclusion

We have solved numerically a crack problem in an elastic medium by a finite element method relying on a *fictitious domain* approach. Rates of convergence have been computed and physical tests have been performed for underlying the accuracy of our approach.

The direct problem we have considered is the first step toward the study of an inverse problem: The goal is to recover information on the crack (position, shape, stress) from surface measurements. The interest of the *fictitious domain* approach lies in an algorithmic framework in which the position of the sought crack would have to be

updated; In that case no re-meshing would be needed, only local re-assembling of stiffness matrices would be required, and so we hope to obtain a gain of efficiency in terms of computation time and resources.

Acknowledgements This research was financed by the French Government Laboratory of Excellence initiative no. ANR-10-LABX-0006, the Région Auvergne and the European Regional Development Fund. This is Laboratory of Excellence ClerVolc contribution number 139.

References

1. Bodart, O., Cayol, V., Court, S., Koko, J.: Xfem based fictitious domain method for linear elasticity model with crack. SIAM J. Sci. Comput. **38**(2), B219–B246 (2016). http://arxiv.org/abs/1502.03148
2. Choi, Y.J., Hulsen, M.A., Meijer, H.E.H.: Simulation of the flow of a viscoelastic fluid around a stationary cylinder using an extended finite element method. Comput. Fluids **57**, 183–194 (2012). doi:10.1016/j.compfluid.2011.12.020. http://dx.doi.org/10.1016/j.compfluid.2011.12.020
3. Court, S., Fournié, M., Lozinski, A.: A fictitious domain approach for the stokes problem based on the extended finite element method. Int. J. Numer. Methods Fluids **74**(2), 73–99 (2014). doi:10.1002/fld.3839. http://dx.doi.org/10.1002/fld.3839
4. Ern, A., Guermond, J.L.: Theory and Practice of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, New York (2004)
5. Gerstenberger, A., Wall, W.A.: An extended finite element method/Lagrange multiplier based approach for fluid-structure interaction. Comput. Methods Appl. Mech. Eng. **197**(19–20), 1699–1714 (2008). doi:10.1016/j.cma.2007.07.002. http://dx.doi.org/10.1016/j.cma.2007.07.002
6. Girault, V., Glowinski, R., Pan, T.W.: A fictitious-domain method with distributed multiplier for the Stokes problem. In: Applied Nonlinear Analysis, pp. 159–174. Kluwer/Plenum, New York (1999)
7. Haslinger, J., Renard, Y.: A new fictitious domain approach inspired by the extended finite element method. SIAM J. Numer. Anal. **47**(2), 1474–1499 (2009). doi:10.1137/070704435. http://dx.doi.org/10.1137/070704435
8. Moës, N., Dolbow, J., Belytschko, T.: A finite element method for crack growth without remeshing. Int. J. Numer. Methods Eng. **46**(1), 131–150 (1999). doi:10.1002/(SICI)1097-0207(19990910)46:1<131::AID-NME726>3.0.CO;2-J

Geophysical Changes in Hydrothermal-Volcanic Areas: A Finite-Difference Ghost-Point Method to Solve Thermo-Poroelastic Equations

Armando Coco, Gilda Currenti, Ciro Del Negro, Joachim Gottsmann, and Giovanni Russo

Abstract We propose a finite-difference ghost-point method for the numerical solution of thermo-poroelastic equations. The method is applied to evaluate deformation, gravity and thermomagnetic changes in Campi Flegrei area caused by hydrothermal fluid circulation during an unrest.

Keywords Finite differences • Hydrothermal fluid circulation

1 Introduction

The increasing combined use of satellite and ground-based geophysical observations in volcanic areas has dramatically enhanced our ability to detect and track complex and multifaceted volcanic processes that are often difficult to reconcile using models of elastic mechanical behavior of Earth's upper crust [4]. Usually, magma accumulation and intrusion are modelled as volume and pressure changes within the crust. However, the interaction between magma and host rocks such as the heating and expansion of hydrothermal fluids may also induce measurable changes in geophysical signals [6, 10, 11, 14]. A thermo-poroelastic numerical model is proposed to jointly evaluate ground deformation, magnetic and gravity changes caused by hydrothermal fluid circulation in complex media with surface topography

A. Coco (✉) • J. Gottsmann
Bristol University, Queen's Road, Bristol BS8 1RJ, UK
e-mail: Armando.Coco@bristol.ac.uk; J.Gottsmann@bristol.ac.uk

G. Currenti • C.D. Negro
Istituto Nazionale di Geofisica e Vulcanologia, Piazza Roma 2, Catania, Italy
e-mail: currenti@ct.ingv.it; delnegro@ct.ingv.it

G. Russo
Universita' di Catania, Viale Andrea Doria 6, Catania, Italy
e-mail: russo@dmi.unict.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_81

and mechanical heterogeneities. The aim is to provide a numerical framework for a more realistic assessment of geophysical observations associated with sub-volcanic processes.

2 Thermo-Poroelastic Model

The mathematical model is based on the governing equations of the thermo-poro-elasticity theory, which describe the elastic response of a porous medium to the propagation of hot fluid through pores. Assuming that the deformation occurs slowly, the rock is in static equilibrium and the displacement can be found by solving the equations of equilibrium coupled with thermo-poroelastic extension of the Hooke's law [8], giving the following set of equations [5]:

$$\nabla \cdot \boldsymbol{\sigma} = 0, \quad \boldsymbol{\sigma} = \lambda tr(\boldsymbol{\epsilon})\mathbf{I} + 2\mu\boldsymbol{\epsilon} + \alpha \Delta P \mathbf{I} + K\beta\Delta T\mathbf{I}, \quad \boldsymbol{\epsilon} = \frac{1}{2}\left(\nabla \mathbf{u} + (\nabla \mathbf{u})^T\right) \tag{1}$$

where $\boldsymbol{\sigma}$ and $\boldsymbol{\epsilon}$ are the stress and strain tensors, respectively, $\mathbf{u}$ is the deformation vector and λ and μ are the Lame's elastic medium parameters and K the bulk modulus. To the elastic stress tensor of the Hooke's law for elastic media, two terms are added: the ΔP pore-pressure change from poroelasticity theory through the α Biot-Willis coefficient and the ΔT temperature change from thermo-elasticity theory through the volumetric thermal expansion coefficient β. Fluid circulation, temperature and pore-pressure changes necessarily alter the density distribution and the magnetization of the porous media, which, in turn, affects the gravity and the magnetic fields, respectively. The gravity change Δg can be calculated by solving the following boundary value problem for the gravitational potential ϕ_g [3]:

$$\nabla^2 \phi_g = -4\,\pi\, G\, \Delta\rho, \quad \phi_g = 0 \text{ at infinity}, \quad \Delta g = -\frac{\partial \phi_g}{\partial z} \tag{2}$$

where G is the gravitational constant and $\Delta\rho$ is the density distribution change. As for the magnetic field changes, thermomagnetic effect due to thermal demagnetization/remagnetization is considered, since it generally yields larger magnetic changes with respect to piezomagnetic and electrokinetic effects, which may be also associated with volcanic activity. The thermomagnetic field can be described through the scalar potential formulation [12]:

$$\nabla^2 \phi_m = 4\,\pi\, \nabla \cdot \mathbf{J}, \quad \frac{\partial \phi_m}{\partial n} = 0 \text{ at infinity}, \quad \mathbf{B} = -\nabla \phi_m \tag{3}$$

where ϕ_m is the thermomagnetic scalar potential, **J** the thermal magnetization change, **B** the magnetic field. Pressure, temperature and density changes induced by hydrothermal fluid circulation are computed using the HYDROTHERM numerical code [7]. Starting from these quantities, ground deformation, gravity and thermomagnetic changes are solved through Eqs. (1)–(3). In this work we do not take into account the effects of displacements back to the hydrological properties (permeability and porosity), though its implementation is under consideration.

2.1 Numerical Method

Equations (1)–(3) are discretized by the finite-difference ghost-cell approach for unbounded domains and complex geometries [1, 2]. The ground surface is implicitly described by a level-set function, computed as the minimum (signed) distance between each node of the Cartesian grid and the nodes defining the topography. Using a coordinate transformation method, the unbounded domain is mapped into a bounded one $[-1, 1]^2$:

$$(r, z) \longrightarrow \left(\chi^{-1}(r), \chi^{-1}(z)\right), \text{ with } \chi(x) = \frac{cx}{(1 - x^2)}$$

where r and z are the radial and vertical coordinate, respectively, and c is a parameter regulating the length scale of the computational grid. The bounded domain is then discretized by uniform Cartesian grid, which automatically results in a quasi-uniform grid for the original domain, with a finer resolution close to the axis of symmetry, smoothly decreasing toward infinity. In this way we avoid artefact introduced when using artificial truncation of the domain. The original equations are transformed as well, following the transformation of differential operators $\partial_x = (\chi')^{-1}\partial_X$, where x and X are the general coordinate in the unbounded and bounded domain, respectively. The transformed equations are then discretized in the bounded domain by finite-difference method. Discretization on a generic grid node requires to know the values on surrounding grid nodes. Some of these grid nodes may lie outside the computational domain (*ghost point*) and a suitable value must be defined on it. Referring to Fig. 1, we define the value in the ghost point G in such a way the biquadratic interpolation on the nine-point stencil matches the boundary condition on the projection point B, which is computed by the signed distance function.

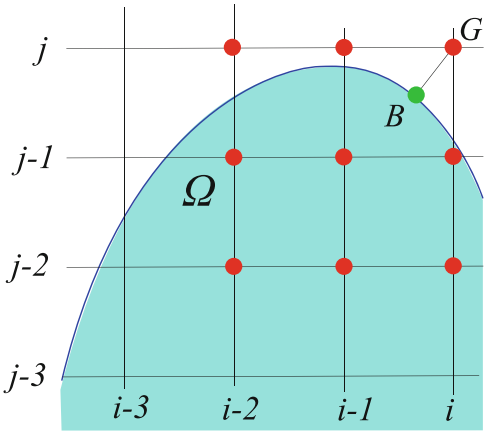


Fig. 1 Nine-point stencil used to compute the extrapolation value in G of the boundary condition on B

3 Hydrothermal Model

The hot fluid circulation in the hydrothermal system is simulated by the HYDROTHERM software, which solves the mass and energy balance equations for a multiphase ground-water flow [7]. The equations can be resumed as follows:

$$\frac{\partial Q_\alpha}{\partial t} + \nabla \cdot F_\alpha - q_\alpha = 0, \tag{4}$$

where Q is the accumulation term, F the flux and q the source (or sink) term, while the subscript $\alpha = M$ or E refers to the mass or energy balance equation. The accumulation term for mass balance equation is described by $Q_M = \phi \sum_\beta \rho_\beta S_\beta$, where the subscript $\beta = l$ or g refers to the liquid or gas phase, ϕ is the porosity, ρ_β the density and S_β the saturation. The fluid flux is described by the Darcy's law, so that $F_M = \sum_\beta F_\beta$, with $F_\beta = \frac{KK_{r\beta}\rho_\beta}{\mu_\beta}(\nabla P_\beta - \rho_\beta \hat{g})$, where K and $K_{r\beta}$ are the absolute and relative permeability, μ_β the viscosity, P_β the fluid pressure, $\hat{g}$ the gravitational vector. For the energy balance equation, the accumulation term is $Q_E = \phi \sum_\beta (\rho_\beta h_\beta S_\beta) + (1 - \phi)\rho_r h_r$, where h_β is the specific enthalpy of the phase β, while ρ_r and h_r are the density and specific enthalpy of the porous-matrix solid phase, respectively. The heat flux is $F_E = -\lambda \nabla T + \sum_\beta h_\beta F_\beta$, where λ is the thermal conductivity of the bulk porous medium and T the temperature.

4 Campi Flegrei Unrest Simulation

The method is applied at the Campi Flegrei area to simulate an unrest caused by a deep injection of hot fluid in the caldera [10, 11, 14]. Without loosing generality the multiparametric model is designed in axi-symmetric formulation. The

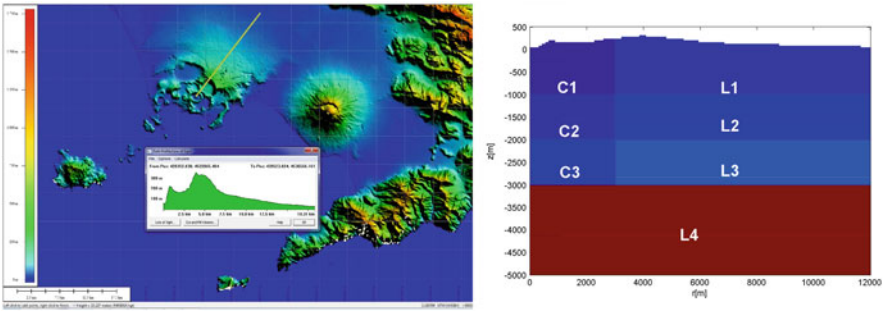


Fig. 2 *Left*: digital elevation model from the 90 m Shuttle Radar Topography Mission data. *Right*: axis-symmetric model geometry and elastic medium heterogeneity of the Campi Flegrei area

computational domain is a vertical section (Fig. 2), and for the hydrothermal model (Sect. 3) it extends up to a radial distance of 10 km and a depth of 1.5 km computed from the intersection between the ground surface and the axis of symmetry. It consists of 2400 cells, with a horizontal resolution of 61.33 m along the axis of symmetry, which decreases as the radial distance increases. The vertical resolution is 45.5 m in the whole domain. The computational nodes are cell-centered. In the thermo-poroelastic model and for the gravity (Eq. (2)) and magnetic (Eq. (3)) problems (Sect. 2) the domain extends toward infinity, as described in Sect. 2.1, with a resolution of 38.76 m close to the axis of symmetry. The computational grid is vertex-centered.

Firstly, HYDROTHERM is run to simulate the fluid injection into the system for 3 years at a constant rate of 70 kg/s and at a temperature of 350 °C from a point source located in the lower part of the caldera ($r = 0$ km; $z = -1.5$ km). Initial conditions are obtained by simulating a 10 thousands year long quiet phase with a deep injection of hot fluids in the same area at the same temperature but at a lower constant rate of 27 kg/s. The ground surface is at atmospheric pressure (0.1 MPa) and temperature (20 °C) during the whole simulation. Right and bottom boundaries are assumed to be impervious and adiabatic. Values of the hydrological and thermal properties of rocks are defined on the basis of literature data [10, 14] (Table 1). During the simulation of the unrest, temperature, pressure and density variations with respect to their initial distributions are computed and fed into the thermo-poroelastic solver. Due to the different grids adopted, these quantities are interpolated from the HYDROTHERM grid to the thermo-poroelastic solver nodes. The elastic medium properties are defined on the basis of tomographic studies [15], which depict the heterogeneity of the shallow structure of Campi Flegrei (Table 2). Following [13], the shallow area of the medium is divided into three horizontal layers having a thickness of 1 km. The inner caldera is modelled as a 3 km wide cylinder, coaxial with the axis of symmetry, with an internal variation of the elastic parameters. A further layer extends from 3 km depth to the bottom of the (infinite) computational domain (Fig. 2). The thermal expansion coefficient β is $10^{-5}K^{-1}$ and

Table 1 Rock properties used in calculations of the hydrothermal model

Density	2000 kg/m^3
Permeability	10^{-14} m^2
Conductivity	2.8 W m^{-1} K^{-1}
Porosity	0.2
Specific heat	1000 J kg^{-1} K^{-1}

Table 2 Elastic properties of the medium

Region	Rigidity [GPa]	Poisson ratio
L1	4	0.25
C1	2.3	0.33
L2	5	0.25
C2	3.6	0.33
L3	6.5	0.25
C3	5	0.33
L4	20	0.25

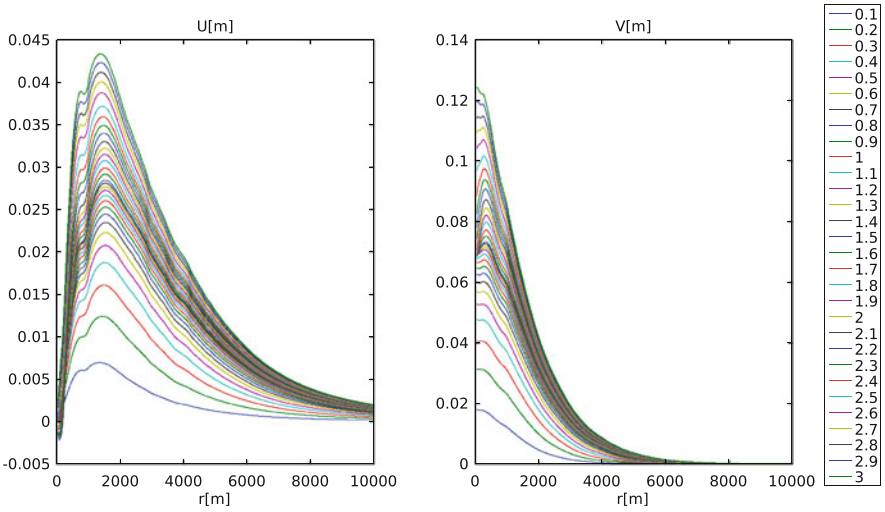


Fig. 3 Time evolution of the horizontal (*right*) and vertical (*left*) deformations at the ground surface during the unrest phase. Deformations increase over time

the Biot-Willis coefficient α is $1 - K/K_s$, where K is the isothermal drained bulk modulus (5 GPa) and K_s is the bulk modulus of the solid constituent (30 GPa). The deformation pattern (Fig. 3) computed solving Eq. (1) is in agreement with the one obtained in [10, 11]. The contribution of the thermo-elastic term is negligible for this simulation (Fig. 4).

To solve Eq. (2), density changes in the medium are computed with respect to the initial density distribution $\Delta\rho = \rho_i - \rho_0$, where $\rho_i = \phi(\rho_{li}S_{li} + \rho_{si}(1 - S_{li}))$ with ϕ the porosity, S the saturation, ρ_s and ρ_l the density of the steam and liquid phases. The gravity changes expected at the ground surface over time (Fig. 5, right) reach

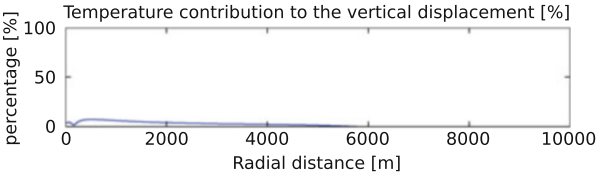


Fig. 4 Percentage of the temperature contribution to the vertical deformation, computed as $(v_T/v) \cdot 100$, where v_T is the vertical displacement computed taking $\alpha = 0$ in Eq. (1)

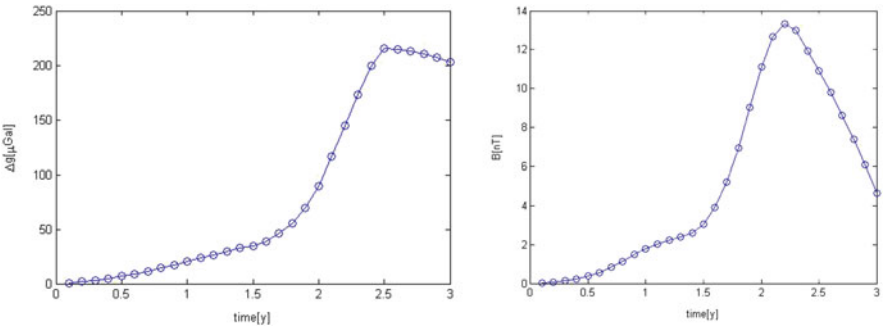


Fig. 5 Gravity (*left*) and thermomagnetic (*right*) changes computed at the ground surface in r=0

about 210 μGal in 3 years in agreement with those obtained in [11] for the half space domain. Rock magnetization in Eq. (3) is related to temperature changes on the basis of laboratory experiment on rock sample [9]. Magnetic changes reach about 14 nT in 2 years. Future investigations, mainly concentrated on a full coupling between the thermo-poroelastic and the hydrothermal solvers is currently underway, with the aim to extend the simulation to more realistic case of multi-component fluids.

Acknowledgements This work has been supported by the VUELCO and MEDSUV projects, which are funded by the EC FP7 under contracts #282759 and #308665. The research has been carried out within the framework of TecnoLab, the Laboratory for Technological Advance in Volcano Geophysics, at the INGV in Catania.

References

1. Coco, A., Russo, G.: Finite-difference ghost-point multigrid methods on cartesian grids for elliptic problems in arbitrary domains. J. Comput. Phys. **241**, 464–501 (2013)
2. Coco, A., Currenti, G., Del Negro, C., Russo, G.: A second order finite-difference ghost-point method for elasticity problems on unbounded domains with applications to volcanology. Commun. Comput. Phys. **16**, 983–1009 (2014)
3. Currenti, G.: Numerical evidences enabling to reconcile gravity and height changes in volcanic areas. Geophys. J. Int. (2014). doi:10.1093/gji/ggt507

4. Del Negro, C., Currenti, G., Solaro, G., Greco, F., Pepe, A., Napoli, R., Pepe, S., Casu, F., Sansosti, E.: Capturing the fingerprint of Etna volcano activity in gravity and satellite radar data. Sci. Rep. **3**, 3089 (2013)
5. Fung, Y.: Foundations of Solid Mechanics. Prentice-Hall, Englewood Cliffs (1965)
6. Gottsmann, J., Camacho, A.G., Tiampo, K.F., Fernández, J.: Spatiotemporal variations in vertical gravity gradients at the Campi Flegrei caldera (Italy): a case for source multiplicity during unrest? Geophys. J. Int. **167**, 1089–1096 (2006)
7. Hayba, D., Ingebritsen, S.: Multiphase groundwater flow near cooling plutons. J. Geophys. Res. **102**, 12235–12252 (1994)
8. Jaeger, J., Cook, N., Zimmerman, R.: Fundamentals of Rock Mechanics, 4th edn. Blackwell Publishing, Oxford (2007)
9. Okubo, A., Kanda, W., Ishiara, K.: Numerical simulation of volcanomagnetic effects due to Hydrothermal activity. Ann. Disas. Prev. Res. Inst. Kyoto Univ. **49**, 211–218 (2006)
10. Rinaldi, A., Todesco, M., Bonafede, M.: Hydrothermal instability and ground displacement at the Campi Flegrei caldera. Phys. Earth Planet. Inter. **178**, 155–161 (2010)
11. Rinaldi, A., Todesco, M., Bonafede, Vandemeulebrouck, M.J., Revil, A.: Electrical conductivity, ground displacement, gravity changes, and gas flow at Solfatara crater (Campi Flegrei caldera, Italy): results from numerical modeling. J. Volcanol. Geotherm. Res. **207**, 93–105 (2011)
12. Sasai, Y.: Integrals of Lipschitz-Hankel type for solving potential problems with axial symmetry. In: Proceedings Conductivity Anomaly Workshop, pp. 110–121 (1991)
13. Trasatti, E., Giunchi, C., Bonafede, M.: Structural and rheological constraints on source depth and overpressure estimates at the Campi Flegrei caldera, Italy. J. Volcanol. Geotherm. Res. **144**, 105–118 (2005)
14. Troiano, A., Di Giuseppe, M., Petrillo, Z., Troise, C., De Natale, G.: Ground deformation at calderas driven by fluid injection: modelling unrest episodes at Campi Flegrei (Italy). Geophys. J. Int. **187**, 833–847 (2011)
15. Zollo, A., Judenherc, S., Auger, E., D'Auria, L., Virieux, J., Capuano, P., Chiarabba, C., De Franco, R., Makris, J., Michelini, A., Musacchio, G.: Evidence for the buried rim of Campi Flegrei caldera from 3-d active seismic imaging. Geophys. Res. Lett. **30**, 25–45 (2003)

Numerical Simulation Applied to the Solfatara-Pisciarelli Shallow Hydrothermal System

A. Troiano, M.G. Di Giuseppe, A. Fedele, R. Somma, C. Troise, and G. De Natale

Abstract The Solfatara-Pisciarelli area represents the most active zone within the Campi Flegrei caldera (CFc) in terms of hydrothermal manifestations and local seismicity. Periodic injections of hot CO_2-rich fluids at the base of a relatively shallow hydrothermal system has already been correlated to ground uplift in a wide range of numerical modelling of the CFc unrests, that highlight a strong correlation between chemical composition of the Solfatara and Pisciarelli fumaroles, seismicity and ground movements. In particular, a new simulation has been realised via the coupling of TOUGH2©and Comsol Multiphysics©. Recent uplift episodes in the in the centre of Pozzuoli Bay have been reconstructed imposing fluid flows in the system as experimentally recorded. Numerical studies, geochemical data and Magnetotelluric (MT) survey have been integrated, to guess the main features of the shallower part of the hydrothermal system of the Solfatara-Pisciarelli area.

Keywords Multiphysics • Volcano geophysics

1 Solfatara and Pisciarelli Settings

Campi Flegrei caldera (CFc) has been formed by huge eruptions, occurred 39 and 15 Ky B.p., which have been the largest ones occurred in the Mediterranean since the beginning of mankind [19]. Up and down ground movements with rates from centimetres to meters per year characterize the dynamics of this area also during quiescent periods [10] Since 1969, the area started a new phase of uplift after several centuries of subsidence dating back to 1538 AD, when the last eruption occurred in the area. Recent studies on the interpretation of such uplift episodes point out the active role played by the geothermal system, which is characterized by hydrothermal manifestations such as distributed degassing zones and fumaroles

A. Troiano (✉) • M.G. Di Giuseppe • A. Fedele • R. Somma • C. Troise • G. De Natale
Istituto Nazionale di Geofisica e Vulcanologia Osservatorio Vesuviano, Naples, Italy
e-mail: antonio.troiano@ov.ingv.it; mariagiulia.digiuseppe@ov.ingv.it; alessandro.fedele@ov.ingv.it; renato.somma@ov.ingv.it; claudia.troise@ov.ingv.it; giuseppe.denatale@ov.ingv.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*, Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_82

[3, 9]. The Solfatara-Pisciarelli area represents the most active zone within the CFc in terms of hydrothermal manifestations and nowadays local seismicity. The Solfatara volcano is located inside the CFc, about 2 km east-northeast of the town of Pozzuoli. It is a tuff cone formed about 3.7–3.9 Ky B.p., which generated in 1198 AD a low-magnitude hydromagmatic explosive eruption that ejected tephra over a small area ($<1\,km^2$). The crater has a roughly elliptical shape with the two axes of 580 and 770 m, and a maximum elevation of 199 m asl. The Solfatara crater is located very close to the area of maximum ground uplift during the last unrest crises. It hosts large and spectacular fumarole vents, with maximum temperatures in the range 150–160 °C at the Bocca Grande (BG) and Bocca Nuova (BN) and about 100 °C at Le Stufe (LS) and La Fangaia (LF) ones [5] . Systematic measurements of the gas fluxes from the soil evidenced up to 1500 tonnes/day of CO_2 emission which are well aligned with the main fault system and temperature up to 95 °C far from the fumaroles [7, 13]. During the first 16 years of systematic monitoring of the geochemical composition of the BG and BN fumaroles, spanning from 1984 to 2000, the CO_2/H_2O has shown three clear anomalous ratios, occurred in 1986, 1991 and 1995–1996, with molar ratio respectively of 0.30, 0.26 and 0.34 over a background average value of 0.17, peaked about 1 year later from the corresponding unrest ground deformation. Since 2000 the CO_2/H_2O has progressively increased with a nearly linear trend from the background value of 0.17 up to about 0.32 [6]. The Pisciarelli area is located outside the south-east side of the Solfatara crater. It extends from the eastern slopes of the Solfatara volcano to the western margin of the nearby Agnano crater. The Pisciarelli area is characterised by a fumarole field, which is affected by near-surface secondary processes of seasonal character that seem to mask the deeper signals related to the temperature-pressure changes occurring in the hydrothermal system, clearly observed, instead, inside the Solfatara crater at the BG and BN fumarole vents [7]. Starting from 2003, the Pisciarelli field has experienced an evident increase of activity, which has been marked by a sequence of temperature peaks of the fumaroles above the average background temperature of 95 °C, each lasting up to half a year until early 2011, and exceptionally about 1 year, from mid 2011 to mid 2012, the last recorded peak. Furthermore, a nearly linear trend of the peak temperatures, from about 97 °C up to around 112 °C, has been recorded from 2003 up to date. The increase of activity has also been marked by the opening of new vigorous vents and degassing pools, also accompanied by intense local seismic activity. Continuous monitoring of such phenomena is on-going, by permanent networks for seismic, ground deformation and geochemical measurements. Geophysical surveys have so far allowed a quite good knowledge of the subsurface structure of the CFc volcanic system.

2 Electromagnetic Evidences

A 1 km long, nearly W-E directed CSAMT-MT profile crossing the fumaroles field was realised [21], carried out with the aim of deducting an EM model of the structural setting of the hydrothermal system in the first 3 km depth of the Solfatara-Pisciarelli area. The results allow us to identify three EM zones (Fig. 1).

The first EM zone (A) is characterized by a very shallow, electrically conductive body localized beneath the westernmost segment of the profile, which, within a short distance of about 100 m, dips westwards from near surface down to some hundred metres depth. This shallow zone has been ascribed to a water-saturated, high-pressurized geothermal reservoir. The second EM zone (B), which has been localized below the west-central portion of the EM transect, appears as a composite body made of a nearly vertical plume-like structure arising from about 2 km depth to the top edge of the east side of a presumably horizontal plate-like body. Such plume-like structure, centered in correspondence of the Solfatara fumaroles field, rises up to the free surface whereas the plate-like structure deepens at least down to the 3 km of maximum EM exploration depth. The plume-like portion is likely associated with a steam/gas-saturated column and the plate-like portion to a high temperature (>300 °C), over-pressurized, gas-saturated reservoir. Finally, a third EM zone (C), which has been localized beneath the eastern half of the EM transect, corresponding to the Pisciarelli area, is also characterized by the lowest resistivity values (1–10 Ωm) from about 1.2 km down to about 3 km of depth. As it is known, in a volcano- geothermal coastal environment a highly conductive body can indicate either a hydrothermally mineralized, clay-rich layer [23], or a cold seawater-bearing layer [12], or a highly hydrothermalized water-bearing rock [18]. In order to decide which of this hypothesis is the most reliable, we consider that in all of the deep wells drilled by AGIP, during the eighties, at the west border (Mofete area) and north border (San Vito area) of the caldera, the effects of a strong hydrothermal paragenesis have been detected. Abundance of semi- conducting minerals (e.g. pyrrhotite, pyrite, magnetite) and presence of thick argillitic layers, are, in fact, documented at temperatures ranging between 250 and 350 °C, in the depth range between 1 and 3 km, which was the maximum depth reached by the wells [4]. Therefore we are tentatively allowed to associate the very low resistivity zone (C), under the Pisciarelli area, with a hydrothermally mineralized, clay-rich body. Alternatively, we cannot exclude the presence of a deep hydrothermal aquifer, although we know from previous drillings that critical temperature is reached in the whole caldera at depths higher than 3 km. Further consideration arise from the analysis of the seismic P-wave velocity (v_P) and the P-wave/S-wave velocity ratio (v_p/v_s) in the same zones [1]. The conductive C-zone almost completely coincides with a low v_p/v_s area ($v_p/v_s \sim 1.73$). The reason for assuming v_p and v_p/v_s as test parameters resides in the relationship existing between their variations and reservoir fluid phases. In detail, low v_p/v_s values are related to a decrease of v_p in areas with low pore pressure, high heat flow, fracturing and steam/gas saturation in reservoirs, while high v_p/v_s values are found in liquid-saturated high-pressure fields

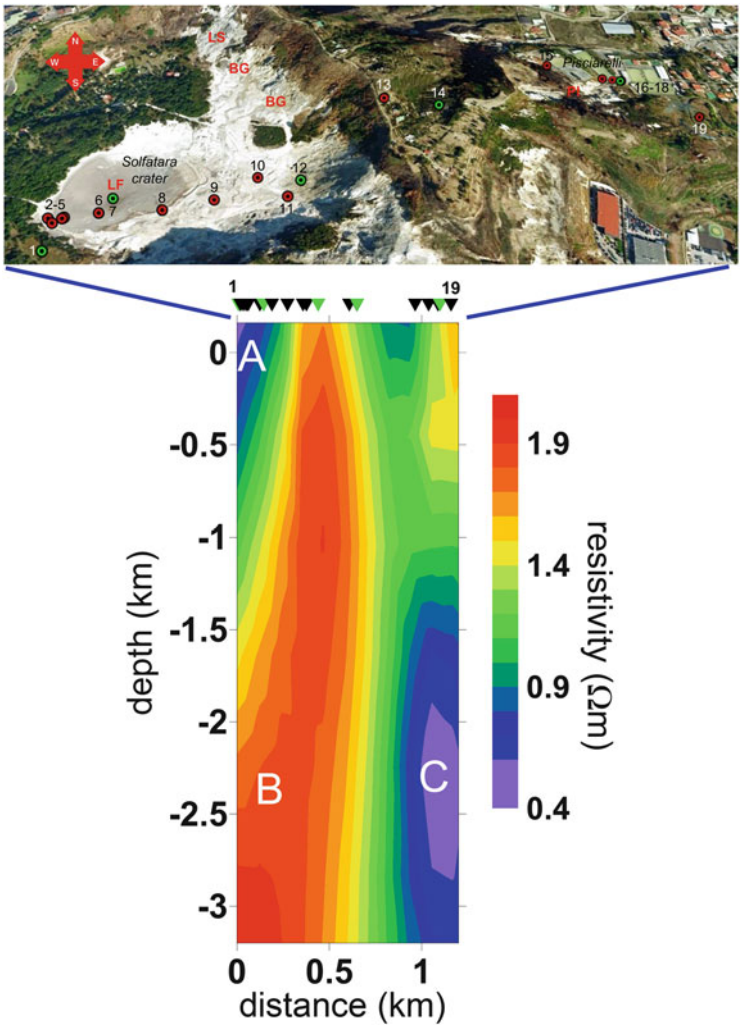


Fig. 1 *Top*: aerial view of the Solfatara crater and surrounding urbanized areas. The *white area* inside the crater is the vegetation-free degassing area. BG, BN, LS and LF indicate the Bocca Grande, Bocca Nuova, Le Stufe and La Fangaia main fumaroles, respectively, located inside the Solfatara crater. PI indicates the Pisciarelli main fumaroles, located outside the crater. *Red and green circlets* indicate the CSAMT and combined CSAMT-MT sounding stations, respectively. *Bottom*: resistivity model obtained from the 1D inversion of the MT data, along the Solfatara-Pisciarelli profile. A common logarithmic scale is used for the resistivity. *Black and green triangles* along the distance scale indicate the CSAMT and combined CSAMT-MT stations, respectively

[15]. It is well established, in fact, that the presence of steam/gas in rocks generally changes the rock compressibility with a v_p decrease, whereas waters in rock voids do not sustain shear stress and decrease the v_s without any v_p variation [22]. It has also been ascertained that the v_p/v_s ratio increases with pressure increase and temperature decrease from vapour-saturated to liquid-saturated conditions [14], and that v_p is affected by the degree of water saturation [16, 17] . The lowest resistivity values that characterize this zone, combined with the seismic evidences, allow us to exclude a water-saturated reservoir, but very likely to admit the presence of a dry and impermeable hydrothermally mineralized, clay-rich body.

3 Geochemical Evidences

An important issue for further discussion is the implication that this EM model, correlated with the evidences emerging from geochemical analysis, can have on the understanding of the fluids up- lift in the Solfatara-Pisciarelli area. According to [2] the peaks of the CO_2/H_2O concentration ratio, occurred in 1986, 1991 and 1995–1996 at the Solfatara crater a few months later an uplift of the ground [6], reflect the increased component of magmatic gases in the composition of the fumaroles, probably due to episodes of intense degassing of magma at depth. The Pisciarelli area is also characterized by emission of gases and fluids through fractures mostly trending N110-120E and mainly NWSE and NE-SW. The main component of the fumaroles is H_2O followed by CO_2 and H_2S and with a range of temperature between 100–110 °C [6]. Fedele [11] During field surveys in the Pisciarelli made during the year 2006 were observed, compared to similar surveys conducted in the past (the year 2005), changes emission style of gases and fluids. Particularly the first are characterised by several point sources of emission while, along the eastern side of the small hill to the east, it is a mud boiling characterised by a diffuse and active degassing zone. A on-line gas monitoring station was localised close the fumaroles field (100 m) during the period May 16–30th 2012, June 1st–5th, 2012. The main relationships of good tracer of magmatic fluids injection such us CO_2/CH_4 and H_2S/CO_2 was reconstructed due to this continuous monitoring [11]. In particular, the CO_2/CH_4 is a good tracer of magmatic fluids injection because CO_2 concentration increased, due to its the higher content of the magmatic component, and CH_4, a gas species formed within the hydrothermal system, is lowered both by dilution and by the more oxidizing, transient conditions caused by the arrival of SO_2 into the hydrothermal system [6, 8]. This opposite behaviour causes rapid increases of the CO_2/CH_4 ratio in fumarolic fluids like it showed by the Fig. 2. This trend seems to be confirmed by the data of GPS ground deformation that show a general tendency to uplift with an acceleration of the phenomenon in the period spanning from June to August 2012 (25 mm/month in average) and increasing during the last month beginning on December 2012 (10 mm/month), as also shown in Fig. 2.

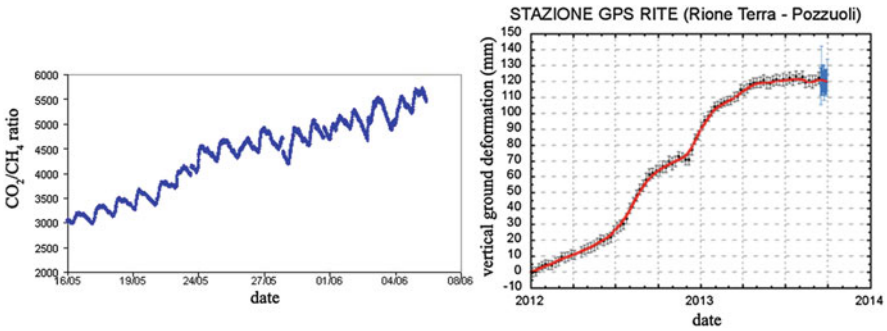


Fig. 2 On the *left* CO_2/CH_4 ratio from 16/05/2012 to 05/06/2012 measured by Quadrupole Mass Spectrometer. On the *right* ground deformation (from Osservatorio Vesuviamo website)

4 Discussion and Conclusion

In effect, periodic injections of hot CO_2-rich fluids at the base of a relatively shallow hydrothermal system has been correlated to ground uplift in a wide range of numerical modelling of the CFc unrests, that highlight a strong correlation between chemical composition of the Solfatara and Pisciarelli fumaroles, seismicity and ground movements [20]. In particular, a new simulation has been realised via the coupling of TOUGH2©and Comsol Multiphysics©. Recent uplift episodes in the in the centre of Pozzuoli Bay have been reconstructed imposing fluid flows in the system as experimentally recorded. The comparison between numerical simulation, geochemical data and EM survey highlight the main features of the shallower part of the hydrothermal system of the Pisciarelli area. The high CO_2/CH_4 ratio indicate a plausible magmatic component. For such magmatic origin, the plume identified in the MT imaging below the Solfatara crater seems to contribute also to fluid flow uplift below Pisciarelli. The low resistivity values under Pisciarelli, that indicate a strong local fluid circulation, support this kind of hypothesis. The fluid flow patterns reconstructed by our numerical simulations enforce this interpretation (Fig. 3). Fluids migrate, in the upper part of our model, from its central part, ideally placed below the Solfatara crater, toward an area localised some hundreds of meters away, fitting the Pisciarelli zone. The clear evidence that the thermodynamic condition of the system in the shallower part results compatible with the presence of convective cells enforce the idea that the degassing of the magma batch localised under the Solfatara crater contribute also to fluid circulation under Pisciarelli.

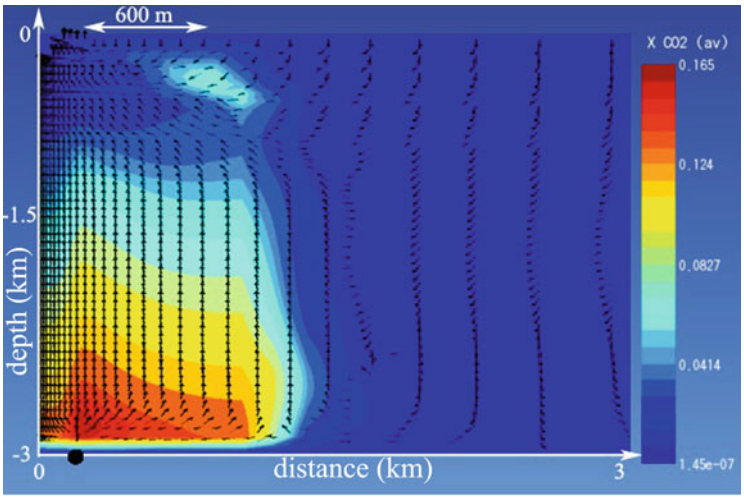


Fig. 3 Fluid flows patterns as reconstructed by numerical simulation. *Black arrows* show the CO_2 fluxes migrating from the injection point, placed below the Solfatara crater, towards the surface, ending in the Pisciarelli area. *Colour* contours shows the CO_2 mass fraction

References

1. Battaglia, J., Zollo, A., Virieux, J., Iacono, D.D.: Merging active and passive data sets in traveltime tomography: the case study of Campi Flegrei caldera (Southern Italy). Geophys. Prospect. **56**(4), 555–573 (2008)
2. Caliro, S., Chiodini, G., Moretti, R., Avino, R., Granieri, D., Russo, M., Fiebig, J.: The origin of the fumaroles of la solfatara (Campi Flegrei, South Italy). Geochim. Cosmochim. Acta **71**(12), 3040–3055 (2007)
3. Carlino, S., Somma, R., Troise, C., De Natale, G.: The geothermal exploration of Campanian volcanoes: historical review and future development. Renew. Sust. Energ. Rev. **16**(1), 1004–1030 (2012)
4. Chelini, W., Sbrana, A.: Phlegraean Fields. Subsurface Geology, vol. 9. Consiglio Nazionale Delle Ricerche, Roma (1987)
5. Chiodini, G., Frondini, F., Cardellini, C., Granieri, D., Marini, L., Ventura, G.: CO_2 degassing and energy release at Solfatara volcano, Campi Flegrei, Italy. J. Geophys. Res. Solid Earth (1978–2012) **106**(B8), 16213–16221 (2001)
6. Chiodini, G., Caliro, S., Cardellini, C., Granieri, D., Avino, R., Baldini, A., Donnini, M., Minopoli, C.: Long-term variations of the Campi Flegrei, Italy, volcanic system as revealed by the monitoring of hydrothermal activity. J. Geophys. Res. Solid Earth **115**(B3), 2156–2202 (2010)
7. Chiodini, G., Avino, R., Caliro, S., Minopoli, C.: Temperature and pressure gas geoindicators at the Solfatara fumaroles (Campi Flegrei). Ann. Geophys. **54**, 151–160 (2011)
8. Chiodini, G., Caliro, S., De Martino, P., Avino, R., Gherardi, F.: Early signals of new volcanic unrest at Campi Flegrei caldera? Insights from geochemical data and physical simulations. Geology **40**(10), 943–946 (2012)
9. De Natale, G., Pingue, F., Allard, P., Zollo, A.: Geophysical and geochemical modelling of the 1982–1984 unrest phenomena at Campi Flegrei Caldera (Southern Italy). J. Volcanol. Geotherm. Res. **48**(1–2), 199–222 (1991)

10. Dvorak, J.J., Mastrolorenzo, G.: The mechanism of recent vertical crustal movements in Campi Flegrei caldera. Southern Italy, Geological Society of America, Special Papers 263 (1991)
11. Fedele, A.: Continuous geochemical monitoring by mass-spectrometer in the Campi Flegrei geothermal area. An application at Pisciarelli-Solfatara (diffuse and fumarolic gases) and at the mud gases during drilling of the CFDDP pilot hole. Ph.D. thesis, Alma Mater Studiorium Universitàă degli Studi di Bologna (2013)
12. Goldman, M., Gilad, D., Ronen, A., Melloul, A.: Mapping of seawater intrusion into the coastal aquifer of Israel by the time domain electromagnetic method. Geoexploration **28**(2), 153–174 (1991)
13. Granieri, D., Avino, R., Chiodini, G.: Carbon dioxide diffuse emission from the soil: ten years of observations at Vesuvio and Campi Flegrei (Pozzuoli), and linkages with volcanic activity. Bull. Volcanol. **72**(1), 103–118 (2010)
14. Ito, H., De Vilbiss, J., Nur, A.: Compressional and shear waves in saturated rock during water-steam transition. J. Geophys. Res. Solid Earth **84**(B9), 4731–4735 (1979)
15. Mormone, A., Tramelli, A., Di Vito, M.A., Piochi, M., Troise, C., De Natale, G.: Secondary hydrothermal minerals in buried rocks at the Campi Flegrei caldera, Italy: a possible tool to understand the rock-physics and to assess the state of the volcanic system. Periodico Mineral. **80**, 385 (2011)
16. O'Connell, R.J., Budiansky, B.: Seismic velocities in dry and saturated cracked solids. J. Geophys. Res. **79**(35), 5412–5426 (1974)
17. O'Connell, R.J., Budiansky, B.: Viscoelastic properties of fluid-saturated cracked solids. J. Geophys. Res. **82**(36), 5719–5735 (1977)
18. Patella, D., Rossi, A., Tramacere, A.: First results of the application of the dipole electrical sounding method in the geothermal area of Travale-Radicondoli (Tuscany). Geothermics **8**(2), 111–134 (1979)
19. Rosi, M., Sbrana, A.: Phlegrean Fields, vol. 9. Consiglio Nazionale Delle Ricerche, Roma (1987)
20. Troiano, A., Di Giuseppe, M.G., Petrillo, Z., Troise, C., De Natale, G.: Ground deformation at calderas driven by fluid injection: modelling unrest episodes at Campi Flegrei (Italy). Geophys. J. Int. **187**(2), 833–847 (2011)
21. Troiano, A., Di Giuseppe, M.G., Patella, D., Troise, C., De Natale, G.: Electromagnetic outline of the Solfatara–Pisciarelli hydrothermal system, Campi Flegrei (Southern Italy). J. Volcanol. Geotherm. Res. **277**(0), 9–21 (2014)
22. Vanorio, T., Prasad, M., Patella, D., Nur, A.: Ultrasonic velocity measurements in volcanic rocks: correlation with microtexture. Geophys. J. Int. **149**(1), 22–36 (2002)
23. Ward, S.H.: Resistivity and induced polarization methods. Geotech. Environ. Geophys. **1**, 147–189 (1990)

MS 25
MINISYMPOSIUM: OPTIMIZATION AND OPTIMIZATION-BASED CONTROL METHODS FOR INDUSTRIAL APPLICATIONS

Organizers

Kathrin Flaßkamp[1] and Timm Faulwasser[2]

Speakers

Jürgen Pannek[3]
Logistics Driven Requirements and Limitations to Model Predictive Control

Milan Korda[4], Faran Ahmed Qureshi[5], Tomasz T. Gorecki[6] and Colin N. Jones[7]
Periodic Stochastic Model Predictive Control Applied to Building Temperature Control

[1]Kathrin Flaßkamp, University of Paderborn, Germany.
[2]Timm Faulwasser, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
[3]Jürgen Pannek, University of Bremen, Germany.
[4]Milan Korda, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
[5]Faran Ahmed Qureshi, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
[6]Tomasz T. Gorecki, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
[7]Colin N. Jones, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_83

Karl Worthmann[8], Christopher Kellett[9], Philipp Braun[10], Lars Grüne[11] and Steven Weller[12]
Distributed Model Predictive Control for a Smart Grid Application

Michael Dellnitz[13], Julian Eckstein[14], Kathrin Flaßkamp[1], Patrick Friedel[15], Christian Horenkamp[16], Ulrich Köhler[17], Sina Ober-Blöbaum[18], Sebastian Peitz[19] and Sebastian Tiemeyer[20]
Multiobjective Optimal Control Methods for the Development of an Intelligent Cruise Control

Johann C. Dauer[21], Timm Faulwasser[2] and Sven Lorenz[22]
Computational Aspects of Optimization-Based Path Following of an Unmanned Helicopter

Valeria Artale[23], Cristina Milazzo[24], Calogero Orlando[25] and Angela Ricciardello[26]
Mathematical Applications to an Unmanned Aerial Vehicle

Sabrina Fiege[27], Andreas Griewank[28] and Andrea Walther[29]
A Optimization Method for Piecewise Linear Problems

[8]Karl Worthmann, TU Ilmenau, Germany.
[9]Christopher Kellett, University of Newcastle, Australia.
[10]Philipp Braun, University of Bayreuth, Germany.
[11]Lars Grüne, University of Bayreuth, Germany.
[12]Steven Weller, University of Newcastle, Australia.
[13]Michael Dellnitz, University of Paderborn, Germany.
[14]Julian Eckstein, Hella KGaA Hueck & Co., Germany.
[15]Patrick Friedel, Hella KGaA Hueck & Co., Germany.
[16]Christian Horenkamp, University of Paderborn, Germany.
[17]Ulrich Köhler, Hella KGaA Hueck & Co., Germany.
[18]Sina Ober-Blöbaum, University of Paderborn, Germany.
[19]Sebastian Peitz, University of Paderborn, Germany.
[20]Sebastian Tiemeyer, Hella KGaA Hueck & Co., Germany.
[21]Johann C. Dauer, DLR Braunschweig, Germany.
[22]Sven Lorenz, DLR Braunschweig, Germany.
[23]Valeria Artale, Cittadella Universitaria, Italy.
[24]Cristina Milazzo, Cittadella Universitaria, Italy.
[25]Calogero Orlando, Cittadella Universitaria, Italy.
[26]Angela Ricciardello, Cittadella Universitaria, Italy.
[27]Sabrina Fiege, University of Paderborn, Germany.
[28]Andreas Griewank, Humboldt University, Germany.
[29]Andrea Walther, University of Paderborn, Germany.

Jan Kuátko[30] and Stefan Ratschan[31]
Global Multiple Shooting for Hybrid Dynamical Systems

Keywords

Optimal control
Optimization
Optimization-based control

Short Description

The increasing need for resource and energy efficiency in industrial applications leads to manifold scientific and technical challenges in the design and operation of industrial systems and processes. Typical examples of such challenges are, for instance, energy-efficient operation close to safety-critical constraints, online determination of optimal operation policies, and the need for an optimal trade-off between conflicting objectives. Mathematical optimization and optimal control methods are key enabling technologies to tackle these challenges. They play a crucial role in various fields of applications ranging from aeronautics and aerodynamics, automotive, space mission design, and robotics to biomechanics and process industries. This mini-symposium provides an overview of current trends of mathematical optimization and optimization-based control methods for industrial applications.

[30] Jan Kuátko, Academy of Sciences of the Czech Republic, Czech Republic.
[31] Stefan Ratschan, Academy of Sciences of the Czech Republic, Czech Republic.

Computational Aspects of Optimization-Based Path Following of an Unmanned Helicopter

Johann C. Dauer, Timm Faulwasser, and Sven Lorenz

Abstract This paper considers the path following of unmanned helicopters based on dynamic optimization. We assume that the helicopter is equipped with a flight control system that provides an approximation of its closed-loop dynamics. The task at hand is to compute inputs for this flight control system in order to track a geometrically specified path. A concise problem formulation and a discussion of an efficient implementation are presented. The implementation achieves computation times below the flight duration of the path by exploiting differential flatness properties of parts of the dynamics. Finally, we present quantitative results with respect to convergence and required iterations for a challenging nonlinear path. We show that the proposed optimization based approach is capable of tackling nonlinear path following for unmanned helicopters in an efficient and practicable manner.

Keywords Approximation of closed-loop dynamics • Dynamic optimization • Flight control • Path following

1 Introduction

In this contribution the problem of *path following* of small unmanned helicopters such as the one shown in Fig. 1 is considered. Here, path following is defined as the task to fly along a geometrically specified space curve. The time-wise progress on the path is not a priori known or specified. Rather, we allow mission based requirements such as, for instance, a desired velocity along the path. It is assumed that a mathematical representation of a path is available, provided either directly in the mission specification or by a path planner, as presented e.g. in [1]. Furthermore, we assume that there exists a flight control system handling the stabilization of the helicopter and allows to derive an approximation of the closed-loop dynamics.

J.C. Dauer (✉) • S. Lorenz
German Aerospace Center (DLR e.V.), Institute of Flight Systems, Braunschweig, Germany
e-mail: johann.dauer@dlr.de; sven.lorenz@dlr.de

T. Faulwasser
Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: timm.faulwasser@epfl.ch

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_84

Fig. 1 Automated helicopter midiARTIS of the German Aerospace Center, maximum-take-off-weight 14 kg, rotor-diameter 2 m

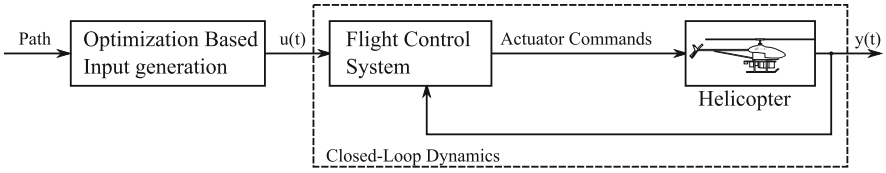


Fig. 2 Simplified block diagram of the optimization-based generation of inputs for the flight control system

Our task is to find suitable inputs for the flight control system, which steers the helicopter along the desired path. These inputs have to satisfy dynamic constraints of the vehicle as well as limitations of the control system implementation which would otherwise cause path deviations. These aspects are shown in the simplified block diagram of Fig. 2, where the block considered here is called "Optimization Based Input Generation". As shown in [3], a problem formulation like this can be tackled by dynamic optimization using a receding horizon approach. The closed-loop dynamics of the figure is not known exactly. However, the control concept presented in [11] is based on a reference model, which can be regarded as an approximation of the closed-loop. A good starting point for a literature review of alternative approaches can be found in the surveys [2, 9].

In the present paper, the results of [3] are extended by details of the numerical implementation of the problem. We present an implementation capable to solve this kind of path following problem. The implementation is based on the open-source project ACADO Toolkit [8]. A nonlinear representation of the closed-loop behavior of the helicopter is considered as well as nonlinear paths that do not correspond to paths created by trimmed trajectories.

2 Problem Formulation

This section gives an overview on the problem formulation of path following for an unmanned helicopter. A space curve is defined, which the helicopter is supposed to track. An optimal control problem (OCP) is formulated afterwards.

The closed-loop approximation of the flight control system considered here can be represented in state space by

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{1a}$$

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t)), \tag{1b}$$

where the state vector $\mathbf{x}$ contains position and velocity of the helicopter, as well as rotational and engine states. The inputs $\mathbf{u}$ are the input channels of the flight control system. They contain a velocity command $\mathbf{u}_v$ and a command for a desired scalar azimuth u_ψ. The azimuth is the third component of an Euler representation of the helicopter's attitude corresponding to the nomenclature defined in ISO 1151. The outputs are defined by the position of the helicopter ($\mathbf{r}$) in Cartesian north-east-down (NED) frame and the azimuth (ψ), thus

$$\mathbf{u}(t) = (\mathbf{u}_v(t)^T, u_\psi(t))^T \in \mathbb{R}^4, \tag{2a}$$

$$\mathbf{y}(t) = (\mathbf{r}(t)^T, \psi(t))^T \in \mathbb{R}^4. \tag{2b}$$

Representations (1) and (2) result in a consistent problem formulation for missions in obstacle occupied environments for short distance missions. Alternative formulations can also be considered, for example velocity in respect to wind as well as the sideslip angle, which partly defines the attitude of the helicopter with respect to aerodynamic inflow.

In this paper we focus on the structure of the problem. The information needed to reconstruct the complete set of equations can be found in [3]. For brevity, the following paragraphs are limited to the general idea and the required links to [3] by providing the physical meaning of some of the variables. The state vector can be subdivided into four sub-vectors: the translation dynamics $\mathbf{x}_t \in \mathbb{R}^6$, states of the engine $\mathbf{x}_e \in \mathbb{R}^2$, states for the rotational rates $\mathbf{x}_r \in \mathbb{R}^6$ and states of the quaternion based representation of the attitude $\mathbf{x}_q \in \mathbb{R}^4$. The closed-loop dynamics can be represented under the following structure

$$\dot{\mathbf{x}} = \begin{pmatrix} \dot{\mathbf{x}}_e(t) \\ \dot{\mathbf{x}}_r(t) \\ \dot{\mathbf{x}}_q(t) \\ \dot{\mathbf{x}}_t(t) \end{pmatrix} = \begin{pmatrix} \mathbf{A}_e\mathbf{x}_e(t) + \mathbf{g}_e(\mathbf{x}(t), \mathbf{u}_v(t)) \\ \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{g}_r(\mathbf{x}(t), \mathbf{u}_v(t), u_\psi(t)) \\ \mathbf{f}_q(\mathbf{x}_q(t), \mathbf{x}_r(t)) \\ \mathbf{f}_t(\mathbf{x}(t)) \end{pmatrix}, \quad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^{18}. \tag{3}$$

Note that the two components $\mathbf{x}_e$ and $\mathbf{x}_r$ have linear state mapping and nonlinear input functions, the components $\mathbf{x}_q$ and $\mathbf{x}_t$ have nonlinear dynamics but are not directly influenced by the inputs. The optimization problem can be simplified by exploiting the fact that the engine dynamics $\mathbf{x}_e$ are differentially flat with the thrust of the main rotor as flat output if the remaining states are considered as parameters; see [7] for details on differential flatness. Thus it is possible to calculate the input $\mathbf{u}_v$ if a sufficiently smooth thrust trajectory and its derivatives are given. An equivalent argumentation holds for the dynamics of $\mathbf{x}_r$ considering the remaining states and $\mathbf{u}_v$ as parameters. The flat outputs in this case are the body-fixed rotation rates.

Thus, these flat subsystems allow to define simple integrator chains of sufficient order with new inputs for the thrust u_T and rotation rates $\mathbf{u}_\omega$, in the following represented using the sparse system matrices $\mathbf{A}_{1,2}$ and input matrices $\mathbf{B}_{1,2}$ containing only a small number of ones. It is thus possible to derive the state trajectories for $\mathbf{x}_e$ and $\mathbf{x}_r$ only by integration of the new inputs and using the algebraic relations of the flat outputs, which results in modifications of $\mathbf{f}_q$ and $\mathbf{f}_t$. The original system inputs can be calculated in the same way. Finally a modified structure is obtained

$$\dot{\mathbf{x}}_m = \begin{pmatrix} \dot{\mathbf{x}}_{e,m}(t) \\ \dot{\mathbf{x}}_{r,m}(t) \\ \dot{\mathbf{x}}_q(t) \\ \dot{\mathbf{x}}_t(t) \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1\mathbf{x}_{e,m}(t) + \mathbf{B}_1 u_t(t)) \\ \mathbf{A}_2\mathbf{x}_{r,m}(t) + \mathbf{B}_2\mathbf{u}_\omega(t) \\ \mathbf{f}_{q,m}(\mathbf{x}_q(t), \mathbf{x}_{r,m}(t)) \\ \mathbf{f}_{t,m}(\mathbf{x}_m(t)) \end{pmatrix}, \quad \mathbf{x}_m(0) = \mathbf{x}_{m,0} \in \mathbb{R}^{18}. \tag{4}$$

The control system also imposes constraints on the optimization problem. These constraints are defined by the so-called envelope protection. We omit details here and represent the constraint sets of the states as X and of the inputs as U. The details can be found in [3], where constraints for accelerations, velocities, in both body-fixed as well as in NED frame, and actuator deflections are introduced.

The path that has to be tracked by the helicopter is a four dimensional parametric curve depending on a scalar parameter θ. It is defined in the output space of (1)

$$\mathscr{P} = \left\{ p(\theta) \in \mathbb{R}^4 \mid \theta \in [\theta_0, \theta_1] \mapsto \left(\mathbf{r}^T(\theta), \psi(\theta)\right)^T \right\}. \tag{5}$$

For the time-wise evolution of the position on the path, we introduce artificial dynamics with an input v which augment the dynamics of the system [5] and is chosen to be a double integrator

$$\dot{\mathbf{z}}(t) = \begin{pmatrix} \dot{z}_1 \\ \dot{z}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \mathbf{z}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} v(t) =: \mathbf{l}(\mathbf{z}(t), v(t)), \quad \mathbf{z}(0) = \mathbf{z}_0, \tag{6a}$$

$$\theta(t) = z_1(t). \tag{6b}$$

Higher degrees of the path-dynamics would increase the smoothness of the evolution along the path. It would, however, increase the computational burden as well.

The second order integrator allows us to specify a desired velocity along the path later on.

Now, the optimal control problem can be defined: Given the closed-loop approximation according to (1) and a path of the form (5), calculate the evolution $\theta : t \in [t_0, t_1] \mapsto \theta(t) \in [\theta_0, \theta_1]$ and the inputs $\mathbf{u}$, such that (a) the constraints are satisfied, (b) the helicopter moves forward on the path ($\dot{\theta} > 0$) and (c) a cost function is minimized:

$$\underset{\mathbf{u}_\omega(\cdot),u_t(\cdot),v(\cdot)}{\text{minimize}} \quad \int_{t_k}^{t_k+T} \underbrace{\left\|(\mathbf{e}^T(t), \dot{\mathbf{e}}^T(t))\right\|_{\mathbf{Q}_e}^2}_{\text{path deviation}} + \underbrace{\|\mathbf{z}(t) - \mathbf{z}_r(t)\|_{\mathbf{Q}_z}^2}_{\text{reference behavior}} + \underbrace{\left\|(\mathbf{u}^T(t), v(t))\right\|_{\mathbf{R}}^2}_{\text{regularization}} dt, \tag{7a}$$

subject to the dynamics and constraints

$$\dot{\mathbf{x}}_m(t) = \mathbf{f}_m(\mathbf{x}_m(t), u_T(t), \mathbf{u}_\omega(t)), \quad \mathbf{x}_m(0) = \mathbf{x}_{m,0} \in \mathbb{R}^{18} \tag{7b}$$

$$\dot{\mathbf{z}}(t) = \mathbf{l}(\mathbf{z}(t), v(t)), \quad \mathbf{z}(0) = \mathbf{z}_0 \in \mathbb{R}^2 \tag{7c}$$

$$\dot{\mathbf{e}}(t) = \frac{\partial \mathbf{h}}{\partial \mathbf{x}_m}\mathbf{f}_m(\mathbf{x}_m(t), u_T(t), \mathbf{u}_\omega(t)) - \frac{\partial \mathbf{p}}{\partial \theta}\dot{\theta}, \quad \mathbf{e}(0) = \mathbf{e}_0 \in \mathbb{R}^4 \tag{7d}$$

$$\mathbf{x}(t) \in X, \mathbf{u}(t) \in U. \tag{7e}$$

The path error $\mathbf{e}$ leads to tracking of the reference path and its derivative avoids solutions oscillating around it. The reference behavior term in (7a) allows us to specify dynamic requirements like a desired velocity along the path and finally the regulation enforces certain smoothness on the derived state trajectories. Optimization on a whole path can be very computational extensive. In order to limit the computation time we thus apply a receding horizon approach with prediction horizon T, i.e. for each point in time $t_k = \delta k, k \in \mathbb{N}$ we solve (7) over the horizon $[t_k, t_k + T]$.

3 Implementation and Computational Results

The path optimization described in the previous section has been carried out with the help of the open-source project ACADO Toolkit [8]. An advantage of this project is what the authors refer to as code generation. A piece of self-contained code is generated based on the mathematical problem formulation. This code contains an efficient implementation of the optimization problem based on a tailored discretization. This code can then either be used separately, interfacing with MATLAB or directly integrated into another piece of software for the desired application.

The optimization is performed using a direct multiple shooting approach [12]. The length of the prediction horizon T is chosen taking the computation time and a minimal stopping distance into account. The minimal stopping distance can be transferred into a minimal prediction horizon using the velocity maximally allowed and the deceleration limits which are implemented in the flight control system.

The prediction horizon is subdivided into equal shooting intervals. The solution of the differential equations over these multiple-shooting intervals is normally performed using adaptive step size integration algorithms. Using adaptive integrators has the advantage that the integration grid does not have to be specified a priori. However, it comes at the cost of non-deterministic discretization. This is why in [8] the use of fixed-step size integrators is proposed. The integration algorithm and its step-size has to be determined a priori, e.g. heuristically. By doing so, it is possible to tailor a discretization that is deterministic in calculation time and allows the generation of efficient code exploiting aspects like static memory allocation.

By these means, a nonlinear program (NLP) is formulated for each shift of the prediction horizon that, if feasible, directly solves the optimization problem. The NLP is solved in an iterative process known as sequential quadratic programming (SQP). In each sequential step a quadratic approximation of the cost function is created as well as affine approximations of the constraints thus creating a quadratic program (QP). There exist powerful methods to solve QP. Here, the qpOASES Package [6] is used for that purpose. Using the solution of the QP, the original NLP is approximated again and this process is repeated until a certain residual of the KKT-conditions is sufficiently small. A comprehensive tutorial on this process can be found in [4].

At the beginning of the path an initial guess is used, which corresponds to the hover states of the helicopter. These conditions can be determined by trim calculations [10] or simulation experiments. Using this initial guess, the first NLP is solved over the prediction horizon $T = 2.5\,\mathrm{s}$. Each prediction horizon is subdivided into 25 shooting intervals, of which each is solved using an implicit Runge-Kutta integrator of second order with three discretization steps. The number of SQP formulations that are required to achieve residuals of the KKT-conditions that we define to be less than 10^{-4} are referred to as SQP iterations in the following. Each SQP iteration requires the solution of a QP that again is solved iteratively and is called QP iterations in the following. After convergence, the inputs on the first shooting interval is used for the final solution trajectory. The remaining intervals serve as initial guess after shifting the prediction horizon by one shooting interval.

The path shown in Fig. 3 shall serve as an example. It is a clover leaf with a height profile, which is generated using

$$\mathbf{r}_p = \begin{pmatrix} \hat{r}\cos\left(3\theta - \frac{3}{2}\pi\right)\cos(\theta) \\ \hat{r}\cos\left(3\theta - \frac{3}{2}\pi\right)\sin(\theta) \\ \hat{h}\sin(4\theta) \end{pmatrix}, \quad \hat{r} = 25\,\mathrm{m}, \quad \hat{h} = 3\,\mathrm{m}, \tag{8}$$

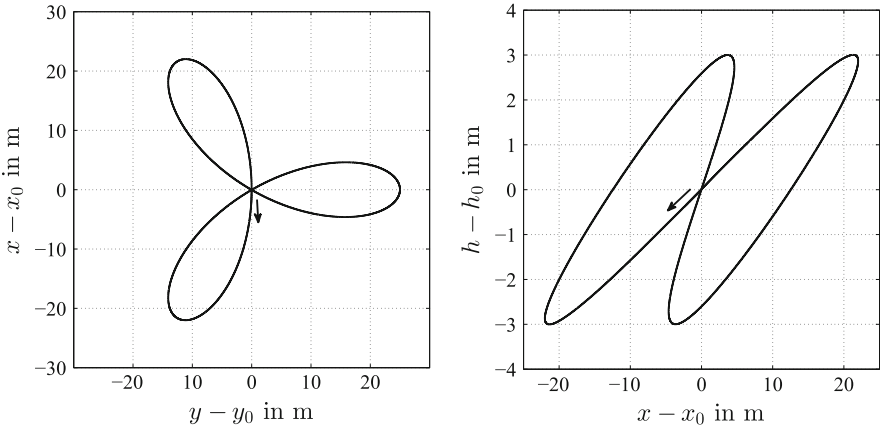


Fig. 3 Example path: clover *leaf* with sinusoidal height profile. The arrows indicate the start and direction of flight, while the *left* shows the *top-view* and *right* the height over north direction

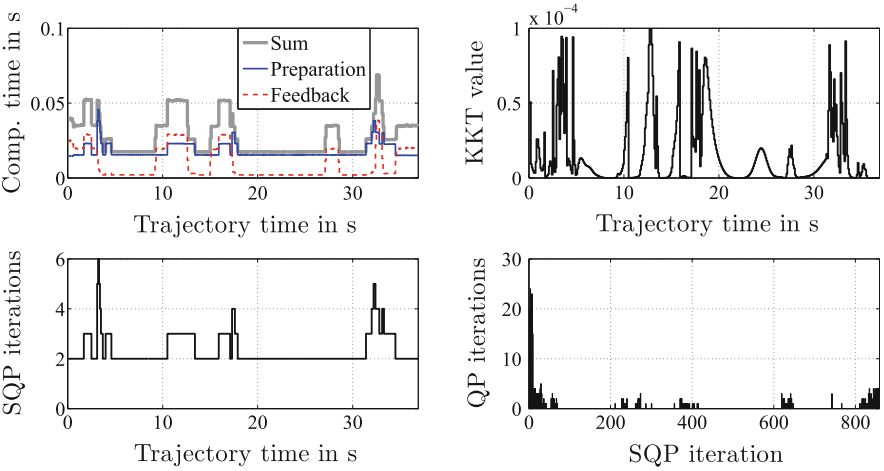


Fig. 4 Calculation characteristics of the example path

where $\hat{r}$ is the clover leaf's radius and $\hat{h}$ the amplitude of the height profile. The fourth component of the path, the azimuth, has been defined such that the helicopter is always oriented tangential to the movement resulting in flight without sideslip angle in the wind-free case.

Figure 4 shows the computation characteristics of the optimization; corresponding input and state trajectories can be found in [3] for similar paths. The upper left plot presents the required computation time over the timeline of the solution trajectory. The plot shows how long it takes to compute one NLP that is to optimize once over the prediction horizon. Additionally, the computation time of the preparation is shown, which contains the discretization and the setup of the QP

Problem. The feedback time refers to the time needed to solve the QP. The lower left plot depicts how many QP approximations are necessary in total to solve each NLP. The lower right plot shows the number of iterations necessary to solve a certain QP, which is the only graph not having the trajectory time as abscissa. As for each NLP multiple QP have to be solved the number of iterations for each SQP step are shown.

From the achieved KKT values (upper right plot) it appears that the formulated problem is always feasible since the maximum value allowed is never exceeded. The desired velocity was set to be 8 m/s resulting in a duration of less than 40 s flight time. The overall computation on a standard desktop computer without parallelization takes around 10 s. This alone is a good result, as it means that the optimization over the complete path takes less than a third of the time it takes to fly it.

There are several regions of higher computational burden that can be seen in the of the upper left plot. These increase in computational cost is caused by an increased active set of constraints. Most of these regions are significantly influenced by the number of SQP iteration. However, one region is created by the QP iterations alone. The discretization of the problem thus gives a lower bound on the needed computation time, which is in this case around 0.02 s for an prediction horizon of 2.5 s. Nevertheless, the solution of the QP problems has significant impact on the overall computational burden as well and lies around the same order of magnitude.

4 Conclusion

The proposed approach is well capable to tackle the problem of path following for unmanned helicopters where an approximation of the closed-loop dynamics is available. It is shown that trajectories can be generated faster than it would take to fly them. This fact makes the presented approach very powerful as a great variety of paths and reference variations can be solved in reasonable time. Applications can be mission planning purposes or generation of controller inputs that can be stored in a maneuver database. However, an upper bound of the time needed for prediction is not guaranteed. Future work will thus investigate possibilities to enable any-time capability which would render the approach also applicable for online purposes. The results of this paper show that an adaption has to consider both, the SQP iterations and QP iterations as both have significant impact on the computation time.

References

1. Adolf, F.M., Andert, F.: Rapid multi-query path planning for a vertical take-off and landing unmanned aerial vehicle. J. Aerosp. Comput. Inf. Commun. **8**(11), 310–327 (2011)
2. Dadkhah, N., Mettler, B.: Survey of motion planning literature in the presence of uncertainty: considerations for UAV guidance. J. Intell. Robot. Syst. **65**(1–4), 233–246 (2012). doi: 10.1007/s10846-011-9642-9

3. Dauer, J.C., Faulwasser, T., Lorenz, S., Findeisen, R.: Optimization-based feedforward path following for model reference adaptive control of an unmanned helicopter. In: Proceedings of the AIAA Guidance, Navigation, and Control Conference 2013, AIAA 2013–5002, Boston, MA (2013). doi:10.2514/6.2013-5002
4. Diehl, M., Bock, H., Diedam, H., Wieber, P.B.: Fast direct multiple shooting algorithms for optimal robot control. In: Diehl, M., Mombaur, K. (eds.) Fast Motions in Biomechanics and Robotics. Lecture Notes in Control and Information Sciences, vol. 340, pp. 65–93. Springer, Berlin/Heidelberg (2006)
5. Faulwasser, T.: Optimization-Based Solutions to Constrained Trajectory-Tracking and Path-Following Problems. Shaker, Aachen (2013). doi:10.2370/9783844015942
6. Ferreau, H., Bock, H., Diehl, M.: An online active set strategy to overcome the limitations of explicit MPC. Int. J. Robust Nonlinear Control **18**(8), 816–830 (2008)
7. Fliess, M., Lévine, J., Martin, P., Rouchon, P.: Flatness and defect of non-linear systems: introductory theory and examples. Int. J. Control. **61**(6), 1327–1361 (1995)
8. Houska, B., Ferreau, H., Diehl, M.: ACADO toolkit – an open-source framework for automatic control and dynamic optimization. Opt. Control Appl. Methods **32**(3), 298–312 (2011)
9. Huang, G.Q., Lu, Y.P., Nan, Y.: A survey of numerical algorithms for trajectory optimization of flight vehicles. Sci. China Technol. Sci. **55**(9), 2538–2560 (2012). doi:10.1007/s11431-012-4946-y
10. Leishmann, J.G.: Principles of Helicopter Aerodynamics, 2nd edn. Cambridge University Press, Cambridge (2006)
11. Lorenz, S.: Open-loop reference system for nonlinear control applied to unmanned helicopters. J. Guid. Control Dyn. **35**(1), 259–269 (2012). doi:10.2514/1.52033
12. Nocedal, J., Wright, S.: Numerical Optimization, 2nd edn. Springer Series in Operations Research and Financial Engineering. Springer, New York (2000)

Model Predictive Control of Residential Energy Systems Using Energy Storage and Controllable Loads

Philipp Braun, Lars Grüne, Christopher M. Kellett, Steven R. Weller, and Karl Worthmann

Abstract Local energy storage and smart energy scheduling can be used to flatten energy profiles with undesirable peaks. Extending a recently developed model to allow controllable loads, we present Centralized and Decentralized Model Predictive Control algorithms to reduce these peaks. Numerical results show that the additional degree of freedom leads to improved performance.

Keywords Centralized model predictive control • Decentralized model predictive control • Model predictive control

1 Introduction

Widespread uptake of local electricity generation technologies such as solar photovoltaics and wind turbines are leading to undesirable voltage swings in electricity distribution networks. Large variations in the grid profile, resulting from periods of high local energy generation followed by periods of high power demand require significant network infrastructure and can lead to a degradation of power quality and even outages. In response to these challenges local energy storage is increasingly considered to reduce the peak demand [4, 5]. Additionally, a recent study [1] suggests that up to 60 % of the consumption of a household, in the form of appliances such as air conditioners and refrigerators, is elastic or schedulable.

P. Braun (✉) • L. Grüne
Mathematisches Institut, Universität Bayreuth, 95440 Bayreuth, Germany
e-mail: philipp.braun@uni-bayreuth.de; lars.gruene@uni-bayreuth.de

C.M. Kellett • S.R. Weller
School of Electrical Engineering and Computer Science at the University of Newcastle, Callaghan, NSW 2308, Australia
e-mail: chris.kellett@newcastle.edu.au; steven.weller@newcastle.edu.au

K. Worthmann
Institut für Mathematik, Technische Universität Ilmenau, 99693 Ilmenau, Germany
e-mail: karl.worthmann@tu-ilmenau.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*, Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_85

Therefore, an alternate, but complementary, approach to the use of energy storage devices to reduce the grid variations involves energy consumption scheduling [3, 7].

We consider a small, neighborhood-level, electricity network consisting of several residences. Each residence comprises a Residential Energy System (RES), consisting of a residential load, a local energy storage element, and solar photovoltaic panels. Each RES is connected to the wider electricity network. For the sake of simplicity, we refer to the storage element as a battery though fuel cells also satisfy our proposed energy storage model and constraints. The important contribution with respect to our previous work [8] is the extension of the model to handle controllable or elastic loads. While the extension of the model is trivial, the resulting constraints are not obvious.

The paper is organized as follows. The extended model is introduced in Sect. 2 together with two performance metrics. Section 3 introduces two Model Predictive Control algorithms and shows how to incorporate controllable loads in a receding horizon algorithm. The paper concludes with numerical results in Sect. 4.

2 The Residential Energy System

Let $\mathscr{I} \in \mathbb{N}$ be the number of RESs connected in the local area under consideration. A simple model of the RES of user $i \in \{1, \ldots, \mathscr{I}\}$ is:

$$\begin{aligned} x_i(k+1) &= x_i(k) + Tu_{i_1}(k), \\ z_i(k) &= w_i(k) + u_{i_1}(k) + u_{i_2}(k) \end{aligned} \tag{1}$$

where x_i is the state of charge of the battery in kWh, u_{i_1} is the battery charge/discharge rate in kW, u_{i_2} is the controllable load in kW, w_i is the static load minus the local generation in kW, and z_i is the power supplied by/to the grid in kW. Here, T represents the length of the sampling interval in hours; e.g., $T = 0.5$ corresponds to 30 min. The RES network is then defined by the following discrete-time system

$$x(k+1) = f(x(k), u(k)), \tag{2}$$

$$z(k) = h(u(k), w(k)) \tag{3}$$

where $x, w \in \mathbb{R}^{\mathscr{I}}$, $u \in \mathbb{R}^{2\mathscr{I}}$, and the definitions of f and h are obvious from (1). We assume constraints on the battery capacity and charge/discharge rates are given by $C_i, \overline{u}_i \in \mathbb{R}_{>0}$ and $\underline{u}_i \in \mathbb{R}_{<0}$ so that for each RESs i, $i \in \{1, \ldots, \mathscr{I}\}$:

$$0 \le x_i(k) \le C_i \quad \text{and} \quad \underline{u}_{i_1} \le u_{i_1}(k) \le \overline{u}_{i_1} \quad \forall k \in \mathbb{N}_0. \tag{4}$$

Note that this model adequately captures elements of fuel cells as energy storage devices since the conversion of electricity to hydrogen, and vice versa, is rate-limited and fuel cells have a fixed storage capacity.

We assume that the load can be split into two parts: controllable and static load. The static load is included in w. The controllable loads $\{w_c(k)\}_{k\in\mathbb{N}} \subset \mathbb{R}^{\mathscr{I}}$ must be scheduled during a certain time window. More precisely $w_{c_i}(k+\overline{N}-1)$ can be scheduled during the time interval from $\max\{0,k\}$ to $k+\overline{N}-1$ for a given $\overline{N}\in\mathbb{N}$. This leads to the time-dependent constraints

$$\sum_{j=0}^{k} w_{c_i}(j) - \sum_{j=0}^{k-1} u_{i_2}(j) \le u_{i_2}(k) \le \sum_{j=0}^{k+\overline{N}-1} w_{c_i}(j) - \sum_{j=0}^{k-1} u_{i_2}(j) \quad \forall k \in \mathbb{N}_0 \tag{5}$$

for each RES $i \in \{1,\dots,\mathscr{I}\}$. Observe that at time k, $u_{i_2}(j)$ is fixed for all $j<k$, rather than a control variable, since it is a control action that was already applied. We introduce upper and lower bounds on u_{i_2} reflecting the fact that only a certain amount of the controllable load can be scheduled in one time step, i.e., for each RES i, $i\in\{1,\dots,\mathscr{I}\}$, and given $\underline{w}_{c_i}$, $\overline{w}_{c_i}\in\mathbb{R}$,

$$\underline{w}_{c_i} \le u_{i_2}(k) \le \overline{w}_{c_i} \quad \forall k \in \mathbb{N}_0. \tag{6}$$

We assume that the $\underline{w}_{c_i}$, $\overline{w}_{c_i}$ are chosen such that conditions (5) and (6) can be simultaneously satisfied.

Our goal is to flatten the performance output z. We introduce two relevant performance metrics. The average power demand at time k defined as $\Pi(k) := \frac{1}{\mathscr{I}}\sum_{i=1}^{\mathscr{I}} z_i(k)$ and let $\mathscr{N}$ denote the simulation length in number of samples. The performance metric of peak-to-peak (PTP) variation of the average demand of all RESs is given by

$$\left(\max_{k\in\{0,\dots,\mathscr{N}\}} \Pi(k)\right) - \left(\min_{k\in\{0,\dots,\mathscr{N}\}} \Pi(k)\right). \tag{PTP}$$

The second performance metric is the root-mean-square (RMS) deviation from the average; i.e., the average $\Upsilon := \frac{1}{\mathscr{N}\mathscr{I}}\sum_{k=0}^{\mathscr{N}-1}\sum_{i=1}^{\mathscr{I}}(w_i(k)+w_{c_i}(k))$ is calculated and the respective deviations are quadratically penalized:

$$\sqrt{\frac{1}{\mathscr{N}}\sum_{k=0}^{\mathscr{N}-1}(\Pi(k)-\Upsilon)^2}. \tag{RMS}$$

3 Model Predictive Control Approaches

We present two Model Predictive Control (MPC) algorithms for the control of a network of RESs. The first approach is a Centralized MPC algorithm. This scheme requires full communication of all relevant variables for the entire network as well as a known model of the network. The second approach is a Decentralized MPC

approach where each RES implements its own local MPC controller. This requires no communication or cooperation between RESs. Both schemes use a receding horizon controller.

MPC iteratively minimizes an optimization criterion with respect to predicted trajectories and implements the first part of the resulting optimal control sequence until the next optimization is performed (see, e.g., [6] or [2]). We propose such a predictive controller for (1). In order to do this, we assume that we have predictions of the residential load and generation some time into the future that is coincident with the horizon of the predictive controller. In other words, given a prediction horizon $N \in \mathbb{N}$, we assume knowledge of $w_i(j)$, $w_{c_i}(j)$ for all $j \in \{k, \dots, k+N-1\}$, where $k \in \mathbb{N}_0$ is the current time.

Before defining the cost function for the MPC approaches, we rewrite the constraints (5) in a receding horizon fashion. The constraints on $u_{i_2}(j)$ in the prediction horizon are captured by

$$\lambda_i^q(k) := \sum_{j=0}^{k+q} w_{c_i}(j) - \sum_{j=0}^{k-1} u_{i_2}(j) \le \sum_{j=k}^{k+q} u_{i_2}(j) \tag{7a}$$

$$\Lambda_i^q(k) := \sum_{j=0}^{k+\min\{q+\overline{N},N\}-1} w_{c_i}(j) - \sum_{j=0}^{k-1} u_{i_2}(j) \ge \sum_{j=k}^{k+q} u_{i_2}(j) \tag{7b}$$

for $q \in \{0, \dots, N-1\}$ and $i \in \{1, \dots, \mathscr{I}\}$. The term $\min\{q+\overline{N}, N\}$ reflects that we predict only N steps ahead and therefore only controllable load with a deadline during the prediction horizon is considered. Observe that the bounds can be easily updated by $\lambda_i^q(k+1) = \lambda_i^q(k) + w_{c_i}(k+q+1) - u_{i_2}(k)$ and $\Lambda_i^q(k+1) = \Lambda_i^q(k) + w_{c_i}(k+\min\{q+\overline{N}, N\}) - u_{i_2}(k)$.

3.1 Centralized Model Predictive Control

Define the predicted average power usage for the ith RES as

$$\zeta_i(k) := \frac{1}{N}\left(\lambda_i^0(k) - w_{c_i}(k) + \sum_{j=k}^{k+N-1} (w_i(j) + w_{c_i}(j))\right). \tag{8}$$

To implement the Centralized MPC algorithm, we compute the overall average on the prediction horizon by $\bar{\zeta}(k) := \frac{1}{\mathscr{I}} \sum_{i=1}^{\mathscr{I}} \zeta_i(k)$ and then minimize the joint cost function

$$\min_{\hat{u}(\cdot)} \sum_{j=k}^{k+N-1} \Big(\bar{\zeta}(k) - \frac{1}{\mathscr{I}} \sum_{i=1}^{\mathscr{I}} \underbrace{(w_i(j) + \hat{u}_{i_1}(j) + \hat{u}_{i_2}(j))}_{\hat{z}_i(j)}\Big)^2 \tag{9}$$

with respect to the predicted control input $\hat{u}(k), \hat{u}(k+1), \ldots, \hat{u}(k+N-1)$ with $\hat{u}(\cdot) = (\hat{u}_1(\cdot), \hat{u}_2(\cdot), \ldots, \hat{u}_{\mathscr{I}}(\cdot))^T$ subject to the system dynamics (1), the current state $x(k) = (x_1(k), \ldots, x_{\mathscr{I}}(k))^T$, and the constraints (4), (6), and (7) for all $i \in \{1, \ldots, \mathscr{I}\}$.

Here, and in what follows, we denote predicted controls and outputs in the MPC algorithm by hats; i.e., for the ith RES at time j the predicted control is $\hat{u}_i(j)$ and the predicted performance output is $\hat{z}_i(j)$.

3.2 Decentralized Model Predictive Control

The Centralized MPC approach presented above requires a significant amount of communication overhead. A further drawback of the Centralized MPC approach is that the central entity requires full knowledge of the network model, in particular (4), (6), and (7) for each $i \in \{1, \ldots, \mathscr{I}\}$. Therefore, any change in the network such as the addition of new generation or storage resources requires an update of the central model. As a remedy we propose a decentralized control approach that alleviates the communication and computation difficulties.

A straightforward option in order to flatten the energy profile of the ith RES is to penalize deviations from its (anticipated) average usage defined in (8). With a quadratic cost function, this leads to the finite-horizon optimal control problem

$$\min_{\hat{u}_i(\cdot)} \sum_{j=k}^{k+N-1} (\zeta_i(k) - \underbrace{(w_i(j) + \hat{u}_{i_1}(j) + \hat{u}_{i_2}(j))}_{\hat{z}_i(j)})^2$$

subject to the system dynamics (1), the current state of charge $x_i(k)$ of the energy storage, the constraints (4), (6), and (7) corresponding to the controllable loads. With each RES solving its own optimization problem with no reference to the rest of the network, the aforementioned communication and computation difficulties of Centralized MPC are not present in the Decentralized MPC algorithm.

4 Numerical Results

In this section, we compare the discussed controllers and the impact of controllable load in the model by considering the load and generation profiles for a group of 20 customers drawn from the Australian electricity distribution company Ausgrid. The data from these customers was collected as part of the *Smart Grid, Smart City* project. We use 2 weeks starting on 1 March 2011. As already mentioned in the introduction, motivated by Barker et al. [1], we split the given load profile into 60 % static load and 40 % controllable load. Figure 1 visualizes the impact of the energy storage and the controllable loads on the uncontrolled grid profile. Table 1

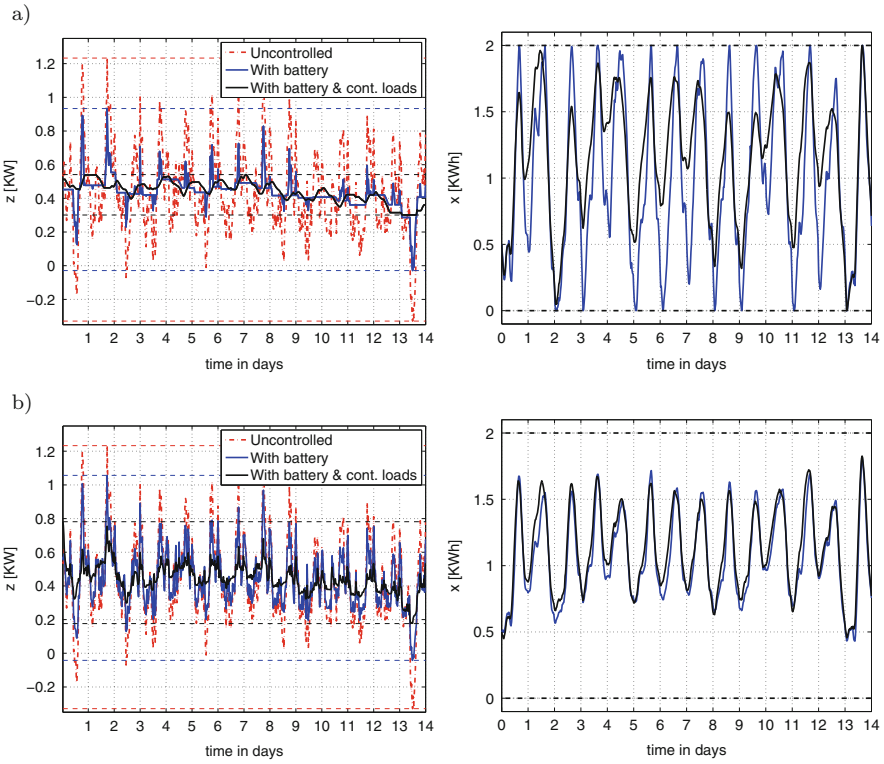


Fig. 1 Australian data for 20 systems. Flattened aggregated grid profile using an energy storage and controllable loads with $\overline{N} = 12$. (**a**) Centralized MPC: aggregated grid (*left*) and battery (*right*) profiles. (**b**) Decentralized MPC: aggregated grid (*left*) and battery (*right*) profiles

Table 1 Australian data: Peak-to-peak variation and RMS deviation from the average for 20 RESs and a simulation length of 2 weeks. Results without Controllable Load (C.L.) and for differing controllable load horizons $\overline{N}$

	Without C. L.		$\overline{N} = 12$		$\overline{N} = 24$		$\overline{N} = 36$	
	PTP	RMS	PTP	RMS	PTP	RMS	PTP	RMS
No battery storage	1.5621	0.2506						
Decentralized MPC	1.0986	0.1663	0.6050	0.0845	0.5555	0.0777	0.5555	0.0777
Centralized MPC	0.9621	0.0952	0.2401	0.0609	0.2915	0.0594	0.2915	0.0594

summarizes the results with respect to the introduced metrics. All simulations use the prediction horizon $N = 48$, $T = 0.5$ and initial battery state $x_i = 0.5$ for all $i \in \{1, \ldots, \mathscr{I}\}$. For simplicity we assume that all systems have the same box constraints which for the simulations means $-\underline{u}_i = \overline{u}_i = 0.3$, $C_i = 2$, $\underline{w}_{c_i} = 0$ and $\overline{w}_{c_i} = 1.25$ for all $i \in \{1, \ldots, \mathscr{I}\}$.

The addition of controllable loads yields the expected improvement in the defined performance metrics. Additionally Centralized MPC outperforms Decentralized MPC due to the lack of global coordination in the decentralized setting. $\overline{N}$ seems to play a minor role (assuming that $\overline{N}$ is big enough to call the loads controllable). In the centralized setting we obtain the smallest PTP variation for $\overline{N} = 12$, an observation that requires further investigation.

5 Conclusion

In this paper we have extended our earlier Residential Energy System (RES) model introduced in [8] by adding controllable loads. Numerically we have shown that the additional degree of freedom leads to the expected improvements with respect to the grid profile.

References

1. Barker, S., Mishra, A., Irwin, D., Shenoy, P., Albrecht, J.: SmartCap: flattening peak electricity demand in smart homes. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communications. Lugano, Switzerland (2012)
2. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control. Theory and Algorithms. Springer, London (2011)
3. Mohsenian-Rad, A.H., Leon-Garcia, A.: Optimal residential load control with price prediction in real-time electricity pricing environments. IEEE Trans. Smart Grid **1**(2), 120–133 (2010)
4. Nykamp, S., Molderink, A., Hurink, J.L., Smit, G.J.M.: Storage operation for peak shaving of distributed PV and wind generation. In: Proceedings of the IEEE PES Innovative Smart Grid Technologies (2013)
5. Nykamp, S., Bosman, M.G.C., Molderink, A., Hurink, J.L., Smit, G.J.M.: Value of storage in distribution grids—competition or cooperation of stakeholders? IEEE Trans. Smart Grid **4**(3), 1361–1370 (2013)
6. Rawlings, J.B., Mayne, D.Q.: Model Predictive Control: Theory and Design. Nob Hill Publishing, Madison (2009)
7. Samadi, P., Mohsenian-Rad, H., Wong, V.W.S., Schober, R.: Tackling the load uncertainty challenges for energy consumption scheduling in smart grid. IEEE Trans. Smart Grid **4**(2), 1007–1016 (2013)
8. Worthmann, K., Kellett, C.M., Braun, P., Grüne, L., Weller, S.R.: Distributed and decentralized control of residential energy systems incorporating battery storage. IEEE Trans. Smart Grid (2015). doi:10.1109/TSG.2015.2392081

Particle Swarm Optimization Applied to Hexarotor Flight Dynamics

Valeria Artale, Cristina L.R. Milazzo, Calogero Orlando, and Angela Ricciardello

Abstract In this work, results obtained by the flight control simulations of a prototype of hexarotor Unmanned Aerial Vehicle (UAV) are shown. The mathematical model and control of the hexacopter airframe are presented. To stabilize the entire system, Linear Quadratic Regulator (LQR) control is used in such a way to set both Proportional Derivative (PD) and Proportional Integral Derivative (PID) controls. Particle Swarm Optimization has been used to set the optimal coefficient matrices of the LQR control algorithm. The simulations are performed to show how LQR tuned PD and PID controllers lead to zero the error of the position along gravity acceleration direction, stop the rotation of UAV around body axes and stabilize the hexarotor. Moreover, the obtained LQR-PD and LQR-PID controllers have been tested by comparing the response to impulse disturbances of the nonlinear dynamical system with the response of the linearized one.

Keywords Flight control • Linear quadratic regulator control • Particle swarm optimization

1 Introduction

In this work a prototype of Unmanned Aerial Vehicle (UAV) is considered in agreement with the design requirements of a Project supported by the PO. FESR 2007/2013 whose objective is the realization of a multirotor system for the environmental survey. The peculiarity of UAV consists in its versatility and therefore in its use in missions with different aims, from surveying to the rescue, from inspection in inaccessible areas to emergency action [1]. Nevertheless of the configuration, the UAV should be equipped with a robust control system in order to autonomously manage its flight and to achieve and steadily maintain a desired position [2, 3]. With this aim two control algorithms are investigated in this work, namely the Proportional-Derivative (PD) and the Proportional-Integral-Derivative

V. Artale • C.L.R. Milazzo (✉) • C. Orlando • A. Ricciardello
Faculty of Engineering and Architecture, Cittadella Universitaria, Enna, Italy
e-mail: valeria.artale@unikore.it; cristina.milazzo@unikore.it; calogero.orlando@unikore.it; angela.ricciardello@unikore.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_86

(PID) controls. Both control schemes are written as Linear Quadratic Regulator (LQR) problems by properly defining the state space vector of the linearized model. Particle Swarm Optimization (PSO) is used to compute the optimal coefficients of the LQR problem for the chosen objective function. Results are presented for the linear and the non-linear models stabilized by means of the both the LQR-PD and the LQR-PID controllers.

2 Hexarotor Dynamics

The hexacopter considered is assumed as a rigid body and it is equipped with six fixed pitch propellers; the structure of the UAV is symmetrical with respect to the body reference frame. In this formulation gyroscopic moments as well as aerodynamic drag forces are not taken into account. The equations of the motion of the hexarotor are deduced from Newton-Euler equations [4] which allow to decompose the motion into translational and rotational components. Let $(0,\ X_E,\ Y_E,\ Z_E)$ be the reference frame fixed to the inertial space and $(G,\ X_B,\ Y_B,\ Z_B)$ be the body frame fixed to the hexacopter and centered in the body center of mass G. Both Z_E and Z_B axes point downward while $\phi,\ \theta$ and ψ are the Euler angles that determine the orientation of the body frame with respect to the inertial one. As the translational kinematic concerns, the total force acting on the aircraft takes into account the gravitational action mg along the Z_E direction, where m is the constant mass and g the gravitational acceleration, and the total thrust T along the Z_B axis; on the other hand, the external torques acting on the drone are $\tau_\phi,\ \tau_\theta$ and τ_ψ and are generated by the six propellers. The non-linear model that describes the dynamics of the hexacopter is

$$\begin{bmatrix} \ddot{X}_E \\ \ddot{Y}_E \\ \ddot{Z}_E \end{bmatrix} = \begin{bmatrix} T\left(\cos(\psi)\sin(\theta)\cos(\phi) + \sin(\psi)\sin(\phi)\right)/m \\ T\left(\sin(\psi)\sin(\theta)\cos(\theta) - \cos(\psi)\sin(\theta)\right)/m \\ g - T\cos(\theta)\cos(\phi)/m \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} \dot{p} \\ \dot{q} \\ \dot{r} \end{bmatrix} = \begin{bmatrix} (I_{yy} - I_{zz})\, q\, r/I_{xx} \\ (I_{zz} - I_{xx})\, p\, r/I_{yy} \\ (I_{xx} - I_{yy})\, p\, q/I_{zz} \end{bmatrix} + \begin{bmatrix} \tau_\phi/I_{xx} \\ \tau_\theta/I_{yy} \\ \tau_\psi/I_{zz} \end{bmatrix} \tag{2}$$

in which $I_{xx},\ I_{yy},\ I_{zz}$ are the inertial moments with respect to $X_B,\ Y_B$ and Z_B. In Eq. (2) $p,\ q$ and r are the angular velocity components of the hexarotor about the body axes and are related [5] to the Euler angle velocities as

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\sin\theta \\ 0 & \cos\phi & \cos\theta\cos\phi \\ 0 & -\sin\phi & \cos\theta\cos\phi \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix}$$

3 Hexarotor Control

In this section the LQR-PD and LQR-PID controller are presented, then PSO algorithm is briefly introduced to compute the LQR coefficients on the basis of the chosen objective function. The LQR is a feedback control technique that can be applied to linear system of the form

$$\dot{\mathbf{X}} = \mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{U} \tag{3}$$

where $\mathbf{X}$ and $\mathbf{U}$ are the state and input vectors, respectively, while $\mathbf{A}$ and $\mathbf{B}$ are the dynamic and input matrices. The LQR algorithm requires that the optimal control input $\mathbf{U} = -\mathbf{K}\mathbf{X}$ should be computed in such a way to minimize a chosen cost function J defined as

$$J = \lim_{t \to +\infty} \int_0^t \left(\mathbf{X}^{\mathbf{T}}\, \mathbf{Q}\, \mathbf{X} + \mathbf{U}^{\mathbf{T}}\, \mathbf{R}\, \mathbf{U} \right) \tag{4}$$

in which $\mathbf{Q}$ and $\mathbf{R}$ are positive definite matrices, set by the user to amplify or reduce the influence of the entries of the $\mathbf{X}$ and $\mathbf{U}$ vectors. The gain matrix $\mathbf{K}$ is obtained by solving the associated Lurie-Riccati equation [6, 7] which depends on $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Q}$ and $\mathbf{R}$. On the other hand, both of PD and PID control techniques are based on the minimization of the error value $e(t) = x_d(t) - x(t)$, relative to a specific variable to be controlled $x(t)$ with respect to a desired target $x_d(t)$ [8, 9]. The PD control assumes that control signal $u(t)$ is the sum of a term proportional to $e(t)$, that accounts for the current error, plus a term proportional to the error time derivative $\dot{e}(t)$, which accounts for a prediction of future errors. In the PID case a term proportional to the integral of the error is also added to the control input and accounts for the accumulation of past errors. Both of PD and PID controller can be written as a state feedback LQR problem. This is obtained by suitably defining the state space problem Eq. (3) representing Eq. (1) linearized around the equilibrium point. In this work the equilibrium point is the hovering configuration and represents the desired state. More particularly, let $\mathbf{x} = [z\ \phi\ \theta\ \psi]$ be the vector that collects the heave, roll, pitch and yaw deviation from the hovering configuration to be zeroed, in the LQR-PD control problem the state vector is defined as $\mathbf{X} = [\mathbf{x}\ \ \dot{\mathbf{x}}]^T$ while in the PID-LQR case it writes as $\mathbf{X} = \left[\int \mathbf{x}dt\ \mathbf{x}\ \ \dot{\mathbf{x}}\right]^T$. The input vector is defined as $\mathbf{U} = \left[T\ \tau_\phi\ \tau_\theta\ \tau_\psi\right]^T$ independently of LQR-PD or LQR-PID control. The dynamic and input matrices $\mathbf{A}$ and $\mathbf{B}$ are defined accordingly to the state vector $\mathbf{X}$ but are not reported here for the sake of conciseness. It follows that the optimal control vector for the LQR-PD and for the LQR-PID specify as

$$\mathbf{U}_{\mathrm{LQR-PD}} = -\mathbf{K}\mathbf{X} = -\left[\,\mathbf{K}_P\ \mathbf{K}_D\,\right]\mathbf{X} = -\mathbf{K}_P\mathbf{x}\,(t) - \mathbf{K}_D\dot{\mathbf{x}}\,(t)$$

$$\mathbf{U}_{\mathrm{LQR-PID}} = -\mathbf{K}\mathbf{X} = -\left[\,\mathbf{K}_I\ \mathbf{K}_P\ \mathbf{K}_D\,\right]\mathbf{X} = -\mathbf{K}_P\mathbf{x}\,(t) - \mathbf{K}_I\int_0^t \mathbf{x}\,(\tilde{t})\,d\tilde{t} - \mathbf{K}_D\dot{\mathbf{x}}\,(t)$$

where $\mathbf{K_P}$, $\mathbf{K_I}$ and $\mathbf{K_D}$ are 4×4 diagonal matrices whose entries represent the proportional, integral and derivative gains associated to the components of the vector $\mathbf{x}$, i.e. to the heave, roll, pitch and yaw deviation from hovering configuration. The last point to be addressed is the computation of the optimal $\mathbf{Q}$ and $\mathbf{R}$ coefficients of the LQR control problem. This is achieved by using the Particle Swarm Optimization technique, an evolutionary computational scheme based on simplified social model [10]. Each individual (Particle) has a number of qualities, called position variables, and represents an *n-dimensional* vector in problem space, i.e. a candidate problem solution. The position of a set of particles (Swarm) is initialized randomly at first. Then the performance of each particle is evaluated on the basis of the objective function. At successive steps particles tend to emulate the success of neighboring individuals, i.e. particles accelerate toward the objective. More in details, in this work both of the matrices, namely $\mathbf{Q}$ and $\mathbf{R}$, are assumed to be diagonal and their entries are the position variables of particles. This implies that the number of particles' qualities is $n = 12$ for the LQR-PD problem; otherwise, the number of variables of each particle is $n = 16$ for the LQR-PID case. In particular, being h_i, $i = 1 \dots n$ the particle's variables, the $\mathbf{Q}$ and $\mathbf{R}$ matrices write as

$$\mathbf{Q} = diag\left\{h_1 \cdots h_{n-4}\right\}; \quad \mathbf{R} = diag\left\{h_{n-3} \cdots h_n\right\}. \tag{5}$$

Each position component of a particle is initialized randomly at the iteration step $\lambda = 1$ and is then updated as

$$h_i(\lambda + 1) = h_i(\lambda) + dh_i(\lambda + 1) \tag{6}$$

where dh_i represents the velocity of the i-th particle's variable on a unitary step increment and is computed as [10]

$$dh_i(\lambda + 1) = \mu dh_i(\lambda) + c_1 r_1(h_{ib} - h_i(\lambda)) + c_2 r_2(h_b - h_i(\lambda)), \tag{7}$$

in which μ is a coefficient called inertia weight, c_1 and c_2 are the acceleration coefficients, called cognitive and social constant, respectively, $r_1, \ r_2 \in [0, 1]$ randomly. In Eq. (7), h_{pb} represents the previous personal best position of a particle while h_{gb} is the global best position of the entire swarm. In details, the inertia weight varies as

$$\mu = \mu_{max} - \frac{\mu_{max} - \mu_{min}}{\lambda_{Max}} \lambda \tag{8}$$

with μ_{min} and μ_{max} the minimum and maximum value that μ might assume and λ_{Max} the maximum number of iterations.

4 Numerical Results

In this section results obtained by using PSO-LQR-PD and PSO-LQR-PID methods are discussed and, in particular, they have been tested and compared in both linear and non linear model. As presented in previous section, the goal of this paper is to control z, ϕ, θ, ψ variables. At this aim, in the implementation of the PSO algorithm, the time settling of controlled variables, denoted by ts_z, ts_ϕ, ts_θ, ts_ψ, are minimized. With other words, the objective function has been chosen as

$$f_{OBJ} = ts_z + ts_\phi + ts_\theta + ts_\psi . \tag{9}$$

Thus, the control of the overshoot of variables as well as the control of input actions are neglected in this formulation. Nevertheless, in the non linear model the total thrust and the roll, pitch and yaw moments are adjusted by means of saturation function. This guaranties that values of acting forces could be actually achieved by the propulsive system. The parameters for the PSO problem are: number of swarm set to 1; number of particles set to 50; cognitive constant $c_1 = 2.05$; social constant $c_2 = 4.05$; minimum of inertia weight $\mu_{min} = 0.05$; maximum of the inertia weight $\mu_{max} = 0.995$; maximum number of iterations $\lambda_{Max} = 50$. As discussed in previous section, the PSO algorithm has been used in order to estimate the **Q** and **R** matrices of the LQR technique and then the parameters of PD and PID regulators. The values obtained by simulations are reported in Table 1. In order to validate the presented method, a test with the presence of a disturbance moving the drone far from its hovering position has been carried out with the scope of evaluating the time settling of each controlled variables. The disturbance is an impulse function characterized by a unitary amplitude and a pulse width 0.2 s that influences the hexarotor altitude and attitude rates, i.e. $\dot{z}$, $\dot{\phi}$, $\dot{\theta}$, $\dot{\psi}$. The trend of heave, roll, pitch and yaw motion versus time is depicted in Fig. 1, as linearised model concerns, and in Fig. 2 the non linear case. In each figure the LQR-PD regulator (continuous) and LQR-PID regulator (dashed) can be distinguished. Linear model time histories, see Fig. 1, show that overshoot of variables controlled via the LQR-PID is less than that obtained using the LQR-PD scheme. This behavior is confirmed by the non linear results shown in Fig. 2. Moreover, it is worth noting that linear and non linear results are mostly

Table 1 Parameters of LQR-PD and LQR-PID control obtained by means of PSO algorithm

–	i	$K_{i,P}$	$K_{i,I}$	$K_{i,D}$
LQR-PD	z	−70711	–	−1518.1
	ϕ	70711	–	365.73
	θ	70711	–	365.73
	ψ	70711	–	491.11

–	i	$K_{i,P}$	$K_{i,I}$	$K_{i,D}$
LQR-PID	z	-10^5	−1	−1805.3
	ϕ	10^5	1	434.93
	θ	10^5	1	434.93
	ψ	10^5	1	584.03

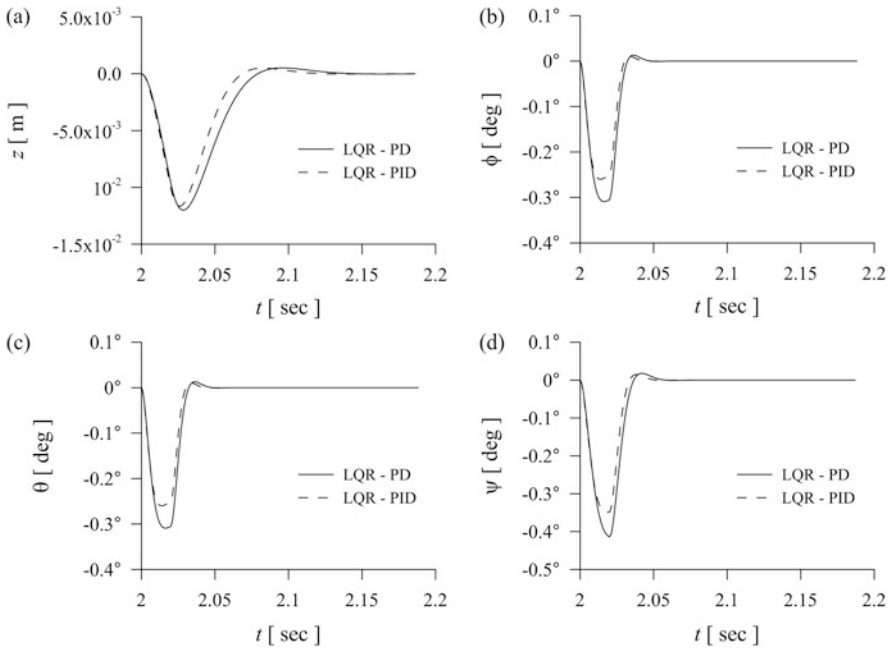


Fig. 1 Linear model time history: (**a**) heave motion; (**b**) roll motion; (**c**) pitch motion; (**d**) yaw motion

comparable for the attitude time history, however it stems from Fig. 2a that the non linear heave transient behaviour is one order of magnitude lower than the linear one, see Fig. 2a. This appears to be a consequence of the fact that in the linear model the total thrust increases unbounded while the saturation function in the non linear model bounds the total thrust in $[-180,\ 0]$ N and the roll, pitch, yaw moment in $[-60, 60]$ N/m. Furthermore, for sake of clarity, the time settling of depicted quantities are summarized in Table 2. This latter shows that in the linear case the LQR-PID method minimizes the objective function faster than the LQR-PD, both in linear and non linear model. Here again, it can be noticed that the time settling associated with altitude decreases of one order of magnitude from linear model to non linear one, whether or not LQR-PD or LQR-PID are taken into account. It also appears to be a consequence of the reduction of the z overshoot due to the bounds of total thrust T.

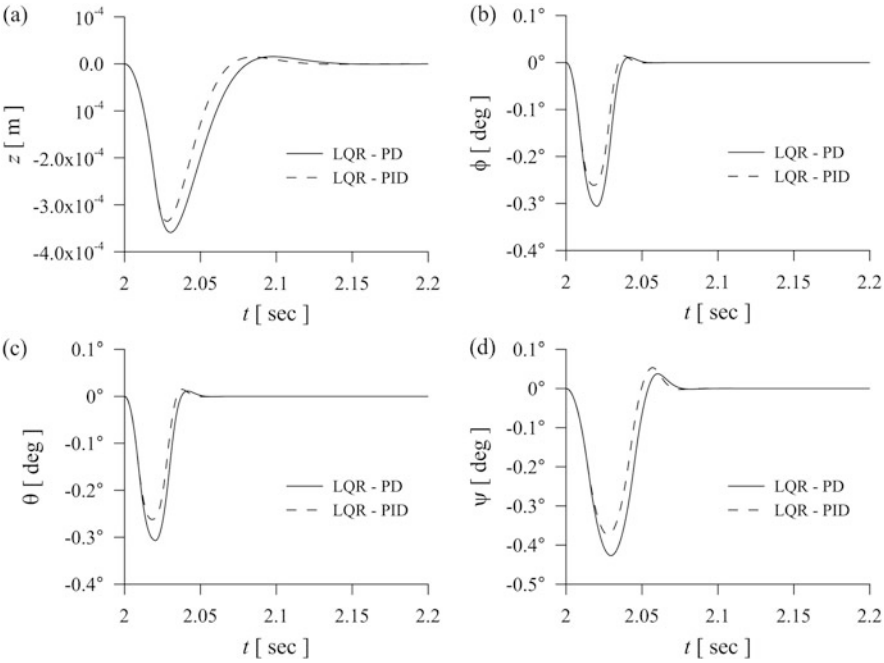


Fig. 2 Non-Linear model time history: (**a**) heave motion; (**b**) roll motion; (**c**) pitch motion; (**d**) yaw motion

Table 2 Time settling of controlled variables

		Linear model	Non linear model
PD	z	0.1216	0.049
	ϕ	0.0376	0.0358
	θ	0.0376	0.0358
	ψ	0.0466	0.0695

		Linear model	Non linear model
PID	z	0.1041	0.0434
	ϕ	0.029	0.04
	θ	0.029	0.04
	ψ	0.0418	0.066

5 Conclusions

In this paper the PSO algorithm has been introduced in order to estimate the parameters of LQR-PD and LQR-PID regulators with the aim of stabilizing the altitude and attitude of a hexacopter around its hovering (equilibrium) configuration, by minimizing the time settling of controlled variables. The different approaches applied on both linearised and non linear dynamical model are able to control the hexacopter dynamics also in presence of the considered disturbances.

Acknowledgements This work is supported by the PO. FESR 2007/2013 subprogram 4.1.1.1, Prog. "Mezzo Aereo a controllo remoto per il Rilevamento del TErritorio—MARTE" Grant No. 10772131.

References

1. Rango, A., Laliberte, A., Herrick, J., Winters, C., Havstad, K., Steele, C., Browning, D.: Unmanned aerial vehicle-based remote sensing for rangeland assessment, monitoring and management. J. Appl. Remote Sens. **3**(1), 033542 (2009)
2. Mian, A., Daoboo, W.: Modeling and backstepping-based nonlinear control strategy for a 6 DOF quadrotor helicopter. Chin. J. Aeronaut. **21**, 261–268 (2008)
3. Zhang, R., Quan, Q., Cai, K.: Attitude control of quadrotor aircraft subject to a class of time-varying disturbances. IET Control Theory Appl. **5**, 1140–1146 (2011)
4. Artale, V., Milazzo, C., Ricciardello, A.: Mathematical modeling of hexacopter. Appl. Math. Sci. **7**(97–100), 4805–4811 (2013)
5. Artale, V., Milazzo, C., Ricciardello, A.: A quaternion based simulation of multirotor dynamics. Int. J. Model. Simul. Sci. Comput. **6**, 1550009 (2015). doi:10.1142/S1793962315500099
6. Alaimo, A., Artale, V., Milazzo, C., Ricciardello, A.: PID controller applied to hexacopter flight. J. Intell. Robot. Syst. **73**(1–4), 261–270 (2014)
7. Artale, V., Barbaraci, G., Milazzo, C., Orlando, C., Ricciardello, A.: Dynamic analysis of a hexacopter controlled via LQR-PI. AIP Conf. Proc. **1558**, 1212–1215 (2013)
8. Artale, V., Milazzo, C., Ricciardello, A.: An example of quaternion parameterization for dynamical simulations. J. Phys. Conf. Ser. **490**(1), 012005 (2014)
9. Alaimo, A., Artale, V., Milazzo, C., Ricciardello, A.: Comparison between Euler and quaternion parametrization in UAV dynamics. AIP Conf. Proc. **1558**, 1228–1231 (2013)
10. Umarani, R., Selvi, V.: Particle swarm optimization evolution, overview and applications. Int. J. Eng. Sci. Technol. **2**(7), 2802–2806 (2010)

Multiobjective Optimal Control Methods for the Development of an Intelligent Cruise Control

Michael Dellnitz, Julian Eckstein, Kathrin Flaßkamp, Patrick Friedel, Christian Horenkamp, Ulrich Köhler, Sina Ober-Blöbaum, Sebastian Peitz, and Sebastian Tiemeyer

Abstract During the last years, alternative drive technologies, for example electrically powered vehicles (EV), have gained more and more attention, mainly caused by an increasing awareness of the impact of CO2 emissions on climate change and by the limitation of fossil fuels. However, these technologies currently come with new challenges due to limited lithium ion battery storage density and high battery costs which lead to a considerably reduced range in comparison to conventional internal combustion engine powered vehicles. For this reason, it is desirable to increase the vehicle range without enlarging the battery. When the route and the road slope are known in advance, it is possible to vary the vehicles velocity within certain limits in order to reduce the overall drivetrain energy consumption. This may either result in an increased range or, alternatively, in larger energy reserves for comfort functions such as air conditioning.

In this presentation, we formulate the challenge of range extension as a multiobjective optimal control problem. We then apply different numerical methods to calculate the so-called Pareto set of optimal compromises for the drivetrain power profile with respect to the two concurrent objectives battery state of charge and mean velocity. In order to numerically solve the optimal control problem by means of a direct method, a time discretization of the drivetrain power profile is necessary. In combination with a vehicle dynamics simulation model, the optimal control problem is transformed into a high dimensional nonlinear optimization problem. For the approximation of the Pareto set, two different optimization algorithms implemented in the software package GAIO are used. The first one yields a global optimal solution by applying a set-oriented subdivision technique to parameter space. By construction, this technique is limited to coarse discretizations of the drivetrain power profile. In contrast, the second technique, which is based on an image space

M. Dellnitz • K. Flaßkamp • C. Horenkamp • S. Ober-Blöbaum • S. Peitz (✉)
University of Paderborn and Institute for Industrial Mathematics, Warburger Str. 100, 33098 Paderborn, Germany
e-mail: speitz@math.upb.de

J. Eckstein • P. Friedel • U. Köhler • S. Tiemeyer
HELLA KGaA Hueck and Co., Beckumer Str. 130, 59552 Lippstadt, Germany

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_87

continuation method, is more suitable when the number of parameters is large while the number of objectives is less than five. We compare the solutions of the two algorithms and study the influence of different discretizations on the quality of the solutions.
A MATLAB/Simulink model is used to describe the dynamics of an EV. It is based on a drivetrain efficiency map and considers vehicle properties such as rolling friction and air drag, as well as environmental conditions like slope and ambient temperature. The vehicle model takes into account the traction battery too, enabling an exact prediction of the batterys response to power requests of drivetrain and auxiliary loads, including state of charge.

Keywords Cruise control • Multiobjective optimal control • Pareto set

1 Introduction

Electrically powered vehicles (EV) have gained more and more attention during the last years due to an increasing awareness of the impact of CO_2 emissions on climate change and the limitation of fossil fuels. New research challenges arise due to limited battery storage densities, high battery costs and a considerably reduced range in comparison to conventionally powered vehicles. Therefore, range increasing driving strategies play an important role in electromobility (cf. e.g. [8]).

Different control and optimization strategies have been suggested for vehicle applications in the past, see [16] for an overview [7, 9] for model predictive control for trucks [5] for an application of dynamic programming [13] for indirect or [4] for direct optimal control methods.

In this paper, an "intelligent cruise control" is developed by taking into account topographic data of a given travel route. We formulate the challenge of range extension as a multiobjective optimal control problem, transforming it into a multiobjective optimization problem by using a direct approach. We then use numerical methods to compute the so-called Pareto set of optimal compromises between the concurrent objectives "maximize battery charge" and "maximize driven distance". Pareto optimal accelerator pedal position profiles of the EV are computed by using two different multiobjective optimization methods.

The paper is organized as follows: In Sect. 2, the mathematical problem formulation and solution methods for multiobjective optimal control problems are given. The EV model and computational results are presented in Sect. 3 followed by a conclusion in Sect. 4.

2 Multiobjective Optimal Control

Searching for a control strategy of an EV which maximizes the driving distance is an example of an *optimal control problem*. In general, the technical system is represented by a model, typically of the form $\dot{x}(t) = f(x(t), u(t))$. Further, the objectives subject to optimization have to be modeled. By convention, we always consider minimization problems. Typically, objective functionals are of the form

$$J(x,u) = \int_0^T C(x,u)dt + \Psi(x(T)) \tag{1}$$

with running costs $C(\cdot,\cdot)$ depending on the system's states x and controls u and a final cost $\Psi(\cdot)$ depending on the final state $x(T)$. Finally, there might be different kinds of constraints, e.g. boundary conditions $g(x(0), x(T)) = 0$ and box constraints $b_l \leq \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} \leq b_u$ for all $t \in [0,T]$.

In many applications, there arise several objective functionals that have to be minimized simultaneously. This leads to vector-valued objective functionals, denoted by $J(x,u)$ with $J = (J_1, \ldots, J_k)$, $k \geq 1$ and, for all $i \in \{1, \ldots, k\}$, J_i as in (1). Altogether, we obtain a *multiobjective optimal control problem* (MOCP)

$$\min_u J(x,u) \quad \text{w.r.t. } \dot{x} = f(x,u),\ g(x(0),x(T)) = 0,\ b_l \leq \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} \leq b_u \ \forall t \in [0,T].$$

The minimization of the vector valued functional $J(x,u)$ is understood w.r.t. the partial order $<_p$ on $\mathbb{R}^k$, defined as follows: Let $v, w \in \mathbb{R}^k$, then the vector v is *less than* w ($v <_p w$), if $v_i < w_i$ for all $i \in \{1, \ldots, k\}$. The relation $\leq_p$ is defined analogously. By this relation, we can introduce the concept of dominance and Pareto optimality (cf. [6], for instance).

Definition 1 (Dominated and Pareto Optimal Solutions) Let (x,u) and (x^*,u^*) be admissible points, i.e. they satisfy the restrictions of the MOCP.

a) The point (x,u) is *dominated* by the point (x^*,u^*) w.r.t. $J(x,u)$, if $J(x^*,u^*) \leq_p J(x,u)$ and $J(x,u) \neq J(x^*,u^*)$, otherwise (x,u) is *non-dominated* by (x^*,u^*).
b) The point (x^*,u^*) is called *Pareto optimal* if there exists no admissible (x,u) which dominates (x^*,u^*).
c) The set of all Pareto optimal points (x^*,u^*) is called the *Pareto set* and its image under J the *Pareto front*.

For classical, i.e. single-objective optimal control problems, direct methods have shown to be well suitable in many applications (cf. [1], for instance). Such methods transform the control problem into a high dimensional optimization problem by a time discretization. For solving MOCPs, multiobjective optimization techniques have to be applied to the discretized problem, cf. e.g. [10, 12, 15]. A number of methods exist for the computation of single Pareto points (cf. [6] for an overview).

To approximate the whole Pareto front, methods such as evolutionary algorithms (cf. [2]), set-oriented techniques, or path following methods (cf. [3, 14, 15] and the short overview given below) can be applied.

To transform the MOCP into an optimization problem with multiple objectives, we introduce a discrete time grid $\Delta t = \{t_0 = 0, t_1, \ldots, t_N = T\}$. The control u is approximated by a discrete control $u_d = \{u_k\}_{k=0}^{N-1}$ with u_k being an approximation of u on the interval $[t_k, t_{k+1}]$ for $k = 0, \ldots, N-1$. A discrete state trajectory $x_d = \{x_k\}_{k=0}^{N}$ with $x_k \approx x(t_k)$ can be obtained by a numerical integration scheme, $x_{k+1} = \Phi_{t_k}^{t_{k+1}}(x_k, u_k)$ with $x_0 = x(0)$ and for $k = 0, \ldots, N-1$. Together with an approximation of all objective functionals on the discrete time grid, we obtain a *multiobjective optimization problem*

$$\min_{u_d} J_d(x_d, u_d) = \sum_{k=0}^{N-1} C_d(x_k, u_k) + \Psi_d(x_N), \tag{2}$$

$$\text{w.r.t. } x_{k+1} = \Phi_{t_k}^{t_{k+1}}(x_k, u_k), \forall\, k < N,\ g_d(x_0, x_N) = 0,\ b_l \le \begin{pmatrix} x_k \\ u_k \end{pmatrix} \le b_u \forall\, k \le N. \tag{3}$$

2.1 Set-Oriented Subdivision

The aim of the subdivision method is to approximate the Pareto set by a successive refinement and selection of boxes, cf. [3, 15]. The procedure starts with a box that covers the admissible set of optimization parameters. Then, subdivision and selection steps are applied alternatingly. In a subdivision step, all active boxes are subdivided into smaller boxes. For the selection, a number of test points are chosen in all boxes and the objective functions are evaluated. Then, all boxes not containing any non-dominated test points are deleted and one proceeds with the next subdivision step (cf. Fig. 1, left). This a gradient-free sampling technique which, amongst other set-oriented algorithms, is implemented in the software package GAIO [3, 15].

2.2 Scalarization by Reference Point Techniques

A discretization with a fine time grid Δt leads to a high number of optimization parameters $u_d = \{u_0, \ldots, u_{N-1}\}$ in the transformed MOCP (2), (3). In this case, scalarization techniques have shown to be well suitable, cf. e.g. [10, 14]. More concretely, we apply a reference point method which defines auxiliary scalar optimization problems. To this aim, nonadmissible target points P in image space

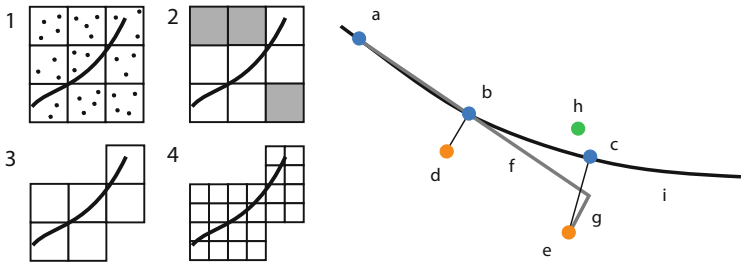


Fig. 1 *Left:* Subdivision method. Alternatingly, dominated boxes are removed and non-dominated boxes are subdivided. *Right:* Image space continuation method. If two points are known (say y_{i-1} and y_i), a target T_{i+1} is calculated. Then, the scalar minimization problem with initial guess x^p_{i+1} yields y_{i+1}

are defined, and the distance to the image of the admissible set is minimized,

$$\min_{u_d} \|J_d(x_d, u_d) - P\| \quad \text{w.r.t. constraints as in (3).}$$

As a result, single Pareto points on the boundary of the admissible set can be found. The target points are defined iteratively by a continuation method in image space as depicted in Fig. 1 to the right. These points are not necessarily Pareto optimal. However, dominated points can be easily eliminated by a subsequent non-dominance test. The auxiliary optimization problems can be efficiently solved by *sequential quadratic programming* (SQP) methods (cf. e.g. [1, Sect. 5.4] and the references therein).

While the set-oriented subdivision method works globally but it is restricted to moderate dimensions of the parameter space, SQP methods are suitable for high numbers of optimization parameters.

3 Application to the Electric Vehicle

A Matlab/SIMULINK model is used to describe the EV dynamics. It is based on a drivetrain efficiency map and considers vehicle properties such as rolling friction and air drag, as well as environmental conditions like slope and ambient temperature. The model holds several state variables such as position, velocity and state of charge. These variables depend on the input variables accelerator pedal position profile u and the inclination profile α. For a more detailed model description, we refer to [4, 11].

Since we aim to compute the Pareto set for the objectives "final state of charge $SOC(T)$" and "driven distance $s(T)$" with a fixed final time T, we set the vector of objective functionals (cf. Eq. (1)) to $J(x, u) = \Psi(x(T)) = (SOC(T), s(T))$. As an example scenario we choose a track with a periodic inclination profile superimposed

by a linear increase:

$$\alpha(s) = 4^\circ \sin\left(360^\circ \frac{s}{2000\,\mathrm{m}}\right) + 1^\circ \frac{s}{2000\,\mathrm{m}}.$$

This defines the height profile h. In this way, we ensure that most of the Pareto points (except for the solutions with very short driven distances) are computed for tracks with both uphill and downhill sections.

To compute the Pareto set we apply the algorithms presented in Sect. 2. The two solutions $u(t) = 0$ and $u(t) = 100$, respectively, correspond to the two endpoints of the Pareto front, where one objective becomes minimal while the other becomes maximal. To improve numerical accuracy, these values have been used to normalize both objectives to the interval $[0, 1]$ with the optimum being 0. For the results shown in the following, the normalization has been reversed. In this case, a maximization of both objectives is desired.

We start the **subdivision algorithm** with a box of dimension n (number of parameters) with the center at 50 and a radius of 50 so that it covers the whole pedal position profile $u_i \in [0, 100],\ i = 1, \ldots, n$. We then apply $4n$ subdivision steps. Figure 2 shows the resulting Pareto front for different pre-image dimensions on the left and one EV simulation with a Pareto optimal pedal position profile and the resulting velocity profile on the right. As has been observed before (cf. [4]), a high engine torque on positive slopes but lower torque on negative slopes is beneficial to the energy consumption.

It is obvious that solutions with a higher pre-image space dimension always have to be at least as good as the lower dimensional solutions (cf. Fig. 2). Additionally, the difference between the solutions is largest in the middle section. This is due to a higher variability in this part while near the ends of the front, the pedal position has to be close to the maximal or minimal value at all times.

When looking at the Pareto points around $SOC(T) \approx 0.745$ (as well as $SOC \approx 0.725$ for $u \in \mathbb{R}^{10}$), one observes a gap which is caused by the EV's recuperation

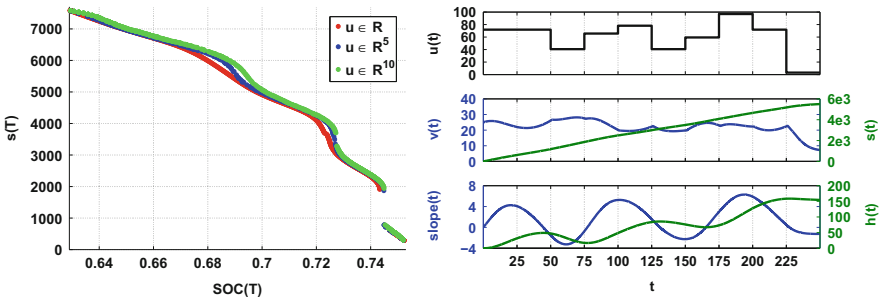


Fig. 2 *Left:* Pareto front computed by the subdivision algorithm for different pre-image dimensions (boxes represented by their center points). *Right:* EV simulation with a Pareto optimal pedal position profile ($u \in \mathbb{R}^{10}$, $SOC(T) = 0.6914$, $s(T) = 5800\,\mathrm{m}$)

technique. The last point at the low distance part of the Pareto front corresponds to a stop at the top of a hill. Increasing the pedal position profile only slightly results in a final position with a negative inclination $\alpha(s(T))$. Since the EV can roll down the slope and recharge its battery via recuperation, a slight reduction of the objective $SOC(T)$ leads to a huge increase of the second objective $s(T)$ which then results in a gap in the front. The varying inclination of the Pareto front is a result of the track slope alternating between positive and negative values.

It should be mentioned that due to the relatively long EV simulation time ($\approx$1 s), the number of testpoints for each box was set to a comparably low value of 30 which may cause boxes to be either eliminated or identified as non-dominated by mistake. The first case leads to spurious gaps in the Pareto front (cf. e.g. the Pareto front of $u \in \mathbb{R}^{10}$ in Fig. 2 for $SOC(T) \approx 0.68$) while the second case leads to boxes apart from the Pareto front.

A comparison of the results of the subdivision and the **image space continuation algorithm** (cf. Fig. 3, left) shows good agreement for the case $u \in \mathbb{R}^{10}$, indicating that the image space continuation method also yields good results despite its local nature. Having shown the continuation algorithm's applicability, Pareto sets of higher dimension are computed (cf. Fig. 3, left).

To improve the simulation time and convergence rate, each Pareto point from a lower pre-image space dimension serves as the initial guess for the next higher dimensional solution. As can be seen in Fig. 3, the resulting improvements become smaller quickly. The choice of the pre-image space dimension should be considered carefully since computation time increases significantly with the number of optimization variables. This effect is even strengthened by an observed decreasing convergence rate for high-dimensional cases, presumably caused by inaccuracies in the numerical differentiation of the EV model required for the SQP method.

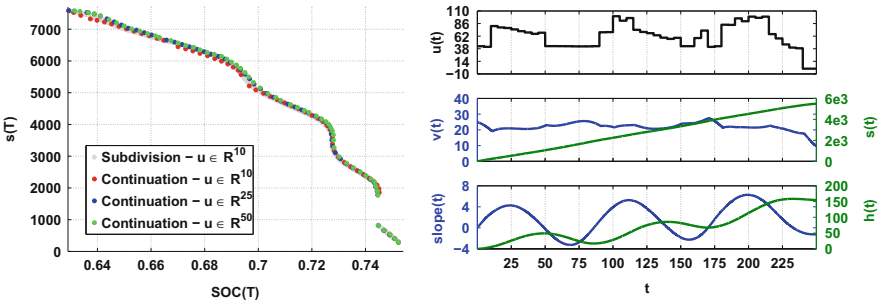


Fig. 3 *Left:* Pareto front computed by the image space continuation algorithm for different dimensions of pre-image space. *Right:* EV simulation with a Pareto optimal pedal position profile ($u \in \mathbb{R}^{50}$, $SOC(T) = 0.6934$, $s(T) = 5750\,\mathrm{m}$)

4 Conclusion

In this paper, we apply two MOCP algorithms for the development of an intelligent cruise control. Pedal position profiles can be chosen as optimal compromises between energy consumption and travel distance.

For future work, it will be interesting to compute the Pareto set with a constant travel distance instead of a constant driving time. Moreover, Model Predictive Control methods can be applied to realize real time optimization.

Acknowledgements This research was partially funded by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster 'Intelligent Technical Systems OstWestfalenLippe' (it's OWL) and managed by the Project Management Agency Karlsruhe (PTKA).

References

1. Binder, T., Blank, L., Bock, H., Bulirsch, R., Dahmen, W., Diehl, M., Kronseder, T., Marquardt, W., Schlöder, J., Stryk, O.: Introduction to model based optimization of chemical processes on moving horizons. In: Grötschel, M., et al. (ed.) Online Optimization of Large Scale Systems: State of the Art, pp. 295–340. Springer, Berlin (2001)
2. Coello Coello, C., Lamont, G., Veldhuizen, D.V.: Evolutionary Algorithms for Solving Multi-Objective Optimization Problems, 2nd edn. Springer, Boston (2007)
3. Dellnitz, M., Schütze, O., Hestermeyer, T.: Covering pareto sets by multilevel subdivision techniques. J. Optim. Theory Appl. **124**(1), 113–136 (2005)
4. Dellnitz, M., Eckstein, J., Flaßkamp, K., Friedel, P., Horenkamp, C., Köhler, U., Ober-Blöbaum, S., Peitz, S., Tiemeyer, S.: Development of an intelligent cruise control using optimal control methods. Proc. Technol. **15**, 285–294 (2014)
5. Dib, W., Serrao, L., Sciarretta, A.: Optimal control to minimize trip time and energy consumption in electric vehicles. In: Vehicle Power and Propulsion Conference (VPPC), pp. 1–8 (2011)
6. Ehrgott, M.: Multicriteria Optimization, 2nd edn. Springer, Berlin (2005)
7. Hellström, E., Åslund, J., Nielsen, L.: Design of an efficient algorithm for fuel-optimal look-ahead control. Control Eng. Pract. **18**(11), 1318–1327 (2010)
8. Keichel, M., Schwedes, O.: Das Elektroauto: Mobilität im Umbruch. Springer Vieweg, Wiesbaden (2013)
9. Li, S., Li, K., Rajamani, R., Wang, J.: Model predictive multi-objective vehicular adaptive cruise control. IEEE Trans. Control Syst. Technol. **19**(3), 556–566 (2011)
10. Logist, F., Houska, B., Diehl, M., Van Impe, J.: Fast Pareto set generation for nonlinear optimal control problems with multiple objectives. Struct. Multidiscip. Optim. **42**(4), 591–603 (2010)
11. Masjosthusmann, C., Köhler, U., Decius, N., Büker, U.: A vehicle energy management system for a battery electric vehicle. In: Vehicle Power and Propulsion Conference (VPPC), pp. 339–344 (2012)
12. Ober-Blöbaum, S., Ringkamp, M., Zum Felde, G.: Solving multiobjective optimal control problems in space mission design using discrete mechanics and reference point techniques. In: 51st IEEE International Conference on Decision and Control, pp. 5711–5716 (2012)
13. Petit, N., Sciarretta, A.: Optimal drive of electric vehicles using an inversion-based trajectory generation approach. In: Proceedings of the 18th IFAC World Congress, pp. 14519–14525 (2011)

14. Romaus, C., Bocker, J., Witting, K., Seifried, A., Znamenshchykov, O.: Optimal energy management for a hybrid energy storage system combining batteries and double layer capacitors. In: Energy Conversion Congress and Exposition (ECCE), 2009, pp. 1640–1647. IEEE, Piscataway (2009)
15. Schütze, O., Witting, K., Ober-Blöbaum, S., Dellnitz, M.: Set oriented methods for the numerical treatment of multiobjective optimization problems. In: Tantar, E., Tantar, A.A., Bouvry, P., Del Moral, P., Legrand, P., Coello Coello, C.A., Schütze, O. (eds.) EVOLVE- A Bridge between Probability, Set Oriented Numerics and Evolutionary Computation. Studies in Computational Intelligence, vol. 447, pp. 187–219. Springer, Berlin/Heidelberg (2013)
16. Sciarretta, A., Guzzella, L.: Control of hybrid electric vehicles. IEEE Control Syst. **27**(2), 60–70 (2007)

MS 26
MINISYMPOSIUM: PARAMETERIZED MODEL ORDER REDUCTION METHODS FOR COMPLEX MULTIDIMENSIONAL SYSTEMS

Organizers

Francesco Ferranti[1] and Wil Schilders[2]

Speakers

Yao Yue[3], Suzhou Li[3], Lihong Feng[3], Andreas Seidel-Morgenstern[3] and Peter Benner[3]
Accelerating Uncertainty Quantification of Linear Simulated Moving Bed Chromatography Models by Krylov-Type (Parametric) Model Order Reduction Methods

Domenico Spina[4], Francesco Ferranti[1], Tom Dhaene[4], Luc Knockaert[4] and Giulio Antonini[5]
Polynomial Chaos and Model Order Reduction for Efficient Variability Analysis

Massimiliano Lupo Pasini[6], Simona Perotto[7] and Alessandro Veneziani[6]
Hi-Mod Reduction Driven by a POD Strategy

[1]Francesco Ferranti, Vrije Universiteit Brussel, Brussels, Belgium.
[2]Wil Schilders, TU Eindhoven, Eindhoven, Netherlands.
[3]Yao Yue, Suzhou Li, Lihong Feng, Andreas Seidel-Morgenstern and Peter Benner, Max Planck Institute Magdeburg, Magdeburg, Germany.
[4]Domenico Spina, Tom Dhaene and Luc Knockaert, Ghent University, Ghent, Belgium.
[5]Giulio Antonini, Università degli Studi dell'Aquila, L'Aquila, Italy.
[6]Massimiliano Lupo Pasini and Alessandro Veneziani, Emory University, Atlanta, USA.
[7]Simona Perotto, Politecnico di Milano, Milano, Italy.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_88

Laura Iapichino[8], Alfio Quarteroni[9], Gianluigi Rozza[10] and Stefan Volkwein[8]
Reduced Basis Method and Domain Decomposition for Viscous Flows in Parametrized Complex Networks

Luca Daniel[11]
Accelerating Optimization, Inverse Problems and Uncertainty Quantification on Complex Systems via Parameterized Model Order Reduction

Olivier Maury[12]
The Use of Parameterized Cells in EDA Softwares

Wil Schilders[2]
EU-MORNET: European Model Reduction Network

Keywords

Dynamical systems
Model order reduction
Parameterized model order reduction
Reduced basis method

Short Description

In recent years, Parameterized Model Order Reduction (PMOR) methods have attracted a lot of attention from several scientific and industrial communities (e.g. electrical, chemical and biomedical engineering) as a powerful tool to significantly speed up analysis and design of complex systems. Several analysis and design tasks, such as design optimization, sensitivity and variability analysis, require multiple simulations of the system behavior for multiple values of the design parameters (e.g. layout features of an electronic system).

Using physics-based solvers (e.g. electromagnetic solvers to solve Maxwell's equations and fluid dynamic solvers to solve Navier-Stokes equations) for these tasks becomes very computationally expensive. PMOR methods are advanced

[8]Laura Iapichino and Stefan Volkwein, University of Konstanz, Konstanz, Germany.
[9]Alfio Quarteroni, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
[10]Gianluigi Rozza, International School for Advanced Studies of Trieste, Trieste, Italy.
[11]Luca Daniel, Massachusetts Institute of Technology, Cambridge, USA.
[12]Olivier Maury, Mentor Graphics, France.

modeling and mathematical tools that allow a very significant reduction of the computational cost of crucial analysis and design tasks, without compromising the accuracy of the results.

This mini-symposium is focused on PMOR methods for complex multidimensional systems. The talks of this mini-symposium discuss state-of-the-art PMOR methods in different domains: chemical processes, nonlinear electronic systems, fluid dynamic systems and delayed differential systems.

Reduced Basis Method for the Stokes Equations in Decomposable Parametrized Domains Using Greedy Optimization

Laura Iapichino, Alfio Quarteroni, Gianluigi Rozza, and Stefan Volkwein

Abstract In this paper we present a reduced order method for the solution of parametrized Stokes equations in domain composed by an arbitrary number of predefined shapes. The novelty of the proposed approach is the possibility to use a small set of precomputed bases to solve Stokes equations in very different computational domains, defined by combining one or more reference geometries. The selection of the basis functions is performed through an optimization greedy algorithm.

Keywords Parameterized model order reduction • Reduced basis method • Stokes equation

1 Introduction

Flow simulations in pipelined channels and several kinds of parametrized configurations have a growing interest in many life sciences and industrial applications. Applications may be found in the analysis of the blood flow in specific compartments of the circulatory system that can be represented as a combination of few deformed vessels from reference ones, e.g. pipes. We propose a solution approach that is particularly suitable for the study of internal flows in hierarchical parametrized geometries. The main motivation is for applications requiring rapid and reliable numerical simulations of problems in domains involving parametrized complex geometries. The classical reduced basis (RB) method is very effective to

L. Iapichino (✉) • S. Volkwein
University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany
e-mail: Laura.Iapichino@uni-konstanz.de; Stefan.Volkwein@uni-konstanz.de

A. Quarteroni
École Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland
e-mail: alfio.quarteroni@epfl.ch

G. Rozza
International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, Italy
e-mail: gianluigi.rozza@sissa.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_89

address viscous flows equations in parametrized geometries (see, e.g., [10]). An interesting alternative foresees a combination of RB with a domain decomposition approach. In this respect, preliminary efforts to reduce the global parametrized problem to local ones have led to the introduction of the so-called reduced basis element method to solve the Stokes problem [6], and more recently to the reduced basis hybrid method [3] and to the static condensation method [7]. In general, we are interested in defining a method able to maintain the flexibility of dealing with arbitrary combinations of subdomains and several geometrical deformations of the latter. A further new contribution to this field is the computation of the reduced basis functions through an optimization greedy algorithm [11].

2 Problem Setting

The method we present is a model order technique for solving a parametrized Stokes problem in a domain Ω defined by an arbitrary non-overlapping union of one or more predefined smaller geometries. For instance, we consider the geometry Λ depicted in the left plot of Fig. 1 representing a stenosis of the longitudinal section of an artery. This geometry can be interpreted as a two-dimensional model of a pipe and its deformation defined through the Boundary Displacement Dependent Transfinite Map $T(\boldsymbol{\mu})$ introduced in [4]. In particular we fix the size of the pipe as well as the position and the length of the occlusion or the dilatation of the pipe. We consider $D = 4$, $L = 1$, $S = 1$, $C = 2.5$, $T = 1$, $H = 1.5$. The parameter vector $\boldsymbol{\mu}_\Lambda = (\mu_1, \mu_2)$ allows to either "inflate" or "compress" the pipe, where $\boldsymbol{\mu}_\Lambda$ belongs to a closed and bounded subset $\mathscr{D}_\Lambda \subset \mathbb{R}^2$. We consider the computational domain of interest Ω as a network representing a channel with curved upper and bottom walls and composed by an arbitrary finite number of stenosed geometries $\Omega_i = T(\boldsymbol{\mu}_i)\Lambda$, $i = 1, \ldots, K$, for instance $K = 4$ in the right plot of Fig. 1. In this example the network is parametrized through eight parameters, two for each stenosed subdomain, $\boldsymbol{\mu} = (\boldsymbol{\mu}^1; \boldsymbol{\mu}^2; \boldsymbol{\mu}^3; \boldsymbol{\mu}^4)$ with $\boldsymbol{\mu}^i = (\mu_1^i, \mu_2^i)$ for $i = 1, \ldots, 4$. We impose homogeneous Dirichlet boundary conditions (BC) on both the upper and bottom walls of the domain, homogeneous Neumann BC on the outflow boundary (on the left) and non-homogeneous Neumann BC on the inflow boundary of the channel (on the right). Let us consider the following steady Stokes problem for a fluid of constant density [8] in the domain $\Omega \subset \mathbb{R}^2$ with mixed boundary conditions

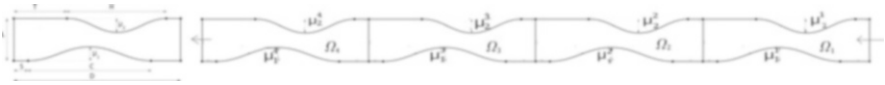


Fig. 1 Stenosis geometry Λ (*left*) and geometrical scheme for the curved channel Ω (*right*)

on $\Gamma = \Gamma_{in} \cup \Gamma_{out} \cup \Gamma_w$:

$$\begin{aligned} &-\nu\Delta\mathbf{u} + \nabla p = \mathbf{f} \text{ in } \Omega, \quad \nabla\cdot\mathbf{u} = 0 \text{ in } \Omega, \quad \mathbf{u} = 0 \text{ on } \Gamma_w, \\ &\boldsymbol{\sigma}_n^{in} := \nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}} - p\mathbf{n} = 1 \text{ on } \Gamma_{in}, \quad \boldsymbol{\sigma}_n^{out} := \nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}} - p\mathbf{n} = 0 \text{ on } \Gamma_{out}, \end{aligned} \tag{1}$$

where $\mathbf{u} = \mathbf{u}(\mathbf{x}) \in \mathbb{R}^2$ is the fluid velocity, $p = p(\mathbf{x})$ the pressure, $\mathbf{f} = \mathbf{f}(\mathbf{x}) \in \mathbb{R}^2$ a force field (e.g. gravity), $\nu > 0$ a kinematic viscosity and $\mathbf{n} = \mathbf{n}(\mathbf{x}) \in \mathbb{R}^2$ the normal outward unit vector to the domain boundary; Γ_{in} and Γ_{out} represent the inflow and outflow, respectively, while Γ_w is a boundary-wall. On Ω we introduce the velocity space and the pressure space, respectively, as $\tilde{Y} = \left\{\mathbf{v} \in (H^1(\Omega))^2 : \mathbf{v}\,|_{\Gamma_w} = \mathbf{0}\right\}$, $\tilde{M} = L^2(\Omega)$. Now, (1) in weak formulation reads: find $(\mathbf{u}, p) \in \tilde{Z} = (\tilde{Y} \times \tilde{M})$:

$$\tilde{a}(\mathbf{u}, \mathbf{v}; \boldsymbol{\mu}) + \tilde{b}(\mathbf{v}, p) = \tilde{f}(\mathbf{v}; \boldsymbol{\mu}), \quad \tilde{b}(\mathbf{u}, q; \boldsymbol{\mu}) = 0 \quad \forall (\mathbf{v}, q) \in \tilde{Z}. \tag{2}$$

As shown in [6], the continuously differentiable parametric map $T(\boldsymbol{\mu}_i)$ and its Jacobian $\boldsymbol{J}_i$ allow the definition of the bilinear and linear forms on the deformed subdomains, $\Omega_i = T(\boldsymbol{\mu}_i)\Lambda$, through the evaluation of the corresponding forms in the reference domain $\Lambda \subset \mathbb{R}^2$:

$$\tilde{a}(\mathbf{v}, \mathbf{w}; \boldsymbol{\mu}) = \sum_{i=1}^{K} \nu \int_{\Omega_i} \nabla\mathbf{v} : \nabla\mathbf{w}\, d\Omega_i = \sum_{i=1}^{K} \nu \int_{\Lambda} \boldsymbol{J}_i^{-\top}\nabla\mathbf{v} : \boldsymbol{J}_i^{-\top}\nabla\mathbf{w}|\boldsymbol{J}_i|\, d\Lambda, \tag{3a}$$

$$\tilde{b}(\mathbf{v}, q; \boldsymbol{\mu}) = -\sum_{i=1}^{K} \int_{\Omega_i} q\nabla\cdot\mathbf{v} d\Omega_i = -\sum_{i=1}^{K} \int_{\Lambda} q\nabla\cdot(\boldsymbol{J}_i^{-1}\mathbf{v})|\boldsymbol{J}_i| d\Lambda, \tag{3b}$$

where $|\boldsymbol{J}_i|$ denote the determinants of $\boldsymbol{J}_i$, $i = 1, \ldots, K$. For the right-hand-side let

$$\tilde{f}(\mathbf{v}; \boldsymbol{\mu}) = \sum_{i=1}^{K} \int_{\Lambda} \mathbf{f}\cdot\mathbf{v}|\boldsymbol{J}_i| d\hat{\Omega} + \int_{\hat{\Gamma}_i^{in} \cup \hat{\Gamma}_i^{out}} \boldsymbol{\sigma}_n \cdot \mathbf{v}|\boldsymbol{J}_i| d\hat{\Gamma}_{\hat{\Omega}_i}, \tag{3c}$$

where $\hat{\Gamma}_i^{in}$ and $\hat{\Gamma}_i^{out}$ stand for the inflow and outflow boundary, respectively, of the transformed domain $\hat{\Omega} = \cup_{i=1}^{K} \overline{T^{-1}(\boldsymbol{\mu}_i)\Omega_i}$. Since the bilinear forms $\tilde{a}(\cdot, \cdot; \boldsymbol{\mu})$, $\tilde{b}(\cdot, \cdot; \boldsymbol{\mu})$ are continuous and $\tilde{a}(\cdot, \cdot; \boldsymbol{\mu})$ is coercive, problem (2) admits a unique solution; see, e.g., [9]. We collect the contributions from the transposed inverse Jacobians and the Jacobian determinants in the tensors $\tilde{\boldsymbol{\nu}}$ and $\tilde{\boldsymbol{\chi}}$, for viscous and pressure terms, respectively, and use the elements of these tensors as the parameter dependent functions: $\tilde{\boldsymbol{\nu}}(\hat{\mathbf{x}}, \boldsymbol{\mu}_i) = \boldsymbol{J}_i^{-1}(\hat{\mathbf{x}})\boldsymbol{J}_i^{-\top}(\hat{\mathbf{x}})|\boldsymbol{J}_i(\hat{\mathbf{x}})|$ and $\tilde{\boldsymbol{\chi}}(\hat{\mathbf{x}}, \boldsymbol{\mu}_i) = \boldsymbol{J}_i^{-1}(\hat{\mathbf{x}})|\boldsymbol{J}_i(\hat{\mathbf{x}})|$. Since the tensors $\tilde{\boldsymbol{\nu}}$, $\tilde{\boldsymbol{\chi}}$ and the determinants $|\boldsymbol{J}_i|$ are non-affine (due to the use of a transfinite map) for $i = 1, \cdots, K$, we apply the empirical interpolation procedure [1] in order to approximate them into affine functions defined as sums of some parameter dependent coefficients $\Theta^m(\boldsymbol{\mu})$, $\Phi^m(\boldsymbol{\mu})$, $\Psi^m(\boldsymbol{\mu})$ and functions $\boldsymbol{\nu}^m$, $\boldsymbol{\chi}^m$, $\boldsymbol{j}^m$ depending

only on spatial coordinates [10], e.g. $\tilde{\boldsymbol{\nu}}(\hat{\mathbf{x}}, \boldsymbol{\mu}) \approx \sum_{m=1}^{M_a} \Phi^m(\boldsymbol{\mu}) \boldsymbol{\nu}^m(\hat{\mathbf{x}})$. Thanks to these interpolations, we can approximate (3a) with

$$a(\mathbf{v}, \mathbf{w}; \boldsymbol{\mu}) = \sum_{i=1}^{K} \sum_{m=1}^{M_a} \Theta^m(\boldsymbol{\mu}_i) \nu \int_{\Lambda} \boldsymbol{\nu}^m \nabla \mathbf{v} : \nabla \mathbf{w} \, d\Lambda. \tag{4}$$

The forms $b(\mathbf{v}, q; \boldsymbol{\mu})$ and $f(\mathbf{v}; \boldsymbol{\mu})$ approximating $\tilde{b}(\mathbf{v}, q; \boldsymbol{\mu})$ and $\tilde{f}(\mathbf{v}; \boldsymbol{\mu})$ are defined similarly. These recovered affine decompositions are crucial even for a classical discretization technique, e.g. finite element (FE) method. Once the FE scheme is defined, this decoupling allows to split all the computations not involving the parameters (concerning discretization) in an offline stage. In the online stage we can easily assemble the forms $a(\mathbf{v}, \mathbf{w}; \boldsymbol{\mu}), b(\mathbf{v}, q; \boldsymbol{\mu})$ and $f(\mathbf{v}; \boldsymbol{\mu})$ by summing the fast evaluations of the parametric functions and the integrals already computed; see (4). Once the computations of all the integrals are done, for every new $\boldsymbol{\mu}$ and for any number K of stenosed subdomains in the network Ω, we define the correspondent reference domain $\hat{\Omega}$ as non-overlapping union of the K reference domains $\hat{\Omega} = \cup_{i=1}^{K} \overline{T^{-1}(\boldsymbol{\mu}_i)\Omega_i}$ and the Stokes problem can be efficiently assembled and written as: find $(\mathbf{u}(\boldsymbol{\mu}), p(\boldsymbol{\mu})) \in Z = Y \times M$ such that

$$a(\mathbf{u}(\boldsymbol{\mu}), \mathbf{v}, \boldsymbol{\mu}) + b(\mathbf{v}, p(\boldsymbol{\mu}), \boldsymbol{\mu}) = f(\mathbf{v}, \boldsymbol{\mu}), \quad b(\mathbf{u}(\boldsymbol{\mu}), q, \boldsymbol{\mu}) = 0 \quad \forall (\mathbf{v}, q) \in Z, \tag{5}$$

where we set $Y = \{\mathbf{v} \in (H^1(\hat{\Omega}))^2 : \mathbf{v} \mid_{\hat{\Gamma}_w} = 0\}$, $M = L^2(\hat{\Omega})$ and $\hat{\Gamma}_w$ denotes the boundary-wall of the transformed domain $\hat{\Omega}$. Even if some computations can be performed in one offline parameter independent stage, the solution of (5) for a many different parameters using a classical numerical technique (e.g. FE) requires many solutions of a typically large linear systems.

3 The Reduced Basis (RB) Method for Decomposable Domains

The reduced scheme we propose in this paper consists in approximating the spaces Y and M with small dimensional spaces Y_N and M_N (the so called RB spaces) where the solution of system (5) is looked for. In particular, the RB spaces Y_N and M_N are generated by the direct sum of the subspaces $Y_L^i = \text{span}\{\mathbf{w}_j^i, j = 1, \ldots, L\}$ and $M_N^i = \text{span}\{q_j^i, j = 1, \ldots, N\}$ for $i = 1, \ldots, K$, respectively, representing small sets of basis functions with support on the subdomains $\Lambda_i = T^{-1}(\boldsymbol{\mu}_i)\Omega_i$ of $\hat{\Omega}$ (compare (7)): $Y_L = Y_L^1 \oplus \cdots \oplus Y_L^K$ and $M_N = M_N^1 \oplus \cdots \oplus M_N^K$. The RB approximation of problem (5) reads: find $(\mathbf{u}^L(\boldsymbol{\mu}), p^N(\boldsymbol{\mu})) \in Z_{LN} = Y_L \times M_N$ such that

$$a(\mathbf{u}^L(\boldsymbol{\mu}), \mathbf{v}, \boldsymbol{\mu}) + b(\mathbf{v}, p^N(\boldsymbol{\mu}), \boldsymbol{\mu}) = f(\mathbf{v}, \boldsymbol{\mu}), \quad b(\mathbf{u}^L(\boldsymbol{\mu}), q, \boldsymbol{\mu}) = 0 \quad \forall (\mathbf{v}, q) \in Z_{LN}. \tag{6}$$

In terms of computational effort, the method consists in defining, during the offline stage, for $i = 1, \ldots, K$, the reduced basis functions $\mathbf{w}_j^i$, $j = 1, \ldots, L$ and q_j^i, $j = 1, \ldots, N$ and the $\boldsymbol{\mu}$ independent part of system (6). The latter consists in assembling the matrices containing the evaluations of the integrals of the linear and bilinear forms involving the functions $\mathbf{w}_j^i$ and q_j^i; see [2]. During the offline stage of the method, we store K small matrices each one of dimensions $L \times L$, $L \times N$ and $L \times 1$, respectively. In the online stage, we sum up these matrices, with the respective parametric function and we solve a system that is much smaller the ones needed for a classical numerical discretization, precisely $K(L+N) \times K(L+N)$.

Basis Functions Computations In this section we illustrate the procedure to compute for $i = 1, \ldots, K$ the basis functions $\mathbf{w}_l^i$ and q_j^i, for $l = 1, \ldots, L$ and $j = 1, \ldots, N$. They are defined as follows:

$$\mathbf{w}_l^i|_{\Lambda_i} = \boldsymbol{\xi}_l,\ \mathbf{w}_l^i|_{\hat{\Omega}\setminus\Lambda_i} = \mathbf{0} \quad \text{and} \quad q_j^i|_{\Lambda_i} = \eta_j,\ q_j^i|_{\hat{\Omega}\setminus\Lambda_i} = 0. \tag{7}$$

As we are considering the simplified case with only one reference geometry (the stenosis of Fig. 1), such that $T^{-1}(\Omega_i, \boldsymbol{\mu}_i) = \Lambda$ for every $i = 1, \ldots, K$, the functions $\boldsymbol{\xi}_l$ and η_j are the same for every $i = 1, \ldots, K$ and are defined through only one local problem. We consider the following Stokes problem defined in Λ:

$$\begin{aligned} &-\nu\Delta\mathbf{v}(\boldsymbol{\lambda}) + \nabla q(\boldsymbol{\lambda}) = \mathbf{f} \text{ in } \Lambda,\ \ \nabla\cdot\mathbf{v}(\boldsymbol{\lambda}) = 0 \text{ in } \Lambda,\ \ \mathbf{v}(\boldsymbol{\lambda}) = 0 \text{ on } \Gamma_w, \\ &\nu\frac{\partial\mathbf{v}(\boldsymbol{\lambda})}{\partial\mathbf{n}} - q(\boldsymbol{\lambda})\mathbf{n} = \boldsymbol{\lambda}^{in} \text{ on } \Gamma_{in}^{\Lambda},\ \ \nu\frac{\partial\mathbf{v}(\boldsymbol{\lambda})}{\partial\mathbf{n}} - q(\boldsymbol{\lambda})\mathbf{n} = \boldsymbol{\lambda}^{out} \text{ on } \Gamma_{out}^{\Lambda}, \end{aligned} \tag{8}$$

where ν and $\mathbf{f}$ are the same as in (1) and Γ_{in}^{Λ} and Γ_{out}^{Λ} denote the inflow and outflow boundary of Λ, respectively. Furthermore, $\boldsymbol{\lambda}^{in}(\mathbf{x}) = \sum_{i=1}^{N^{in}} \mu_i^{in}\phi_j(\mathbf{x})$ and $\boldsymbol{\lambda}^{out}(\mathbf{x}) = \sum_{i=1}^{N^{out}} \mu_i^{out}\tilde{\phi}_j(\mathbf{x})$ are distributed parameter functions in $L^2(\Gamma_{in}^{\Lambda})$ and $L^2(\Gamma_{out}^{\Lambda})$ defining the BCs of the problem. Problem (8) is a parametrized Stokes problem, whose parameter is $\boldsymbol{\lambda} = (\boldsymbol{\mu}_\Lambda, \boldsymbol{\mu}^{in}, \boldsymbol{\mu}^{out})$ and the correspondent parameter space $\mathscr{D} = \{\boldsymbol{\lambda} = (\boldsymbol{\mu}_\Lambda, \boldsymbol{\mu}^{in}, \boldsymbol{\mu}^{out}), \boldsymbol{\mu}_\Lambda \in \mathscr{D}_\Lambda, \boldsymbol{\mu}^{in} \in [\boldsymbol{\mu}_a^{in}, \boldsymbol{\mu}_b^{in}] \subset \mathbb{R}^{N^{in}}, \boldsymbol{\mu}^{out} \in [\boldsymbol{\mu}_a^{out}, \boldsymbol{\mu}_b^{out}] \subset \mathbb{R}^{N^{out}}\}$. Upon introducing the velocity space and the pressure space, respectively, as $V = \{\mathbf{v} \in (H^1(\Lambda))^2 : \mathbf{v}|_{\Gamma_w} = \mathbf{0}\}$, $Q = L^2(\Lambda)$, the weak formulation of (8) reads: find $(\mathbf{v}(\boldsymbol{\lambda}), q(\boldsymbol{\lambda})) \in X = V \times Q$:

$$\mathscr{A}(\mathbf{v}(\boldsymbol{\lambda}), \mathbf{w}; \boldsymbol{\lambda}) + \mathscr{B}(\mathbf{w}, q(\boldsymbol{\lambda}); \boldsymbol{\lambda}) = \mathscr{F}(\mathbf{w}; \boldsymbol{\lambda}),\ \mathscr{B}(\mathbf{v}(\boldsymbol{\lambda}), q; \boldsymbol{\lambda}) = 0\ \forall (\mathbf{w}, q) \in X, \tag{9}$$

where the linear and bilinear forms are defined as done in the previous section. We use (9) to define the reduced basis spaces and select of small set of parameter values (described in the next section), $S_N = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N\}$. The solutions $(\mathbf{v}(\boldsymbol{\lambda}_j), q(\boldsymbol{\lambda}_j))$, $j = 1, \ldots, N$ of (9) found by using a classical numerical technique (e.g. FE) and in correspondence of the parameter values of set S_N will represent the first sets of basis functions needed. In order to guarantee the approximation stability of the reduced basis scheme, we need to fulfill the inf-sup condition [10].

This is achieved by enriching the velocity subspace with some additional basis functions as follows. For every pressure solution $q(\boldsymbol{\lambda}_j)$, we introduce $\mathbf{w}(\boldsymbol{\lambda}_j) = \arg\sup_{\mathbf{v}\in V} \mathscr{B}(\mathbf{v}, q(\boldsymbol{\lambda}_j); \boldsymbol{\lambda}_j)/\|\mathbf{v}\|_V$. Now we define the basis functions $\boldsymbol{\xi}_l$ and η_j, $l = 1, \ldots, L = 2N, j = 1, \ldots, N$ of (7) as the orthonormal bases of the two spaces $V_N = \text{span}\,\{\mathbf{v}(\boldsymbol{\lambda}_j), \mathbf{w}(\boldsymbol{\lambda}_j), j = 1, \ldots, N\}$ and $Q_N = \text{span}\,\{q(\boldsymbol{\lambda}_j), j = 1, \ldots, N\}$.

Selection of the Parameter Set Using Greedy Optimization We suppose that we have defined the first N parameter values, the corresponding basis functions and the initial reduced basis spaces V_N and Q_N. We define now the local reduced approximation of problem (9): find $\mathbf{v}_N(\boldsymbol{\lambda}) \in V_N$, $q_N(\boldsymbol{\lambda}) \in Q_N$ such that

$$\begin{cases} \mathscr{A}(\mathbf{v}_N(\boldsymbol{\lambda}), \mathbf{w}; \boldsymbol{\lambda}) + \mathscr{B}(\mathbf{w}, q_N(\boldsymbol{\lambda}); \boldsymbol{\lambda}) = \mathscr{F}(\mathbf{w}; \boldsymbol{\lambda}) & \forall \mathbf{w} \in V_N, \\ \mathscr{B}(\mathbf{v}_N(\boldsymbol{\lambda}), q; \boldsymbol{\lambda}) = 0 & \forall q \in Q_N. \end{cases} \tag{10}$$

Thus, we can define the space $X_N = V_N \times Q_N$, its dual $X_N' = V_N' \times Q_N'$ and operators $\mathscr{K}(\cdot, \cdot; \boldsymbol{\lambda}) \in L(X_N \times X_N, X_N')$, $\mathscr{R}(\cdot; \boldsymbol{\lambda}) \in L(X_N, X_N')$ so that (10) can be written in the compact form: find $\mathbf{z}_N(\boldsymbol{\lambda}) = (\mathbf{v}_N(\boldsymbol{\lambda}), q_N(\boldsymbol{\lambda})) \in X_N$

$$\mathscr{K}\left(\mathbf{z}_N(\boldsymbol{\lambda}), \boldsymbol{\psi}; \boldsymbol{\lambda}\right) = \mathscr{R}(\boldsymbol{\psi}; \boldsymbol{\lambda}) \quad \forall \boldsymbol{\psi} \in X_N. \tag{11}$$

The next parameter to add to the parameter set S_N will be the solution to (see [11]):

$$\min \hat{J}(\boldsymbol{\lambda}) \quad \text{subject to} \quad \boldsymbol{\lambda} \in \mathscr{D}, \tag{12}$$

where the cost functional is $\hat{J}(\boldsymbol{\lambda}) = -\|\mathscr{K}(\mathbf{z}_N(\boldsymbol{\lambda}), \cdot; \boldsymbol{\lambda}) - \mathscr{R}(\cdot; \boldsymbol{\lambda})\|_{X'}^2/2$ and $\mathbf{z}_N(\boldsymbol{\lambda})$ denotes the solution to (11) defined with the already selected basis functions. Of course, the space X has to be discretized to evaluate the dual norm in our numerical realization. We have not introduced a high dimensional (truth) approximation to simplify the presentation of the reduced basis approach and the greedy optimization algorithm. Since the transfinite map T is continuously differentiable, the cost J is continuously differentiable as well. Thus, we can characterise a local solution to (12) by first-order necessary optimality conditions; see, e.g., [11]. Therefore, we apply the projected gradient method combined with a line search based on the Armijo rule (see [5, Sect. 5.4]). The gradient of $\hat{J}$ at a given $\lambda \in \mathscr{D}$ is $\hat{J}'(\boldsymbol{\lambda}) = \mathscr{K}_{\boldsymbol{\lambda}}(\mathbf{z}_N(\boldsymbol{\lambda}), \mathbf{r}_N(\boldsymbol{\lambda}) + \mathbf{p}_N(\boldsymbol{\lambda}); \boldsymbol{\lambda}) - \mathscr{R}_{\boldsymbol{\lambda}}(\mathbf{r}_N(\boldsymbol{\lambda}) + \mathbf{p}_N(\boldsymbol{\lambda}); \boldsymbol{\lambda})$, where $\mathbf{p}_N = \mathbf{p}_N(\boldsymbol{\lambda}) \in X_N$ is the unique solution to the adjoint equation

$$\mathscr{K}(\boldsymbol{\psi}, \mathbf{p}_N; \boldsymbol{\lambda}) = -\mathscr{K}(\boldsymbol{\psi}, \mathbf{r}_N(\boldsymbol{\lambda}); \bar{\boldsymbol{\lambda}}) \quad \forall \boldsymbol{\psi} \in X_N.$$

and $\mathbf{r}_N = \mathbf{r}_N(\boldsymbol{\lambda}) \in X$ denotes the Riesz representant of the residual $\mathbf{R}_N = \mathscr{R}(\cdot; \boldsymbol{\lambda}) - \mathscr{K}(\mathbf{z}_N, \cdot; \boldsymbol{\lambda}) \in X'$. As a stopping criterion for the gradient projection method we use $\|\hat{J}'(\boldsymbol{\lambda}^{(k)})\|_{\mathbb{R}^d} \leq \tau_{abs} + \tau_{rel}\|\hat{J}'(\boldsymbol{\lambda}^{(0)})\|_{\mathbb{R}^d}$. We note that (12) may have several local minima (specially for large N), so that a good choice of the initial point is fundamental to reach the global minimum parameter value. In order to define a suitable starting value $\boldsymbol{\lambda}^{(0)}$, we consider a very coarse training set

$\Xi_{train} \subset \mathscr{D}$ and we define the starting value of the gradient projection method by $\boldsymbol{\lambda}^{(0)} = \arg\min_{\boldsymbol{\lambda} \in \Xi_{train}} \hat{J}(\boldsymbol{\lambda})$.

Numerical Results In this section, we present some numerical results obtained by solving problem (1) in the domain Ω introduced in Sect. 2. The FE computations are performed by using Taylor-Hood elements, in particular, in every stenosed subdomain we have 6538 $\mathbb{P}_2$ elements for velocity and supremizer, 850 $\mathbb{P}_1$ for pressure, respectively. Moreover the parameters values are $\mu_1^i \in [-0.2, 0.5]$, $\mu_2^i \in 2[-0.2, 0.3]$ and the parameters defining the local BCs considered for problem (8) are defined between $\boldsymbol{\mu}_a^{out} = \mathbf{0}, \boldsymbol{\mu}_a^{in} = \mathbf{0}$ and $\boldsymbol{\mu}_b^{out} = \mathbf{1}, \boldsymbol{\mu}_b^{in} = \mathbf{1}$, the functions $\phi_j(\mathbf{x}), \tilde{\phi}_j(\mathbf{x}), j = 1, \ldots, 5$ are the Fourier basis functions defined along Γ_{in}^{Λ} and Γ_{out}^{Λ}. In Fig. 2, we show the error decay between the RB solution and the FE one, by increasing the number of basis N used in the reduced scheme (we note that in this test $L = 2N, K = 4$). In Fig. 3, the RB solution for a particular parameter set is plotted and in Table 1 the computational times needed for the online FE and RB solutions are compared, by considering an increasing number of subdomains in Ω. We note that the proposed RB scheme allows to compute accurate solutions at a very low computational times and in many different computational domains.

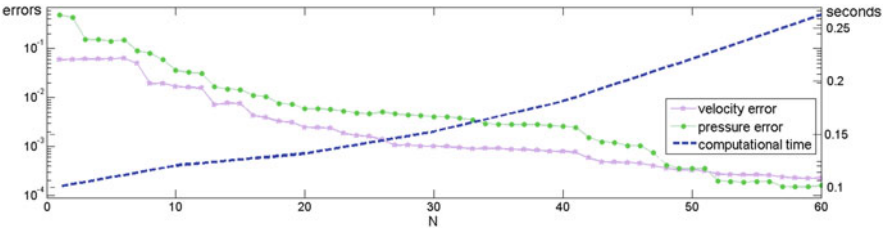


Fig. 2 Errors between the RB solution and the FE one and CPU RB times by increasing N

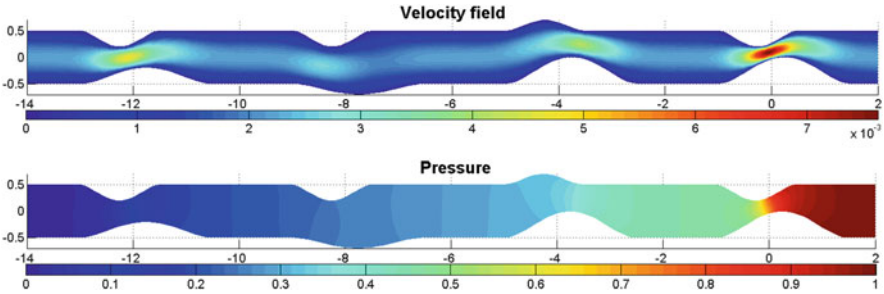


Fig. 3 The reduced basis solution (velocity field on the *top*, pressure on the *bottom*) corresponding to $K = 4$ and $\boldsymbol{\mu} = (0.5, 0.3; 0.5, -0.2; -0.2, 0.3; 0.3, 0.3)$, by using $N = 40$

Table 1 Computational online times (in seconds) needed for the solution computed with the FE method and the RB one (by using $N = 20$) by varying the number of subdomains in Ω

Method	$K = 6$	$K = 9$	$K = 12$	$K = 15$	$K = 18$	$K = 20$
FE online	2.93	4.46	6.64	7.91	10.00	11.14
RB online	0.13	0.20	0.34	0.53	0.70	0.79

References

1. Barrault, M., Maday, Y., Nguyen, N., Patera, A.: An "empirical interpolation" method: application to efficient reduced-basis discretization of partial differential equations. C. R. Math. Acad. Sci. Paris **339**(9), 667–672 (2004)
2. Iapichino, L.: Reduced basis methods for the solution of parametrized PDEs in repetitive and complex networks with application to CFD. Ph.D. thesis, N. 5529, École Polytechnique Fédérale de Lausanne (2012)
3. Iapichino, L., Quarteroni, A., Rozza, G.: A reduced basis hybrid method for the coupling of parametrized domains represented by fluidic networks. Comput. Methods Appl. Mech. Eng. **221–222**, 63–82 (2012)
4. Jäggli, C., Iapichino, L., Rozza, G.: An improvement on geometrical parameterizations by transfinite maps. C. R. Math. **352**(3), 263–268 (2014)
5. Kelley, C.T.: Iterative Methods for Optimization. Frontiers in Applied Mathematics. SIAM, Philadelphia (1999)
6. Løvgren, A., Maday, Y., Rønquist, E.: A reduced basis element method for complex flow systems. In: Wesseling, P., Onate, E., Periaux, J. (eds.) Proceedings of ECCOMAS CFD, TU Delft, The Netherlands (2006)
7. Phuong Huynh, D.B., Knezevic, D.J., Patera, A.T.: A static condensation reduced basis element method: approximation and a posteriori error estimation. ESAIM: Math. Model. Numer. Anal. **47**, 213–251 (2013)
8. Quarteroni, A.: Numerical Models for Differential Problems. Series MS and A, vol. 8. Springer, New York (2013)
9. Quarteroni, A., Valli, A.: Numerical Approximation of Partial Differential Equations, 1st edn. Springer, Berlin-Heidelberg (1994)
10. Rozza, G., Veroy, K.: On the stability of the reduced basis method for Stokes equations in parametrized domains. Comput. Methods Appl. Mech. Eng. **196**(7), 1244–1260 (2007)
11. Urban, K., Volkwein, S., Zeeb, O.: Greedy sampling using nonlinear optimization. In: Quarteroni, A., Rozza, R. (eds.) Reduced Order Methods for Modeling and Computational Reduction, pp. 137–157. Springer, Cham (2014)

MS 27
MINISYMPOSIUM: ROBUST VARIABLE-STRUCTURE APPROACHES FOR CONTROL AND ESTIMATION OF UNCERTAIN DYNAMIC PROCESSES

Organizers

Andreas Rauh[3] and Luise Senkel[1]

Speakers

Luise Senkel[1]
Experimental Validation of State and Parameter Estimation Using Sliding Mode Techniques with Bounded and Stochastic Disturbances

Horst Schulte[2]
Extension of Sliding Mode Observers for Fault Reconstruction: Comparison Between LPV and Takagi-Sugeno Model Approaches

Andreas Rauh[3]
Interval-Based Sliding Mode Control for High-Temperature Fuel Cells Under Actuator Constraints

[1]Luise Senkel, University of Rostock, Chair of Mechatronics, Rostock, Germany.
[2]Horst Schulte, HTW Berlin, Department of Engineering I, Control Engineering, Berlin, Germany.
[3]Andreas Rauh, University of Rostock, Chair of Mechatronics, Rostock, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_90

Piotr Lésniewski[4]
Sliding Mode Data Flow Regulation for Connection-Oriented Networks with Unpredictable Packet Loss Ratio

Keywords

Interval analysis
Linear matrix inequalities
Lyapunov functions
Sliding mode control and estimation
Stability analysis
Uncertainty
Variable-structure approaches

Short Description

In recent years, numerous variable-structure approaches have been developed for control of nonlinear dynamic systems and for the model-based estimation of non-measurable states and parameters. These approaches typically make use of first-order as well as higher-order sliding mode techniques and related procedures. One of their main advantages is the inherent proof of asymptotic stability. This stability proof is either performed offline during the corresponding controller as well as estimator design or online by the real-time evaluation of a suitable candidate for a Lyapunov function.

The methodological framework for variable-structure control and estimation approaches is quite well developed in the case of systems, for which process models are accurately known. Nevertheless, research efforts are still necessary to make the corresponding procedures applicable when only worst-case bounds are available for specific parameters (e.g. due to non-negligible manufacturing tolerances). Moreover, significant stochastic disturbances (e.g. as a result of measurement noise) may act as further system inputs in such applications. To enhance robustness in such cases, it is possible to combine techniques which are for instance based on interval analysis, stochastic differential equations, or linear matrix inequalities with variable-structure approaches. Verified stability analysis is a challenging task for finite-dimensional dynamic systems which are affected by bounded uncertainty. Finally, the adequate consideration of both actuator and state constraints represents a challenging task that is currently intensively investigated.

[4]Piotr Lésniewski, University of Łódź, Poland.

In this Minisymposium, ongoing research activities in the field of robust variable-structure control and estimation are presented. Novel methodological aspects as well as the use of variable-structure techniques in industrial applications including their efficient (software) implementation on hardware for real-time control are explained. Numerical verification and experimental validation for industry-motivated applications in control of fuel cell systems, automotive applications and mechanics as well as mechatronics show the applicability of the mentioned techniques.

Experimental Validation of State and Parameter Estimation Using Sliding-Mode-Techniques with Bounded and Stochastic Disturbances

Luise Senkel, Andreas Rauh, and Harald Aschemann

Abstract Uncertainties—more precisely bounded and stochastic disturbances—play a major role in control and estimation tasks in general. Examples for bounded uncertainty are lack of knowledge about specific parameters and manufacturing tolerances. Moreover, stochastic disturbances have a large influence on dynamic systems, especially on sensor measurements. These issues make it difficult to control a system such that robustness and stability are guaranteed if system parameters are not exactly known and system states cannot be measured with high accuracy due to process and measurement noise. Sliding mode techniques are known for their robustness, so that an extension of classical approaches is presented that accounts for uncertainties and estimates non-measurable states as well as unknown parameters.

Keywords Parameter estimation • Sliding mode control and estimation • Uncertainty

1 Introduction

The principle of sliding mode techniques in control theory is to affect a nonlinear dynamic system such that it tends to a user-defined stable operation mode and always stays in its near surrounding area (called sliding surface). Often, the nonlinear system is divided into a linear part and a (sometimes hardly mathematically describable) nonlinear one including unknown disturbances. Then, the task of sliding mode approaches is to compensate the second part by including a switching term into the control law (for trajectory tracking purposes) or into the observer part (for estimating non-measurable system states or identifying uncertain parameters). That is why, the advantage of existing sliding mode techniques is their finite-time convergence while reaching predefined sliding surfaces [1]. Since technical applications need to be controlled in a robust and stabilizing way, the

L. Senkel (✉) • A. Rauh • H. Aschemann
Chair of Mechatronics, University of Rostock, 18059 Rostock, Germany
e-mail: Luise.Senkel@uni-rostock.de; Andreas.Rauh@uni-rostock.de;
Harald.Aschemann@uni-rostock.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_91

estimation of non-measurable states and unknown parameters is necessary for control procedures. Therefore, a sliding mode observer that copes with uncertain parameters as well as with noisy measurements is proposed for a combined state and parameter estimation [7]. The advantage of the presented observer is robustness against uncertainty based on the usage of suitable candidates for Lyapunov functions as it is usually also done in existing sliding mode approaches to guarantee the system's stability [1]. In contrast to existing sliding mode observers, the presented approach, firstly, is not restricted by matching conditions leading to a more general applicability of the observer to different systems. Secondly, interval descriptions are used for uncertain states and parameters. Finally, not fully known nonlinearities that inevitably influence the stability of the system—as for example friction, wear of mechanic components or remanence of a brake—can be included as stochastic disturbances. Interval arithmetic is helpful to consider unknown influences on the system dynamics. Therefore, intervals for uncertain parameters, inaccurate measurements, and for estimation errors are taken into account. Additionally, the number of switching amplitudes of the presented sliding mode observer is equal to the number of measurements which enables an individual computation of this variable structure part gain. The switching amplitude is evaluated by using a suitable candidate of a Lyapunov function and the Itô differential operator for stochastic processes. Moreover, Pontryagin's maximum principle improves the parameter estimation in terms of an optimal input design.

2 Sliding Mode Techniques for State and Parameter Estimation

In general, state and parameter estimation is always necessary for good trajectory tracking in control purposes. In fact, the principle of estimation by so-called observers is to reconstruct unknown states by the simulation of the mathematically described real system (set ordinary differential state equations (ODEs)) in a parallel way to the real system. Then, the non-measurable states result from minimization of the difference between measurements and the corresponding estimates from the observer by choosing observer gain matrices. This standard procedure is extended in the following by a switching amplitude matrix that compensates additionally the non-modeled unknown influences (e.g. friction) in order to reconstruct the true system states in good accuracy. Parameter estimation is applied in a similar way because the parameters are interpreted as states under the assumption that the time derivatives of the parameters are close to zero. Consequently, the system parameters are assumed to be nearly constant or change only within tolerance bounds (intervals). Therefore, a dynamic system is taken into consideration which is given by the ODEs

$$\mathbf{f}\left(\mathbf{x}(t), \mathbf{p}, \mathbf{u}(t)\right) = \dot{\mathbf{x}}(t) = \mathbf{A}\cdot\mathbf{x}\left(t\right) + \mathbf{B}\cdot\mathbf{u}(t) + \mathbf{S}\cdot\boldsymbol{\xi}\left(\mathbf{x}(t), \mathbf{u}(t)\right), \; \mathbf{y}(t) = \mathbf{C}\cdot\mathbf{x}\left(t\right) \tag{1}$$

where the system, input and output matrices belong to the interval expressions $\mathbf{A} := \mathbf{A}(\mathbf{x}(t), \mathbf{p}) \in [\mathbb{A}]$, $\mathbf{B} := \mathbf{B}(\mathbf{x}(t), \mathbf{p}) \in [\mathbb{B}]$ and $\mathbf{C} := \mathbf{C}(\mathbf{x}(t), \mathbf{p}) \in [\mathbb{C}]$. In (1), the state vector $\mathbf{x}(t)$ includes also uncertain but bounded parameters $\mathbf{p}(t) \in [\mathbf{p}(t)]$. The control input vector is denoted by $\mathbf{u}(t)$. A-priori unknown as well as nonlinear terms are represented by $\mathbf{S} \cdot \boldsymbol{\xi}\,(\mathbf{x}(t), \mathbf{u}(t))$ with $\mathbf{S} \in \mathbb{R}^{n \times q}$ where the condition $\|\boldsymbol{\xi}\,(\mathbf{x}, \mathbf{u})\| \leq \overline{\boldsymbol{\xi}}$ with a fixed upper bound of the vector norm $\overline{\boldsymbol{\xi}}$ has to be fulfilled [5, 6]. For system (1), a classical sliding mode observer can be formulated as it is derived in [5, 6]. Based on this, the modified observer ODEs considering uncertainty become

$$\begin{aligned}\hat{\dot{\mathbf{f}}}(\hat{\mathbf{x}}\,(t)\,, [\mathbf{p}], \mathbf{u}(t)) =& \hat{\mathbf{f}}(\hat{\mathbf{x}}\,(t)\,, [\mathbf{p}], \mathbf{u}(t)) + \mathbf{P}^{+}[\hat{\mathbb{C}}]^{T} \cdot \mathbf{H}_s \cdot \operatorname{sign}(\mathbf{e}_m + [\Delta \mathbf{y}_m]) \\ :=& [\hat{\mathbb{A}}] \cdot \hat{\mathbf{x}}\,(t) + [\hat{\mathbb{B}}] \cdot \mathbf{u}\,(t) + \mathbf{H}_p \cdot [\mathbf{e}_m] + \mathbf{P}^{+}[\hat{\mathbb{C}}]^{T} \cdot \mathbf{H}_s \cdot \operatorname{sign}(\mathbf{e}_m + [\Delta \mathbf{y}_m]) \\ \hat{\mathbf{y}}_m :=& [\hat{\mathbb{C}}] \cdot \hat{\mathbf{x}}\,(t) \end{aligned} \tag{2}$$

with the component-wise defined measurement error interval vector $\mathbf{e}_m(t) \in [\mathbf{e}_m] = \mathbf{y}_m - \hat{\mathbf{y}}_m + [\Delta \mathbf{y}_m]$. It accounts for deviations between measured and estimated system outputs described by bounded uncertainty with the measurement error interval $[\Delta \mathbf{y}_m]$. Equation (2) can be interpreted as the combination of a locally valid linear system model denoted by $\hat{\mathbf{f}}(\mathbf{x}\,(t)\,, [\mathbf{p}], \mathbf{u}(t))$ and a variable structure part that handles uncertainty and nonlinearities without destabilizing the error dynamics (see [5]). Moreover, uncertainty in parameters and measurements can be considered instead of their nominal values during design and implementation using interval arithmetic [2]. Therefore, the nominal system, input and output matrices become interval matrices $[\hat{\mathbb{A}}]$, $[\hat{\mathbb{B}}]$ and $[\hat{\mathbb{C}}]$ denoting the interval evaluations $\hat{\mathbf{A}}(\hat{\mathbf{x}}\,(t)\,, [\mathbf{p}]) \in [\hat{\mathbb{A}}]$, $\hat{\mathbf{B}}(\hat{\mathbf{x}}\,(t)\,, [\mathbf{p}]) \in [\hat{\mathbb{B}}]$ and $\hat{\mathbf{C}}(\hat{\mathbf{x}}\,(t)\,, [\mathbf{p}]) \in [\hat{\mathbb{C}}]$, respectively. In addition, interval specifications for control, estimation and measurement errors $[\Delta \mathbf{x}_c]$, $[\Delta \mathbf{x}_e]$ and $[\Delta \mathbf{y}_m]$ are included. Then, the switching amplitudes of the variable structure observer, and hence, chattering as well as actuator wear in closed-loop control can be reduced efficiently. Besides the usage of intervals and the consideration of process and measurement noise, states and parameters are estimated simultaneously even if they are coupled in a multiplicative way in the system model. Parameter estimation is especially useful if these values are not constant, e.g. velocity dependent friction coefficients. The task of the observer gain matrix $\mathbf{H}_p$ is to stabilize the error dynamics of the linear part in an underlying way. This matrix can be determined, for example, using pole assignment, linear matrix inequalities or by minimizing a quadratic cost function (see [7]). The matrix $\mathbf{P}$ results from solving the Lyapunov equation $\tilde{\mathbf{A}} \cdot \mathbf{P} + \mathbf{P} \cdot \tilde{\mathbf{A}}^{T} + \mathbf{Q} = \mathbf{0}$ with $\tilde{\mathbf{A}} = \hat{\mathbf{A}} - \mathbf{H}_p \cdot \hat{\mathbf{C}}$ of the linear observer part. In classical sliding mode approaches, the switching amplitude needs to be defined in advance. This often leads to unnecessarily large chattering. Here, this issue is replaced by an online evaluation of the switching amplitudes in each time step. To consider stochastic disturbances, the Itô differential operator ($V = \frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^{T}\mathbf{P}(\mathbf{x} - \hat{\mathbf{x}})$)

$$L(V(t)) = \frac{\partial V}{\partial t} + \left(\frac{\partial V}{\partial \mathbf{e}}\right)^{T} \cdot \left(\mathbf{f} - \hat{\dot{\mathbf{f}}}\right) + \frac{1}{2}\operatorname{trace}\left\{\mathbf{G}^{T}\frac{\partial^2 V}{\partial \mathbf{e}^2}\mathbf{G}\right\} \tag{3}$$

with $\mathbf{f} := \mathbf{f}(\mathbf{x}(t), [\mathbf{p}], \mathbf{u}(t))$ and $\hat{\tilde{\mathbf{f}}} := \hat{\tilde{\mathbf{f}}}(\hat{\mathbf{x}}(t), [\mathbf{p}], \mathbf{u}(t))$ is used under consideration of Eqs. (1) and (2). Moreover, the standard deviation of both process as well as measurement noise is denoted by $\mathbf{G} = [\mathbf{G}_p \quad -\mathbf{H}_p\mathbf{G}_m]$ which aims at the simulation of neglected nonlinear phenomena as well as inaccurate sensor measurements. Introducing worst-case estimation errors $[\Delta\mathbf{x}_e]$ prevents the observer from switching in regions, where the positive or negative sign of $\mathbf{e}_m(t)$ cannot be determined—usually in regions around zero—as it is described in [5]. Normally, sliding mode techniques show chattering of the sliding variable (here $\mathbf{e}_m$) around zero caused by noise, discretization errors, etc. Taking into account the additional interval, chattering can be reduced. Using an element-wise non-negative defined stability margin $\mathbf{q} \geq \mathbf{0}$, the switching amplitude can be calculated by $L(V(t)) \overset{!}{<} -\mathbf{q}^T |[\mathbf{e}_m]|$. Applying this condition to Eq. (2), the matrix $\mathbf{H}_s$ follows as a diagonal matrix from $\mathbf{H}_s = \mathrm{diag}(\mathbf{h}_s) \in \mathbb{R}^{n_y \times n_y}$ (number of measured system states n_y). Substituting all terms into (2), the condition $[\dot{V}_a] - [\mathbf{e}]^T \cdot \mathbf{P}\mathbf{P}^+ \hat{\mathbb{C}}^T \mathbf{H}_s \cdot \mathrm{sign}([\mathbf{e}_m]) + \frac{1}{2}\mathrm{trace}\left\{\mathbf{G}^T \frac{\partial^2 V}{\partial \mathbf{e}^2}\mathbf{G}\right\} < -\mathbf{q}^T |[\mathbf{e}_m]|$ with $[\mathbf{e}] = [\mathbf{x}] - [\hat{\mathbf{x}}]$, $[\mathbf{x}] = \mathbf{x} + [\Delta\mathbf{x}_c]$, $[\hat{\mathbf{x}}] = \hat{\mathbf{x}} + [\Delta\mathbf{x}_e]$, $[\mathbf{e}_m] = \mathbf{y}_m(t) - \hat{\mathbf{y}}_m(t) + [\Delta\mathbf{y}_m]$, and $[\dot{V}_a] = [\mathbf{e}]^T\mathbf{P} \cdot ([\mathbf{f}] - [\hat{\mathbf{f}}] - \mathbf{H}_p^{(k)} \cdot \mathbf{e}_m^{(k)})$ is obtained (time arguments are omitted). Then, the switching amplitudes are determined component-wise according to

$$\mathbf{h}_s = \begin{cases} \mathbf{0}, & \text{if } [\delta] \subseteq [\mathbf{e}_m]^T\,[\mathbf{e}_m] \\ \sup\left(|[\mathbf{e}_m]|^+ \cdot \left([\dot{V}_a] + \frac{1}{2} \cdot \mathrm{trace}\left\{\mathbf{G}^T \frac{\partial^2 V}{\partial \mathbf{e}^2}\mathbf{G}\right\}\right) + \mathbf{q}^T\right), & \text{else} \end{cases} \tag{4}$$

where sup denotes the upper bound of the corresponding interval. In (4), a small interval $[\delta]$ around zero is used to prevent a division by zero [7] and to reduce, both, the value of the calculated switching amplitudes and chattering. Additionally, the interval pseudo inverse $|[\mathbf{e}_m]|^+ = \left(|[\mathbf{e}_m]|^T\,|[\mathbf{e}_m]|\right)^{-1} \cdot |[\mathbf{e}_m]|^T$ is taken into account. Here, the absolute value of the difference between measured and estimated states $|[\mathbf{e}_m]|$ and the definition of the sign function are given by

$$\left|[e_{m,i}]\right| = \begin{cases} \left[-\overline{e}_{m,i}\ ;\ -\underline{e}_{m,i}\right] & \text{for } \overline{e}_{m,i} \leq 0 \\ \left[\underline{e}_{m,i}\ ;\ \overline{e}_{m,i}\right] & \text{for } \underline{e}_{m,i} \geq 0 \\ \left[0\ ;\ \max\{|\underline{e}_{m,i}|, |\overline{e}_{m,i}|\}\right] & \text{else} \end{cases}, \quad \mathrm{sign}([e_{m,i}]) = \begin{cases} 1 & \text{if } \inf([e_{m,i}]) > 0 \\ -1 & \text{if } \sup([e_{m,i}]) < 0 \\ 0 & \text{else}. \end{cases} \tag{5}$$

for each vector component i. The stability proof (only in simulation) is successful, if the evaluation of Eq. (3) is less than zero which corresponds to the negative definiteness of the time derivative of the Lyapunov function. The presented sliding mode observer considers bounded and stochastic uncertainty simultaneously. It is illustrated in the following by a cascaded structure for a simplified model for the longitudinal dynamics of a vehicle in simulation and experiment.

3 Application Scenario: Simulative and Experimental Validation

In this section, the described sliding mode strategy is validated in simulation and experiment for a simultaneous state and parameter estimation at a test rig available at the Chair of Mechatronics, University of Rostock as it is depicted in Fig. 1. This nonlinear system can be described by the state-space model

$$\mathbf{f}(\mathbf{x}(t), [\mathbf{p}], \mathbf{u}(t)) := \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ \alpha \cdot x_2(t) + \beta \cdot u(t) \end{bmatrix} \quad \text{and} \quad y(t) = x_1(t) \tag{6}$$

with the two not a-priori known parameters $\alpha = -\frac{d}{J} \in [\alpha]$ and $\beta = \frac{1}{J} \in [\beta]$ (mass moment of inertia J and velocity-proportional friction coefficient d). System (6) is nonlinear due to the multiplicative coupling between parameters and the time-varying system state x_2 (angular velocity) as well as the input u (motor torque). In general, it is assumed that the system parameters are located in the intervals $\alpha \in [\alpha] = -3 \cdot [0.5, 1.5]$ and $\beta \in [\beta] = 60 \cdot [0.5, 1.5]$. The sliding mode observer is implemented using s-functions in Matlab/Simulink with Intlab [4] and the C++ library C-XSC [3] for interval arithmetics. The considered stochastic noise, on the one hand, simulates inaccurate measured data and, on the other hand, can be interpreted as a random disturbance. Such influences play a role due to discretization errors and friction, which both cannot be quantified easily but have to be taken into account to implement later on robust control strategies that also require knowledge about system parameters [7]. For the input trajectory, Pontryagin's maximum principle is used to calculate an optimal input torque that improves the parameter estimation [6]. The cascaded structure depicted in Fig. 1 is used due to the multiplicative coupling of states and parameters in (6). Because of the assumption that system states change faster than parameters, S_1 estimates the

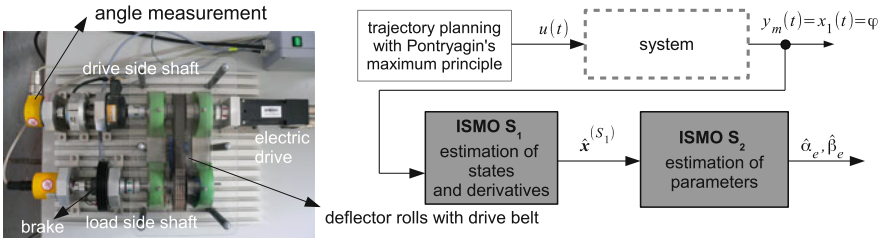


Fig. 1 Test rig and cascaded structure of the interval sliding mode observers (ISMO)

angle $x_1(t)$ and its four time derivatives using an integrator chain. Both models $\mathbf{f}^{(S_1)}$ and $\hat{\mathbf{f}}^{(S_1)}$ are given by

$$\mathbf{f}^{(S_1)} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{z}_1 \end{bmatrix} = \begin{bmatrix} [x_2] \\ [\alpha]\,[x_2] + S([\alpha], u, 0) \\ [\alpha]^2\,[x_2] + S([\alpha], u, 1) \\ [\alpha]^3\,[x_2] + S([\alpha], u, 2) \\ [\alpha]^4\,[x_2] + S([\alpha], u, 3) \end{bmatrix}, \quad \hat{\mathbf{f}}^{(S_1)} = \begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \\ \dot{\hat{x}}_4 \\ \dot{\hat{z}}_1 \end{bmatrix} = \begin{bmatrix} [\hat{x}_2] \\ [\hat{x}_3] \\ [\hat{x}_4] \\ [\hat{z}_1] \\ [\Xi] \end{bmatrix} \tag{7}$$

with $[x_2] = x_{2,d} + \left[\Delta x_{c,2}^{(S_1)}\right]$, where the desired trajectory for the second system state $x_{2,d}$ is used (the real state x_2 is not directly measured). Moreover, $[\hat{x}_j] = \hat{x}_j + \left[\Delta x_{e,j-1}^{(S_1)}\right]$ for $j = 1, \ldots, 3$, the model error $[\hat{z}_1] = \hat{z}_1 + \left[\Delta x_{e,4}^{(S_1)}\right]$, $[\Xi] = 0 + \left[\Delta x_{e,5}^{(S_1)}\right]$ and $S([\alpha], u, n) := [\beta] \cdot \sum_{i=0}^{n} [\alpha]^{n-i} \cdot u^{(i)}$, where (i) denotes the i-th derivative and the powers $n-i$ of the interval $[\alpha]$ are included. The outputs are $y_m^{(S_1)} = x_1$ and $\hat{y}_m^{(S_1)} = \hat{x}_1$. The estimates for the two system parameters $\alpha_e \in [\delta_\alpha]$ and $\beta_e \in [\delta_\beta]$ (point-values) are determined in S_2, where the estimated states (velocity and acceleration) provided by S_1 are used as virtual measurements. Analogously, subsystem S_2 is defined by

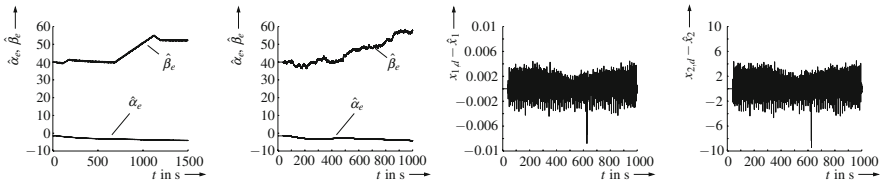
$$\mathbf{f}^{(S_2)} = \begin{bmatrix} [\alpha][x_2] + [\beta]u + [z_2] \\ [\alpha]^2[x_2] + [\alpha][\beta]u + [\beta]\dot{u} \\ 0 + \left[\Delta x_{c,3}^{(S_2)}\right] \\ 0 + \left[\Delta x_{c,4}^{(S_2)}\right] \\ 0 + \left[\Delta x_{c,5}^{(S_2)}\right] \end{bmatrix}, \quad \hat{\mathbf{f}}^{(S_2)} = \begin{bmatrix} [\delta_\alpha] \cdot [\hat{x}_2] + [\delta_\beta]u + [\hat{z}_2] \\ [\delta_\alpha]^2 \cdot [\hat{x}_2] + [\delta_\alpha][\delta_\beta]u + [\delta_\beta]\dot{u} \\ 0 + \left[\Delta x_{e,3}^{(S_2)}\right] \\ 0 + \left[\Delta x_{e,4}^{(S_2)}\right] \\ 0 + \left[\Delta x_{e,5}^{(S_2)}\right] \end{bmatrix} \tag{8}$$

with $\mathbf{f}^{(S_2)} = \left[\dot{x}_2\ \dot{x}_3\ \dot{\alpha}_e\ \dot{\beta}_e\ \dot{z}_2\right]^T$, $\hat{\mathbf{f}}^{(S_2)} = \left[\dot{\hat{x}}_2\ \dot{\hat{x}}_3\ \dot{\hat{\alpha}}_e\ \dot{\hat{\beta}}_e\ \dot{\hat{z}}_2\right]^T$. The interval definitions are for the velocity $[x_2] = \left(x_2 + \left[\Delta x_{c,1}^{(S_2)}\right]\right)$, for the model error $[z_2] = \left(z_2 + \left[\Delta x_{c,5}^{(S_2)}\right]\right)$, for the estimated velocity $[\hat{x}_2] = (\hat{x}_2 + [\Delta x_{e,1}^{(S_2)}])$, for the estimated model error $[\hat{z}_2] = \left(\hat{z}_2 + \left[\Delta x_{e,5}^{(S_2)}\right]\right)$ (assumed to be slowly varying). The enlarged intervals for the estimated parameters are $[\delta_\alpha] = \left([\alpha] + \left[\Delta x_{e,3}^{(S_2)}\right]\right)$ as well as $[\delta_\beta] = \left([\beta] + \left[\Delta x_{e,4}^{(S_2)}\right]\right)$. The output vectors of S_2 are defined as $\mathbf{y}_m^{(S_2)} = [x_2, x_3, z_2]^T$ and

$\hat{\mathbf{y}}_m^{(S_2)} = [\hat{x}_2, \hat{x}_3, \hat{z}_2]^T$. Then, the ISMOs for both subsystems $k = \{S_1, S_2\}$ are given by

$$\dot{\hat{\mathbf{x}}}^{(k)} = \mathbf{A}_O^{(k)} \cdot \mathbf{x}^{(k)} + \mathbf{b}_O^{(k)} \cdot u + \mathbf{H}_p^{(k)} \cdot \mathbf{e}_m + (\mathbf{P}^{(k)})^+ (\mathbf{C}^{(k)})^T \mathbf{H}_s^{(k)} \cdot \text{sign}([\mathbf{e}_m^{(k)}]) \,, \tag{9}$$

with $\mathbf{A}_O^{(S_1)} = \begin{bmatrix} 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&1 \\ 0&0&0&0&0 \end{bmatrix}$, $\mathbf{x}^{(S_1)} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ z_1 \end{bmatrix}$, $\mathbf{A}_O^{(S_2)} = \begin{bmatrix} 0 & 0 & \hat{x}_{2,w} & 0 & 1 \\ \alpha_{e,w}^2 & 0 & 0 & \dot{u}_w & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$, $\mathbf{x}^{(S_2)} = \begin{bmatrix} x_2 \\ x_3 \\ \alpha_e \\ \beta_e \\ z_2 \end{bmatrix}$ and the defined working points $\hat{x}_{2,w}$, $\alpha_{e,w}$, $\dot{u}_w$. Note, $\mathbf{H}_p^{(S_1)}$ becomes a vector and the switching term $\mathbf{H}_s^{(S_1)} \cdot \text{sign}([e_m^{(S_1)}]) = \mathbf{H}_s^{(S_1)} \cdot \text{sign}([e_m^{(S_1)}])$ in Eq. (9) is scalar. The input vectors for S_1 and S_2 are given by $\mathbf{b}_O^{(S_1)} = [\mathbf{0}^{5x1}]$ and $\mathbf{b}_O^{(S_2)} = \left[\beta_e, \alpha_e \cdot \beta_e, 0, 0, 0\right]^T$. The matrix $\mathbf{P}^{(S_2)}$ can be determined as for subsystem S_1 and stabilizes together with $\mathbf{H}_p^{(S_2)}$ the linear observer part where a working set point is chosen for the matrix $\mathbf{A}_O^{(S_2)}$. Here, the switching term $\mathbf{H}_s^{(S_2)} \cdot \text{sign}(\mathbf{e}_m^{(S2)} + [\Delta \mathbf{y}_m^{(S_2)}])$ is a vector with 3 components. Note, the switching amplitude matrices for both subsystems are calculated separately according to Sect. 2. Simulative and experimental results are depicted in Fig. 2, where one driving cycle takes 8 s. that is repeated periodically. It can be seen, that both, simulation and experiment, are able to estimate the system parameters with good accuracy. The experimental results using the cascaded ISMO structure, where the adaptation of parameters and states takes place at each discretization step, are much better than the results of a least-squares parameter identification, where the estimates are constant for time intervals of 8 s. This can be seen from Table 1, where the root mean square errors are compared. Therefore, system (6) was evaluated separately including the estimated parameters of both, the ISMO structure and the least squares estimation.



(a) Estimates $\hat{\alpha}_e$, $\hat{\beta}_e$. (b) Estimates $\hat{\alpha}_e$, $\hat{\beta}_e$. (c) Estimation error x_1. (d) Estimation error x_2.

Fig. 2 Simulative (**a**) and experimental results ((**b**)–(**d**)), measurement noise of $y_m^{(S_1)}$: $G_m^{(S_1)} = 0.005$

Table 1 Results: standard deviations of the estimation errors compared to the simulated measurement noise (left) and comparison to least-squares (LS) parameter identification (right)

Measurement noise	$x_{1,d} - \hat{x}_1$	$x_{2,d} - \hat{x}_2$
0.005	0.0080	0.2362

	LS	ISMO	Improvement
x_1	$\Delta_{x_1,LS} = 2730$	$\Delta_{x_1,ISMO} = 261.36$	90.43 %
x_2	$\Delta_{x_2,LS} = 4.79$	$\Delta_{x_2,ISMO} = 4.51$	5.85 %

4 Conclusions and Outlook

In this paper, an interval sliding mode observer considering bounded and stochastic disturbances for a simultaneous state and parameter estimation has been described and validated in experiment. In future work, the ISMO will be applied to other test rigs with an equivalent interval sliding mode control.

References

1. Bartoszewicz, A., Nowacka-Leverton, A.: Time-Varying Sliding Modes for Second and Third Order Systems. Lecture Notes in Control and Information Sciences, vol. 382. Springer, Berlin (2009)
2. Jaulin, L., Kieffer, M., Didrit, O., Walter, É.: Applied Interval Analysis. Springer, London (2001)
3. Krämer, W.: XSC Languages (C-XSC, PASCAL-XSC) — Scientific Computing with Validation, Arithmetic Requirements, Hardware Solution and Language Support (n.a.) (2014). www.math.uni-wuppertal.de/ xsc/
4. Rump, S.M.: Interval computations with IntLab. Braz. Electron. J. Math. Comput. **1**, 818–823 (1999)
5. Senkel, L., Rauh, A., Aschemann, H.: Interval-based sliding mode observer design for nonlinear systems with bounded measurement and parameter uncertainty. In: Proceedings of IEEE International Conference on Methods and Models in Automation and Robotics. Miedzyzdroje, Poland (2013)
6. Senkel, L., Rauh, A., Aschemann, H.: Optimal input design for online state and parameter estimation using interval sliding mode observers. In: Proceedings of 52nd IEEE Conference on Decision and Control, CDC 2013, Florence, Italy (2013)
7. Senkel, L., Rauh, A., Aschemann, H.: Robust sliding mode techniques for control and state estimation of dynamic systems with bounded and stochastic uncertainty. In: Proceedings of Second International Conference on Vulnerability and Risk Analysis and Management, Liverpool (2014)

Interval-Based Sliding Mode Control for High-Temperature Fuel Cells Under Actuator Constraints

Andreas Rauh, Luise Senkel, and Harald Aschemann

Abstract Interval-based sliding mode controllers can be used efficiently for a robust stabilization of systems with bounded uncertainty. The real-time implementation of these procedures makes use of software libraries that provide functionalities for interval analysis and algorithmic differentiation. This paper gives an overview of possible extensions of such control procedures for the reliable stabilization of the thermal behavior of high-temperature solid oxide fuel cell systems. During the real-time stabilization, limitations of the range of state and control variables are treated by constraints implemented in a barrier Lyapunov function approach.

Keywords Interval-based sliding mode control • Lyapunov function • Real-time stabilization • Uncertainty

1 Introduction

In previous work, different approaches have been derived by the authors for the reliable control of the thermal behavior of high-temperature solid oxide fuel cell stacks (SOFC stacks) under consideration of uncertain parameters and a-priori unknown load variations. These approaches comprise model-predictive, feedback linearizing, and sliding mode techniques. It was shown that both predictive and sliding mode techniques can be applied successfully to systems with interval parameters [4].

In contrast to the predictive control approach, state and actuator constraints cannot be handled directly within a classical sliding mode design for this type of application. However, the sliding mode-type control design provides an inherent stability proof, which is not directly available for the predictive technique. For these

A. Rauh (✉) • L. Senkel • H. Aschemann
Chair of Mechatronics, University of Rostock, 18059 Rostock, Germany
e-mail: Andreas.Rauh@uni-rostock.de; Luise.Senkel@uni-rostock.de;
Harald.Aschemann@uni-rostock.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_92

reasons, this contribution removes the before-mentioned drawback concerning the handling of constraints by a reformulation of the stability conditions resulting from the variable structure control design. This reformulation is based on a stability-preserving gain adaptation. It becomes active as soon as saturation limits are reached for the control inputs. For that purpose, actuator limits are treated as hard constraints that must not be exceeded, whereas state limitations are described by soft constraints. These are introduced as penalty terms inspired by the concept of barrier Lyapunov functions [6]. While handling the actuator constraints, it becomes possible to adapt the gain value of the variable-structure control part as well as the parameters characterizing the sliding surface during system operation [5]. These adaptations directly lead to a modification of the system dynamics during the transient reaching phase. This phase describes the system dynamics as long as the operating conditions do not fully comply with those system states for which the sliding condition is fulfilled.

The prerequisite for the design of the before-mentioned control strategies is the modeling of the thermal behavior of the SOFC stack. In previous work, a spatial semi-discretization of the temperature distribution in the interior of the SOFC stack has been performed. It is based on the assumption that temperature variations can be described by a set of ordinary differential equations (ODEs), where the temperature is piecewise homogeneous in each finite volume element. This model is coupled with the lag dynamics of the gas preheaters (providing preheated air (CG) to the cathode and a mixture of hydrogen (H_2), nitrogen (N_2) and water vapor (H_2O) to the anode) according to Fig. 1. Here, the time constants of the preheaters have to be accounted for during the control design to avoid undesired chattering [4].

According to [3, 4], a set of ODEs is obtained in the input-affine form

$$\dot{\mathbf{x}}(t) = \mathbf{f}\left(\mathbf{x}(t), \mathbf{p}, \mathbf{v}_\mathrm{d}(t)\right) = \boldsymbol{\psi}_1\left(\mathbf{x}(t), \mathbf{p}\right) + \boldsymbol{\Psi}_2\left(\mathbf{x}(t), \mathbf{p}\right) \cdot \mathbf{v}_\mathrm{d}\left(\mathbf{u}(t)\right) \tag{1}$$

with the input vector $\mathbf{v}_\mathrm{d}(t) = \left[v_{\mathrm{H_2,d}}(t)\ v_{\mathrm{N_2,d}}(t)\ v_{\mathrm{H_2O,d}}(t)\ v_{\mathrm{CG,d}}(t)\right]^T$ of the anode and cathode gas. This vector consists of products of the corresponding gas mass flows $\dot{m}_\chi(t) = \dot{m}_{\mathrm{ffl,d}}(t) = \dot{m}_{\mathrm{ffl,in}}(t)$ for each $\chi \in \{\mathrm{H_2}, \mathrm{N_2}, \mathrm{H_2O}, \mathrm{CG}\}$ and the

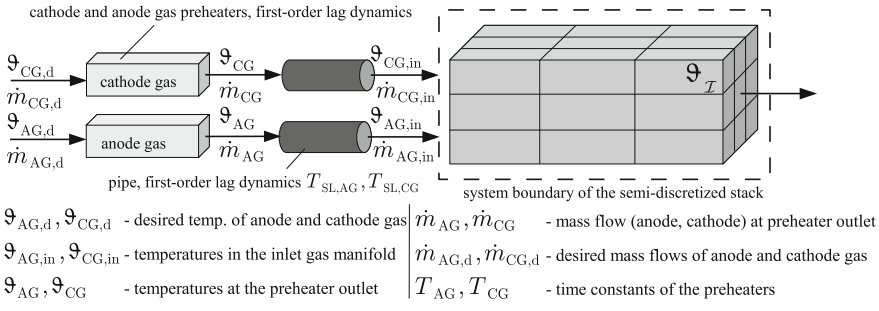


Fig. 1 Semi-discretization of the fuel cell stack module with gas preheaters

desired preheater temperatures $\vartheta_{\mathrm{CG,d}}(t)$ as well as $\vartheta_{\mathrm{AG,d}}(t) = \vartheta_{\mathrm{H_2,d}}(t) = \vartheta_{\mathrm{N_2,d}}(t) = \vartheta_{\mathrm{H_2O,d}}(t)$.

If it is assumed that the time constants for changes in the gas mass flows are significantly smaller than those for the thermal behavior, the relation $\frac{d\dot{m}_{\chi}(t)}{dt} = \frac{d\dot{m}_{\mathrm{fll,d}}(t)}{dt} = \frac{d\dot{m}_{\mathrm{fll,in}}(t)}{dt} = 0$ can be substituted into Eq. (1). Assuming additionally that the anode gas properties are predefined by a subsidiary controller, the ODEs (1) turn into

$$\dot{\mathbf{x}}(t) = \underbrace{\boldsymbol{\phi}_1(\mathbf{x}(t), \mathbf{p}) + \boldsymbol{\Phi}_{2,\mathrm{AG}}(\mathbf{x}(t), \mathbf{p}) \cdot \begin{bmatrix} v_{\mathrm{H_2,d}}(t) \\ v_{\mathrm{N_2,d}}(t) \\ v_{\mathrm{H_2O,d}}(t) \end{bmatrix}}_{=: \mathbf{f}_1(\mathbf{x}(t), \mathbf{p})} + \underbrace{\boldsymbol{\phi}_{2,\mathrm{CG}}(\mathbf{x}(t), \mathbf{p}) \cdot v_{\mathrm{CG,d}}(t)}_{=: \mathbf{f}_2(\mathbf{x}(t), \mathbf{p}) \cdot v_{\mathrm{CG,d}}(t)} \tag{2}$$

with $\boldsymbol{\psi}_1(\mathbf{x}(t), \mathbf{p}) = \boldsymbol{\phi}_1(\mathbf{x}(t), \mathbf{p})$ and the state-dependent term $\mathbf{f}_1(\mathbf{x}(t), \mathbf{p})$. Moreover, the system input is characterized by $\mathbf{f}_2(\mathbf{x}(t), \mathbf{p}) = \left[\mathbf{0}_{1\times 9}\ \frac{1}{T_{\mathrm{CG}}}\ 0\ 0\ \mathbf{0}_{1\times n_x}\right]^T$, where $\mathbf{0}_{i\times j}$ is a zero matrix of dimension $i \times j$ (n_x: number of finite volume elements in the SOFC stack). Due to the above-mentioned simplifying assumptions, the equality $\frac{\partial \mathbf{f}_1(\mathbf{x}(t),\mathbf{p})}{\partial v_{\mathrm{CG,d}}} = \mathbf{0}$ holds for all operating points. Choosing $v_{\mathrm{CG,d}}(t)$ as the primary input justifies the use of $\frac{d}{dt}\left[v_{\mathrm{H_2,d}}(t),\ v_{\mathrm{N_2,d}}(t),\ v_{\mathrm{H_2O,d}}(t)\right]^T \approx \mathbf{0}$ during control synthesis. Errors introduced by this simplification can be taken into consideration by an additive disturbance variable in the following state-space transformation. This transformation replaces the ODEs (2) by a nonlinear controller canonical form. According to [4], it is reasonable to perform the transformation by computing the required Lie derivatives in the following section by means of algorithmic differentiation [1]. This significantly reduces the complexity that results from the excessive lengths of the symbolic expressions for the state transformation that arise if larger dimensions (typically $n_x \geq 3$ volume elements) are used for the semi-discretization approach.

2 Interval-Based Sliding Mode Control Design

Assume that the system output is defined by $y(t) = h(\mathbf{x}(t)) = \vartheta_{\mathcal{I}^*}$, $\mathbf{x}(t) \in \mathbb{R}^{\mathcal{N}}$, as the stack temperature in the segment $\mathcal{I}^*$. This segment temperature should be controlled by means of $v_{\mathrm{CG,d}}(t)$. Then, a successive computation of the Lie derivatives

$$\frac{d^r y(t)}{dt^r} = y^{(r)}(t) = L_{\mathbf{f}}^r h(\mathbf{x}(t)) = L_{\mathbf{f}}\left(L_{\mathbf{f}}^{r-1} h(\mathbf{x}(t))\right)\ ,\quad r = 1, \ldots, \delta \tag{3}$$

is performed up to the system's relative degree δ, which is defined as

$$\frac{\partial L_{\mathbf{f}}^{r} h(\mathbf{x}(t))}{\partial v_{\mathrm{CG,d}}} \equiv 0 \;\text{ for all }\; r = 0, \ldots, \delta - 1 \;\text{ with }\; \frac{\partial L_{\mathbf{f}}^{\delta} h(\mathbf{x}(t))}{\partial v_{\mathrm{CG,d}}} \neq 0 \;. \tag{4}$$

Using the new state vector $\boldsymbol{\xi}(t) = \left[h(\mathbf{x}(t)),\, L_{\mathbf{f}}h(\mathbf{x}(t)),\, \ldots,\, L_{\mathbf{f}}^{\delta-1}h(\mathbf{x}(t))\right]^T \in \mathbb{R}^{\delta}$ with $\xi_1(t) = y(t) = h(\mathbf{x}(t)) =: h(\mathbf{x})$, the ODEs (2) can be transformed into

$$\begin{aligned}\left[\,\dot{\boldsymbol{\xi}}^T(t) \,\vdots\, \dot{\boldsymbol{\zeta}}^T(t)\,\right]^T &= \left[\,L_{\mathbf{f}}h(\mathbf{x}),\, \ldots,\, L_{\mathbf{f}}^{\delta-1}h(\mathbf{x}),\, L_{\mathbf{f}}^{\delta}h(\mathbf{x}) \,\vdots\, L_{\mathbf{f}}^{\delta+1}h(\mathbf{x}),\, \ldots,\, L_{\mathbf{f}}^{\mathcal{N}}h(\mathbf{x})\,\right]^T \\ &= \left[\,\xi_2(t),\, \ldots,\, \xi_{\delta}(t),\, \tilde{a}(\mathbf{x}(t), \mathbf{p}, d(t)) \,\vdots\, \mathbf{a}^{\diamondsuit}(\mathbf{x}(t), \mathbf{p}, d(t))\,\right]^T \\ &\quad + \left[\,0,\, \ldots,\, 0,\, \tilde{b}(\mathbf{x}(t), \mathbf{p}) \cdot v_{\mathrm{CG,d}}(t) \,\vdots\, \mathbf{b}^{\diamondsuit}(\mathbf{x}(t), \mathbf{p}, d(t), v_{\mathrm{CG,d}}(t), \dot{v}_{\mathrm{CG,d}}(t), \ldots)\,\right]^T\end{aligned} \tag{5}$$

with the interval parameters $\mathbf{p} \in [\mathbf{p}]$ and the additive disturbance $d(t) \in [d] = \left[\underline{d}\,;\,\overline{d}\right]$ in $\tilde{a}(\mathbf{x}(t), \mathbf{p}, d(t)) := L_{\mathbf{f}}^{\delta}h(\mathbf{x}(t)) - \tilde{b}(\mathbf{x}(t), \mathbf{p}) \cdot v_{\mathrm{CG,d}} + d(t)$. As shown in [4], the original and transformed states $\mathbf{x}(t)$ as well as $\boldsymbol{\xi}(t)$ and $\boldsymbol{\zeta}(t)$, respectively, can be estimated by different model-based techniques as well as by algebraic or low-pass filtered derivative approximations.

During the design of the sliding mode control strategy, the input $v_{\mathrm{CG,d}}(t)$ is determined in such a way that asymptotic stability of the closed-loop dynamics is guaranteed despite interval parameters $[\mathbf{p}]$ and $[d]$. This requires that the tracking errors of all components of $\boldsymbol{\xi}(t)$ converge to zero with certainty. The corresponding error signals are given by $\tilde{\xi}_1^{(r)}(t) = \xi_1^{(r)}(t) - \xi_{1,\mathrm{d}}^{(r)}(t)$ with $r = 0, \ldots, \delta - 1$ and $\vartheta_{\mathcal{I}^*}(t) =: \xi_1(t) = \xi_1^{(0)}(t)$. These error signals are summarized in the vector

$$\tilde{\boldsymbol{\xi}}(t) = \left[\xi_1(t) - \xi_{1,\mathrm{d}}(t),\;\; \xi_1^{(1)}(t) - \xi_{1,\mathrm{d}}^{(1)}(t),\;\; \ldots,\;\; \xi_1^{(\delta-1)}(t) - \xi_{1,\mathrm{d}}^{(\delta-1)}(t)\right]^T \in \mathbb{R}^{\delta}, \tag{6}$$

where the desired trajectory is denoted by $\xi_{1,\mathrm{d}}(t)$ with the time derivatives $\xi_{1,\mathrm{d}}^{(r)}(t)$. According to [3, 4], perfect trajectory tracking corresponds to states that are located on the sliding surface

$$s := s\left(\tilde{\boldsymbol{\xi}}(t)\right) = \tilde{\xi}_1^{(\delta-1)}(t) + \alpha_{\delta-2} \cdot \tilde{\xi}_1^{(\delta-2)}(t) + \ldots + \alpha_0 \cdot \tilde{\xi}_1^{(0)}(t) = 0 \;. \tag{7}$$

To guarantee asymptotic stability for $s = 0$, the parameters $\alpha_0, \ldots, \alpha_{\delta-2}$ have to be chosen as coefficients of a Hurwitz polynomial of the order $\delta - 1$. Furthermore, stabilization of $\boldsymbol{\xi}(t)$ towards the sliding surface is required if $s \neq 0$ holds. In this case, called reaching phase, a variable structure control is employed. It can be derived from the Lyapunov function candidate $V = \frac{1}{2}s^2 > 0$ with $\dot{V} = s \cdot \dot{s} < 0$ for $s \neq 0$. Replacing $\dot{V}$ by a linear convergence rate in s leads to $\dot{V} = s \cdot \dot{s} \leq -\eta \cdot |s|$, $\eta > 0$.

Then, the stability requirement can be satisfied by the interval-based control law (time arguments are omitted for brevity)

$$[v_{\mathrm{CG,d}}] := \frac{-\tilde{a}\left(\mathbf{x}, [\mathbf{p}], [d]\right) + \xi_{1,\mathrm{d}}^{(\delta)} - \alpha_{\delta-2} \cdot \tilde{\xi}_1^{(\delta-1)} - \ldots - \alpha_0 \tilde{\xi}_1^{(1)} - \tilde{\eta} \cdot \mathrm{sign}\{s\}}{\tilde{b}\left(\mathbf{x}, [\mathbf{p}]\right)} \tag{8}$$

with $\tilde{\eta} > \eta > 0$ and the interval evaluation of $\tilde{a}\left(\mathbf{x}(t), \mathbf{p}, d(t)\right)$ and $\tilde{b}\left(\mathbf{x}(t), \mathbf{p}\right)$ for all $\mathbf{p} \in [\mathbf{p}]$ and $d(t) \in [d]$. In a real-time environment, the expression (8) is evaluated by means of C-XSC [2] with $\tilde{\xi}^{(r)}(t) \in \left[\tilde{\xi}^{(r)}\right](t)$. According to [3], point values $\hat{d}(t)$ for the term $d(t)$ can be estimated by a suitable observer. To express the level of confidence in these values, they are inflated to the interval $[d] := \hat{d}(t) + \Delta d \cdot [-1\ ;\ 1]$ with $\Delta d > 0$. To specify the final control law in such a manner that it can be applied regardless of the sign of $\tilde{b}(\mathbf{x}(t), \mathbf{p})$, $0 \notin \tilde{b}(\mathbf{x}(t), [\mathbf{p}])$, a stabilizing point-valued control $v_{\mathrm{CG,d}}(t)$ is chosen by checking the sign of $\dot{V}$ for all candidates from the set

$$\mathcal{V}_{\mathrm{CG,d}} := \left\{\underline{v}_{\mathrm{CG,d}}(t) - \epsilon, \underline{v}_{\mathrm{CG,d}}(t) + \epsilon, \overline{v}_{\mathrm{CG,d}}(t) - \epsilon, \overline{v}_{\mathrm{CG,d}}(t) + \epsilon\right\} \tag{9}$$

with the control infimum $\underline{v}_{\mathrm{CG,d}}(t) := \inf\{[v_{\mathrm{CG,d}}(t)]\}$, the supremum $\overline{v}_{\mathrm{CG,d}}(t) := \sup\{[v_{\mathrm{CG,d}}(t)]\}$ and some small $\epsilon > 0$.

For the control of the SOFC towards a fixed maximum stack temperature, the output temperature $y(t) = \vartheta_{\mathcal{I}^*}(t)$ is determined in each time step according to $\mathcal{I}^* = \arg\max_{\mathcal{I}}\{\vartheta_{\mathcal{I}}(t)\}$, where $\vartheta_{\mathcal{I}}(t)$ denotes the temperatures in all stack segments.

3 One-Sided Barrier Lyapunov Function Constraints, Offline Trajectory Planning and Online Gain Adaptation

To prevent large overshoots of the output $y(t) = \vartheta_{\mathcal{I}^*}(t)$ over the desired temperature $\theta_{\max}$, the safety margin $\Delta\theta_{\max} > 0$ is introduced with $\bar{\theta}_{\max} := \theta_{\max} + \Delta\theta_{\max}$. Using the value $\bar{\theta}_{\max}$, a one-sided barrier Lyapunov function candidate

$$\tilde{V} = V + \rho_{\mathrm{V}} \cdot \sum_{i \in \{\mathcal{I}\}} \ln\left(\frac{\bar{\theta}_{\max}}{\bar{\theta}_{\max} - \vartheta_i}\right) > 0 \ \text{ for } \ s \neq 0 \ \text{ with } \ V = \frac{1}{2}s^2\ ,\ \ \rho_{\mathrm{V}} > 0\ , \tag{10}$$

can be defined, which serves two purposes. On the one hand, the term V stabilizes the dynamics of the uncertain SOFC model. On the other hand, the additive second term introduces a soft constraint $\vartheta_{\mathcal{I}} \overset{!}{\leq} \theta_{\max}$ by means of a strict barrier $\vartheta_{\mathcal{I}} < \bar{\theta}_{\max}$. Here, the factor ρ_{V} has to be chosen so that control constraints are not violated and that the term V has dominating influence in a neighborhood of the desired trajectory,

corresponding to $s = 0$. Computing the time derivative $\dot{\tilde{V}} = \dot{V} + \rho_{\mathrm{V}} \cdot \sum_{i \in \{\mathcal{I}\}} \left(\frac{\dot{\vartheta}_i}{\bar{\theta}_{\max} - \vartheta_i} \right)$ of the extended Lyapunov function candidate leads to the modified interval control

$$[\tilde{v}_{\mathrm{CG,d}}] := [v_{\mathrm{CG,d}}] - \frac{s}{s^2 + \tilde{\epsilon}} \cdot \rho_{\mathrm{V}} \cdot \frac{1}{\tilde{b}\,(\mathbf{x}, [\mathbf{p}])} \cdot \sum_{i \in \{\mathcal{I}\}} \left(\frac{\dot{\vartheta}_i}{\bar{\theta}_{\max} - \vartheta_i} \right) \ , \ \tilde{\epsilon} > 0 \ , \qquad (11)$$

where the term $\frac{s}{s^2+\tilde{\epsilon}}$ is an approximation of the expression $\frac{1}{s}$ for $s \neq 0$. This latter modification guarantees that violations of the constraint $\vartheta_{\mathcal{I}} \overset{!}{\leq} \theta_{\max}$ are penalized by the modified control law, while the adaptation becomes inactive for $s = 0$ and simultaneously avoids singularities at $s = 0$. Point-valued control signals are again chosen as in (9). Note that the denominator term $\tilde{b}\,(\mathbf{x}, [\mathbf{p}])$ in (11) remains unchanged by the introduction of the logarithmic barrier function if the derivatives $\dot{\vartheta}_{\mathcal{I}}$ do not explicitly depend on the system input $v_{\mathrm{CG,d}}$. This is true as long as the preheater dynamics are included in the ODEs (2). If their dynamics were neglected, the denominator term had to be adjusted by solving the inequality $\dot{\tilde{V}} \leq -\eta \cdot |s|$ in a rigorous way.

To make sure that the guaranteed stabilizing control strategy can be evaluated on a real test rig, it is important to account for hard actuator constraints. For that purpose, the control variable $v_{\mathrm{CG,d}}$ (or $\tilde{v}_{\mathrm{CG,d}}$, resp.) is decomposed into a product of admissible gas mass flows and desired preheater temperatures as described in [3]. This decomposition relies on an online optimization procedure, which penalizes high-frequent variations of both system inputs as well as deviations from their desired set-points. To make this optimization routine applicable, it is necessary that the system input is compliant with the actual actuator constraints of the system with both nominal (i.e., point-valued) and uncertain (i.e., interval) parameters.

First of all, a nominal output $\xi_{1,\mathrm{d}}(t)$ is determined for a fixed output segment (as a time-dependent polynomial) with a predefined composition and temperature of the anode gas mixture. This trajectory (or resp. steady-state operating point) has to be selected in such a manner that it does not exceed any of the input saturations.

In a second stage, the interval control signals $[v_{\mathrm{CG,d}}]$ and $[\tilde{v}_{\mathrm{CG,d}}]$ are split up into continuous and variable structure parts according to

$$\begin{aligned} [v_{\mathrm{CG,d}}] &= [v_{\mathrm{CG,d,I}}] + \tilde{\eta} \cdot [v_{\mathrm{CG,d,II}}] \subseteq [v_{\mathrm{CG,max}}] \quad \text{and} \\ [\tilde{v}_{\mathrm{CG,d}}] &= [\tilde{v}_{\mathrm{CG,d,I}}] + \tilde{\eta} \cdot [\tilde{v}_{\mathrm{CG,d,II}}] \subseteq [v_{\mathrm{CG,max}}] \ , \end{aligned} \qquad (12)$$

where the choice between $[v_{\mathrm{CG,d}}]$ and $[\tilde{v}_{\mathrm{CG,d}}]$ is made depending on whether state constraints should be handled or not. Here, a suitable set of asymptotically stable eigenvalues is selected for the dynamics on the sliding surface. With these eigenvalues and intervals $\tilde{\xi}^{(r)}(t) \in \left[\tilde{\xi}^{(r)}\right](t)$ for the operating range, the intervals $[v_{\mathrm{CG,d,I}}]$ and $[\tilde{v}_{\mathrm{CG,d,I}}]$ are evaluated offline such that both are true subsets $[v_{\mathrm{CG,d,I}}] \subset [v_{\mathrm{CG,max}}]$ and $[\tilde{v}_{\mathrm{CG,d,I}}] \subset [v_{\mathrm{CG,max}}]$ of the maximum possible input range $[v_{\mathrm{CG,max}}]$.

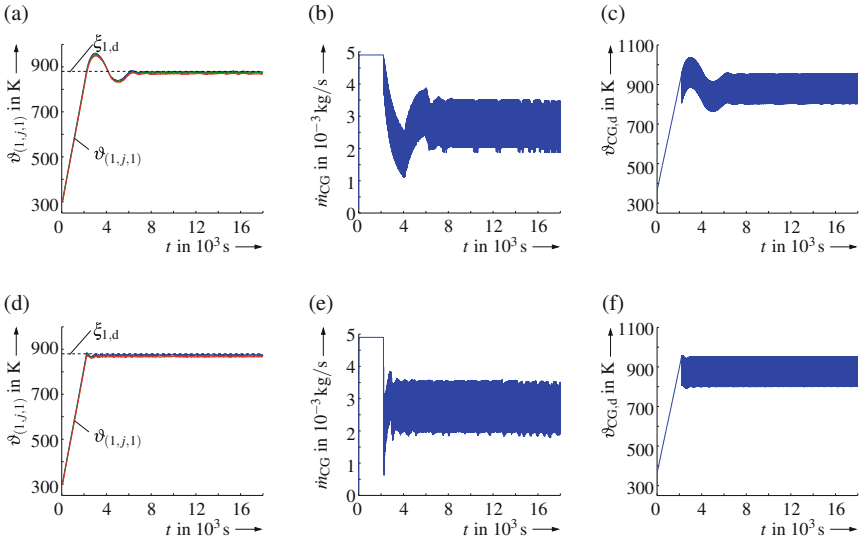


Fig. 2 Interval-based sliding mode control without (**a**–**c**) and with barrier Lyapunov function (**d**–**f**). (**a**) Stack module temperatures. (**b**) Mass flow (cathode) $\dot{m}_{CG}$. (**c**) Desired temperature $\vartheta_{CG,d}$. (**d**) Stack module temperatures. (**e**) Mass flow (cathode) $\dot{m}_{CG}$. (**f**) Desired temperature $\vartheta_{CG,d}$

Finally, the parameter $\tilde{\eta}$ is chosen according to the desired dynamics during the reaching phase. This value is kept as long as the input is compliant with the actuator constraints. Otherwise, the positive value $\tilde{\eta}$ is adapted so that the point-valued control lies (in the interior or) on the boundary of the possible input range [5].

Figure 2 contains a comparison of the simulation results of the dynamic system without and with the extension by the one-sided barrier Lyapunov function. The interval evaluation of the control law was performed for the same parameters that were used for the unconstrained case in [4]. It can be shown numerically that using saturation values for the input until $t \approx 2200$ s leads to a guaranteed stabilization of the system towards $\theta_{max} = \vartheta_d = 880\,\mathrm{K} = \text{const}$. After this point of time, the use of the extended control procedure helps to reduce the overshoot over this set-point and hence improves both the efficiency and practical usability of the control strategy.

4 Conclusions and Outlook on Future Work

In this paper, a robust sliding mode procedure has been presented and extended by a one-sided barrier Lyapunov function approach to prevent the violation of state constraints. The implementation of this control procedure in rapid control prototyping environments makes use of interval analysis. In such a way, it can be guaranteed that the considered dynamic system is robustly stabilized despite

bounded uncertainty in parameters as well as in measured and estimated state variables.

Future work will extend the presented gain scheduling approach towards a real-time modification of the parameters α_i of the sliding surface. In addition to the presented control parameterization, this provides a further degree of freedom in the choice of the variable structure gain if the sum of the control parts (I) and (II) in (12) is not yet equal to one of the saturation limits. Besides the requirement that input constraints must not be violated, criteria for robustness against measurement noise and non-modeled, but bounded, disturbances will be taken into consideration.

References

1. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. SIAM, Philadelphia (2008)
2. Krämer, W.: XSC languages (C-XSC, PASCAL-XSC) — scientific computing with validation, arithmetic requirements, hardware solution and language support (2012). http://www.math.uni-wuppertal.de/~xsc/. C-XSC 2.5.3
3. Rauh, A., Senkel, L., Kersten, J., Aschemann, H.: Verified stability analysis for interval-based sliding mode and predictive control procedures with applications to high-temperature fuel cell systems. In: Proceedings of 9th IFAC Symposium on Nonlinear Control Systems, Toulouse (2013)
4. Rauh, A., Senkel, L., Kersten, J., Aschemann, H.: Reliable control of high-temperature fuel cell systems using interval-based sliding mode techniques. IMA J. Math. Control. Inf. **33**(2), 457–484 (2014).
5. Rauh, A., Senkel, L., Aschemann, H.: Interval-based sliding mode control design for solid oxide fuel cells with state and actuator constraints. IEEE Trans. Ind. Electron. **62**(8), 5208–5217 (2015)
6. Tee, K.P., Ge, S.S., Tay, E.H.: Barrier Lyapunov functions for the control of output-constrained nonlinear systems. Automatica **45**(4), 918–927 (2009)

Sliding Mode Data Flow Regulation for Connection-Oriented Networks with Unpredictable Packet Loss Ratio

Piotr Lesniewski and Andrzej Bartoszewicz

Abstract In this paper we propose a discrete time sliding mode congestion controller for a single virtual circuit in connection-oriented communication networks. The circuit is characterized by the non-negligible propagation delay, the maximum link capacity and unknown, time-varying data loss rate. The proposed controller generates non-negative and limited transmission rates, ensures upper bounded queue length in the bottleneck link buffer and may guarantee full utilization of the link capacity. In order to ensure fast reaction to the unpredictable data loss and unknown changes of the available bandwidth, the controller employs the dead-beat sliding hyperplane. However, straightforward application of the dead-beat paradigm could lead to unacceptably big transmission rates. Therefore, the controller is designed using the concept of the reaching law, which helps to attenuate the excessive magnitude of control signal at the beginning of the transmission process.

Keywords Connection-oriented communication network • Dead-beat • Sliding mode control and estimation • Virtual circuit

1 Introduction

Congestion control in connection-oriented data transmission networks is an important and up to date research topic [1, 2, 4–7]. The difficulty of the congestion control is caused by long propagation delays, rapidly changing bandwidth and unpredictable packet losses. When congestion of a specific link is detected, an appropriate communique must be sent to all the sources transmitting data through this link. Delivery of this communique involves feedback propagation delays. Then data sources adjust their flow rates in order to counteract the congestion, however, the adjusted rates begin to affect the congested link after the forward propagation delay. Therefore, in the modern data transmission networks characterized by significant bandwidth and delays the need for the proper flow regulation cannot be ignored.

P. Lesniewski (✉) • A. Bartoszewicz
Institute of Automatic Control, Technical University of Lodz, 18/22 Bohdana Stefanowskiego St., 90-924 Lodz, Poland
e-mail: piotr.lesniewski2@gmail.com; andrzej.bartoszewicz@p.lodz.pl

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_93

In this work, we design a discrete time sliding mode congestion controller for connection-oriented networks. In the design process not only we take into account propagation delays and inevitable bandwidth changes, but also we explicitly consider unpredictable data losses. Therefore, we propose a controller which ensures robustness of the closed loop system with respect to the a priori unknown and time-varying packet losses.

2 Network Model

We consider a virtual circuit of a connection-oriented network, that consists of a data source, some intermediate nodes and a destination. We assume that there is a single bottleneck in the network. A congestion controller is placed at the bottleneck node, and it generates a signal (denoted by u) that determines the transmission rate of the source. The source receives this signal after the backward delay T_B, and sends the requested amount of data, which is passed from node to node, until it reaches the bottleneck after the forward delay T_F. It is assumed, that during transmission some data packets are lost so that only αu data arrives at the congested node, where $0 < \alpha_{min} \leq \alpha \leq \alpha_{max} \leq 1$. The delay between generating the control signal and the requested data arrival at the bottleneck node, known as the round trip time (RTT), can be expressed as the sum of the forward and backward propagation delays, $RTT = T_B + T_F$. We denote the discretization period by T and the bottleneck queue length at time kT is represented by $y(kT)$. The buffer is empty prior to data transmission, i.e. $y(kT < 0) = 0$. We assume, that the round trip time is a multiple of the discretization period, i.e. $RTT = mT$, where m is a natural number. The control signal at time kT is represented by $u(kT)$. The first data arrives at the queue after RTT, therefore $y(kT \leq RTT) = 0$.

The bottleneck link available bandwidth is modeled as a non-negative, a priori unknown function of time $d(kT)$. We assume, that only the maximum value of this function, d_{max} is known. We also introduce a function $h(kT)$, that corresponds to the amount of data actually leaving the buffer at time kT. This value cannot exceed the available bandwidth, but it can be smaller if there is not enough data ready to send in the congested node. Therefore, $0 \leq h(kT) \leq d(kT) \leq d_{max}$ for any $k \geq 0$

The queue length can be represented as the difference between incoming and outgoing amounts of data, i.e.

$$y(kT) = \alpha \sum_{j=0}^{k-1} u(jT - RTT) - \sum_{j=0}^{k-1} h(jT) = \alpha \sum_{j=0}^{k-m_{RTT}-1} u(jT) - \sum_{j=0}^{k-1} h(jT). \quad (1)$$

We can also express the system using the standard state space notation

$$\begin{aligned} \boldsymbol{x}[(k+1)T] &= \boldsymbol{A}\boldsymbol{x}(kT) + \Delta\boldsymbol{A}\boldsymbol{x}(kT) + \boldsymbol{b}u(kT) + \boldsymbol{o}h(kT) \\ y(kT) &= \boldsymbol{r}^T\boldsymbol{x}(kT), \end{aligned} \tag{2}$$

where $\boldsymbol{x}(kT) = [x_1(kT) \;\; x_2(kT) \;\; \cdots \;\; x_n(kT)]^T$ is the state vector, $y(kT) = x_1(kT)$ is the queue length, and the remaining state variables are the delayed values of the control signal, i.e. $x_i(kT) = u[(k-n+i-1)T]$, for $i = 2, \ldots, n$. $\boldsymbol{A}$ is a n x n state matrix and $\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{r}$ are n x 1 vectors

$$\boldsymbol{A} = \begin{bmatrix} 1 & \alpha_{max} & 0 & & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{o} = \begin{bmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{3}$$

$\Delta\boldsymbol{A}$ is a n x n model uncertainty matrix where $\Delta a_{12} = \delta\alpha \in [\alpha_{min} - \alpha_{max}, 0]$, and the remaining elements of $\Delta\boldsymbol{A}$ are equal to zero. The desired state of the system is denoted by $\boldsymbol{x_d} = [y_d \;\; 0 \;\; \cdots \;\; 0]^T$, where y_d is the demand bottleneck queue length.

3 Non-switching Reaching Law Based SM Controller

In this section we design a non-switching reaching law based sliding mode controller. We begin by selecting the sliding variable as

$$s(kT) = \boldsymbol{c}^T\boldsymbol{e}(kT), \tag{4}$$

where $\boldsymbol{e}(kT) = \boldsymbol{x_d} - \boldsymbol{x}(kT)$ denotes the closed loop system error. With this choice of variable s the sliding hyperplane is determined by the equation $s(kT) = 0$. The elements of vector $\boldsymbol{c}$ are selected so that $\boldsymbol{c}^T\boldsymbol{b} \neq 0$ and the closed loop system has the desired performance. As we want to obtain finite time error convergence to zero, we choose the vector $\boldsymbol{c}$ in such a way, that the closed loop system exhibits dead-beat dynamics. We begin by calculating the control signal that satisfies $s[(k+1)T] = 0$ and substitute it into (2). In this way we obtain the closed loop system matrix $\boldsymbol{A_c} = [\boldsymbol{I_n} - \boldsymbol{b}(\boldsymbol{c}^T\boldsymbol{b})^{-1}\boldsymbol{c}^T]\boldsymbol{A}$. The matrix $\boldsymbol{A_c}$ has the following characteristic polynomial

$$\det(z\boldsymbol{I_n} - \boldsymbol{A_c}) = z^n + \frac{c_{n-1} - c_n}{c_n}z^{n-1} + \cdots + \frac{c_2 - c_3}{c_n}z^2 + \frac{\alpha_{max}c_1 - c_2}{c_n}z. \tag{5}$$

In order to ensure dead-beat characteristics, polynomial (5) must have the form $\det(z\boldsymbol{I_n} - \boldsymbol{A_c}) = z^n$. We find, that this is achieved with the following choice of $\boldsymbol{c}$

$$\boldsymbol{c} = [1/\alpha_{max} \;\; 1 \;\; 1 \;\; \cdots \;\; 1]^T. \tag{6}$$

We assume, that the aim of the controller is to decrease the value of $|s(kT)|$ until it reaches a band around $s(kT) = 0$, further in the paper called the quasi-sliding band. After reaching this band, the sliding variable should remain inside it. In contrast to some previous works [3], in our definition crossing the hyperplane during the quasi-sliding mode is allowed but not required.

Having calculated the appropriate sliding hyperplane parameters we now propose the following reaching law, that describes the desired sliding variable evolution

$$s[(k+1)T] = \{1 - q[s(kT)]\}s(kT) - \tilde{F}(kT) - \tilde{S}(kT) + F_1, \tag{7}$$

where

$$\tilde{F}(kT) = \boldsymbol{c}^T \boldsymbol{o} h(kT) = -h(kT)/\alpha_{max} \tag{8}$$

is the influence of the disturbance (in our case the amount of outgoing data) on the sliding variable, and

$$\tilde{S}(kT) = \boldsymbol{c}^T \Delta \boldsymbol{A}\boldsymbol{x}(kT) = \delta\alpha x_2(kT)/\alpha_{max} \tag{9}$$

represents the effect of the model uncertainty (the unknown and varying transmission losses). The term $F_1 = -\frac{d_{max}}{2\alpha_{max}}$ is used to compensate the mean value of $\tilde{F}(kT)$. Contrary to some previous works [3] we omit the term S_1 used to compensate the mean value of the model uncertainty, as it could lead to generating a negative control signal, which is not feasible in the considered system. The term $q[s(kT)]$ is given by

$$q[s(kT)] = s_0/[s_0 + |s(kT)|], \tag{10}$$

where $s_0 > \frac{d_{max}(2\alpha_{max} - \alpha_{min})}{\alpha_{max}\alpha_{min}}$ is a design parameter, that allows to tune the controller so that it exhibits fast convergence to the vicinity of $s(kT) = 0$, while not exceeding the maximum transmission rate of the source.

We now derive the control signal, that ensures, that the sliding variable evolution is indeed described by (7). We begin by using (2) to rewrite (4) as

$$s[(k+1)T] = \boldsymbol{c}^T \boldsymbol{x_d} - \boldsymbol{c}^T[\boldsymbol{A}\boldsymbol{x}(kT) + \Delta \boldsymbol{A}\boldsymbol{x}(kT) + \boldsymbol{b}u(kT) + \boldsymbol{c}^T \boldsymbol{o} h(kT)]. \tag{11}$$

Comparing (7) and (11) we arrive at

$$u(kT) = (\boldsymbol{c}^T\boldsymbol{b})^{-1}\{q[s(kT)]s(kT) + d_{max}/2\alpha_{max} - \boldsymbol{c}^T(\boldsymbol{A} - \boldsymbol{I_n})\boldsymbol{x}(kT)\}. \tag{12}$$

We now observe, that by selecting $\boldsymbol{c}$ according to (6) we have obtained $\boldsymbol{c}^T(\boldsymbol{A} - \boldsymbol{I_n}) = [0 \quad \ldots \quad 0]$. This, together with (3) and (10) allows us to express (12) as

$$u(kT) = s_0 s(kT)/[s_0 + |s(kT)|] + d_{max}/2\alpha_{max}. \tag{13}$$

This completes the design of the reaching law based sliding-mode flow controller.

4 Properties of the System

In this section we demonstrate important properties of the considered system, that are guaranteed with the application of our controller. We start by showing, that once the value of $s(kT)$ reaches the quasi-sliding band, it never leaves it again.

Theorem 1 *Once the following inequalities are satisfied, they remain true for the remainder of the control process*

$$\frac{-d_{max}s_0}{2\alpha_{max}s_0 - d_{max}} \leq s(kT) \leq \frac{s_0\left[(\alpha_{max} - \alpha_{min})\left(\frac{y_d s_0}{s_0\alpha_{max}+y_d} + \frac{d_{max}}{2\alpha_{max}}\right) + \frac{d_{max}}{2}\right]}{\alpha_{max}s_0 - \left[(\alpha_{max} - \alpha_{min})\left(\frac{y_d s_0}{s_0\alpha_{max}+y_d} + \frac{d_{max}}{2\alpha_{max}}\right) + \frac{d_{max}}{2}\right]}. \tag{14}$$

In the next theorem we show, that the proposed controller always generates data requests that are non-negative and upper bounded by some a priori known value. As the transmission rate of the source cannot exceed the bandwidth of its outgoing link (and evidently cannot be negative), these properties are important for application in a real network.

Theorem 2 *The control signal, for any $k \geq 0$, satisfies the following inequalities*

$$0 \leq u(kT) \leq \frac{y_d s_0}{s_0\alpha_{max} + y_d} + \frac{d_{max}}{2\alpha_{max}}. \tag{15}$$

Buffer overflows lead to data losses and therefore are undesirable in communication networks. In the next theorem we calculate the value, which the bottleneck queue length never exceeds. If the bottleneck buffer capacity is equal to or greater than this value, then the risk of overflows is completely eliminated.

Theorem 3 *The queue length is always upper bounded by the following value*

$$y(kT) \leq y_d + \frac{d_{max} s_0}{2s_0 - d_{max}/\alpha_{max}}. \tag{16}$$

An efficient flow control algorithm should ensure the greatest possible network throughput. In the last theorem, we derive the minimum value of the demand queue length that ensures, that the queue length never drops to zero, after the first data reach it. This is equivalent to 100 % utilization of the available bandwidth.

Theorem 4 *If the demand queue length satisfies*

$$y_d > \frac{\alpha_{max} s_0 d_{max}(2\alpha_{max} - \alpha_{min})}{2\alpha_{max}\alpha_{min} s_0 - d_{max}(2\alpha_{max} - \alpha_{min})} + \frac{\alpha_{max} d_{max}(n-1)}{\alpha_{min}}, \tag{17}$$

then the queue length is strictly positive for any $k \geq n$.

5 Simulation Results

In order to verify the properties of the proposed control strategy, computer simulations were performed. The discretization period $T = 1$ms and the round trip time $RTT = 11$ms. Therefore, $m = 11$ and $n = 12$. The bounds of the packet loss ratio are $\alpha_{min} = 0.8$ and $\alpha_{max} = 0.98$. The actual transient of the loss ratio is as follows: $\alpha = 0.8$ for $kT \in [0, 0.05)$s, $\alpha = 0.92$ for $kT \in [0.05, 0.1)$s, $\alpha = 0.98$ for $kT \in [0.1, 0.15)$s, $\alpha = 0.86$ for $kT \in [0.15, 0.2]$s. Parameter $d_{max} = 30$ kb, and $d(kT) = 30$ kb for $kT \in [0, 0.05)$s, $d(kT) = 8$ kb for $kT \in [0.05, 0.1)$s, $d(kT) = 0$ kb for $kT \in [0.1, 0.15)$s, $d(kT) = 22$ kb for $kT \in [0.15, 0.2]$s. We assume, that the source can send a maximum of 40 kb of data in a single discretization period. Therefore, we select $s_0 = 25.8$ kb and $y_d = 562$ kb, as this combination satisfies condition (17), ensures convergence to the vicinity of $s(kT) = 0$ and guarantees generating a control signal which will never exceed 40 kb. The simulation results are shown in Figs. 1, 2, and 3. Figure 1 depicts the control signal. As predicted by Theorem 2, it is always non-negative and never exceeds 40 kb. The bottleneck queue length is shown in Fig. 2. It never exceeds the value of 599 kb predicted by Theorem 3 and never drops to zero, after the first data reach it. Therefore, the risk of overflow is eliminated and full bandwidth utilization is ensured. Figure 3 depicts the sliding variable. As we can observe, once it enters the quasi-sliding mode band calculated in Theorem 1 (shown with dashed lines) it never leaves it again.

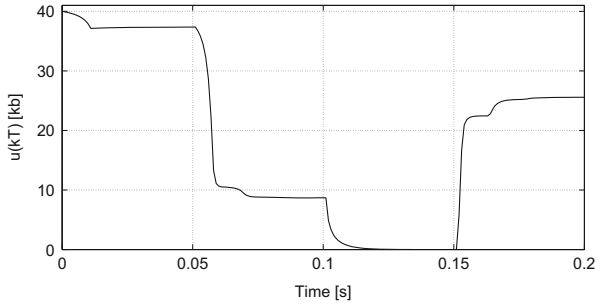


Fig. 1 Control signal

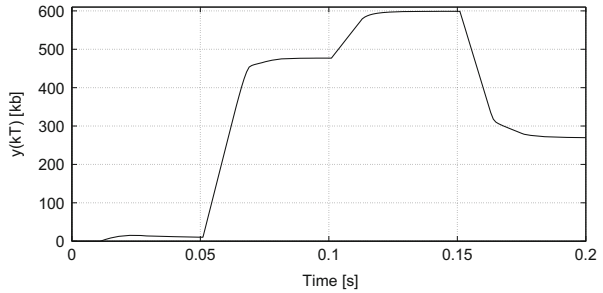


Fig. 2 Bottleneck node queue length

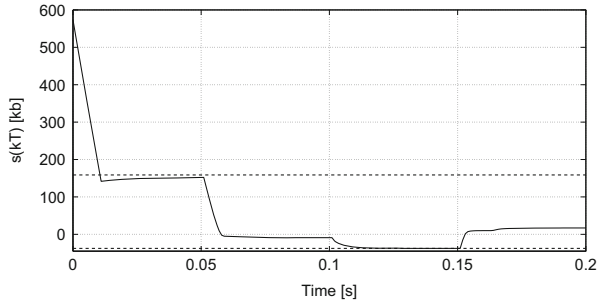


Fig. 3 Sliding variable evolution

6 Conclusions

In this paper a robust sliding mode control strategy for a single virtual circuit in connection oriented communication networks has been proposed. The strategy is designed using the dead-beat sliding mode control paradigm and the reaching law approach. It ensures favorable performance of the circuit even when the available bandwidth and the actual loss rate ratio change with time and are highly unpredictable.

Acknowledgements This work has been performed in the framework of a project "Optimal sliding mode control of time delay systems"İ financed by the National Science Centre of Poland decision number DEC 2011/01/B/ST7/02582. Kind support provided by the Foundation for Polish Science under Mistrz grant is also acknowledged.

References

1. Bartoszewicz, A.: Nonlinear flow control strategies for connection oriented communication networks. Proc. IEE Part D Control Theory Appl. **153**(1), 21–28 (2006)
2. Bartoszewicz, A., Lesniewski, P.: A new reaching law based sliding mode flow controller for connection-oriented data transmission networks. In: 13th International Workshop on Variable Structure Systems, Nantes (2014)
3. Gao, W., Wang, Y., Homaifa, A.: Discrete-time variable structure control systems. IEEE Trans. Ind. Electron. **42**(2), 117–122 (1995)
4. Jagannathan, S., Talluri, J.: Predictive congestion control of ATM networks: multiple sources/single buffer scenario. Automatica **38**(5), 815–820 (2002)
5. Jing, Y., Yu, N., Kong, Z., Dimirovski, G.: Active queue management algorithm based on fuzzy sliding model controller. In: Proceedings of the 17th IFAC World Congress, Seoul, pp. 6148–6153 (2008)
6. Laberteaux, K., Rohrs, C., Antsaklis, P.: A practical controller for explicit rate congestion control. IEEE Trans. Autom. Control **47**(6), 960–978 (2002)
7. Quet, P., Ataslar, B., Iftar, A., Ozbay, H., Kalyanaraman, S., Kang, T.: Rate-based flow controllers for communication networks in the presence of uncertain time-varying multiple time-delays. Automatica **38**(6), 917–928 (2002)

MS 28
MINISYMPOSIUM: SELECTED TOPICS IN SEMI-CLASSICAL AND QUANTUM TRANSPORT MODELING

Organizer

Dragica Vasileska[1]

Speakers

Massimo Fischetti[2]
Pseudopotential-Based Study of Electron Transport in Low-Dimensionality Nanostructures

Neophytos Neophytou[3] and Hans Kosina[4]
Full-Band Calculations of Thermoelectric Properties of Si Nanowires and Thin Layers

Dmitri Osintsev[5], Viktor Sverdlov[6] and Siegfried Selberherr[7]
Electron Momentum and Spin Relaxation in Silicon Films

[1]Dragica Vasileska, Arizona State University, Tempe, AZ, USA.
[2]Massimo Fischetti, University of Texas, Dallas, TX, USA.
[3]Neophytos Neophytou, University of Warwick, Coventry, UK.
[4]Hans Kosina, Technical University, Vienna, Austria.
[5]Dmitri Osintsev, Technical University, Vienna, Austria.
[6]Viktor Sverdlov, Technical University, Vienna, Austria.
[7]Siegfried Selberherr, Technical University, Vienna, Austria.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_94

Philippe Dollfus[8], Viet Hung Nguyen[9], Salim Berrada[10], Van Truong Tran[11], Mai Chung Nguyen[12], Arnaud Bournel[13] and Jérôme Saint-Martin[14]
Strategies to Improve the Control of Current in Graphene Transistors

Ivan Dimov[15], Mihail Nedjalkov[16], Jean Michel Sellier[17] and Siegfried Selberherr[7]
Neumann Series Analysis of the Wigner Equation Solution

Salvatore Maria Amoroso[18], Louis Gerrer[19], Vihar Georgiev[20] and Asen Asenov[21]
Simulation of Oxide Reliability of Nanoscaled MOSFETs Using Drift-Diffusion, Monte Carlo and Full-Quantum Techniques

Zlatan Stanojevic[22], Lidija Filipovic[23], Oskar Baumgartner[24], Markus Karner[25], Christian Kernstock[26] and Hans Kosina[4]
Advanced Numerical Methods for Semiclassical Transport Simulation in Ultra-Narrow Channels

Daniel Brinkman[27], Klemens Fellner[28] and Peter Markowich[29]
Drift-Diffusion Models for Organic Photovoltaics: Deriving a Unipolar Model

[8]Philippe Dollfus, University Paris-Sud, Paris, France.
[9]Viet Hung Nguyen, University Paris-Sud, Paris, France.
[10]Salim Berrada, University Paris-Sud, Paris, France.
[11]Van Truong Tran, University Paris-Sud, Paris, France.
[12]Mai Chung Nguyen, University Paris-Sud, Paris, France.
[13]Arnaud Bournel, University Paris-Sud, Paris, France.
[14]Jérôme Saint-Martin, University Paris-Sud, Paris, France.
[15]Ivan Dimov, Bulgarian Academy of Sciences, Sofia, Bulgaria.
[16]Mihail Nedjalkov, Technical University, Vienna, Austria.
[17]Jean Michel Sellier, Bulgarian Academy of Sciences, Sofia, Bulgaria.
[18]Salvatore Maria Amoroso, University of Glasgow, Glasgow, UK.
[19]Louis Gerrer, University of Glasgow, Glasgow, UK.
[20]Vihar Georgiev, University of Glasgow, Glasgow, UK.
[21]Asen Asenov, University of Glasgow, Glasgow, UK.
[22]Zlatan Stanojevic, Technical University, Vienna, Austria.
[23]Lidija Filipovic, Technical University, Vienna, Austria.
[24]Oskar Baumgartner, Technical University, Vienna, Austria.
[25]Markus Karner, Technical University, Vienna, Austria.
[26]Christian Kernstock, Technical University, Vienna, Austria.
[27]Daniel Brinkman, Arizona State University, Tempe, AZ, USA.
[28]Klemens Fellner, University of Graz, Graz, Austria.
[29]Peter Markowich, University of Cambridge, Cambridge, UK.

Keywords

Nanoelectronics
Quantum transport
Semi-classical transport
Semiconductor

Short Description

The field of nanoelectronics has entered every pore of our everyday life starting from cell-phones and computers to complicated medical equipment. Indeed, in the last twenty years, the progress in nanotechnology has revolutionized the medical field in terms of the development of very sophisticated diagnostic tools. The development of nanotechnology has been made possible with the application of Moore's law and transistor scaling into shorter and shorter channel lengths that have allowed more functions to be put on a single chip. Unfortunately, nanotechnology is getting to a point at which further miniaturization of transistors has become more difficult. This is due to the fact that nanoscale feature sizes can not be successfully achieved using standard optical lithography processes. To reduce the cost of arriving at optimal device designs, simulation is becoming more and more of a essential avenue to be pursued in both Industry and Academia. Simulation offers possibilities that are not attainable via experiments, such as looking into internal variables, like the electric field profile, that can not be measured experimentally but definitely has significant impact on the operation of a device. Also, simulation is cheaper and allows one to use a range of simulation models starting from compact models to semi-classical transport approaches to quantum-mechanical density matrix, Wigner Function and Green's function approaches.

This special session contained papers with topics that range from compact modeling and fluctuations in device characteristics down to fully quantum mechanical and atomistic modeling needed for describing the operation of the nanostructure devices of today. The papers are written by leading experts in the field and the Session Organizer is thankful for their invaluable contribution.

Advanced Numerical Methods for Semi-classical Transport Simulation in Ultra-Narrow Channels

Zlatan Stanojević, Oskar Baumgartner, Markus Karner, Lidija Filipović, Christian Kernstock, and Hans Kosina

Abstract In this work we present a semi-classical modeling and simulation approach for ultra-narrow channels that has been implemented as part of the Vienna Schrödinger-Poisson (VSP) simulation framework (Baumgartner, J Comput Electron 12:701–721, 2013; http://www.globaltcad.com/en/products/vsp.html (2014)) over the past few years. Our research has been driven by two goals: maintaining high physical accuracy of the models while producing a computationally efficient and flexible simulation code.

Keywords Device design • Semi-classical transport • Ultra-narrow channels

1 Introduction

The first commercialization of the FinFET sparked interest in non-planar, ultra-narrow channels among researchers and manufacturers alike. Questions as to what are the design parameters of such a device or whether the existing FinFET process is optimal have arisen. The influence of strain and usage of materials other than silicon are also hotly debated. As a consequence, the interest in advanced modeling and simulation tools, that could help understand the physics of ultra-narrow channels, is rising. But CMOS is not the only application. There are other fields of research that are investigating the benefits of ultra-narrow channels in various applications, such as silicon thermoelectrics, photovoltaics, or nano-electromechanical systems. These too profit from a sound modeling and simulation framework that can help understand effects and guide experimental efforts.

In novel device designs it is important to fully capture the effects of geometry, crystal orientation, material composition, doping, and strain [11]. To do so, we

Z. Stanojević (✉) • O. Baumgartner • L. Filipović • H. Kosina
Institute for Microelectronics, TU Wien, Gußhausstraße 27-29/E360, 1040 Vienna, Austria
e-mail: stanojevic@iue.tuwien.ac.at; lidijafilipovic@iue.tuwien.ac.at; baumgartner@iue.tuwien.ac.at; kosina@iue.tuwien.ac.at

M. Karner • C. Kernstock
Global TCAD Solutions, Landhausgasse 4/1a, 1010 Vienna, Austria
e-mail: m.karner@globaltcad.com; c.kernstock@globaltcad.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_95

developed a simulation work-flow consisting of a electronic structure calculation based on **k**·**p**-theory, self-consistently coupled with electrostatics, semi-classical modeling of scattering processes, and transport modeling based on the linearized Boltzmann transport equation, which allows extraction of channel mobility and transconductance. The novelty of our approach is that we mostly rely on numerical solutions to each of the subproblems in the work-flow, thereby gaining flexibility. On the one hand, a numerical treatment allows to avoid some analytical approximations, such as the momentum relaxation time approximation [4], but on the other hand results in a generally increased computational workload. This is mitigated through several innovations that allow us to use computational resources more economically. The particular innovations for each of the sub-problems will be addressed in the following sections.

2 Electronic Structure

Our electronic structure model is based on **k**·**p** theory, although the methods shown here also apply to other electronic structure models, such as tight binding. Rather than focusing on a particular **k**·**p** Hamiltonian, we formulate a generic n-band effective Hamiltonian [1],

$$\mathbf{H} = \begin{pmatrix} H_{11} & H_{12} & \cdots \\ & H_{22} & \cdots \\ \text{c.c.} & & \ddots \end{pmatrix}, \quad H_{lm} = \frac{\hbar^2}{2}\mathbf{k}\cdot\mathbf{m}_{lm}^{-1}\cdot\mathbf{k} + \hbar\mathbf{v}_{lm}\cdot\mathbf{k} + U_{lm} + D_{lm}^{\xi\eta}\varepsilon^{\xi\eta}. \tag{1}$$

The model parameters are the coupling mass tensors $\mathbf{m}_{lm}$, Fermi velocities $\mathbf{v}_{lm}$, coupling potentials U_{lm}, and deformation potentials $D_{lm}^{\xi\eta}$ for each strain component $\varepsilon^{\xi\eta}$. Using this template we can construct a variety of different **k**·**p** Hamiltonians, such as the 3×3 and 6×6 Dresselhaus-Kip-Kittel [3] (DKK) Hamiltonian for valence bands in diamond and zinc-blende crystals, the 4×4 Hamiltonian due to Kane [6], the 2×2 Hensel-Hasegawa-Nakayama Hamiltonian for electrons in silicon [5], or the 4×4 Hamiltonian for lead-salts due to Dimmock and Wright [2].

To obtain the electronic structure of nano-structures the effective **k**·**p** Schrödinger equation is solved by separating the wave function into a confined state and a plane wave. To compute the confined state, we replace the **k**-vector by the operator $-\mathrm{i}\boldsymbol{\nabla}_\perp + \mathbf{k}_\parallel$, discretize the Schrödinger equation on an unstructured mesh [1], and calculate its eigenenergies and eigenstates. Thus we obtain the $E(\mathbf{k}_\parallel)$-relation, i.e. the subband dispersion.

The number of required subbands is not known beforehand and generally depends on the nano-structure geometry, size, and electrostatic potential. However, a cut-off energy can be defined beyond which the occupation numbers of the states are negligible. This energy is conveniently placed several $k_\mathrm{B}T$ away from the Fermi energy. The task is now to find all electronic states between the band edge and the cut-off energy. Standard packages for iterative eigenvalue algorithms,

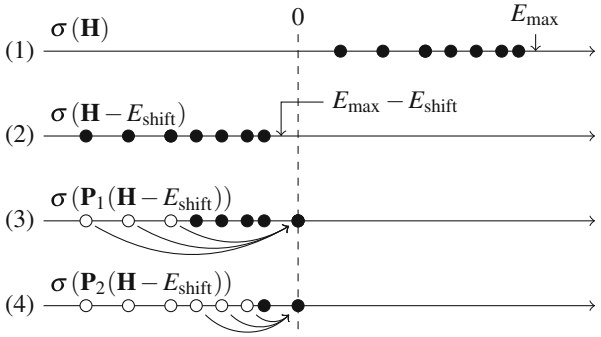


Fig. 1 Searching for eigenvalues up to $E_{\max}$; the spectrum of the Hamiltonian $\mathbf{H}$ is shifted to the left by $E_{\text{shift}} > E_{\max}$ and a subspace solver is invoked to find the first few eigenenergies. A projection matrix selectively shifts the found eigenenergies to 0. The process is repeated until all eigenenergies $E_i < E_{\max}$ are found

such as ARPACK [7], cannot provide all the eigenvalues in an interval, so we developed a search algorithm that can be used on top of an iterative eigenvalue solver. The algorithm, shown in Fig. 1, shifts the spectrum of the Hamiltonian to the left by $E_{\text{shift}} > E_{\max}$ and the first n_{ev} eigenenergies are computed by a subspace solver (e.g. ARPACK). A projection matrix $\mathbf{P}_1 = \mathbf{I} - \mathbf{v}_i\mathbf{v}_i^{\text{H}}$ is then constructed from the eigenstates $\mathbf{v}_i$. The subspace solver is invoked again on the projected system $\mathbf{P}_1(\mathbf{H}-E_{\text{shift}})$. The projection selectively shifts the found eigenenergies to 0, preventing the solver to converge on already found eigenenergies, which calculates the next n_{ev} eigenenergies instead. The process is repeated until all eigenenergies $E_i < E_{\max}$ are found. Since subspace solvers do not require matrices to be provided explicitly, relying instead on matrix-vector multiplications performed externally, the operation $\mathbf{P}_n(\mathbf{H}-E_{\text{shift}})$ can be done implicitly, which results in little computational overhead.

3 Carrier Scattering

The electronic states are used to evaluate the transition rates due to scattering, which are needed for any kind of semi-classical transport simulation in nano-structures. Here, we give the expressions for non-polar phonon scattering, Coulomb scattering and surface roughness scattering in a dimension-independent way, i.e. all expressions are valid for one and two-dimensional carrier gases (1DEG, 2DEG), and, of course, also for bulk (3DEG).

For acoustic phonon scattering the squared matrix element for a transition from state $(n, \mathbf{k})$ to $(n', \mathbf{k}')$ reads

$$|H_{n,n';\mathbf{k},\mathbf{k}'}|^2 = \frac{\pi k_{\text{B}} T D_{\text{A}}^2}{\hbar \rho_{\text{m}} c_{\text{l}} L^d} \int |\psi_{n,\mathbf{k}}(\mathbf{r})|^2 |\psi_{n',\mathbf{k}'}(\mathbf{r})|^2 \mathrm{d}^d r, \tag{2}$$

where d is the dimension of the channel cross-section implying a carrier gas of dimension $(3-d)$ [9]. For Coulomb scattering, the squared matrix element reads

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = q_0^2 \int |U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})|^2 N_{\text{imp}}(\mathbf{r}) \mathrm{d}^d r, \tag{3}$$

with the matrix element for a single point charge,

$$U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r}) = q_0 \int \psi^*_{n,\mathbf{k}}(\mathbf{r}') \psi_{n',\mathbf{k}'}(\mathbf{r}') G_{\mathbf{k}-\mathbf{k}'}(\mathbf{r},\mathbf{r}') \mathrm{d}^d r'. \tag{4}$$

$G_{\mathbf{k}-\mathbf{k}'}(\mathbf{r},\mathbf{r}')$ is the reduced electrostatic Green's function which already contains a linearized screening term.

Providing a squared matrix element for surface roughness scattering is more involved. The Prange-Nee squared matrix element [8],

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \frac{C(\mathbf{q})}{A} |F_{n,n';\mathbf{k},\mathbf{k}'}|^2, \tag{5}$$

is extended to obtain [10]

$$\left\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \right\rangle = \frac{1}{2\pi L} \int_{\mathbb{R}} |\tilde{f}_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)|^2 C(\mathbf{q}) \mathrm{d}q_\perp, \tag{6}$$

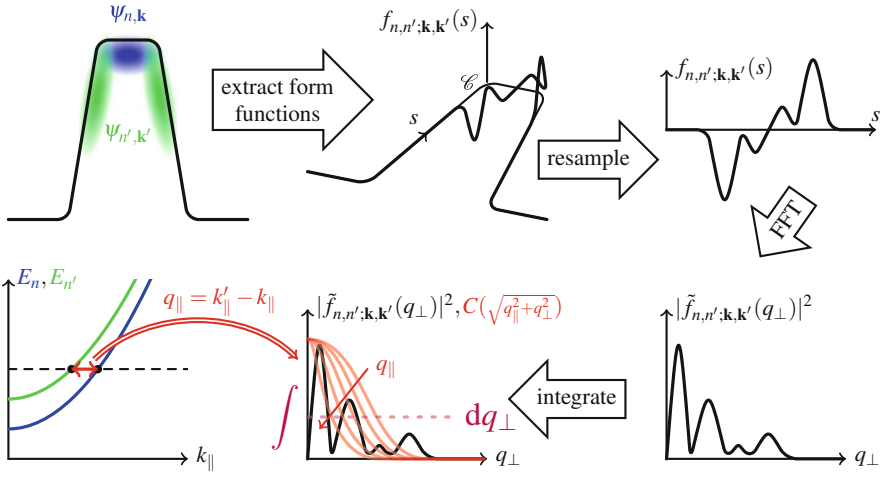


Fig. 2 Two wavefunctions $\psi_{n,\mathbf{k}}$ and $\psi_{n',\mathbf{k}'}$ in a fin-cross-section interact through surface roughness. The corresponding form-function $f_{n,n';\mathbf{k},\mathbf{k}'}(s)$ is computed and interpolated onto an equidistant s-grid. The spectral form-function $\tilde{f}_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ is then computed using the fast Fourier transform (FFT). For a given energy the difference of axial k-vectors is evaluated which represents the axial momentum transfer $q_\parallel$. The roughness *power spectrum* $C(q)$ is offset using $\sqrt{q_\parallel^2+q_\perp^2}$ and its product with the spectral form-function $\tilde{f}_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ is integrated to obtain the square matrix element

where $C(\mathbf{q})$ is the roughness power spectrum, and $F_{n,n';\mathbf{k},\mathbf{k}'}$ are form-factors. The function $\tilde{f}_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ is called a *spectral form-function*. It is the Fourier transform of the form-function, i.e. the form-factor dependent on the position along the curve $\mathscr{C}$ which represents the ideal channel surface in the cross-section plane. The squared matrix element is obtained by integrating the product of the squared spectral form function and the in-plane component of the roughness power spectrum, which is dependent on $q_\parallel = \mathbf{k}'_\parallel - \mathbf{k}_\parallel$. For planar channels the spectral form function is reduced to $|\tilde{f}_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)|^2 = 2\pi/L\delta(q_\perp)|F_{n,n';\mathbf{k},\mathbf{k}'}|^2$ reproducing the Prange-Nee expression (Fig. 2).

4 Low-Field Transport

The linearized Boltzmann transport equation for a small homogeneous driving field $\mathbf{F} = -q_0\mathbf{E}$ and elastic or quasi-elastic scattering processes,

$$\sum_{n',\mathbf{k}'} S_{n,n'}(\mathbf{k},\mathbf{k}')\left[f_n^1(\mathbf{k}) - f_{n'}^1(\mathbf{k}')\right] = -\mathbf{F}\cdot\mathbf{v}_n(\mathbf{k})\frac{\mathrm{d}f^0}{\mathrm{d}E},$$

is discretized in $\mathbf{k}$-space, giving

$$\sum_{\nu'} S_{\nu,\nu'} w_{\nu,\nu'}\left[f_\nu^1 - f_{\nu'}^1\right] = -\mathbf{F}\cdot\mathbf{v}_\nu\frac{\mathrm{d}f^0}{\mathrm{d}E}W_{\mathbf{k}},$$

where ν represents the global index of a state $(n,\mathbf{k})$, $w_{\nu,\nu'}$ is a coupling weight between state ν and ν', and $W_{\mathbf{k}}$ is the $\mathbf{k}$-grid cell volume. The weights are computed by piecewise integration of the density-of-states product, $g(E)_\nu g(E)_{\nu'}$, along the equi-energy lines for a 2DEG, or points for a 1DEG (see Fig. 3). Due to energy conservation, most of the weights are zero, and square matrix elements are computed for the non-zero ones only. Sorting all states ν by their absolute energy produces a symmetric skyline matrix, allowing dense date storage and efficient memory access. In a realistic device the number of non-zero elements can still reach several millions and most computation time is spent in the evaluation of the square matrix elements for each transition. Fortunately, this task can be parallelized very effectively.

Having evaluated the elements of the scattering operator's matrix, the linearized Boltzmann transport equation is readily solved in $\mathbf{k}$-space using iterative methods. The resulting linear distribution responses for a p-type Si MOS and 5 nm thick p-type Si nanowire are shown in Fig. 4.

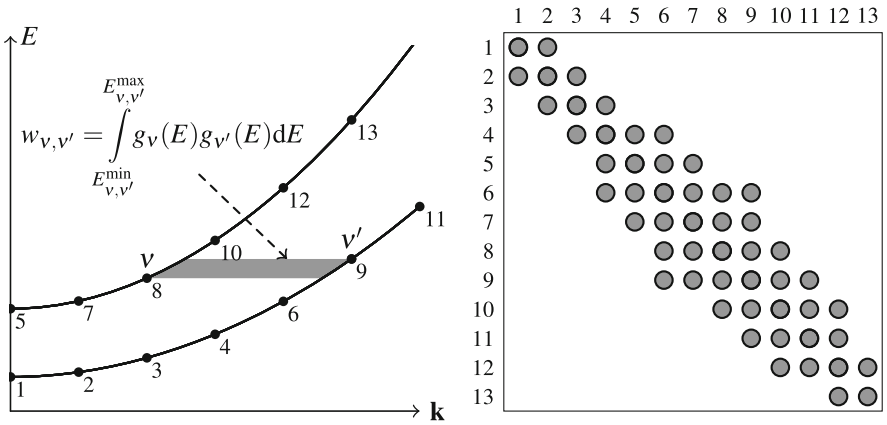


Fig. 3 *Left*: calculation of the coupling weights for an elastic scattering operator; $w_{\nu,\nu'}$ is obtained by integrating the product of the density of states of state ν and ν' over the energy interval where ν and ν' overlap. Multiplied by a scattering rate it gives the probability flux between ν and ν'. *Right*: the resulting non-zero pattern of the discretized scattering operator; sorting all states by absolute energy produces dense symmetric skyline matrix, thus eliminating storage overhead

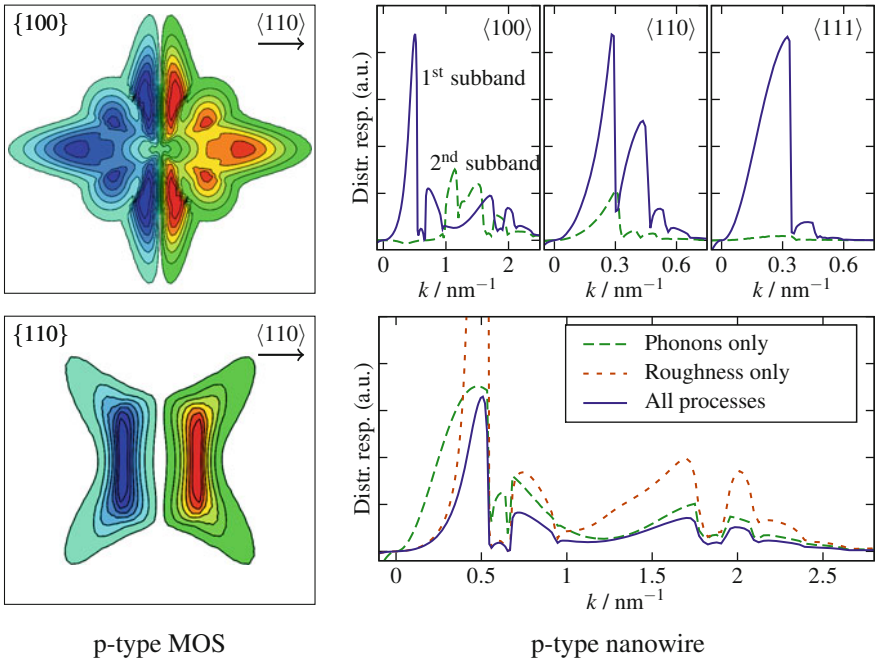


Fig. 4 *Left*: calculated **k**-space hole distribution response for a Si p-type MOS channel at 1 MV/cm for different substrate orientations; *top right*: hole distribution response in a cylindrical 5 nm thick p-type Si nanowire for different growth orientations; *bottom right*: the influences of phonon and roughness scattering on the hole response distribution for the {100}-nanowire

5 Conclusion

We presented several improvements to the numerical calculation of electronic structure, scattering rates, and low-field transport in planar and non-planar semiconductor channels. The improvements target to increase computational efficiency while avoiding any sacrifice of accuracy within the semi-classical theory. The presented modeling framework provides great flexibility with respect to dimensionality, the form of the Hamiltonian, and the cross-section geometry of the channel, effectively providing a physics-based TCAD simulation tool for semi-classical carrier transport.

Acknowledgements This work has been supported by the Austrian Science Fund through contracts F2509 and I841-N16.

References

1. Baumgartner, O., Stanojevic, Z., Schnass, K., Karner, M., Kosina, H.: VSP–a quantum-electronic simulation framework. J. Comput. Electron. **12**, 701–721 (2013). doi: 10.1007/s10825-013-0535-y. http://dx.doi.org/10.1007/s10825-013-0535-y
2. Dimmock, J.O., Wright, G.B.: Band edge structure of PbS, PbSe, and PbTe. Phys. Rev. **135**, A821–A830 (1964). doi: 10.1103/PhysRev.135.A821. http://link.aps.org/doi/10.1103/PhysRev.135.A821
3. Dresselhaus, G., Kip, A.F., Kittel, C.: Cyclotron resonance of electrons and holes in silicon and germanium crystals. Phys. Rev. **98**, 368–384 (1955). doi: 10.1103/PhysRev.98.368. http://link.aps.org/doi/10.1103/PhysRev.98.368
4. Fischetti, M.V., Ren, Z., Solomon, P.M., Yang, M., Rim, K.: Six-band $k{\cdot}p$ calculation of the hole mobility in silicon inversion layers: dependence on surface orientation, strain, and silicon thickness. J. Appl. Phys. **94**(2), 1079–1095 (2003). doi: http://dx.doi.org/10.1063/1.1585120. http://scitation.aip.org/content/aip/journal/jap/94/2/10.1063/1.1585120
5. Hensel, J.C., Hasegawa, H., Nakayama, M.: Cyclotron resonance in uniaxially stressed silicon. II. Nature of the covalent bond. Phys. Rev. **138**(1A), A225–A238 (1965). doi: 10.1103/PhysRev.138.A225
6. Kane, E.O.: Energy band structure in p-type germanium and silicon. J. Phys. Chem. Solids **1**(1–2), 82–99 (1956). doi: http://dx.doi.org/10.1016/0022-3697(56)90014-2. http://www.sciencedirect.com/science/article/pii/0022369756900142
7. Lehoucq, R., Sorensen, D., Yang, C.: ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods (1998)
8. Prange, R.E., Nee, T.W.: Quantum spectroscopy of the low-field oscillations in the surface impedance. Phys. Rev. **168**, 779–786 (1968). doi: 10.1103/PhysRev.168.779. http://link.aps.org/doi/10.1103/PhysRev.168.779
9. Ramayya, E.B., Vasileska, D., Goodnick, S.M., Knezevic, I.: Electron transport in silicon nanowires: the role of acoustic phonon confinement and surface roughness scattering. J. Appl. Phys. **104**(6), 063711 (2008). doi: 10.1063/1.2977758. http://link.aip.org/link/?JAP/104/063711/1
10. Stanojevic, Z., Kosina, H.: Surface-roughness-scattering in non-planar channels – the role of band anisotropy. In: International Conference on Simulation of Semiconductor Processes and Devices, pp. 352–355 (2013)
11. Stanojevic, Z., Karner, M., Kosina, H.: Exploring the design space of non-planar channels: shape, orientation, and strain. In: International Electron Device Meeting, pp. 332–335 (2013). doi: 10.1109/IEDM.2013.6724618

Electron Momentum and Spin Relaxation in Silicon Films

D. Osintsev, V. Sverdlov, and S. Selberherr

Abstract Semiconductor spintronics is promising, because it allows creating microelectronic elements which are smaller and consume less energy than present charge-based devices. Silicon is the main element of modern charge-based electronics, thus, understanding the peculiarities of spin propagation in silicon is the key for designing novel devices. We investigate the electron momentum and the spin relaxation in thin (001) oriented SOI films using a $\mathbf{k} \cdot \mathbf{p}$-based approach with spin degree of freedom properly included. We demonstrate that shear strain routinely used to enhance the electron mobility can boost the spin lifetime by an order of magnitude.

Keywords Charge-based electronics • Semiconductor • Silicon films

1 Introduction

Growing technological challenges and soaring costs are gradually bringing MOSFET scaling to an end. This intensifies the search of alternative technologies and computational principles. The electron spin attracts attention as a possible candidate to be used in future electron devices for complementing or even replacing the charge degree of freedom employed in MOSFETs. The spin state is characterized by the two spin projections on a given axis and it thus has a potential in digital information processing. In addition, only a small amount of energy is needed to flip the spin orientation. Silicon is an ideal material for spintronic applications due to the long spin lifetime in the bulk. The spin lifetime is determined by spin-flip scattering between the valleys located on different crystallographic axes [1, 2]. This mechanism is suppressed in thin films; however, large spin relaxation in gated silicon structures was observed [3]. Understanding the spin relaxation mechanisms and identifying ways to boost the spin lifetime in confined electron systems is urgently needed.

D. Osintsev (✉) • V. Sverdlov • S. Selberherr
Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, A-1040 Wien, Austria
e-mail: Osintsev@iue.tuwien.ac.at; Sverdlov@iue.tuwien.ac.at; Selberherr@iue.tuwien.ac.at

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_96

2 Model and Results

We investigate spin relaxation in (001) silicon structures by taking into account surface roughness and electron-phonon interaction induced momentum scattering and spin relaxation. The two interfaces of the film are assumed to be independent. The surface roughness scattering matrix elements are proportional to the product of the corresponding subband wave functions' derivatives at each interface [4]. To find the wave functions and matrix elements we use the effective $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian written at the X-point for the two relevant valleys along the OZ-axis with shear strain and the spin degree of freedom included [5]. We generalize the deformation potential based electron-phonon scattering theory to include the shear strain deformation potential and the deformation potential due to spin-orbit interaction responsible for spin relaxation in confined systems [6].

In the two valleys' plus two spin projections' basis the subband wave functions possess four components. These wave functions are written as ($k_x = 0$)

$$\Psi_1 = \begin{pmatrix} \Psi_{1,1} \\ \Psi_{1,2} \\ \Psi^*_{1,1} \\ -\Psi^*_{1,2} \end{pmatrix} \Psi_2 = \begin{pmatrix} -\Psi_{1,2} \\ \Psi_{1,1} \\ \Psi^*_{1,2} \\ \Psi^*_{1,1} \end{pmatrix} \Psi_3 = \begin{pmatrix} \Psi_{2,2} \\ \Psi_{2,1} \\ -\Psi^*_{2,2} \\ \Psi^*_{2,1} \end{pmatrix} \Psi_4 = \begin{pmatrix} -\Psi_{2,1} \\ \Psi_{2,2} \\ -\Psi^*_{2,1} \\ -\Psi^*_{2,2} \end{pmatrix}, \qquad (1)$$

where $\Psi_{1(3)}$ and $\Psi_{2(4)}$ are the up- and down-spin wave functions for the first (second) subband. Wave functions with opposite spin in the same valley are orthogonal. The dominant components are $\Psi_{1,1}$ and $\Psi_{2,2}$ for $\Psi_{1(2)}$ and $\Psi_{3(4)}$, respectively. Thus, Ψ_1 and Ψ_3 are considered as up-spin wave functions, while Ψ_2 and Ψ_4 are the down-spin wave functions. The small components of the wave functions are the result of the spin-orbit interaction taken into account with the $\tau_y \otimes \Delta_{SO}(k_x\sigma_x - k_y\sigma_y)$ term, where $\Delta_{SO} = 1.27\,\text{meVnm}$ [2, 5], τ_y is the y-Pauli matrix in the valley degree of freedom, and σ_x and σ_y are the spin Pauli matrices.

Without spin-orbit interaction included the wave function conserves the spin projection which is assumed along the OZ-axis. The large components of the wave functions are well described by $\Psi_{1,1(2,2)} = e^{ik_0 z}\sin\left(\frac{\pi z}{t}\right)$ (Fig. 1) and their conjugates. This expression corresponds to the usual envelope quantization function. Under shear strain ε_{xy} the degeneracy between the two unprimed subbands is lifted which results in slightly different envelope functions $\Psi_{1,1}$ and $\Psi_{2,2}$ (Fig. 2).

The small components of the four-components' wave function are proportional to the spin-orbit interaction strength. The amplitude of these components shown in Fig. 3 for an unstrained film of 4 nm thickness for $k_x = 0$ strongly depends on the value of k_y. For $k_y = 1\,\text{nm}^{-1}$ the small components of the wave functions are pronounced, while decreasing k_y value makes the small components vanish.

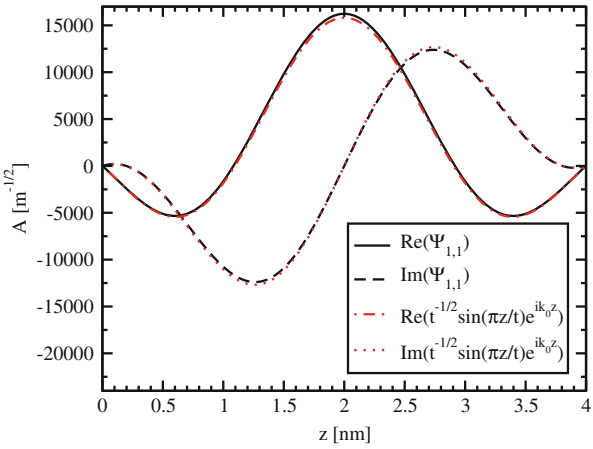


Fig. 1 The large component of the wave function of the lowest unprimed subband in an unstrained film located in the valley centered at k_0

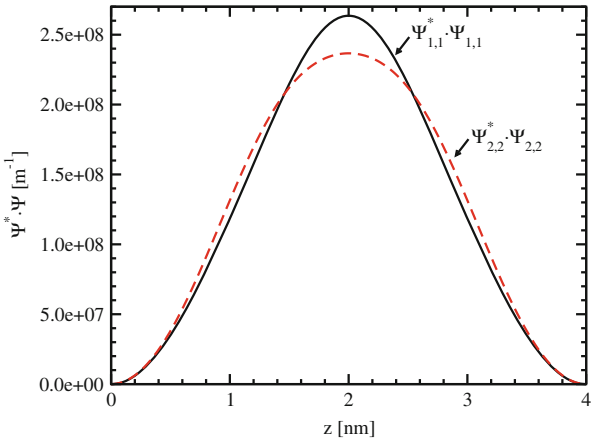


Fig. 2 The large components of the two unprimed subbands with ε_{xy}=0.05 %

Shear strain ε_{xy} greatly suppresses the small components as shown in Fig. 4. $\Psi_{1,2}$ for the strain value of 1 % is almost vanished, while for the film the wave function component is significant (Fig. 4). Vanishing values of the small components decrease the spin mixing between the states with the opposite spin projections, which results in longer spin lifetime.

Surface roughness limited spin lifetime and momentum relaxation time as a function of temperature are shown in Fig. 5. For the chosen electron concentrations the spin and momentum relaxation times decrease with temperature [1]. As a confirmation of the Elliot-Yafet spin relaxation mechanism, the spin lifetime remains proportional to the momentum relaxation time (Fig. 5).

Under shear strain the spin lifetime is enhanced much stronger than the momentum relaxation time (Fig. 6) due to the small components' suppression (Fig. 4). An

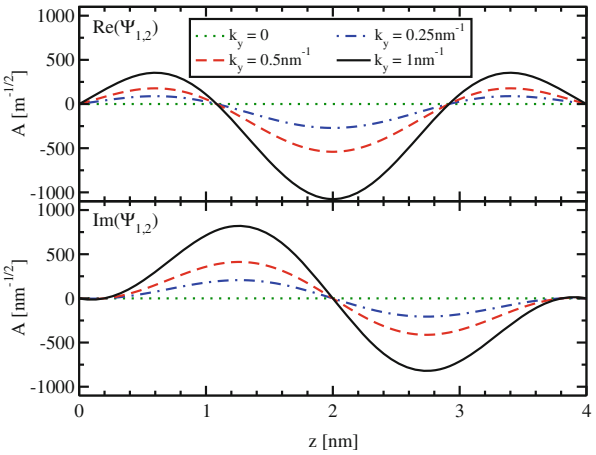


Fig. 3 The small components are proportional to the strength of the spin-orbit interaction

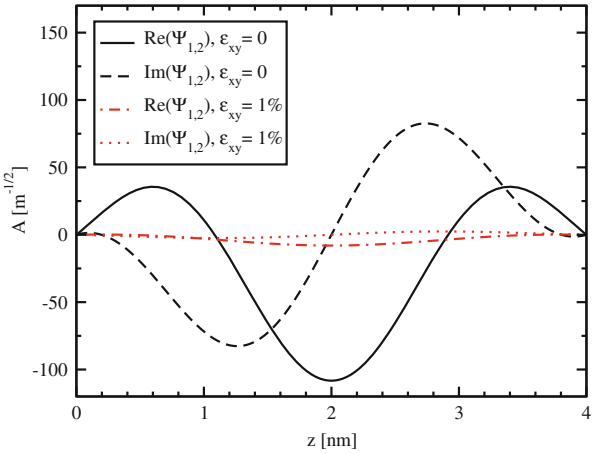


Fig. 4 The small components are considerably suppressed by tensile shear strain

extensive code parallelization and optimization allowed us to extend the method [6] for a larger set of parameters, including the film thickness and the electron concentration. The ratio of the spin to the momentum relaxation time (inset in Fig. 6) demonstrates the significant enhancement.

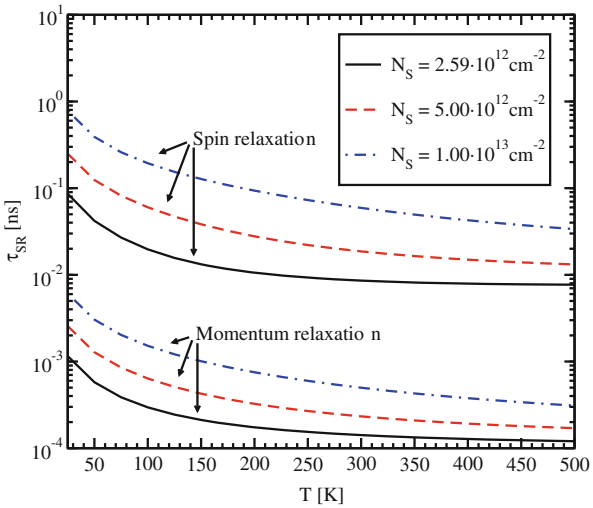


Fig. 5 The spin lifetime is proportional to the momentum relaxation time as function of temperature. This is an indication of the Elliot-Yafet spin relaxation mechanism [1]

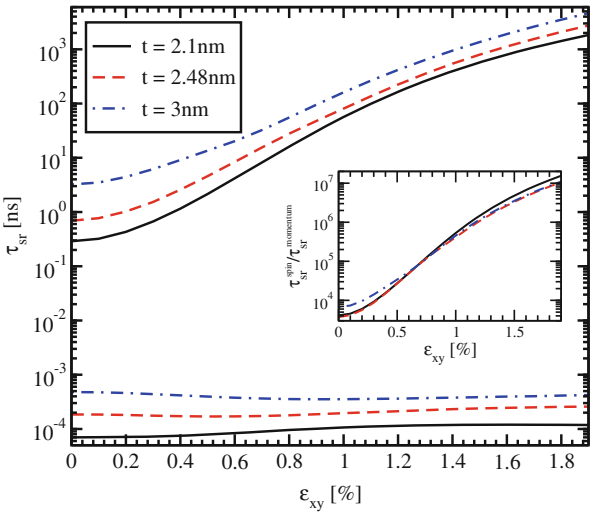


Fig. 6 Dependence of the spin lifetime and the momentum relaxation time on shear strain for different thicknesses. *Inset*: ratio of the spin to the momentum relaxation time

3 Conclusion

We have used a $\mathbf{k} \cdot \mathbf{p}$ approach to evaluate the momentum relaxation time and the spin lifetime in strained thin silicon films. We have shown that the small components of four-component wave functions vanish with strain. Thus, the spin lifetime is enhanced much stronger by shear strain than the momentum relaxation time. Tensile shear strain boosts both the electron mobility and the spin lifetime in silicon films.

Acknowledgements This work is supported by the European Research Council through the grant #247056 MOSILSPIN. The computational results have been achieved in part using the Vienna Scientific Cluster (VSC).

References

1. Song, Y., Dery, H.: Analysis of phonon-induced spin relaxation processes in silicon. Phys. Rev. B **86**, 085201 (2012)
2. Li, P., Dery, H.: Spin-orbit symmetries of conduction electrons in silicon. Phys. Rev. Lett. **107**, 107293 (2011)
3. Li, J., Appelbaum, I.: Lateral spin transport through bulk silicon. Appl. Phys. Lett. **100** (16) (2012), http://dx.doi.org/10.1063/1.4704802
4. Fischetti, M.V., Ren, Z., Solomon, P.M., Yang, M., Rim, K.: Six-band $\mathbf{k} \cdot \mathbf{p}$ calculation of the hole mobility in silicon inversion layers: dependence on surface orientation, strain, and silicon thickness. J. Appl. Phys. **94**(2), 1079–1095 (2003)
5. Osintsev, D., Baumgartner, O., Stanojevic, Z., Sverdlov, V., Selberherr, S.: Subband splitting and surface roughness induced spin relaxation in (001) silicon SOI MOSFETs. Solid State Electron. **90**, 34–38 (2013)
6. Osintsev, D., Sverdlov, V., Selberherr, S.: Reduction of momentum and spin relaxation rate in strained thin silicon films. In: Proceedings of the 43rd European Solid-State Device Research Conference (ESSDERC), pp. 334–337 (2013)

Neumann Series Analysis of the Wigner Equation Solution

I. Dimov, M. Nedjalkov, J.M. Sellier, and S. Selberherr

Abstract The existence and uniqueness of the electron transport Wigner equation solution, determined by boundary conditions, is analyzed in terms of the Neumann series expansion of the integral form of the equation, obtained with the help of Newton's trajectories. For understanding of the peculiarities of Wigner-quantum electron transport in semiconductor structures such mathematical issues can not be separated from the physical attributes of the solution. In the presented analysis these two sides of the problem mutually interplay.

The problem is first formulated from a physical point of view, where the stationary solution is considered as the long time limit of the general evolution problem posed by both initial and boundary conditions. The proof of convergence relies on the assumption for reasonable local conditions which may be specified for the kernel and on the fact that the Neumann series expansion corresponds to an integral equation of Volterra type with respect to the time variable.

Keywords Electron transport • Neumann series analysis • Semiconductor • Wigner equation

1 Introduction

The existence and uniqueness of the solution of the Wigner equation (WE) is subject of an active research interest [1–3] since the rising importance of a quantum description of the electron transport in the novel semiconductor nanoelectronics. An analysis of the regularity of the Wigner function and the existence and uniqueness of the solution of the by initial conditions posed evolution problem relevant for single-dimensional nanostructures is presented in [1] and used to proof the convergence of the suggested operator-splitting method. The analysis has been further augmented

I. Dimov (✉) • J.M. Sellier
IICT, Bulgarian Academy of Sciences, Acad. G. Bonchev 25 A, 1113 Sofia, Bulgaria
e-mail: ivdimov@bas.bg; jeanmichel.sellier@gmail.com

M. Nedjalkov • S. Selberherr
Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, 1040 Vienna, Austria
e-mail: mixi@iue.tuwien.ac.at; Selberherr@TUWien.ac.at

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_97

to account for the existence of boundary conditions characterizing the contacts of such structures. The well posedness of the transient problem, associated with time-dependent inflow boundary conditions has been shown in a mathematically rigorous way [2]. The integral form of the Wigner equation based on classical Newtonian trajectories for transient (posed by an initial condition, (IC)) and stationary (posed by boundary conditions, (BC)) problems has been used to investigate the corresponding Neumann expansion of the solution in connection with convergence proofs of the developed quantum Monte Carlo methods [4, 5]. In both cases the equation is of Volterra type with respect to the evolution time or the time to the boundary, so that the trajectory approach is straightforwardly generalized to the typical multidimensional structures of modern nanoelectronics. In a recent work [6] it has been shown that the stationary Wigner equation can be expressed as a Volterra type integral equation with respect to the spatial variable. It is argued that moving the boundaries arbitrary close, or imposing arbitrary inflow BCs on them, may lead to non-unique and unphysical solutions [6]. However, another recent work shows the well-posedness of the problem within the interval of periodicity $\Omega = [-l/2, l/2]$ of a certain class of periodic potentials, under arbitrary inflow BCs specified at $-l/2$ ($v > 0$) and $l/2$ ($v < 0$) [3]. Thus under certain physical settings the solution of the stationary Wigner equation is well defined by the boundary conditions, while in other circumstances the physical soundness of the problem becomes questionable. Alternatively stated, there circumstances where the stationary Wigner equation is of practical importance, while in other occasions the equation is of academic importance only.

Here we present an analysis, in which mathematical and physical aspects of the problem mutually interplay. This imposes a rather physical way of presentation with an accent on the application aspects of the results, on the expense of the mathematical rigor. The single-dimensional Wigner equation is considered, however the analysis holds for three-dimensional transport as in the case of classical transport [7, 8].

The problem is first formulated from a physical point of view, where the stationary solution is considered as the long time limit of the general evolution problem posed by both ICs and BCs. This implies the existence of a generic solution, determined partially by the ICs and partially by the BCs. The latter are interpreted in this scheme as a known part of the generic solution, which is complementary to the part corresponding to the IC's. If the contribution from the IC's does not vanish with time and only BCs are considered, the problem remains not well formulated already from a physical point of view. It follows that the time-dependent component of the field-less Liouville operator can not be neglected a priori, so that a restriction to the stationary WE can be relevant only after existence of physical arguments for that. As a matter of fact, physically relevant models in the Wigner formulation are the stargenvalue problem and the time-dependent Wigner equation, thoroughly discussed in [9–11].

These considerations are apart from the practical aspects of Wigner transport. Boundary conditions are known explicitly only in rare cases,. An exception is the equilibrium Wigner function, which is well known. Thus equilibrium conditions are routinely assumed at the boundaries. However, then the domain of the equation must be extended to infinity to avoid correlations with the non-equilibrium central region of the structure, where the electron flow occurs.

2 Integral Representation

The presented analysis is dimension-independent, so that for the sake of simplicity the single- dimensional formulation of the problem is considered. The equation for the Wigner function f reads:

$$\frac{\partial f(x,k,t)}{\partial t} + v(k)\frac{\partial f(x,k,t)}{\partial x} = \int dk' V_w(x, k-k') f(x,k',t), \tag{1}$$

where $v(k) = \hbar k/m$ and m are the electron velocity and effective mass, and V_w is the Wigner potential:

$$V_w(x,k) = \frac{1}{i\hbar 2\pi}\int ds e^{-iks}(V(x+s/2) - V(x-s/2)), \tag{2}$$

with $V(x)$ the electric potential of the structure determining the kernel of the equation. The differential component of (1) is given by the Liouville operator, whose characteristics are the field-less Newton trajectories.

$$x(t') = x - v(k)(t-t'); \qquad k(t') = k \tag{3}$$

The trajectory (3) is initialized by x, m, t and parameterized backwards in time by $t' < t$. An important property of Newton trajectories is that they do not cross in the phase space, so that (3) is uniquely determined by the initialization point.

The Liouville operator becomes a full time differential over given characteristics, so that it is now possible to rewrite Eq. (1) as a set of equations parametrized by t', which can be furthermore integrated on t' in the limits $[0, t]$, giving rise to:

$$\begin{aligned} f(x,k,t) = &\int_0^t dt' \int dk' V_w(x(t'), k(t') - k') f(x(t'), k', t) \\ &+ f_i(x(0), k(0))\theta_\Omega(x(0)) + f_b(x(t_b), k(t_b))\theta(t_b). \end{aligned} \tag{4}$$

Here the domain indicator θ_Ω is unity, if the argument belongs to the closed interval Ω, and is zero otherwise, t_b is the time needed for $x(t')$ to reach the boundary.

Finally (3) has been used to set $x(t) = x$, $k(t) = k$. The solution is sought in the interval Ω, where the initial condition (IC) $f_i(x,k)$ is known at time $t = 0$ and the BCs $f_b(-l/2,k,t)$, $k > 0$, $f_b(l/2,k,t)$, $k < 0$ are known at any time $t > 0$ (and zero at $t = 0$). Here we assume stationary physical conditions, in particular the BCs and the potential profile V are time independent. Furthermore boundaries, usually associated with certain physical interfaces, have now the meaning of points, where the function f, the unique solution of a generic evolution problem, is known.

3 Convergence

The second kind Fredholm integral equation (1) has a free term given by the IC and BCs. The solution can be presented as a Neumann series of the consecutive iteration of the kernel on the free term and is uniquely determined by the latter provided the series converges. The proof of the convergence relies on the fact that (1) is of Volterra type with respect to the variable t. This allows to rewrite the equation as

$$f(x,k,t) = \int_{t_0}^{t} dt' \int dk' V_w(x(t'),k(t')-k')f(x(t'),k',t) + f_1(x,k,t_0), \tag{5}$$

where itself the free term

$$f_1(x,k,t_0) = f(x(t_0),k(t_0),t_0) \tag{6}$$

of (5) satisfies Eq. (4) at $t_0 = t - \Delta t_1$, which is a time of the past with respect to the initialization time, $t > t_0$. Under the assumption that f_1 is known, reasonable local conditions may be specified for the kernel, in order to guarantee the convergence of the series. In [7] the necessary conditions for the convergence of such a kind of iterative expansion are given. These conditions concern the kernel of the equation V_w. We consider a typical condition for V_w after the following remark. Frequently authors use the term *mild conditions*. However, since one is interested in computational convergence, we also need to have a mild condition number. If the solution convergence is mild, then the solution can be confidently declared as non-singular. Since the convergence behavior and the condition number can be affected by poor scaling, the definition of *mild* is problem dependent. Simply speaking, mildness means confidence in the convergence to the true non-singular solution [12]. Now, one sufficient condition for convergence is the boundedness of the Wigner potential, $|V_w| < C$, where C is a constant. Indeed, if Δt_1 is small enough, the iterative terms have an upper limit given by the corresponding terms of a geometric

progression defined by

$$C\Delta t_1 < 1. \tag{7}$$

In this way the solution f of (5) is uniquely determined by the free term f_1.

The procedure can be repeated for f_1, which introduces the free term f_2 and so on, giving a decomposition of the backward evolution into the time intervals Δt_i. It is important to show that these intervals can cover the whole evolution interval, which ensures that the initial time is reached. The next estimation addresses this problem. By assuming that the Fourier transform $\tilde{V}$ of the electric potential V is bounded by a constant $\hbar C/4$ and using the definition (2) it may be shown that:

$$|V_w(x,k)| < C, \tag{8}$$

Thus, it is sufficient to request that the potential V is an absolutely integrable function, as the Fourier transform of such a function is bounded and continuous. The result (8) used in (7), shows the existence of an infimum of the set Δt_i, which can be used as a global decomposition time Δt.

Finally, this procedure links f to the free term in (4): the initial and the boundary conditions, which uniquely determine the solution of the equation.

4 Physical Analysis

The physical aspects of this proof may be associated to the Markovian character of the Wigner evolution. Furthermore we note that the solution has two complementary contributions from the IC and the BCs. In general, for small evolution times t the main contribution to the solution in an internal point of Ω is given by the IC. For large times (3) encounters the boundary, so that the BCs determine the solution. Moreover, since the trajectory evolves backward in time, the function f outside Ω contributes to the solution inside Ω by these values of k only, which guarantee the injecting character of f_b.

An important conclusion follows from this analysis: In the case when the initial condition 'leaks' through the boundaries: $f_i = 0$ after given time t_s, the electron system enters into a stationary regime and it is legitimate to consider the stationary equation as physically relevant. However, from a physical point of view it is clear that if there are electronic states which remain insulated away from the boundaries, they can not be controlled by the boundaries and the time dependent factor in the Wigner equation can not be neglected. Such are the bound eigenstates of the Hamiltonian related e.g. to periodic in time solutions or electrons with zero momenta. From a mathematical point of view, states which commute with the system Hamiltonian give rise to the 'bound state problem' for the von Neumann or Wigner equations [13]. The particular manifestation of this problem within the developed approach, is the fact that such states can not be

associated with trajectories which reach the boundaries: the wave vector of bounded states is undefined. In particular, zero momentum electrons are routinely neglected in the mathematical approaches. Indeed they have zero contribution to certain physical mean values like velocity, energy, and current, however, they affect the electrostatics.

The requirement for V to be an absolute integrable function is satisfied by a large class of potentials. Indeed the physical quantities are usually assumed to be smooth functions of their variables. In particular the existence of the first derivative, the electric force, guarantees the continuity of V almost everywhere, besides the fact that discontinuities are considered as convenient for the mathematical treatment of limiting cases. Furthermore one must assume that V approaches zero far away from the structure, which correctly accounts for the recovery of the equilibrium [14].

Finally almost everywhere continuous functions which become zero at infinity are absolutely integrable, showing that this condition does not restrict, but rather characterize the physically relevant potentials.

We conclude with a remark concerning the fact that boundaries are considered as a part of a global Wigner function. There are conditions for both pure and mixed states, which, if satisfied, allow to interpret a phase space function as a physically acceptable quasi-distribution, or Wigner function [13]. In this respect, an inconsistent change of the values at the boundaries will lead to unphysical results.

Acknowledgements This work has been supported by the EC FP7 Project AComIn (FP7-REGPOT-2012-2013-1), the Austrian Science Fund Project FWF-P21685-N22, as well as the Bulgarian Science Fund under grant DFNI I 02/20.

References

1. Arnold, A., Ringhofer, C.: An operator splitting method for the wigner' Poisson problem. SIAM J. Numer. Anal. **33**, 1622–1643 (1996)
2. Manzini, C., Barletti, L.: An analysis of the Wigner' Poisson problem with inflow boundary conditions. Nonlinear Anal. **60**, 77–100 (2005)
3. Li, R., Lu, T., Sun, Z.: Stationary Wigner equation with inflow boundary conditions: will a symmetric potential yield a symmetric solution? SIAM J. Appl. Math. **74**(3), 885–897 (2014)
4. Nedjalkov, M., Dimov, I., Rossi, F., Jacoboni, C.: Convergency of the Monte Carlo algorithm for the Wigner quantum transport equation. J. Math. Comput. Model. **23**, 159–166 (1996)
5. Nedjalkov, M., Kosina, H., Selberherr, S., Ringhofer, C., Ferry, D.: Unified particle approach to Wigner-Boltzmann transport in small semiconductor devices. Phys. Rev. B **70**, 115319–115335 (2004)
6. Rosati, R., Dolcini, F., Iotti, R.C., Rossi, F.: Wigner-function formalism applied to semiconductor quantum devices: failure of the conventional boundary condition scheme. Phys. Rev. B **88**, 035401 (2013)
7. Dimov, I.: Monte Carlo Methods for Applied Scientists, 291 pp. World Scientific, Singapore (2008)
8. Nedjalkov, M., Vasileska, D., Dimov, I., Arsov, G.: Mixed initial-boundary value problem in particle modeling of microelectronic devices. Monte Carlo Methods Appl. **13**, 299–331 (2007)

9. Moyal, J.E.: Quantum mechanics as a statistical theory. In: Proceedings of the Cambridge Philosophical Society, vol. 45, pp. 99–124 (1949)
10. Groenewold, H.J.: On the Principles of Elementary Quantum Mechanics. Doctoral thesis (1946)
11. Dias, N.C., Prata, J.N.: Admissible states in quantum phase space. Ann. Phys. **313**, 110–146 (2004)
12. Nashed, M., Wahba, G.: Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind. Math. Comput. **28**, 69–80 (1974)
13. Nedjalkov, M., Querlioz, D., Dollfus, P., Kosina, H.: Wigner function approach. In: Vasileska, D., Goodnick, S.M. (eds.) Nano-Electronic Devices: Semiclassical and Quantum Transport Modeling. Springer, Berlin (2011)
14. Nedjalkov, M., Selberherr, S., Ferry, D.K., Vasileska, D., Dollfus, P., Querlioz, D., Dimov, I., Schwaha, P.: Physical scales in the Wigner-Boltzmann equation. Ann. Phys. **328**, 220–237 (2013)

MS 29
MINISYMPOSIUM: SEMICLASSICAL AND QUANTUM TRANSPORT IN SEMICONDUCTORS AND LOW DIMENSIONAL MATERIALS

Organizers

Mariano Alvaro[1], Luis L. Bonilla[2], Orazio Muscato[3], and Vittorio Romano[4]

Speakers

Wolfgang Wagner[5]
Heat Generation in the Electrothermal Monte Carlo Method

Valentina Tozzini[6], Dario Camiola[7], Riccardo Farchioni[8], Antonio Rossi[9], Tommaso Cavallucci[10] and Vittorio Pellegrini[11]
Multi-Scale Simulations of Rippled Hydrogenated Graphene

[1]Mariano Alvaro, Universidad Carlos III de Madrid, Leganes, Spain.
[2]Luis Bonilla, Universidad Carlos III de Madrid, Leganes, Spain.
[3]Orazio Muscato, Università di Catania, Catania, Italy.
[4]Vittorio Romano, Università di Catania, Catania, Italy.
[5]Wolfgang Wagner, WIAS, Berlin, Germany.
[6]Valentina Tozzini, CNR-NEST Pisa, Italy.
[7]Dario Camiola, CNR-NEST Pisa, Italy.
[8]Riccardo Farchioni, CNR-NEST Pisa, Italy.
[9]Antonio Rossi, CNR-NEST Pisa, Italy.
[10]Tommaso Cavallucci, CNR-NEST Pisa, Italy.
[11]Vittorio Pellegrini, IIT Genova, Italy.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_98

Mariano Alvaro[1], Luis L. Bonilla[2] and Manuel Carretero[12]
Modulated Bloch Waves in Semiconductor Superlattices

Luis Bonilla[2], Mariano Alvaro[1] and Manuel Carretero[12]
Spontaneous Chaos, Random Number Generators and Electron Transport in Multiquantum Wells at Room Temperature

Manuel Carretero[12], Mariano Alvaro[1], and Luis Bonilla[2]
Characterizing Spontaneous Chaos in Semiconductor Multi-Quantum Wells at Room Temperature

Luigi Barletti[13]
Derivation of a Hydrodynamic Model for Electron Transport in Graphene via Entropy Maximization

Armando Majorana[14] and Vittorio Romano[4]
Deterministic Solutions of the Transport Equation for Charge Carrier in Graphene

Dario Camiola[7] and Vittorio Romano[4]
Hydrodynamic Model for Charge Transport in Graphene

Paola Pietra[15] and Clement Jourdana[16]
Modeling and Simulation of Electron Transport in a CNTFET

Davide Cagnoni[17], Marco Bellini[18], Jan Vobecky[19], Marco Restelli[20] and Carlo de Falco[21]
Mixed Mode TCAD for Large Scale Power Electronics Devices

[12]Manuel Carretero, Universidad Carlos III de Madrid, Leganes, Spain.
[13]Luigi Barletti, Università di Firenze, Florence, Italy.
[14]Armando Majorana, Università di Catania, Catania, Italy.
[15]Paola Pietra, Università di Pavia, Pavia, Italy.
[16]Clement Jourdana, UJF, Grenoble, France.
[17]Davide Cagnoni, MOX – Politecnico di Milano, Milano, Italy.
[18]Marco Bellini, ABB Corporate Research, Dättwil, Switzerland.
[19]Jan Vobecky, ABB Switzerland Ltd, Semiconductors, Lenzburg, Switzerland.
[20]Marco Restelli, NMPP – Numerische Methoden in der Plasmaphysik, Max-Planck-Institüt für Plasmaphysik, Garching, Germany.
[21]Carlo de Falco, MOX – Politecnico di Milano, Milano, Italy.

Keywords

Electron transport
Low dimensional materials
Semiconductor

Short Description

Appropriate descriptions of electron transport in semiconductor micro and nano devices are crucial for many current and future technologies able to generate great industrial and economic activity. This minisymposium will explore a wide variety of topics in modern electron transport including numerical and stochastic methods for electron transport equations, thermal effects, sub-band models, quantum dots, wires, wells and superlattices, and semiclassical and quantum transport in graphene and carbon nanotubes.

An Algorithm for Mixed-Mode 3D TCAD for Power Electronics Devices, and Application to Power *p-i-n* Diode

D. Cagnoni, M. Bellini, J. Vobecký, M. Restelli, and C. de Falco

Abstract Cutting edge semiconductor devices for power electronic applications, such as Phase Control Thyristors (PCTs) or Bimode Insulated Gate Transistor (BIGTs), present large area and complex 3D geometry, thus requiring full scale 3D models for their simulation. Moreover, sensitivity to temperature variations and complex loading conditions call for mixed mode simulation of distributed devices coupled to external controlling circuits. In this work, we describe a strategy for coupled simulation of 3D devices and lumped circuit networks, with particular emphasis on efficient iterative solution strategies for nonlinear equations. The algorithm presented is tested on a *p-i-n* power diode, for which quasi-static on-state and transient switching (reverse recovery) simulations are performed.

Keywords Power electronic devices • p-i-n power diode • Semiconductor

1 Introduction

Using numerical simulation of power devices is becoming more and more common in the semiconductor industry because of the need to analyze both performance [4, 8, 10] and failures [5, 9, 13, 16], while avoiding expensive prototyping and testing. The design of state of the art devices uses complex, three-dimensional

D. Cagnoni (✉) • C. de Falco
MOX - Politecnico di Milano, Milano, Italy
e-mail: davide.cagnoni@polimi.it; carlo.defalco@polimi.it

M. Bellini
ABB Corporate Research, Dättwil, Switzerland
e-mail: marco.bellini@ch.abb.com

J. Vobecký
ABB Switzerland Ltd, Semiconductors, Lenzburg, Switzerland
e-mail: jan.vobecky@ch.abb.com

M. Restelli
NMPP - Numerische Methoden in der Plasmaphysik, Max-Planck-Institüt für Plasmaphysik, Garching, Germany
e-mail: marco.restelli@ipp.mpg.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_99

structures and doping profiles in an attempt to minimize both on-state and switching losses [12] and to improve high-voltage and high-temperature operation capability [18]. Such complex devices require full 3D simulations [4, 8, 10]. Mainstream commercial 3D TCAD simulators [7, 14, 15] are focused on the accurate physical modeling of nanoscale CMOS devices, rather than on providing the computational efficiency required for the simulation of large devices with complex geometries. This study is part of an ongoing effort to develop a specialized simulator, targeting the peculiar needs of large scale TCAD in the power electronics industry. In addition to size and geometric complexities, challenges posed by the simulation of 3D power devices are due to extremely high doping densities, and by the wide range of temperatures under which correct operation has to be guaranteed, both of which must be taken into account when choosing material parameter models. Finally, as accurate lumped parameter models are often unavailable for power electronic devices, mixed mode simulation of devices coupled to both electrical [1, 2] and thermal [3, 6, 9] networks is often needed.

The focus of the present work is two-fold: on one hand we describe the algorithm being used in our simulator for time stepping and nonlinear iterations, while on the other hand we verify the accuracy of the temperature dependent coefficient models on a relevant benchmark problem. This work is structured as follows: Sect. 2 introduces the mathematical model and the employed algorithm, while in Sect. 3 simulation results for an irradiated power diode, both in quasi-static and fast transient switching regime, are presented to benchmark the simulator performance. The effect of adjusting carrier lifetimes to account for deep levels created via the irradiation process (see [3]) is discussed. In Sect. 4 conclusions are drawn and future perspectives discussed.

2 Equations for the Coupled Model

In order to introduce the system of Partial-Differential-Algebraic equations modeling a semiconductor device coupled with an external controlling circuit, let $\Omega \subset \mathbb{R}^3$ be the device domain, let $\Gamma^N \subset \partial\Omega$ denote insulated or artificial boundaries, and let $\Gamma_i^D \subset \partial\Omega$ denote each of the $\mathcal{N}_c$ contacts. Denote by F_i the voltage applied at the i-th contact. Assume that the lumped circuit network, modeled via Modified Nodal Analysis (MNA), has $\mathcal{N}_c$ pins, and that its state vector x consists of $\mathcal{N}_f$ degrees of freedom. Charge transport within the device is described by the following system of conservation laws

$$
\begin{aligned}
&-\nabla\cdot(\varepsilon\nabla\phi) + q\,(n-p-D) = 0\\
&\dot{n} + \nabla\cdot J_n + R(n,p,\nabla\phi) = 0;\ \dot{p} + \nabla\cdot J_p + R(n,p,\nabla\phi) = 0\\
&J_n = -\mu_n\,(V_{th}\nabla n - n\nabla\,(\phi+\phi_{\mathrm{BGN}}))\,;\ J_p = -\mu_p\,(V_{th}\nabla p + p\nabla\,(\phi-\phi_{\mathrm{BGN}}))\,;\ J = J_p - J_n
\end{aligned}
\tag{1}
$$

ϕ being the electrostatic potential, n and p electron and hole densities, and D the doping density. $R(n, p, \phi)$ is the net recombination rate, J_n, J_p are the electron and hole fluxes, μ_n and μ_p are the mobilities, V_{th} is the thermal voltage, and ϕ_{BGN} accounts for bandgap narrowing. The set od PDEs (1) is usually called *drift-diffusion approximation* and may be derived by moment expansion of Boltzmann transport equation. At the boundaries the following conditions are enforced:

$$\phi - F_i = \phi_{\text{built-in}}(D); \quad D + p - n = 0; \quad pn - n_i^2 = 0 \quad \text{on } \Gamma_i^D \tag{2}$$

$$\nabla\phi \cdot \nu = 0; \quad J_n \cdot \nu = 0; \quad J_p \cdot \nu = 0 \quad \text{on } \Gamma^N \tag{3}$$

ν being the outward normal unit vector. The device-circuit coupling conditions read

$$A\dot{x} + C(x) + rI = 0; \quad F = r^T x; \quad I_i = \int_{\Gamma_i^D} (-\partial(\epsilon\nabla\phi)/\partial t + qJ) \cdot \nu_i^D \mathrm{d}\Gamma, \tag{4}$$

where the matrix A and the vector valued nonlinear function C may be assembled via standard MNA modelling of the controlling circuit network, while $I = [I_i]^T$, $i = 1 \ldots \mathcal{N}_c$ and $F = [F_i]^T$, $i = 1 \ldots \mathcal{N}_c$ are vectors of device currents and corresponding voltages, respectively; the incidence matrix r accounts for attaching each device contact to a circuit node by adding the corresponding current to the correct KCL equation in the MNA system. To solve the system comprised of (1)–(4) a suitable numerical method consisting of a mix of different ingredients has to be employed. For spatial discretization of a strategy based on the OSC [11] algorithm and on the well known exponential fitting stabilization method is employed, in order to guarantee a discrete maximum principle with weak regularity requirements on the 3D mesh. For time discretization an implicit Euler method is then used, so that the problem of solving (1)–(4) is reduced to that of solving a sequence of nonlinear-algebraic problems, one for each time step t, of the form:

$$\mathbf{G}^{(t)}(S^{(t)}) = 0; \tag{5}$$

Where $S^{(t)} = [\boldsymbol{\phi}^{(t)}, \mathbf{n}^{(t)}, \mathbf{p}^{(t)}, \mathbf{F}^{(t)}, \mathbf{I}^{(t)}]^T$ denotes the system state vector at time t and $\mathbf{G}^{(t)}(S^{(t)}) = [G_\phi^{(t)}(S^{(t)}), G_n^{(t)}(S^{(t)}), G_p^{(t)}(S^{(t)}), G_F^{(t)}(S^{(t)}), G_I^{(t)}(S^{(t)})]^T$.

In order to initialize the quasi-Newton algorithm for the nonlinear problem solution, an initial guess is constructed by extrapolating data at previous time steps. Such extrapolated guess is also used for truncation error estimation, and for time step control. As a termination criterion for quasi-Newton iterations, we consider the convergence reached if either the residual $G(S)$ or the increment $\mathrm{d}S$ satisfy:

$$\|G_u(S)\|_\infty < \sigma_u \tau_u \quad \text{or} \quad \|\mathrm{d}u\|_\infty < (\|u\|_\infty + \rho_u)\theta_u$$

where $u = \phi, F, I$ represents the generic field in the state vector, τ_u, θ_u are the respective tolerances, while σ_u, ρ_u are suitably chosen reference values. For $u = n, p$

a slightly different criterion is employed, namely

$$\|G_u(S)\|_\infty < \sigma_u \tau_u \quad \text{or} \quad \|\mathrm{d}V_{th} \log(u/n_i)\|_\infty < (\|V_{th} \log(u/n_i)\|_\infty + \rho_u)\theta_u$$

so that all reference values are dimensionally consistent. The outline of the solution algorithm is provided in Algorithm 1.

Algorithm 1 Outline of the time-adaptive implicit Euler modified quasi-Newton solution algorithm

procedure CGDD($S_J^{(0)}, S_J^{(-1)}, S_J^{(-2)}, t^{(0)}, t^{(-1)}, t^{(-2)}, \mathrm{d}t^{(1)}$)
  $k \leftarrow 1$
  repeat ▷ Time loop
    $j \leftarrow 0, t^{(k)} \leftarrow t^{(k-1)} + \mathrm{d}t^{(k)}$
    $S_0^{(k)} \leftarrow$ EXTRAPOLATE($S_J^{(k-3)}, S_J^{(k-2)}, S_J^{(k-1)}, \mathrm{d}t^{(k)}$)
    while $j < J_{\max}$ **do** ▷ Fix maximum qN iterations
      $C \leftarrow$ COEFFICIENTS_UPDATE($S_j^{(k)}$)
      $G_{j+1} \leftarrow$ RESIDUAL_UPDATE($S_j^{(k)}, S_J^{(k-1)}, C, \mathrm{d}t^{(k)}, t^{(k)}$)
      $J \leftarrow$ JACOBIAN_UPDATE($S_j^{(k)}, C, \mathrm{d}t^{(k)}, t^{(k)}$)
      $\mathrm{d}S_{j+1} \leftarrow$ SOLVE_LINEAR_SYSTEM($J, -G_{j+1}$)
      $\alpha =$ COMPUTE_DAMPING_FACTOR($S_j^{(k)}, \mathrm{d}S_{j+1}$) ; $S_{j+1}^{(k)} \leftarrow S_j^{(k)} + \alpha \mathrm{d}S_{j+1}$
      if $\|f(S_0^{(k)} - S_{j+1}^{(k)})\|_\infty > \Delta$ or ($\|G_{j+1}\|_\infty > \|G_{j-a}\|_\infty$ and $\|\mathrm{d}S_{j+1}\|_\infty > \|\mathrm{d}S_{j-a}\|_\infty$) **then**
        DECREASE_DT($\mathrm{d}t^{(k)}$), BREAK_Q-N()
      else if $\|G_{j+1}\|_\infty < (\|S_{j+1}^{(k)}\|_\infty + \sigma)\tau$ or $\|\mathrm{d}f(S_{j+1}^{(k)})\|_\infty < (\|f(S_{j+1}^{(k)})\|_\infty + \rho)\theta$ **then**
        $S_J^{(k)} \leftarrow S_{j+1}^{(k)}$, $\mathrm{d}t^{k+1} \leftarrow$ ESTIMATE_DT($S_0^{(k)}, S_J^{(k)}$), $k \leftarrow k+1$, BREAK_Q-N()
      end if
      $l \leftarrow 0, S_l \leftarrow S_{j+1}^{(k)}$
      while $j < J_{\max}^M$ **do** ▷ Modified qN iterations
        $C \leftarrow$ COEFFICIENTS_UPDATE(S_l)
        $G_l \leftarrow$ RESIDUAL_UPDATE($S_l, S_J^{k-1}, C, \mathrm{d}t^{(k)}, t^{(k)}$)
        $\mathrm{d}S_{l+1} \leftarrow$ SOLVE_LINEAR_SYSTEM($J, -G_l$)
        $S_{l+1} \leftarrow S_{l+1} + \mathrm{d}S_l$
        if $\|G_j\|_\infty > \|G_{j-a}\|_\infty$ or $\|\mathrm{d}S_j\|_\infty > \|\mathrm{d}S_{j-a}\|_\infty$ **then**
          BREAK_MODIFIED_QUASI-NEWTON()
        else if $\|G_l\|_\infty < (\|S_l\|_\infty + \sigma)\tau$ or $\|\mathrm{d}f(S_l)\|_\infty < (\|f(S_l)\|_\infty + \rho)\theta$ **then**
          $S_J^{(k)} \leftarrow S_l$, $\mathrm{d}t^{k+1} \leftarrow$ ESTIMATE_DT($S_0^{(k)}, S_J^{(k)}$), $k \leftarrow k+1$, BREAK_Q-N()
        end if
        $l \leftarrow l + 1$ ▷ Next modified quasi-Newton step
      end while
      $j \leftarrow j + 1$; $S_{j+1}^{(k)} \leftarrow S_l$ ▷ Accept modified Newton result
    end while
  until $t^{(k)} < T_{\max}$ ▷ Fix time range
end procedure

3 Benchmark

As a benchmark test case, we consider the power diode studied in [3]. Such diodes are irradiated with 1–5 MeV electrons at a dose between 5 and 20 kGy and 5–12 MeV He at doses ranging between 10^{10}–10^{11} cm^{-2} and annealed at a temperature below 300 °C. In these conditions the dominant deep levels are the vacancy-oxygen pair (V-O) at $\simeq$ EC − 0.16 eV and the divacancy (V-V) at $\simeq$ EC − 0.42 eV. As a result, an accurate modeling of the generation-recombination processes via these deep levels is necessary, in order to precisely reproduce the reverse recovery characteristics of the diode. Complete deep levels models are computationally expensive and degrade convergence; thus, an effective carrier lifetime profile was obtained via optimization with a commercial simulator [15], and introduced within the conventional SRH framework. The schematic of testing circuit used for reverse recovery measurements is shown in Fig. 1. The inductance is tuned to match the $^{dI}/_{dt}$ of the measurements. The simulations are performed over a wide temperature range (300–413 K), and the switch is modeled as a time varying resistor, with the conductance ramping smoothly from 10^{-3} to 10^{3} S in 10 µs (the time derivative of conductance is continuous). Figure 2 shows the computed discharge profiles. The effect of lifetime controlling results in a prolonged and increased discharge of the power diode, at all temperatures, due to an increased charge buildup.

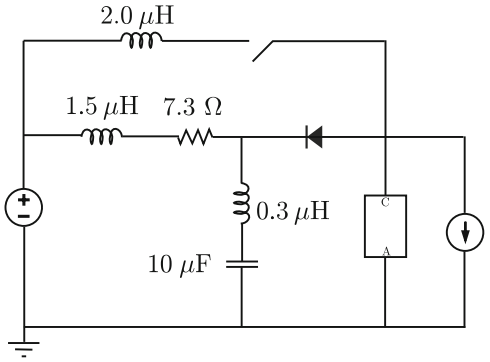


Fig. 1 *Up*: schematic structure of the simulated circuit. *Down*: the diode and switch models used for the nonlinear circuit simulation

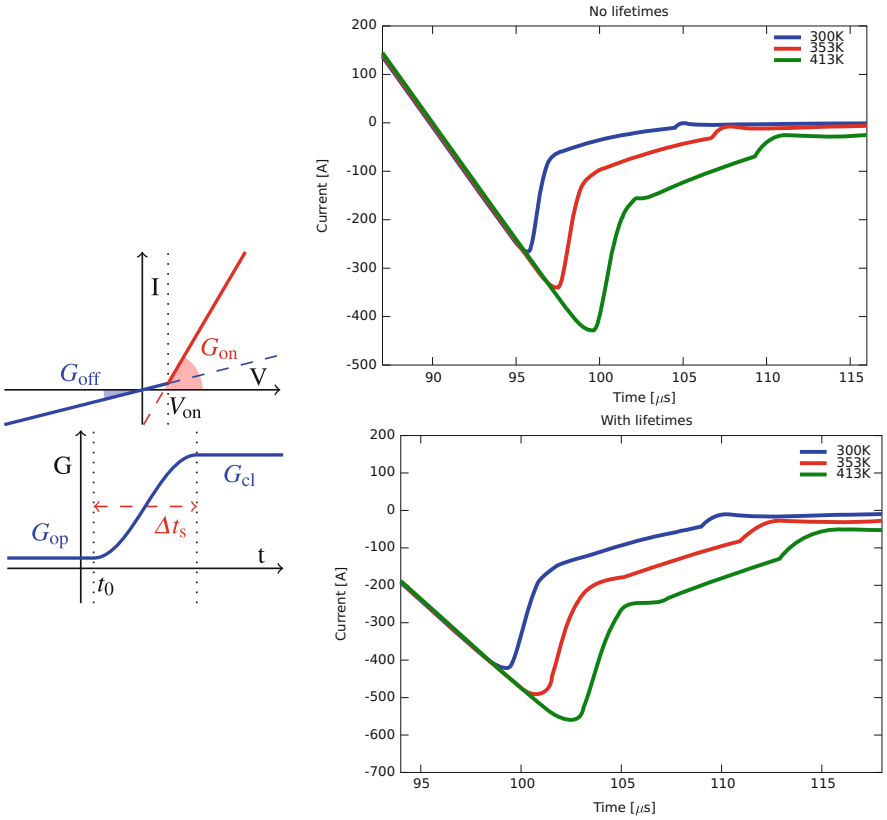


Fig. 2 Reverse recovery characteristics at 27, 80, and 140 °C. *Up*, without lifetime optimization. *Down*, with optimized lifetimes

Figure 3 shows a detailed view of the computed forward IV characteristic, both with and without the computed lifetimes. The importance of introducing the optimized lifetimes is particularly evidenced in high-injection regime, where the crossing of characteristic curve typical of irradiated devices is correctly reproduced by the optimized carrier lifetimes. Low injection regime characteristics, visible in the log-scale graphs, do not present substantial differences.

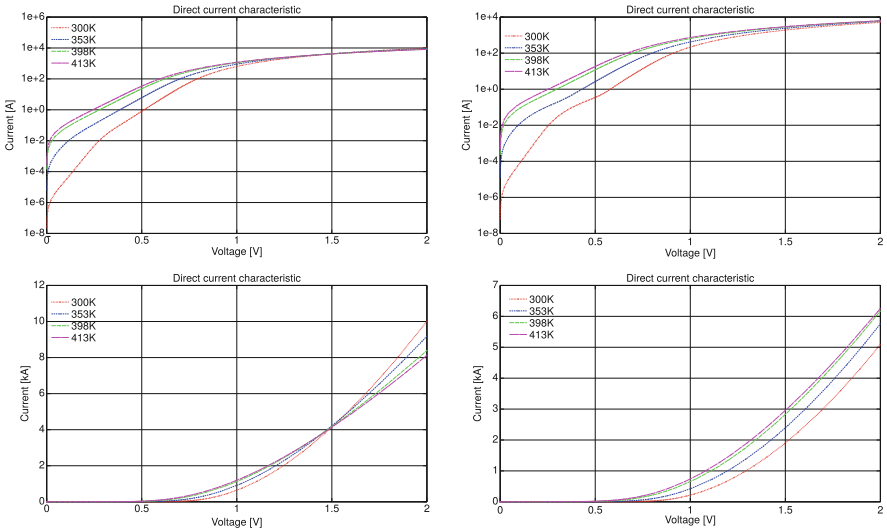


Fig. 3 Direct current characteristic in 0–2 V bias range. *Left*, locally optimized lifetimes; *right*, standard doping-dependent model [17]; *top*, logarithmic scale, *bottom*, linear scale

4 Conclusions and Perspectives

In this work, a complete strategy for the coupled simulation of 3D semiconductor devices and lumped circuit elements has been presented. Even being as simple as reasonably possible, the simulated model features all the characteristics also present in more complex models: steep variations in spatial input data as well as in the solution, complex physical models of nonlinear effects, abruptly fast transients are successfully dealt with. The *p-i-n* power diode benchmark results demonstrate the proposed algorithm's ability to correctly reproduce the device behavior over a wide range of operation temperature and conditions.

Striving for a successful use of TCAD in the simulation of more geometrically and operationally complex power electronic devices, in turn leading to bigger and stiffer problems, future goals include: the intertwining of the linear and nonlinear solvers in Newton-Krylov-like methods, to increase solver efficiency, the application of domain-decomposition to enable parallel solution, the introduction of the heat equation and a thermal circuit, to relax the assumption on temperature uniformity and achieve more precise simulation results.

Acknowledgements Carlo de Falco's work was partially funded by the "Start-up Packages and PhD Program project", co-funded by Regione Lombardia through the "Fondo per lo sviluppo e la coesione 2007–2013", formerly FAS program.

References

1. Alí, G., Bartel, A., Culpo, M., de Falco, C.: Analysis of a PDE thermal element model for electrothermal circuit simulation. In: Roos, J., Costa, L.R. (eds.) Scientific Computing in Electrical Engineering SCEE 2008. Mathematics in Industry 2010, pp. 273–280. Springer, Berlin (2010)
2. Alí, G., Bartel, A., Brunk, M., Schöps, S.: A convergent iteration scheme for semiconductor/circuit coupled problems. In: Michielsen, B., Poirier, J.R. (eds.) Scientific Computing in Electrical Engineering SCEE 2010. Mathematics in Industry 2012, pp. 233–242. Springer, Berlin (2011)
3. Bellini, M., Vobecky, J.: TCAD simulations of irradiated power diodes over a wide temperature range. In: 2011 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), pp. 183–186. IEEE, New York (2011)
4. Bellini, M., Vobecky, J.: Large-scale 3D TCAD study of the impact of shorts in phase controlled thyristors. In: 2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), pp. 265–268. IEEE, New York (2014)
5. Breglio, G., Irace, A., Napoli, E., Riccio, M., Spirito, P.: Experimental detection and numerical validation of different failure mechanisms in IGBTs during unclamped inductive switching. IEEE Trans. Electron Devices **60**(2), 563–570 (2013)
6. Castellazzi, A., Funaki, T., Kimoto, T., Hikihara, T.: Thermal instability effects in SiC power MOSFETs. Microelectron. Reliab. **52**(9), 2414–2419 (2012)
7. Cogenda Pte, Ltd.: Genius User's Guide, Version 1.74 (2010)
8. Lophitis, N., Antoniou, M., Udrea, F., Nistor, I., Arnold, M., Wikstrom, T., Vobecky, J.: Experimentally validated three dimensional GCT wafer level simulations. In: 2012 24th International Symposium on Power Semiconductor Devices and ICs (ISPSD), pp. 349–352. IEEE, New York (2012)
9. Morand, S., Miller, F., Austin, P., Poirot, P., Gaillard, R., Carriere, T., Buard, N.: Temperature effects on power MOSFET and IGBT sensitivities toward single events. In: 2011 12th European Conference on Radiation and Its Effects on Components and Systems (RADECS), pp. 109–114. IEEE, New York (2011)
10. Phung, L.V., Planson, D., Brosselard, P., Tournier, D., Brylinski, C.: 3D TCAD simulations for more efficient SiC power devices design. ECS Trans. **58**(4), 331–339 (2013)
11. Putti, M., Cordes, C.: Finite element approximation of the diffusion operator on tetrahedra. SIAM J. Sci. Comput. **19**(4), 1154–1168 (1998)
12. Rose, M., Krupar, J., Hauswald, H.: Adaptive dv/dt and di/dt control for isolated gate power devices. In: 2010 IEEE Energy Conversion Congress and Exposition (ECCE), pp. 927–934. IEEE, New York (2010)
13. Sandow, C., Baburske, R., Niedernostheide, F.J., Pfirsch, F., Töchterle, C.: Exploring the limits of the safe operation area of power semiconductor devices. In: 2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), pp. 49–52. IEEE, New York (2014)
14. Silvaco, Inc: Atlas. (2014), www.silvaco.com/products/tcad
15. Synopsys, Inc: Sentaurus device user guide, Version D-2010.03 (2010)
16. Töchterle, C., Pfirsch, F., Sandow, C., Wachutka, G.: Analysis of the latch-up process and current filamentation in high-voltage trench-IGBT cell arrays. In: 2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), pp. 296–299. IEEE, New York (2013)
17. Tyagi, M.S., Van Overstraeten, R.: Minority carrier recombination in heavily-doped silicon. Solid State Electron. **26**(6), 577–597 (1983)
18. Wang, J., Sung, W., Liu, Y., Baliga, B.J.: Smart grid technologies. IEEE Ind. Electron. Mag. **3**(2), 16–23 (2009)

An Electro-Thermal Hydrodynamical Model for Charge Transport in Graphene

V. Dario Camiola, Giovanni Mascali, and Vittorio Romano

Abstract A hydrodynamical model for the charge and the heat transport in graphene is presented. The state variables are moments of the electron, hole and phonon distribution functions, and their evolution equations are derived from the respective Boltzmann equations by integration. The closure of the system is obtained by means of the maximum entropy principle and all the main scattering mechanisms are taken into account. Numerical simulations are presented in the case of a suspended graphene monolayer.

Keywords Charge transport • Electro-thermal hydrodynamical model • Maximum entropy principle

1 Introduction

Graphene is among the most promising materials for future applications in nanoelectronics devices. It is two dimensional and consists of a single layer of carbon atoms arranged into a honeycomb hexagonal lattice. Graphene has very good mechanical properties and is an excellent heat and electricity conductor. In order to formulate comprehensive transport models it is necessary to take into account the electronic and phonon bandstructure and the most relevant scattering mechanisms

V.D. Camiola
NEST, Istituto Nanoscienze - CNR, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56127 Pisa, Italy
e-mail: dario.camiola@nano.cnr.it

G. Mascali
Department of Mathematics and Computer Science, University of Calabria and INFN-Gruppo c., 87036 Cosenza, Rende, Italy
e-mail: giovanni.mascali@unical.it

V. Romano (✉)
Dipartimento di Matematica e Informatica, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy
e-mail: romano@dmi.unict.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_100

between electrons and phonons. The case of a suspended sheet of graphene is considered here.

2 Kinetic Description

Electrons which contribute to the charge transport in graphene are those in the conduction and valence band, and it is preferable to treat the latter as holes for insuring integrability of the distribution function. Electrons and holes mostly populate the states near to the K and K' Dirac points situated at the boundary of the first hexagonal Brillouin zone, the respective neighborhoods being called K and K' valleys. In these valleys, the energies ϵ_i, $i = e, h$, (e and h respectively stay for electrons and holes) are, with a good approximation, linear in the wave vector $\mathbf{k}$: $\epsilon_i = \hbar v_F |\mathbf{k}|$, $\mathbf{k} \in \mathbb{R}^2$, $i = e, h$, $\hbar$ being the reduced Planck constant, and v_F the Fermi velocity. K and K' valleys will be treated as equivalent.

A semiclassical kinetic description of the charge transport in graphene is based on the two Boltzmann equations for electrons and holes (approaches which make use of the Wigner transport equations are also present in the literature, for example, see [1])

$$\frac{\partial f_i}{\partial t} + \mathbf{v}_i \cdot \nabla_{\mathbf{r}} f_i + \frac{\mathsf{e}_i}{\hbar} \mathbf{E} \cdot \nabla_{\mathbf{k}} f_i = \mathscr{C}_i, \quad i = e, h, \tag{1}$$

where $f_i(\mathbf{r}, \mathbf{k}, t)$, $i = e, h$, represent the state occupation numbers of electrons and holes at position $\mathbf{r}$, time t and with wave-vector $\mathbf{k}$. $\nabla_{\mathbf{r}}$ and $\nabla_{\mathbf{k}}$ are the gradients with respect to the position and the wave vector respectively, e_i, $i = e, h$, are the particle charges (negative for electrons and positive for holes), and $\mathbf{E}$ is the electric field obtained by the Poisson equation, which must be coupled with the above system. The group velocity $\mathbf{v}$ is related to the band energy by $\mathbf{v} = \frac{1}{\hbar} \nabla_{\mathbf{k}} \varepsilon_i = v_F \frac{\mathbf{k}}{|\mathbf{k}|}$. $\mathscr{C}_i$, $i = e, h$, are the scattering operators representing both the intra and inter-band interactions of electrons and holes with acoustic and optical phonons. Its complete expression is rather involved, here, for simplicity, we report only the generic contribution relative to the intra-conduction band scattering and refer the interested readers to [2, 3]

$$\mathscr{C}_e(\mathbf{k}) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \Big[\underbrace{w_{ee}(\mathbf{k}', \mathbf{k}) f_e(\mathbf{k}') \left(1 - f_e(\mathbf{k})\right)}_{gain} - \underbrace{w_{ee}(\mathbf{k}, \mathbf{k}') f_e(\mathbf{k}) \left(1 - f_e(\mathbf{k}')\right)}_{loss} \Big] d\mathbf{k}',$$

where $w_{ee}(\mathbf{k}', \mathbf{k})$ is the transition rate from the state $\mathbf{k}$ to the state $\mathbf{k}'$. In this case, the detailed balance principle implies $w_{ee}(\mathbf{k}, \mathbf{k}') = e^{(\varepsilon - \varepsilon')/k_B T} w_{ee}(\mathbf{k}', \mathbf{k})$, with k_B the Boltzmann constant and T the lattice temperature.

We consider interactions with acoustic phonons, longitudinal optical phonons (Γ-LO), transversal optical phonons (Γ-TO), and K-phonons.

In the elastic approximation, the production term relative to acoustic phonon (intraband) transitions simplifies into

$$\mathscr{C}_i(\mathbf{k}) = \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} A^{(ac)}(1+\cos\theta'')\delta(\varepsilon_i'-\varepsilon_i)\left(f_i(\mathbf{k}') - f_i(\mathbf{k})\right)\, d\mathbf{k}'.$$

where $A^{(ac)}$ can be found in [4–6] and θ'' is the angle between $\mathbf{k}$ and $\mathbf{k}'$.

For the optical and the K-phonons, in the Einstein approximation ($\hbar\omega = \text{cost}$, with ω phonon frequency), one has

$$w_{ee}^s(\mathbf{k}',\mathbf{k}) = s_{ee}^s(\mathbf{k}',\mathbf{k})\left[\underbrace{\left(g_s^- + 1\right)\delta\left(\varepsilon_e - \varepsilon_e' + \hbar\omega_s\right)}_{emission} + \underbrace{g_s^+\delta\left(\varepsilon_e - \varepsilon_e' - \hbar\omega_s\right)}_{absorption}\right],$$
$$s = LO, TO, K,$$

with $s_{ee}^K(\mathbf{k}',\mathbf{k}) = A_K D_K^2\,(1-\cos\theta'')$ for the K-phonons and $s_{ee}^\Gamma(\mathbf{k}',\mathbf{k}) = A_\Gamma D_\Gamma^2\big(1 \mp \cos(\theta+\theta')\big)$ respectively for the LO and TO phonons. A^K and A^Γ can be found in [5, 6], and θ and θ' respectively denote the angle between $\mathbf{k}$ and $\mathbf{k}'-\mathbf{k}$ and that between $\mathbf{k}'$ and $\mathbf{k}'-\mathbf{k}$.

If phonons are considered as a thermal bath at the constant temperature T_L

$$g_s^\pm \approx \left[e^{\hbar\omega_s/k_B T_L} - 1\right]^{-1}, \quad \text{equilibrium Bose-Einstein,}$$

otherwise $g_s^\pm = g_s(\mathbf{r}, t, \mathbf{q}^\pm)$, with the phonon wave vector given by $\mathbf{q}^\pm = \pm(\mathbf{k}'-\mathbf{k})$, in agreement with the momentum conservation.

Moreover, if we consider the phonon dynamics, the evolution of the phonon occupation number is governed by the following Boltzmann equations

$$\frac{\partial g_s}{\partial t} + \underbrace{\nabla_{\mathbf{q}}\omega_s(\mathbf{q})}_{\approx 0}\cdot\nabla_{\mathbf{r}} g_s = \mathscr{C}_s, \quad s = LO, TO, K,$$

$$\frac{\partial g_{ac}}{\partial t} + \nabla_{\mathbf{q}}\omega_{ac}(\mathbf{q})\cdot\nabla_{\mathbf{r}} g_{ac} = \mathscr{C}_{ac},$$

$$\mathscr{C}_s = -\frac{\left(g_s - g_s^0\right)}{\tau_{OA}} + \sum_{ij}\mathscr{C}_s^{ij},\ i,j = e,h,\ s = LO, TO, K,$$

$$\mathscr{C}_{ac} = -\frac{3}{\tau_{OA}}\left(g_{ac} - g_{ac}^0\right) + \sum_i \mathscr{C}_{ac}^i, \quad i = e,h,$$

where τ_{OA} is the relaxation time for the decay of an optical phonon into two acoustic phonons, and g_s^0, $s = LO,\ TO,\ K$, g_{ac}^0 are the equilibrium occupation number of the optical and acoustic phonons corresponding to the temperature they would have if

they were at the local equilibrium relative to their total average energy [7]. In the acoustic phonon scattering, normal and umpklapp types of intra-mode interactions as well as interactions with defects/impurities should also be considered.

Direct simulations based on the above-written semiclassical kinetic equations have been performed by MC methods, see e.g. [5], or with suitable numerical schemes [6], but they, even if very accurate, are too heavy from a computational point of view. Therefore models based on integrated quantities are preferable for computer aided design (CAD) purposes in view of a possible use of graphene in electron devices like MOSFETs or DG-MOSFETs.

3 Carrier Moment Equations

Macroscopic quantities can be defined as moments of the distribution functions with respect to some suitable weight functions $\psi(\mathbf{k})$, assuming a sufficient regularity for the existence of the involved integrals. In particular for electrons and holes we propose a set of moment equations consisting of the balance equations of the quantities ($i = e, h$)

$$\text{average density} \quad \rho_i = \frac{4}{(2\pi)^2}\int_{\mathbb{R}^2} f_i(\mathbf{r},\mathbf{k},t)\, d\mathbf{k},$$

$$\text{average velocity} \quad \rho_i \mathbf{V}_i = \frac{4}{(2\pi)^2}\int_{\mathbb{R}^2} f_i(\mathbf{r},\mathbf{k},t)\, \mathbf{v}\, d\mathbf{k},$$

$$\text{average energy} \quad \rho_i W_i = \frac{4}{(2\pi)^2}\int_{\mathbb{R}^2} f_i(\mathbf{r},\mathbf{k},t)\, \varepsilon\, d\mathbf{k},$$

$$\text{average energy-flux} \quad \rho_i \mathbf{S}_i = \frac{4}{(2\pi)^2}\int_{\mathbb{R}^2} f_i(\mathbf{r},\mathbf{k},t)\, \varepsilon \mathbf{v}\, d\,\mathbf{k},$$

where the factor 4 arises from taking into account both the spin states and the two equivalent valleys.

By integrating the Boltzmann equations with respect to $\mathbf{k}$, one has the following balance equations for the above-defined macroscopic quantities

$$\frac{\partial}{\partial t}\rho_i + \nabla_{\mathbf{r}} \cdot (\rho_i\ \mathbf{V}_i) = \rho_i\, C_i,$$

$$\frac{\partial}{\partial t}(\rho_i\, \mathbf{V}_i) + \nabla_{\mathbf{r}} \cdot \left(\rho_i\, \mathbf{F}_i^{(0)}\right) - \mathsf{e}_i\, \rho_i \mathbf{G}_i^{(0)} \cdot \mathbf{E} = \rho_i C_{\mathbf{V}_i},$$

$$\frac{\partial}{\partial t}(\rho_i W_i) + \nabla_{\mathbf{r}} \cdot (\rho_i\, \mathbf{S}_i) - \mathsf{e}_i \rho_i \mathbf{E} \cdot \mathbf{V}_i = \rho_i C_{W_i},$$

$$\frac{\partial}{\partial t}(\rho_i\, \mathbf{S}_i) + \nabla_{\mathbf{r}} \cdot \left(\rho_i \mathbf{F}_i^{(1)}\right) - \mathsf{e}_i \rho_i \mathbf{G}_i^{(1)} \cdot \mathbf{E} = \rho_i C_{\mathbf{S}_i},$$

where the G's and F's are extra-fluxes and the terms at the right hand sides are productions [3].

4 The Phonon Moment System

Similarly for each type of phonons we have

$$\frac{\partial}{\partial t}W_p + \nabla_{\mathbf{r}} \cdot \mathbf{Q}_p = C_{W_p}, \text{ energy balance equation,}$$
$$\frac{\partial}{\partial t}\mathbf{Q}_p + \nabla_{\mathbf{r}} \cdot \mathbf{T}_p = C_{\mathbf{Q}_p}, \text{ energy-flux balance equation,}$$

where for each phonon mode

$$W_p = \int_{\mathscr{B}} \hbar\omega_p g_p \, d\mathbf{q}, \quad \text{average energy,}$$
$$\mathbf{Q}_p = \int_{\mathscr{B}} \hbar\omega_p \mathbf{v}_p g_p \, d\mathbf{q}, \quad \text{average energy-flux,}$$

$\mathscr{B}$ is the hexagonal Brillouin zone and $p = LO, TO, K, ac$, the $\mathbf{T}$'s are extra-fluxes, and the terms at the right hand sides are productions [3]. More general moment systems can be considered, that we consider here is the minimal one for a reasonable description of the thermo-electrical effects.

5 The Closure Problem

The extra fluxes and the production terms are additional unknown quantities. For them constitutive relations in terms of the fundamental variables are needed in order to get a closed system of balance equations. A well theoretically founded way to get the desired closure relations is to resort to the Maximum Entropy Principle (MEP) [8], according to which the electron, hole and phonon distribution functions can be estimated by the distributions $f_{e,MEP}, f_{h,MEP}$, $g_{p,MEP}$ which solve the following problem:

$$(f_{e,MEP}, f_{h,MEP}, g_{p,MEP}) = \max_{f_e, f_h, g_p \in L^1(\mathbb{R}^2)} S[f_e, f_h, g_p],$$

under the constraints

$$\begin{pmatrix} \rho_i \\ \rho_i W_i \end{pmatrix} = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix} f_i(\mathbf{r}, \mathbf{k}, t)\, d\mathbf{k},$$

$$\begin{pmatrix} \rho_i \mathbf{V}_i \\ \rho_i \mathbf{S}_i \end{pmatrix} = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f_i(\mathbf{r}, \mathbf{k}, t) \begin{pmatrix} \mathbf{v} \\ \varepsilon \mathbf{v} \end{pmatrix} d\mathbf{k},$$

$$W_p = \int_{\mathbb{R}^2} \hbar \omega_p g_p \, d\mathbf{q}, \quad \mathbf{Q}_p = \int_{\mathbb{R}^2} \hbar \omega_p \mathbf{v}_p g_p \, d\mathbf{q},$$

where $S[f_e, f_h, g_p]$ is the total entropy of the system given by

$$-k_B \left\{ \frac{4}{(2\pi)^2} \int_{\mathbb{R}^2} \left[f^e \ln f^e + (1 - f^e) \ln (1 - f^e) \right] d\mathbf{k} + \frac{4}{(2\pi)^2} \int_{\mathbb{R}^2} \left[f^h \ln f^h + \right.\right.$$

$$\left.\left. \left(1 - f^h\right) \ln \left(1 - f^h\right) \right] d\mathbf{k} + \sum_p y^p \int_{\mathscr{B}} \left(1 + \frac{g^p}{y^p}\right) \ln \left(1 + \frac{g^p}{y^p}\right) \right] d\mathbf{q} \right\},$$

y^p being the phonon densities of states and $L^1(\mathbb{R}^2)$ the usual Banach space.

By solving the above maximization problem we get

$$f_i = \frac{1}{1 + \exp\left(\lambda_i + \lambda_{W_i} \varepsilon_i + \mathbf{v}_i \cdot (\lambda_{\mathbf{V}_i} + \varepsilon_i \lambda_{\mathbf{S}_i})\right)}, \; g_p = \frac{1}{\exp\left(\lambda_{W_p} \varepsilon_p + \varepsilon_p \, \mathbf{v}_p \cdot \lambda_{\mathbf{Q}_p}\right) - 1}.$$

As in [9–12] we linearize the distributions around their anisotropic part, obtaining

$$f_i \approx \frac{1}{e^{\lambda_i + \lambda_{W_i} \varepsilon_i} + 1} \left[1 - \frac{e^{\lambda_i + \lambda_{W_i} \varepsilon_i}}{e^{\lambda_i + \lambda_{W_i} \varepsilon_i} - 1} \mathbf{v}_i \cdot (\lambda_{\mathbf{V}_i} + \varepsilon_i \lambda_{\mathbf{S}_i}) \right], \; i = e, h,$$

$$g_p \approx \frac{1}{e^{\lambda_{W_p} \varepsilon_p} - 1} \left[1 - \frac{e^{\lambda_{W_p} \varepsilon_p}}{e^{\lambda_{W_p} \varepsilon_p} - 1} \epsilon_p \, \mathbf{v}_p \cdot \lambda_{\mathbf{Q}_p} \right], \; p = LO, TO, K, ac,$$

where the λ's are Lagrange multipliers which have to be expressed as functions of the state variables by taking into account the constraints.

After that, these distributions are inserted into the kinetic definitions of the additional variables, so closing the system of the balance equations. For example, for the optical phonons we obtain $C_{W_s} = \sum_{ij} C^{ij}_{W_s} + C^{ac}_{W_s}$, where the sum is for

$(i,j) \in \{(e,h),(e,e),(h,h)\}$, and

$$C^{eh}_{W_s} = \frac{2D_s^2}{2\pi^2\rho\hbar^3 v_F^4}\int_0^{\epsilon_s} \varepsilon(\epsilon_s-\varepsilon)\chi_s^{eh}\Big(\frac{\epsilon_s-\varepsilon}{\varepsilon}\Big)g^0(\epsilon_s)$$

$$\times \mathscr{F}^e_{FD}(\varepsilon)\mathscr{F}^h_{FD}(\epsilon_s-\varepsilon)\Big[e^{\epsilon_s\lambda_{W_s}} - e^{\lambda_e+\lambda_h+\lambda_{W_e}\varepsilon}\, e^{\lambda_{W_h}(\epsilon_s-\varepsilon)}\Big]d\varepsilon,$$

$$C^{ee}_{W_s} = \frac{1}{\pi^2\rho\hbar^3 v_F^4}D_s^2\int_0^{\infty} \varepsilon(\varepsilon+\epsilon_s)\chi_s^{ee}\Big(\frac{\epsilon_s+\varepsilon}{\varepsilon}\Big)g^0(\epsilon_s)$$

$$\times \mathscr{F}^e_{FD}(\varepsilon)\mathscr{F}^e_{FD}(\varepsilon+\epsilon_s)\Big[e^{\epsilon_s\lambda_{W_s}+\lambda_e+\lambda_{W_e}\varepsilon} - e^{\lambda_e+\lambda_{W_e}\varepsilon}\, e^{\lambda_{W_e}\epsilon_s}\Big]d\varepsilon,$$

$$C^{ac}_{W_s} = \frac{A\,\epsilon_s}{\tau_{OA}}\Big[g^0\Big(\epsilon_s,\frac{1}{k_B T_{OA}}\Big) - g^0\Big(\epsilon_s,\lambda_{W_s}\Big)\Big],$$

ρ being the area density of graphene, A the area of the first Brillouin zone, $\mathscr{F}_{FD}$ the equilibrium Fermi-Dirac occupation number, T_{OA} the phonon local equilibrium temperature, ϵ_s the optical phonon energy, while the functions χ_s^{ij}, $i,j = e,h$, $s = LO,TO,K$, and the relaxation time τ_{OA} can be found in [3]. Neglecting the acoustic phonon dynamics, the simplest way to study the effect of lattice heating is to use a temperature T which empirically depends on the total current, that is $T = T_L + \gamma\frac{IU}{L}$, where I is the total current, U the applied voltage bias, L the device length, and γ can be found in [6].

6 Numerical Simulations

In the literature there are several values for the coupling constants entering into the collision terms. For example for the acoustic deformation potential one can find values ranging from 2.6 to 29 eV. Similar degree of uncertainty is found for the optical and K phonon coupling constants as well. We have performed numerical simulations of a suspended graphene monolayer by considering the parameters used in [13], see Figs. 1 and 2.

For moderate applied fields the asymptotic value of the electron velocity increases with the applied field, while for high electric fields the velocity decreases (negative differential conductivity) but there is no velocity saturation. The results are consistent with the Monte Carlo simulations presented in [14].

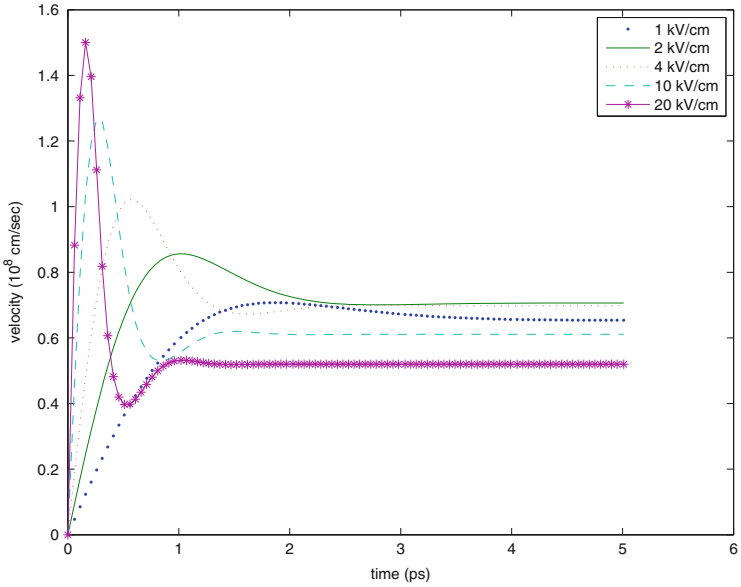


Fig. 1 Average velocity for the electric fields $E = 1\,\text{kV/cm}$, $E = 2\,\text{kV/cm}$, $E = 4\,\text{kV/cm}$, $E = 10\,\text{kV/cm}$, $E = 20\,\text{kV/cm}$ by using the same values of the scattering parameters as in [13], by considering a constant lattice temperature of 300 K and a carrier density equal to 10^{12} cm^{-2}

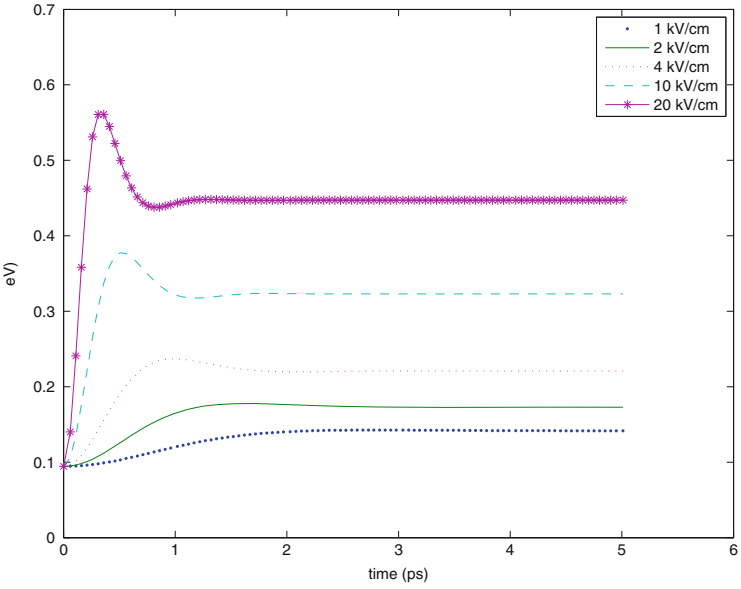


Fig. 2 Average energy for the electric fields $E = 1\,\text{kV/cm}$, $E = 2\,\text{kV/cm}$, $E = 4\,\text{kV/cm}$, $E = 10\,\text{kV/cm}$, $E = 20\,\text{kV/cm}$ by using the same values of the scattering parameters as in [13], by considering a constant lattice temperature of 300 K and a carrier density equal to 10^{12} cm^{-2}

References

1. Zamponi, N., Barletti, L.: Quantum electronic transport in graphene: a kinetic and fluiddynamical approach. Math. Methods Appl. Sci. **34**, 807–818 (2011)
2. Camiola, V.D., Romano, V.: Hydrodynamical model for charge transport in graphene. J. Stat. Phys. **157**(6), 1114–1137 (2014)
3. Mascali, G., Romano, V.: A comprehensive hydrodynamical model for charge transport in graphene. In: 2014 International Workshop on Computational Electronics, Paris (2014). doi: 10.1109/IWCE.2014.6865866
4. Castro Neto, A.H., Guinea, F., Peres, N.M.R., Novoselov, K.S., Geim, A.K.: The electronic properties of graphene. Rev. Modern Phys. **81**, 109 (2009)
5. Fang, T., Konar, A., Xing, H., Jena, D.: High-field transport in two-dimensional graphene. Phys. Rev. B **84**, 125450 (2011)
6. Lichtenberger, P., Morandi, O., Schürrer, F.: High-field transport and optical phonon scattering in graphene. Phys. Rev. B **84**, 045406 (2011)
7. Mascali, G.: A hydrodynamic model for silicon semiconductors including crystal heating. Eur. J. Appl. Math. **26**(4), 477–496 (2015)
8. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. B **106**, 620 (1957)
9. Alì, G., Mascali, G., Romano, V., Torcasio, R.C.: A hydrodynamical model for covalent semiconductors with a generalized energy dispersion relation. Eur. J. Appl. Math. **25**, 255–276 (2014)
10. Muscato, O., Di Stefano, V.: Modeling heat generation in a submicrometric n+-n-n+ silicon diode. J. Appl. Phys. **104**, 124501 (2008)
11. Alì, G., Mascali, G., Romano, V., Torcasio, R.C.: A hydrodynamic model for covalent semiconductors with applications to GaN and SiC. Acta Applicandae Maetamicae **122**, 335 (2012)
12. Camiola, V.D., Mascali, G., Romano, V.: Simulation of a double-gate MOSFET by a nonparabolic hydrodynamical subband model for semiconductors based on the maximum entropy principle. Math. Comput. Model. **58**, 321 (2013)
13. Borysenko, K.M., Mullen, J.T., Barry, E.A., Paul, S., Semenov, Y.G., Zavada, J.M., Buongiorno Nardelli, M., Kim, K.W.: First-principles analysis of electron–phonon interactions in graphene. Phys. Rev. B **11**, 121412(R) (2010)
14. Rengel, R., Couso, C., Martin, M.J.: A Monte Carlo study of electron transport in suspended monolayer graphene. In: Spanish Conference on Electron Devices (CDE) 2013, IEEEXplore

Derivation of a Hydrodynamic Model for Electron Transport in Graphene via Entropy Maximization

L. Barletti

Abstract In this contribution, which is based on the results published in Barletti (J Math Phys 55:083303, 2014) and Barletti et al. (Tr Inst Mat 11:11–29, 2014), we apply the maximum entropy closure technique in order to derive equations of hydrodynamic type for a system of particles with spin-orbit interaction, with particular focus on the case of electrons on a graphene sheet.

Keywords Electron transport • Hydrodynamic model • Maximum entropy closure technique

1 Phase-Space Description of Spin-Orbit Particles

Let us consider a rather general spin-orbit Hamiltonian of the form

$$H(\mathbf{x},\mathbf{p}) = [h_0(\mathbf{p}) + V(\mathbf{x})]\,\sigma_0 + \mathbf{h}(\mathbf{p})\cdot\sigma, \tag{1}$$

where $\mathbf{x}\in\mathbb{R}^d$, $\mathbf{p}\in\mathbb{R}^d$, $\sigma = (\sigma_1,\sigma_2,\sigma_3)$, and

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Hamiltonians of this kind describe, among others, the Rashba spin-orbit interaction [2], the two-band K·P model [3, 4], Dirac massless particles [9] and electrons in graphene [1, 6].

The main semiclassical quantities associated with (1) are the the two energy bands

$$E_\pm(\mathbf{p}) = h_0(\mathbf{p}) \pm |\mathbf{h}(\mathbf{p})|$$

L. Barletti (✉)
Dipartimento di Matematica e Informatica "Ulisse Dini", Viale Morgagni 67/A, 50134 Firenze, Italy
e-mail: luigi.barletti@unifi.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_101

(eigenvalues of H with $V = 0$), the corresponding eigenprojectors

$$P_\pm(\mathbf{p}) = \frac{1}{2}(\sigma_0 \pm \boldsymbol{\nu}(\mathbf{p}) \cdot \sigma), \qquad \boldsymbol{\nu}(\mathbf{p}) = \frac{\mathbf{h}(\mathbf{p})}{|\mathbf{h}(\mathbf{p})|},$$

the semiclassical velocities

$$\mathbf{v}_\pm(\mathbf{p}) = \nabla_{\mathbf{p}} E_\pm(\mathbf{p})$$

and the effective-mass tensor

$$\mathbb{M}_\pm^{-1}(\mathbf{p}) = \nabla_{\mathbf{p}} \otimes \mathbf{v}_\pm(\mathbf{p}) = \nabla_{\mathbf{p}} \otimes \nabla_{\mathbf{p}} E_\pm(\mathbf{p}).$$

The phase-space description of a statistical population of electrons with Hamiltonian (1) is provided by the Wigner matrix [4]

$$F(\mathbf{x}, \mathbf{p}, t) = \sum_{k=0}^{3} f_k(\mathbf{x}, \mathbf{p}, t)\sigma_k,$$

which has the fundamental property that the expected value of an observable with symbol $A = \sum_{k=0}^{3} a_k(\mathbf{x}, \mathbf{p})\sigma_k$ is given by the classical-looking formula

$$\mathbb{E}_F[A] = \int \mathrm{tr}(FA) d\mathbf{x}\, d\mathbf{p} = 2 \int \sum_{k=0}^{3} a_k(\mathbf{x}, \mathbf{p}) f_k(\mathbf{x}, \mathbf{p}, t)\, d\mathbf{x}\, d\mathbf{p}.$$

When applying it to the band projectors $P_\pm(\mathbf{p})$ we obtain

$$\mathbb{E}_F[P_\pm] = \int (f_0 \pm \boldsymbol{\nu} \cdot \mathbf{f})\, d\mathbf{x}\, d\mathbf{p}$$

and it is therefore natural to interpret the functions

$$f_\pm = f_0 \pm \boldsymbol{\nu} \cdot \mathbf{f} \tag{2}$$

as the phase-space densities of electrons having energies, respectively, in the upper and lower band.

Following [1, 3] (see also [5] where additional moments are considered), we shall write equations for the hydrodynamic moments

$$n_\pm = \langle f_\pm \rangle, \qquad \text{(band-densities)},$$

$$n_\pm \mathbf{u}_\pm = \langle \mathbf{v}_\pm f_\pm \rangle \qquad \text{(average velocities)},$$

where we put

$$\langle f\rangle(\mathbf{x},t)=\frac{1}{(2\pi\hbar)^d}\int f(\mathbf{x},\mathbf{p},t)\,d\mathbf{p}.$$

The (semiclassical) dynamics of F is provided by the Wigner equation for the Hamiltonian (1) [4], from which the following equations for the band-Wigner functions f_+ and f_- is obtained:

$$\big(\partial_t+\mathbf{v}_\pm\cdot\nabla_\mathbf{x}-\nabla_\mathbf{x}V\cdot\nabla_\mathbf{p}\big)f_\pm=-\nabla_\mathbf{x}\cdot\mathbf{f}_\perp\pm\boldsymbol{\nu}\cdot(\nabla_\mathbf{x}V\cdot\nabla_\mathbf{p})\mathbf{f}_\perp,$$

where the terms containing $\mathbf{f}_\perp:=(\boldsymbol{\nu}\times\mathbf{f})\times\boldsymbol{\nu}$ are responsible for quantum interference between the two bands [7].

2 Maximum Entropy Closure

We assume that the system is in a state F^{me} of maximum entropy, according to the

Maximum Entropy Principle (MEP) F^{me} is the most probable microscopic state with the observed macroscopic moments $n_\pm$ and $\mathbf{u}_\pm$.

For an electron population in thermal equilibrium with a phonon bath, at constant temperature $T>0$, the most probable microscopic state F^{me} minimizes of the *free-energy*

$$\mathscr{E}(F)=\int_{\mathbb{R}^4}\mathrm{tr}\,[k_BT\,s(F)+HF]\,d\mathbf{p}\,d\mathbf{x},$$

where $s(F)=F\log F+(1-F)\log(1-F)$ is (minus) the Fermi-Dirac entropy function. Moreover, F^{me} is subject to the macroscopic constraints

$$\langle f_\pm^{me}\rangle=n_\pm,\qquad\langle\mathbf{v}_\pm f_\pm^{me}\rangle=n_\pm\mathbf{u}_\pm.$$

It can be proven that

$$f_\pm^{me}=\frac{1}{1+\exp\left(\frac{1}{k_BT}E_\pm(\mathbf{p})-\mathbf{v}_\pm(\mathbf{p})\cdot\mathbf{B}_\pm-A_\pm\right)},\qquad\mathbf{f}_\perp^{me}=\mathbf{0}.$$

where $A_\pm$ and $\mathbf{B}_\pm=(B_1,\dots,B_d)_\pm$ are Lagrange multipliers (functions of $\mathbf{x}$ and t). Thus, the MEP state corresponds to two local and independent Fermi-Dirac distributions in the two energy bands with no interference terms. Substituting $F\mapsto F^{me}$ in the equation for $f_\pm$ yields the decoupled equations

$$\big(\partial_t+\mathbf{v}_\pm\cdot\nabla_\mathbf{x}-\nabla_\mathbf{x}V\cdot\nabla_\mathbf{p}\big)f_\pm^{me}=0.\tag{3}$$

Notation Since the upper-band and lower-band equations are decoupled, we can suppress the $\pm$ labels and use them exclusively for those terms that are different in the two equations.

Taking $\langle\cdot\rangle$ and $\langle\mathbf{v}_\pm\cdot\rangle$ both sides of Eq. (3) we obtain the moment equations

$$\begin{cases} \partial_t n + \partial_i(nu_i) = 0, \\ \partial_t(nu_i) + \partial_j P_{ij}^{\pm} + Q_{ij}^{\pm}\partial_j V = 0, \end{cases} \tag{4}$$

where $\partial_t = \partial/\partial t$, $\partial_i = \partial/\partial x_i$ and

$$P_{ij}^{\pm} = \langle v_i^{\pm} v_j^{\pm} f^{me}\rangle, \qquad Q_{ij}^{\pm} = \langle \tfrac{\partial v_i^{\pm}}{\partial p_j} f^{me}\rangle = \langle (\mathbb{M}_{\pm}^{-1})_{ij} f^{me}\rangle. \tag{5}$$

Thanks to the MEP, the moment system is implicitly closed by the relations

$$\langle f^{me}\rangle = n, \quad \langle \mathbf{v}_\pm f^{me}\rangle = n\mathbf{u}, \quad f^{me} = \frac{1}{1+\exp\left(\frac{1}{k_B T}E_\pm(\mathbf{p}) - \mathbf{v}_\pm(\mathbf{p})\cdot\mathbf{B} - A\right)}, \tag{6}$$

linking the Lagrange multipliers $(A, \mathbf{B})$ to the moments $(n, \mathbf{u})$.

It can be proven that the mapping $(A, \mathbf{B}) \mapsto (n, \mathbf{u})$, provided by Eq. (6), is globally invertible [1, 3], which implies that the MEP states $f_\pm^{me}$ can be viewed as being parametrized by $(n, \mathbf{u})$, instead of $(A, \mathbf{B})$, and this means that the moment equations (4) is a closed system in the unknowns n and $\mathbf{u}$. Such system shares many properties with other models obtained by entropy minimization, the most relevant being the existence of a local entropy and the consequent hyperbolicity [1].

3 The Case of Graphene

Electrons in a single-layer graphene sheet are described by the Hamiltonian (1) with

$$d = 2, \qquad h_0(\mathbf{p}) = 0, \qquad \mathbf{h}(\mathbf{p}) = c\mathbf{p}$$

(where $c \approx 10^6\,\mathrm{m/s}$ is the Fermi velocity). Therefore, the energy bands, the eigenprojections, the semiclassical velocities and the effective-mass are given by

$$E_\pm(\mathbf{p}) = \pm c|\mathbf{p}|, \qquad P_\pm(\mathbf{p}) = \frac{1}{2}(\sigma_0 \pm \boldsymbol{\nu}(\mathbf{p})\cdot\sigma)$$

$$\mathbf{v}_\pm(\mathbf{p}) = \pm c\,\boldsymbol{\nu}(\mathbf{p}), \qquad \mathbb{M}_\pm^{-1}(\mathbf{p}) = \frac{c}{|\mathbf{p}|}\,\boldsymbol{\nu}_\perp(\mathbf{p})\otimes\boldsymbol{\nu}_\perp(\mathbf{p})$$

where

$$\boldsymbol{\nu}(\mathbf{p}) = \frac{\mathbf{p}}{|\mathbf{p}|}, \qquad \boldsymbol{\nu}_\perp = (-\nu_2, \nu_1).$$

Note that $\pm\boldsymbol{\nu}(\mathbf{p})$ is the electron *direction*. The energy bands are the well-known Dirac cones and, since the lower band is unbounded from below, we have to change a little the theory developed in the previous sections and describe the lower-band population in terms of electron vacancies, i.e. *holes*. This is achieved by means of the substitution

$$f_-(\mathbf{x}, \mathbf{p}, t) \longmapsto 1 - f_-(\mathbf{x}, -\mathbf{p}, t),$$

which brings the transport equation, Eq. (3), into

$$\left(\partial_t + c\,\boldsymbol{\nu} \cdot \nabla_{\mathbf{x}} \mp \nabla_{\mathbf{x}} V \cdot \nabla_{\mathbf{p}}\right) f^{me} = 0.$$

Note that the only difference between electrons and holes is the charge sign. Moreover, the MEP-states for electrons and holes have now the same form

$$f^{me} = \frac{1}{1 + \exp\left(\frac{c}{k_B T}|\mathbf{p}| - \boldsymbol{\nu}(\mathbf{p}) \cdot \mathbf{B} - A\right)},$$

(note that both upper-cone electrons and lower-cone holes have positive energies). Moreover, we slightly change the definition of $\mathbf{u}$ to be the average *direction*

$$n\mathbf{u} = \langle \boldsymbol{\nu} f \rangle, \qquad 0 \le |\mathbf{u}| \le 1,$$

which differs from average velocity just for the constant factor c (in fact, for both electrons and holes in graphene $\mathbf{v}_+ = \mathbf{v}_- = c\boldsymbol{\nu}$). The inequality $|\mathbf{u}| \le 1$ is a direct consequence of Jensen inequality.

The moment equations (4), in the specific case of graphene, read as follows:

$$\begin{cases} \partial_t n + \partial_i (n u_i) = 0, \\ \partial_t (n u_i) + \partial_j P_{ij} \pm Q_{ij} \partial_j V = 0, \end{cases} \tag{7}$$

where $P_{ij} = \langle \nu_i \nu_j f^{me} \rangle$ and $Q_{ij} = \langle \frac{1}{|\mathbf{p}|} \nu_i^\perp \nu_j^\perp f^{me} \rangle$.

We now intend to find an (as much as possible) explicit expression for the dependence of the Lagrange multipliers A and $\mathbf{B} = (B_1, B_2)$ in terms of the moments n and $\mathbf{u} = (u_1, u_2)$, as resulting from the constraint equations

$$\langle f^{me} \rangle = n, \qquad n \langle \boldsymbol{\nu} f^{me} \rangle = \mathbf{u}. \tag{8}$$

To this end we recall the *Fermi integral* of order $s > 0$:

$$\phi_s(z) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{t^{s-1}}{e^{t-z}+1}\, dt$$

and introduce the following functions:

$$\mathscr{I}_N^s(A,B) = \frac{1}{\pi} \int_0^\pi \cos(N\theta)\, \phi_s(A + B\cos\theta)\, d\theta,$$

where $A \in \mathbb{R}$, $B \geq 0$, $s > 0$ and N is an integer. From the constraint equations (8) we obtain that $\mathbf{B}$ has the same direction as $\mathbf{u}$ and, as far as A and the modulus $B = |\mathbf{B}|$ are concerned, we have

$$\mathscr{I}_0^2(A,B) = \frac{n}{n_T}, \qquad \frac{\mathscr{I}_1^2(A,B)}{\mathscr{I}_0^2(A,B)} = |\mathbf{u}| \tag{9}$$

where $n_T = (k_B T)^2/2\pi\hbar^2 c^2$. In [1] the following result is proven, showing that the tensors $P_{ij}^{\pm}$ and $Q_{ij}^{\pm}$ can be expressed in terms of the moments n and $\mathbf{u}$, and of the two scalar Lagrange multipliers $A \in \mathbb{R}$ and $B \geq 0$ through the functions $\mathscr{I}_N^s$.

Theorem 1 *The following equalities hold:*

$$P_{ij} = \frac{n}{|\mathbf{u}|^2}\big(P\, u_i u_j + P_\perp\, u_i^\perp u_j^\perp\big), \quad Q_{ij} = \frac{c}{k_B T}\frac{n}{|\mathbf{u}|^2}\big(Q\, u_i u_j + Q_\perp\, u_i^\perp u_j^\perp\big), \tag{10}$$

where the scalar functions $P(A,B)$, $P_\perp(A,B)$, $Q(A,B)$, $Q_\perp(A,B)$ are given by

$$P = \frac{\mathscr{I}_0^2 + \mathscr{I}_2^2}{2\,\mathscr{I}_0^2}, \quad P_\perp = 1 - P, \quad Q = \frac{\mathscr{I}_0^1 - \mathscr{I}_2^1}{2\,\mathscr{I}_0^2}, \quad Q_\perp = \frac{\mathscr{I}_0^1 + \mathscr{I}_2^1}{2\,\mathscr{I}_0^2}.$$

4 Asymptotic Regimes

The expressions (10) of $P_{ij}^{\pm}$ and $Q_{ij}^{\pm}$ are still not explicit, as functions of n and $\mathbf{u}$. Nevertheless, we can say more in some particular regime of physical interest. Such regimes correspond to different asymptotic regions [1] in the half plane $(A,B) \in \mathbb{R} \times [0,\infty)$.

4.1 Diffusive Regime

The diffusive limit corresponds to completely spread directions, i.e. to $|\mathbf{u}| \to 0$, which is equivalent to $B \to 0$. In order to obtain nontrivial limit equations, we have to rescale the time by putting $\mathbf{u} \mapsto \tau_0 \mathbf{u}$ and $t \mapsto t/\tau_0$, and adding a relaxation term $-n\mathbf{u}/\tau_0$. As $\tau_0 \to 0$, we obtain the nonlinear drift-diffusion equation

$$\partial_t n = \frac{\tau_0 c^2}{2} \nabla \cdot \left[\nabla n \pm \frac{n_T}{k_B T} \phi_1\Big(\phi_2^{-1}\Big(\frac{n}{n_T}\Big)\Big) \nabla V \right]. \tag{11}$$

4.2 Maxwell-Boltzmann Regime ($T \to +\infty$)

The limit $T \to +\infty$ corresponds to $A^2 + B^2 \to \infty$ with $A < -B$. In this case, we can use the approximation [1]

$$\mathscr{I}_N^s(A, B) \sim \mathrm{e}^A I_N(B), \tag{12}$$

where I_N are the modified Bessel functions of the first kind. The constraint equations (9) become

$$\mathrm{e}^A I_0(B) = \frac{n}{n_T}, \qquad \frac{I_1(B)}{I_2(B)} = |\mathbf{u}|,$$

from which we finally get the explicit form of the tensors P_{ij} and Q_{ij}:

$$\begin{aligned} P_{ij} &= \frac{n}{|\mathbf{u}|^2} \left[X(|\mathbf{u}|)\, u_i u_j + (1 - X(|\mathbf{u}|))\, u_i^\perp u_j^\perp \right], \\ Q_{ij} &= \frac{c}{k_B T} \frac{n}{|\mathbf{u}|^2} \left[X(|\mathbf{u}|)\, u_i^\perp u_j^\perp + (1 - X(|\mathbf{u}|))\, u_i u_j \right], \end{aligned} \tag{13}$$

where

$$X(|\mathbf{u}|) = \frac{I_0(B) + I_2(B)}{2 I_0(B)}, \qquad B = \Big(\frac{I_1}{I_0}\Big)^{-1}(|\mathbf{u}|).$$

If we perform the diffusive limit $|\mathbf{u}| \to 0$ within the Maxwell-Boltzmann regime, we obtain a linear (although nonconventional) drift-diffusion equation:

$$\partial_t n = \frac{\tau_0 c^2}{2} \nabla \cdot \left(\nabla n \pm \frac{1}{k_B T} n \nabla V \right). \tag{14}$$

4.3 Collimation Regime

On the opposite side with respect to the diffusive limit, the collimation limit corresponds to the absence of spread in the particle directions, i.e. to $|\mathbf{u}| \to 1$. It can be shown that this corresponds to $A^2 + B^2 \to \infty$ with $A/B \to -1$. However, there is a completely different behavior when the critical line $A = -B$ is approached from below (Maxwell-Boltzmann collimation) of from above (degenerate-gas collimation). In the first case, the hydrodynamic system reduces to

$$\partial_t u_i + c u_j \partial_j u_i \pm \frac{c}{k_B T} u_i^{\perp} u_j^{\perp} \partial_j V = 0. \tag{15}$$

This equation reveals that collimated electrons in graphene have the properties of a geometrical-optics system, with "refractive index" $N(\mathbf{x}) = \exp\left(\mp \frac{1}{k_B T} V(\mathbf{x}) \right)$ (see Refs. [1] and [8] for a detailed discussion).

The second case (degenerate-gas collimation) yields the "trivial" equation

$$\begin{cases} \partial_t n + c \partial_i (n u_i) = 0, \\ \partial_t (n u_i) + c \partial_j \left(n u_i u_j \right) = 0, \end{cases}$$

since $Q \to 0$ and $Q_\perp \to 0$, as (A, B) approaches the critical line from above.

4.4 Degenerate Gas Regime ($T \to 0$)

This limit corresponds to $A^2 + B^2 \to \infty$ with $A > -B$, in which case we can use the approximation [1]

$$\mathscr{I}_N^s(A, B) \sim \frac{1}{\pi \Gamma(s+1)} \int_0^{C(A,B)} \cos(N\theta)(A + B\cos\theta)^s d\theta,$$

where

$$C(A, B) = \begin{cases} \arccos(-A/B), & \text{if } -B < A < B, \\ \pi, & \text{if } A \geq B. \end{cases}$$

The tensors P_{ij} and Q_{ij} for a degenerate gas are therefore given by

$$\begin{aligned} P_{ij} &= \frac{n}{|\mathbf{u}|^2} \left[Y(|\mathbf{u}|) u_i u_j + (1 - Y(|\mathbf{u}|)) u_i^{\perp} u_j^{\perp} \right], \\ Q_{ij} &= \frac{\sqrt{n}}{\hbar \sqrt{\pi} |\mathbf{u}|^2} \left[Z(|\mathbf{u}|) u_i u_j + Z_\perp(|\mathbf{u}|) u_i^{\perp} u_j^{\perp} \right], \end{aligned} \tag{16}$$

where

$$Y(|\mathbf{u}|) = \frac{\mathscr{F}_0^2(\psi) + \mathscr{F}_2^2(\psi)}{2\mathscr{F}_0^2(\psi)}, \qquad Z(|\mathbf{u}|) = \frac{\mathscr{F}_0^1(\psi) - \mathscr{F}_2^1(\psi)}{2\sqrt{2\mathscr{F}_0^2(\psi)}},$$

$$Z_\perp(|\mathbf{u}|) = \frac{\mathscr{F}_0^1(\psi) + \mathscr{F}_2^1(\psi)}{2\sqrt{2\mathscr{F}_0^2(\psi)}}, \qquad \psi = \left(\frac{\mathscr{F}_1^2}{\mathscr{F}_0^2}\right)^{-1}(|\mathbf{u}|),$$

and we put $\mathscr{F}_N^s(\psi) \sim R^{-s}\,\mathscr{I}_N^s(R\cos\psi, R\sin\psi)$.

If, moreover, we perform the diffusive limit $|\mathbf{u}| \to 0$ we obtain the nonlinear drift-diffusion equation

$$\partial_t n = \frac{\tau_0 c}{2}\,\nabla\cdot\left(c\nabla n \pm \frac{1}{\hbar\sqrt{\pi}}\,\sqrt{n}\,\nabla V\right). \tag{17}$$

References

1. Barletti, L.: Hydrodynamic equations for electrons in graphene obtained from the maximum entropy principle. J. Math. Phys. **55**, 083303 (2014)
2. Barletti, L., Méhats, F.: Quantum drift-diffusion modeling of spin transport in nanostructures. J. Math. Phys. **51**, 053304 (2010)
3. Barletti, L., Borgioli, G., Frosali, G.: Semiclassical hydrodynamics of a quantum Kane model for semiconductors. Tr. Inst. Mat. **11**, 11–29 (2014)
4. Barletti, L., Frosali, G., Morandi, O.: Kinetic and hydrodynamic models for multi-band quantum transport in crystals. In: Ehrhardt, M., Koprucki, T. (eds.) Multi-Band Effective Mass Approximations: Advanced Mathematical Models and Numerical Techniques. Springer, Berlin (2014)
5. Camiola, V.D., Romano, V.: Hydrodynamical model for charge transport in graphene. J. Stat. Phys. **157**, 1114–1137 (2014)
6. Castro Neto, A.H., Guinea, F., Peres, N.M.R., Novoselov, K.S., Geim, A.K.: The electronic properties of graphene. Rev. Mod. Phys. **81**, 109–162 (2009)
7. Morandi, O.: Wigner-function formalism applied to the Zener band transition in a semiconductor. Phys. Rev. B **80**, 02430 (2009)
8. Morandi, O., Barletti, L.: Particle dynamics in graphene: collimated beam limit. J. Comput. Theor. Transp. **43**, 1–15 (2014)
9. Thaller, B.: The Dirac Equation. Springer, Berlin (1992)

Deterministic Solutions of the Transport Equation for Charge Carrier in Graphene

Armando Majorana and Vittorio Romano

Abstract The aim of this work is to use a numerical scheme based on the discontinuous Galerkin method for finding deterministic (non stochastic) solutions of the electron Boltzmann transport equation in graphene. The same methods has been already successfully applied to a more conventional semiconductor material like Si (Cheng et al., Comput Methods Appl Mech Eng 198(37–40):3130–3150, 2009; Cheng et al., Boletin de la Sociedad Espanola de Matematica Aplicada 54:47–64, 2011). A n-type doping or equivalently a high value of the Fermi potential is considered. Therefore we neglect the inter band scatterings but retain all the main electron-phonon scatterings. Simulations in graphene nano-ribbons are presented and discussed.

Keywords Charge carrier • Discontinuous Galerkin method • Electron Boltzmann equation • Electron transport

1 The Mathematical Model

Graphene is a gapless semiconductor made of a sheet composed of a single layer of carbon atoms arranged into a honeycomb hexagonal lattice [1]. In view of application in graphene-based electron devices, it is crucial to understand the basic transport properties of this material.

A physically accurate model is given by a semiclassical transport equation whose scattering terms have been deeply analyzed recently [2–4]. Due to the computational difficulties, the most part of the available solutions have been obtained by direct Monte Carlo simulations. A different approach has been employed in [5]. Macroscopic models can be found in [6–8].

The aim of this work is to use a numerical scheme based on the discontinuous Galerkin method for finding deterministic (non stochastic) solutions of the electron Boltzmann equation in graphene. The same methods has been already successfully applied to a more conventional semiconductor material like Si [9, 10].

A. Majorana (✉) • V. Romano
Department of Mathematics and Computer Science, Viale A. Doria 6, 95125 Catania, Italy
e-mail: majorana@dmi.unict.it; romano@dmi.unict.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_102

The electron energy in graphene depends on a two dimensional wave vector $\mathbf{k}$ belonging to a bi-dimensional Brillouin zone which has an hexagonal shape. The most part of electrons are in the valleys, around the vertexes of the Brillouin zone, called Dirac points or K and K' points. Usually the three K-valley are treated as a single equivalent one and similarly the three K'-valleys.

In a semiclassical kinetic setting, the charge transport in graphene is described by four Boltzmann equations, one for electrons in the valence (π) band and one for electrons in the conductions (π^*) band, that in turn can belong to the K or K' valley,

$$\frac{\partial f_{\ell,s}(\mathrm{t},\mathbf{x},\mathbf{k})}{\partial t} + \mathbf{v}_{\ell,s}\cdot\nabla_{\mathbf{x}} f_{\ell,s}(\mathrm{t},\mathbf{x},\mathbf{k}) - \frac{e}{\hbar}\mathbf{E}\cdot\nabla_{\mathbf{k}} f_{\ell,s}(\mathrm{t},\mathbf{x},\mathbf{k}) = \left.\frac{df_{\ell,s}}{dt}(\mathrm{t},\mathbf{x},\mathbf{k})\right|_{e-ph}, \qquad (1)$$

where $f_{\ell,s}(\mathrm{t},\mathbf{x},\mathbf{k})$ represents the distribution function of charge carriers in the valley ℓ (K or K'), band π or π^* ($s=-1$ or $s=1$) at position $\mathbf{x}$, time t and wave-vector $\mathbf{k}$. We denote by $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{k}}$ the gradients with respect to the position and wave vector, respectively. The microscopic velocity $\mathbf{v}_{\ell,s}$ is related to the energy band $\varepsilon_{\ell,s}$ by

$$\mathbf{v}_{\ell,s} = \frac{1}{\hbar}\nabla_{\mathbf{k}}\,\varepsilon_{\ell,s}\,.$$

With a very good approximation [1] a linear dispersion relation holds for the energy bands $\varepsilon_{\ell,s}$ around the equivalent Dirac points; so that $\varepsilon_{\ell,s} = s\,\hbar\,v_F\,|\mathbf{k}-\mathbf{k}_\ell|$, where v_F is the (constant) Fermi velocity, $\hbar$ the Planck constant divided by $2\,\pi$, and $\mathbf{k}_\ell$ is the position of the Dirac point ℓ. The elementary (positive) charge is denoted by e, and $\mathbf{E}$ is the electric field obtained by the Poisson equation, which must be coupled with the above system. The right hand side of Eq. (1) is the collision term representing the interaction of electrons with acoustic, optical and K phonons. Acoustic phonon scattering is intra-valley and intra-band. Optical phonon scattering is intra-valley and can be longitudinal optical (LO) and the transversal optical (TO); it can be intra-band, that is leaves the electron in the same band, or inter-band pushing the electron from an initial band to the other one. Scattering with optical phonon of type K pushes electrons from a valley to a neighbor one (inter-valley scattering). We assume that phonons are at thermal equilibrium. Hence, the general form of the collision term can be written as

$$\left.\frac{df_{\ell,s}}{dt}(\mathrm{t},\mathbf{x},\mathbf{k})\right|_{e-ph} = \sum_{\ell',s'}\left[\int S_{\ell',s',\ell,s}(\mathbf{k}',\mathbf{k})\,f_{\ell',s'}(\mathrm{t},\mathbf{x},\mathbf{k}')\,(1-f_{\ell,s}(\mathrm{t},\mathbf{x},\mathbf{k}))\,d\mathbf{k}'\right.$$
$$\left. - \int S_{\ell,s,\ell',s'}(\mathbf{k},\mathbf{k}')\,f_{\ell,s}(\mathrm{t},\mathbf{x},\mathbf{k})\,\left(1-f_{\ell',s'}(\mathrm{t},\mathbf{x},\mathbf{k}')\right)\,d\mathbf{k}'\right]$$

where the total collision term is given by the sum of the contributions of several types of scatterings

$$S_{\ell',s',\ell,s}(\mathbf{k}',\mathbf{k}) = \sum_{\nu}\left|G^{(\nu)}_{\ell',s',\ell,s}(\mathbf{k}',\mathbf{k})\right|^2\left[\left(n^{(\nu)}_{\mathbf{q}}+1\right)\delta\left(\varepsilon_{\ell,s}(\mathbf{k})-\varepsilon_{\ell',s'}(\mathbf{k}')+\hbar\,\omega^{(\nu)}_{\mathbf{q}}\right)\right.$$
$$\left. + n^{(\nu)}_{\mathbf{q}}\,\delta\left(\varepsilon_{\ell,s}(\mathbf{k})-\varepsilon_{\ell',s'}(\mathbf{k}')-\hbar\,\omega^{(\nu)}_{\mathbf{q}}\right)\right]. \qquad (2)$$

The index ν labels the νth phonon mode, $G^{(\nu)}_{\ell',s',\ell,s}(\mathbf{k}',\mathbf{k})$ is the scattering rate, which describes the scattering mechanism, due to phonons ν, between electrons belonging to valley ℓ' and band s', and electron belonging to valley ℓ and band s. The symbol δ denotes the Dirac distribution function, $\omega^{(\nu)}_{\mathbf{q}}$ the νth phonon frequency, $n^{(\nu)}_{\mathbf{q}}$ is the Bose-Einstein distribution for the phonon of type ν

$$n^{(\nu)}_{\mathbf{q}} = \frac{1}{e^{\hbar\,\omega^{(\nu)}_{\mathbf{q}}/k_B T} - 1}\,,$$

k_B is the Boltzmann constant and T the constant graphene lattice temperature. When, for a phonon ν_*, $\hbar\,\omega^{(\nu_*)}_{\mathbf{q}} \ll k_B T$, then the scattering with the phonon ν_* can be assumed elastic. In this case, we eliminate in Eq. (2) the term $\hbar\,\omega^{(\nu_*)}_{\mathbf{q}}$ inside the delta distribution and we use the approximation $n^{(\nu_*)}_{\mathbf{q}} + 1 \approx n^{(\nu_*)}_{\mathbf{q}}$.

1.1 The Model with Only One Distribution Function

In this paper we consider a numerical no stochastic technique, based on the discontinuous Galerkin method, for solving the kinetic model described in Sect. 1. In this first application, we study the case of a single distribution function f. This corresponds to a physical case, where a n-type doping or equivalently a high value of the Fermi potential is considered, and the electrons, belonging to a conduction band, do not move to the valence band. Moreover K and K' are considered equivalent. A reference frame centered in the K-point will be used. Of course, we simplify the notation, omitting the indexes s and ℓ. Now, we write the scattering rates used in our simulations, explicitly.

For acoustic phonons, usually one considers the elastic approximation, and

$$2\,n^{(ac)}_{\mathbf{q}}\left|G^{(ac)}(\mathbf{k}',\mathbf{k})\right|^2 = \frac{1}{(2\,\pi)^2}\,\frac{\pi\,D^2_{ac}\,k_B\,T}{2\hbar\,\sigma_m\,v^2_p}\,(1+\cos\vartheta_{\mathbf{k},\mathbf{k}'})\,, \tag{3}$$

where D_{ac} is the acoustic phonon coupling constant, v_p is the sound speed in graphene, σ_m the graphene areal density, and $\vartheta_{\mathbf{k},\mathbf{k}'}$ is the convex angle between $\mathbf{k}$ and $\mathbf{k}'$.

There are three relevant optical phonon scatterings: the longitudinal optical (LO), the transversal optical (TO) and the K (K) phonons. The scattering rates are

$$\left|G^{(LO)}(\mathbf{k}',\mathbf{k})\right|^2 = \frac{1}{(2\,\pi)^2}\,\frac{\pi\,D^2_O}{\sigma_m\,\omega_O}\,(1-\cos(\vartheta_{\mathbf{k},\mathbf{k}'-\mathbf{k}}+\vartheta_{\mathbf{k}',\mathbf{k}'-\mathbf{k}})) \tag{4}$$

$$\left|G^{(TO)}(\mathbf{k}',\mathbf{k})\right|^2 = \frac{1}{(2\,\pi)^2}\,\frac{\pi\,D^2_O}{\sigma_m\,\omega_O}\,(1+\cos(\vartheta_{\mathbf{k},\mathbf{k}'-\mathbf{k}}+\vartheta_{\mathbf{k}',\mathbf{k}'-\mathbf{k}})) \tag{5}$$

$$\left|G^{(K)}(\mathbf{k}',\mathbf{k})\right|^2 = \frac{1}{(2\,\pi)^2}\,\frac{2\pi\,D^2_K}{\sigma_m\,\omega_K}\,(1-\cos\vartheta_{\mathbf{k},\mathbf{k}'})\,, \tag{6}$$

where D_O is the optical phonon coupling constant, ω_O the optical phonon frequency, D_K is the K-phonon coupling constant and ω_K the K-phonon frequency. The angles $\vartheta_{\mathbf{k},\mathbf{k}'-\mathbf{k}}$ and $\vartheta_{\mathbf{k}',\mathbf{k}'-\mathbf{k}}$ denote the convex angles between $\mathbf{k}$ and $\mathbf{k}'-\mathbf{k}$ and between $\mathbf{k}'$ and $\mathbf{k}'-\mathbf{k}$, respectively.

2 The Numerical Method

We look for spatially homogeneous solutions to Eq. (1) with a constant electric field. Now, the Boltzmann equation reduces to

$$\frac{\partial f(\mathrm{t},\mathbf{k})}{\partial t} - \frac{e}{\hbar}\,\mathbf{E}\cdot\nabla_{\mathbf{k}} f(\mathrm{t},\mathbf{k}) = \int S(\mathbf{k}',\mathbf{k}) f(\mathrm{t},\mathbf{k}')\,(1-f(\mathrm{t},\mathbf{k}))\,d\mathbf{k}' - \int S(\mathbf{k},\mathbf{k}') f(\mathrm{t},\mathbf{k})\,\big(1-f(\mathrm{t},\mathbf{k}')\big)\,d\mathbf{k}'\,. \tag{7}$$

We take a Fermi-Dirac distribution, as initial condition,

$$f(0,\mathbf{k}) = \frac{1}{1+\exp\left(\dfrac{\varepsilon(\mathbf{k})-\mu}{k_B\,T}\right)},$$

where $T = 300\,\mathrm{K}$, and μ is the chemical potential, that is determined by choosing the initial charge density

$$\rho(0) = \frac{2}{(2\,\pi)^2}\int f(0,\mathbf{k})\,d\mathbf{k}\,. \tag{8}$$

Equation (7) is discretized by adopting a discontinuous Galerkin scheme. We choose a bounded domain $\Omega \subset \mathbb{R}^2$ such that $f(\mathrm{t},\mathbf{k}) \approx 0$ for every $\mathbf{k} \notin \Omega$ and $\mathrm{t} > 0$, and we introduce a finite decomposition $\{C_\alpha\}$ of Ω, with C_α appropriate open set, such that

$$C_\alpha \cap C_\beta = \emptyset \quad \text{if } \alpha \neq \beta, \text{ and } \quad \bigcup_{\alpha=1}^{N} \overline{C_\alpha} = \Omega\,.$$

We assume that the distribution function is constant in each cell C_α. If we denote by $\chi_\alpha(\mathbf{k})$ the characteristic function over the cell C_α, then the approximation of the distribution function f is given by

$$f(\mathrm{t},\mathbf{k}) \approx f^\alpha(\mathrm{t}) \quad \forall\,\mathbf{k}\in C_\alpha \iff f(\mathrm{t},\mathbf{k}) \approx \sum_{\alpha=1}^{N} f^\alpha(\mathrm{t})\,\chi_\alpha(\mathbf{k}) \quad \forall\,\mathbf{k}\in\bigcup_{\alpha=1}^{N} C_\alpha\,.$$

This assumption replaces the unknown f, which depends on the two variables t and $\mathbf{k}$, in a set of N unknowns f^α, which depend only on time t. In order to obtain a set of N equations for the new unknowns f^α, we integrate Eq. (7) with respect to $\mathbf{k}$ over every cell C_α and replace f with its approximation. The derivative of f with respect to the time is treated easily. We have

$$\int_{C_\alpha} \frac{\partial f(\mathrm{t},\mathbf{k})}{\partial t}\, d\mathbf{k} \approx M_\alpha \frac{df^\alpha}{dt}$$

where M_α is the measure of the cell C_α. It is clear that the numerical method yields a system of ordinary differential equations. This is achieved by discretizing the collision operator and the drift term.

2.1 Discretization of the Collision Operator

Since, for each $\mathbf{k} \in C_\alpha$, we have

$$\begin{aligned}
&\int S(\mathbf{k}',\mathbf{k}) f(\mathrm{t},\mathbf{k}') \left(1 - f(\mathrm{t},\mathbf{k})\right) d\mathbf{k}' - \int S(\mathbf{k},\mathbf{k}') f(\mathrm{t},\mathbf{k}) \left(1 - f(\mathrm{t},\mathbf{k}')\right) d\mathbf{k}' \\
&\approx \sum_{\beta=1}^{N} \left[\int_{C_\beta} S(\mathbf{k}',\mathbf{k}) f^\beta(\mathrm{t}) \left(1 - f^\alpha(\mathrm{t})\right) d\mathbf{k}' - \int_{C_\beta} S(\mathbf{k},\mathbf{k}') f^\alpha(\mathrm{t}) \left(1 - f^\beta(\mathrm{t})\right) d\mathbf{k}' \right] \\
&= \sum_{\beta=1}^{N} \left[f^\beta(\mathrm{t}) \left(1 - f^\alpha(\mathrm{t})\right) \int_{C_\beta} S(\mathbf{k}',\mathbf{k})\, d\mathbf{k}' - f^\alpha(\mathrm{t}) \left(1 - f^\beta(\mathrm{t})\right) \int_{C_\beta} S(\mathbf{k},\mathbf{k}')\, d\mathbf{k}' \right].
\end{aligned}$$

Now, if we define

$$A^{\alpha,\beta} = \int_{C_\alpha} \left[\int_{C_\beta} S(\mathbf{k},\mathbf{k}')\, d\mathbf{k}' \right] d\mathbf{k}\,, \tag{9}$$

then we obtain

$$\begin{aligned}
&\int_{C_\alpha} \left[\int S(\mathbf{k}',\mathbf{k}) f(\mathrm{t},\mathbf{k}') \left(1 - f(\mathrm{t},\mathbf{k})\right) d\mathbf{k}' - \int S(\mathbf{k},\mathbf{k}') f(\mathrm{t},\mathbf{k}) \left(1 - f(\mathrm{t},\mathbf{k}')\right) d\mathbf{k}' \right] d\mathbf{k} \\
&\approx \sum_{\beta=1}^{N} \left[A^{\beta,\alpha} \left(1 - f^\alpha(\mathrm{t})\right) f^\beta(\mathrm{t}) - A^{\alpha,\beta} f^\alpha(\mathrm{t}) \left(1 - f^\beta(\mathrm{t})\right) \right].
\end{aligned}$$

So, the integral collision operator is replaced by quadratic polynomials. We note that the numerical coefficients $A^{\alpha,\beta}$ depend only on the scattering terms and the domain decomposition.

2.2 Discretizaton of the Force Term

We must approximate the term

$$-\frac{e}{\hbar}\,\mathbf{E}\cdot\int_{C_\alpha}\nabla_{\mathbf{k}} f(\mathrm{t},\mathbf{k})\,d\mathbf{k} = -\frac{e}{\hbar}\,\mathbf{E}\cdot\int_{\partial C_\alpha} f(t,\mathbf{k})\,\mathbf{n}\,d\sigma$$

where $\mathbf{n}$ is the normal to the boundary ∂C_α of the cell C_α. Since, due to the Galerkin method, the approximation of f is not defined on the boundary of the cells, we must introduce a *numerical flux*, that furnishes reasonable values of f on every ∂C_α, depending on the values of the approximation of f in the nearest neighborhood of the cell C_α and on the sign of $\mathbf{E}\cdot\mathbf{n}$. In Fig. 1 we show a simple picture of the cells that can be involved to find the numerical flux. The simplest numerical flux is given by the *upwind rule*, that use only four nearest adjacent cells.

3 Numerical Simulations

We consider a circle as domain Ω. We used the same physical parameters of [3]. The charge density is taken equal to $10^{12}\,\mathrm{cm}^{-2}$. A TVD third Runge-Kutta scheme is used to solve the resulting ODE system. The numerical scheme is very similar to [11]. We remark that the numerical scheme guarantees the mass conservation. We solve Eq. (7) for different value of the applied electric field. In Fig. 2 we show the macroscopic velocity and energy, defined by

$$\frac{2}{(2\,\pi)^2\,\rho(0)}\int f(t,\mathbf{k})\,v_F\,\frac{\mathbf{k}}{|\mathbf{k}|}\,d\mathbf{k}\,,\qquad \frac{2}{(2\,\pi)^2\,\rho(0)}\int f(t,\mathbf{k})\,\varepsilon(\mathbf{k})\,d\mathbf{k}\,.$$

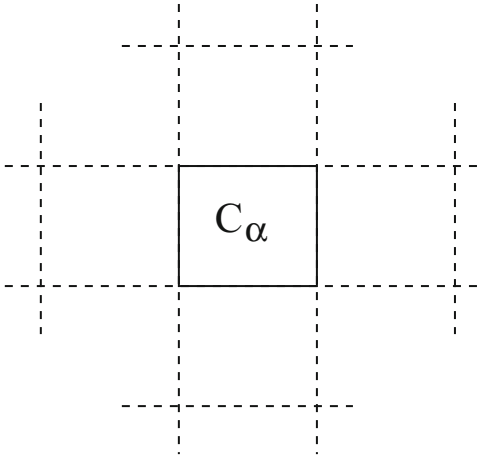


Fig. 1 Cells employed for the numerical flux in the case of a simple rectangular grid

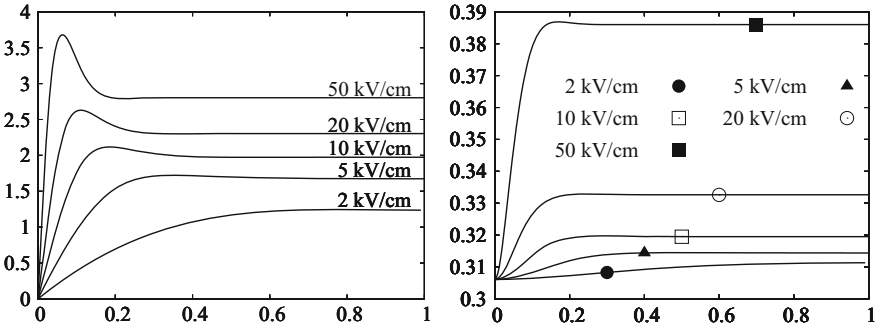


Fig. 2 *Left figure*: the mean velocity in 10^7 cm/s versus time (in ps). *Right figure*: the mean energy in eV versus time (in ps)

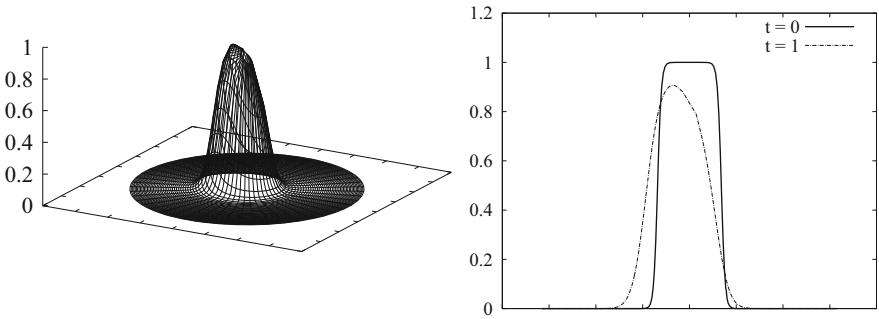


Fig. 3 *Left figure*: The distribution function (electric field equal to 50 kV/cm) at 1 ps. *Right figure*: the section at $k_y = 0$ of the distribution function (electric field equal to 50 kV/cm) at the initial time and at 1 ps

We note that the asymptotic mean velocity and energy increase by increasing the applied voltage. In Fig. 3 we show the distribution function f for the highest electric field.

References

1. Castro Neto, A.H., Guinea, F., Peres, N.M.R., Novoselov, K.S., Geim, A.K.: The electronic properties of graphene. Rev. Mod. Phys. **81**, 109–162 (2009)
2. Shishir, R.S., Ferry, D.K.: Velocity saturation in intrinsic graphene. J. Phys. Condens. Matter **21**, 344201 (2009)
3. Fang, T., Konar, A., Xing, H., Jena, D.: High-field transport in two-dimensional graphene. Phys. Rev. B **84**, 125450 (2011)
4. Tomadin, A., Brida, D., Cerullo, G., Ferrari, A.C, Polini, M.: Nonequilibrium dynamics of photoexcited electrons in graphene: collinear scattering, Auger processes, and the impact of screening. Phys. Rev. B **88**, 035430 (2013)

5. Lichtenberger, P., Morandi, O., Schürrer, F.: High-field transport and optical phonon scattering in graphene. Phys. Rev. B **84**, 045406 (2011)
6. Zamponi, N., Barletti, L.: Quantum electronic transport in graphene: a kinetic and fluid-dynamical approach. Math. Methods Appl. Sci. **34**, 807 (2011)
7. Camiola, V.D., Romano, V.: Hydrodynamical model for charge transport in graphene. J. Stat. Phys. **157**, 1114 (2014)
8. Mascali, G., Romano, V.: A comprehensive hydrodynamical model for charge transport in graphene. In: 978-1-4799-5433-9/14/$31.00 © 2014 IEEE, IWCE-2014, Paris (2014)
9. Cheng, Y., Gamba, I.M., Majorana, A., Shu, C.-W.: A discontinuous Galerkin solver for Boltzmann-Poisson systems in nano devices. Comput. Methods Appl. Mech. Eng. **198**(37–40), 3130–3150 (2009)
10. Cheng, Y., Gamba, I.M., Majorana, A., Shu, C.-W.: A brief survey of the discontinuous Galerkin method for the Boltzmann-Poisson equations. Boletin de la Sociedad Espanola de Matematica Aplicada **54**, 47–64 (2011)
11. Galler, M., Majorana, A.: Deterministic and stochastic simulations of electron transport in semiconductors. Bull. Inst. Math. Acad. Sin. N. S. **2**, 349–365 (2007)

Modulated Bloch Waves in Semiconductor Superlattices

M. Alvaro, L.L. Bonilla, and M. Carretero

Abstract We show that in a semiconductor superlattice with long scattering times, damping of Bloch oscillations due to scattering is so small that nonlinearities may compensate it and Bloch oscillations persist even in the hydrodynamic regime. In order to demonstrate this, we propose a Boltzmann-Poisson transport model of miniband superlattices with inelastic collisions and we derive by singular perturbation methods hydrodynamic equations for electron density, electric field, and the complex amplitude of the Bloch oscillations. Numerical solutions of these equations show stable Bloch oscillations with spatially inhomogeneous field, charge, current density, and energy density profiles. These Bloch oscillations disappear as scattering times become sufficiently short. For sufficiently low lattice temperatures (70 K), Bloch and Gunn type oscillations mediated by electric field, current, and energy domains coexist for a range of voltages. For larger lattice temperatures (300 K), there are only Bloch oscillations with stationary amplitude and electric field profiles.

Keywords Bloch oscillations • Modulated Bloch waves • Semiconductor • Semiconductor superlattices

1 Introduction

Bloch oscillations (BOs) are coherent oscillations of the position of electrons inside energy bands of a crystal under an applied constant electric field $-F$. Their frequency is $\omega_B = eFl/\hbar$ (l lattice constant), and therefore it can be tuned by an applied voltage. BOs were predicted by Zener in 1934 [1]. To observe BOs, their period has to be shorter than the scattering time τ, and therefore the applied field has to surpass the value $\hbar/(el\tau)$, which is too large for most natural materials, in which l is of Ångström size. In 1970, Esaki and Tsu suggested to create an artificial crystal, called superlattice (SL) [2]. Damped Bloch oscillations were first observed in 1992 in semiconductor SLs [3]. Besides their interest for theoretical

M. Alvaro • L.L. Bonilla (✉) • M. Carretero
Gregorio Millan Institute for Fluid Dynamics, Nanoscience and Industrial Mathematics, Universidad Carlos III de Madrid, Avenida de la Universidad 30, E28911 Leganes, Spain
e-mail: mariano.alvaro@uc3m.es; bonilla@ing.uc3m.es; manuel.carretero@uc3m.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*, Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_103

physics, BOs have attracted the attention of many physicists and engineers because of their potential for designing infrared detectors, emitters or lasers which can be tuned in the THz frequency range simply by varying the applied electric field [4]. However no electrically driven devices based on BOs have been realized, because their applications are severely limited by scattering which rapidly damps BOs and, for a dc voltage biased SL, favors the formation of electric field domains (EFDs) whose dynamics yields self-sustained oscillations of lower frequency (GHz) [5, 6]. To understand the role of EFD formation in the observation of BOs we propose a model in which BOs and EFDs are both possible solutions of the governing equations. Therefore, we consider materials with long-lived BOs corresponding to almost elastic collisions, and, also, we consider that the local equilibrium in the Boltzmann-BGK kinetic theory depends on electron density, electron current density and mean energy and the collision term preserves charge but dissipates momentum and energy [7], which is the crucial feature if we want to derive a hydrodynamic regime that allows BOs.

2 Model

The model equations are [8]:

$$\partial_t f + v(k)\,\partial_x f + eF\hbar^{-1}\partial_k f = Q[f] \equiv -\nu(f - f^B), \tag{1}$$

$$\varepsilon\,\partial_x F = el^{-1}(n - N_D), \tag{2}$$

$$f^B(k; n, J_n, E) = n\,\frac{\pi\, e^{\tilde{u}kl + \tilde{\beta}\cos kl}}{\int_0^\pi e^{\tilde{\beta}\cos K}\cosh(\tilde{u}K)\,dK}, \tag{3}$$

$$n = \tfrac{l}{2\pi}\int_{-\pi/l}^{\pi/l} f(x,k,t)dk = \tfrac{l}{2\pi}\int_{-\pi/l}^{\pi/l} f^B dk. \tag{4}$$

Here n, N_D, ε, $-e < 0$, m^*, ν, and $-F$ are the 2D electron density, the 2D doping density, the permittivity, the electron charge, the effective mass of the electron, the constant collision frequency and the electric field, respectively. $v(k) = \Delta l \sin(kl)/(2\hbar)$ is the group velocity corresponding to the miniband tight binding dispersion relation $\mathscr{E}(k) = \Delta(1 - \cos kl)/2$. We have assumed a Boltzmann local equilibrium (3). The distribution functions f and f^B have the same units as n and are $2\pi/l$-periodic in k (the function $\tilde{u}kl$ in (3) is extended periodically outside $-\pi < kl \le \pi$).

The dimensionless multipliers $\tilde{\beta}(x,t)$ and $\tilde{u}(x,t)$ depend on $J_n{=}e\int_{-\pi/l}^{\pi/l} v(k) f\, dk/(2\pi)$ (electron current density) and on $E{=}l\int_{-\pi/l}^{\pi/l}[\Delta/2 - \mathscr{E}(k)]f dk/(2\pi n)$ (mean energy). They are found by solving

$$\begin{aligned} \tfrac{e}{2\pi}\int_{-\pi/l}^{\pi/l} v(k) f^B\, dk &= (1-\alpha_j)J_n, \\ \tfrac{l}{2\pi n}\int_{-\pi/l}^{\pi/l}\left(\tfrac{\Delta}{2} - \mathscr{E}\right) f^B dk &= \alpha_e E_0 + (1-\alpha_e)E. \end{aligned} \tag{5}$$

The restitution coefficients α_j and α_e take values on the interval $[0, 1]$ and measure the dissipation due to collisions in current density and energy, respectively. For $\alpha_{e,j} = 0$ the collisions conserve energy and momentum (elastic limit). To simplify matters, we shall assume that α_j and α_e are constant. E_0 is the mean energy at the lattice temperature of the global equilibrium which will be reached in the absence of bias and contact with external reservoirs. At the lattice temperature, $T_0 = \Delta/(2k_B\tilde{\beta}_0)$, $\tilde{u} = 0$, $E = E_0$, and (5) yields $2E_0/\Delta = I_1(\tilde{\beta}_0)/I_0(\tilde{\beta}_0)$, where $I_s(x)$, $s = 0, 1$, are modified Bessel functions.

3 Hydrodynamic Equations

In the hyperbolic limit in which the collision and Bloch frequencies are comparable and dominate all other terms in (1), it is possible to derive closed equations for nondimensional n, F and A (the complex envelope of the BO solution), provided the collisions are almost elastic. The small dimensionless parameter $\delta = e^2 N_D l \Delta/(2\varepsilon\hbar^2\nu^2)$ is the ratio between the scattering time and the dielectric relaxation time and the restitution coefficients are assumed to scale with it, $\alpha_{e,j} = \delta\gamma_{e,j}$. $f_j(x, \theta, t; \delta)$ are the Fourier coefficients of $f(x, k, t; \delta) = \sum_{j=-\infty}^{\infty} f_j e^{ijk}$ and f_1 is given by

$$f_1 = nE - iJ_n = A(x,t)e^{-i\theta} + f_{1,S}(x,t)\,, \tag{6}$$

where $\theta = \frac{1}{\delta}\int_0^t F(x,s)\,ds$ is the rapidly varying phase of the BO. The nondimensional equations are

$$\frac{\partial F}{\partial t} + \frac{\delta}{F^2 + \delta^2\gamma_j\gamma_e}\left[\gamma_e E_0 nF + \frac{F}{2}\frac{\partial}{\partial x}\mathrm{Im}\,\frac{f_{2,0}^{B(0)}}{1+2iF}\right.$$
$$\left. - \frac{\delta\gamma_e}{2}\frac{\partial}{\partial x}\left(n - \mathrm{Re}\,\frac{f_{2,0}^{B(0)}}{1+2iF}\right) - F\mathrm{Re}\,h_S + \delta\gamma_e \mathrm{Im}\,h_S\right] = J(t), \tag{7}$$

$$\frac{\partial F}{\partial x} = n - 1, \tag{8}$$

$$\frac{\partial A}{\partial t} = -\frac{\gamma_e + \gamma_j}{2}A + \frac{1}{2i}\frac{\partial}{\partial x}\left(\frac{f_{2,-1}^{B(0)}}{1+iF}\right), \tag{9}$$

$$h_S = \frac{f_{1,Su}}{n}\frac{\partial \mathrm{Im} f_{1,Su}}{\partial x} + (J + \mathrm{Im} f_{1,Su})\frac{\partial f_{1,Su}}{\partial F}, \tag{10}$$

$$f_{1,Su} = \frac{\delta\gamma_e nE_0(\delta\gamma_j - iF)}{\delta^2\gamma_e\gamma_j + F^2},$$

$$f_{1,S} = nE_S - iJ_{n,S}$$
$$= \frac{\delta}{F^2+\delta^2\gamma_j\gamma_e}\left[\gamma_e n E_0(\delta\gamma_j - iF) - (\delta\gamma_j - iF)\mathrm{Re}h_S - (F + i\delta\gamma_e)\mathrm{Im}h_S\right.$$
$$\left. + \frac{F+i\delta\gamma_e}{2}\frac{\partial}{\partial x}\left(n - \mathrm{Re}\,\frac{f_{2,0}^{B(0)}}{1+i2F}\right) + \frac{\delta\gamma_j - iF}{2}\,\mathrm{Im}\frac{\partial}{\partial x}\left(\frac{f_{2,0}^{B(0)}}{1+i2F}\right)\right]. \tag{11}$$

The dimensionless multipliers $\tilde{\beta}$ and $\tilde{u}$ in f^B are functions of the rapidly varying BO phase θ due to (6) and therefore, we can expand f^B in (3) in powers of δ, $f^B \sim f^{B(0)} + \delta f^{B(1)}$. The $f^{B(m)}$ $(m = 1, 2)$ are now 2π-periodic functions of θ and k. Then we have the Fourier coefficients

$$f_{j,m}^{B(0)} = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f^{B(0)}(k; n, f_1)\, e^{-ijk-im\theta}\,\frac{dk\,d\theta}{(2\pi)^2}, \tag{12}$$

in which we set $f_1 = A\,e^{-i\theta}$ ignoring $O(\delta)$ terms in (6).

To derive (7)–(9), we start from the equations for the moments f_j which can be obtained from (1) by integration over k [7]:

$$\frac{\partial f_0}{\partial t} - \mathrm{Im}\frac{\partial f_1}{\partial x} = 0, \tag{13}$$
$$\left(\delta\frac{\partial}{\partial t} + iF\right)f_1 = \delta\left[\gamma_e f_0 E_0 - \frac{\gamma_e + \gamma_j}{2}f_1 - \frac{\gamma_e - \gamma_j}{2}f_1^* - \frac{1}{2i}\frac{\partial}{\partial x}(f_0 - f_2)\right], \tag{14}$$

where $f_0 = n, f_1 = nE - iJ_n$, and there are similar equations for higher moments. From (13) and the Poisson equation $\partial F/\partial x = n - 1$, we find Ampère's law for F: $\partial F/\partial t = J(t) - J_n$, where $J(t)$ is the total current density. We shall assume that the second moment f_2 is a known function of f_0 and f_1, $f_2 = g(f_0, f_1)$. Then we find equations for n, F and A in (6) by a method of nonlinear multiple scales with time scales θ and t. To obtain the function g, we carry out a Chapman-Enskog expansion [9] for (1) with a time derivative given by $F\,\partial f/\partial\theta + \delta\,\partial f/\partial t$, according to (6). The distribution function is supposed to be periodic in k and in θ. This procedure gives approximate formulas for $g = f_2$ from which (7)–(9) are obtained.

The hydrodynamic equations (7)–(9) have the spatially uniform solutions, $n = 1$, $J = \delta\gamma_e E_0 nF/(\delta^2\gamma_e\gamma_j + F^2)$, and $A = A_0 e^{-(\gamma_e+\gamma_j)t/2}$. Inserting the latter formula in (6), we see that this corresponds to a damped BO whose amplitude relaxes to 0. Even when we manage to prepare the initial state with a coherent BO of complex amplitude A_0, ignoring space dependence will lead to disappearance of the BOs after a relaxation time $2/(\gamma_e + \gamma_j)$. Stabilization of the BOs may be caused only by the spatially dependent second term on the right hand side of (9).

4 Results

We now solve numerically the hydrodynamic equations with the boundary conditions [6]

$$\frac{\partial F}{\partial t} + \sigma_0 F\Big|_{x=0} = J, \quad \frac{\partial F}{\partial t} + \sigma_1 nF\Big|_{x=L} = J, \tag{15}$$

$$\frac{1}{L}\int_0^L F(x,t)\,dx = \phi, \tag{16}$$

$$\frac{\partial A}{\partial x} = 0, \quad \text{at } x = 0. \tag{17}$$

Initially, the mean energy density is $E_0 = 0.5501$ and the profiles of A and F are uniform, taking on values of $A_0 = 0.5501$ and $\phi = 0.05$, respectively. We use a small parameter $\delta = 0.0053$, and in order to obtain undamped BOs, we have used $\gamma_{e,j} = 1.1269$ so that $(\gamma_e + \gamma_j)/2 < \gamma_{\text{crit}}$. We consider a 50-period dc voltage biased GaAs-AlAs SL with lattice temperature 70 K and dimensionless contact conductivities $\sigma_0 = 1$ and $\sigma_1 = 0.25$.

For a voltage bias $\phi = 0.05$ ($V = 0.166$ V) and after a short transient that depends on the initial conditions, we observe coexisting BOs of frequency 0.36 THz and Gunn type oscillations of frequency 11 GHz. BOs are stable because $(\gamma_e + \gamma_j)/2 < \gamma_{\text{crit}}$ and Gunn type oscillations are a consequence of the periodic recycling and motion of electric field pulses from the cathode to the anode. Figure 1 illustrates the total current density of the coexisting 0.36 THz Bloch and 11 GHz Gunn type oscillations, respectively. For each lattice temperature, there is a critical curve in the plane of restitution coefficients such that, for $(\gamma_e + \gamma_j)/2 > \gamma_{\text{crit}}$, BOs disappear after a relaxation time but they persist for smaller values of $(\gamma_e + \gamma_j)$. Figure 2

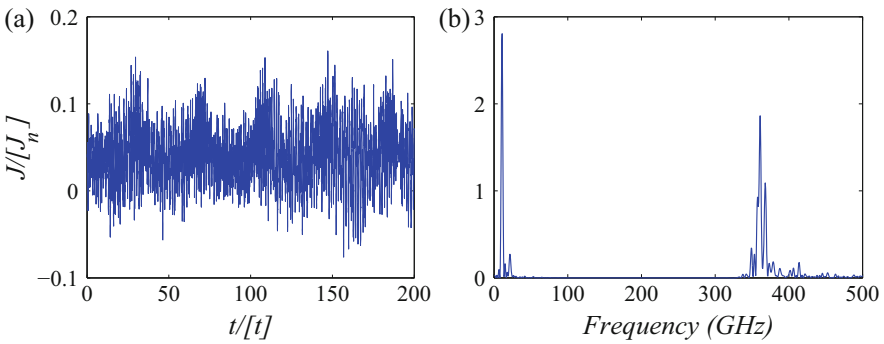


Fig. 1 (**a**) Total current density vs time during coexisting Bloch and Gunn type oscillations at 70 K. (**b**) Fourier transform of the total current density showing two peaks corresponding to coexisting Bloch (0.36 THz) and Gunn type (11 GHz) oscillations

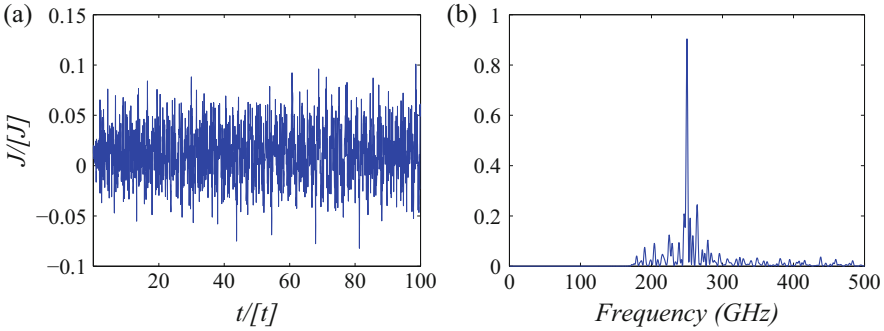


Fig. 2 (**a**) Total current density vs time during Bloch oscillations at 300 K. (**b**) Fourier transform of the total current density showing only one peak corresponding to BOs (0.26 THz). The zero-frequency constant corresponding to the time average of the total current density has been subtracted

depicts the total current density at temperature 300 K, with $E_0 = 0.1529$, for the same values of $\gamma_{e,j}$ and the other parameters. We find BOs but not the slower Gunn type oscillations. Whether Bloch and Gunn type oscillations coexist depends on the lattice temperature. There is a critical temperature below which the electric field pulses are periodically recycled when they reach the anode, originating the Gunn type oscillations. For larger temperatures, Bloch and Gunn type oscillations cannot occur simultaneously: When the electric field pulse reaches the anode, it remains stuck there and the electric field profile becomes stationary. Note that the largest peak in the current spectrum in Fig. 2b occurs at a lower frequency (0.26 THz) than in the case of lower lattice temperature shown in Fig. 1b.

5 Conclusions

We have analyzed the Boltzmann-BGK-Poisson equations with local equilibrium depending on the electron density, current density and energy density in the hyperbolic limit in which the BO period is much shorter than the dielectric relaxation time and collisions are almost elastic. In the long-time scale, there is a hydrodynamic regime described by coupled equations for the electric field, the electron density and the BO complex amplitude. When the restitution coefficients (equivalently the inverse of the scattering times) are sufficiently small and the initial state has been prepared so that there is a nonzero Bloch oscillation, there are undamped BOs coexisting with Gunn-type oscillations for low temperatures. For larger temperatures the only persisting oscillations are undamped BOs.

Acknowledgements This work has been supported by the Spanish Ministerio de Economía y Competitividad grant FIS2011-28838-C02-01.

References

1. Zener, C.: Proc. R. Soc. Lond. Ser. A **145**, 523 (1934)
2. Esaki, L., Tsu, R.: IBM J. Res. Dev. **14**, 61 (1970)
3. Feldmann, J., Leo, K., Shah, J., Miller, D.A.B., Cunningham, J.E., Meier, T., von Plessen, G., Schulze, A., Thomas, P., Schmitt-Rink, S.: Phys. Rev. B **46**, 7252 (1992)
4. Leo, K.: High-Field Transport in Semiconductor Superlattices. Springer Tracts in Modern Physics, vol. 187. Springer, Berlin (2003)
5. Hofbeck, K., Grenzer, J., Schomburg, E., Ignatov, A.A., Renk, K.F., Pavel'ev, D.G., Koschurinov, Yu., Melzer, B., Ivanov, S., Schaposchnikov, S., Kop'ev, P.S.: Phys. Lett. A **218**, 349 (1996)
6. Bonilla, L.L., Grahn, H.T.: Rep. Prog. Phys. **68**, 577 (2005)
7. Bonilla, L.L., Carretero, M.: In: Simos, T.E., Psihoyios, G., Tsitouras, Ch. (eds.) Numerical Analysis and Applied Mathematics, vol. 1048, p. 9. American Institute of Physics Proceedings, Melville (2008)
8. Bonilla, L.L., Alvaro, M., Carretero, M.: Phys. Rev. B **84**, 155316 (2011)
9. Bonilla, L.L., Escobedo, R., Perales, A.: Phys. Rev. B **68**, 241304(R) (2003)

MS 30
MINISYMPOSIUM: SHAPE AND SIZE IN BIOMEDICINE, INDUSTRY AND MATERIALS SCIENCE: AN ECMI SPECIAL INTEREST GROUP

Organizer

Alessandra Micheletti[1]

Speakers

Sergei Zuyev[2]
Optimisation on Measures with Applications to Material Design and Telecommunications

Alessandra Micheletti[1]
Mathematical Morphology Applied to the Study of Dual Phase Steel Formation

François Willot[3]
Mathematical Morphology of Mesoporous Alumina: Design from TEM Micrographs of a 3D Random Set Model

Katharina Losch[4]
Stochastic Modeling of Engineering Materials for Prediction of Spatial Mechanical Characteristics

[1]Alessandra Micheletti, Universita' degli Studi di Milano, Milano, Italy.
[2]Sergei Zuyev, Chalmers University of Technology, Gotheborg, Sweden.
[3]François Willot, Mines ParisTech, Centre for Mathematical Morphology, Fontainebleau, France.
[4]Katarina Loesch, Technical University of Kaiserslautern, Kaiserslautern, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_104

Keywords

Applications to biomedicine and materials science
Mathematical morphology
Statistical shape analysis
Stochastic geometry

Short Description

Statistical Shape Analysis and Stochastic Geometry deal with the geometrical information of families of objects in presence of stochasticity. Thanks to the development of information technologies, the last decades have seen a considerable growth of interest in the statistical theory of shape and its application to many and diverse scientific areas.

Often the diagnosis of a pathology, or the description of a biological process mainly depend on the shapes present in images of cells, organs, biological systems, etc., and mathematical models which relate the main features of these shapes with the correct outcome of the diagnosis, or with the main kinetic parameters are often still not present.

In material sciences, and industrial applications optimization for quality control requires mathematical models from Stochastic Geometry and the related statistical estimation procedures, and methods of Statistical Shape Analysis for comparison of different random geometrical patterns.

From the mathematical point of view, Shape Analysis and Stochastic Geometry use a variety of mathematical tools from differential geometry, geometric measure theory, stochastic processes, etc., dealing with both direct and inverse problems.

As far as applications are concerned, in this minisymposium topics which are relevant in biomedicine and material sciences will be emphasized.

Mathematical Morphology Applied to the Study of Dual Phase Steel Formation

Alessandra Micheletti, Junichi Nakagawa, Alessio A. Alessi, Vincenzo Capasso, Davide Grimaldi, Daniela Morale, and Elena Villa

Abstract Dual Phase steel (DP steel) has shown high potential for automotive and other applications, due to its remarkable combined properties of high strength and good formability. The mechanical properties of the material are strictly related to the spatial distribution of the two steel phases, ferrite and martensite, and with their stochastic geometry. Unfortunately the experimental costs to obtain images of sections of steel samples are very high, so that one important industrial problem is to reduce the required number of 2D sections in order to either reconstruct the 3D geometry of the material, or to simulate realistic ones. In this work we will present a germ-grain statistical model which can be used for a best fitting of the main geometric characteristics of the martensite phase. The parameters of the model are estimated on the basis of morphological characteristics of the images of about 150 tomographic sections taken from a real sample. After optimization or tuning of the relevant parameters, the statistical model can then be used to identify the minimum number of sections of the sample which are needed to estimate the parameters in a reliable way.

Keywords Dual phase steel • Germ-grain model • Mathematical morphology

A. Micheletti (✉)
Department of Mathematics and ADAMSS Center, Università degli Studi di Milano, Milano, Italy
e-mail: alessandra.micheletti@unimi.it

J. Nakagawa
Mathematical Science & Technology Research Lab, Nippon Steel & Sumitomo Metal, Tokyo, Japan

A.A. Alessi • D. Grimaldi • D. Morale • E. Villa
Department of Mathematics, Università degli Studi di Milano, Milano, Italy

V. Capasso
Gregorio Millan Institute, Escuela Politecnica Superior, Universidad Carlos III de Madrid, Leganes, Spain

ADAMSS Center, Università degli Studi di Milano, Milano, Italy

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_105

1 Introduction

Dual Phase steels (DP steels) have shown high potential for many applications due to their remarkable property combination between high strength and good formability.

Here we consider a sample of steel formed by martensite and ferrite. The relative position and geometric structure of the two phases is responsible of the mechanical properties of the material, thus it is particularly important to provide statistical models which may reproduce the main geometric characteristics of the two phases. Our results are based on images of about 150 tomographic sections taken from a lab sample of steel.

The formation of the two phases of the material starts after a cooling phase of the melted alloy of iron and carbon, during which austenite is formed, followed by a rolling phase, transforming slabs of steel into thin metal foils.

A further cooling phase follows the rolling; during this phase the formation of ferrite starts. Crystals of ferrite nucleate mainly from the interfaces of the rolled (and thus deformed) austenite, and grow up to impingement with other crystals of ferrite, driven by the evolving field of carbon concentration. After a fixed time interval the formation of ferrite is stopped by a sudden quenching, during which the material still not transformed into ferrite becomes martensite. The final result is a dual phase steel formed by ferrite and martensite, having a stochastic geometric structure.

In order to define a dynamical model able to reproduce the complete geometric structure of the material, a stochastic birth and growth process coupled with the evolution of the carbon field should be used (see [1] for a similar model applied to polymer crystallization). A first model which goes in this direction, though facing the problem at only a macroscopic scale, has been studied in [4].

In the following we will introduce a germ-grain model which may reproduce the main mean geometric characteristics of the martensite contained in the real sample.

As from a confidentiality agreement with Nippon Steel & Sumitomo Metal, who provided the real data, the images of the real sample will not be shown.

2 Structure of the Austenite Phase

First of all we studied the geometric structure of the interfaces of austenite after rolling, since nucleation of ferrite happens mainly on such interfaces, so that the location of the final ferrite and martensite crystals depends on the location of such interfaces.

The shape of the crystals of austenite before rolling is quite close to a 3D Voronoi tessellation. The rolling reduces the thickness of the rolled slab to 1/50 of the original thickness, but preserving the width and the total volume. The result is a long thin foil (see Fig. 1).

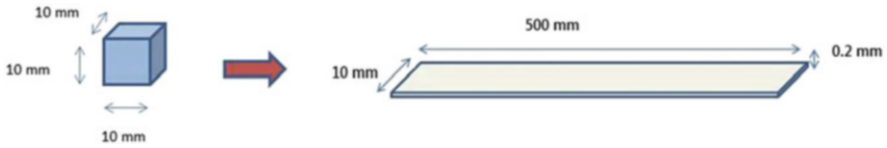


Fig. 1 Deformation applied to the cube containing the 3D Voronoi tessellation formed by austenite crystals

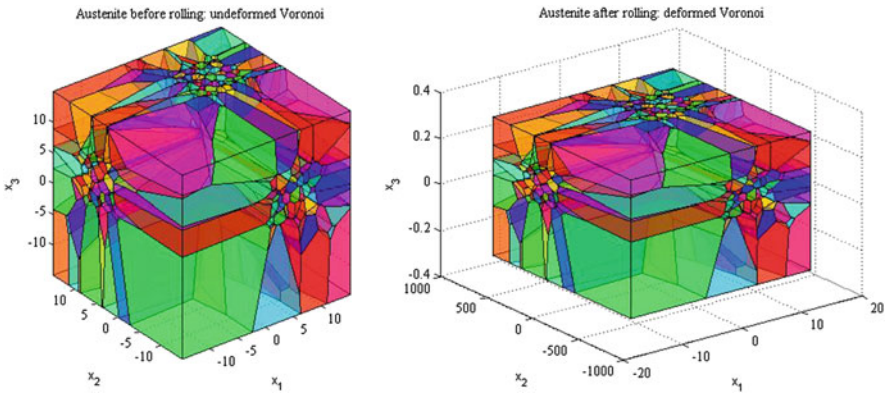


Fig. 2 Voronoi tessellation of the 3D space with 3000 crystals in a cube. On the *left*: before rolling, on the *right*: after rolling. The different axes scales in the two images evidentiate the deformation along the three axes, due to rolling

If we apply a deformation to a 3D Voronoi tessellation, maintaining the proportions used in the real experimental situation, we obtain the results shown in Fig. 2.

Since the real sample is composed by a very small portion of the rolled metal foil, extracted from the middle of the foil, we sectioned a small cube of the deformed tessellation, with dimensions proportional to the real sample, to look at its internal geometric structure. Two sections in the x_1x_3 and x_2x_3 direction, respectively, are reported in Fig. 3.

It is evident that, at the scales relevant for our application, the effect of rolling is to transform the interfaces of the 3D Voronoi tessellation into approximately parallel planes, having random quotes along the x_3 axis. The distribution of the quotes looks quite regular. This is in accordance with the images of sections of austenite after rolling, which shows a typical pancake structure (see Fig. 4).

We thus modelled the interfaces of austenite as parallel horizontal planes, i.e. parallel to direction x_1x_2, with randomly distributed quotes along the x_3 axis. The quotes have been distributed according to a 1-dimensional hard core process [3, Sect. 5.4], that is a point process with an inhibition distance between points. Figure 4 on the right shows a simulated realization of such a process.

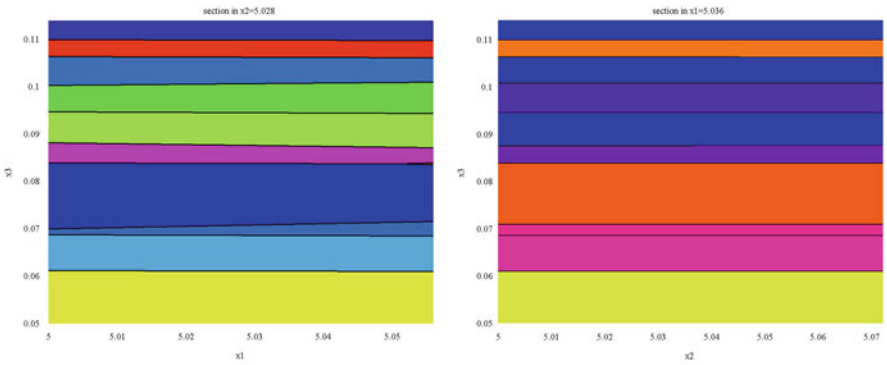


Fig. 3 Two sections of the deformed Voronoi tesselation on the *right* of Fig. 2, taken in the center of the parallelepiped. On the *left*: section parallel to the x_1x_3 plane; on the *right*: section parallel to the x_2x_3 plane. *Different colours* correspond to different crystals

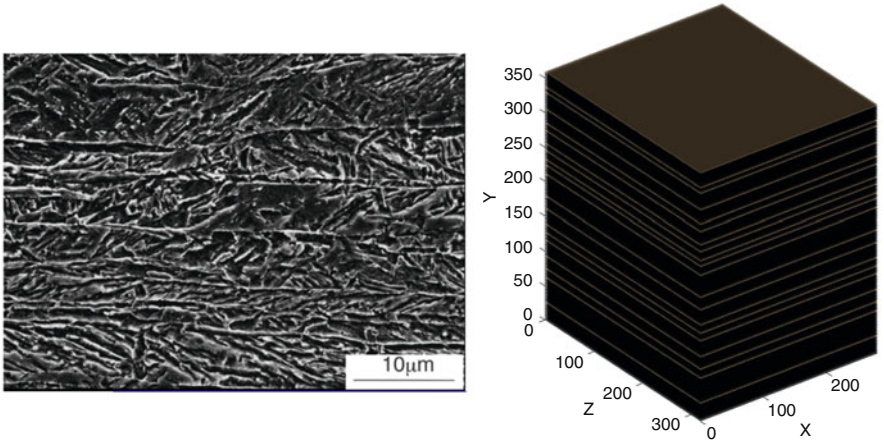


Fig. 4 On the *left*: pancake structure of austenite after rolling and before the birth and growth of ferrite; on the *right*: simulated random parallel planes, with an inhibition distance, from which the nucleation of ferrite starts

In Fig. 5 a simulated sample of the two phases which resembles the real one is reported. The black region, occupied by martensite, can be represented as the free space between different crystals of ferrite at the moment of quenching. Since the crystals of ferrite nucleate on the parallel planes representing the interfaces of austenite, martensite will have a tendency to concentrate in between two adjacent parallel planes.

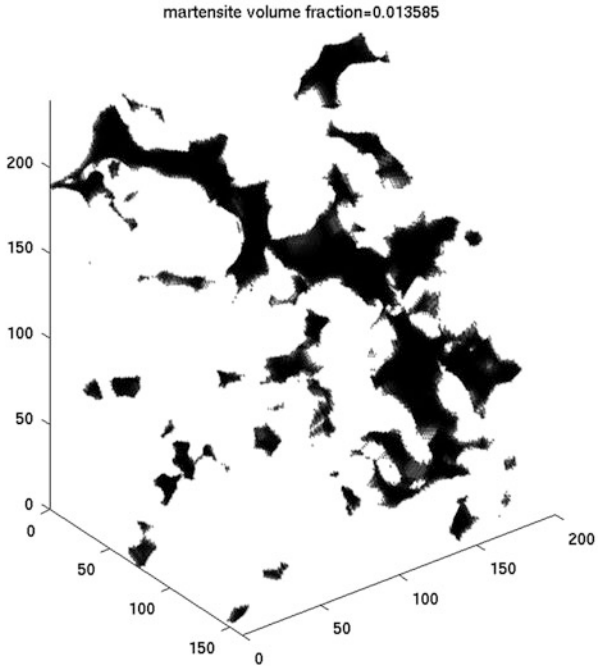


Fig. 5 A simulated sample of the two phases: the region occupied by martensite is depicted in *black*, while ferrite is in *white*

3 A Germ-Grain Model

In order to set up a statistical model able to reproduce the mean geometric structure of martensite, we have defined a germ-grain model with spherical grains, depending on a small set of unknown parameters.

A germ-grain model is a random closed set $\Xi \subseteq \mathbb{R}^d$ defined as

$$\Xi = \cup_{i \in \mathbb{N}} \Theta_i \oplus x_i$$

where $\{x_i\}$ are points in $\mathbb{R}^d$ forming a locally finite point process, called *germs*; Θ_i are i.i.d. uniformly bounded random closed sets (usually containing the origin) called *grains*, and $\oplus$ denotes the Minkowski sum between sets, thus $\Theta_i \oplus x_i = \{y + x_i | y \in \Theta_i\}$ (for more details see [2, 3]).

We modelled the point process of germs as a Neyman-Scott clustered point process, taking into account that martensite is formed in between the nucleation planes of ferrite, and also observing that martensite in the real sample exhibits a cluster structure.

The Neyman-Scott point process [3, Sect. 5.3] is formed by generating a spatial Poisson point process of *parents* having intensity $\lambda_p(x)$ and then surrounding

the parents by a random number of *daughter points*, scattered independently and identically distributed around the parents. The parents are then removed and the Neyman-Scott process is formed just by the daughter points.

The germs in our model have thus been generated according to the following algorithm:

Algorithm 1

Input:
n_{planes} = Number of parallel planes,
σ_{vert}= standard deviation of the daughters' distribution in vertical direction,
σ_{hor}=standard deviation of the daughters' distribution in horizontal direction,

1. randomly generate a set of parallel planes, as in Fig. 4, from which ferrite nucleates;
2. build up a set of "virtual" parallel planes, from which martensite originates, each located in the middle of two adjacent planes of the previous set;
3. distribute the parent germs uniformly on the "virtual" parallel planes;
4. distribute the daughter germs around the parents according to a 3-variate normal distribution having diagonal covariance matrix given by

$$\Sigma = diag(\sigma_{hor}^2, \sigma_{hor}^2, \sigma_{vert}^2).$$

The grains have been modeled as independent spheres of random radius $R = L{\cdot}\rho$, where L is a constant representing the maximum possible radius of the spheres and ρ is a random variable distributed as a $Beta(2, b)$ with $b > 2$. The reason of this choice is that in this way smaller spheres are privileged, providing to the germ grain model an aspect closer to the real sample.

The number of germs in the model is random, and has been generated according to the following procedure. Since the volume fraction occupied by martensite in the real sample is 0.027, we generated iteratively new germs and grains up to when the volume fraction occupied by the germ-grain model overcame 0.027. We made this choice since usually the percentage of martensite contained in steel is a known parameter, which can be measured even without sectioning the material.

In order to avoid edge effects, the simulation of the model has been performed in a window of observation enlarged by L on each side, and then only the central portion of the window with dimensions equal to the real sample has been considered. All the simulations have been performed in Matlab.

Thus our germ grain model is based on the following five parameters:

$$(n_{planes}, \sigma_{hor}, \sigma_{vert}, L, b) \in \mathbb{N} \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{N} \times [2, +\infty) \tag{1}$$

4 Parameters Estimate

In order to estimate the parameters of the model we considered the 2D images of the about 150 sections of the real sample and for each section we computed the following geometric characteristics:

1. A = the area of the region occupied by martensite, measured in pixels;
2. P = the perimeter of the region occupied by martensite, computed by calculating the distance between each adjoining pair of pixels around the border of the region;
3. A_{hull} = area of the convex hull of martensite;
4. D = diameter of a circle with the same area of martensite;
5. E = Euler number i.e. the number of connected components of martensite in the region minus the number of holes in the components;
6. O = orientation, i.e. the angle between the x-axis and the major axis of the ellipse that has the same second-moments as the region occupied by martensite;
7. M = length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region occupied by martensite;
8. m = length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region occupied by martensite.

For each set of the parameters (1) we performed 30 simulations of the germ grain model, and we sectioned the simulated images into the same number of parts of the real sample. On each simulated section we computed the eight geometric characteristics listed above.

The parameters can be estimated by minimizing a suitable distance between the values of the geometric characteristics measured on the real sample and the mean geometric characteristics of the simulated germ-grain model. Since we measured 8 variables for each section, resulting in a huge number of variables, we reduced the dimension of the problem by applying a principal component analysis to the $8 \times (n.\ of\ sections)$ variables, retaining only the first three principal components, which explain about the 40 % of the total variance of the simulations.

Let us denote by $\underline{G}_{real} \subseteq \mathbb{R}^3$, the first three PC's computed on the real sample, by $\bar{\underline{G}}_{sim}(\underline{p}) \subseteq \mathbb{R}^3$ the means, over 30 simulations, of the first three PC's computed on the simulated samples, and by S the sample covariance matrix of the first three PC's computed on the simulations.

For minimizing with respect to the parameters $\underline{p} = (n_{planes}, \sigma_{hor}, \sigma_{vert}, L, b)$, we have adopted the Mahalanobis distance

$$\Delta(\underline{G}_{real}, \bar{\underline{G}}_{sim}(\underline{p})) = \sqrt{(\underline{G}_{real} - \bar{\underline{G}}_{sim}(\underline{p}))^T S^{-1} (\underline{G}_{real} - \bar{\underline{G}}_{sim}(\underline{p}))}, \tag{2}$$

Table 1 Optimal values of the parameters. The corresponding computed Mahalanobis distance is 0.09. The simulations have been performed in a parallelepiped formed by $160 \times 200 \times 240$ voxels, where the side of each voxels is $0.25\,\mu$m long. The unit of measure of the parameters is a voxel side length

Parameter	Optimal value
n_{planes}	32
σ_{hor}	14.69
σ_{vert}	70.33
L	9
b	3.8

since it weights the distance from the mean of the simulations with the variance, taking thus into account the variability related with each set of parameters. Note that the distance (2) is stochastic, being based on the outcomes of random simulations.

Since both parameters n_{planes} and L are integers (the maximum side of the spheres has been defined in the simulation program in terms of number of voxels), we needed to apply an optimization algorithm to a function which is not expressed in algebraic form and depending upon mixed integer and real parameters, so that we have decided to apply a genetic algorithm for the best fitting

$$\underline{\hat{p}} = \arg\min_{\underline{p}} \Delta(\underline{G}_{real}, \bar{\underline{G}}_{sim}(\underline{p})).$$

The estimated parameters are reported in Table 1.

In Fig. 6 the following results obtained by simulating the germ grain model are reported: the top figure represents a "cloud" of 50 points with coordinates equal to the first three principal components of $\underline{G}_{sim}$, each computed on a different simulation performed with the optimal parameters. The black bold dot represents the first three principal components of $\underline{G}_{real}$, it is almost in the center of the cloud and very close to the mean of the principal components of the simulated patterns, showing thus a good agreement between the fitted model and the real data. The two bottom figures visualize the results of one single simulation of the germ grain model performed with the optimal parameters, and a section of the simulated pattern in direction orthogonal to the Z axis. These images must be compared with the corresponding ones of the real sample. At a first visual inspection, we may observe a rather good agreement between the simulation and the real sample, thus confirming the effectiveness of the adopted method.

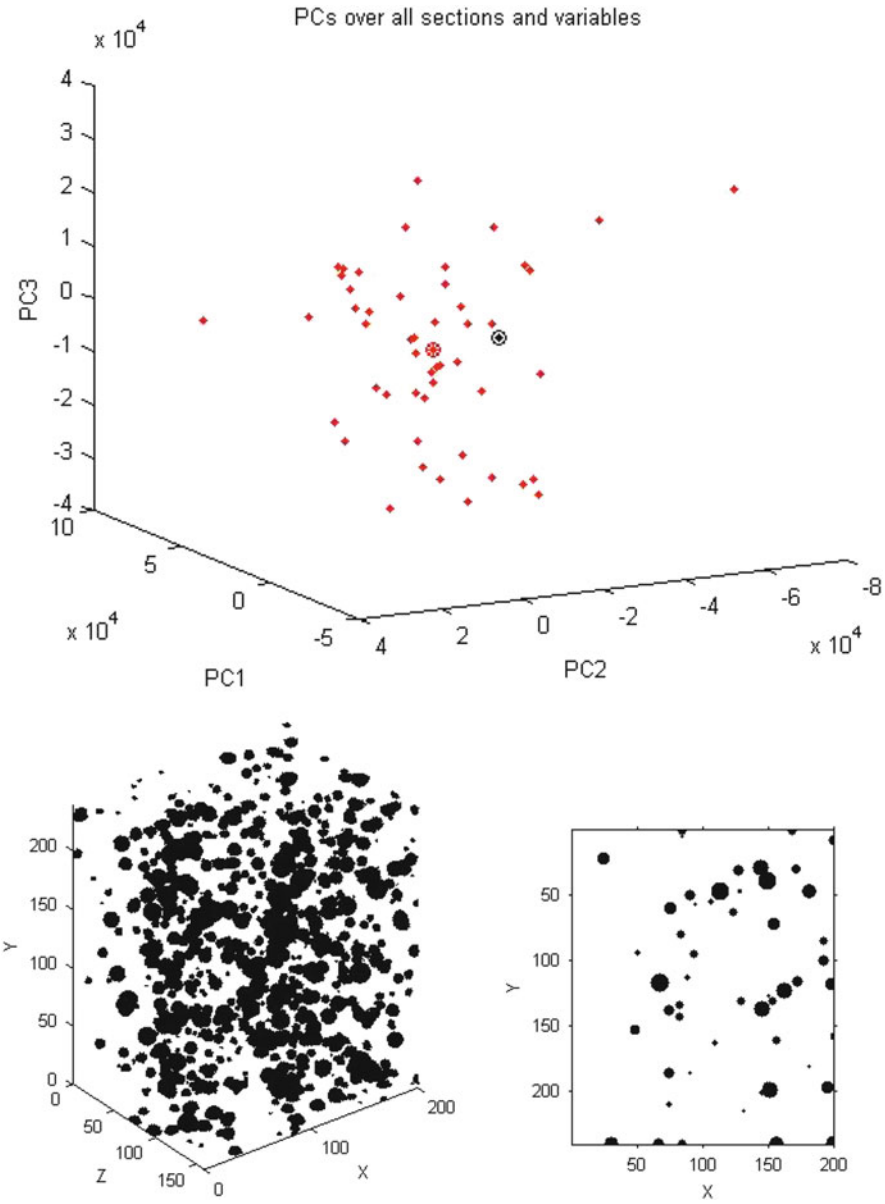


Fig. 6 *Top figure*: the first three PC's of 50 simulations performed with the optimal parameters (*red dots*), their mean (*red bold dot*), and the first three PC's of the real sample (*black bold dot*); *bottom left*: a simulation of the germ grain model with the optimal parameters; *bottom right*: a section orthogonal to Z direction of the simulation

References

1. Capasso, V. (ed.): Mathematical Modelling for Polymer Processing. Polymerization, Crystallization, Manufacturing. Mathematics in Industry, vol. 2. Springer, Heidelberg (2003)
2. Matheron, G.: Random Sets and Integral Geometry. Wiley, New York (1975)
3. Stoyan, D., Kendall, W.S., Mecke, J.: Stochastic Geometry and Its Application. Wiley, New York (1995)
4. Suwanpinij, P., Togobytska, N., Prahl, U., Weiss, W., Hömberg, D., Bleck, W.: Numerical cooling strategy design for hot rolled dual phase steel. Steel Res. Int. **81**, 1001–1009 (2010)

MS 31
MINISYMPOSIUM: SIMULATION AND OPTIMIZATION OF SOLAR TOWER POWER PLANTS

Organizers

Martin Frank[1] and Pascal Richter[2]

Speakers

Peter Schöttl[4], Raymond Branke[3], Tom Fluri[4], Anna Heimsath[5] and Peter Nitz[6]
Simulation of Solar Power Towers: A Techno-Economical Approach

Andreas Reinholz[7], Peter Schwarzboezl[8], Nils Ahlbrink[9] and Amadeus Rong[10]
Heliostat Aim Point Optimization During Operation of Solar Tower Power Plants
Optical Simulation of Solar Tower Power Plant

[1]Martin Frank, RWTH Aachen University, Center for Computational Engineering Science, Aachen, Germany.

[2]Pascal Richter, RWTH Aachen University, Center for Computational Engineering Science, Aachen, Germany.

[3]Raymond Branke, Fraunhofer ISE, Freiburg, Germany.

[4]Tom Fluri, Fraunhofer ISE, Freiburg, Germany.

[5]Anna Heimsath, Fraunhofer ISE, Freiburg, Germany.

[6]Peter Nitz, Fraunhofer ISE, Freiburg, Germany.

[7]Andreas Reinholz, German Aerospace Center (DLR), Solar Research, Cologne, Germany.

[8]Peter Schwarzboezl, German Aerospace Center (DLR), Solar Research, Cologne, Germany.

[9]Nils Ahlbrink, German Aerospace Center (DLR), Solar Research, Cologne, Germany.

[10]Amadeus Rong, German Aerospace Center (DLR), Solar Research, Cologne, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_106

Pascal Richter[2], Martin Frank[1] and Erika Ábrahám[3]
Multi-Objective Optimization of Solar Tower Heliostat Fields

Keywords

Solar towers
Sustainable energy

Short Description

Solar towers use many flat mirrors to concentrate sun light on the absorber, which is mounted on a tower. The simulation of solar tower power plants play an important role in the planning stage of a project. The goal is to find the most efficient arrangement of mirrors that balances power production against construction costs.

Unfortunately, the high computational costs associated with accurate enough simulation tools has in the past made optimizations impractical in most settings. However, recent advances in computer resources and numerical algorithms are making optimization processes possible. The trend of using more and more simulation/optimization tools in the field of solar thermal power plants is expected to continue in the future.

The purpose of this mini-symposium is to report on the continuing progress of simulation and optimization strategies for solar tower plants. It brings together researchers from applied mathematics, computational science, physics, and engineering communities and is designed specifically as a forum for researchers in earlier stages of their career to discuss their work and exchange ideas.

Multi-Objective Optimization of Solar Tower Heliostat Fields

Pascal Richter, Martin Frank, and Erika Ábrahám

Abstract We introduce a model to compute the annual performance of a heliostat field. We take into account topography, tracking errors, and the position and intensity of the sun. An approach is introduced, which improves on the otherwise expensive pairwise comparison to calculate shading and blocking. Because the computational time is reduced significantly, the presented implementation is sufficiently fast to allow for heliostat field layout optimization within a couple of hours. The optimization is executed via a genetic algorithm, which optimizes the heliostat positioning parameters as well as other design parameters, e.g. receiver tilt angle. A novel approach is used to reduce the search domain. Because the search domain delivers several local optima with comparable values of the objective function, the objective function is augmented. We use smoothing functionals to disperse the local optima. A field layout is optimized on a hilly ground in South Africa, with additional constraints on the heliostat positions.

Keywords Heliostat fields • Multiobjective optimization • Solar towers

1 Introduction

Solar tower plants generate electric power from sunlight by focusing concentrated solar radiation on a tower-mounted receiver, see Fig. 1. The collector system uses hundreds or thousands of sun-tracking mirrors called heliostats, to reflect the incident sunlight onto the receiver where a fluid is being heated up. Today's receiver types use water/steam, air or molten salt to transport the heat. Usually, the heat of the fluid is exchanged into steam which powers a turbine to generate electricity.

P. Richter (✉) • M. Frank
Center for Computational Engineering Sciences, RWTH Aachen University, Schinkelstrasse 2, 52062 Aachen, Germany
e-mail: richter@mathcces.rwth-aachen.de; frank@mathcces.rwth-aachen.de

E. Ábrahám
Chair of Computer Science 2, RWTH Aachen University, Ahornstrasse 55, 52074 Aachen, Germany
e-mail: abraham@informatik.rwth-aachen.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_107

Fig. 1 Solar tower plant PS10, 11 MW in Andalusia, Spain [source: flickr]

Solar tower plants are not yet cost-competitive [6]. Therefore, concentrating solar thermal power plant markets and projects today only evolve where a political framework ensures financial incentives. For commercial solar tower project developments, a conceptual plant design has to be determined in an early planning stage. Designing commercial power plants aims always at finding the most economic plant design under a given set of constraints.

In this paper a model and optimisation algorithm for heliostat field layout is introduced. The underlying solar tower model is presented in Sect. 2. Because the model is used in an optimisation process, a computationally efficient calculation of insolation with enough accuracy is needed.

2 Ray-Tracing Model

A solar field is given by N heliostats H_i, each with an area A_i. For the time-dependent solar angles θ_{solar} and γ_{solar}, and the direct normal irradiation I_{DNI}, the ray-tracing model computes the received optical radiation over a year, while taking cosine effects $\eta_{\cos}$, shading and blocking η_{sb}, heliostat reflectivity η_{ref}, atmospheric attenuation η_{aa} and spillage losses η_{spl} into account. For each heliostat H_i the time dependent received optical radiation is defined by

$$P^i(t,d) = A_i \cdot I_{\text{DNI}}(t,d) \cdot \eta_{\cos,i}(t,d) \cdot \eta_{\text{sb},i}(t,d) \cdot \eta_{\text{ref},i}(t,d) \cdot \eta_{\text{aa},i}(t,d) \cdot \eta_{\text{spl},i}(t,d), \tag{1}$$

at time t of the day d. At the end of this section the annual received radiation of the full plant is computed, which depends on the sunrise and sunset of every day in the year. We essentially use the same model as [7], the main difference being that we use a hierarchical ray-tracing method, and speed-up the computation of shading and blocking effects.

2.1 Hierarchical Ray-Tracing Method

The rays have their origin in the sun, hit the surface of a heliostat and are reflected in direction of the receiver. We are interested in the reflected power of a heliostat, which is hitting the receiver. To detect the optical flux over the heliostat's surface we are using a hierarchical approach of ray-tracing methods [1, 7], where the complete flux is computed by numerical integration with the use of Gauss-Legendre quadrature rule. Thus, the surface is partitioned in a number of regions, each with a representative ray. The influence on the reflection by shading, blocking and ray interception at the receiver is determined just for this single ray as representative for the whole region. Each area is weighted by the irradiance of its representative ray. Finally all values are summed to get the power of the heliostat. The number of representative rays per heliostat is given by the selected order of the Gaussian quadrature rule.

2.2 Shading and Blocking

For each representative ray of a heliostat, shading and blocking effects by neighbouring heliostats or the tower must be detected. This is the most expensive part of a simulation. The brute-force approach of a pairwise comparison of each ray with all heliostats is computationally expensive. The computational complexity can be reduced by only considering a subset of heliostats that can potentially shade or block a heliostat [1]. To determine this subset, a data structure is needed which is fast in nearest-neighbour search and in range-search.

Therefore, for better performance, a two-dimensional bitboard index structure is used. The idea is to cover the two-dimensional $\boldsymbol{x}$-$\boldsymbol{y}$ space with an equidistant grid such that the space is sub-divided in distinct quadratic cells. Inside those cells the information is stored if nearby is a heliostat.

For a nearest-neighbour search, just the surrounding cells around a cell have to be checked, instead of all heliostats. The same holds for a range-search, where e.g. all heliostats in one direction are wanted. Just the containing cells of the range have to be checked. In some test cases we could accelerate the simulation by factor 100.

2.3 Annual Received Optical Radiation

The annual received optical radiation of the whole power plant is given by the sum of the annual received optical radiation of all heliostats H_i,

$$E_{\text{year}} = \sum_{i=1}^{N} E_{\text{year}}^{i} = \sum_{i=1}^{N} \sum_{d=1}^{365} \left(\int_{\text{sunrise}}^{\text{sunset}} P_i(t, d) \, \mathrm{d}t \right), \tag{2}$$

with power P_i given in Eq. (1). The sunrise and the sunset depend on the day d. The value of the received optical radiation over a year E_{year}, is the basis for each objective function in the optimisation process, see Sect. 3. For each different configuration of the solar field, this value has to be computed by a simulation.

The time integral from sunrise to sunset in (2) is solved numerically. In common practice, an iteration with constant time step [3, 5, 10] is used. Noone et al. [7] propose an iteration with constant solar angle step, which allows the same accuracy but needs fewer iterations. Both approaches approximate the time integral with midpoint rule. For higher accuracy other numerical quadrature rules are recommended. The herein proposed Gauss-Legendre quadrature rule uses non-constant time (or solar angle) steps:

$$\int_{a:=\text{sunrise}}^{b:=\text{sunset}} P_i(t, d) \, \mathrm{d}t \approx \frac{b-a}{2} \sum_{i=1}^{n} w_i \cdot f\left(\frac{b-a}{2} t_i + \frac{a+b}{2} \right), \tag{3}$$

with n Gaussian time-abscissas t_i and Gaussian weights w_i. Additionally the sum of the days can be approximated by using a sort of trapezoidal rule with just $m \in \{1, 2, \ldots, 365\}$ days.

3 Optimisation

Various effects—cosine effects, shading and blocking of heliostats (presented in Sect. 2)—reduce the efficiency of the solar tower. An objective of an optimisation is to discover an optimal positioning of the heliostats in the field. In the literature, the general structure of the heliostat arrangement is predefined by assumptions, e.g. radial staggered, circles or spirals [7–10]. In these cases, an optimisation means to find an assignment of about two to four parameters which define the structure, e.g. radius or angle of a spiral. However, the assumption of the structure leads to many comparable local optima [9]. In addition, these optimisations generate a regular or symmetric structure which could be suitable for nearly flat areas but not for a hilly topology.

In this work, we introduce an approach where the heliostats' alignment does not depend on any structure. Namely, the heliostats obtain the highest amount of

freedom in order to find their optimal position. For that purpose, we use a genetic algorithm [2, 4] with a novel genotype-representation which reduces the search domain of the algorithm. The only restriction of the approach is that neighbouring heliostats must be separated by a minimum distance in order to prevent a collision.

3.1 Genetic Algorithm

The functionality of a genetic algorithm is inspired by the biological evolution. A population of candidate solutions, called individuals, evolves in order to provide better solutions for an optimisation problem. Each individual has a set of properties, called genotypes or genomes. Usually, a genotype is represented as an array of several genes such that a unique assignment of gene and property exists. Two or more individuals are combined by mixing the genotypes gene by gene in order to generate a new population. Using this approach for the position of heliostats, the sets of heliostats from different heliostats are "ordered" by an artificial identifier. This identifier is defined by the position of the corresponding gene in the array. To generate a new population of individuals from a set of evaluated individuals, four main operations are used by the genetic algorithm: First, two or more individuals are randomly selected by roulette-wheel method from the old population according to their fitness values. The properties of the selected individuals are combined according to the fitness value of their heliostats. Therefore the heliostats of all parent individuals are sorted in descending order according to this value. Successively the best heliostats are picked for the new individual. If any selected heliostat causes a conflict, it is neglected and the next best heliostat is picked. In case that there are no more heliostats, the remaining heliostats are generated by random. Afterwards, the heliostats are mutated by locally change their position. The whole population is simulated to get the fitness values for the new individuals. The algorithm terminates if a stop criterion is satisfied, e.g. maximum simulation time or maximum number of generations.

3.2 Objective Functions

The purpose of our optimisation is to find an individual which has a high efficiency. But additionally we aim to reward a solution which looks "nice", which means, that the distribution of the heliostat positions are somehow smooth. We scalarise our objective function, i.e. we look for a solution

$$\max_{\mathscr{I}\in\mathscr{P}} F(\mathscr{I}) = \max_{\mathscr{I}\in\mathscr{P}} \left(\sum_{j=1}^{n} w_j \cdot \frac{f_j(\mathscr{I}) - \min\limits_{\mathscr{I}\in\mathscr{P}} f_j(\mathscr{I})}{\max\limits_{\mathscr{I}\in\mathscr{P}} f_j(\mathscr{I}) - \min\limits_{\mathscr{I}\in\mathscr{P}} f_j(\mathscr{I})} \right), \tag{4}$$

All objective functions f_j are normalised by the minimum and maximum value of the whole population $\mathscr{P}$, so that the normalised value lies in the range between 0 and 1. The weights of the objectives $w_j > 0$, with $\sum_{j=1}^{n} w_j = 1$, are the parameters of the scalarisation. Every objective function f_j has to be maximised. If there is an objective function $\hat{f}_j$ that should be minimised, we set the corresponding objective function as $f_j := -\hat{f}_j$ which is maximised.

The model described in Sect. 2 delivers in Eq. (2) the annual received optical radiation, which is used as objective function

$$f_1(\mathscr{I}) := E_{\text{year}}(\mathscr{I}). \tag{5}$$

To reward solutions, which are looking "nice", additional smoothing functionals are created. The variance of the k-nearest-neighbour distance is given by,

$$f_2(\mathscr{I}) = -\iint |\nabla \mathsf{KNN}|^2 \,\mathrm{d}x\,\mathrm{d}y \approx -\sum_{T\in\mathscr{T}} A_T \cdot \left(\frac{\partial \mathsf{KNN}(T)}{\partial x} + \frac{\partial \mathsf{KNN}(T)}{\partial y}\right)^2, \tag{6}$$

$\mathscr{T}$ denotes the triangulation of the heliostats H_i in the x-y plane. A_T is the area of a triangle $T \in \mathscr{T}$. For each heliostat H_i the k-nearest-neighbour distance is given by

$$\mathsf{KNN}_i = \sum_{H_\ell \in \mathscr{N}_k(H_i)} |\boldsymbol{p}_i - \boldsymbol{p}_\ell| \tag{7}$$

where $\mathscr{N}_k(H_i)$ is the set of the k nearest neighbours of H_i and $\boldsymbol{p}_i$ and $\boldsymbol{p}_\ell$ are the positions of the heliostats. By linear interpolation the k-nearest-neighbour distance is piecewise defined for each triangle $T \in \mathscr{T}$ which is denoted by $\mathsf{KNN}(T)$.

Another smoothing functional is the density distribution, which is given by

$$f_3(\mathscr{I}) = -\iint |\nabla \rho|^2 \,\mathrm{d}x\,\mathrm{d}y \approx -\sum_{T\in\mathscr{T}} A_T \cdot \left(\frac{\partial \rho^r(T)}{\partial x} + \frac{\partial \rho^r(T)}{\partial y}\right)^2. \tag{8}$$

For each heliostat H_i the density is given by

$$\rho_i^r = \left|\{H_\ell \in \mathscr{I} \;\mid\; |\boldsymbol{p}_i - \boldsymbol{p}_\ell| \le r\}\right| \tag{9}$$

where ρ_i^r is the number of H_i-neighbouring heliostats inside a defined radius r. Again by linear interpolation the density is piecewise defined for each triangle $T \in \mathscr{T}$ which is denoted by $\rho^r(T)$.

The variance of the kNN distance and the density distribution functionals aim to create a field of equally distributed or dense heliostats. The importance of the smoothing functionals can be adjusted by using the weights described in Eq. (4).

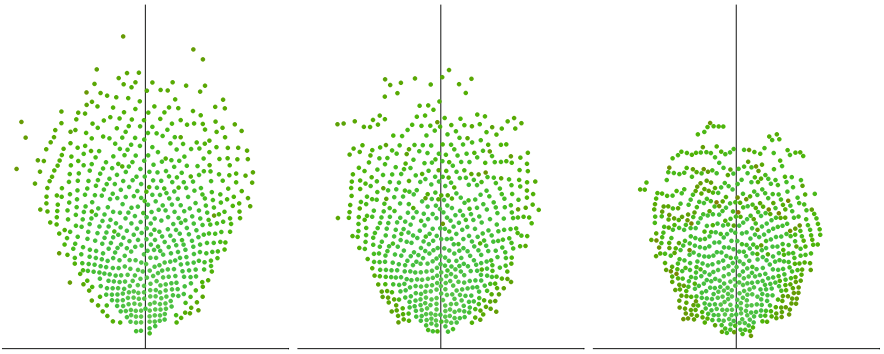


Fig. 2 Comparison of the best power plant configurations after optimisation with different weights for the density and kNN functional. The color gradient from *green* to *red* shows the annual received optical radiation for each heliostat

3.3 Testing the Genetic Algorithm

By combining the three functionals of energy, kNN and density, different results are reached, see Fig. 2. The produced energy is high for all combinations. By taking kNN and density into account, the optimisations yields a field in which the heliostats stand closer together and are evenly distributed. The single outliers that occur due to mutation could be eliminated during a post-processing step.

Using the smoothing functionals, it is possible to create "nicer looking" solutions that result in comparable energy production. For further fine-tuning one could also try to gradually adjust the weight of the functional during the course of the optimisation.

4 Application

With the above introduced optimisation algorithm a solar power plant can be optimized. To optimize the heliostats alignment of a planned pilot plant in South Africa we had to extend the model in such a way, that the heliostats can be grouped by a joint pod system, where they are positioned on an arbitrary truss construction. So, instead of positioning single heliostats, groups of heliostats with fixed relative positions are placed on the field. The pod systems are not allowed to touch each other, this includes all heliostats and the truss construction. Figure 3 shows the distribution of the heliostats before and after the optimisation.

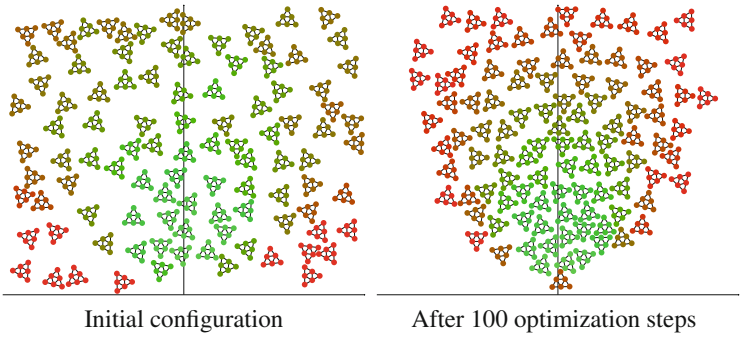


Fig. 3 Optimisation of a solar power plant with pod systems

References

1. Belhomme, B., Pitz-Paal, R., Schwarzbözl, P., Ulmer, S.: A fast ray tracing tool for high-precision simulation of heliostat fields. Sol. Energy Eng. Trans. ASME **131**(3), 031002 (2009)
2. De Jong, K.: An analysis of the behavior of a class of genetic adaptive systems. Ph.D. thesis, University of Michigan (1975)
3. Elsayed, M., Allah, H., Al-Rabghi, O.: Yearly-averaged daily usefulness efficiency of heliostat surfaces. Sol. Energy **49**(2), 111–121 (1992)
4. Holland, J.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press, Cambridge (1975)
5. Kistler, B.: A user's manual for DELSOL3: a computer code for calculating the optical performance and optimal system design for solar thermal central receiver plants. Tech. rep., Sandia National Labs., Livermore, CA (1986)
6. Morin, G.: Techno-economic design optimization of solar thermal power plants. Ph.D. thesis, Technische Universität Braunschweig (2010)
7. Noone, C., Torrilhon, M., Mitsos, A.: Heliostat field optimization: a new computationally efficient model and biomimetic layout. Sol. Energy **86**, 792–803 (2012)
8. Noone, C.J., Ghobeity, A., Slocum, A.H., Tzamtzis, G., Mitsos, A.: Site selection for hillside central receiver solar thermal plants. Sol. Energy **85**(5), 839–848 (2011)
9. Pitz-Paal, R., Botero, N., Steinfeld, A.: Heliostat field layout optimization for high-temperature solar thermochemical processing. Sol. Energy **85**(2), 334–343 (2011)
10. Yao, Z., Wang, Z., Lu, Z., Wei, X.: Modeling and simulation of the pioneer 1mw solar thermal central receiver system in china. Renew. Energy **34**(11), 2437–2446 (2009)

MS 32
MINISYMPOSIUM: SIMULATION AND OPTIMIZATION OF WATER AND GAS NETWORKS

Organizers

Gerd Steinebach[1], Oliver Kolb[2] and Jens Lang[3]

Speakers

Jens Lang[4], Pia Domschke[5], Oliver Kolb[6]
Adaptive Modelling, Simulation and Optimization of Water and Gas Supply Networks

Raul Borsche[7] and Jochen Kall[8]
ADER Schemes on Networks of Hyperbolic PDEs

[1]Gerd Steinebach, Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany.
[2]Oliver Kolb, University of Mannheim, Germany.
[3]Jens Lang, Technical University of Darmstadt, Germany.
[4]Jens Lang, Technical University of Darmstadt, Germany.
[5]Pia Domschke, Technical University of Darmstadt, Germany.
[6]Oliver Kolb, University of Mannheim, Germany.
[7]Raul Borsche, University of Kaiserslautern, Germany.
[8]Jochen Kall, University of Kaiserslautern, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_108

Oliver Kolb[9], Pia Domschke[5], Jens Lang[10], Björn Geißler[11], Antonio Morsi[12], Alexander Martin[13]
Combination of Linear and Nonlinear Programming Techniques for Optimization Problems in Gas and Water Supply Networks

Gerd Steinebach[14]
From River Rhine Alarm Model to Water Supply Network Simulation by the Method of Lines

Alfredo Bermúdez[15], Julio González-Díaz[16], Francisco J. González-Diéguez[17] and Ángel M. González-Rueda[18]
Modeling and Optimization of Gas Networks

Yi Lu[19], Nicole Marheineke[20] and Jan Mohring[21]
MOR Via Quadratic-Linear Representation of Nonlinear-Parametric PDEs

Tim Jax[22] and Gerd Steinebach[23]
ROW Methods Adapted to Network Simulation for Fluid Flow

Tanja Clees[24], Kläre Cassirer[25], Bernhard Klaaßen[26], Igor Nikitin[27], Lialia Nikitina[28]
Simulation and Analysis of Gas Networks with MYNTS

[9]Oliver Kolb, University of Mannheim, Germany.
[10]Jens Lang, Technical University of Darmstadt, Germany.
[11]Björn Geißler, Friedrich-Alexander-University Erlangen-Nürnberg, Germany.
[12]Antonio Morsi, Friedrich-Alexander-University Erlangen-Nürnberg, Germany.
[13]Alexander Martin, Friedrich-Alexander-University Erlangen-Nürnberg, Germany.
[14]Gerd Steinebach, Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany.
[15]Alfredo Bermúdez, University of Santiago de Compostela, Spain.
[16]Julio González-Díaz, University of Santiago de Compostela, Spain.
[17]Francisco J. González-Diéguez, University of Santiago de Compostela, Spain.
[18]Ángel M. González-Rueda, University of Santiago de Compostela, Spain.
[19]Yi Lu, Friedrich-Alexander-University Erlangen-Nürnberg, Germany.
[20]Nicole Marheineke, Friedrich-Alexander-University Erlangen-Nürnberg, Germany.
[21]Jan Mohring, Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany.
[22]Tim Jax, Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany.
[23]Gerd Steinebach, Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany.
[24]Tanja Clees, Fraunhofer SCAI, Sankt Augustin, Germany.
[25]Kläre Cassirer, Fraunhofer SCAI, Sankt Augustin, Germany.
[26]Bernhard Klaaßen, Fraunhofer SCAI, Sankt Augustin, Germany.
[27]Igor Nikitin, Fraunhofer SCAI, Sankt Augustin, Germany.
[28]Lialia Nikitina, Fraunhofer SCAI, Sankt Augustin, Germany.

Keywords

Energy supply and consumption
Gas networks
Water supply networks

Short Description

Large networks for fluid flow of water and gas can consist of rivers or channels, gas and water supply or sewer systems. The flow in one single network element is usually modelled by hyperbolic conservation laws or some simplifications. All single flow reaches must be coupled by appropriate coupling and boundary conditions. This approach leads to PDAEs (partial differential algebraic equations) and requires very robust and efficient numerical methods for their solution. Moreover, the optimization of the network operation with respect to the security of supply or energy consumption is of importance. Suitable optimization methods for these requirements are an active field of research.

This symposium will brought together researchers from across Europe who develop and apply mathematical models and numerical solution techniques in this context. The Symposium provided a platform for new methodologies, ideas and applications.

From River Rhine Alarm Model to Water Supply Network Simulation by the Method of Lines

Gerd Steinebach

Abstract In this paper an overview on modelling techniques and numerical methods applied to problems in water network simulation is given. The considered applications cover river alarm systems (Rentrop and Steinebach, Surv Math Ind 6:245–265, 1997), water level forecast methods (Steinebach and Wilke, J CIWEM 14(1):39–44, 2000) up to sewer and water supply networks (Steinebach et al., Mathematical Optimization of Water Networks Martin. Springer, Basel, 2012).

The hyperbolic modelling equations are derived from mass and momentum conservation laws. A typical example are the well known Saint-Venant equations. For their numerical solution a conservative semi-discretisation in space by finite differences is proposed. A new well-balanced space discretisation scheme is presented which improves the local Lax-Friedrichs approach applied so far. Higher order discretisations are achieved by WENO methods (Kurganov and Levy, SIAM J Sci Comput 22(4):1461–1488, 2000).

Together with appropriate boundary and coupling conditions this method of lines approach leads to an index-one DAE system. Efficient solution of the DAE system is the topic of Jax and Steinebach (ROW methods adapted to network simulation for fluid flow, in preparation).

Keywords Method of lines • River alarm systems • Water supply networks

1 Introduction to Water Network Simulation

Many practical problems exist where simulation of water networks is a key issue. These problems cover questions concerning water quality and water quantity. At first a water quality problem is considered, which became important after an accident in a chemical plant at river Rhine in 1986, the so called Sandoz accident. The aim was to develop a river alarm model in order to compute the transport and diffusion

G. Steinebach (✉)
Department 03 (EMT), Hochschule Bonn-Rhein-Sieg, Grantham-Allee 20, 53757 Sankt Augustin, Germany
e-mail: gerd.steinebach@h-brs.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_109

of chemical substances released to the river or its tributaries. In order to set up a mathematical model, information about the flow velocity of the river system and the diffusion and decay of the substances must be known. The latter parameters were estimated from tracer experiments which took place in rivers Rhine, Moselle and Elbe in the early nineties [6]. One possibility to compute flow velocities is to set up a separate water quantity model. A suitable modelling approach is to consider mass and momentum conservation in one space dimension for the whole river system. This approach leads to the well know Saint-Venant equations [4]. Beside the computation of flow velocities these equations can be applied to water level forecast models as well. Such forecast models are important for flood warning or low water situations, where navigation is suffering from [14].

These concepts for river system simulation can be transferred to sewer and water supply networks. In contrast to free surface flow in rivers, in water supply pipes pressure flow is dominating. In Sect. 2 a modelling approach covering both flow types according to [5] is presented. A suitable numerical solution strategy is given in Sect. 3. This solution approach is an improvement of the successfully applied methods in river alarm models and water level forecast models of rivers Rhine, Danube and Odra of the German Federal Institute of Hydrology [11, 17].

In order to preserve steady state solutions of the conservation equations a new well-balanced enhancement of the local Lax-Friedrichs method is proposed. Finally, in Sect. 4 numerical examples are presented.

2 Modelling Equations

River alarm models rely on the convection-diffusion-reaction equation for concentration $c(x,t)$ of the chemical substance. Tracer experiments suggest an improvement through the coupling with dead-zones, where no transport takes place. These assumptions lead to the model equations [7]

$$\partial_t c = -u\partial_x c + \frac{1}{A}\partial_x(DA\partial_x c) - Kc + \frac{A_0}{A}\tilde{K}(m-c)\ , \tag{1}$$

$$\partial_t m = \tilde{K}(c-m)\ , \tag{2}$$

with flow velocity u, flooded cross sections A in the main stream and A_0 in the dead zones, diffusion coefficient D, linear decay rate K, dead zone concentration $m(x,t)$ and exchange coefficient $\tilde{K}$.

In order to compute the velocities $u(x,t)$ and flooded cross sections $A(x,t)$ the Saint-Venant equations [4] are considered:

$$\partial_t A + \partial_x Q = 0\ , \tag{3}$$

$$\partial_t Q + \partial_x\left(\frac{Q^2}{A}\right) + gA\partial_x z = -gAS_f\ . \tag{4}$$

$Q(x,t) = u \cdot A$ denotes the flow discharge, $z(x,t)$ the water surface elevation above some reference level, g the gravitational constant and expression S_f is called friction slope. An empirical formula (by Manning-Strickler) reads $S_f = \frac{u \cdot |u|}{K_{St}^2 h^{4/3}}$ with water depth h and friction coefficient K_{St}. The surface level z must be computed from A and the river bed elevation $S_0(x)$ and thus is a function of A and x, i.e. $z(x,t) = f_z(x, A(x,t))$.

For free surface flow the density ρ of water is constant. To consider pressure flow in a pipe, a non constant density $\rho(x,t)$ is assumed. Furthermore, flooded cross sectional area A does not depend on time and is identical to the whole pipe cross section denoted by $\bar{A}$, which may vary with space ($A = \bar{A}(x)$). The idea of Bourdarias and Gerbi [3] is to consider pressure p as a combination of hydrostatic and overload pressure due to the assumed compressibility of water:

$$p = \rho g h + \frac{1}{\beta} \frac{\rho - \rho_0}{\rho_0} .$$

β denotes isothermal compressibility and ρ_0 density of water under free flowing conditions. It is assumed that $\rho_0 = 1000$ and $\beta = 5 \cdot 10^{-10}$ [1].

With these assumptions the equations for pressure and free surface flow read [16]

$$\partial_t(\rho A) + \partial_x(\rho Q) = 0 , \tag{5}$$

$$\partial_t(\rho Q) + \partial_x \left(\frac{(\rho Q)^2}{\rho A} \right) + g \rho A \partial_x z + \frac{A}{\beta \rho_0} \partial_x \rho = -g \rho A S_f . \tag{6}$$

If $\frac{(\rho A)}{\rho_0} > \bar{A}$, pressure flow exists with density $\rho = \frac{(\rho A)}{\bar{A}}$. Otherwise free surface flow is assumed with cross sectional area $A = \frac{(\rho A)}{\rho_0}$. Therefore, the computation of A and ρ from state variable (ρA) leads to the non smooth functions

$$\rho = \begin{cases} \frac{(\rho A)}{\bar{A}} & \text{if } \frac{(\rho A)}{\rho_0} > \bar{A} \\ \rho_0 & \text{otherwise} \end{cases} \quad ; \quad A = \frac{(\rho A)}{\rho} . \tag{7}$$

Depart from the diffusion term in (1), systems [(1), (2)], [(3), (4)] and [(5), (6)] can be expressed as a mixture of a conservative and a quasi-linear system of type

$$\partial_t q + M(q) \frac{d}{dx} f(x,q) = S(x,q) . \tag{8}$$

For (5), (6) $q = (\rho A, \rho Q)^T$, $f(x,q) = (\rho Q, \frac{(\rho Q)^2}{\rho A}, z, \rho)^T$, $M(q) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & g\rho A & \frac{A}{\beta \rho_0} \end{pmatrix}$, $S(q) = (0, -g\rho A S_f)^T$.

3 Numerical Solution Approach

The considered hyperbolic systems of type (8) must be completed with initial values (IVs) and boundary conditions (BCs):

$$\partial_t q + M(q)\frac{d}{dx}f(x,q) = S(x,q)\ ; \quad x \in [x_L, x_R]\ , \quad t > t_0\ , \tag{9}$$

$$\text{BC: } g(t, q(x_L,t), q(x_R,t)) = 0\ ; \quad \text{IV: } q(x,t_0) = q_0(x)\ . \tag{10}$$

3.1 Conservative Finite Difference Space Discretisation and DAE Solver

In the common method of lines (MOL) approach with a conservative finite difference space discretisation, at first the space interval is discretised into N cells according to $x_L = x_{1/2} < x_{3/2} < \ldots < x_{N+1/2} = x_R$ with constant width $\Delta x = x_{i+1/2} - x_{i-1/2}$ and cell centers $x_i = \frac{1}{2}(x_{i-1/2} + x_{i+1/2})$.

For each cell center $i = 1, \ldots, N$ the variable $q_i(t) = q(x_i, t)$ is defined and Eq. (9) leads to the ODE system

$$\frac{d}{dt}q_i + M(q_i)\frac{1}{\Delta x}(f_{i+1/2} - f_{i-1/2}) = S(x_i, q_i)\ . \tag{11}$$

The numerical fluxes $f_{i\pm1/2}$ are approximations to $h(x_{i\pm1/2}, t)$ of a function $h(x,t)$. This function h is implicitly defined by

$$f(x, q(x,t)) = \frac{1}{\Delta x}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} h(\xi, t)d\xi\ , \tag{12}$$

which leads to

$$\frac{d}{dx}f(x, q(x,t)) = \frac{1}{\Delta x}(h(x + \frac{\Delta x}{2}, t) - h(x - \frac{\Delta x}{2}, t)).$$

The values $f_{i\pm1/2} \approx h(x_{i\pm1/2}, t)$ can be computed via the common recovery method which is used in finite volume methods, see [9, 12]. The cell means $\bar{h}_i$ of function h coincide with the cell center values of f:

$$\bar{h}_i(t) = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} h(\xi, t)d\xi = f(x_i, q(x_i, t)).$$

Therefore the numerical fluxes $f_{i\pm1/2}$ can be recovered from cell means $\bar{h}_j$, $j = i - k_1, \ldots, i + k_2$ by polynomial interpolation. The choice of k_1, k_2 defines the stencil. For hyperbolic problems a stable numerical scheme should propagate its information in direction of the characteristics [12]. This upwinding is discussed later.

Finally, all unknowns are collected in a new vector $Y = (q_{1/2}, \underbrace{q_1, \ldots, q_N}_{Y_1}, q_{N+1/2})^T$ of dimension $2N + 4$. The semi-discretised PDE (11) together with the boundary conditions from (10) yields a semi-explicit DAE system assumed to be of index one:

$$Y_1' = F(t, Y) \,, \tag{13}$$

$$0 = G(t, Y) \,. \tag{14}$$

Usually, the physical BCs must be complemented by some artificial numerical conditions. The most simple idea is a linear extrapolation of the interior values $q_1, \ldots, q_N$ to the boundaries $x_{1/2}$ and $x_{N+1/2}$. Physical BCs and artificial conditions are summarised in function G of dimension 4.

The formulation of the BCs as algebraic equations allows conveniently to implement the coupling of several single pipe or channel sections [15]. As an example, assume the junction of two channels 1 and 2 into channel 3, see Fig. 1. Then, for the boundary variables $q_R^1 = q_{N_1+1/2}^1$, $q_R^2 = q_{N_2+1/2}^2$, $q_L^3 = q_{1/2}^3$ three

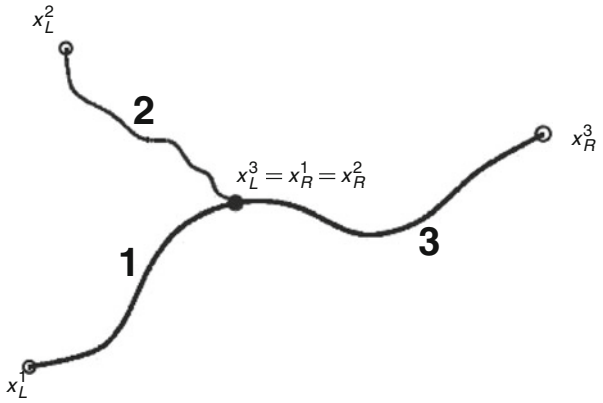


Fig. 1 Junction of two channels

coupling and three numerical boundary conditions are required:

$$\begin{aligned}
(\rho Q)^1_R + (\rho Q)^2_R &= (\rho Q)^3_L \,, \\
h^1_{p,R} &= h^3_{p,L} \,, \\
h^2_{p,R} &= h^3_{p,L} \,, \\
(\rho Q)^3_L &= \frac{3}{2}(\rho Q)^3_1 - \frac{1}{2}(\rho Q)^3_2 \,, \\
h^1_{p,R} &= \frac{3}{2}h^1_{p,N_1} - \frac{1}{2}h^1_{p,N_1-1} \,, \\
h^2_{p,R} &= \frac{3}{2}h^2_{p,N_2} - \frac{1}{2}h^2_{p,N_2-1} \,.
\end{aligned}$$

The physical conditions are mass conservation and identities of piezometric heads $h_p(x,t) = S_0(x) + \dfrac{p(x,t)}{\rho_0 g}$. Possible numerical conditions are obtained by the forementioned extrapolation of these quantities.

The solution of (13), (14) can then be performed by an appropriate DAE solver and is discussed in [8] in more detail.

3.2 *Local Lax-Friedrichs and Well Balanced Upwind Technique*

It remains to compute the inner fluxes $f_{i+1/2}$, $i = 0, \ldots, N$. Considering a conservative system $\partial_t q + \partial_x f(q) = 0$ the flux function is splitted into a positive and a negative part according to

$$f(q) = \underbrace{\frac{1}{2}(f(q) + |\lambda^*|q)}_{f_{up}} + \underbrace{\frac{1}{2}(f(q) - |\lambda^*|q)}_{f_{do}} \,. \tag{15}$$

Choosing constant λ^* as largest absolute eigenvalue of $J = f'(q)$ ensures only positive eigenvalues of the Jacobian of f_{up} and only negative eigenvalues of the Jacobian of f_{do}. This leads to the upwind approximation $f_{i+1/2} = f^-_{up,i+1/2} + f^+_{do,i+1/2}$. In the local Lax-Friedrichs or Rusanov approach, $\lambda^* = \lambda^*_{i+1/2}$ is chosen only locally around $x_{i+1/2}$ [2].

Here, upwind reconstruction of a function h is denoted by $h^-_{i+1/2}$ using a stencil with one more point to the left and downwind reconstruction is $h^+_{i+1/2}$ using a stencil with one more point to the right [12]. A first order approximation $h^-_{i+1/2} = h_i$, $h^+_{i+1/2} = h_{i+1}$ yields $f_{i+1/2} = \frac{1}{2}(f_i + f_{i+1} - |\lambda^*|(q_{i+1} - q_i))$. The term

$-\frac{1}{2}|\lambda^*|(q_{i+1}-q_i)$ can be interpreted as a flux correction to the system proportional to $\partial_{xx}q$.

In order to get space discretisations higher than order one, the values $h^+_{i+1/2}$, $h^-_{i+1/2}$ of function h in (12) must be interpolated from its cell averages $\bar{h}_1,\ldots,\bar{h}_n$. Weighted essentially non-oscillatory (WENO) schemes use a linear combination of interpolation polynomials such that weights are chosen according to its smoothness. A third order approximation according to [10] is applied.

Splitting of flux function (15) can be applied to system (8), too:

$$M(q)\frac{d}{dx}f(x,q) = \frac{1}{2}\left(M(q)\frac{d}{dx}f(x,q) + |\lambda^*|\partial_x q\right) + \frac{1}{2}\left(M(q)\frac{d}{dx}f(x,q) - |\lambda^*|\partial_x q\right).$$

In order to get the desired eigenvalues, λ^* has to be chosen as largest absolute eigenvalue of $M \cdot \partial_q f(x,q)$.

This local Lax-Friedrichs (LLF) approach does not preserve stationary solutions of (5), (6) with $\partial_x(\rho Q) = 0$ in general. The reason is that the flux correction in mass conservation equation $\partial_t(\rho A) + \partial_x(\rho Q) = 0$ corresponds to a term of size $\epsilon \cdot \partial_{xx}(\rho A)$, which will not be zero, especially near a discontinuity due to a change in flow conditions. This non-balanced property can be observed in the numerical tests of Sect. 4.

To avoid this disadvantage, an additional splitting approach (WBLLF) for Eqs. (5), (6) is introduced:

$$M(q)\frac{d}{dx}f(x,q) = \frac{1}{2}M_{up}(q)\frac{d}{dx}\tilde{f}(x,q) + \frac{1}{2}M_{do}(q)\frac{d}{dx}\tilde{f}(x,q)\,, \tag{16}$$

with $\tilde{f}(x,q) = (\rho A, \rho Q, \frac{(\rho Q)^2}{\rho A}, z, \rho)^T$,

$$M_{up} = \begin{pmatrix} 0 & u_p & 0 & 0 & 0 \\ -\frac{u_p-u_m}{2}|\lambda^*|^2 & 2|\lambda^*| & 1 & g\rho A & \frac{A}{\beta\rho_0} \end{pmatrix},$$

$$M_{do} = \begin{pmatrix} 0 & u_m & 0 & 0 & 0 \\ \frac{u_p-u_m}{2}|\lambda^*|^2 & -2|\lambda^*| & 1 & g\rho A & \frac{A}{\beta\rho_0} \end{pmatrix}.$$

and locally chosen constants $u_p = (1 + sign(u))$, $u_m = (1 - sign(u))$. The eigenvalues of $M_{up}(q)\partial_q\tilde{f}_{up}(x,q)$ are non-negative and those of $M_{do}(q)\partial_q\tilde{f}_{do}(x,q)$ are non-positive. Moreover, the splitting (16) preserves stationary solutions with $\partial_x(\rho Q) = 0$, since a term like $\epsilon \cdot \partial_{xx}(\rho A)$ does not appear in the mass conservation equation.

4 Numerical Examples

In order to show the well balanced property of the WBLLF splitting a steady free surface flow problem with constant non zero mass flow is considered. A channel consisting of two flat parts connected by one steep part of total length $L = 2200$ and constant width $B = 2$ is discretised into $N = 220$ cells. Figure 2 shows the initial values and the final solutions computed with LLF and WBLLF splitting. The BCs have been chosen according to $Q(0,t) = 5$, $z(2200,t) = -6.5$ and friction coefficient is $K_{St} = 50$. Obviously, LLF splitting is not able to get the right volume flow $Q \equiv 5$ near the transition points from sub- to supercritical flow. Splitting by WBLLF delivers the exact solution $Q \equiv 5$.

In a second example pressure flow is considered. A pipe of length $L = 1000$ and bottom elevation $S_0(x) = 10 - \frac{x}{100}$ is assumed. The diameter of the pipe is given by $D(x) = 1$ for $x \leq 400$ or $x \geq 600$ and $D(x) = \frac{1}{2}$ for $x \in [410, 590]$. In the range $x \in (400, 410)$ and $x \in (590, 600)$ $D(x)$ is obtained by linear interpolation. A simplified friction formula $S_f = \frac{10^{-3}u}{gA}$ has been used and the number of cells is $N = 200$. At both ends the pipe is connected to a storage basin, each of cross sectional area $A_S = 200$. The coupling of the storage basins to the pipe ends left and

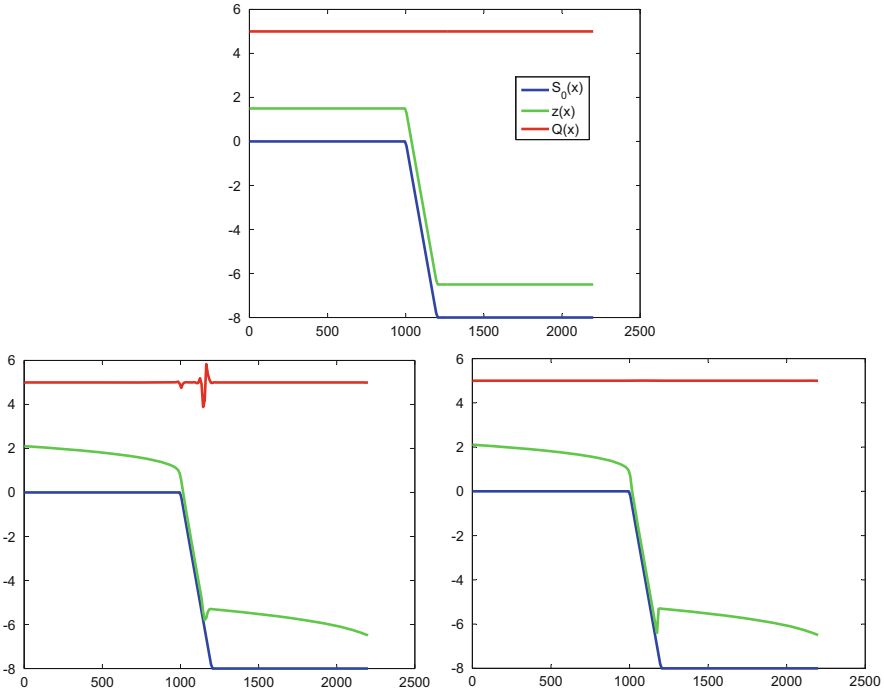


Fig. 2 Initial and final solutions computed with LLF (*left*) and WBLLF splitting (*right*)

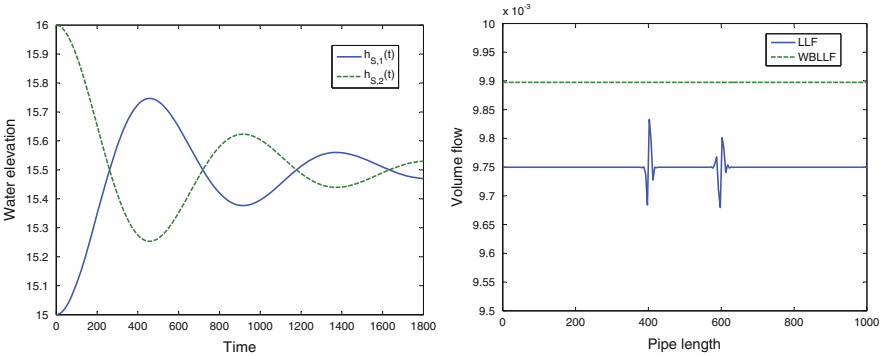


Fig. 3 Water elevation in storage basins and final volume flow in the pipe

Table 1 Number of successful (NSUC) and failed (NFAIL) time steps, function evaluations (NFCN) and CPU-time

		NSUC	NFAIL	NFCN	CPU
Channel flow	LLF	404	56	8764	115
Channel flow	WBLLF	380	50	8230	103
Pipe flow	LLF	45	4	1055	10.3
Pipe flow	WBLLF	48	4	1124	10.5

right are modelled by:

$$\begin{aligned} \frac{d}{dt} h_{S,1}(t) &= -\frac{Q_L(t)}{A_{S,1}} \,, \\ 0 &= h_{S,1}(t) - h_{p,L}(t) \,, \\ \frac{d}{dt} h_{S,2}(t) &= \frac{Q_R(t)}{A_{S,2}} \,, \\ 0 &= h_{S,2}(t) - h_{p,R} \,. \end{aligned}$$

$h_{S,1}$, $h_{S,2}$ denote the water elevation in the storage basins and Q_L, Q_R, $h_{p,L}$, $h_{p,R}$ the left and right boundary values of volume flow and piezometric heads in the pipe. Initial conditions are chosen as $h_{S,1}(0) = 15$, $h_{S,2}(0) = 16$. The initial piezometric heads in the pipe are obtained by linear interpolation between these values and the initial volume flow is zero. During simulation time $t \in [0, 1800]$ the water movement through the pipe oscillates between the basins with a decreasing amplitude caused by friction. Figure 3 shows the water elevation in the storage basins and the volume flows $Q(x, t_{end})$ in the pipe at final time obtained by the LLF and WBLLF approaches. Again it can be seen that the LLF discretisation produces oscillations in the neighbourhood of the contraction of the pipe and WBLLF is preferable.

Finally Table 1 shows that the numerical efforts for solving the final DAE system are comparable for both approaches. All computations have been performed with relative and absolute tolerance of 10^{-4} by the integrator `rodasp`[13].

References

1. Bohl, W.: Technische Strömungslehre. Vogel Fachbuch, Würzburg (1998)
2. Bouchut, F.: Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources. Frontiers in Mathematics. Birkhäuser, Basel (2004)
3. Bourdarias, C., Gerbi, S.: A finite volume scheme for a model coupling free surface and pressurised flows in pipes. J. Comput. Appl. Math. **209**(1), 109–131 (2007)
4. de Saint-Venant, B.: Théorie du Mouvement Non-Permanent des Eaux aves Application aux Crues des Rivières et à l'introduction des Marées dans leur lit. C. R. Acad. Sci. Paris **73**, 148–154, 237–240 (1871)
5. Guinot, V.: Godunov-Type Schemes, an Introduction for Engineers. Elsevier, Amsterdam (2003)
6. Hanisch, H.-H., Eidner, R., Grigo, J., Lippert, D.: Planung und Durchführung des 1. Tracerversuches Elbe. Dtsch. Gewässerkundliche Mitt. **41**(5), S.212–215 (1997)
7. Hilden, M., Steinebach, G.: ENO-discretizations in MOL-applications, some examples in river hydraulics. Appl. Numer. Math. **28**, 293–308 (1998)
8. Jax, T., Steinebach, G.: ROW methods adapted to network simulation for fluid flow. In: Russo, G., Capasso, V., Nicosia, G., Romano, V. (eds.) Progress in Industrial Mathematics at ECMI 2014. Springer (2017)
9. Jiang, G.-S., Shu, C.-W.: Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**, 202–228 (1996)
10. Kurganov, A., Levy, D.: A third-order semidiscrete central scheme for conservation laws and convection-diffusion equations. SIAM J. Sci. Comput. **22**(4), 1461–1488 (2000)
11. Rentrop, P., Steinebach, G.: Model and numerical techniques for the alarm system of river Rhine. Surv. Math. Ind. **6**, 245–265 (1997)
12. Shu, C.-W.: High order weighted essentially nonoscillatory schemes for convection dominated problems. SIAM Rev. **51**(1), 82–126 (2009)
13. Steinebach, G., Rentrop, P.: An adaptive method of lines approach for modelling flow and transport in rivers. In: Vande Wouver, A., Sauces, Ph., Schiesser, W.E. (eds.) Adaptive Method of Lines, pp. 181–205. Chapman & Hall/CRC, Boca Raton (2001)
14. Steinebach, G., Wilke, K.: Flood forecasting and warning on the River Rhine. Water Environ. Manag. J. CIWEM **14**(1), 39–44 (2000)
15. Steinebach, G., Rademacher, S., Rentrop, P., Schulz, M.: Mechanisms of coupling in river flow simulation systems. J. Comput. Appl. Math. **168**(1–2), 459–470 (2004)
16. Steinebach, G., Rosen, R., Sohr, A.: Modeling and numerical simulation of pipe flow problems in water supply systems. In: Klamroth, A., Lang, K., Leugering, J., Morsi, G., Oberlack, A., Ostrowski, M., Rosen, R. (eds.) Mathematical Optimization of Water Networks Martin. International Series of Numerical Mathematics, vol. 162, pp. 3–15. Springer, Basel (2012)
17. von Dalwigk, V., Steinebach, G.: Mathematische Modelle in der Gewässerkunde - Stand und Perspektiven. BfG-Mitteilungen Nr.19, Koblenz (1999)

MOR via Quadratic-Linear Representation of Nonlinear-Parametric PDEs

Yi Lu, Nicole Marheineke, and Jan Mohring

Abstract This work deals with the model order reduction (MOR) of a nonlinear-parametric system of partial differential equations (PDEs). Applying a semidiscretization in space and replacing the nonlinearities by introducing new state variables, we set up quadratic-linear differential algebraic systems (QLDAE) and use a Krylov-subspace MOR. The approach is investigated for gas pipeline modeling.

Keywords Gas networks • Krylov-subspace method • Model order reduction • Quadratic-linear differential algebraic system

1 Quadratic Linearization and Krylov-Subspace MOR for QLDAE

MOR for large-scale systems is a recent topic in research [1, 2]. For nonlinear differential algebraic systems (DAE), quadratic linearization and Krylov subspace MOR for the resulting QLDAE have been investigated, see e.g. [3]. The basic idea to get rid of the nonlinearities is the **polynomialization and quadratic linearization** of the system by taking derivatives or adding polynomial algebraic equations for new state variables. Depending on the choice of the new variables, various quadratic-linear representations can be obtained that are equal in the phase (state) space, but differ in the frequency space. In view of MOR via a Krylov subspace method that approximates the transfer function and the moments, this difference in

Y. Lu (✉) • N. Marheineke
FAU Erlangen-Nürnberg, Lehrstuhl Angewandte Mathematik 1, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: yi.lu@math.fau.de; marheineke@math.fau.de

J. Mohring
Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
e-mail: jan.mohring@itwm.fraunhofer.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_110

the frequency space effects the outcome of the reduced models. Hence, in modeling, the set-up of an appropriate quadratic-linear representation is crucial and problem-dependent. Considering the QLDAE that is quadratic in the states $\mathbf{x}(t) \in \mathbb{R}^n$ and bilinear in the input $u(t) \in \mathbb{R}$ for $t \in \mathbb{R}$ with Kronecker product $\otimes$ and coefficient matrices $\mathbf{E}, \mathbf{G}_1, \mathbf{D}_1 \in \mathbb{R}^{n,n}$, $\mathbf{G}_2, \mathbf{D}_2 \in \mathbb{R}^{n,n^2}$ and $\mathbf{b} \in \mathbb{R}^n$

$$\mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{G}_1\mathbf{x} + \mathbf{G}_2\mathbf{x} \otimes \mathbf{x} + (\mathbf{D}_1\mathbf{x} + \mathbf{D}_2\mathbf{x} \otimes \mathbf{x} + \mathbf{b})\, u,$$

the **Krylov-subspace MOR** provides a reduced model in the sense of moment matching. The quality of the reduced model is thereby determined by two errors coming from (a) the asymptotic expansion of the QLDAE in m linear subsystems and (b) the approximation of the transfer functions and moments up to order ℓ.

(a) Set-up of Linear Subsystems Consider a small input αu, $\alpha < 1$ and assume that the response $\mathbf{x}$ of the QLDAE can be expanded in a regular power series in α, i.e. $\mathbf{x} = \sum_{i=1}^{\infty} \alpha^i \mathbf{x}^{(i)}$. Plugging this expansion into the QLDAE yields homogeneous subsystems with the responses $\mathbf{x}^{(k)}$ in the different orders $\mathscr{O}(\alpha^k)$ (variational approach [6]). The subsystems can be solved iteratively. In every order it is a linear time invariant system with known input (being determined from the lower orders),

$$\mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^{(k)} = \mathbf{G}_1\mathbf{x}^{(k)} + \mathbf{R}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k-1)}, u).$$

In an expansion up to $\mathscr{O}(\alpha^{m+1})$, we deal consequently with m subsystems. For example, in case $m = 2$ they are

$$\mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^{(1)} = \mathbf{G}_1\mathbf{x}^{(1)} + \mathbf{b}u, \qquad \mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^{(2)} = \mathbf{G}_1\mathbf{x}^{(2)} + \mathbf{G}_2\mathbf{x}^{(1)} \otimes \mathbf{x}^{(1)} + \mathbf{D}_1\mathbf{x}^{(1)}u.$$

(b) Approximation of Transfer Functions and Moments To the m subsystems an associated Krylov space $\mathscr{K}^m_{\ell+1}$, $\ell \in \mathbb{N}$ can be set up, if $\mathbf{G}_1$ is regular. Projecting the QLDAE in the Krylov space yields the reduced model that matches the moments of the m subsystems up to order ℓ, e.g. $\mathscr{K}^2_{\ell+1}(\mathbf{G}^{-1}\mathbf{E}, \mathbf{G}^{-1}([\mathbf{b}, \mathbf{G}_2, \mathbf{D}_1]))$. Since the dimension of the Krylov space exceeds the dimension of the underlying QLDAE, a minimization is aimed for. Therefore, a multi-variable Laplace transformation is applied, i.e. for any $\mathbf{f} : \mathbb{R}^k_{\geq 0} \to \mathbb{R}^n$

$$\tilde{\mathbf{f}}(s_1, \ldots, s_k) = \int_{\mathbb{R}^k_{\geq 0}} \mathbf{f}(t_1, \ldots, t_k) \prod_{i=1}^{k} \mathrm{e}^{-s_i t_i}\, \mathrm{d}t_1 \cdots \mathrm{d}t_k.$$

Defining $\tilde{\mathbf{x}}^{(k)}(t_1,\ldots,t_k) = \mathbf{x}^{(k)}(t)|_{t_1=\cdots=t_k=t}$ for the subsystem responses, the Laplace transformation gives for the derivative

$$\int_{\mathbb{R}^n_{\geq 0}} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^{(k)}(t) \prod_{i=1}^{k} \mathrm{e}^{-s_i t_i}\, \mathrm{d}t_1 \cdots \mathrm{d}t_k = \sum_{i=1}^{k} s_i \tilde{\mathbf{x}}^{(k)}(s_1,\ldots,s_k).$$

Furthermore, the Kronecker products transform to arithmetic averages for the terms of all possible permutations over the frequency variables $s = (s_1,\ldots,s_k)$ in the Kronecker product [5]. So, the subsystems $k = 1,\ldots,m$ become

$$\left(\sum_{i=1}^{k} s_i \mathbf{E} - \mathbf{G}_1\right)\tilde{\mathbf{x}}^{(k)} = \tilde{\mathbf{R}}(\tilde{\mathbf{x}}^{(1)},\ldots,\tilde{\mathbf{x}}^{(k-1)},\tilde{u}), \qquad \tilde{\mathbf{x}}^{(k)} = \mathbf{H}^{(k)} \prod_{i=1}^{k} \tilde{u}(s_i)$$

if $(\sum_{i=1}^{k} s_i\mathbf{E} - \mathbf{G}_1)$ is regular. The transfer functions $\mathbf{H}^{(k)} : \mathbb{R}^k \to \mathbb{R}^n$ for $k = 1, 2$ are

$$\begin{aligned}\mathbf{H}^{(1)}(s) &= (s\mathbf{E} - \mathbf{G}_1)^{-1}\mathbf{b}\\ \mathbf{H}^{(2)}(s_1,s_2) &= ((s_1+s_2)\mathbf{E} - \mathbf{G}_1)^{-1}\frac{1}{2}\Big[\mathbf{D}_1\left(\mathbf{H}^{(1)}(s_1) + \mathbf{H}^{(1)}(s_2)\right)\\ &\qquad + \mathbf{G}_2\left(\mathbf{H}^{(1)}(s_1)\otimes\mathbf{H}^{(1)}(s_2) + \mathbf{H}^{(1)}(s_2)\otimes\mathbf{H}^{(1)}(s_1)\right)\Big]\end{aligned}$$

The ℓ-th moments associated to the transfer functions of the m subsystems are the coefficients of the term $\prod_{i=1}^{m}(s_i + s_0)^{j_i}$, $\sum j_i = \ell$ for $\mathbf{H}^{(1)},\ldots,\mathbf{H}^{(m)}$ that can be computed by performing a Taylor expansion around the frequency s_0. They yield the desired subspace $\mathscr{S}^m_{\ell+1}$ for the reduced model. Using the Neumann series $(\sum_{i=1}^{k} s_i\mathbf{E} - \mathbf{G}_1)^{-1} = \sum_{j=0}^{\infty}\mathbf{A}_k^j(-\mathbf{G}_1^{(k)})^{-1}(\prod_{i=1}^{k}(s_i - s_0))^j$ with $\mathbf{A}_k = (\mathbf{G}_1^{(k)})^{-1}\mathbf{E}$, $\mathbf{G}_1^{(k)} = \mathbf{G}_1 - ks_0\mathbf{E}$ and $\mathbf{r} = (-\mathbf{G}_1^{(1)})^{-1}\mathbf{b}$ we particularly obtain for the subspace $\mathscr{S}^2_{\ell+1}$ to $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$

$$\begin{aligned}\mathscr{S}^2_{\ell+1} = &\left\{\mathbf{A}_1^i\mathbf{r},\ i \leq \ell\right\} \cup \left\{\mathbf{A}_2^i(-\mathbf{G}_1^{(2)})^{-1}\mathbf{D}_1\mathbf{A}_1^j\mathbf{r},\ i+j \leq \ell\right\}\\ &\cup \left\{\mathbf{A}_2^i(-\mathbf{G}_1^{(2)})^{-1}\mathbf{G}_2(\mathbf{A}_1^j\mathbf{r})\otimes(\mathbf{A}_1^h\mathbf{r}),\ i+j+h \leq \ell,\ h \leq j\right\}.\end{aligned}$$

The dimension of $\mathscr{S}^m_{\ell+1}$ is smaller than the dimension of the Krylov space $\mathscr{K}^m_{\ell+1}$ if $\ell < n$ (n dimension of the QLDAE). Let $\mathbf{V}$ be an orthonormal basis of $\mathscr{S}^m_{\ell+1}$. Then the projected QLDAE (reduced model)

$$\hat{\mathbf{E}}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{z} = \hat{\mathbf{G}}_1\mathbf{z} + \hat{\mathbf{G}}_2\mathbf{z}\otimes\mathbf{z} + \left(\hat{\mathbf{D}}_1\mathbf{z} + \hat{\mathbf{D}}_2\mathbf{z}\otimes\mathbf{z} + \hat{\mathbf{b}}\right)u, \qquad \mathbf{z} = \mathbf{V}^{\mathrm{T}}\mathbf{x} \tag{1}$$

with $\hat{\mathbf{b}} = \mathbf{V}^{\mathrm{T}}\mathbf{b}$, $\hat{\mathbf{E}} = \mathbf{V}^{\mathrm{T}}\mathbf{E}\mathbf{V}$, $\hat{\mathbf{G}}_1 = \mathbf{V}^{\mathrm{T}}\mathbf{G}_1\mathbf{V}$, $\hat{\mathbf{G}}_2 = \mathbf{V}^{\mathrm{T}}\mathbf{G}_2\mathbf{V} \otimes \mathbf{V}$, $\hat{\mathbf{D}}_1 = \mathbf{V}^{\mathrm{T}}\mathbf{D}_1\mathbf{V}$ and $\hat{\mathbf{D}}_2 = \mathbf{V}^{\mathrm{T}}\mathbf{D}_2\mathbf{V} \otimes \mathbf{V}$ matches the moments of the m subsystems up to the order ℓ [3].

2 Application to Nonlinear PDE for Gas Pipeline Modeling

Gas flowing in a horizontal pipeline of length L and diameter D can be described in terms of the pressure p and the flow rate q by the one-dimensional conservation laws (instationary Euler equations) [4], i.e. for $x \in [0, L]$, $t \in [0, \tau]$

$$\frac{1}{RT}\frac{\partial}{\partial t}\frac{p}{z} + \frac{\partial}{\partial x}q = 0, \qquad \frac{\partial}{\partial t}q + RT\frac{\partial}{\partial x}\frac{zq^2}{p} + \frac{\partial}{\partial x}p + \frac{RT}{2D}\frac{\lambda z q|q|}{p} = 0$$

with temperature T, gas constant R, Darcy friction λ and the gas compressibility given by $z(p) = 1 + c_z p$, $p = 0.257/p_c - 0.533T_c/(p_c T)$ with the pseudo-critical pressure p_c and temperature T_c. To study the MOR approach for the hyperbolic system and avoid shocks (or flow inversions), we assume a strictly monotonic pressure along the pipe. We prescribe the outflow pressure at $x = L$ as $P = const$ and treat the higher inflow pressure as time-dependent input u. The system is initialized with the stationary solution p_s and $Q = const$.

Using the reference quantities L, τ, P and Q, the dimensionless system is hence given by—note that we keep the original terminology—

$$\frac{\partial}{\partial t}p + \delta z^2\frac{\partial}{\partial x}q = 0, \qquad \beta\frac{\partial}{\partial t}q + \alpha\frac{qz}{p}\frac{\partial}{\partial x}q + \left(1 - \alpha\frac{q^2}{p^2}\right)\frac{\partial}{\partial x}p + \gamma\frac{zq^2}{p} = 0$$

$$p(0,t) = u(t), \quad p(1,t) = 1, \quad u(t) > 1, \qquad p(x,0) = p_s(x), \quad q(x,0) = 1$$

with parameters $\alpha = RTQ^2/P^2$, $\beta = QL/(P\tau)$, $\gamma = \alpha\lambda L/(2D)$ and $\delta = \alpha/\beta$.

We transfer the PDE system into a nonlinear DAE by applying a spatial semi-discretization with central differences on a staggered grid: midpoints $x_i = i\Delta x$, $i = 0, \ldots, N$ and edges $x_{j+1/2} = x_j + \Delta x/2$, $j = 0, \ldots, N-1$ with $\Delta x = L/N$, $N \in \mathbb{N}$ number of cells. We consider the pressure p and the first PDE at the midpoints x_i and the gas flow rate q and the second PDE at the edges $x_{i+1/2}$. We use a second order interpolation to evaluate p and q at the respective other positions, yielding a second order approximation for the differential operators $\frac{\partial}{\partial x}[\cdot]$. Consequently, we obtain the parametric nonlinear input system

$$-\frac{\mathrm{d}}{\mathrm{d}t}p_i = \frac{\delta}{\Delta x}z_i^2(q_{i+1/2} - q_{i-1/2}),$$

$$-\beta\frac{\mathrm{d}}{\mathrm{d}t}q_{i-1/2} = \frac{1}{\Delta x}\left[\frac{\alpha q_{i-1/2}\hat{z}_i}{\hat{p}_i}\hat{\delta}_{q_{i-1/2}} + \left(1 - \frac{\alpha q_{i-1/2}^2}{\hat{p}_i^2}\right)(p_i - p_{i-1})\right] + \frac{\gamma\hat{z}_i q_{i-1/2}^2}{\hat{p}_i}$$

$$\hat{\delta}_{q_{i-1/2}} = \begin{cases} -3q_{i-1/2} + 4q_{i+1/2} - q_{i+3/2} & i = 1 \\ q_{i+1/2} - q_{i-3/2}, & i = 2, \dots, N-1 \\ q_{i-5/2} - 4q_{i-3/2} + 3q_{i-1/2}, & i = N \end{cases}$$

$$\hat{p}_i = \begin{cases} \frac{1}{8}(3p_{i-1} + 6p_i - p_{i+1}), & i = 1, \dots, N-1 \\ \frac{1}{8}(6p_{N-1} + 3p_N - p_{N-2}), & i = N \end{cases} \qquad \hat{z}_i = z(\hat{p}_i)$$

with the states $\mathbf{x} = (p_1, \dots, p_{N-1}, q_{1/2}, \dots q_{N-1/2}) \in \mathbb{R}^{2N-1}$ and the input $u = p_0$.

We perform the quadratic linearization by applying the following substitutions $v_{1,i} = p_i^2$, $v_{2,i} = q_{i-1/2}/\hat{p}_i$ and $v_{3,i} = v_{2,i}^2$ in two different ways. The resulting two quadratic-linear representations (QLDAE1) and (QLDAE2) have the same size and just differ in the sign of one of the additional equations. However, this difference effects the reduced models as we will see.

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t} p_i &= \frac{\delta}{\Delta x} \left(1 + 2c_z p_i + c_z^2 v_{1,i}\right) (q_{i+1/2} - q_{i-1/2}) \\ -\beta \frac{\mathrm{d}}{\mathrm{d}t} q_{i-1/2} &= \frac{1}{\Delta x} \left[\alpha(v_{2,i} + c_z q_{i-1/2})\delta_{q_{i-1/2}} + (1 - \alpha v_{3,i})\,(p_i - p_{i-1})\right] \\ &\quad + \gamma(v_{2,i} + c_z q_{i-1/2}) q_{i-1/2} \\ v_{1,i} - p_i &= 0, \qquad v_{3,i} - v_{2,i}^2 = 0 \\ q_i - v_{2,i}\hat{p}_i &= 0 \quad \text{(QLDAE1)} \qquad\qquad v_{2,i}\hat{p}_i - q_i = 0 \quad \text{(QLDAE2)} \end{aligned}$$

In the gas QLDAEs we have $\mathbf{D}_2 = \mathbf{0}$. Moreover, the coefficient matrix $\mathbf{G}_1$ is singular, therefore we use one of the 10 % smallest modes of the stationary solution as expansion point s_0 for the transfer functions in MOR.

Figure 1 shows the results for a Krylov-space MOR with $m = 2$ subsystems and moment order $\ell = 1, 2, 9$. The relative error of the gas pressure and the flow rate in comparison to the original nonlinear DAE (after semi-discretization) is visualized. As expected, QLDAE1 and QLDAE2 yield equivalent results that coincide with the reference solution. The approximations of the respective reduced models of order $\ell = 9$ are very well. In lower order the reduced models of the two quadratic-linear representations behave differently. Mostly, the reduced models to QLDAE1 show better results. However, for $q = 1, 2$ the errors are not acceptable in both cases. This emphasizes the difficulty in substituting the nonlinearities and setting up an appropriate quadratic-linear representation, already a sign can change the outcome. Considering the computational effort, the reduced models $q = 9$ require one-third of the time of the nonlinear DAE system in the online modus. The respective projection matrix $\mathbf{V}$ is computed offline. The results are promising, in particular in view of larger systems.

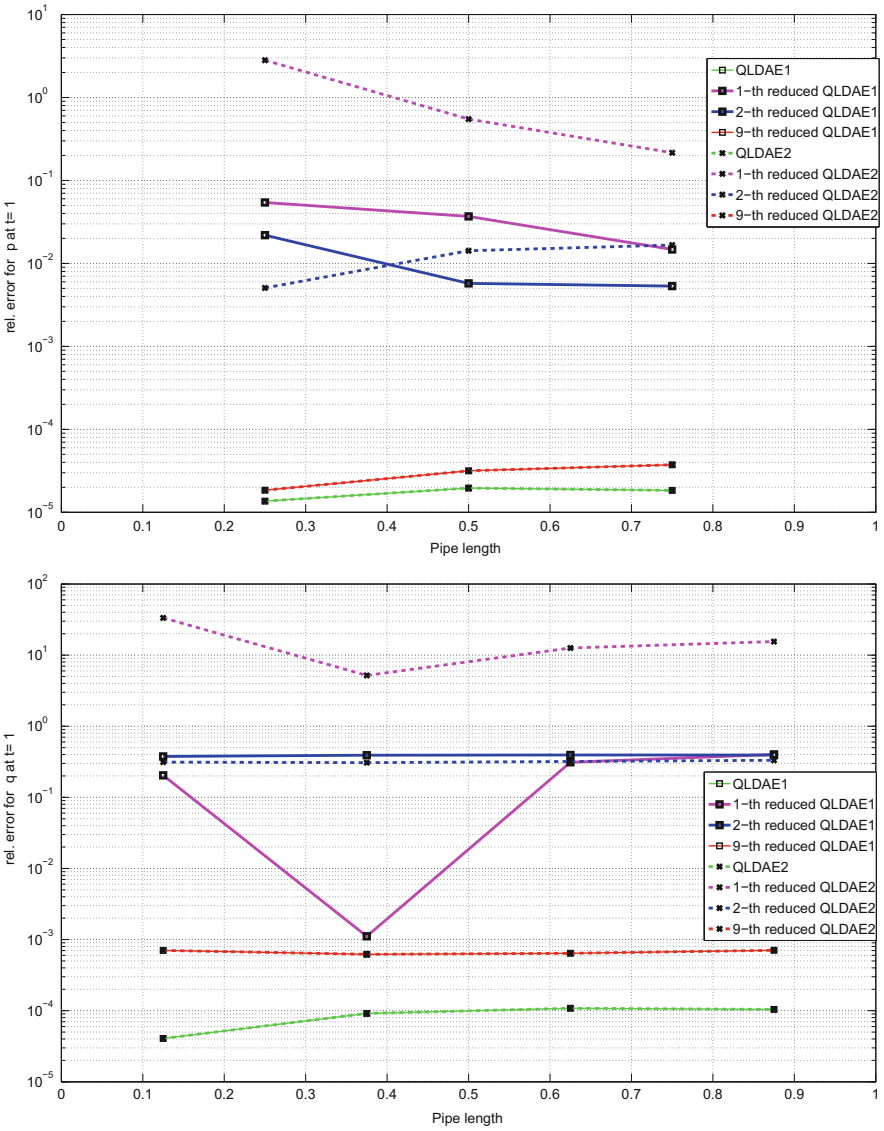


Fig. 1 Relative error for gas pressure p (*top*) and flow rate q (*bottom*) at $t = 1$ – comparison of QLDAE1, QLDAE2, associated reduced models of order $\ell = 1, 2, 9$ with the original nonlinear DAE. Simulation set-up: $\alpha = 6.8{\cdot}10^{-4}$, $\beta = 2.0{\cdot}10^{-3}$, $\gamma = 24$, $u(t) = 1.5{-}0.0625(1{-}\cos(\pi t))$, $N = 4$, size of the reduced models d_ℓ and QLDAE $d_\star$ $(d_1, d_2, d_9, d_\star) = (3, 4, 11, 18)$. Simulations are performed with MATLAB using the DAE-solver ode15s.m

Acknowledgement The support of the DFG, CRC TRR 154, Project C02 is acknowledged.

References

1. Antoulas, A.: Approximation of Large-Scale Dynamical Systems. Advances in Design and Control. SIAM, Philadelphia (2005)
2. Benner, P., Mehrmann, V., Sorensen, D.C. (eds.): Dimension Reduction of Large-Scale Systems. Springer, Berlin, Heidelberg (2005)
3. Gu, C.: MOR: a projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. IEEE Trans. Comput-Aided Des. Integr. Circ. Syst. **30**, 1307–1320 (2011)
4. Herty, M., Mohring, J., Sachers, V.: A new model for gas flow in pipe networks. Math. Methods Appl. Sci. **33**(7), 845–855 (2009)
5. Peng, L., Pileggi, L.T.: Compact reduced-order modeling of weakly nonlinear analog and RF circuits. IEEE Trans. Comput.-Aided Des. Integr. Circ. Syst. **24**(2), 184–203 (2005)
6. Rugh, W.: Nonlinear System Theory: The Volterra-Wiener Approach. Johns Hopkins University Press, Baltimore, MD (1981)

ROW Methods Adapted to Network Simulation for Fluid Flow

Tim Jax and Gerd Steinebach

Abstract Simulating free-surface and pressurised flow is important to many fields of application, especially in network approaches. Modelling equations to describe flow behaviour arising in these problems are often expressed by one-dimensional formulations of the hyperbolic shallow water equations. One established approach to realise their numerical computation is the method of lines based on semi-discretisation in space (Steinebach and Rentrop, An adaptive method of lines approach for modeling flow and transport in rivers. In: Vande Wouwer, Saucez, Schiesser (eds) Adaptive method of lines, pp 181–205. Chapman & Hall/CRC, Boca Raton, London, New York, Washington, DC, 2001; Steinebach and Weiner, Appl Numer Math 62:1567–1578, 2012; Steinebach et al., Modeling and numerical simulation of pipe flow problems in water supply systems. In: Martin, Klamroth, et al. (eds) Mathematical optimization of water networks. International series of numerical mathematics, vol 162, pp 3–15. Springer, Basel, 2012). It leads to index-one DAE systems as algebraic constraints are required to realise coupling and boundary conditions of single reaches.
Linearly implicit ROW schemes proved to be effective to solve these DAE systems (Steinebach and Rentrop, An adaptive method of lines approach for modeling flow and transport in rivers. In: Vande Wouwer, Saucez, Schiesser (eds) Adaptive method of lines, pp 181–205. Chapman & Hall/CRC, Boca Raton, London, New York, Washington, DC, 2001). However, under certain conditions an extended partial explicit time-integration of the shallow water equations could be worthwhile to save computational effort. To restrict implicit solution by ROW schemes to stiff components while using explicit solution by RK methods for remaining terms, we adapt ROW method ROS34PRW (Rang, J Comput Appl Math 262:105–114, 2014) to an AMF and IMEX combining approach (Hundsdorfer and Verwer, Numerical solution of time-dependent advection-diffusion-reaction equations. Springer, Berlin, Heidelberg, New York, 2003). Applied to first test problems regarding open channel flow, efficiency is analysed with respect to flow behaviour. Results prove to be advantageous especially concerning dynamical flow.

T. Jax (✉) • G. Steinebach
Department of EMT, Hochschule Bonn-Rhein-Sieg, Grantham-Allee 20, 53757 Sankt Augustin, Germany
e-mail: tim.jax@h-brs.de; gerd.steinebach@h-brs.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_111

Keywords Network simulation • ROW methods • Water supply systems

1 Introduction

Simulating fluid flow in large networks is relevant to numerous fields in industry and environment. Applications cover free-surface flow in open channels as well as pressurised flow in closed pipes that arise for instance when planning water supply and sewerage in urban infrastructure [10]. Significant parameters in these applications often reduce to space and time dependent water surface elevation, volume flow and pressure distribution in one single direction. Hence, modelling equations are generally expressed by the one-dimensional shallow water equations (SWEs).

Considering network simulation, one established approach to solve the system of hyperbolic partial differential equations (PDEs) given by the SWEs is the method of lines (MOL). Based on semi-discretisation in space, it approximates space derivatives by appropriate discretisation schemes while time derivatives remain. That way the SWEs are converted into a system of differential-algebraic equations (DAEs), as in networks regarding fluid flow boundary and coupling conditions of single reaches are generally expressed by additional algebraic constraints [8, 10].

DAEs can be interpreted as infinitely stiff ordinary differential equations (ODEs). Moreover, friction is assumed to introduce some stiffness to the SWEs [9]. So the resulting DAE system is finally solved by implicit schemes for time-integration. These are more efficient regarding stiff problems than explicit schemes due to stability reasons. In this context, linearly implicit Rosenbrock-Wanner (ROW) methods are among the schemes that prove to be effective solving the SWEs after space-discretisation [8, 10].

However, in contrast to explicit methods implicit schemes show an increased effort when applied to non-stiff problems. Regarding the SWE system, components assumed to be stiff reduce to algebraic constraints and friction terms. Even more, frictional effects might be too small to affect stiffness properties significantly. Therefore, restricting implicit schemes to stiff parts while applying explicit methods to remaining non-stiff terms might improve computational efficiency.

Methods combining implicit and explicit time-integration are known as implicit-explicit (IMEX) schemes. In this paper, IMEX schemes are realised by adapting the ROW method ROS34PRW [6]. Their implementation is related to approximate matrix factorisation (AMF) that enables to solve non-stiff parts by explicit Runge-Kutta (RK) schemes using approximations to the given Jacobian [2, 3]. Although applying additive splitting of the DAE system's right hand side, the resulting schemes consider additional separation of unknowns into stiff and non-stiff components. Particularly as given equations of mass conservation are regarded as non-stiff.

In order to analyse efficiency of this AMF-IMEX approach in terms of flow behaviour, we simulate test problems of open channel flow. To investigate frictional effects on stiffness and thus time-integration in more detail, explicit and implicit solution of different friction parameters is tested.

In the following, Sect. 2 defines SWE modelling equations, Sect. 3 describes the AMF-IMEX approach, Sect. 4 shows numerical results and Sect. 5 gives a conclusion.

2 SWE Modelling Equations

We consider the SWEs for simulating free-surface flow, given by an adapted formulation of the Saint-Venant equations:

$$z_t + \frac{1}{B} q_x = 0 \tag{1}$$

$$q_t + \left(\frac{q^2}{A} \right)_x + gAz_x = -gAS_f \, . \tag{2}$$

Equations (1) and (2) describe a nonlinear, hyperbolic PDE system based on conservations of mass and momentum. Here, z denotes water surface elevation and q equals volume flow. Further parameters are width of water surface B, flooded cross-sectional area A as well as gravitational constant g.

Parameter S_f denotes friction slope that describes effects due to turbulence, viscosity and friction. According to empirical formula by Manning-Strickler it is expressed by

$$S_f = u|u| k_{St}^{-2} r_{hydr}^{-4/3} \tag{3}$$

with mean velocity u, hydraulic radius r_{hydr} and Strickler coefficient k_{St}. Hydraulic radius r_{hydr} defines ratios of flooded cross-sectional area to wetted perimeter. The Strickler coefficient k_{St} describes a friction parameter with values ranging from $20\,\mathrm{m}^{1/3}/\mathrm{s}$ for rough surfaces to $80\,\mathrm{m}^{1/3}/\mathrm{s}$ for smooth surfaces [1].

Due to their nonlinear hyperbolic character, the SWEs tend to generate discontinuous solutions in the form of shock waves. Therefore, solution by MOL requires specific schemes for space discretisation that ensure accuracy even at discontinuities. One efficient approach are finite volumes with numerical fluxes according to local Lax-Friedrichs or Harten, Lax and van Leer, supplemented by a central third-order WENO interpolation as introduced by Kurganov and Levy [4, 5, 10].

In order to realise coupling and boundary conditions without perturbing conservative properties of the SWE system, linear and nonlinear algebraic constraints are applied. Semi-discretisation in space combined with these algebraic constraints yields a DAE system assumed to be of index one. Restricting to autonomous problems, it can be expressed by

$$My'(t) = f(y(t)) \tag{4}$$

where M denotes a singular mass matrix.

3 AMF-IMEX Approach

DAE systems are generally solved by implicit time-integration, due to their stiff properties. Regarding the SWEs after space-discretisation, linearly implicit ROW schemes proved to be effective [8, 9]. In autonomous formulation they read [8]:

$$y_1 = y_0 + \sum_{i=1}^{s} b_i k_i \quad , \quad \left(M - h\gamma f_y\right) k_i = hf\left(y_0 + \sum_{j=1}^{i-1} a_{ij} k_j\right) + hf_y \sum_{j=1}^{i-1} \gamma_{ij} k_j \; . \tag{5}$$

Here, parameters k_i denote stage values. Weights b_i and coefficients a_{ij}, γ_{ij} (with $\gamma_{ii} = \gamma$) for $i = 1, \ldots, s$, $j = 1, \ldots, i-1$ are defined in corresponding sets of coefficients. Expression $f_y = f_y(y_0)$ denotes the Jacobian. It mainly determines computational effort of ROW schemes that can be reduced using sparse structures.

Every linearly implicit ROW scheme includes an underlying explicit RK method. Related to IMEX approach, this enables to combine both methods in order to restrict implicit solution to stiff components while applying explicit solution to remaining non-stiff terms. For this purpose, function f of DAE system (4) is assumed to be split into a stiff part f_S and a non-stiff part f_N, such that $f = f_N + f_S$ [3].

This splitting of f yields a corresponding splitting of the Jacobian given by $f_y = (f_N + f_S)_y$. By neglecting the Jacobian's non-stiff parts as related to AMF approach [2, 3], standard ROW method (5) leads to an adapted scheme given by:

$$y_1 = y_0 + \sum_{i=1}^{s} b_i k_i \quad , \quad \left(M - h\gamma f_{S_y}\right) k_i = hf\left(y_0 + \sum_{j=1}^{i-1} a_{ij} k_j\right) + hf_{S_y} \sum_{j=1}^{i-1} \gamma_{ij} k_j \; . \tag{6}$$

Regarding computation of stage values k_i, this formulation applies a ROW method to stiff terms but reduces to a RK method for non-stiff terms. Hence, stiff components are evaluated implicitly while non-stiff components are calculated explicitly. As there is just one stiff and one non-stiff part of the Jacobian considered, AMF and IMEX can be assumed to be equivalent [3]. Hence, we refer to the adapted ROW method as AMF-IMEX approach.

This approach corresponds to ROW methods with approximated Jacobians. As standard ROW methods require exact Jacobians, realising the described scheme demands coefficients of W methods that fulfil special order conditions enabling to use non-exact matrices [7]. In addition, supplemental constraints must be considered to evaluate DAE problems. For this reason, we applied coefficients of third-order ROW method ROS34PRW given in [6] that satisfies these conditions (see Table 1).

As implicitly solved stiff components are defined by entries of the Jacobian f_{S_y}, the AMF-IMEX approach enables to adapt time-integration to stiffness properties and special requirements of a problem. In this context, we generally considered algebraic constraints of the resulting DAE system to be stiff. However, in order to analyse frictional effects on stiffness, two different schemes were applied

Table 1 Coefficient set for ROS34PRW [6]

γ	=	$4.3586652150845900E-1$			
α_{21}	=	$8.7173304301691801E-1$	γ_{21}	=	$-8.7173304301691801E-1$
α_{31}	=	$1.4722022879435914E+0$	γ_{31}	=	$-1.2855347382089872E+0$
α_{32}	=	$-3.1840250568090289E-1$	γ_{32}	=	$5.0507005541550687E-1$
α_{41}	=	$8.1505192016694938E-1$	γ_{41}	=	$-4.8201449182864348E-1$
α_{42}	=	$5.0000000000000000E-1$	γ_{42}	=	$2.1793326075422950E-1$
α_{43}	=	$-3.1505192016694938E-1$	γ_{43}	=	$-1.7178529043404503E-1$
b_1	=	$3.3303742833830591E-1$	$\hat{b}_1$	=	$2.5000000000000000E-1$
b_2	=	$7.1793326075422947E-1$	$\hat{b}_2$	=	$7.4276119608319180E-1$
b_3	=	$-4.8683721060099439E-1$	$\hat{b}_3$	=	$-3.1472922970066219E-1$
b_4	=	$4.3586652150845900E-1$	$\hat{b}_4$	=	$3.2196803361747034E-1$

concerning the given friction term by Manning-Strickler: One solves friction implicitly as additional stiff component, referred to as AMIEs. Another one solves friction explicitly as non-stiff component, referred to as AMIEn. In both schemes, all remaining terms of (1) and (2) are evaluated explicitly by the underlying RK method.

4 Numerical Results and Discussion

In this section, the adapted ROW methods were applied to test problems dam break and shallow water flow as introduced in [9]. Both describe open channel flows characterised by different flow behaviour (see Fig. 1): Dam break considers dynamical discontinuous solutions with propagating shock waves. Shallow water flow considers stationary smooth solutions, given after a dynamical transition phase.

Our aim is to analyse efficiency of the presented AMF-IMEX approach in terms of this different flow behaviour and to investigate frictional effects on stiffness and thus time-integration. For this purpose, both test problems were simulated using friction slope according to Manning-Strickler, considering Strickler coefficients $k_{St} = 80\,\mathrm{m}^{1/3}/\mathrm{s}$ (low friction) and $k_{St} = 20\,\mathrm{m}^{1/3}/\mathrm{s}$ (high friction). We expect rising friction to increase stiffness within the system. Thus, solving friction terms implicitly by AMIEs and explicitly by AMIEn should be superior regarding higher and lower friction, respectively.

To compare efficiency of AMF-IMEX schemes AMIEn and AMIEs based on coefficient set ROS34PRW to standard ROW schemes, the same set of coefficients was also applied as ROW method with exact Jacobian. In addition, fourth-order ROW method RODASP [8] was taken into account that is proven to be quite efficient and robust solving the SWEs [9, 10]. Efficiencies were analysed plotting error against CPU-time on logarithmic scales. Errors were evaluated as defined in [9], comparing numerical solutions that consider absolute and relative tolerances

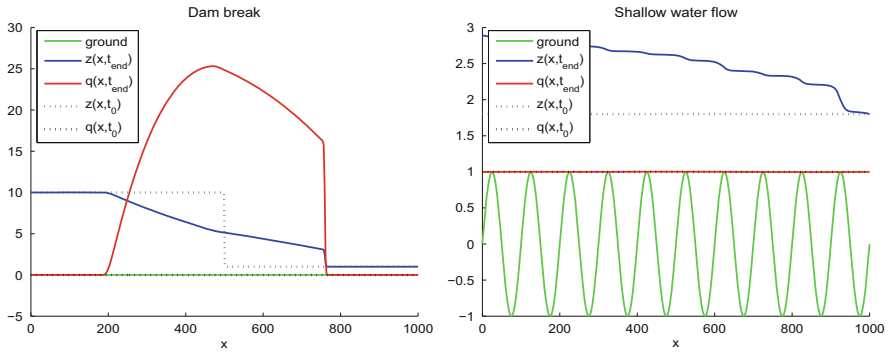


Fig. 1 Initial and final profile of dam break (*left*) and shallow water flow (*right*)

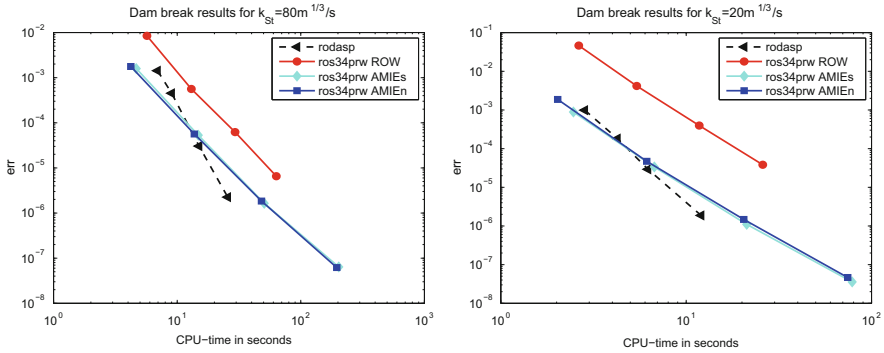


Fig. 2 Efficiencies of dam break

$\texttt{tol} = 10^{-3}, \ldots, 10^{-6}$ for step-size control with more accurate numerical reference solutions determined by MATLAB integrator `ode15s`and $\texttt{tol} = 10^{-12}$.

We regard dam break problem first. Efficiency plots for friction parameters $k_{St} = 80\,\mathrm{m}^{1/3}/\mathrm{s}$ and $k_{St} = 20\,\mathrm{m}^{1/3}/\mathrm{s}$ are given in Fig. 2. They show that both AMF-IMEX schemes based on ROS34PRW are more effective than standard ROW formulation realised by the same set of coefficients for low as well as high friction. The applied third-order schemes based on AMF-IMEX approach even catch up with efficiency of fourth-order ROW method RODASP, performing better results for low accuracy. However, contrary to our expectations, solving friction implicitly by AMIEs and explicitly by AMIEn shows no significant differences even for high friction.

Advantages of the AMF-IMEX approach are due to exploiting properties of explicit schemes and sparser Jacobians, saving effort computing small-step sizes required for dynamical problems. Missing differences solving friction implicitly and explicitly even for high friction might indicate that the considered friction slope has no significant effect on stiffness of the SWEs within the applied range.

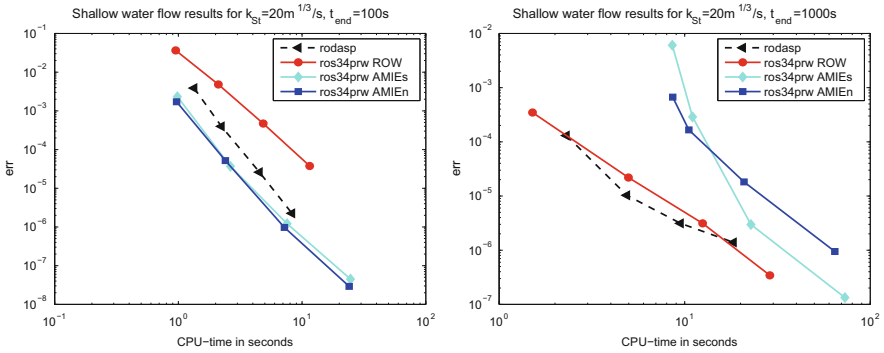


Fig. 3 Efficiencies of shallow water flow

Simulating shallow water flow, different time-intervals were considered to analyse efficiency and frictional effects with respect to dynamical transition phase as well as subsequent stationary smooth solution. In Fig. 3 results for $k_{St} = 20\,\mathrm{m}^{1/3}/\mathrm{s}$ with respect to time-intervals $t = [0, 100]$ and $t = [0, 1000]$ are compared. For $t = [0, 100]$, results prove to be similar to dam break problem: Again, AMF-IMEX approaches are superior to ROW schemes based on ROS34PRW and RODASP. However, for $t = [0, 1000]$ the AMF-IMEX schemes offer a significant loss in efficiency, consequently being inferior to standard ROW formulation.

A possible explanation for this behaviour is given by the dynamical transition phase before finding the stationary solution. For $k_{St} = 20\,\mathrm{m}^{1/3}/\mathrm{s}$ stationary solution seems to be found within $t = [0, 1000]$ but not within $t = [0, 100]$. As dynamical problems require small step-sizes that are performed advantageously by explicit integration, applying the AMF-IMEX approach seems to be superior in transition phase. However, step-sizes can become nearly arbitrarily large simulating stationary solutions such that linearly implicit ROW schemes become advantageous. The fact that standard ROW methods are more efficient solving stationary cases indicates that the AMF-IMEX schemes considered have step-size restrictions similar to explicit RK methods.

Further experiments with $k_{St} = 80\,\mathrm{m}^{1/3}/\mathrm{s}$ showed that loss of efficiency for AMF-IMEX approaches occurs significantly when simulating time-intervals beyond $t = [0, 1000]$. Therefore, duration of transition phase depends on friction parameters, being more abbreviated the higher frictional effects become.

Regarding implicit and explicit solution of friction, the AMF-IMEX schemes show almost no differences within transition phase. But for stationary case, solving friction implicitly by AMIEs is superior to AMIEn for small tolerances. However, these differences were less distinct for $k_{St} = 10\,\mathrm{m}^{1/3}/\mathrm{s}$. So there seem to be no general frictional effects on stiffness within the considered range of parameters.

5 Conclusion

We adapted ROW method ROS34PRW to stiffness properties of DAEs describing the one-dimensional hyperbolic SWEs. For this purpose an AMF and IMEX based approach was used to combine ROW and RK methods in order to implement the partial implicit and explicit integration required. The resulting AMF-IMEX schemes led to increased efficiencies simulating dynamical problems. However, for stationary solutions standard ROW schemes proved to be more effective. The findings indicate that adjusting time-integration to flow behaviour within single reaches of network approaches by adaptive methods can be worthwhile. Our results showed no general frictional effects on stiffness, notably for dynamical problems. So we assume that friction terms of the SWEs could be just mildly-stiff within the range of parameters considered.

Acknowledgements The corresponding author is indebted to the Graduate-Institute of the Bonn-Rhein-Sieg University of Applied Sciences for financial support by PhD scholarship. As well, the authors would like to thank the members of the Chair of Applied Mathematics/Numerical Analysis at University Wuppertal for their fruitful comments and discussions.

References

1. Chanson, H.: The Hydraulics of Open Channel Flow: An Introduction, 2nd edn. Elsevier, Oxford, Burlington (2004)
2. Gerisch, A., Verwer, J.G.: Operator splitting and approximate factorization for taxis-diffusion-reaction models. Appl. Numer. Math. **42**, 159–176 (2002)
3. Hundsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Springer, Berlin, Heidelberg, New York (2003)
4. Kurganov, A., Levy, D.: A third-order semidiscrete central scheme for conservation laws and convection-diffusion equations. SIAM J. Sci. Comput. **22**, 1461–1488 (2000)
5. Kurganov, A., Noelle, S., Petrova, G.: Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton-Jacobi equations. SIAM J. Sci. Comput. **23**, 707–740 (2001)
6. Rang, J.: An analysis of the Prothero-Robinson example for constructing new DIRK and ROW methods. J. Comput. Appl. Math. **262**, 105–114 (2014)
7. Steihaug, T., Wolfbrandt, A.: An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations. Math. Comput. **33**, 521–534 (1979)
8. Steinebach, G., Rentrop, P.: An adaptive method of lines approach for modeling flow and transport in rivers. In: Vande Wouwer, A., Saucez, Ph., Schiesser, W.E. (eds.) Adaptive Method of Lines, pp. 181–205. Chapman & Hall/CRC, Boca Raton, London, New York, Washington, DC (2001)
9. Steinebach, G., Weiner, R.: Peer methods for the one-dimensional shallow-water equations with CWENO space discretization. Appl. Numer. Math. **62**, 1567–1578 (2012)
10. Steinebach, G., Rosen, R., Sohr, A.: Modeling and numerical simulation of pipe flow problems in water supply systems. In: Martin, A., Klamroth, K., et al. (eds.) Mathematical Optimization of Water Networks. International Series of Numerical Mathematics, vol. 162, pp. 3–15. Springer, Basel (2012)

MS 33
MINISYMPOSIUM: SIMULATION ISSUES FOR NANOELECTRONIC COUPLED PROBLEMS

Organizers

Jan ter Maten[1] and Caren Tischendorf[2]

Speakers

Rick Jannsen[3] Jan ter Maten[1] and Caren Tischendorf[2]
The European Project nanoCOPS for Nanoelectronic Coupled Problems Solution

Wim Schoenmaker[4], Olivier Dupuis[5] Bart De Smedt[6] and Peter Meuris[7]
Fully-Coupled Electro-Thermal Power Device Fields

David Duque Guerra[8] and Sebastian Schöps[9]
Fast and Reliable Simulations of the Heating of Bond Wires

[1]Jan ter Maten, Bergische Universität Wuppertal, Wuppertal, Germany.
[2]Caren Tischendorf, Humboldt-Universität zu Berlin, Berlin, Germany.
[3]Rick Jannsen, NXP Semiconductors, Eindhoven, The Netherlands.
[4]Wim Schoenmaker, MAGWEL NV, Leuven, Belgium.
[5]Oliver Dupuis, MAGWEL NV, Leuven, Belgium.
[6]Bart De Smedt, MAGWEL NV, Leuven, Belgium.
[7]Peter Meuris, MAGWEL NV, Leuven, Belgium.
[8]David Duque Guerra, TU Darmstadt, Darmstadt, Germany.
[9]Sebastian Schöps, TU Darmstadt, Darmstadt, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_112

Lihong Feng[10], Athanasios C. Antoulas[11] and Peter Benner[12]
Automatic Generation of Reduced Order Models for Linear Parametric Systems

Caren Tischendorf[2]
Dynamic Coupled Electromagnetic Field Circuit Simulation

Keywords

Coupled problems
Model order reduction
Multirate methods
Partial differential algebraic equations
Time-domain simulation

Short Description

This minisymposium addresses simulation issues for the design development in nanoelectronics. In order to meet the challenge of large-size simulation problems involving EM-circuit-heat couplings the following mathematical topics shall be covered: co-simulation and coupled monolithic simulation methods for partial differential algebraic equation, multirate envelope methods, reduced basis methods for stochastic partial differential equations and parameterized model order reduction approaches.

We present recent advances of the European collaborative research project nanoCOPS in which eight academic partners, two large-scale semiconductor companies and two SMEs develop, extend and validate design tools for nanoelectronic IC and EM simulation.

[10]Lihong Feng, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

[11]Athanasios C. Antoulas, Rice University, Houston, USA.

[12]Peter Benner, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

Automatic Generation of Reduced-Order Models for Linear Parametric Systems

Lihong Feng, Athanasios C. Antoulas, and Peter Benner

Abstract Parametric modeling as well as parametric model order reduction (PMOR) of parametric systems are being widely researched in many micro- and nano-electrical(-mechanical) problems as well as in coupled micro- and nano-electro-thermal problems. We propose an adaptive technique for automatically implementing PMOR, so as to automatically construct the reduced-order models. The adaptive technique is based on a posteriori error estimation and is realized through a greedy algorithm which uses the error estimation as a stopping criteria.

Keywords Model order reduction • Multi-moment-matching • Parametric model order reduction

1 Introduction

Geometrical and physical variations are becoming unavoidable in many micro- and nano-electrical(-mechanical) problems as well as in coupled micro- and nano-electro-thermal problems. For design purposes, it is desired to extract a parametric model, where geometrical or physical variations appear as parameters in the system. For very large-scale parametric systems, parametric model order reduction (PMOR) attracts more and more attention due to its great potential of reducing the simulation time. Through PMOR, all the parameters are preserved in the reduced-order models as symbolic quantities. The goal is that a single reduced-order model is capable of replacing the original large-scale systems.

PMOR methods are often the extensions of standard model order reduction methods for non-parametric systems. Till now, there have been various PMOR methods proposed, such as multi-moment-matching PMOR methods [3], methods

L. Feng (✉) • P. Benner
Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany
e-mail: feng@mpi-magdeburg.mpg.de; benner@mpi-magdeburg.mpg.de

A.C. Antoulas
Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005-1892, USA
e-mail: aca@rice.edu

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_113

based on $\mathscr{H}_2$-optimal interpolation [1], methods based on proper orthogonal decomposition (POD) and interpolation [4], the Loewner approach to parametric model reduction [6], as well as the reduced basis methods (RBM) [7].

In micro- and nanoelectronics and MEMS design, the multi-moment-matching PMOR methods for linear parametric systems are still the most popular approaches used in practical applications since they are easy to implement and require few assumptions on system properties. However, constructing the reduced-order models using the multi-moment-matching PMOR method is still not automatic, for example, the expansion points need to be heuristically determined before implementation. How to automatically construct the reduced-order models for parametric systems, especially using the multi-moment-matching PMOR method, is rarely discussed by far. In this paper, we propose a technique of adaptively implementing the multi-moment-matching PMOR method. As a result, it automatically generates the reduced-order model, which is desired in design automation in real applications. The basic idea of the technique is to use an a posteriori output error bound for the reduced-order model as a stopping criteria in a greedy algorithm, which finally enables adaptive implementation of the multi-moment-matching PMOR method.

The paper is organized as follows. In Sect. 2, we review the basic idea of multi-moment-matching PMOR methods. Section 3 proposes an a posteriori output error bound for the transfer function of the reduced-order model. Section 4 presents an algorithm for adaptively implementing the multi-moment-matching PMOR method. In Sect. 5, a reduced-order model is automatically obtained for a parametric model of a silicon-nitride membrane. Conclusions are given in the end.

2 Multi-Moment-Matching PMOR

In this section multi-moment-matching PMOR methods for model order reduction of linear parametric systems are discussed. Especially, the robust multi-moment-matching PMOR method in [3] is reviewed. A linear parametrized system can be written as,

$$\begin{aligned} E(\tilde{\mu})\tfrac{dx}{dt} &= A(\tilde{\mu})x + Bu(t), \\ y(t,\tilde{\mu}) &= Cx, \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the vector of unknowns, $u(t) \in \mathbb{R}^{m_1}$ is the input signal and $y(t,\tilde{\mu}) \in \mathbb{R}^{m_2}$ is the output response. $E \in \mathbb{R}^{n\times n}, A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m_1}, C \in \mathbb{R}^{m_2\times n}$ are the system matrices. $\tilde{\mu} \in \mathbb{R}^{p-1}$ is a vector of parameters. The number of degrees of freedom n is usually very large.

Through PMOR, a reduced-order model is obtained as

$$\begin{aligned} \hat{E}(\tilde{\mu})\tfrac{dz}{dt} &= \hat{A}(\tilde{\mu})z + \hat{B}u(t), \\ \hat{y}(t,\tilde{\mu}) &= \hat{C}z, \end{aligned} \tag{2}$$

where $\hat{E} = V^T E(\mu) V$, $\hat{A} = V^T A(\mu) V$, $\hat{B} = V^T B$, $\hat{C} = CV$, and $z \in \mathbb{R}^r$, with $r \ll n$.

For multi-moment-matching PMOR methods, the matrix $V \in \mathbb{R}^{n \times r}$ is computed based on the series expansion of the state vector x in the frequency domain. Applying the Laplace transform to the original system in (1), and taking the input $u(t)$ as the impulse input, we get the following system in frequency domain (assuming $x(\mu)|_{t=0} = 0$)

$$\begin{aligned} G(\mu)x(\mu) &= B, \\ y(\mu) &= Cx(\mu). \end{aligned} \tag{3}$$

Here $G(\mu) = sE(\tilde{\mu}) - A(\tilde{\mu})$, and $\mu = (\tilde{\mu}, s) =: (\mu_1, \ldots, \mu_p)$ is the vector of parameters in the frequency domain. The variable $s \in \mathbb{C}$ is the Laplace variable with imaginary part $\mathrm{Im}(s) = 2\pi f$, f being the frequency in Hz.

Assume that $G(\mu) \in \mathbb{C}^{n \times n}$ has the following affine form with respect to the parameters, i.e.

$$G(\mu) = E_0 + \mu_1 E_1 + \ldots + \mu_p E_p,$$

where E_i are constant matrices and are independent of the parameters. Given an expansion point $\boldsymbol{\mu}^0 = [\mu_1^0, \mu_2^0, \cdots, \mu_p^0]$, $x(\mu)$ in (3) can be expanded as

$$\begin{aligned} x &= [I - (\sigma_1 M_1 + \ldots + \sigma_p M_p)]^{-1} B_M \\ &= \sum_{i=0}^{\infty} (\sigma_1 M_1 + \ldots + \sigma_p M_p)^i B_M, \end{aligned} \tag{4}$$

where $\sigma_i = \mu_i - \mu_i^0$, $B_M = [G(\mu^0)]^{-1} B$, and $M_i = -[G(\mu^0)]^{-1} E_i$, $i = 1, 2, \ldots, p$.

When the number of the parameters p in a parametrized system is larger than 2, it is desired that multiple point expansion is used, such that the size of the reduced-order model can be kept small. Given a group of expansion points $\mu^i, i = 0, \ldots, k$, a matrix V_{μ^i} can be computed for each μ^i as

$$\mathrm{range}\{V_{\mu^i}\} = \mathrm{span}\{R_0, R_1, \ldots, R_q\}_{\mu^i}.$$

Here $R_j = [M_1, \ldots, M_p] R_{j-1}$, $j = 1, \ldots, q$, and $R_0 = B_M$. For each μ^i, M_i, and B_M are defined as above by replacing μ^0 with μ^i. The final projection matrix V is a combination (orthogonalization) of all the matrices V_{μ^i},

$$V = \mathrm{orth}\{V_{\mu^0}, \ldots, V_{\mu^k}\}. \tag{5}$$

The question of how to properly select the expansion points μ^i is still open in general, and will be addressed in this paper. An algorithm for adaptively selecting the expansion points is proposed in Sect. 4, where the a posteriori error bound $\Delta(\mu)$ plays a crucial role. In the next section we propose an a posteriori error bound for the transfer function of the reduced-order model in (2).

3 A Posteriori Error Bound

In order to derive the error bound for the transfer function $\hat{H}(\mu)$ of the reduced-order model (2), we need to define a primal system and a dual system. The primal system is defined as in (3). Its output $y(\mu)$ is exactly the transfer function $H(\mu)$ of the original system in (1), since the system (3) is derived using the impulse input $u(t) = \delta(t)$, where $\delta(t)$ is the δ function. The dual system is defined as

$$\begin{aligned} G^*(\mu)x^{du}(\mu) &= -C^T, \\ y^{du}(\mu) &= B^T x^{du}(\mu). \end{aligned} \tag{6}$$

Here, $G^*(\mu)$ is the conjugate transpose of $G(\mu)$. We also need the residuals caused by the reduced-order models for the primal and the dual systems. The reduced-order model of the primal system is obtained using V from (2) by

$$\begin{aligned} V^T G(\mu)Vz(\mu) &= V^T B, \\ \hat{y}(\mu) &= CVz(\mu), \end{aligned} \tag{7}$$

where $\hat{x}(\mu) = Vz(\mu)$ approximates $x(\mu)$. It can be easily seen that the transfer function $\hat{H}(\mu)$ of the reduced-order model in (2) equals $\hat{y}(\mu)$. The reduced-order model of the dual system is

$$\begin{aligned} (V^{du})^T G^*(\mu)V^{du}z^{du}(\mu) &= -(V^{du})^T C^T, \\ \hat{y}^{du}(\mu) &= B^T V^{du} z^{du}(\mu), \end{aligned} \tag{8}$$

where $\hat{x}^{du}(\mu) = V^{du}z^{du}(\mu)$ is the approximation of $x^{du}(\mu)$. The two residuals are $r^{pr}(\mu) = B - G(\mu)\hat{x}(\mu)$ and $r^{du}(\mu) = -C^T - G^*(\mu)\hat{x}^{du}(\mu)$.

Notice that the matrix V^{du} can be computed similarly as in (5), only by replacing the matrices $G(\mu^i)$ with $G^*(\mu^i)$, $i = 0, \ldots, k$, B with $-C^T$, and E_i with E_i^T, $i = 0, \ldots, p$. More specifically, defining, $C_M = -[G^*(\mu^i)]^{-1}C^T$, we compute

$$\text{range}\{V^{du}_{\mu^i}\} = \text{span}\{R_0^{du}, R_1^{du}, \ldots, R_q^{du}\}_{\mu^i},$$

where $R_j^{du} = [M_1^{du}, \ldots, M_p^{du}]^j R_{j-1}^{du}, j = 0, \ldots, q$. Here $M_i^{du} = -[G^*(\mu^i)]^{-1}E_i^T$ and $R_0^{du} = C_M$.

Defining two new variables $e(\mu) = (\hat{x}^{du}(\mu))^* r^{pr}(\mu)$ and $\tilde{y}(\mu) = \hat{y}(\mu) - e(\mu)$ and assuming that $G(\mu)$ satisfies

$$\inf_{\substack{w\in\mathbb{C}^n \\ w\neq 0}} \sup_{\substack{v\in\mathbb{C}^n \\ v\neq 0}} \frac{w^* G(\mu)v}{||w||_2||v||_2} = \beta(\mu) > 0, \tag{9}$$

we have the following theorem.

Theorem 1 *For a single-input single-output (SISO) linear parametric system as in (1), if $G(\mu)$ satisfies (9), then $|y(\mu) - \tilde{y}(\mu)| \leq \tilde{\Delta}(\mu)$, $\tilde{\Delta}(\mu) := \frac{||r^{du}(\mu)||_2 ||r^{pr}(\mu)||_2}{\beta(\mu)}$. As a result,*

$$|H(\mu) - \hat{H}(\mu)| = |y(\mu) - \hat{y}(\mu)| \leq \Delta(\mu),$$

where $\Delta(\mu) := \tilde{\Delta}(\mu) + |e(\mu)|$.

Remark The proof of the theorem and the analysis of the error bound, as well as the extension to multiple-input multiple-output (MIMO) system cannot be presented here due to space limitations, but is detailed in [5].

Remark By simple calculation, it can be seen that $\beta(\mu)$ is the smallest singular value of $G(\mu)$, so that the error bound $\Delta(\mu)$ is computable. When $G(\mu)$ is very large, the smallest singular value of the projected matrix $V^T G(\mu) V$ could be a heuristic approximation of $\beta(\mu)$.

By combing a greedy algorithm proposed for the reduced basis method [7] with the robust PMOR algorithm in [3], one can use the error bound to adaptively select the parameters, so as to automatically construct the reduced-order model. The algorithm of automatic generation of the reduced-order model is described in the next section.

4 Automatic Generation of the Reduced-Order Models

The algorithm in this section follows the idea of the greedy algorithm widely used in the reduced basis community. A large sample space Ξ_{train} of the parameters μ, covering the whole interesting parameter domain, must be initially given. During each step of the algorithm, a point $\hat{\mu}$ in Ξ_{train}, which causes the largest error [indicated by the error bound $\Delta(\mu)$], is chosen as the next expansion point. The process continues until the error bound is smaller than an acceptable error tolerance $\epsilon_{tol} (< 1)$. The matrix V is used to construct the reduced-order model in (2). The matrix V_{du} only aids the computation of the error estimation $\Delta(\mu)$, and is not used in the final reduced-order model.

5 Simulation Results

We use a thermal model of a silicon-nitride membrane to illustrate the process of automatically generating the reduced-order model according to Algorithm 1. The physical description of the model can be found in [2]. One can also refer to the MORwiki (www.modelreduction.org) for the details of the model. It is a system

Table 1 $V_{\mu^i} = \text{span}\{B_M, R_1\}_{\mu^i}$, $\epsilon^{re}_{tol} = 10^{-2}$, $n = 60,020$, $r = 8$

Iteration	ε^{re}_{true}	$\Delta^{re}(\mu^i)$
1	1×10^{-3}	3.44
2	1×10^{-4}	4.59×10^{-2}
3	2.80×10^{-5}	4.07×10^{-2}
4	2.58×10^{-6}	2.62×10^{-5}

with four parameters as described in (10),

$$\begin{aligned}(E_0 + \rho c_p E_1)dx/dt + (K_0 + \kappa K_1 + hK_2)x &= bu(t)\\ y &= Cx,\end{aligned} \tag{10}$$

where $\rho \in [3000, 3200]$, $c_p \in [400, 750]$, $\kappa \in [2.5, 4]$, $h \in [10, 12]$. Here the mass density ρ in kg/m^3, the specific heat capacity c_p in J/kg/K, the thermal conductivity in W/m/K, and the heat transfer coefficient h in $\text{W/m}^2\text{/K}$. The size of the system is $n = 60,020$.

In Table 1, the true relative error is defined as $\varepsilon^{re}_{true} = \max_{\mu \in \Xi_{train}} |H(\mu) - \hat{H}(\mu)|/|H(\mu)|$. We use the error bound for the relative error defined as $\Delta^{re}(\mu) = \Delta(\mu)/|\hat{H}(\mu)|$, to estimate the true relative error. In the table, we show that after four iteration steps in Algorithm 1, a reduced-order model satisfying the error tolerance $\varepsilon^{re}_{tol} = 10^{-2}$ is finally derived. The error bound in the final step is very close but above the true error, which is to be expected from a rigorous error bound.

For the train sample space Ξ_{train} in Algorithm 1, we have used three samples for κ, ten samples for the frequency. In Fig. 1, we use 16 samples for κ, and 51 samples for the frequency to further check the accuracy of the reduced-order model. The plot shows that the error of the reduced-order model at every sample is below the error tolerance. The size of the reduced-order model is also very small, $r = 8$, showing that the automatically derived reduced-order model meets both the requirements of accuracy and compactness.

Algorithm 1 Automatic generation of the reduced-order model by adaptively selecting expansion points $\hat{\mu}$ for parametrized Linear time-invariant (LTI) systems

1: $V = []$; $V^{du} = []$;
2: $\epsilon = 1$; ϵ_{tol}: acceptable error of the reduced-order model;
3: Initial expansion point: $\hat{\mu}$;
4: Ξ_{train}: a large set of samples of μ, taken over the interesting domain of the parameters;
5: **while** $\epsilon > \epsilon_{tol}$ **do**
6: $\quad \text{range}(V_{\hat{\mu}}) = \text{span}\{R_0, R_1, \ldots, R_q\}_{\hat{\mu}}$;
7: $\quad \text{range}(V^{du}_{\hat{\mu}}) = \text{span}\{R^{du}_0, R^{du}_1, \ldots, R^{du}_q\}_{\hat{\mu}}$;
8: $\quad V = \text{orth}\{V, V_{\hat{\mu}}\}$;
9: $\quad V^{du} = \text{orth}\{V^{du}, V^{du}_{\hat{\mu}}\}$;
10: $\quad \hat{\mu} = \underset{\mu \in \Xi_{train}}{\text{argmax}}\ \Delta(\mu)$;
11: $\quad \epsilon = \Delta(\hat{\mu})$;
12: **end while**.

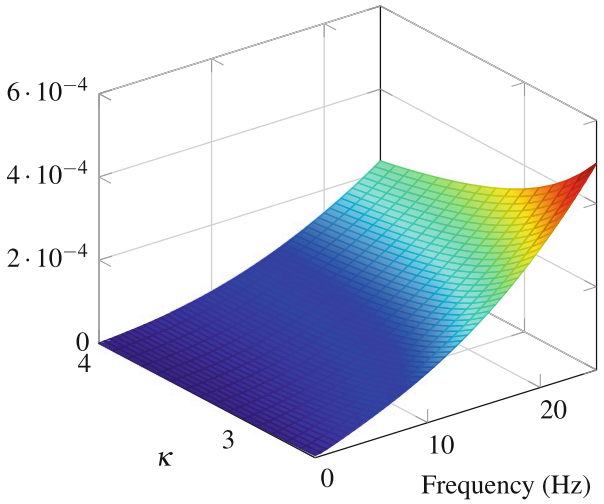


Fig. 1 Relative error of the final ROM over a fine space of samples

6 Conclusions

In this paper we have proposed an a posteriori error bound for reduced-order modeling of linear parametric systems. Guided by the error bound, reduced-order models can be automatically derived by the multi-moment-matching PMOR method. The simulation results show that the proposed algorithm automatically constructs the reduced-order models according to the given accuracy requirement. This provides a promising way of design automation for compact modeling of parametric problems.

Acknowledgements This work is supported by the collaborative project nanoCOPS, Nanoelectronic COupled Problems Solutions, supported by the European Union in the FP7-ICT-2013-11 Program under Grant Agreement Number 619166.

References

1. Baur, U., Benner, P., Beattie, C., Gugercin, S.: Interpolatory projection methods for parameterized model reduction. SIAM J. Sci. Comput. **33**, 2489–2518 (2011)
2. Bechtold, T., Hohlfeld, D., Rudnyi, E.B., Guenther, M.: Efficient extraction of thin-film thermal parameters from numerical models via parametric model order reduction. J. Micromech. Microeng. **20**, 045030 (2010)
3. Benner, P., Feng, L.: A robust algorithm for parametric model order reduction based on implicit moment-matching. In: Quarteroni, G.R.A. (ed.) Reduced Order Methods for Modeling and Computational Reduction. MS&A, vol. 9, pp. 159–186. Springer, Cham (2014)

4. Bui-Thanh, T., Willcox, K., Ghattas, O.: Model reduction for large-scale systems with high-dimensional parametric input space. SIAM J. Sci. Comput. **30**, 3270–3288 (2008)
5. Feng, L., Benner, P., Antoulas, A. C.: An a posteriori error bound for reduced order modeling of micro-and nano-electrical (-mechanical) systems. In: SCEE-2014 (Scientific Computing in Electrical Engineering), Wuppertal, Germany (2014)
6. Lefteriu, S., Antoulas, A.C., Ionita, A.C.: Parametric model reduction in the Loewner framework. In: Proceedings of 18th IFAC World Congress, pp. 12752–12756 (2011)
7. Patera, A.T., Rozza, G.: Reduced basis approximation and a posteriori error estimation for parametrized partial differential equations. MIT Pappalardo Graduate Monographs in Mechanical Engineering, Version 1.0, Copyright MIT 2006 (2007)

Fast and Reliable Simulations of the Heating of Bond Wires

David Duque and Sebastian Schöps

Abstract We present an extended analytic formulation for the determination of the temperature distribution along a bond-wire within a package in order to extract the maximum allowable current to not exceed a specific temperature. The closed-form formula involves the essential physical parameters that define a package, i.e., moulding compound material and dimensions, bond-wire characteristics, etc. This is very important if one wants to assess the influence of (randomly distributed) parameter variations on the current capacity of the wire by means of uncertainty quantification methods.

Keywords Bond wires • Moulding compound • Temperature distribution

1 Introduction

Among the several available techniques to provide electric connection between the chip and the lead frame (or pins) during device assembly, wire-bonding is still the most cost-effective one [2–6]. This techniques uses fine aluminium, copper, or gold wires to establish an electric path between the chip and its package. As chip miniaturisation becomes inevitable, the wire diameter must also decrease. Since the electric power for the chip has to be supplied through the wires, these are heated up and their temperature can increase substantially because high current densities may occur. As a matter of fact, fusing or melting of wire-bonds is one potential source of failure in integrated circuit (IC) devices [5]. From the afore-described situation, one would like to have available a simple formula that enables to predict the safe operation range of a given bond-wire in a particular application. This calculation should be done in a expedient manner and must involve the important physical parameters defining a package. Several simplistic analytic formulations for the estimation of current capacities in bond-wires have already been proposed [2–6]. Nonetheless, in their attempt to simplify the resulting partial differential

D. Duque (✉) • S. Schöps
Technische Universität Darmstadt, Graduate School of Computational Engineering and Institut für Theorie Elektromagnetischer Felder, Schloßgartenstr. 8, 64289 Darmstadt, Germany
e-mail: duque@gsc.tu-darmstadt.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_114

equation (PDE) so as to cope with the non-linearities introduced by the thermal dependence of the wire parameters, the resulting formulations end up lacking the variables that define a package as a geometrical shape.

In this paper, we address these typical deficiencies by developing an analytic formulation for the determination of the wire temperature which involves the essential physical parameters that define a package, i.e., moulding compound material and dimensions, bond-wire characteristics, etc., by using an appropriate set of heat transfer boundary conditions (BCs).

2 Problem Formulation

We show a simple diagram of a classic IC lead-frame package in Fig. 1a. Because of the often complicate geometric arrangement of the conductors within the package, simplifications are necessary in order to formulate the relevant heat transfer problem. In Fig. 1b, we depict a suitable bond-wire thermal problem, which

consists of a rectangular piece of moulding compound of height H_m and width W_m that encapsulates the bond-wire. The compound is characterised by a thermal conductivity κ_m, specific heat $c_{\mathrm{e;m}}$, and mass density ρ_m. Similarly, the bond-wire of length L_w is characterised by its specific heat $c_{\mathrm{e;w}}$, mass density ρ_w, and linearised thermal conductivity and electrical resistivity

$$\kappa_\mathrm{w}(\widetilde{T}_\mathrm{w}) := \kappa_\mathrm{o}\left(1+\alpha_\kappa \widetilde{T}_\mathrm{w}\right), \qquad \text{and} \qquad \rho_{\mathrm{e;w}}(\widetilde{T}_\mathrm{w}) := \rho_{\mathrm{e;o}}\left(1+\alpha_\rho \widetilde{T}_\mathrm{w}\right), \tag{1}$$

respectively, with $\widetilde{T}_\mathrm{w} = T_\mathrm{w} - T_\mathrm{o}$, where T_w denotes the wire temperature, T_o the reference (ambient) temperature, α_κ the thermal conductivity temperature coefficient, and α_ρ the electrical resistivity temperature coefficient.

The wire is heated up, during a time t_p, by the action of an electric current $i(t) = I_0$. We want to determine the temperature T_w as function of time. We impose suitable BCs on the domain boundaries, i.e., On the rightmost wall, we assume that heat

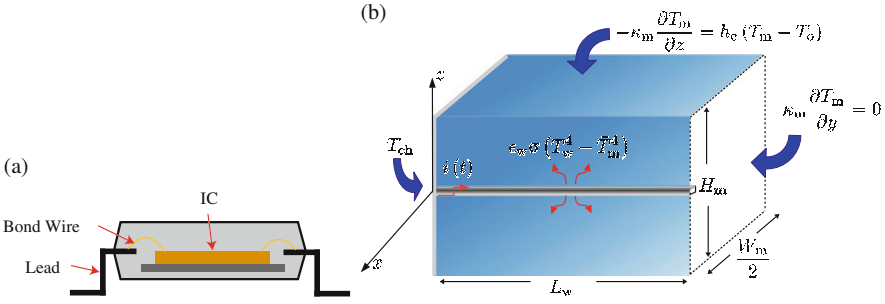


Fig. 1 Diagram of a classic IC lead-frame package and bond-wire heat transfer problem configuration. (**a**) IC lead-frame package. (**b**) Heat transfer problem configuration

flux mainly occurs through the wire, thus we require the vanishing of the heat flux throughout the moulding compound section of this wall, while the wire remains at the lead temperature T_{ld}, viz.

$$-\kappa_{\mathrm{m}}\frac{\partial}{\partial y}T_{\mathrm{m}}\left(x, L_{\mathrm{w}}, z, t\right) = 0, \quad T_{\mathrm{w}}\left(x, L_{\mathrm{w}}, z, t\right) = T_{\mathrm{ld}}, \tag{2}$$

where T_{m} is the compound temperature. On the leftmost wall, we assume a constant chip temperature T_{ch}, viz.

$$T_{\mathrm{m}}\left(x, 0, z, t\right) = T_{\mathrm{ch}}, \quad T_{\mathrm{w}}\left(x, 0, z, t\right) = T_{\mathrm{ch}}. \tag{3}$$

On the remaining lateral walls, we assume convective heat transfer [1], viz.

$$-\kappa_{\mathrm{m}}\frac{\partial}{\partial z}T_{\mathrm{m}}\left(x, y, \pm\frac{H_{\mathrm{m}}}{2}, t\right) = h_{\mathrm{c}}\left(T_{\mathrm{m}}(x, y, \pm\frac{H_{\mathrm{m}}}{2}, t) - T_{\mathrm{o}}\right), \tag{4}$$

$$-\kappa_{\mathrm{m}}\frac{\partial}{\partial x}T_{\mathrm{m}}\left(\pm\frac{W_{\mathrm{m}}}{2}, y, z, t\right) = h_{\mathrm{c}}\left(T_{\mathrm{m}}(\pm\frac{W_{\mathrm{m}}}{2}, y, z, t) - T_{\mathrm{o}}\right), \tag{5}$$

where h_{c} is the convective heat transfer coefficient [1]. Finally, on the wire surface, we assume that thermal radiation takes place, viz.

$$-\int_{S_{\mathrm{w}}}\kappa_{\mathrm{w}}\nabla_{\mathrm{T}}T_{\mathrm{w}}\cdot\mathrm{d}\boldsymbol{S} = \int_{S_{\mathrm{w}}}\epsilon_{\mathrm{w}}\sigma\left(T_{\mathrm{w}}^4 - T_{\mathrm{o}}^4\right)\mathrm{d}S, \tag{6}$$

where $\nabla_{\mathrm{T}}T_{\mathrm{w}}$ is the temperature *transverse* gradient, that is along the xz-plane, S_{w} is the wire surface, ϵ_{w} is the wire emissivity [1], and σ is the Stefan-Boltzmann constant [1].

3 Bond-Wire Heat Transfer Problem

The heat equation for the stationary and incompressible wire of constant mass density reads [1]

$$\rho_{\mathrm{w}}c_{\mathrm{e;w}}\frac{\partial T_{\mathrm{w}}}{\partial t} = \nabla\cdot\left(\kappa_{\mathrm{w}}\nabla T_{\mathrm{w}}\right) + \dot{q}_{\mathrm{i}}, \tag{7}$$

where $\dot{q}_{\mathrm{i}}$ is the *impressed* volume thermal power density generated within the wire, which is

$$\dot{q}^{\mathrm{i}} = \frac{I_0^2\rho_{\mathrm{e;o}}}{A_{\mathrm{w}}^2} + \frac{I_0^2\rho_{\mathrm{e;o}}\alpha_\rho\widetilde{T}_{\mathrm{w}}}{A_{\mathrm{w}}^2}, \tag{8}$$

where A_w is the cross-section area of the wire. The wire heat equation is obtained by using (6), expanding the gradient operator as $\nabla = \partial_y \boldsymbol{a_y} + \nabla_\mathrm{T}$, and assuming that the wire Biot number [1] is small along the xz-plane, viz.

$$\rho_\mathrm{w} c_{\mathrm{e;w}} \frac{\partial}{\partial t} T_\mathrm{w}(y,t) = \frac{\partial}{\partial y}\left(\kappa_\mathrm{w} \frac{\partial}{\partial y} T_\mathrm{w}(y,t)\right) - \epsilon_\mathrm{w}\sigma\left(T_\mathrm{w}^4(y,t) - T_\mathrm{o}^4\right)\frac{C_\mathrm{w}(y)}{A_\mathrm{w}(y)} + \dot{q}_\mathrm{i}, \tag{9}$$

where C_w is the cross-section perimeter of the wire. We linearise (9) by first expanding the radiation term as follows

$$T_\mathrm{w}^4 - T_\mathrm{o}^4 = \left(T_\mathrm{w}^3 + T_\mathrm{w}^2 T_\mathrm{o} + T_\mathrm{w} T_\mathrm{o}^2 + T_\mathrm{o}^3\right)(T_\mathrm{w} - T_\mathrm{o}) \cong \chi_\mathrm{w}(y,t)\,(T_\mathrm{w} - T_\mathrm{o}), \tag{10}$$

where $\chi_\mathrm{w}(y,t)$ is a linearising function whose formal calculation entails the solution of an integral equation that ensues from the continuity of temperature and heat flux across the wire-mould interface. Since, we are aiming at a simple and fast method for determining the wire temperature, we deem it convenient to regard $\chi_\mathrm{w}(y,t) \equiv \chi_\mathrm{w}$ a constant in keeping with the observation that temperature continuity along the interface cannot be rigorously imposed. This approximation is reasonable insofar as thermal radiation is not the heat transfer dominant term. Next, we employ the following transformation

$$\tilde{\theta}_\mathrm{w}\left(\widetilde{T}_\mathrm{w}\right) := \frac{1}{\kappa_\mathrm{o}} \int_0^{\widetilde{T}_\mathrm{w}} \kappa_\mathrm{w}(s)\,\mathrm{d}s, \tag{11}$$

which entails that $\tilde{\theta}_\mathrm{w} = \widetilde{T}_\mathrm{w} + \alpha_\kappa/2\,\widetilde{T}_\mathrm{w}^2$, and $\partial_y \tilde{\theta}_\mathrm{w} = \kappa_\mathrm{w}/\kappa_\mathrm{o}\,\partial_y \widetilde{T}_\mathrm{w}$. Additionally, to keep the transient term in (9) linear, we employ $\partial_t \tilde{\theta}_\mathrm{w} = \partial_t \widetilde{T}_\mathrm{w}$ as an approximation. These steps yield

$$\rho_\mathrm{w} c_{\mathrm{e;w}} \frac{\partial}{\partial t}\tilde{\theta}_\mathrm{w} = \kappa_\mathrm{o} \frac{\partial^2}{\partial y^2}\tilde{\theta}_\mathrm{w} - F_{\mathrm{o;w;r}}\tilde{\theta}_\mathrm{w} + G_{\mathrm{o;w}} + \frac{1}{2}H_{\mathrm{o;w;r}}, \tag{12}$$

with

$$G_{\mathrm{o;w}} = \frac{I_0^2 \rho_{\mathrm{e;o}}}{A_\mathrm{w}^2}, \quad F_{\mathrm{o;w;r}} = \epsilon_\mathrm{w}\sigma\chi_\mathrm{w}\frac{C_\mathrm{w}}{A_\mathrm{w}}, \quad H_{\mathrm{o;w;r}} = \frac{2I_0^2\rho_{\mathrm{e;o}}\alpha_\rho \widetilde{T}_{\mathrm{w;e}}}{A_\mathrm{w}^2} + \epsilon_\mathrm{w}\sigma\chi_\mathrm{w}\frac{C_\mathrm{w}}{A_\mathrm{w}}\alpha_\kappa \widetilde{T}_{\mathrm{w;e}}^2, \tag{13}$$

where $\widetilde{T}_{\mathrm{w;e}}$ is the wire *effective* temperature. We solve (12) by assuming $\tilde{\theta}_\mathrm{w}(y,t) = \tilde{\theta}_{\mathrm{w};1}(y,t) + \tilde{\theta}_{\mathrm{w};2}(y)$, thus yielding

$$\begin{aligned}\tilde{\theta}_\mathrm{w}(y,t) = \sum_k C_{\mathrm{w};k;\mathrm{r}}^{\mathrm{t}} e^{-\frac{\kappa_\mathrm{o}}{\rho_\mathrm{w} c_{\mathrm{e;w}}}\lambda_{y;\mathrm{w},k}^2 t} e^{-\frac{F_{\mathrm{o;w;r}}}{\rho_\mathrm{w} c_{\mathrm{e;w}}}t} \sin\left(\lambda_{y;\mathrm{w},k} y\right) + C_{1;y;\mathrm{w;r}}^{\mathrm{s}} \cosh\left(\sqrt{\frac{F_{\mathrm{o;w;r}}}{\kappa_\mathrm{o}}}y\right) \\ + C_{2;y;\mathrm{w;r}}^{\mathrm{s}} \sinh\left(\sqrt{\frac{F_{\mathrm{o;w;r}}}{\kappa_\mathrm{o}}}y\right) + \frac{1}{2}\frac{H_{\mathrm{o;w;r}}}{F_{\mathrm{o;w;r}}} + \frac{G_{\mathrm{o;w}}}{F_{\mathrm{o;w;r}}};\ \lambda_{y;\mathrm{w},k} = \frac{k\pi}{L_\mathrm{w}},\ k > 0.\end{aligned} \tag{14}$$

The set $\{C^{\mathrm{t}}_{\mathrm{w};k;\mathrm{r}}, C^{\mathrm{s}}_{1;y;\mathrm{w};\mathrm{r}}, C^{\mathrm{s}}_{2;y;\mathrm{w};\mathrm{r}}\}$ is determined by means of the initial condition $T_{\mathrm{w}} = T_{\mathrm{o}}$ at $t = 0$, and BCs (2) and (3), respectively. Thus, we finally arrive at

$$\widetilde{T}_{\mathrm{w}}(y,t) \cong \frac{\sqrt{2\alpha_\kappa \tilde{\theta}_{\mathrm{w}}(y,t)+1}}{\alpha_\kappa} - \frac{1}{\alpha_\kappa}. \tag{15}$$

We observe in (15) that if $|\alpha_\kappa| \ll 1$, then $\widetilde{T}_{\mathrm{w}} \cong \tilde{\theta}_{\mathrm{w}}$; moreover, we still need to determine $\widetilde{T}_{\mathrm{w;e}}$ and χ_{w}.

4 Moulding Compound Heat Transfer Problem

The moulding compound heat equation can be expressed as

$$\rho_{\mathrm{m}} c_{\mathrm{e;m}} \frac{\partial}{\partial t} T_{\mathrm{m}} = \kappa_{\mathrm{m}} \nabla^2 T_{\mathrm{m}} + \epsilon_{\mathrm{w}} \sigma \chi_{\mathrm{w}} \widetilde{T}_{\mathrm{w}} C_{\mathrm{w}} \delta(x)\delta(z), \tag{16}$$

with $\delta(.)$ a Dirac delta. Equation (16) is the heat equation of a homogeneous compound with an impressed heat source, whose solution involves the relevant heat kernel (Green's function). If the impressed source were zero in (16), the temperature T_{m} would be established by the chip. Once, the impressed source is activated, another temperature component originates. We solve for the *first* component by defining $\widetilde{T}_{\mathrm{m}} \equiv T_{\mathrm{m}} - T_{\mathrm{o}}$, assuming $\widetilde{T}_{\mathrm{m}}(x,y,z,t) \equiv \widetilde{T}_{\mathrm{m;1}}(x,y,z,t) + \widetilde{T}_{\mathrm{m;2}}(x,y,z)$, and using BCs (2)–(5). The heat kernel is obtained by defining $\widetilde{G}_{\mathrm{m}} \equiv G_{\mathrm{m}} - T_{\mathrm{o}}$, assuming $\widetilde{G}_{\mathrm{m}}(x,y,z,t) \equiv \widetilde{G}_{\mathrm{m;1}}(x,y,z,t) + \widetilde{G}_{\mathrm{m;2}}(x,y,z)$, and using similar BCs. Subsequently, we can express T_{m} as

$$\begin{aligned} T_{\mathrm{m}}(x,y,z,t) = T_{\mathrm{o}} &+ \sum_n \sum_p C^{\mathrm{s}}_{\mathrm{m};n,p} e^{\lambda_{y;\mathrm{m};n,p} y} \left(1 + e^{2\lambda_{y;\mathrm{m};n,p}(L_{\mathrm{w}}-y)}\right) \cos(\lambda_{x;\mathrm{m},n} x) \\ &\cos\left(\lambda_{z;\mathrm{m},p} z\right) + \sum_n \sum_m \sum_p C^{\mathrm{t}}_{\mathrm{m};n,m,p} e^{-\frac{\kappa_{\mathrm{m}}}{\rho_{\mathrm{m}} c_{\mathrm{e;m}}}\left(\lambda^2_{x;\mathrm{m},n}+\lambda^2_{y;\mathrm{m},m}+\lambda^2_{z;\mathrm{m},p}\right)t} \\ &\cos(\lambda_{x;\mathrm{m},n} x) \sin\left(\lambda_{y;\mathrm{m},m} y\right) \cos\left(\lambda_{z;\mathrm{m},p} z\right) \\ &+ \epsilon_{\mathrm{w}} \sigma \chi_{\mathrm{w}} C_{\mathrm{w}} \int_0^t \int_{y'} G_{\mathrm{m}}\left(x,y,z,t-\tau,y'\right) \widetilde{T}_{\mathrm{w}}\left(y',\tau\right) \mathrm{d}y' \mathrm{d}\tau, \end{aligned} \tag{17}$$

with $\lambda_{x;\mathrm{m},n}\tan\left(\lambda_{x;\mathrm{m},n}W_\mathrm{m}/2\right) := h_\mathrm{c}/\kappa_\mathrm{m}$, $\lambda_{z;\mathrm{m},p}\tan\left(\lambda_{z;\mathrm{m},p}H_\mathrm{m}/2\right) := h_\mathrm{c}/\kappa_\mathrm{m}$, $\lambda_{y;\mathrm{m},m} := (2m+1)\,\pi/2L_\mathrm{w};\ m \geq 0$, $\lambda^2_{y;\mathrm{m};n,p} := \lambda^2_{x;\mathrm{m},n} + \lambda^2_{z;\mathrm{m},p}$, and

$$G_\mathrm{m}\left(x,y,z,t,y'\right) = T_\mathrm{o} + \sum_n\sum_p C^\mathrm{s}_{\mathrm{g};n,p}e^{\lambda_{y;\mathrm{m};n,p}y}\left(1+e^{2\lambda_{y;\mathrm{m};n,p}(L_\mathrm{w}-y)}\right)\cos\left(\lambda_{x;\mathrm{m},n}x\right)$$
$$\cos\left(\lambda_{z;\mathrm{m},p}z\right) + \sum_n\sum_m\sum_p C^\mathrm{t}_{\mathrm{g};n,m,p}\left(y'\right)e^{-\frac{\kappa_\mathrm{m}}{\rho_\mathrm{m}c_{\mathrm{e};\mathrm{m}}}\left(\lambda^2_{x;\mathrm{m},n}+\lambda^2_{y;\mathrm{m},m}+\lambda^2_{z;\mathrm{m},p}\right)t}$$
$$\cos\left(\lambda_{x;\mathrm{m},n}x\right)\sin\left(\lambda_{y;\mathrm{m},m}y\right)\cos\left(\lambda_{z;\mathrm{m},p}z\right). \tag{18}$$

The constant function χ_w is calculated by imposing

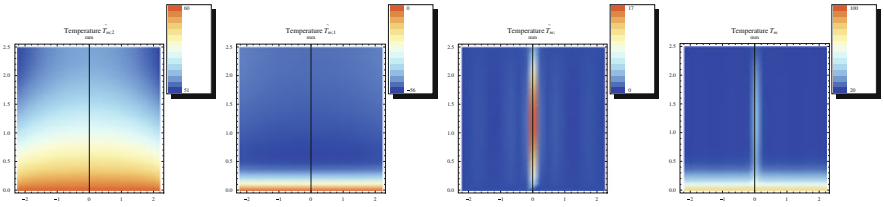
$$\int_0^{t_\mathrm{p}}\int_0^{L_\mathrm{w}}\lim_{z\to 0}\lim_{x\to 0}\widetilde{T}_\mathrm{m}\left(x,y_\mathrm{o},z,t\right)\mathrm{d}y\mathrm{d}t = \int_0^{t_\mathrm{p}}\int_0^{L_\mathrm{w}}\widetilde{T}_\mathrm{w}\left(y_\mathrm{o},t\right)\mathrm{d}y\mathrm{d}t,\ y_\mathrm{o}\in[0,L_\mathrm{w}]\,. \tag{19}$$

Above, the moulding and wire average temperatures should be equal at the interface point y_o. It now becomes evident why the *constant* χ_w does not allow for a rigorous temperature continuity across the moulding-wire interface. The answer is that a constant does not provide enough variability. The formal calculation of $\chi_\mathrm{w}\left(y,t\right)$ in (10) requires it within the convolution in (17), thus stating an integral equation via (19). Yet, by assuming a constant χ_w, we facilitate the estimation of the influence of the moulding in T_w by means of the radiation term. We can make the following observations about χ_w. (1) It is not unique. As matter of fact, it depends on the point y_o; (2) it exhibits a maximum (minimum) where T_w is maximum (minimum); (3) it can be chosen as the average value of these maximum and minimum. Thus, we estimate χ_w by means of (17) and (19), and approximating $\widetilde{T}_\mathrm{w}$ with its *linear*[1] counterpart $\tilde{\theta}_{\mathrm{w};0}$, where the subscript $_0$ denotes (14) when $\chi_{\mathrm{w};\mathrm{e},0} = \left(\bar{T}^3_{\mathrm{w};\mathrm{e}} + \bar{T}^2_{\mathrm{w};\mathrm{e}}T_\mathrm{o} + \bar{T}_{\mathrm{w};\mathrm{e}}T^2_\mathrm{o} + T^3_\mathrm{o}\right)$, with $\bar{T}_{\mathrm{w};\mathrm{e}} = \widetilde{T}_{\mathrm{w};\mathrm{e}} + T_\mathrm{o}$, is used to compute the terms in (13).

5 Numerical Results

We have implemented our approach in a Mathematica™ package, and have performed several test for wires of Gold (Au), Copper (Cu), and Aluminium (Al) with diameters $D_\mathrm{w} = \{0.8, 1.0, \ldots, 1.8, 2.0\}$ mil, and length $L_\mathrm{w} = 2.5$ mm. The current

[1]The *linear* adjective comes from the fact that $\tilde{\theta}_\mathrm{w}$ is the solution of the linearised wire heat transfer equation.



(a) $\widetilde{T}_{\mathrm{m};2}$ (b) $\widetilde{T}_{\mathrm{m};1}$ (c) $\widetilde{T}_{\mathrm{m};\mathrm{g}}$ (d) T_{m}

Fig. 2 Steady temperature component (**a**), transient temperature component (**b**), heat kernel temperature component (**c**) and moulding compound temperature (**d**)

pulse amounts to $I_0 = \{0, \ldots, 6\}$ A with $t_{\mathrm{p}} = 50$ ms, and the moulding compound, with dimensions $W_{\mathrm{m}} = 4.45$ mm and $H_{\mathrm{m}} = 1.48$ mm, is made of an Epoxy resin with $\kappa_{\mathrm{m}} = 0.870\,\mathrm{W/(m\,K)}$, $c_{\mathrm{e};\mathrm{m}} = 882\,\mathrm{J/(kg\,K)}$, and $\rho_{\mathrm{m}} = 1860\,\mathrm{kg/m^3}$.

To check the correctness of the approach, we have computed (17) at $z = 0$ by considering an Au-wire of $D_{\mathrm{w}} = 2$ mil carrying a current $I_0 = 3.7$ A during 50 ms. We have assumed $T_{\mathrm{ch}} = 80\,^{\circ}\mathrm{C}$, $T_{\mathrm{ld}} = 40\,^{\circ}\mathrm{C}$, $T_{\mathrm{o}} = 20\,^{\circ}\mathrm{C}$, and $h_{\mathrm{c}} = 25\,\mathrm{W/(m^2\,K)}$. Figure 2 shows each component in (17) at $t_{\mathrm{o}} = 50$ ms. Figure 2a and b show the steady (first) and transient (second) component of (17). By looking at the temperatures values along the x-axis, which are $\widetilde{T}_{\mathrm{m};2} = 60\,^{\circ}\mathrm{C}$ and $\widetilde{T}_{\mathrm{m};1} = 0\,^{\circ}\mathrm{C}$, we infer that BCs are satisfied. Figure 2c shows the heat kernel (third) component of (17). This component has been computed with a χ_{w} obtained by imposing (19) at $y_{\mathrm{o}} = L_{\mathrm{w}}$, where $T_{\mathrm{w}} = T_{\mathrm{ld}}$ is minimum. Consequently, thermal radiation is underestimated, and yet the maximum heat flux occurs at the wire mid-point and decreases towards the extremes. Figure 2d shows T_{m}; we can see that $T_{\mathrm{m}} = T_{\mathrm{ch}} = 80\,^{\circ}\mathrm{C}$ along the x-axis, and exhibits a minimum that equals $T_{\mathrm{o}} = 20\,^{\circ}\mathrm{C}$. Here, we must recall that T_{m} in Fig. 2d should be interpreted as a figure of merit which accounts for the compound effect on the wire temperature, and not as the *real* compound temperature.

Figure 3 shows the current capacity (T_{w} vs I_0) of the Al-, Au-, and Cu-wires before melting temperature is reached. T_{w} is calculated at the wire mid-point where maximum temperature is reached. The results in Fig. 3 are in the ballpark range with results from literature [6] under similar settings. Figure 3a–c demonstrates the capabilities of the approach to provide a safe range of operation for the bond-wires before melting temperature or moulding deterioration is reached. The approach is also capable of delimiting the wire current capacity by computing the wire mid-point temperature, for a given current amplitude, with the maximum and minimum constant χ_{w}.

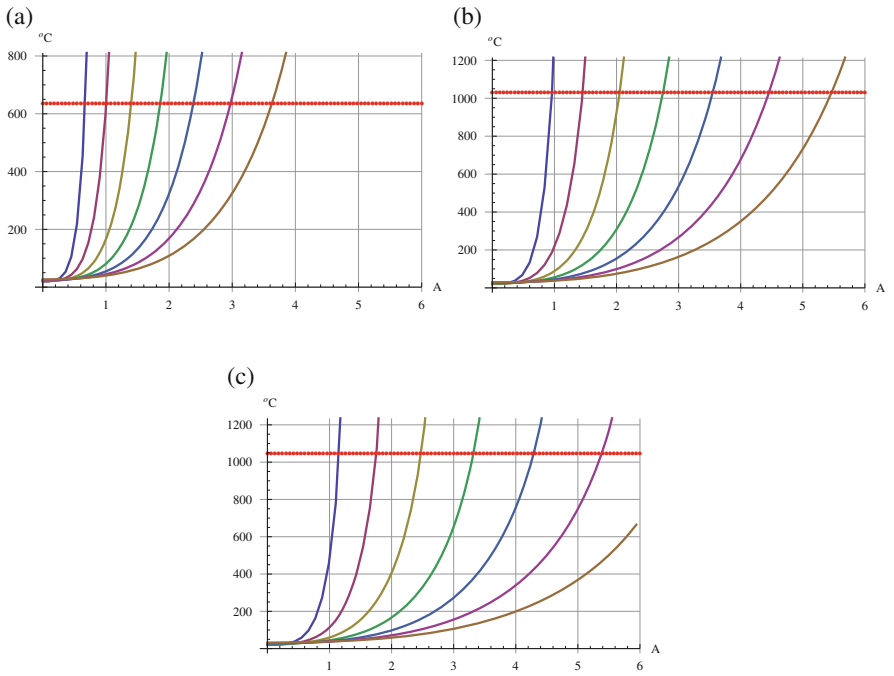


Fig. 3 Bond-wire current capacities for diameters $D_{\mathrm{w}} = \{0.8, 1.0, \ldots, 1.8, 2.0\}$ mil and $L_{\mathrm{w}} = 2.5$ mm: (**a**) Al-wire; (**b**) Au-wire; (**c**) Cu-wire

6 Conclusions

We have developed a simple analytic model for the evaluation of current capacities in bond-wires. The model uses suitable BCs which permit to take into account all parameters defining the heat transfer problem while still retaining the geometric shape of the package. The model permits to determine a safe range of operation for the bond-wires to not exceed a predefined temperature

Acknowledgements This work is a part of the project 'Nanoelectronic Coupled Problems Solutions' (nanoCOPS) funded by the European Union within FP7-ICT-2013 (grant no. 619166). The second author is also supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universitaet Darmstadt.

References

1. Incropera, F., DeWitt, D.: Introduction to Heat Transfer. Wiley, Hoboken (1985)
2. Loh, E.: Physical analysis of data of fused-open bond wires. IEEE Trans. Comp. Hybrids Manuf. Technol. **6**, 209–217 (1983)
3. Loh, E.: Heat transfer of fine-wire fuse. IEEE Trans. Comp. Hybrids Manuf. Technol. **7**, 264–267 (1984)
4. Mallik, A., Stout, R.: Simulation methods for predicting fusing current and time for encapsulated wire bonds. IEEE Trans. Electron. Packag. Manuf. **33**, 255–261 (2010)
5. Mertol, A.: Estimation of aluminum and gold bond wire fusing current and fusing time. IEEE Trans. Comp. Hybrids Manuf. Technol. **18**, 210–214 (1995)
6. Nöbauer, G.T., Moser, H.: Analytical approach to temperature evaluation in bonding wires and calculation of allowable current. IEEE Trans. Adv. Packag. **23**, 426–435 (2000)

Fully-Coupled Electro-Thermal Power Device Fields

Wim Schoenmaker, Olivier Dupuis, Bart De Smedt, and Peter Meuris

Abstract This paper presents a new solution method to deal with thermal effects in power designs. The new ingredients are: (1) the treatment of the electric and thermal fields are done fully self-consistent, (2) the dealing with (fragments of) the transistor fingers by using table models.

Keywords Coupled problem • Electric and thermal fields • Power device

1 Introduction

Power devices are very challenging from a designer's perspective. Whereas their basic operation principles are rather straightforward, numerous complications can arise due to less than optimal balancing of the current distributions, local heating effects and ultimately failure due to positive feed back loops. Until now thermal issues have been usually addressed by adding a 'thermal verification cycle' to the electrical design flow. This way of working has been 'justified' by the conviction that the thermal response takes place on a much larger time scale than the electrical one. As a consequence, the thermal variation is only noticeable at a much larger length scale and as a consequence, if we want an impression of the thermal field, a coarse-grain thermal mesh suffices to characterize the thermal response. However, we have found (using in-depth and detailed electro-thermal field solving) that this picture can not be sustained. In particular, there are fine-grained variations observed in the thermal plots, thereby falsifying above view point that thermal variations need to be incorporated only on a coarse-grain level. Moreover, the local variations not only impact the local current densities because the electrical conductance depends on the local temperature, but the thermal fields must be determined in a self-consistent way with the electric field intensities, since the latter directly provide the local heat generation. To summarize: power device characterization is only complete if the thermal response is incorporated in a fully consistent way with the electrical

W. Schoenmaker (✉) • O. Dupuis • B. De Smedt • P. Meuris
MAGWEL, Martelarenplein 13, 3000 Leuven, Belgium
e-mail: wim.schoenmaker@magwel.com; olivier.dupuis@magwel.com;
bart.desmet@magwel.com; peter.meuris@magwel.com

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_115

response. Popular means to address different problem areas (here: electrical and thermal) is by performing co-simulation. The basic idea is that through an iterative process one visits a series a simulation tools and feeding the latest finding of the prior sub-system simulation into the next one. This process is repeated until convergence (no noticeable updates) is found. Unfortunately, this approach is only applicable if the feedback of one tool on (one of) the other tool(s) is limited. In other words: if the physical coupling is weak. The latter is the case for long range correlations, but as we have argued, the electrical and thermal interfere on quite a local scale. The correlations are short-range and as a consequence co-simulation requires many cycles in order to reach convergence provided it is reached at all and not hooked up a limit cycle or divergence. Therefore, we propose (and present) an alternative to co-simulation, which we may view as a "holistic" or integrated simulation approach. The key idea is to deal with all the degrees of freedom at an integrated level. The cross coupling between sub groups of the degrees of freedom (electrical and thermal) are fully included. These couplings induce flow patterns is the state space which are not reachable in the co-simulation approach and thereby the number of iterations towards the solution is much smaller than in the co-simulation approach. Of course the holistic approach is less generic than the co-simulation approach because the data structures inside 3rd party software tools are usually not accessible. Therefore the couplings can not be determined and prohibiting a holistic solution strategy. As a consequence, the holistic solver must be constructed from scratch. This will be done in the next section.

2 The Integrated ET Solver

2.1 Electric Field Solver

The electrical part of the holistic field solver addresses the current-continuity equation

$$\nabla \cdot J + \frac{\partial \rho}{\partial t} = 0, \;\; J = \sigma E, \;\; \mathbf{E} = -\nabla V, \;\; \rho = -\nabla \left(\epsilon \nabla V\right) \tag{1}$$

In the present modeling, we will not consider local charging effects, since the electrical time response scale is assumed to several order of magnitude smaller than the thermal time constant. Therefore, the current-continuity equation reduces to a Poisson problem for conductive domains, being interconnects and active devices, i.e. the fingers of the power transistors. On the scale of the die, these fingers are truly microscopic and solving the current continuity equation in a detailed manner inside the device channel (TCAD) would require meshes with a prohibitively large number of nodes. Therefore, one must refrain from addressing sub-micron device details in the modeling and replace the transistor fingers (or fragments there off) by compact models as far as there current–voltage response in concerned. This

approach is common in power transistor modeling, where the active channels are replaced by an on-resistance Ron. Provided that local heating is not a major concern, this approach suffices. However, it should be remembered that the conductance, $\sigma = \sigma(T)$, is a temperature-dependent material parameter and if the temperature varies over position and time, T=T(x,t), this will effect the solution of (1). So it becomes mandatory to determine T(x,t).

2.2 Thermal Field Solver

The thermal part of the holistic field solver addressed the heat equation

$$\nabla \cdot Q + \frac{\partial w(T)}{\partial t} = \sum\nolimits_H, \quad Q = -\kappa \nabla T, \quad w(T) = C_T \left(T - T^{ref}\right) \tag{2}$$

In here, Q is the heat flux, κ is the thermal conductance and w the local energy storage characterized by the thermal capacitance C_T and T^{ref} is a typical reference (operational, environment) temperature.

The solution of this equation provides the desired temperature information to be fed into (1). However, the solution is only computable, provided that the heat source is known. The source may consist of several contributions. Energy may be converted to heat by radiative absorption. The boundaries of the simulation domain may contain heat-injecting or extracting properties. Besides these sources the Joule self heating is of particular interest.

$$\sum\nolimits_{SH} = E \cdot J \tag{3}$$

Note that this term is determined by (1) and therefore, is it mandatory to solve (1) and (2) simultaneously. Just as for the active devices in the electrical part of the system, we also apply a compact model for the self heating of the devices. For transistor structures, the self heating is determined as a function of the source-drain voltage, the gate-source voltage and the local temperature. Here we assume that to each gate finger fragment we may assign a unique temperature value, which may vary in going from one fragment or finger to another.

$$\sum\nolimits_{SH}^{device} = V_{DS} I_{DS}(V_{DS}, V_{GS}, T) \tag{4}$$

The detailed current–voltage–temperature characteristic (4) is obtained from transmission line pulse (TLP) measurements or TCAD device characterization.

This (almost) completes the definition of our holistic electro-thermal approach. In the present stage we have implemented two kinds of boundary conditions: at metallic contacts we can select electrical voltage boundary conditions or current boundary conditions or a (primitive) circuit may be selected from a build-in library of circuits. The metallic contacts also serve as heat sinks/ sources meaning that fixed

temperature boundary conditions can be selected. Alternatively, one may option for thermal-current boundary conditions. The side walls of the simulation domain are dealt with using Neumann boundary conditions. This corresponds to ideal thermally and electrically insulating walls.

Finally we note that the boundary conditions can be time-dependent. Even when the voltages adapt instantaneously to the time-dependent contact voltage: no charge effects are considered in Eq. (1), we still deal with a transient problem because of the thermal capacitive term in Eq. (2). Thus our solver will be able to explore in a fully self-consistent way the occurrence of thermal runaway. Of course, predictive simulation requires the availability of accurate compact models over a sufficiently wide temperature range.

2.3 Compact Device Representation

Despite the fact that we address the electrical and thermal variables from a field perspective we do not sustain this practice all the way down into the active devices. Doing so, would mean imply that the field solving approach not only must be applied at a much smaller length scale (sub micron) but moreover, new degrees of freedom (the electron and hole Fermi levels) must be considered. The purpose of underlying scheme is not to contribute to progress in process and device technology but to provide an EDA tool suitable optimize designs. Just because the active devices are very limited size, we may replace them by entities with negligible volume. The active devices are then only 'visible' through their contacts to which we may assign compact models.

3 Simulation of Self Heating in Power Switches

In the section we will apply our self-consistent ET solver to a multi-finger power transistor. The simulation confirms our statement made in the introduction that temperature variations appear at a true micro-scale (e.g. at a scale between different transistor fingers). On this power device, a step function with rise time of 1 μs is applied. The voltages and temperatures are calculated self-consistently over a period of 1 ms. Simulation time is 1000 s using a state-of-the art single-core simulation server. In the middle of the active area, the maximum temperature as a function of time is given in Fig. 1. The overall temperature rise can hide local differences, as shown in Fig. 2. In particular, this figure shows that the finger sections give rise to local temperature differences.

The dependence of electric and thermal conductivity on the local temperature is taken into account, by using a power law model, valid over the full temperature range.

$$\kappa = \kappa_0 T^{-a}, \quad \sigma = \sigma_0 T^{-b} \tag{5}$$

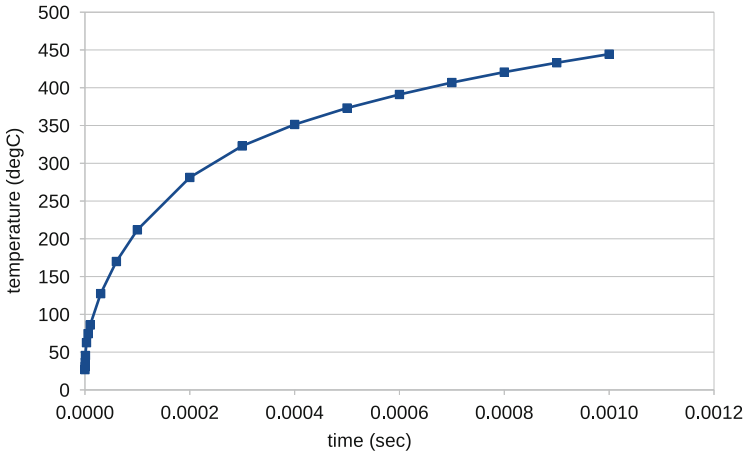


Fig. 1 Time dependence of maximum temperature inside the device

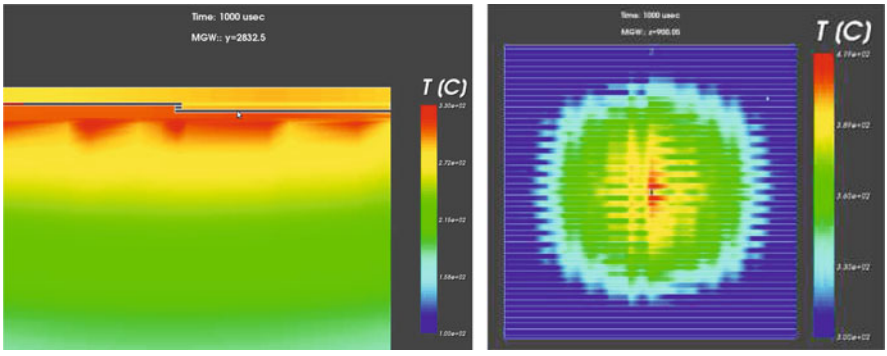


Fig. 2 Cross section, local temperature variations (*left*) and local temperatures in active area (*right*)

The power is dissipated in the different transistor fingers. However we can and have included the dissipation as well as in metalization. The transistor fingers are modeled by using a table model that describes the nonlinear I–V characteristics of the device. The table is created from the dedicated compact model net list model for the device. A successful use of the table is possible provided that it can be used over a sufficient wide range for the input parameters. The reason for this is that the self-consistent solver is using an iterative method for finding the solution and therefore may wander in the search space temporarily into the wrong direction. finding the solution and therefore may wander in the search space temporarily into the wrong direction.

4 Conclusions

The default TCAD [1] approach is not feasible for realistic powerMOS designs containing up to several thousands of device fingers. We presented a new method for dealing with electro-thermal simulation. In this approach all device meshing details that come with computing the junction physics, are avoided. The CPU resources are now fully available for the meshing of the back-end. The self consistency is restricted to merely two degrees of freedom per mesh node. We also demonstrated that self consistency is a necessary aspect for dealing with self heating since local temperature is variable at a very local scale.

Acknowledgements Part of this work is supported by the EU funded project nanoCOPS, FP7 619166. We would like to acknowledge numerous inspiring discussions with Prof. Martin Pfost, University of Reutlingen, Germany.

Reference

1. Burenkov, A., Lorenz, J.: Self-heating effects in nano-scaled MOSFETs and thermal aware compact models. In: THERMINIC 2011 (2011)

The European Project nanoCOPS for Nanoelectronic Coupled Problems Solutions

H.H.J.M. Janssen, P. Benner, K. Bittner, H.-G. Brachtendorf, L. Feng, E.J.W. ter Maten, R. Pulch, W. Schoenmaker, S. Schöps, and C. Tischendorf

Abstract The project nanoCOPS (http://www.fp7-nanocops.eu) is a collaborative research project within the FP7-ICT research program funded by the European Union. The consortium comprises experts in mathematics and electrical engineering from seven universities (BU Wuppertal, HU Berlin, Brno UT, TU Darmstadt, FH OÖ Hagenberg, U Greifswald, KU Leuven), one research institute (MPI Magdeburg), two industrial partners (NXP Semiconductors Netherlands, ON Semiconductor Belgium) and two SMEs (MAGWEL—Belgium, ACCO Semiconductor—France).

We present an overview of the project subjects addressing the "bottlenecks" in the currently-available infrastructure for nanoelectronic design and simulation. In particular, we discuss the issues of an electro-thermal-stress coupled simulation for Power-MOS device design and of simulation approaches for transceiver designs at high carrier frequencies and baseband waveforms such as OFDM (Orthogonal Frequency Division Multiplex).

Keywords Coupled problems • Nanoelectronics

H.H.J.M. Janssen (✉)
NXP Semiconductors, High Tech Campus 46, 5656 AE Eindhoven, The Netherlands
e-mail: Rick.Janssen@nxp.com

P. Benner • K. Bittner • H.-G. Brachtendorf • L. Feng • E.J.W. ter Maten • R. Pulch • W. Schoenmaker • S. Schöps • C. Tischendorf
Max-Planck-Institut, Magdeburg, Germany

University of Applied Sciences Upper Austria, Hagenberg im Mühlkreis, Austria

Bergische Universität Wuppertal, Wuppertal, Germany

Ernst-Moritz-Arndt-Universität Greifswald, Greifswald, Germany

Magwel NV, Leuven, Belgium

Technische Universität Darmstadt, Darmstadt, Germany

Humboldt Universität zu Berlin, Berlin, Germany

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_116

1 Introduction

Designs in nanoelectronics often lead to large-size simulation problems and include strong feedback couplings. Industry demands the provisions of variability to guarantee quality and yield. It also requires the incorporation of higher abstraction levels to allow for system simulation in order to shorten the design cycles, while at the same time preserving accuracy. The nanoCOPS project addresses the simulation of two technically and commercially important problem classes identified by the industrial partners:

- Power-MOS devices, with applications in energy harvesting, that involve couplings between electromagnetics (EM), heat, and stress, and
- RF-circuitry in wireless communication, which involves EM-circuit-heat coupling and multirate behaviour, together with analogue-digital signals.

To meet market demands, the scientific challenges are to:

- create efficient and robust simulation techniques for strongly coupled systems, that exploit the different dynamics of sub-systems and that allow designers to predict reliability and ageing;
- include a variability capability such that robust design and optimization, worst case analysis, and yield estimation with tiny failures are possible (including large deviations like 6-sigma);
- reduce the complexity of the sub-systems while ensuring that the parameters can still be varied and that the reduced models offer higher abstraction models that are efficient to simulate.

Our solutions are

- to develop advanced co-simulation/multirate/monolithic techniques, combined with envelope/wavelet approaches;
- to produce new generalized techniques from Uncertainty Quantification (UQ) for coupled problems, tuned to the statistical demands from manufacturability;
- to develop enhanced, parameterized Model Order Reduction techniques for coupled problems and for UQ.

The best (efficient, robust) algorithms produced are currently being implemented and transferred to SME partner MAGWEL. Validation is conducted on industrial designs provided by the industrial partners. A thorough comparison to measurements on real devices will be made.

2 Coupled Problems, Co-simulation, Multirate

The coupling of various physical effects in nanoelectronics plays an important role in the operational reliability, at both circuits and systems level. This is the case for high-performance applications (CPUs, RF-circuits) as well as applications in

hostile environments (e.g., such as high voltages and/or high currents in automotive applications, RF Power and Base Stations applications). Various types of coupled phenomena exist. For example, electro-thermal coupling is a key concern during operational cycles in industry where a substantial amount of heat is generated that (1) will affect the voltage and current distributions and (2) will indirectly impact the sources of the heat itself. The extent and impacts of electro-thermal-stress coupling is studied in the modelling of power-MOS devices in DC and in the transient regime (time domain), taking environmental aspects like metal stack and package into account. The determination of both reliability and ageing needs to be more effectively addressed by the combined simulation of these coupled effects. Another challenging coupling mechanism concerns Radio Frequency (RF) designs that have to involve with circuit-EM-heat couplings, where parasitic long-range electromagnetic (EM) effects induce substantial distortion at the circuit level, which can lead to the sudden malfunction of the circuit. In order to address both these types of problems, companies need to have a capability for the simulation of multi-physics with dynamics involving different time scales.

Co-simulation techniques are natural approaches in efficiently solving coupled problems. Field-circuit couplings have been considered in [1, 2]. Using *source coupling*, the current $\mathbf{i}$ of an equivalent current source is calculated from the electromagnetic fields and becomes input for the circuit equations . Next, the circuit excites the electromagnetic fields by a time-dependent voltage source. Alternatively, using *inductive coupling*, the current source for the circuit is replaced by a resistor in series with a time-dependent inductor with an inductance that is fitted to the field quantities. This is a more preferred option. The complete problem is now described as follows. The eddy-current field problem on Ω is

$$\sigma \partial_t \mathbf{a}^{(n)} + \nabla \times \left(\nu(|\nabla \times \mathbf{a}^{(n)}|) \, \nabla \times \mathbf{a}^{(n)} \right) = \boldsymbol{\chi} \mathbf{j}^{(n)},$$

where $\mathbf{a}^{(n)}$ is the magnetic vector potential after the n-th iteration (with homogeneous Dirichlet conditions), σ and ν are conductivity and reluctivity, respectively and the winding functions $\boldsymbol{\chi} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_k, \ldots, \boldsymbol{\chi}_K]^\top$ are functions of space that distribute the lumped currents $\mathbf{j}$ in the 3D domain. The circuit coupling is established via integration

$$\partial_t \int_\Omega \boldsymbol{\chi}_k \mathbf{a}^{(n)} \, \mathrm{d}x + R_k j_k^{(n)} = v_k^{(n-1)} \qquad k = 1, \ldots, K$$

to the circuit system of differential algebraic equations

$$\begin{aligned} \mathbf{A}_\mathrm{C} \partial_t \mathbf{q}_\mathrm{C}(\mathbf{A}_\mathrm{C}^\mathrm{T} \mathbf{u}^{(n)}, t) + \mathbf{A}_\mathrm{R} \mathbf{g}_\mathrm{R}(\mathbf{A}_\mathrm{R}^\mathrm{T} \mathbf{u}, t) + \mathbf{A}_\mathrm{L} \mathbf{i}_\mathrm{L}^{(n)} \\ + \mathbf{A}_\mathrm{M} \mathbf{j}^{(n)} + \mathbf{A}_\mathrm{V} \mathbf{i}_\mathrm{V}^{(n)} + \mathbf{A}_\mathrm{I} \mathbf{i}_\mathrm{s}(t) = 0, \\ \partial_t \boldsymbol{\Phi}_\mathrm{L}(\mathbf{i}_\mathrm{L}^{(n)}, t) - \mathbf{A}_\mathrm{L}^{\mathrm{Tu}} = 0, \\ \mathbf{A}_\mathrm{V}^{\mathrm{Tu}} - \mathbf{v}_\mathrm{s}(t) = 0, \end{aligned}$$

with incidence matrices $\mathbf{A}_*$ where $\mathbf{v}_* = \mathbf{A}_*^{\mathrm{T}}\mathbf{u}$ and constitutive laws for conductances, inductances and capacitances (functions with subscripts R, L and C), independent sources $\mathbf{i}_{\mathrm{s}}$ and $\mathbf{v}_{\mathrm{s}}$, unknowns are the potentials $\mathbf{u}$ and currents $\mathbf{i}_{\mathrm{L}}$ and $\mathbf{i}_{\mathrm{V}}$.

Apart from this we deal with a field-mechanical coupling in cavities, a field-thermal coupling and with a thermal-mechanical problem [3]. Dynamic iteration is performed at each time step. In [4], for the field-thermal coupling this is combined with a time-averaging for the heat source, thus exploiting *multirate* difference in the dynamics between the field and the heat quantities.

Multirate time integration for circuit simulation has been studied for circuit decomposition as well as for signals with a broad difference in the frequency domain. When different signal shapes are present in the circuit, these may be approximated more efficiently if individual grids are used for each of the signals. As an example we consider a chain of five frequency dividers (as part of a PLL). In each step the frequency is reduced by a factor 2 as one can see in Fig. 1. Obviously, for the low frequency signals towards the end of the divider chain a much sparser grid is sufficient for an accurate representation, in comparison to the high frequency input signal. In the approach, the problem is cast into a multi-time problem using a slowing varying time scale τ_1 and a second timescale τ_2 for a highly periodic problem. The Rothe method is used for time integration along τ_1. Spline wavelets are used to solve the periodic problems along τ_2. Very efficient discretizations in τ_2 are obtained, that vary with τ_1. From the solution in (τ_1, τ_2)-space, a 1-dimensional solution depending on $(t, \phi(t))$ (for a suitable phase-function ϕ) can be constructed, which provides an envelope solution. Recently, the method has been extended to deal with circuit partitions as well [5, 6]. Currently, one considers coupling with heat as well.

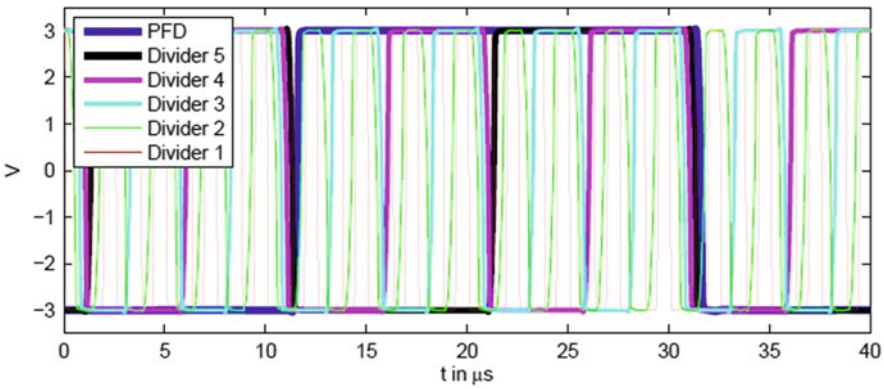


Fig. 1 Several signals in a frequency divider chain as part of a PLL

3 Model Order Reduction, Uncertainty Quantification

In [7] a robust algorithm for *parametrized Model Order Reduction* (pMOR)based on implicit moment matching has been derived, for linear systems based on state-space formulations, which directly applies to circuit equations. In [8] the method has been extended to second order systems coming from electromagnetic field discretizations. Additionally an a posteriori output error bound for reduced order models of micro- and nano-electrical(-mechanical) systems is derived. The error bound is independent of the discretization method (finite difference, finite element, finite volume) applied to the original PDEs. Secondly, the error bound can be directly used in the discretized vector space, without going back to the PDEs, and especially to the bilinear form (weak formulation) associated with the finite element discretization, which must be known a priori for deriving/using the error bound for the reduced basis method. The error bound enables automatic generation of the reduced models computed by parametric model reduction methods based on approximation (interpolation) of the transfer function, e.g., Krylov subspace based methods. Although established for parametrized systems, the error bound is also well-grounded for linear time invariant (LTI) systems without parameters, since it considers the non-parametric LTI systems as a special case [8].

For parameters coming from geometry, the expressions are not always easily obtainable (for instance, meshing for electromagnetic problems is done in the CAD environment) and thus the expansion may be even cumbersome [9]. Here a strategy is to use, for a given parameter, the expression handler in the CAD-environment before starting the simulation to evaluate the p-dependent sub-parts of all expressions (usually also circuit simulators have such an internal step in their expression handler) and then apply a MOR-projection for this parameter.

In [10] MOR for linear, coupled systems was derived based on low-rank approximations of the coupling matrices. When having obtained MOR models for subsystems an interesting application arises for multirate simulation or in use with dynamic iteration and thus provides a link to Sect. 2. The lumped inductor coupling can be seen as a first MOR-model, used for coupling. Dynamic iteration with MOR can be much more robust than iteration with interpolation/extrapolation of values at simple interphases.

In [11, 12] methods for *Uncertainty Quantification* (UQ) via generalized Polynomial Chaos (gPC) expansions have been proposed. These methods can greatly benefit when being combined with methods for pMOR [13]. Assuming that the discretization of the underlying structure of the electromatic problem is fixed, in [14, 15] UQ-results are obtained involving parameterized MOR. For three parameters a full model of ca 30k dofs was compared to ROM of 40 dofs, for different quadrature formulas in Stochastic Collocation. Popular is the so-called Stroud-3 rule [16] to compute the collocation points. One can also use a Hermite

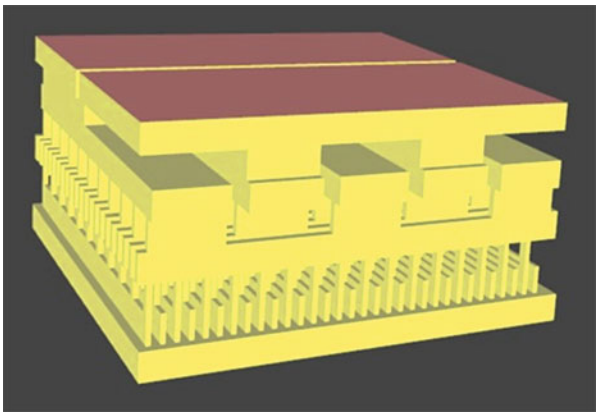


Fig. 2 Typical layout of the power transistor (stretched vertical direction) showing its complex geometry

Genz-Keister [14, 17] sparse grid that yields normally distributed sample points and weights in the quadrature rule. In [18] the sensitivity of the variance with respect to parameters is considered. This gives an indication of dominant parameters, see also [13]. Clearly, MOR should preserve the main statistical properties of the full model.

In [19] stochastically varying domains are considered, leading to topology optimization for a permanent magnet (PM) synchronous machine with material uncertainties. The variations of the non-linear material characteristics are modeled by the gPC method. During the iterative optimization process, the shapes of the rotor poles, represented by zero-level sets, are simultaneously optimized by redistributing the iron and magnet material over the design domain. The gradient directions of the multi-objective function with constraints, composed of the mean and the standard deviation, is evaluated by utilizing the continuous sensitivity equation approach and the Stochastic Collocation Method. Combined with the level set method this yields designs by using already existing deterministic solvers. Finally, a two-dimensional numerical result demonstrates that the proposed method is robust and effective. This example has already a non-trivial geometry. However, there still are a lot steps to be taken. For one of our industrial use cases, a Power-MOS, Fig. 2 shows a complex geometry, for which a lot of coupled effects have to be efficiently determined. Also UQ for large parameter variations is a point of attention [20].

Acknowledgements We acknowledge the support from the project nanoCOPS, Nanoelectronic COupled Problems Solutions (FP7-ICT-2013-11/619166), http://www.fp7-nanoCOPS.eu/.

References

1. Bartel, A., Brunk, M., Günther, M., Schöps, S.: Dynamic iteration for coupled problems of electric circuits and distributed devices. SIAM J. Sci. Comput. **35**(2), B315–B335 (2013)
2. Tischendorf, C., Schoenmaker, W., De Smedt, B., Meuris, P., Baumanns, S., Matthes, M., Jansen, L., Strohm, C.: Dynamic coupled electromagnetic field circuit simulation. Presented in Minisymposium on "Simulation Issues for Nanoelectronic Coupled Problems" at ECMI-2014, 18th European Conference on Mathematics for Industry, Taormina, Sicily, June 11, 2014
3. Schöps, S.: Iterative Schemes for Coupled Multiphysical Problems in Electrical Engineering. Invited talk SCEE-2014 (Scientific Computing in Electrical Engineering), Wuppertal, 2014. IMACM Report 2014-28, pp. 11–12, Bergische Universität Wuppertal. http://www.imacm.uni-wuppertal.de/fileadmin/imacm/preprints/2014/imacm_14_28.pdf (2014)
4. Kaufmann, C., Günther, M., Klagges, D., Knorrenschild, M., Richwin, M., Schöps, S., ter Maten, E.J.W.: Efficient frequency-transient co-simulation of coupled heat-electromagnetic problems. J. Math. Ind. **4**, 1 (2014). http://www.mathematicsinindustry.com/content/4/1/1
5. Bittner, K., Brachtendorf, H.-G.: Adaptive multi-rate wavelet method for circuit simulation. Radioengineering **23**(1), 300–307 (2014). http://www.radioeng.cz/fulltexts/2014/14_01_0300_0307.pdf
6. Bittner, K., Brachtendorf, H.-G.: Fast algorithms for adaptive free knot spline approximation using nonuniform biorthogonal spline wavelets. SIAM J. Sci. Comput. **37**(2), B283–B304 (2015)
7. Benner, P., Feng, L.: A robust algorithm for parametric model order reduction based on implicit moment matching. In: A. Quarteroni, G. Rozza (eds.) Reduced Order Methods for Modeling and Computational Reduction. MS &A Series, vol. 9, pp. 159–186. Springer, Cham (2014)
8. Feng, L., Benner, P., Antoulas, A.C.: An a posteriori error bound for reduced order modeling of micro- and nano-electrical(-mechanical) systems. Presented at SCEE-2014 (Scientific Computing in Electrical Engineering), Wuppertal, 2014. IMACM Report 2014–28, pp. 99–100, Bergische Universität Wuppertal. http://www.imacm.uni-wuppertal.de/fileadmin/imacm/preprints/2014/imacm_14_28.pdf (2014)
9. Stavrakakis, K.K.: Model order reduction methods for parameterized systems in electromagnetic field simulations. PhD thesis, TU-Darmstadt (2012)
10. Lutowska, A.: Model order reduction for coupled systems using low-rank approximations. PhD thesis, TU Eindhoven (2012). http://alexandria.tue.nl/extra2/729804.pdf
11. Le Maître, O.P., Knio, O.M.: Spectral Methods for Uncertainty Quantification, with Applications to Computational Fluid Dynamics. Springer, Science+Business Media B.V., Dordrecht (2010)
12. Xiu, D.: Numerical Methods for Stochastic Computations - A Spectral Method Approach. Princeton University Press, Princeton, NJ (2010)
13. ter Maten, E.J.W., Pulch, R., Schilders, W.H.A., Janssen, H.H.J.M.: Efficient calculation of uncertainty quantification. In: Fontes, M., Günther, M., Marheineke, N. (eds.) Progress in Industrial Mathematics at ECMI 2012. Series Mathematics in Industry, vol. 19, pp. 361–370. Springer, Cham (2014)
14. Benner, P., Schneider, J.: Uncertainty quantification using reduced-order Maxwell's equations. Presented at SCEE-2014 (Scientific Computing in Electrical Engineering), Wuppertal. http://fp7-nanocops.eu/ (2014)
15. Bodendiek, A., Bollhöfer, M.: Adaptive-order rational Arnoldi-type methods in computational electromagnetism. BIT Numer. Math. **15**(2), 1–24 (2013)
16. Stroud, A.H.: Remarks on the disposition of points in numerical integration formulas. Math. Tables Other Aids Comput. **11**(60), 257–261 (1957)
17. Genz, A., Keister, B.D.: Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight. J. Comput. Appl. Math. **71**, 299–309 (1996)

18. Pulch, R., ter Maten, E.J.W., Augustin, F.: Sensitivity analysis and model order reduction for random linear dynamical systems. Math. Comput. Simul. **111**, 80–95 (2015)
19. Putek, P., Gausling, K., Bartel, A., Gawrylczyk, K.M., ter Maten, E.J.W., Pulch, R., Günther, M.: Robust topology optimization of a permanent magnet synchronous machine using multilevel set and stochastic collocation methods. In: Bartel, A., Clemens, M., Günther, M., ter Maten, E.J.W. (eds.) Scientific Computing in Electrical Engineering SCEE 2014. Mathematics in Industry, vol. 23, pp. 233–242. Springer, Berlin (2016)
20. Di Buccianico, A., ter Maten, J., Pulch, R., Janssen, R., Niehof, J., Hanssen, M., Kapora, S.: Robust and efficient uncertainty quantification and validation of RFIC isolation. Radioengineering **23**(1), 308–318 http://www.radioeng.cz/fulltexts/2014/14_01_0308_0318.pdf (2014)

MS 34
MINISYMPOSIUM: SIMULATION, MODEL ORDER REDUCTION AND ROBUST OPTIMIZATION FOR INDUSTRIAL E-MOBILITY APPLICATIONS

Organizers

Sebastian Schöps[1] and Andreas Bartel[2]

Speakers

Alessandro Alla[3] and Michael Hinze[4]
HJB-POD Feedback Control for Navier-Stokes Equations

Zeger Bontinck[5], Herbert De Gersem[6] and Sebastian Schöps[1]
Uncertainty Quantification of Geometric and Material Properties of Permanent Magnetic Synchronous Machines

Kai Gausling[7] and Andreas Bartel[2]
Analysis of the Contraction-Condition in the Co-simulation of a Specific Electric Circuit

[1]Sebastian Schöps, Technische Universität Darmstadt, Darmstadt, Germany.
[2]Andreas Bartel, Bergische Universität Wuppertal, Wuppertal, Germany.
[3]Alessandro Alla, Universität Hamburg, Hamburg, Germany.
[4]Michael Hinze, Universität Hamburg, Hamburg, Germany.
[5]Zeger Bontinck, Technische Universität Darmstadt, Darmstadt, Germany.
[6]Hebert De Gersem, Technische Universität Darmstadt, Darmstadt, Germany.
[7]Kai Gausling, Bergische Universität Wuppertal, Wuppertal, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_117

Oliver Lass[8]
Parameter Identification for Nonlinear Elliptic-Parabolic Systems with Application in Lithium-Ion Battery Modeling

Keywords

Coupled systems
Magnetoquasistatics
Model order reduction
Multiphysics
Optimization
PDAEs
Uncertainty quantification

Short Description

This minisymposium addresses mathematical problems and new methods for the robust design of key components in E-mobility. These key components are for instance: electrical machines, batteries etc. In the overall, multi-physical phenomena have to be modeled. This yields large systems of coupled partial differential algebraic equations. Enhancing these devices, one will also need to further approach their limits. Therefore, the uncertainties of parameters need to be included. In the end, these systems have to be solved accurately and repeatedly (for uncertainty or for optimization). Thus besides modeling, efficient techniques for multiphysical simulation, optimization using reduced order models and uncertainty computations are in our focus. Mathematicians and electrical engineers report about their joint work within the project SIMUROM funded by the German Federal Ministry of Education and Research in the framework 'Mathematics for Innovation in Industry and Services'.

[8]Oliver Lass, Technische Universität Darmstadt, Darmstadt, Germany.

A Meshfree Method for Simulations of Dynamic Wetting

Sudarshan Tiwari, Axel Klar, and Steffen Hardt

Abstract In this paper we present a meshfree Lagrangian particle method for the simulation of dynamic wetting phenomena. The essence of dynamic wetting is that the contact angle between the interface of the immiscible fluids and the solid surface is a dynamic quantity. The dynamic contact angle is modeled as a boundary condition. The two-phase immiscible flow is described by the incompressible Navier-Stokes equations in combination with the continuous surface tension force model. The phases are distinguished by assigning colors to the particles, and the normal vector and curvature of the interface are computed from this color function. Chorin's pressure projection method is used to solve the model equations in a meshfree framework. A two-phase Couette flow is considered, with a capillary bridge spanning the distance between the two walls. The details of the numerical methods can be found in Tiwari and Kuhnert (J Comput Appl Math 203:376–386, 2007), Tiwari et al. (Numerical simulation of wetting phenomena by a meshfree particle method. J Comput Appl Math 292:469–485, 2016. It is shown that the numerical results reproduce the employed empirical law for the dynamic contact angle.

Keywords Dynamic wetting • Meshfree method

1 Introduction

Wetting phenomena play a crucial role in many fields, for example in the area of microfluidics [8, 12], coating technology [14], or oil recovery [1]. The term "wetting" means that an intersection of the solid surface with the interface between the two immiscible fluids exists. This intersecting line is called the three-phase

S. Tiwari (✉) • A. Klar
Department of Mathematics, University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern, Germany
e-mail: tiwari@mathematik.uni-kl.de; klar@mathematik.uni-kl.de

S. Hardt
Center of Smart Interfaces, TU Darmstadt, Alarich-Weiss-Str. 10, 64287 Darmstadt, Germany
e-mail: hardt@csi.tu-darmstadt.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_118

contact line or wetting line. The angle between this interface and the solid surface is called the contact angle. In the case of dynamic wetting the wetting line is a dynamic entity, i.e. it moves over the solid surface. The general wetting behavior of a liquid is determined by the static contact angle θ_s. A wetting liquid has a static contact angle of less than 90°, a non-wetting liquid has a static contact angle larger than 90°.

If external forces are applied, the contact line is usually set into motion, moving with a velocity denoted by U_{cl}. The corresponding dynamic contact angle is denoted by θ_d. It depends on the static contact angle and on the capillary number $Ca = \frac{\mu U_{cl}}{\sigma}$, where μ is the dynamic viscosity of the spreading liquid, and σ is the surface tension coefficient.

Several theoretical and numerical analyses have been reported addressing dynamic contact angles, see [3, 7, 9, 13] and references therein. Widely used numerical methods are based on the classical finite difference or finite volumes approach. In such methods the representation moving interfaces is not straightforward. By contrast, Lagrangian type of methods describe the advection of fluid-fluid interfaces in a natural way. Corresponding reports are available ranging from molecular dynamics [13] to macroscale SPH particle methods [6].

In our earlier work [11] we have presented simulations of wetting phenomena based on the static contact angle. In the current report we have extended the model to account for contact angle dynamics. We use a meshfree Lagrangian particle method which has a similar character as the method of smoothed particle hydrodynamics (SPH) [4], except for the approximation of the spatial derivatives and the treatment of the boundary conditions [10]. We approximate the spatial derivatives with the help of the weighted least squares method, where the smoothing kernel in SPH and the weight in our scheme play similar roles.

2 Mathematical Model

We consider two immiscible fluids, for example liquid and gas, where both of them are incompressible. The two-phase flow is modeled by the incompressible Navier-Stokes equations. The equations are expressed in the Lagrangian form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{v} \tag{1}$$

$$\nabla \cdot \mathbf{v} = 0 \tag{2}$$

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\frac{1}{\rho}\nabla p + \frac{1}{\rho}\nabla \cdot (2\mu D) + \mathbf{g} + \frac{1}{\rho}\mathbf{F}_S, \tag{3}$$

where $\mathbf{v}$ is the fluid velocity vector, ρ is the density, μ is the dynamic viscosity, D is the viscous stress tensor $D = \frac{1}{2}(\nabla \mathbf{v} + \nabla^T \mathbf{v})$, $\mathbf{g}$ is the gravitational acceleration, and $\mathbf{F}_S$ is the surface tension force. In general, ρ and μ are discontinuous across the interface and remain constant in each phase. The surface tension force $\mathbf{F}_S$ is

computed using the classical continuum surface force (CSF) model [2]. It acts in the vicinity of the interface between the fluids. In the CSF model the surface tension force $\mathbf{F}_S$ is defined by

$$\mathbf{F}_S = \sigma\kappa\mathbf{n}_\mathbf{I}\delta_S, \tag{4}$$

where σ is the surface tension coefficient, assumed to be a constant, κ is the curvature of the interface, $\mathbf{n}_I$ is the unit normal vector of the interface and δ_S is a smeared delta function, peaked at the interface.

Equations (1)–(3) are solved with initial and boundary conditions. For this purpose a pressure projection method in a meshfree framework is used. The meshfree scheme has been reported in our earlier papers, see, for example, [10].

2.1 Computation of the Surface Tension Force

Particle methods are suitable to compute the surface tension force and surface tension driven flows, see [10, 11]. The interface can be accurately determined by assigning colors or flags to the particle of each phase. For example, we define the color $c = 1$ for the gas and $c = 2$ for the liquid. The normal vector $\mathbf{n}_I$ at the interface is computed via the gradient of the color function c. Since c is discontinuous across the interface, one has to smooth it. Let $\mathbf{x}$ be the position of an arbitrary particle that has neighbours with function values $c_j = c(\mathbf{x}_j)$. We smooth c at $\mathbf{x}$ from its neighbors with the help of the Shepard interpolation rule given by

$$\tilde{c}(\mathbf{x}) = \frac{\sum_{j=1}^{m} w_j c_j}{\sum_{j=1}^{m} w_j}, \tag{5}$$

where $\mathbf{x}$ is an arbitrary particle position, m is the number of neighbors inside the interaction radius h and w_j is the weight function given by

$$w_j = w(\mathbf{x}_j - \mathbf{x}; h) = \begin{cases} \exp(-\alpha \frac{\|\mathbf{x}_j - \mathbf{x}\|^2}{h^2}), & \text{if } \|\mathbf{x}_i - \mathbf{x}\| \le h \\ 0, & \text{else,} \end{cases} \tag{6}$$

where α is a positive constant. We observe that the gradient of $\tilde{c}$ is non-vanishing only in a region close to the interface. First we compute the unit normal vectors and then compute the curvature by

$$\mathbf{n}_I = \frac{\nabla\tilde{c}}{|\nabla\tilde{c}|}, \qquad \kappa = -\nabla \cdot \mathbf{n}_I. \tag{7}$$

There exist many possible choices for δ_s, but in practice, it is often approximated as $\delta_s \approx |\nabla \check{c}|$. We note that δ_s is non-zero in the vicinity of the interface and zero far from it.

2.2 Boundary Conditions: Dynamic Contact Angle

In the following we consider a two-dimensional problem. The implementation of the contact angle boundary condition is based on the method suggested in [2], where we apply the boundary condition to the unit normal $\mathbf{n}_I$ before computing the curvature from (7). Therefore, the contact angle boundary condition is applied by redefining the interface normals $\mathbf{n}_I$ at the three-phase contact point $\mathbf{x}_w$ and its nearest neighbors within a radius βh $(0 < \beta < 1)$ as

$$\hat{\mathbf{n}}_I = \mathbf{n} \cos\theta_d + \mathbf{n}_{||} \sin\theta_d, \tag{8}$$

where $\mathbf{n}_{||}$ is the unit vector parallel to the wall normal to the three-phase contact line and $\mathbf{n}$ is the outward unit vector normal to the wall at $\mathbf{x}_w$.

There exist several theoretical/empirical models for the dynamic contact angle, see [9] for a comprehensive overview. In this work we use the empirical model suggested by Hoffman [5]. This model captures the general behavior of the contact angle in the entire range $0° < \theta_d < 180°$. Using this model we obtain

$$\theta_d = f_{Hoff}\left(Ca + f_{Hoff}^{-1}(\theta_s)\right), \tag{9}$$

where f_{Hoff} is the Hoffman function given by

$$f_{Hoff}(x) = \arccos\left[1 - 2\tanh\left\{5.16\left(\frac{x}{1 + 1.31x^{0.99}}\right)^{0.706}\right\}\right]. \tag{10}$$

3 Numerical Results

We consider a two-phase Couette flow in a channel of size $[0, 2]\text{m}\times[0, 2]\text{m}$. We define the particles of fluid 1 to initially lie in the rectangle $[0.6, 1.4]\text{m} \times [0, 0.5]\text{m}$, while fluid 2 occupies the rest of the domain. In Fig. 1 the blue (dark grey) particles represent fluid 1, the red (light grey) particles represent fluid 2. The initial number of particles is around 13,500. This number remains approximately constant throughout the simulations, while removing very close particles and adding particles if necessary. Both fluids have the same density $\rho = 1\,\text{kg/m}^3$, viscosity $\mu = 0.1$ kg/ms, and the surface tension is given by $\sigma = 0.2\,\text{N/m}$. For the static contact angle $\theta_s = 90°$ was assumed. Gravitational forces were not taken into

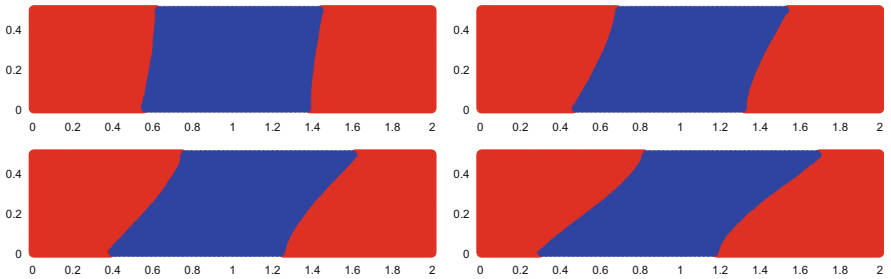


Fig. 1 Particle distribution of both phases for different capillary numbers at time $t = 1.2$ s. *Top left* for $Ca = 0.02$, *top right* for $Ca = 0.06$, *bottom left* for $Ca = 0.1$, and *bottom right* for $Ca = 0.14$

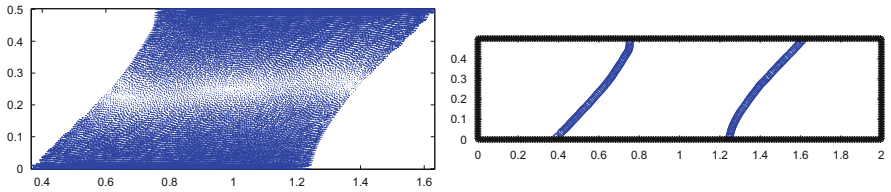


Fig. 2 Velocity field (*left*) and interface points (*right*) of phase 2 for a capillary number Ca = 0.1 at time $t = 1.2$ s

account. Periodic boundary conditions are used at the left and right boundaries of the domain. The top and bottom walls move with opposite velocities of the same magnitude. We fix μ and σ and change the wall velocity U_{cl} to vary the capillary number. Effectively, for such liquid-liquid flows the correlation for the dynamic contact angle may be different from Eq. (10) which was formulated for gas-liquid flows. However, one purpose of this paper is to demonstrate that our numerical scheme also allows implementing a given model for the dynamic contact angle for the more complex situation of liquid-liquid flow.

In Fig. 1 we have plotted the particle distributions for $Ca = 0.02, 0.06, 0.1$ and 0.14 at time $t = 1.2$ s. Moreover, we have plotted the velocity field for $Ca = 0.1$ in Fig. 2 at the same instant in time. We observe the results match qualitatively with the ones obtained from MD simulations [13] and SPH simulations [6].

As a more quantitative check, we evaluate the apparent angle θ_d obtained from the simulations for the case that the fluid-fluid interface advances along the wall. The interface particles are plotted in Fig. 2. We compute the tangent to the interface by considering particles between $y = 0$ to 0.1 for all capillary numbers at time $t = 1.2$s. At this time the apparent angle does no longer change as a function of time, and the contact line moves with the wall velocity. In Fig. 3 we have plotted the

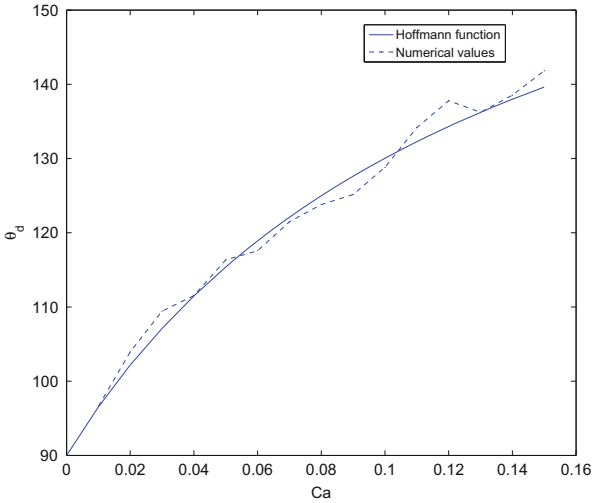


Fig. 3 Dynamic contact angle vs. capillary number. The *solid line* represents the Hoffman function (10), the *dotted lines* represents the numerically obtained angle

Hoffman function (10) together with the numerical approximation of the dynamic contact angle vs. the capillary number. We observe a good agreement between both data sets.

Acknowledgements This work is partially supported by the German research foundation (DFG) grant number KL 1105/201. We would like to thank the DFG for the financial support.

References

1. Babadagli, T.: Dynamics of capillary imbibition when surfactant, polymer, and hot water are used as aqueous phase for oil recovery. J. Colloid Interface Sci. **246**, 203–213 (2002)
2. Brackbill, J.U., Kothe, D.B., Zemach, C.: A continuum method for modeling surface tension. J. Comput. Phys. **100**, 355–354 (1992)
3. Cox, R.G.: The dynamics of the spreading of the liquids on a solid surface, Part I: viscous flow. J. Fluid Mech. **168**, 169–194 (1986)
4. Gingold, R.A., Monaghan, J.J.: Smoothed particle hydrodynamics: theory and application to non-spherical stars. Mon. Not. R. Astron. Soc. **181**, 375–389 (1997)
5. Hoffman, R.: A study of the advancing interface I. Interface shape in liquid-gas systems. J. Colloid Interface Sci. **50**(2), 228–235 (1975)
6. Hu, X.Y., Adams, N.A.: A multi-phase SPH method for macroscopic and mesoscopic flows. J. Comput. Phys. **213**, 844–861 (2006)
7. Schönfeld, S., Hardt, S.: Dynamic contact angles in CFD simulations. Comput. Fluids **38**, 757–764 (2009)
8. Seemann, R., Brinkmann, M., Pfohl, T., Herminghaus, S.: Droplet based microfluidics. Rep. Prog. Phys. **75**, 016601 (2012)

9. Shikhmurzaev, Y.D.: The moving contact lines in liquid/liquid/solid systems. J. Fluid Mech. **334**, 211–249 (1997)
10. Tiwari, S., Kuhnert, J.: Modeling of two phase flows with surface tension by finite pointset method (FPM). J. Comput. Appl. Math. **203**, 376–386 (2007)
11. Tiwari, S., Klar, A., Hardt, S.: Numerical simulation of wetting phenomena by a meshfree particle method. J. Comput. Appl. Math. **292**, 469–485 (2016)
12. Theberge, A.B., Courtois, F., Schaerli, Y., Fischlechner, M., Abell, C., Hollfelder, F., Huck, W.T.S.: Microdroplets in microfluidics: an evolving platform for discoveries in chemistry and biology. Angew. Chem. Int. Ed. **49**, 5846–5868 (2010)
13. Thompson, P.A., Robbins, M.O.: Simulations of contact-line motion: slip and the dynamic contact angle. Phys. Rev. Lett. **63**(7), 766–769 (1989)
14. Weinstein, S.J., Ruschak, K.J.: Coating flows. Annu. Rev. Fluid Mech. **36**, 2953 (2004)

Analysis of the Contraction Condition in the Co-simulation of a Specific Electric Circuit

Kai Gausling and Andreas Bartel

Abstract The convergence for a co-simulation method is commonly based on an error recursion. Usually the contraction condition itself is obtained by some estimations (standard theory). This paper takes a closer look at the coupling structure of co-simulation model for a simple electric circuit. It is shown that with standard theory for our example no contraction could be inferred. However, co-simulation converges. By a detailed analysis, we can prove convergence in this case.

Keywords Co-simulation method • Electric circuit

1 Introduction

Co-simulation is an important method for coupled systems in time domain. In particular, if the monolithic description of a dynamic system is not feasible and/or dedicated simulation tools for the subsystems are available, then it is a relevant option. In practice co-simulation is frequently applied to electrical circuits. Seminal approaches in this field were already specified in [6]. Furthermore this simulation methodology is capable of multirate, multimethod, multiorder (and so on). However, convergence can only be achieved by solving multiple times the subsystems. To enhance convergence, the whole simulation time is split into time windows. Co-simulation applied to coupled ordinary differential equations (ODEs) always convergences [4]. This is not the case for coupled differential-algebraic equations (DAEs). There convergence can only be guaranteed if a contraction condition is fulfilled, see e.g. [1]. It can be shown that the convergence and stability of co-simulation is directly influenced by the order of computation and by the coupling interface, see e.g. [3].

In fact, a co-simulation computes the solutions of the coupled subsystems separately on windows $[T_n, T_n + H]$. We follow the Gauss-Seidel approach. Let

K. Gausling (✉) • A. Bartel
Lehrstuhl für Angewandte Mathematik/Numerische Analysis, Bergische Universität Wuppertal, 42119 Wuppertal, Germany
e-mail: gausling@math.uni-wuppertal.de; bartel@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_119

(k) denote the current iteration, also old iterates $(k-1)$ are involved. Such a co-simulation scheme can be encoded by splitting functions $\mathbf{F}, \mathbf{G}$:

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{f}(\mathbf{y}, \mathbf{z}) & \leftrightarrow \dot{\tilde{\mathbf{y}}} &= \mathbf{F}\left(\tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{z}}^{(k)}, \tilde{\mathbf{y}}^{(k-1)}, \tilde{\mathbf{z}}^{(k-1)}\right) \\ 0 &= \mathbf{g}(\mathbf{y}, \mathbf{z}) & 0 &= \mathbf{G}\left(\tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{z}}^{(k)}, \tilde{\mathbf{y}}^{(k-1)}, \tilde{\mathbf{z}}^{(k-1)}\right) \end{aligned}$$

Then the contraction condition reads:

$$\alpha := \|\mathbf{G}_{\tilde{z}^{(k)}}^{-1} \mathbf{G}_{\tilde{z}^{(k-1)}}\|_2 < 1, \tag{1}$$

where $\mathbf{G}_{\mathbf{z}^{(k)}}$, $\mathbf{G}_{\mathbf{z}^{(k-1)}}$ denote partials Jacobians of $\mathbf{G}$, see e.g. [1, 2].

Our paper is outlined as follows: We consider a linear test system, where the standard contraction condition (1) is not fulfilled. In a numerical treatment, we observe convergence. Then convergence for this test case is proven by an exact fine structure analysis. Finally, we discuss the connection of both types analysis.

2 Circuit Modeling and Test Circuit

Classically, a mathematical model for an electric network can be obtained via modified nodal analysis, see e.g. [5]. This gives a DAE:

$$\mathbf{E}\dot{\mathbf{x}} + \mathbf{A}\mathbf{x} = \mathbf{f}(t),$$

where $\mathbf{E}$ contains the dynamic components, $\mathbf{A}$ static components and $\mathbf{f}$ time depended sources. The unknowns $\mathbf{x}$ are the node voltages and some branch currents.

We investigate the simple RL circuits depicted in Fig. 1. Modified nodal analysis yields an index-1 DAE. By applying the strategy of source coupling (see e.g. [2]), we can model this circuit as two coupled networks as given in Fig. 2. This example serves as our test case for co-simulation. Notice that the monolithic circuit (Fig. 1) is almost the same as the subsystem 2 (Fig. 2). This is due to the fact that we aimed at

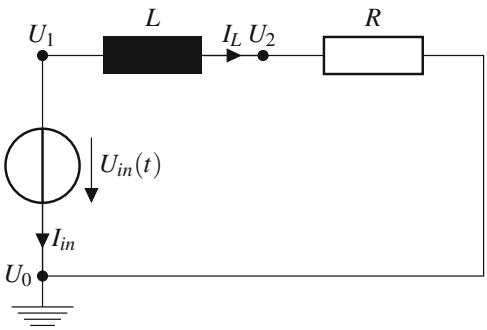


Fig. 1 RL circuit applied by supply voltage $U_{in}(t)$ (reference model)

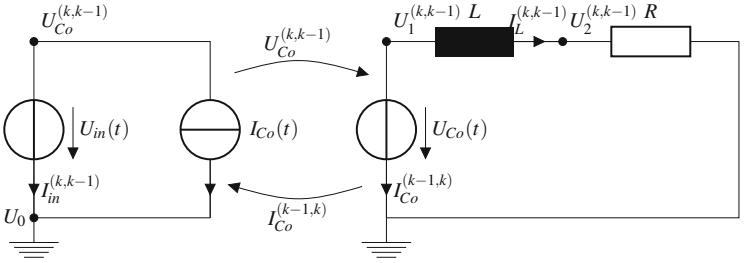


Fig. 2 Decoupled RL network using source-coupling in a co-simulation of Gauss-Seidel type. The first/second notation index denotes the old and new differential and algebraic variables for subsystem 1/subsystem 2 first

an example as simple as possible. This makes our model rather academic, however it shows the divergence between analysis and application of co-simulation, which we want to highlight. Now, the two subsystems for our co-simulation read:

$$\textbf{Subsystem 1:}\quad 0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} U_{Co} \\ I_{in} \end{pmatrix} - \begin{pmatrix} -I_{Co}(t) \\ -U_{in}(t) \end{pmatrix},$$

$$\textbf{Subsystem 2:}\quad 0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & L \end{pmatrix} \begin{pmatrix} \dot{U}_1 \\ \dot{U}_2 \\ \dot{I}_{Co} \\ \dot{I}_L \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & G & 0 & -1 \\ -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ I_{Co} \\ I_L \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ -U_{Co}(t) \\ 0 \end{pmatrix}, \tag{2}$$

with inductance L, conductance $G = 1/R$, given voltage source $U_{in} = U_{in}(t)$, unknown node potentials $U_1,\ U_2,\ U_{Co}$ and unknown currents $I_{in},\ I_L,\ I_{Co}$. U_{Co} and I_{Co} are additional variables needed for the source coupling. The application of a Gauss-Seidel type of co-simulation demands to choose a system, which is computed first.

3 Standard Abstract Co-simulation Analysis

Next we use standard theory [1, 2] to analyze the coupled system (2). To this end, we generalize the system (2) to the following semi-explicit form:

$$0 = \mathbf{g}_1(\mathbf{z}_1, \mathbf{z}_2), \qquad \dot{\mathbf{y}}_2 = \mathbf{f}_2(\mathbf{y}_2, \mathbf{z}_2), \quad 0 = \mathbf{g}_2(\mathbf{z}_1, \mathbf{y}_2, \mathbf{z}_2), \tag{3}$$

where subsystem 1 (subindex '1') is merely a system of linear equations and subsystem 2 is a DAE. The variables of the subsystems are

$$\mathbf{z}_1 := \left[U_{Co},\ I_{in}\right]^T, \quad \mathbf{y}_2 := I_L, \quad \mathbf{z}_2 := \left[U_1,\ U_2,\ I_{Co}\right]^T.$$

Since $\partial \mathbf{g}_i / \partial \mathbf{z}_i$ are not singular in (2), the subsystems and the overall system are index-1. Thus $\mathbf{y_i}$ defines the differential and $\mathbf{z_i}$ the algebraic components. Notice, the semi-explicit form (3) encodes, which type of variables occur in the submodels.

Now, we start co-simulation with the subsystem 1 first and obtain the corresponding splitting functions:

$$\begin{aligned} \mathbf{F}(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}, \mathbf{z}^{(k-1)}) &:= \left[\mathbf{f}_2(0, 0, \mathbf{y}_2^{(k)}, \mathbf{z}_2^{(k)})\right], \\ \mathbf{G}(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}, \mathbf{z}^{(k-1)}) &:= \begin{bmatrix} \mathbf{g}_1(0, \mathbf{z}_1^{(k)}, 0, \mathbf{z}_2^{(k-1)}) \\ \mathbf{g}_2(0, \mathbf{z}_1^{(k)}, \mathbf{y}_2^{(k)}, \mathbf{z}_2^{(k)}) \end{bmatrix}. \end{aligned} \tag{4}$$

Notice the old algebraic iterate $\mathbf{z}_2^{(k-1)}$ ($I_{Co}^{(k-1)}$) enters algebraic equations. The reversed computational order gives us the splitting functions (subsystem 2 first):

$$\begin{aligned} \mathbf{F}(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}, \mathbf{z}^{(k-1)}) &:= \left[\mathbf{f}_2(0, 0, \mathbf{y}_2^{(k)}, \mathbf{z}_2^{(k)})\right], \\ \mathbf{G}(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}, \mathbf{z}^{(k-1)}) &:= \begin{bmatrix} \mathbf{g}_1(0, \mathbf{z}_1^{(k)}, 0, \mathbf{z}_2^{(k)}) \\ \mathbf{g}_2(0, \mathbf{z}_1^{(k-1)}, \mathbf{y}_2^{(k)}, \mathbf{z}_2^{(k)}) \end{bmatrix}. \end{aligned} \tag{5}$$

Also here depends an algebraic constraint on old algebraic iterates (subsystem 2 depends on $\mathbf{z}_1^{(k-1)}$, i.e., $U_{Co}^{(k-1)}$). Thus the contraction factor α does not vanish for both pairs of splitting functions (4) and (5). Consequently, stability and contraction cannot be guaranty without previously estimated contraction factor α. Therefore we calculate the matrices $\mathbf{G}_{z^{(k)}}^{-1}, \mathbf{G}_{z^{(k-1)}}$ needed in (1). Splitting the Jacobian of $\mathbf{G}(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}, \mathbf{z}^{(k-1)})$ into parts of $\mathbf{G}_{\mathbf{y}^{(k)}}, \mathbf{G}_{\mathbf{y}^{(k-1)}}, \mathbf{G}_{\mathbf{z}^{(k)}}$ and $\mathbf{G}_{\mathbf{z}^{(k-1)}}$, we obtain:

$$\textbf{Subsystem 1 first: } \mathbf{G}_{\mathbf{z}^{(k)}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & G & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \end{pmatrix} \Rightarrow \mathbf{G}_{\mathbf{z}^{(k)}}^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & R & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \ \mathbf{G}_{\mathbf{z}^{(k-1)}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\textbf{Subsystem 2 first: } \mathbf{G}_{\mathbf{z}^{(k)}} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & G & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \mathbf{G}_{\mathbf{z}^{(k)}}^{-1} = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & R & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \end{pmatrix}, \ \mathbf{G}_{\mathbf{z}^{(k-1)}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Thus we obtain for the contraction conditions for both splitting schemes:

$$\begin{aligned} &\textbf{Subsystem 1 first}: \quad \|\mathbf{G}_{\mathbf{z}^{(k)}}^{-1}\mathbf{G}_{\mathbf{z}^{(k-1)}}\|_2 = \|\begin{pmatrix}0&0&0&0&1\end{pmatrix}^T\|_2 = 1, \\ &\textbf{Subsystem 2 first}: \quad \|\mathbf{G}_{\mathbf{z}^{(k)}}^{-1}\mathbf{G}_{\mathbf{z}^{(k-1)}}\|_2 = \|\begin{pmatrix}-1&0&0&0&0\end{pmatrix}^T\|_2 = 1, \end{aligned} \tag{6}$$

i.e., stability and contraction cannot be inferred for our co-simulation model directly by using standard theory. Notice that standard theory gives only a rough inside into the co-simulation.

4 Numerical Results

Now we analyze the above RL circuit numerically using MATLAB®.[1] For this purpose, we employ the following parameters: resistance $R = 10\,\mathrm{k\Omega}$, inductance $L = 1\,\mathrm{mH}$, and capacitance $C = 1\,\mathrm{nF}$. The circuit is operated by a supply voltage $U_{in}(t) = 1\mathrm{V} \cdot \cos(\omega t)$ with an angular frequency $\omega = 2\pi \cdot 5 \cdot 10^3\,\mathrm{Hz}$.

To investigate contraction and convergency, a co-simulation is studied in one time window $[t_0, t_0 + H]$ with $t_0 = 0.4\,\mathrm{ms}$ and time window size $H = 10^{-4}$ s. The accuracy of the solutions on the n-th time window after k iterations $\tilde{\mathbf{X}}^{(k)}(t)$ is measured by comparing with a reference solution $\mathbf{X}_m(t)$ computed by a monolithic simulation: $\Delta_n^{(k)}(t) = \mathbf{X}_m(t) - \tilde{\mathbf{X}}_c^{(k)}(t)$, $\delta_n^{(k)} := \|\Delta_n^{(k)}\|_2$. For both splitting schemes (4) and (5), a constant extrapolation of the initial value is employed for the initial guess $\tilde{\mathbf{X}}^{(0)}(t)$ on time window H is used. This is the most common choice for an initial guess.

Figure 3 shows convergence and contraction for both splitting schemes (4) and (5). Thus we have convergence even so the estimate (6) does not indicate this behavior. Additionally, we observe two different convergency orders. For subsystem 1 first, we get order $\mathcal{O}(H)$, whereas for subsystem 2 first $\mathcal{O}(H^2)$ is achieved. This can be explained as follows: Constant extrapolation produces an error of $\mathcal{O}(H)$. For subsystem 1 first, the coupling parameter I_{Co} is constantly extrapolated. Since system 1 is just an algebraic equation, their is no improvement during time integration. For subsystem 2 first U_{Co} is constantly extrapolated. This parameter is coupled to the algebraic unknown U_1. However, subsystem 2 has a dynamic element, which is defined by the coupling current. This current is improved during time integration.

[1] Version: MATLAB R2013a, http://www.mathworks.de.

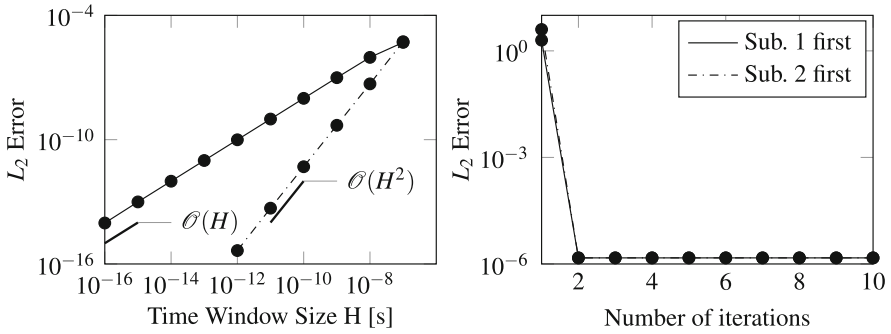


Fig. 3 Convergence and contraction of co-simulation applied to the test circuit in Fig. 2. *Solid lines* indicate subsystem 1 first. *Left*: L^2 error versus window size H for one iteration and one window. *Right*: L^2 error versus the total number of iterations

5 Exact Fine Structure Error Propagation

For our test circuit Fig. 2, we aim at calculating a recursion matrix $\mathbf{K}_e$ explicitly for all unknowns in order to verify the above numerical results. To this end, $\Delta^{(k)}_{\mathbf{X}} X_i := X_i^{(k)}(t) - \tilde{X}_i^{(k)}(t)$ measures the difference of two waveforms on the n-th time window after k iterations. For simplicity of notation the index n is skipped.

We derive the fine structure recursion for our test circuit where the Gauss-Seidel iteration begins with subsystem 1, see (4). For the algebraic variables we find from (2) following the relations to old and new iterates by taking differences

$$\begin{aligned} &\Delta^{(k)}_{\mathbf{z_1}} I_{in} = -\Delta^{(k-1)}_{\mathbf{z_1}} I_{Co}, \quad \Delta^{(k)}_{\mathbf{z_1}} U_{Co} = \Delta U_{in} = 0, \\ &\Delta^{(k)}_{\mathbf{z_2}} I_{Co} = -\Delta^{(k)}_{\mathbf{y_2}} I_L, \qquad \Delta^{(k)}_{\mathbf{z_2}} U_1 = \Delta^{(k)}_{\mathbf{z_1}} U_{Co} = 0, \qquad \Delta^{(k)}_{\mathbf{z_2}} U_2 = \frac{1}{G} \Delta^{(k)}_{\mathbf{y_2}} I_L. \end{aligned} \tag{7}$$

Notice that $U^{(k)}_{Co} = U_{in}(t)$ means that there is no error in the coupling variable $U^{(k)}_{Co}$. From the differential equation for I_L, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\Delta^{(k)}_{\mathbf{y_2}} I_L \right) = \frac{\Delta^{(k)}_{\mathbf{z_2}} U_1 - \Delta^{(k)}_{\mathbf{z_2}} U_2}{L} = \frac{1}{G \cdot L} \Delta^{(k)}_{\mathbf{y_2}} I_L$$

and thus we find for any $t \in [T_n, T_n + H]$

$$|\Delta^{(k)}_{\mathbf{y_2}} I_L(t)| = |\Delta^{(k)}_{\mathbf{y_2}} I_L(t_n)| \cdot e^{(t-t_n)/(G \cdot L)} = |\Delta^{(k-1)}_{\mathbf{y_2}} I_L(t_n)| \cdot e^{(t-t_n)/(G \cdot L)}. \tag{8}$$

Putting (7) and (8) together and using absolute values, we finally find the exact error propagation (for subsystem 1 first):

$$\begin{bmatrix} |\Delta^{(k)}_{\mathbf{y}_2} I_L| \\ |\Delta^{(k)}_{\mathbf{z}_1} I_{in}| \\ |\Delta^{(k)}_{\mathbf{z}_1} U_{Co}| \\ |\Delta^{(k)}_{\mathbf{z}_2} I_{Co}| \\ |\Delta^{(k)}_{\mathbf{z}_2} U_1| \\ |\Delta^{(k)}_{\mathbf{z}_2} U_2| \end{bmatrix} = \underbrace{\begin{bmatrix} 0&0&0&0&0&0 \\ 0&0&0&1&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \end{bmatrix}}_{=\mathbf{K}_e} \begin{bmatrix} |\Delta^{(k-1)}_{\mathbf{y}_2} I_L| \\ |\Delta^{(k-1)}_{\mathbf{z}_1} I_{in}| \\ |\Delta^{(k-1)}_{\mathbf{z}_1} U_{Co}| \\ |\Delta^{(k-1)}_{\mathbf{z}_2} I_{Co}| \\ |\Delta^{(k-1)}_{\mathbf{z}_2} U_1| \\ |\Delta^{(k-1)}_{\mathbf{z}_2} U_2| \end{bmatrix} + e^{(t-t_n)/(G\cdot L)} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ \frac{1}{G} \end{bmatrix} |\Delta^{(k-1)}_{\mathbf{y}_2} I_L(t_n)|. \tag{9}$$

Now, the spectral radius of the recursion matrix is zero $\rho(\mathbf{K}_e) = 0$, since all eigenvalues are zero. Hence, $\mathbf{K}_e$ satisfies the contraction condition, i.e., $\rho(\mathbf{K}_e) < 1$, for splitting scheme (4). An analogous computation verifies contraction for the reversed order of computation. Thus this analysis agrees with our numerical observation of convergence.

Clearly, the relation to the standard theory is the lumping of differential and algebraic components in the error recursion (9). Applying the maximum norm, we obtain the estimate

$$\begin{bmatrix} |\Delta^{(k)}\mathbf{y}| \\ |\Delta^{(k)}\mathbf{z}| \end{bmatrix} \le \mathbf{K} \begin{bmatrix} |\Delta^{(k-1)}\mathbf{y}| \\ |\Delta^{(k-1)}\mathbf{z}| \end{bmatrix} + \gamma := \begin{bmatrix} 0&0 \\ 0&1 \end{bmatrix} \begin{bmatrix} |\Delta^{(k-1)}\mathbf{y}| \\ |\Delta^{(k-1)}\mathbf{z}| \end{bmatrix} + \begin{bmatrix} C \\ C \end{bmatrix} |\Delta^{(k-1)}\mathbf{y}(t_n)|,$$

with $C = (1 + \frac{1}{G})e^{(t-t_n)/(G\cdot L)}$ and $\rho(\mathbf{K}) = 1$. Thus without fine structure analysis, the contraction disappears from the estimate even for our simple test circuit.

6 Conclusions

We have shown that standard co-simulation theory may not always detects convergence. This holds already for a simple electrical circuits, which we have investigated. Therefore we analyzed our model by expressing the exact error propagation (fine structure analysis) and proved stability and thus contraction for our example. In fact convergence holds for both orders of computation.

Clearly, the information about stability and contraction disappeared during lumping, which we have demonstrated for our example. It is a future aim to investigate stability and contraction derived directly from the network structure and thus to generalize convergence results from standard co-simulation theory.

Acknowledgements This work is supported by the German Federal Ministry of Education and Research (BMBF) in the research project SIMUROM project (grant number 05M13PXB). In addition, we acknowledge the support from the project nanoCOPS, Nanoelectronic COupled Problems Solutions (FP7-ICT-2013-11/619166), http://www.fp7-nanoCOPS.eu/.

References

1. Arnold, M., Günther, M.: Preconditioned dynamic iteration for coupled differential-algebraic systems. BIT **41**, 1–25 (2001)
2. Bartel, A., Brunk, M., Günther, M., Schöps, S.: Dynamic iteration for coupled problems of electric circuits and distributed devices. SIAM J. Sci. Comput. **35**(2), B315–B335 (2013)
3. Bartel, A., Brunk, M., Schöps, S.: On the convergence rate of dynamic iteration for coupled problems with multiple subsystems. J. Comput. Appl. Math. **262**, 14–24 (2014)
4. Burrage, K.: Parallel Methods for Systems of Ordinary Differential Equations. Clarendon Press, Oxford (1995)
5. Feldmann, U., Günther, M.: CAD-based electric-circuit modeling in industry I: mathematical structure and index of network equations. Surv. Math. Ind. **8**(2), 97–129 (1999)
6. Lelarasmee, E., Ruehli, A.E., Sangiovanni-Vincentelli, A.: The waveform relaxation method for time domain analysis of large scale integrated circuits. IEEE Comput. Aided Des. Integrated Circuits Syst. **1**, 131–145 (1982)

HJB-POD Feedback Control for Navier-Stokes Equations

Alessandro Alla and Michael Hinze

Abstract In this report we present the approximation of an infinite horizon optimal control problem for the evolutive Navier-Stokes system. The method is based on a model reduction technique, using a POD approximation, coupled with a Hamilton-Jacobi-Bellman (HJB) equation which characterizes the value function of the corresponding control problem for the reduced system. Although the approximation schemes available for the HJB are shown to be convergent for any dimension, in practice we need to restrict the dimension to rather small numbers and this limitation affects the accuracy of the POD approximation. We will present numerical tests for the control of the time-dependent Navier-Stokes system in two-dimensional spatial domains to illustrate our approach and to show the effectiveness of the method.

Keywords Hamilton-Jacobi equations • Navier-Stokes equations • Optimal control • Proper orthogonal decomposition

1 Introduction

In this report we investigate an infinite horizon optimal control problem for the time-dependent Navier-Stokes equations (NSE). The basic ingredient of the method is the coupling between a proper orthogonal decomposition (POD) approximation of the NSE and a Dynamic Programming scheme for the stationary HJB equation characterizing the value function of the optimal control problem. Due to the curse of dimensionality, we need to restrict the dimension of the POD system to a rather small number (typically 4). This limitation naturally affects the accuracy of the POD approximation (see [14]), and, as a consequence, the problem class which we can treat with this technique. It is well known that the solution of the HJB equation is not an easy task from the numerical point of view since viscosity solutions of the HJB equation are usually just Lipschitz-continuous. Optimal control problems for ODEs are solved by Dynamic Programming (DP), both analytically and numerically

A. Alla (✉) • M. Hinze
Department of Mathematics, Universität Hamburg, Bundesstr. 55, 20146 Hamburg, Germany
e-mail: alessandro.alla@uni-hamburg.de; michael.hinze@uni-hamburg.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_120

(see [4] for a general presentation of this theory). From the numerical point of view, this approach has been developed for many classical control problems obtaining convergence results and a-priori error estimates (see the recent book from Falcone and Ferretti [6]). We should mention that a first tentative approach to couple POD and HJB equations is proposed by Atwell and King [3] for the control of the 1D heat equation. Kunisch and Volkwein in [7, 8] extend this approach to diffusion dominated equations and, in particular, Kunisch et al. in [9, 10] apply HJB-POD feedback control to the viscous Burgers equation. We also mention an adaptive POD technique for 1D advection dominated problems proposed by the first author and Falcone in [1, 2].

The novelty in this paper consists in the control of the 2D nonlinear time dependent Navier-Stokes system by means of DP equations and the reduction of the nonlinear term with the *Discrete Empirical Interpolation Method* due to Chaturantabut and Sorensen in [5].

The paper is organized as follows. We first present the optimal control problem in Sect. 2, then we describe the DP equation in Sect. 3. Proper orthogonal decomposition is summarized in Sect. 4 and, finally, the numerical tests are presented in Sect. 5.

2 The Optimal Control Problem

In this section we describe the optimal control problem. The governing equations are the two non-stationary dimensional unsteady Navier-Stokes equations. The flow in the bounded domain $\Omega \subset \mathbb{R}^2$ is characterized by the velocity field $y : \Omega \times [0,T] \to \mathbb{R}^2$ and by the pressure $p : \Omega \times [0,T] \to \mathbb{R}$. The Navier-Stokes equations are given by

$$\left.\begin{aligned} y_t - \nu \Delta y + (y \cdot \nabla y) + \nabla p &= \sum_{i=1}^{N} b_i(x) u_i(t) && \text{in } \Omega \times (0,T], \\ \nabla \cdot y &= 0 && \text{in } \Omega \times (0,T], \\ y(\cdot,0) &= y_0 && \text{in } \Omega, \\ y(\cdot,t) &= y_b && \text{in } \partial\Omega \times (0,T), \end{aligned}\right\} \tag{1}$$

where the viscosity of the flow is given by the parameter $\nu > 0$. The control signals are elements of $\mathscr{U} \equiv \{u : [0,T] \to U,\ u(\cdot) \in L^\infty(0,T)\}$, where U is a compact subset of $\mathbb{R}^m$. Later we take U as a discrete set as explained in [6]. The initial value and the boundary values are denoted by y_0 and y_b, respectively. Finally, the functions $b_i(x) : \Omega \to \mathbb{R}^2$ play the role of the so called shape functions, which model the actions that we can apply to the physical system governed by our PDE.

The cost functional we want to minimize is given by

$$J(u) := \int_0^\infty \left(\|y(\cdot, t; u) - \bar{y}\|^2_{L^2(\Omega)} + \alpha |u(t)|^2 \right) e^{-\lambda t}\, dt, \tag{2}$$

where $\bar{y}$ is the desired state which we choose as the mean flow, $\alpha \in \mathbb{R}^+$ and $\lambda > 0$ is the discount factor. The optimal control problem, then, can be formulated as

$$\min_{u \in \mathscr{U}} J(u) \text{ s. t. } y(u) \text{ satisfies (1)}. \tag{3}$$

We should state, that (1) for a given sufficiently smooth right hand side together with sufficiently smooth initial values and boundary conditions admits a unique solution. We refer to the book of Temam [13] for more details. Whenever we want to emphasize the dependence of the solution on the control u we will write $y = y(u)$.

3 Dynamic Programming Equation

We illustrate the dynamic programming approach for abstract optimal control problems of the form

$$\min_{u \in \mathscr{U}} J_x(u) := \int_0^\infty L(y(t), u(t))\, e^{-\lambda t}\, dt \text{ subject to } \dot{y}(t) = f(y(t), u(t)),\ y(0) = x, \tag{4}$$

with system dynamics in $\mathbb{R}^n$. We assume $\lambda > 0$, and $L(\cdot, \cdot)$ and $f(\cdot, \cdot)$ to be Lipschitz-continuous, bounded functions. Then, it is clear that the optimal control problem (3) fits into the more abstract setting (4).
In this setting, a standard solution tool is the application of the dynamic programming principle, which leads to a characterization of the value function $v(x) := \inf_{u \in \mathscr{U}} J_x(u)$ as a viscosity solution of the HJB equation

$$\lambda v(x) - \inf_{u \in U} \{ Dv \cdot f(x, u) + L(x, u) \} = 0\,. \tag{5}$$

To approximate Eq. (5), we construct a fully-discrete semi-Lagrangian scheme which is based on a discretization of the system dynamics with time step h, and a finite element discretization of the state space with mesh parameter k, leading to a fully discrete approximation $V_{h,k}(x)$ of the value function v satisfying

$$V_{h,k}(x_i) = \min_{u \in U} \{ (1 - \lambda h) I_1[V_{h,k}](x_i + h f(x_i, u)) + L(x_i, u) \}\,, \tag{6}$$

for every element x_i of the discretized spatial domain. In general, the arrival point $x_i + h f(x_i, u)$ is not a node of the state space grid, and therefore the value of $V_{h,k}$ at

this point is approximated by means of a first-order interpolant of the data, denoted by $I_1[V_{h,k}]$ (we refer the reader to [4, Appendix A] for more details).

The goal is to find a feedback control law of the form $u(t) = \Phi(y(t), t)$ which steers the system to the desired trajectory. Φ is called *feedback map*. The computation of feedback maps is almost built in and comes straightforward from the knowledge of the value function. In fact;

$$\Phi(y_x(t)) = u^*(t) = \arg\min_{u \in U} \left\{ L(x,u) + \nabla v(x)^T f(x,u) \right\},$$

and the discrete version may be computed by the semi-Lagrangian scheme already explained.

The characterization of the value function is valid for all classical problems in any dimension and its approximation is based on a-priori error estimates in L^∞.

The request to solve an HJB in high dimensions comes up naturally whenever we want to control evolutive PDEs. However, a direct discretization, in many practically relevant situations, is impossible since the system of ODEs associated to a semi-discretization in time would have the dimension equal to the space dimension where one should solve the HJB equation. Fortunately, at the discrete level, the POD [12, 14] method allows us to obtain low-dimensional reduced models even for complex dynamics, and, thus, presents an opportunity to circumvent the curse of dimensionality in the numerical solution of the HJB equation.

4 POD-Model Reduction for the Controlled Problem

The Reduced Order Modelling (ROM) approach to optimal control problems is based on projecting the nonlinear dynamics onto a low dimensional manifold utilizing projectors that contain informations of the expected controlled flow. A common approach here is based on the snapshot form of POD proposed by Sirovich in [12], which in the present situation works as follows. We compute the snapshots set $y_1, \ldots, y_n$ of the flow corresponding to different time instances $t_1, \ldots, t_n$ and define the POD ansatz of order ℓ for the state y by

$$y^\ell = \bar{y} + \sum_{i=1}^{\ell} w_i \psi_i, \tag{7}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ denotes the mean flow and the basis functions $\{\psi_i\}_{i=1}^{\ell}$ are obtained from the singular value decomposition of the snapshot matrix $Y = [y_1 - \bar{y}, \ldots, y_n - \bar{y}]$, i.e. $Y = \Psi \Sigma V$, and the first ℓ columns of Ψ form the POD basis functions of rank ℓ. Here the SVD is based on the Euclidean inner product. This is reasonable in our situation, since the numerical computations performed in our numerical example for the driven cavity problem are based on a uniform staggered grid. The snapshots are computed on the basis of a stable finite difference

discretization of (1) which leads to a semi-discrete system of ODEs of the form

$$\dot{y} + \nu Ay + Cp = \eta(y) + Bu, \quad y(0) = y_0. \tag{8}$$

Note that we consider $f(y(t), u(t)) = -\nu Ay - Cp + \eta(y) + Bu$ in (4).

The reduced optimal control problem is obtained through replacing (8) by a dynamical system obtained from a Galerkin approximation with basis functions $\{\psi_i\}_{i=1}^{\ell}$ and ansatz (7) for the state.

This leads to a ℓ-dimensional system for the unknown coefficients $\{w_i\}_{i=1}^{\ell}$, namely

$$M^{\ell}\dot{w} + \nu A^{\ell} w = \eta(w) + B^{\ell} u \quad w(0) = w_0. \tag{9}$$

Here the entries of the mass M^{ℓ} and the stiffness A^{ℓ} are given by $\langle \psi_j, \psi_i \rangle$ and $\langle \psi_j, A\psi_i \rangle$, respectively. The reduced shape function is obtained by $(B^{\ell})_i = \langle B, \psi_i \rangle$. The coefficients of the initial condition $y^{\ell}(0) \in \mathbb{R}^{\ell}$ are determined by $w_i(0) = (w_0)_i = \langle y_0 - \bar{y}, \psi_i \rangle$, $1 \le i \le \ell$, and the solution of the reduced dynamical problem is denoted by $w(s) \in \mathbb{R}^{\ell}$. Note that for the reduction of the nonlinear term $\eta(w)$ we use the *Discrete Empirical Interpolation Method* (DEIM, see[5]). The pressure does not appear in the reduced problem (9) since the snapshots are divergence-free. Then, the POD-Galerkin approximation leads to the optimization problem

$$\inf J_{w_0}^{\ell}(u), \tag{10}$$

where $u \in \mathscr{U}$, w solves (9) and the cost functional is defined by

$$J_{w_0}^{\ell}(u) = \int_0^{\infty} L(w(s), u(s), s) e^{-\lambda s} \, ds.$$

The value function v^{ℓ}, defined for the initial state $w_0 \in \mathbb{R}^{\ell}$ is given by

$$v^{\ell}(w_0) = \inf_{u \in \mathscr{U}} J_{w_0}^{\ell}(u),$$

and w solves (9) with the control u and initial condition w_0. HJB equations are defined in $\mathbb{R}^n$, but we need to restrict our numerically domain to a bounded subset of $\mathbb{R}^n$. We refer the interested reader to [1] for a detailed description.

5 Numerical Tests

In this section we consider as numerical example the control of the flow in the lid-driven cavity. In (3) we set: $\Omega = (0,1) \times (0,1)$, $y_0 \equiv 0$, $\nu = 0.01$, $\alpha = 0.01$, $\lambda = 1$, $U = \{-1, 0, 1\}$, $y_b = (1, 0)$ on the top boundary and $y_b = (0, 0)$ on the remaining

boundary segments. The desired configuration is given by $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. In (6) we take $k = 0.2$, $h = 0.04$ whereas the optimal trajectory is obtained with a time stepsize of 0.01.

The control gain of the suboptimal control problem, with the ansatz (7), consists of steering the coefficients w to the origin. For the purpose of this test, we take only three POD and six POD-DEIM basis functions. In our numerical computations the reduction of the nonlinearity with DEIM already yields a considerable computational speedup. Further investigations on the performance of DEIM in relation to the discretization parameters are provided in a subsequent paper. The snapshots are computed with a finite difference scheme from the uncontrolled problem ($u \equiv 0$) in (1) where we use the Matlab code provided in [11].

In Fig. 1 we show the configuration of the flow. On the left we show the mean flow, which is the desired state, in the middle the controlled flow is shown, and on the right the uncontrolled flow is shown. As shape function we use the steady state solution of the Navier-Stokes system.

We can see that at time $t = 0.5$ the suboptimal solution already well approximates the desired state, as confirmed in Table 1, where the L^∞-error of $y^\ell - \bar{y}$ at $t = 0.5$ and $t = 4$ is reported for this shape function. When the time is increasing the solution itself tends to stabilize close to the mean flow, but still the suboptimal solution has a smaller error with respect to the uncontrolled problem. Note that the performance of our method depends on the choice of the shape functions. In Table 2 we display the results obtained with the steady state solution of the Stokes equation

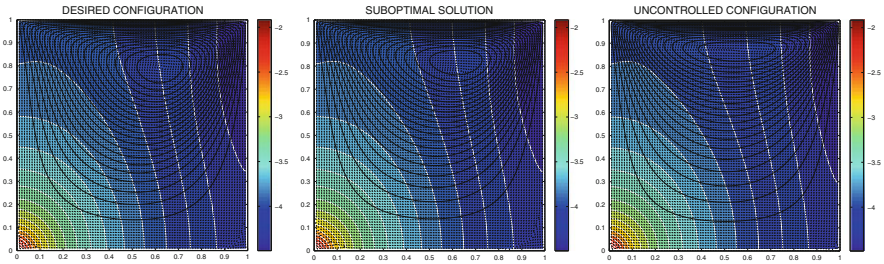


Fig. 1 Mean flow NS (*left*)—controlled configuration at time $t = 0.5$ (*middle*)—uncontrolled configuration at time $t = 0.5$ (*right*)

Table 1 L^∞ error at time $t = 0.5$ and $t = 4$

	$t = 0.5$	$t = 4$
$\|y^\ell(x,t,u^\ell) - \bar{y}\|_\infty$	0.007	0.006
$\|y(x,t;0) - \bar{y}\|_\infty$	0.283	0.048

$\bar{y}$ is the desired state, $y^\ell(x,t;u^\ell)$ is the suboptimal solution, and $y(x,t;0)$ denotes the uncontrolled solution. The shape function is chosen as the steady state solution of the Navier-Stokes equations

Table 2 L^∞ error at time $t = 0.5$ and $t = 4$

	$t = 0.5$	$t = 4$
$\|y^\ell(x,t,u^\ell) - \bar{y}\|_\infty$	0.081	0.022
$\|y(x,t;0) - \bar{y}\|_\infty$	0.283	0.048

$\bar{y}$ is the desired state, $y^\ell(x,t;u^\ell)$ is the suboptimal solution, and $y(x,t;0)$ denotes the uncontrolled solution. The shape function is chosen as the steady state solution of the Stokes equations

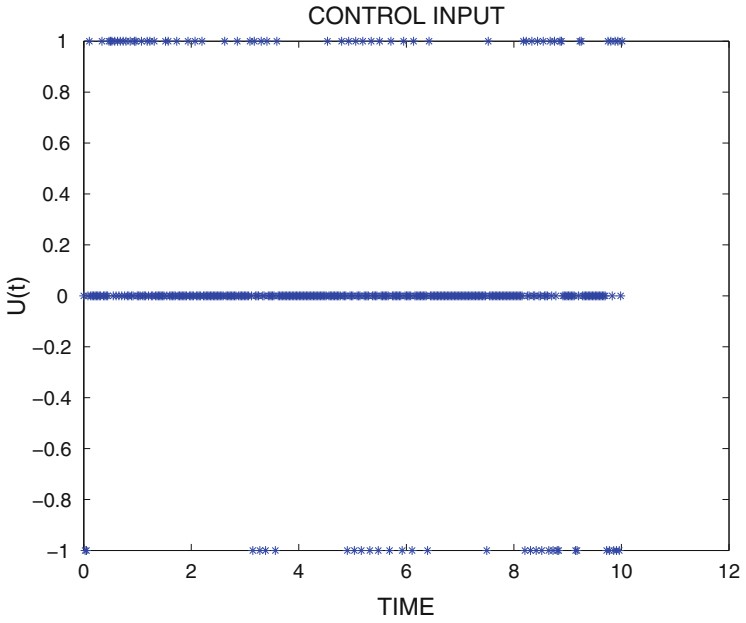


Fig. 2 Control input with three constant controls $\{-1, 0, 1\}$ and one shape function chosen as the steady state solution of the Navier-Stokes equation

as shape function. As expected, the approach works better if we can use the steady state of the Navier-Stokes equation as shape function.

In Fig. 2 we present the control input. The behavior of the control is classical for feedback control, since the system tries to correct step by step the trajectories. The control space is only given by constant values $\{-1, 0, 1\}$.

References

1. Alla, A., Falcone, M.: An adaptive POD approximation method for the control of advection-diffusion equations. In: Control and Optimization with PDE Constraints. International Series of Numerical Mathematics. Birkhauser, Basel (2013)

2. Alla, A., Falcone, M.: A time-adaptive POD method for optimal control problems. In: Proceedings of the 1st IFAC Workshop on Control of Systems Modeled by Partial Differential Equations (2013)
3. Atwell, J.A., King, B.B.: Proper orthogonal decomposition for reduced basis feedback controllers for parabolic equations. Math. Comput. Model. **33**, 1–19 (2001)
4. Bardi, M., Capuzzo Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Birkhauser, Basel (1997)
5. Chaturantabut, S., Sorensen, D.C.: Discrete empirical interpolation for nonlinear model reduction. SIAM J. Sci. Comput. **32**, 2737–2764 (2010)
6. Falcone, M., Ferretti, R.: Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations. SIAM, Philadelphia, PA (2013)
7. Kunisch, K., Volkwein, S.: Control of Burgers' equation by a reduced order approach using proper orthogonal decomposition. J. Optim. Theory Appl. **102**, 345–371 (1999)
8. Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for parabolic problems. Numer. Math. **90**, 117–148 (2001)
9. Kunisch, K., Xie, L.: POD-based feedback control of Burgers equation by solving the evolutionary HJB equation. Comput. Math. Appl. **49**, 1113–1126 (2005)
10. Kunisch, K., Volkwein, S., Xie, L.: HJB-POD based feedback design for the optimal control of evolution problems. SIAM J. Appl. Dyn. Syst. **4**, 701–722 (2004)
11. Seibold, B.: A compact and fast Matlab code solving the incompressible Navier-Stokes equations on rectangular domains (2008). http://math.mit.edu/cse/codes/mit18086-navierstokes.pdf
12. Sirovich, L.: Turbulence and the dynamics of coherent structures. Parts I–II. Q. Appl. Math. **XVL**, 561–590 (1987)
13. Temam, R.: Navier-Stokes Equations: Theory and Numerical Analysis. American Mathematical Society, Philadelphia, PA (2001)
14. Volkwein, S.: Model reduction using proper orthogonal decomposition (2011). www.math.uni-konstanz.de/numerik/personen/volkwein/index.php

MS 35
MINISYMPOSIUM: PARTICLE METHODS AND THEIR APPLICATIONS

Organizers

Giuseppe Bilotta[1] and Alexis Hérault[2]

Speakers

Giuseppe Bilotta[1], Alexander Vorobyev[3], Alexis Hérault[2], Damien VIoleau[4], Ciro Del Negro[5]
SPH for the Simulation of a Dam-Break with Floating Objects

Sudarshan Tiwari[6], Axel Klar[7], Steffen Hardt[8]
Numerical Simulation of Wetting Phenomena by a Meshfree Particle Method

[1]Giuseppe Bilotta, Istituto Nazionale di Geofisica e Vulcanologia, Catania, IT.
[2]Alexis Hérault, Conservatoire National des Arts et Métiers, Paris, FR.
[3]Alexander Vorobyev, Électricité de France, Paris, FR.
[4]Damien Violeau, Électricité de France, Paris, FR.
[5]Ciro Del Negro, Istituto Nazionale di Geofisica e Vulcanologia, Catania, IT.
[6]Sudarshan Tiwari, University of Kaiserslautern, Kaiserslautern, DE.
[7]Axel Klar, University of Kaiserslautern, Kaiserslautern, DE.
[8]Steffen Hardt, Technische Universitaet Darmstadt, Darmstadt, DE.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_121

Eugenio Rustico[9], Béla Sokoray-Varga[10], Giuseppe Bilotta[1], Alexis Hérault[2], Thomas Brudy-Zippelius[11]
Full 3D Numerical Simulation and Validation of a Fish Pass with GPUSPH

Keywords

Meshfree method
Particle method

Short Description

Particle methods such as Smoothing Particle Hydrodynamics (SPH) are Lagrangian, meshless numerical method for Computational Fluid Dynamics (CFD) which has recently seen a growing interest in a wide variety of problems, including hydrodynamics, multi-fluid simulations, thermal problems, lava flow simulations, fluid-structure interaction problems, with applications ranging from oceanography to medicine, from engineering to geophysics.

Particle methods are a powerful and flexible tool for computational fluid-dynamics of great relevance for environmental and industrial applications. The purpose of the mini-symposium is to present the current state-of-the-art in applied particle methods, both for scientific research and in industrial applications, providing an opportunity for researchers and applied mathematicians working on and with particle methods to present their work both to fellow scientists in the same research fields and to a wider audience.

[9]Eugenio Rustico, Bundesanstalt für Wasserbau, Karslruhe, DE.
[10]Béla Sokoray-Varga, Bundesanstalt für Wasserbau, Karslruhe, DE.
[11]Thomas Brudy-Zippelius, Bundesanstalt für Wasserbau, Karslruhe, DE.

Full 3D Numerical Simulation and Validation of a Fish Pass with GPUSPH

Eugenio Rustico, Béla Sokoray-Varga, Giuseppe Bilotta, Alexis Hérault, and Thomas Brudy-Zippelius

Abstract We present a validated fully three-dimensional simulation of a vertical-slot fish pass with GPUSPH, a high-performance CUDA implementation of the Smoothed Particles Hydrodynamics (SPH) numerical method for free-surface flows. The GPUSPH results are compared to flow velocity and water level measurement from a laboratory model with the same geometry. The results show good agreement between the numerical simulations and the experimental data.

Keywords Free-surface flow • Meshfree method • Smoothed particle hydrodynamics method

E. Rustico (✉)
BAW Karlsruhe, Karlsruhe, Germany

INGV, Sezione di Catania, Catania, Italy
e-mail: eugenio.rustico@baw.de

B. Sokoray-Varga
Institute for Water and River Basin Management, Karlsruhe Institute of Technology, Karlsruhe, Germany
e-mail: bela.sokoray-varga@kit.edu

G. Bilotta
INGV, Sezione di Catania, Catania, Italy
e-mail: bilotta@ct.ingv.it

A. Hérault
CNAM, Paris, France

INGV, Sezione di Catania, Catania, Italy
e-mail: alexis.herault@cnam.fr

T. Brudy-Zippelius
BAW Karlsruhe, Karlsruhe, Germany
e-mail: thomas.brudy-zippelius@baw.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_122

1 Introduction

CFD modeling represents an efficient tool to investigate different geometry variants in hydraulic engineering. However, a validation of the models is necessary, due to the complexity and high turbulence level of the flow in a several facilities.

In the present paper we simulate a vertical-slot fish pass with GPUSPH [4], a high-performance CUDA implementation of the Smoothed Particles Hydrodynamics (SPH) numerical model for free-surface flows.

Hydraulic research on vertical slot fish passes has shown that the hydraulic conditions within the pools of such facilities are mostly determined by the pool geometry and the slope of the fish pass. While previous SPH-based approaches to the numerical simulation of fish passes were mostly in two dimensions, the use of SPH allows to run fully three-dimensional simulations, without constraints on the domain shape and with great accuracy. This comes at the cost of significant requirements in terms of memory and computational power, so that only low-resolution attempts have been done so far [6]. To cope with this, we have extended GPUSPH, which previously supported single-node multi-GPU computing, to exploit the computational power of clusters of graphic devices. By exploiting 12 devices across 6 nodes simultaneously we have been able to run a fine-grained simulation with a resolution of about 4 mm per particle and a total of 50 millions particles.

We first present the SPH discretization of the Navier-Stokes equations (Sect. 2), followed by some technical information about GPUSPH, the implementation of SPH on GPU we used for our simulations (Sect. 3). We then introduce the experimental and numerical set-ups (Sect. 4), followed by the comparison between experimental measurements and numerical results (Sect. 5), and some concluding remarks (Sect. 6).

2 SPH Discretization of Navier-Stokes Equations

The motion of a weakly compressible inviscid fluid can be described by the Navier-Stokes equations in the Lagrangian form

$$\frac{D\mathbf{u}}{Dt} = -\frac{\nabla P}{\rho} + \mathbf{g}$$

$$\frac{D\rho}{Dt} = -\rho \nabla \cdot \mathbf{u},$$

where D/Dt denotes the total (Lagrangian) derivative with respect to time, $\mathbf{v}$ the fluid velocity, P the pressure, ρ the density, ν the kinematic viscosity coefficient, and $\mathbf{g}$ the external forces per unit mass (in our case, gravity).

SPH is a meshless Lagrangian numerical method for computational fluid dynamics [8]. Using the standard notation for Weakly-Compressible SPH (WCSPH), we denote by $W(\cdot, h)$ a family of smoothing kernels parametrized by the influence radius h and such that $\lim_{h\to 0} W(\cdot, h) = \delta(\cdot)$, where δ is Dirac's delta and the limit is intended in the sense of distributions. The smoothing kernels are assumed to have compact support and to be positive and with radial symmetry. The fluid body is discretized by a set of particles with average inter-particle distance Δp and we will denoted with $\mathbf{r}_a$ the position of particle a, while $\mathbf{r}_{ab} = \mathbf{r}_b - \mathbf{r}_a$ denotes the distance vector between the particles, r_{ab} its norm and $W_{ab} = W(r_{ab}, h)$.

For the physical properties we will use the usual subscripted convention, with ρ_a being the density of particle a, m_a its mass, V_a its volume, P_a its pressure and $\mathbf{u}_a$ its velocity.

The flow is assumed weakly compressible with an equation of state (EOS) coupling pressure and density. Typically, the Tait EOS is used, in the form $P(\rho) = B((\rho/\rho_0)^\gamma + 1)$ where $\gamma = 7$ is the polytropic constant, ρ_0 the at-rest density of the fluid and $B = \rho_0 c_0^2/\gamma$, c_0 being the numerical speed of sound density of the fluid.

The mass continuity equation is then discretized as

$$\frac{D\rho_i}{Dt} = \rho_i \sum_j \frac{m_j}{\rho_j} \mathbf{u}_{ij} \cdot \nabla_i W_{ij},$$

and the Navier-Stokes momentum equation can be written

$$\frac{D\mathbf{u}_i}{Dt} = -\sum_j m_j \left(\frac{P_j}{\rho_j^2} + \frac{P_i}{\rho_i^2} + \Pi_{ij} \right) \nabla_i W_{ij} + \mathbf{g},$$

where $F(r, h) = (1/r)(\partial W(r, h)/\partial r)$, and the Π_{ij} term is an artificial viscosity that takes the form

$$\Pi_{ij} = \begin{cases} -\alpha \frac{\bar{c}_s}{\bar{\rho}} h \frac{\mathbf{x}_{ij} \cdot \mathbf{u}_{ij}}{r_{ij} + \varepsilon h^2} & \mathbf{x}_{ij} \cdot \mathbf{u}_{ij} < 0, \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{c}_s$ is the average speed of sound computed at the particles i and j, and $\bar{\rho}$ their average density.

In most applications, explicit integration methods are used, so that the timestep is limited by the speed of sound in the EOS. Hence, rather than using the physical speed of sound, a fictitious speed of sound is used, which is at least one order of magnitude higher than the maximum velocity of the fluid in the given problem: this ensures that density fluctuations are kept small (less than 1 %) while allowing for larger timesteps. In our experiments, the expected maximum velocity is 11.7 m/s, so we set $c_0 = 117$ m/s.

3 GPU SPH Implementation

For our numerical tests, we use GPUSPH [4], which is a modular implementation of SPH relying on CUDA-enabled GPUs as high-performance computing devices. GPUSPH is capable of distributing the computation across multiple devices attached to one or more nodes in a network: this allows to run large-scale simulation and/or reduce computational times.

3.1 Integration Scheme

GPUSPH uses a predictor-corrector integration scheme. With accelerations $\mathbf{f}$ and time-step dt, the scheme describing a time-step with the standard SPH formulation can be summarized as follows: compute accelerations $\mathbf{f}^{(n)} = \mathbf{f}(\mathbf{x}^{(n)}, \mathbf{v}^{(n)}, \rho^{(n)})$ and density derivatives $\dot{\rho}^{(n)} = \dot{\rho}(\mathbf{x}^{(n)}, \mathbf{v}^{(n)}, \rho^{(n)})$; compute half-step intermediate positions, velocities, densities: $\mathbf{x}^{(n\star)} = \mathbf{x}^{(n)} + \mathbf{v}^{(n)}\frac{dt}{2}$, $\mathbf{v}^{(n\star)} = \mathbf{v}^{(n)} + \mathbf{f}^{(n)}\frac{dt}{2}$, $\rho^{(n\star)} = \rho^{(n)} + \dot{\rho}^{(n)}\frac{dt}{2}$; compute corrected accelerations $\mathbf{f}^{(n\star)} = \mathbf{f}(\mathbf{x}^{(n\star)}, \mathbf{v}^{(n\star)}, \rho^{(n\star)})$ and density derivatives $\dot{\rho}^{(n\star)} = \dot{\rho}(\mathbf{x}^{(n\star)}, \mathbf{v}^{(n\star)}, \rho^{(n\star)})$; compute new positions, velocities, densities: $\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + (\mathbf{v}^{(n)} + \mathbf{f}^{(n\star)}\frac{dt}{2})dt$, $\mathbf{v}^{(n+1)} = \mathbf{v}^{(n)} + \mathbf{f}^{(n\star)}dt$, $\rho^{(n+1)} = \rho^{(n)} + \dot{\rho}^{(n\star)}dt$.

Stability of this scheme is guaranteed when the time-step is limited by CFL-like conditions in the form $\Delta t \leq Cmin(\sqrt{(h/\|f\|)}, h/c_0)$ for some constant $C < 1$. In our experiments we have $C = 0.1$.

3.2 Neighbors List

For efficiency reasons, GPUSPH keeps track of the neighbors of each particle in a dedicated list. Additionally, since the fluid is normally not subject to large instantaneous deformation, the neighbors list is only rebuilt every n timesteps, with n configurable. In our tests we use the default value $n = 10$.

To speed up neighbor search during the neighbors list construction, particles are indexed by their position with respect to an auxiliary grid with spacing not smaller than the influence radius of the kernel. This ensures that the neighbors of a given particle can be found in cells neighboring the cell to which the particle belongs [2]. Computations are distributed across multiple devices via domain decomposition at the granularity of the auxiliary cell [9, 10].

3.3 Homogeneous Accuracy

As explained in [5], the auxiliary cell grid is used in GPUSPH also to provide homogeneous accuracy in the particle position throughout the domain: instead of

storing the positions of the particles in a global reference frame, GPUSPH stores the cell index (which is used for the neighbor list, and is thus already computed), and the position of the particles with respect to the center of the cell it belongs to.

This ensures that the distance vector between particles can be computed with the same accuracy regardless of the location of the particles with the respect to the origin of the global reference frame, by computing the difference between the local positions of the particles, and adding an offset equal to the cell side when particles are in different cells.

In our application, homogeneous accuracy provides a significant benefit, since the ratio of the particle resolution ($\Delta p = 3.88\,\text{mm}$) to the domain size (15 m in length) is in the order of the single-precision floating-point machine epsilon (10^{-7}, and thus high enough to cause numerical problems with a naive implementation, such as particles near the external parts of the domain clumping together due to their relative distance being computed with only one or two bits of precision at most.

3.4 Boundary Conditions

We model physical domain boundaries using the Lennard-Jones boundary particles method, the classic approach to realize solid boundaries in SPH: physical boundaries are discretized with one layer of equally spaced particles that exert a Lennard-Jones force on fluid particles.

A side effect of using Lennard-Jones boundary particles is the generation of artificial friction of the fluid against the wall. This effect can be reduced by increasing the spatial density of the particles realizing the walls. For simple, convex, plane geometries a better approach is to model the wall as a geometrical plane, and then compute the fluid/plane interaction by finding the projection of the fluid particle on the plane and compute a Lennard-Jones repulsive force from the projection to the actual particle. For particles travelling parallel to the plane, this results in a constant force, and a much smoother motion. Further details on the Lennard-Jones repulsive plane can be found in [3], where the approach was used to locally approximate complex topographies.

4 Modeling a Vertical-Slot Fish Pass with GPUSPH

A laboratory model of the modeled fish pass has already existed in the laboratory of the Federal Waterways Engineering and Research Institute (BAW) as part of ongoing R&D activities.

The laboratory model consists of nine pools with width of 78.5 cm and length of 99 cm installed in a flume. The slot width is 12.2 cm; the slope of the flume is 2.8 %. The side walls and the bottom of the flume are made of plexiglas, the cross-walls and the baffles of the model are made of wood.

The SPH model was designed to reproduce the same geometry, inlet and outlet conditions of the laboratory model. Influx rate, sizes, wall offsets, initial filling and other geometric details are fully parametrized to allow for fast comparative tests and efficient maintenance in case of changes in the physical model.

The physical boundaries of the domains are modeled with a combination of Lennard-Jones repulsive particles and Lennard-Jones repulsive planes. Planes are used to model the side walls of the entire fishpass, as they yield no friction to the fluid stream, and their usage reduces the total number of particles required for the simulation. However, since the current implementation only allows infinite planes, they cannot be used to model other parts of the fishpass (inner walls and floor), for which the standard Lennard-Jones repulsive particles are used instead, with a linear density which is double that of the fluid, in order to reduce the artificial friction introduced by this kind of boundaries.

Inflow is modeled by reserving a section (inlet) at the beginning of the domain for particle generation. These particles are generated with an initial velocity such that the inlet achieves the same inflow rate as the physical model. The inlet itself is divided in two section: the first creates particles and applies the velocity field as is; the second makes a linear interpolation between the velocity imposed by the field and the one which particles would have according to their dynamics computed from the SPH formulation.

A ramp at the end of the outflow channel is used to model the flap gate, which was not moved for this test case. In the computation model, an outflow field is placed right after the freefall (Fig. 1). Its only task is to destroy the particles that fall into it.

Water levels are captured with special "gage particles", which are floating particles with fixed X and Y coordinates. Two gages are set in each pool in the same positions where the water level are measured in the physical model.

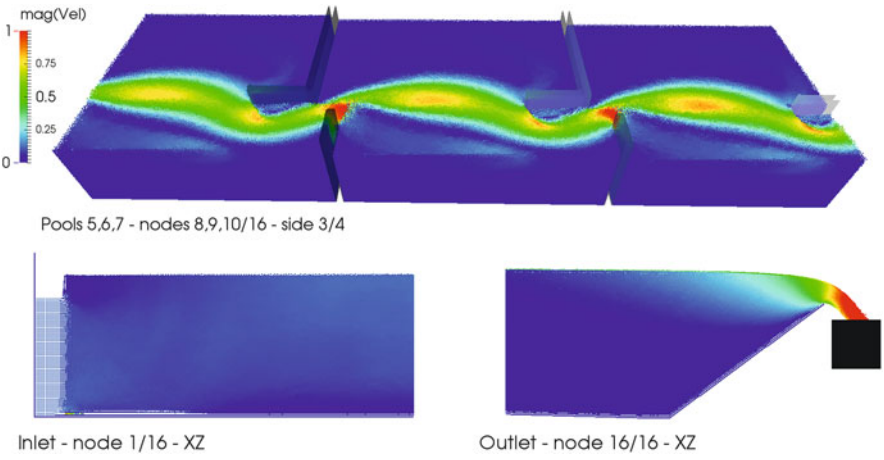


Fig. 1 Three pools, inlet and outlet of a 50 M particles simulation

5 Validation Using Laboratory Measurements

The validation data was provided by measurements performed in the laboratory model described before. The measurements were performed in uniform flow conditions. The measurement duration was 3 min in each point for both the water level sensors and flow velocimeters, as suggested by a preliminary stationarity analysis.

The flow velocity measurements were carried out in the sixth pool of the model 12 cm above bottom level along 134 gridpoints by an Acoustic Doppler Velocimeter (ADV) of type side-looking Vectrino with 200 Hz sampling rate.

For the validation of the flow velocities, velocities were captured at the same gridpoints in the numerical model as in the laboratory model. The flow velocities were recorded for every time-step in the numerical model. The obtained velocity time series show fluctuations, so time averaged values over 10 s were used during the comparison for stationarity reasons. The measured and modeled velocity fields are presented in Fig. 2. It can be observed, that the main flows interconnecting the slots have a similar shape and the magnitudes of the velocities show a good agreement.

The water depths measurements were performed in points A and B in each of the nine pools by ultrasonic water level sensors with a sampling rate of 40 Hz. Water depths in points A and B nearly represent the minimal and maximal water depths within the pools, so that the average of the two values was used as the characteristic water depth in the pool. The water levels captured in the numerical model were transformed to water depth values, and were then averaged over 10 s to get statistical stationary values.

The water depths obtained from the measurements and the numerical model are presented in Fig. 3. It can be observed that the water depths in the numerical model are higher than in the laboratory model, with a larger difference close to the inflow. This indicates that our SPH model produces more hydraulic loss than the laboratory model, which is probably due to a combination of boundary effects, such as the wall

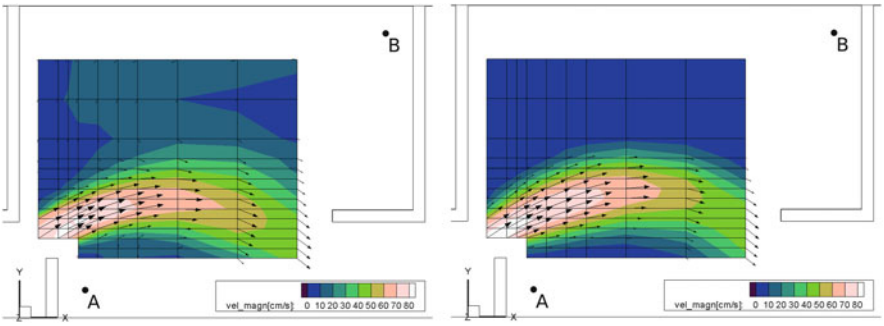


Fig. 2 Velocity field measured in the sixth pool of the laboratory model (*left*) and captured in the sixth pool of the numerical model (*right*)

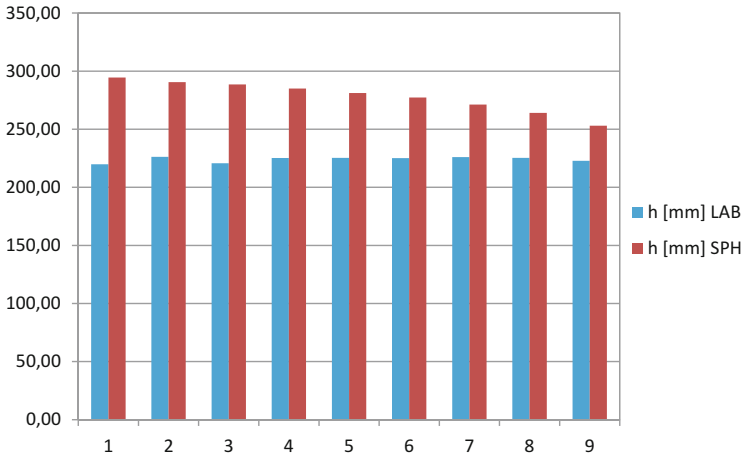


Fig. 3 Characteristic water depths in the pools in the laboratory model and SPH

friction and inflow conditions, as well as to the non-physical viscosity applied in the model.

6 Conclusions

These preliminary tests show good agreement between the numerical simulations and the experimental data.

To improve the results, we need to reduce the energy loss of the fluid, to which end we are testing different viscosity models and we plan to switch to the unified semi-analytical boundary model [1, 7] for the fishpass walls. This boundary model will also allow better modeling of inlet and outlet conditions.

References

1. Ferrand, M., Laurence, D.R., Rogers, B.D., Violeau, D., Kassiotis, C.: Unified semi-analytical wall boundary conditions for inviscid, laminar or turbulent flows in the meshless SPH method. Int. J. Numer. Methods Fluids **71**(4), 446–472 (2013). doi: 10.1002/fld.3666. http://dx.doi.org/10.1002/fld.3666
2. Green, S.: Particle simulation using CUDA (2010)
3. Hérault, A., Bilotta, G., Vicari, A., Rustico, E., Del Negro, C.: Numerical simulation of lava flow using a GPU SPH model. Ann. Geophys. **54**(5) (2011). doi: 10.4401/ag-5341
4. Hérault, A., Bilotta, G., Dalrymple, R.A.: SPH on GPU with CUDA. J. Hydraul. Res. **48**(extra issue), 74–79 (2010)
5. Hérault, A., Bilotta, G., Dalrymple, R.A.: Achieving the best accuracy in an SPH implementation. In: Proceedings of the 9th International SPHERIC Workshop (2014)

6. Marivela, R.: Applications of the SPH model to the design of fishways. In: 33rd IAHR Congress, Vancouver (2009)
7. Mayrhofer, A., Ferrand, M., Kassiotis, C., Violeau, D., Morel, F.X.: Unified semi-analytical wall boundary conditions in SPH: analytical extension to 3-D. In: Numerical Algorithms, pp. 1–20. Springer, Berlin (2014). doi: 10.1007/s11075-014-9835-y. http://dx.doi.org/10.1007/s11075-014-9835-y
8. Monaghan, J.J.: Smoothed particle hydrodynamics. Rep. Prog. Phys. **68**(8), 1703 (2005)
9. Rustico, E., Bilotta, G., Herault, A., Negro, C.D., Gallo, G.: Advances in multi-GPU smoothed particle hydrodynamics simulations. IEEE Trans. Parallel Distrib. Syst. **25**(1), 43–52 (2014). doi:http://doi.ieeecomputersociety.org/10.1109/TPDS.2012.340
10. Rustico, E., Jankowski, J., Herault, A., Bilotta, G., Del Negro, C.: Multi-GPU, multi-node SPH implementation with arbitrary domain decomposition. In: Proceedings of the 9th International SPHERIC Workshop, Paris, pp. 127–133 (2014)

Simulation of a Twisting-Ball Display Cell

Peep Miidla, Jüri Liiv, Aleksei Mashirin, and Toomas Tenno

Abstract A system of differential equations describing the behavior of a single ball in an elementary cell of the twisting-ball information display is considered. Nonlinear ordinary differential equations describe the ball shift and rotation. For efficient practical implementation of the display, the optimal values of ball and cell parameters are needed. To obtain these, the computer simulations were realized. Results of numerical experiments of modeling the balls with different physical parameters are presented. The numerical experiments show that the movement of the elementary particle of the twisting-ball display is extremely sensitive to the physical parameters of the balls but there exist nearly optimal combinations of these parameters. In this case the ball rotation intends towards some complete rotation cycle: if control voltage changes its polarity, the ball rotates nearly 180° and exposes right, black or white, size to the observer and the display works as expected.

Keyword Twisting-ball display cell

1 Introduction

A twisting-ball display is a kind of electrophoretic information display invented at the Xerox Palo Alto Research Center [5–7]. The display consists of a thin layer of transparent silicone plastic in which multiple randomly dispersed bichromal balls. Each ball is an electrical dipole and is placed in the cavity filled with dielectric fluid. The width of the cavity is 10–30 % greater than the diameter of the balls. Depending on the polarity of the control voltage, one or other of their colored hemispheres is exposed to the viewer. The displays based on such physical principles are highly bistable, robust, easy to manufacture and have very low power consumption as they do

P. Miidla (✉)
Institute of Mathematics, Tartu University, J. Liivi 2, 50409 Tartu, Estonia
e-mail: peep.miidla@ut.ee

J. Liiv • A. Mashirin • T. Tenno
Institute of Chemistry, Tartu University, Ravila 14a, 50411 Tartu, Estonia
e-mail: juri.liiv@ut.ee

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_123

not emit light, the image being formed using ambient light, similar to conventional printed paper.

The authors of this work have developed a method for manufacturing the particles of polyvinylidene fluoride (PVDF), which is an electret material with extremely high residual electric field [3]. An electret material is a stable dielectric with a permanently embedded static electric charge, which, owing to the high resistance of the material, will not decay for hundreds of years. Each ball has a monopolar electrical charge and bipolar charge. When the control voltage is constant or zero, the ball is glued to the wall of the cavity due to the electrostatic forces [1]. The ball begins to move towards the opposite wall of the cavity, when the polarity of the control voltage changes. Microscopic asymmetries and other small perturbations cause a deviation of the axis of the electrical dipole from the direction of the electrical field and electrostatic torque causes the ball to rotate.

This paper describes the numerical experiments with a simplified mathematical model of the cell of display and provides the opportunity to determine the performance of the display depending on the physical parameters of the balls[4].

The system of equations was solved using MATLAB solvers with variable time step using step-wise integration. The development of display performance function describing the dependence between luminance and rotation time in cases of different physical parameters is planned as future work.

2 Mathematical Model of the Translation of the Ball

After applying electrostatic force, the ball inside the cavity accelerates at first and reaches a stable velocity determined by the diameter of the ball and the viscosity of the carrier liquid. The equilibrium state of a ball in the cavity is at the cavity wall. If the control voltage changes polarity, the ball begins to move towards the opposite wall of the cavity. Neglect the influence of gravity and buoyant force because of their smallness in comparison with the electrostatic force and viscous drag. The shift $y = y(t)$ of the ball is described by the differential equation:

$$\frac{d^2y}{dt^2} - \frac{F_E - F_L}{m} = 0\,. \tag{1}$$

Here, F_E is an electrostatic force and F_L is a viscous drag, m is the mass of the particle. After simplifications and using Stokes' law to determine the resistance of the fluid, we obtain resulting differential equation

$$\frac{d^2y}{dt^2} + \frac{9 \cdot \eta}{2 \cdot \rho \cdot r^2} \cdot \frac{dy}{dt} - \frac{3 \cdot U \cdot q}{4 \cdot \pi \cdot s \cdot \rho \cdot r^3} = 0\,. \tag{2}$$

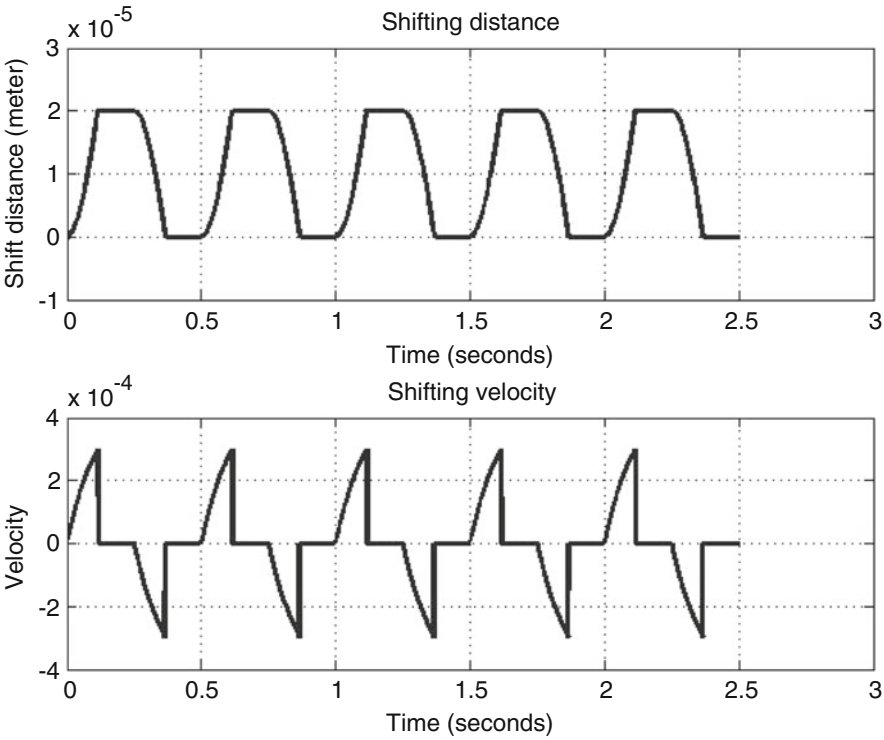


Fig. 1 Movement of the ball in cavity; $q = 6 \times 10^{-16}$ C

q is the monopolar charge of the particle; the other notations and parameters in simulations here and below are the following: the median radius $r = 25 \times 10^{-6}$ m, thickness of the film $R_D = 150 \times 10^{-6}$ m, diameter of the cavity is 70×10^{-6} m, control voltage $U_C = 100$ V, the density of PVDF given by $\rho = 1.75 \times 10^3$ kg/m^3, the viscosity of carrier liquid $\eta = 3 \times 10^{-4}$ N $\times$ s/m^2. The simulated trajectory of the ball governed by the differential equation (2) is shown on Fig. 1. The initial conditions are *y(0) = y'(0) = 0*. On Fig. 1 we see the equilibrium states $y = 0$ and 20×10^{-6} m of the ball shift which are determined by the cavity walls.

3 Mathematical Model of the Rotation of the Ball

Polarized bichromal ball is embedded into a cell filled with dielectric fluid. Rotation of the ball inside the cell cavity as a result of the outer electrical field can be

described through the balance relation:

$$M_J + M_S + M_E = 0. \tag{3}$$

Here M_J is the inertial torque, M_S is the viscous torque and M_E is the electrostatic torque. After transformations, we get the resulting differential equation which describes the rotation $\phi = \phi(t)$ of the ball:

$$-\frac{8}{15} \cdot \pi \rho r^5 \cdot \frac{d^2\phi}{dt^2} - \frac{8}{3} \cdot \pi \eta r^4 \cdot \frac{d\phi}{dt} + \frac{r \cdot q_D \cdot U \cdot \sin\phi}{s} = 0\,. \tag{4}$$

In addition to the notations introduced in previous section, here q_D denotes dipole charge. The magnitude of rotation depends on several parameters. Figure 2 shows a sample movement of a ball free of the cavity. We see that the ball reaches equilibrium state after some periods of damped oscillation. If the ball is in the cavity, physical boundaries are applied. When the polarity of the control voltage changes, the ball begins to move towards the opposite wall of the cavity. The rotation stops

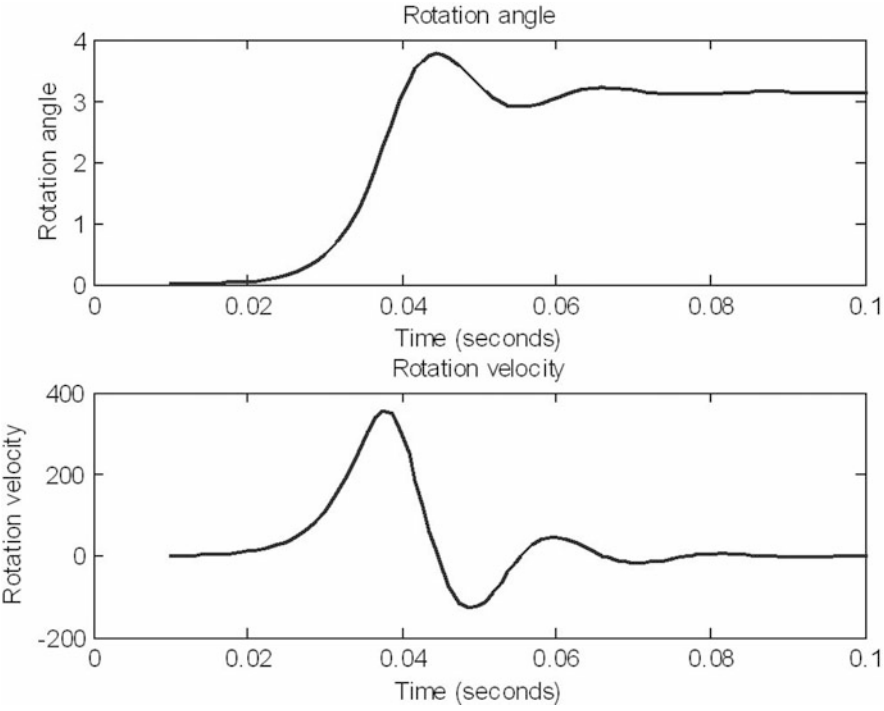


Fig. 2 Sample plot of ball rotation

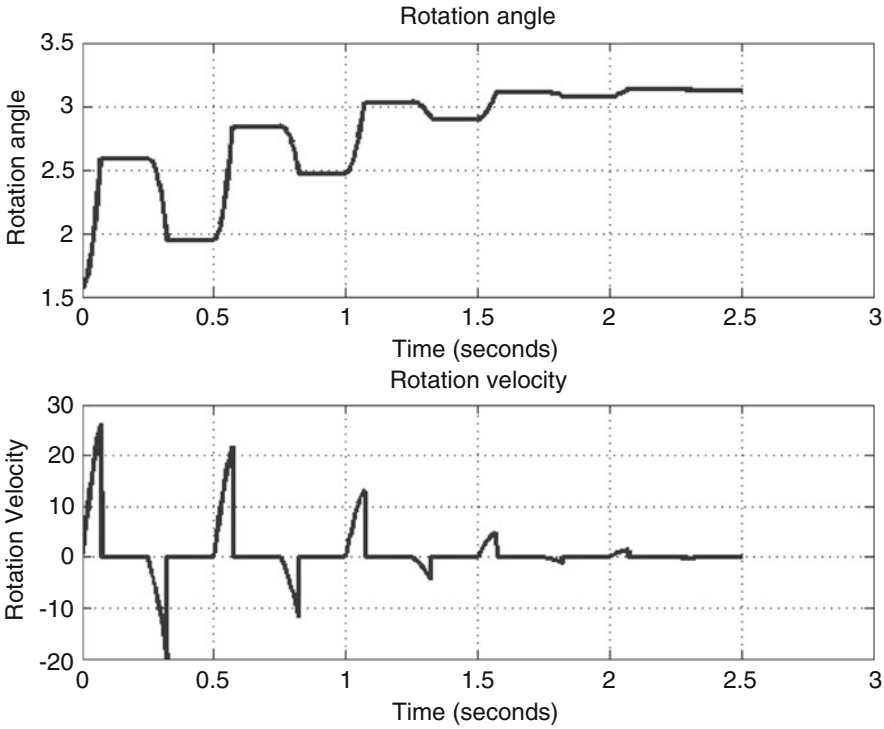


Fig. 3 Rotation angle and velocity of the ball in cavity for $q_D = 1 \times 10^{-18}$ C

at the moment when the ball reaches opposite wall and ball does not obtain the equilibrium state as we saw on Fig. 2. The changes of the angle and velocity of ball rotation corresponding to this situation with different electrical parameters of the ball are shown on Figs. 3, 4 and 5. On Fig. 3 we see that in the case of small dipole charge the ball performs some incomplete rotation cycles and stops in some fixed angle.

On Fig. 4 the ball has nearly optimal dipole charge, other parameters remain unchanged. We see that in this case the ball rotation intends towards some complete rotation cycle. If control voltage changes, the ball rotates nearly 180° and exposes right, black or white size to the observer. Display works as expected.

On Fig. 5 we see that when the dipole charge is too big, over some critical value, then the uniform rotation cycles of the ball are lost and ball performs random rotations and stops in unpredictable states.

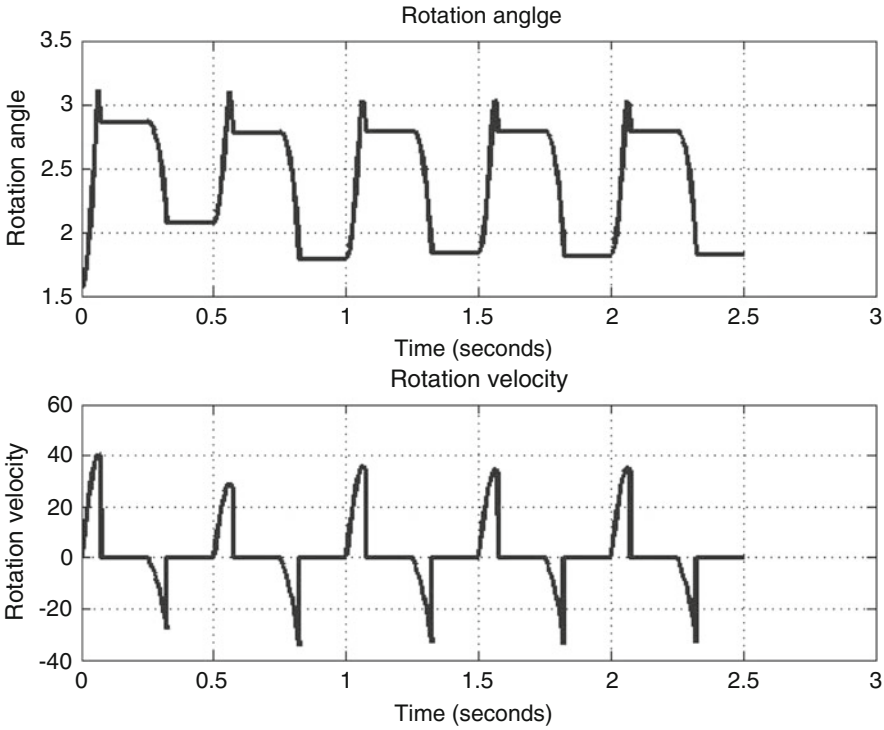


Fig. 4 Rotation angle and velocity of the ball in cavity for $q_D = 2 \times 10^{-18}$ C

4 Conclusions

During numerous laboratory experiments, we discovered unstable behavior of the experimental display for different dipole and monopole charges of balls. In some cases the display worked as expected but in some cases we noticed that the particles acquired random states, not the "white" and "black" as presumed and in other cases the display stopped working at all after some time. Notice that, in the case of unpropitious combination of physical parameters, the rotation of the ball is strictly limited and the final rotation angle is highly undetermined. These situations correspond to the simulations presented on Figs. 3 and 5.

We can conclude that the mathematical model presented in this paper corresponds to proper operation of the display. In our experiments, the dipole charge can be changed by changing the polarization parameters and the monopolar charge can be simply controlled using nonpolar surfactants dissolved in the carrier liquid [2].The proper operation of the display can be achieved only using the strictly predetermined combination of physical characteristics of the particles. Integration of the behavior of the whole electrophoretic display is planned as future work.

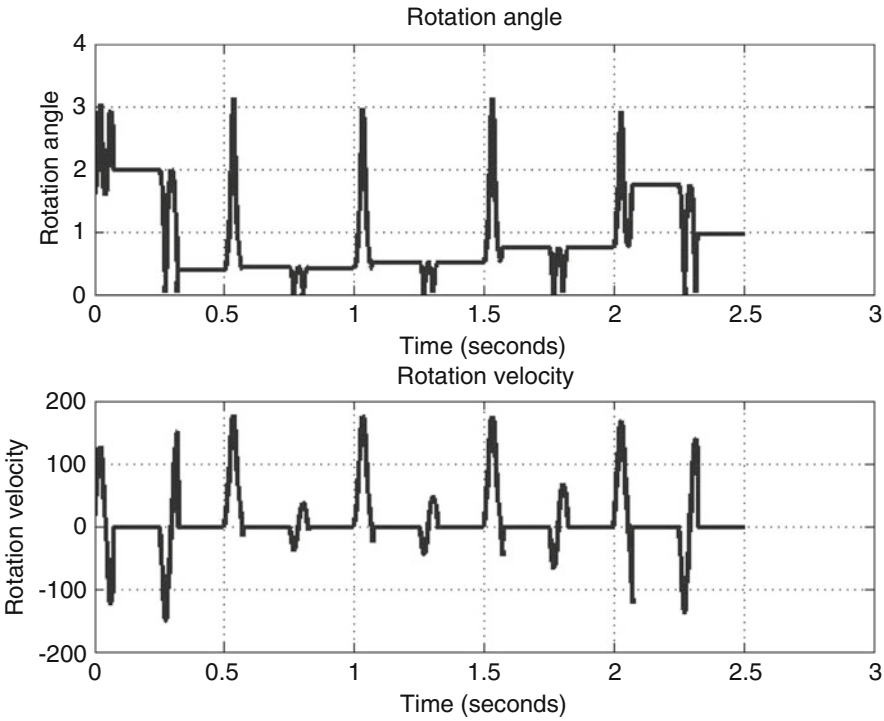


Fig. 5 Rotation angle and velocity of the ball in cavity for $q_D = 1 \times 10^{-16}$ C

Acknowledgements This work was partially supported by Estonian Institutional Research Project IUT20-57 and by FP7-SME-2010-1-286933 Project "E-SIGNAGE. Electronic paper message board for outdoor use with carbon NanoBud display module and GPRS I/O layer".

References

1. Crowley, J.M., Sheridon, N.K., Romano, L.: Dipole moments of gyricon balls. J. Electrost. **55**(3–4), 247–259 (2002). doi: 10.1016/S0304-3886(01)00208-X. http://linkinghub.elsevier.com/retrieve/pii/S030438860100208X
2. Karvar, M., Strubbe, F., Beunis, F., Kemp, R., Smith, A., Goulding, M., Neyts, K.: Transport of charged aerosol OT inverse micelles in nonpolar liquids. Langmuir ACS J. Surf. Colloids **27**(17), 10,386–10,391 (2011)
3. Liiv, J.: Active optical element, method of producing the same. US Patent 8,383,010, 2013. https://www.google.com/patents/US8383010
4. Liiv, J., Mashirin, A., Tenno, T., Miidla, P.: Mathematical modelling of the performance of a twisting-ball display. In: Proceedings of the Estonian Academy of Sciences. Estonian Academy Publishers, Tallinn (2014). doi: 10.3176/proc.2014.4.05
5. Sheridon, N.: Twisting ball panel display. US Patent 4,126,854, 1978. https://www.google.com/patents/US4126854

6. Sheridon, N.K.: Gyricon materials for flexible displays. In: Crawford, G.P. (ed.) Flexible Flat Panel Displays, Chap. 20. Wiley, New York (2005). doi:10.1002/0470870508.ch20. http://onlinelibrary.wiley.com/doi/10.1002/0470870508.ch20/summary
7. Sheridon, N., Richley, E.: The Gyricon rotating ball display. ...information display (2012). http://onlinelibrary.wiley.com/doi/10.1889/1.1985284/abstract

SPH for the Simulation of a Dam-Break with Floating Objects

Giuseppe Bilotta, Alexander Vorobyev, Alexis Hérault, Damien Violeau, and Ciro Del Negro

Abstract We show an application of the Smoothed Particle Hydrodynamics (SPH) method to the simulation of a fully three-dimensional dam-break with floating objects. The simulation is done using GPUSPH, an implementation of the SPH method in CUDA which has been recently extended including support for fully coupled fluid/solid interaction. Boundary conditions are computed using the unified semi-analytical model proposed by Ferrand et al. SPH is also used to compute the total force and torque acting on the floating objects, which are then used to integrate the motion of the objects.

Keywords Floating object • Meshfree method • Smoothed particle hydrodynamics method

1 Introduction

Applications of SPH to real-world problems depend on the correct evaluation of boundary forces, particularly when the domain has a complex shape, or when fluid/object interactions are to be modeled. Additionally, since SPH is a Lagrangian meshless method, implementations can be subject to numerical instabilities when the ratio of the domain size to the resolution is close to machine epsilon.

In this paper we show the approach used to solve these issues in an application of SPH to a dam-break with floating objects on a natural topography. Specifically, we rely on the unified semi-analytical boundary model proposed by Ferrand et al. [2, 8], and on the homogeneous accuracy work by Hérault et al. [4] to solve the numerical precision issues.

G. Bilotta (✉) • A. Vorobyev • C.D. Negro
INGV, Sezione di Catania, Catania, Italy
e-mail: giuseppe.bilotta@ct.ingv.it

A. Hérault
Conservatoire National des Arts et Métiers, Paris, France
INGV, Sezione di Catania, Catania, Italy

D. Violeau
Laboratoire d'Hydraulique Saint-Venant, EDF R&D, Chatou cedex, France

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_124

The simulations are done with GPUSPH [3], an implementation of 3-D SPH using CUDA-enabled GPUs to achieve high computing performance. GPUSPH has been recently extended to include the unified semi-analytical boundary model, and support for floating objects and their fully coupled interactions with the fluid.

We will first briefly summarize the key aspect of the unified semi-analytical boundary model (Sect. 2), and introduce the key aspects of homogeneous precision (Sect. 3), followed by a synthetic description of our approach to the modeling of fluid/solid interaction, which are the theoretical foundations of the simulation of a dam-break with floating objects (debris flow on a spillway, Sect. 5). Conclusions are then drawn in Sect. 6.

2 Unified Semi-analytical Boundary Conditions

The unified semi-analytical boundary model for SPH builds on the standard weakly-compressible SPH formulation (WCSPH), [2, 6, 8]. Following the standard notation, we will denote by $W(\cdot, h)$ a family of smoothing kernels parametrized by the influence radius h and such that $\lim_{h\to 0} W(\cdot, h) = \delta(\cdot)$, where δ is Dirac's delta and the limit is intended in the sense of distributions. The smoothing kernels are assumed to have compact support and to be positive and with radial symmetry. The fluid body is discretized by a set of particles with average inter-particle distance Δp and we will denoted with $\mathbf{r}_a$ the position of particle a, while $\mathbf{r}_{ab} = \mathbf{r}_b - \mathbf{r}_a$ denotes the distance vector between the particles, r_{ab} its norm and $W_{ab} = W(r_{ab}, h)$.

For the physical properties we will use the usual subscripted convention, with ρ_a being the density of particle a, m_a its mass, V_a its volume, P_a its pressure and $\mathbf{u}_a$ its velocity. Additionally we will denote by $\mathscr{F}$ the set of all fluid particles, and by $\mathscr{S}$ the set of all boundary particles.

The flow is assumed weakly compressible with an equation of state (EOS) $P = B((\rho/\rho_0)^\zeta - 1)$ where $\zeta = 7$ is the polytropic constant, ρ_0 the at-rest density of the fluid and $B = \rho_0 c_0^2/\zeta$, c_0 being the numerical speed of sound of the fluid.

2.1 Renormalization Terms

The unified semi-analytical boundary model was first introduced by Kulasegaram et al. [5], suggesting a renormalization for the SPH smoothing kernel near a solid wall. In such a case, the standard WCSPH density summation $\rho_a \simeq \sum_{b\in\mathscr{F}} m_b W_{ab}$ is replaced by

$$\rho_a = \frac{1}{\gamma_a} \sum_{b\in\mathscr{F}} m_b W_{ab}. \tag{1}$$

The renormalization factor γ is defined as

$$\gamma_a \equiv \int_{\Omega \cap \Omega_a} W(|\mathbf{r} - \mathbf{r}_a|) dV, \tag{2}$$

where Ω denotes the fluid domain, Ω_a the support of W centered on particle a, and dV the infinitesimal volume element. Informally, γ_a measures the ratio of the domain volume inside the kernel support to the kernel support itself, and the normalization condition of the smoothing kernel ensures that $\gamma_a = 1$ when the kernel support is fully contained in the fluid domain.

The gradient of γ_a, which appears in the momentum and continuity equations as shown in Sect. 2.2, can be computed analytically as

$$\nabla \gamma_a \equiv \int_{\Omega \cap \Omega_a} \nabla_a W(|\mathbf{r} - \mathbf{r}_a|) dV = \int_{\partial(\Omega \cap \Omega_a)} W(|\mathbf{r} - \mathbf{r}_a|) \mathbf{n} dS, \tag{3}$$

where $\mathbf{n}$ denotes the outer unit normal to the boundary and dS the infinitesimal surface element.

The values of both γ_a and $\nabla \gamma_a$ can only be computed analytical in case of trivial geometries. For more complex cases, Ferrand et al. [2] propose to discretize the boundary into small elements s (segments in the two-dimensional case, extended to triangles in the three-dimensional case [8]) and approximate the integrals by the summation of contributions from each boundary elements located within the influence domain of the particle: $\nabla \gamma_a \simeq \sum_{s \in \mathscr{S}} \nabla \gamma_{as}$.

The formulation by Ferrand et al. also considers vertex particles, located at the vertices of the boundary elements. These particles may be considered as (non-moving) fluid particles attached to the solid wall: as such their mass is a fraction of the mass of standard fluid particles, and depends on the geometry of the wall. For example, a vertex particle for a plane wall would have half the mass of a fluid particle, while at a right corner it would have one fourth of the mass.

In what follows $\mathscr{F}$ will include both moving fluid particles and vertex particles, and we will denote the set of vertex particles by $\mathscr{E}$.

2.2 Wall-Corrected Differential Operators

With kernel renormalization, the pressure gradient approximation takes the form:

$$\nabla_a P \simeq \frac{\rho_a}{\gamma_a} \sum_{b \in \mathscr{F}} \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2} \right) m_b \nabla W_{ab} - \frac{\rho_a}{\gamma_a} \sum_{s \in \mathscr{S}} \left(\frac{P_a}{\rho_a^2} + \frac{P_s}{\rho_s^2} \right) \rho_s \nabla \gamma_{as}. \tag{4}$$

Compared to the classical WCSPH approximation of ∇P, this formulation includes a contribution from the gradient of the renormalization factor, computed as a summation over the neighboring boundary elements. Vertex particles also

contribute to the pressure gradient, as they are included in the first summation. Similarly, the divergence of the velocity is approximated as

$$\nabla_a \cdot \mathbf{u} \simeq -\frac{1}{\gamma_a \rho_a} \sum_{b \in \mathscr{F}} m_b \mathbf{u}_{ab} \cdot \nabla W_{ab} + \frac{1}{\gamma_a \rho_a} \sum_{s \in \mathscr{S}} \rho_s \mathbf{u}_{as} \cdot \nabla \gamma_{as}, \tag{5}$$

where $\mathbf{u}_{ab} = \mathbf{u}_a - \mathbf{u}_b$ and $\mathbf{u}_{as} = \mathbf{u}_a - \mathbf{u}_s$.

Finally, the viscous contribution to the momentum equation can be discretized as:

$$\nabla \cdot \mu \nabla \mathbf{u} \simeq \frac{1}{\gamma_a} \sum_{b \in \mathscr{F}} m_b \frac{\mu_a + \mu_b}{\rho_a \rho_b} \frac{\mathbf{u}_{ab}}{r_{ab}^2} \mathbf{r}_{ab} \cdot \nabla W_{ab} - \frac{1}{\gamma_a \rho_a} \sum_{s \in \mathscr{S}} |\nabla \gamma_{as}| (\mu_a \nabla \mathbf{u}_a + \mu_s \nabla \mathbf{u}_s) \cdot \mathbf{n}_s. \tag{6}$$

The second term is particularly important to correctly model the flow near the boundary. For laminar flow, the term can be rewritten in the form

$$\frac{1}{\gamma_a \rho_a} \sum_{s \in \mathscr{S}} |\nabla \gamma_{as}| \frac{\mu_a + \mu_s}{\delta r_{as}} \mathbf{u}_{\tau a}, \tag{7}$$

where $\mathbf{u}_{\tau a}$ represents the wall-tangential component of the velocity of particle a, computed as $\mathbf{u}_{\tau a} = \mathbf{u}_{as} - (\mathbf{u}_{as} \cdot \mathbf{n})\mathbf{n}$, and δr_{as} is the clipped distance between the particle and the wall, $\delta r_{as} = \max(\Delta p, \mathbf{r}_{as} \cdot n)$, where the lower limit of Δp is imposed to prevent zero values on the denominator.

2.3 Ferrari Correction

Mayhrofer et al. [7] adapted Ferrari et al.'s density correction [1] to the semi-analytical boundary model. The discretized continuity equation then takes the form:

$$\frac{D\rho_a}{dt} = -\frac{\rho_a}{\gamma_a} \sum_{b \in \mathscr{F}} V_b \left(\mathbf{u}_{ba} + K \frac{c_{ab}}{\rho_a} \bar{\rho}_{ab} \frac{\mathbf{r}_{ab}}{r_{ab}} \right) \cdot \nabla_a W_{ab} \frac{\rho_a}{\gamma_a} \sum_{s \in \mathscr{S}} \mathbf{u}_{sa} \cdot \nabla \gamma_{as}, \tag{8}$$

where K is a coefficient to be chosen between 0 and 1, $c_{ab} = \max\{c_a, c_b\}$ is the maximum speed of sound and $\bar{\rho}_{ab}$ is a corrected density difference that takes into account the hydrostatic pressure balance by inverting the EOS:

$$\bar{\rho}_{ab} = \rho_a - \rho_b - \left(\frac{P_a - P_b \rho + \mathbf{F} \cdot \mathbf{r}_{ab}}{B} + 1 \right)^{1/\zeta}, \tag{9}$$

where $\mathbf{F}$ are the external forces (typically, gravity).

2.4 Boundary Pressure and Density

The pressure and density of vertex particles is computed imposing $\partial P/\partial n = 0$ at the boundary, which gives us the following SPH interpolation formulas for the density and pressure of a vertex particle e:

$$\rho_e = \frac{1}{\alpha_e} \sum_{b \in \mathscr{F} \setminus \mathscr{E}} V_b \rho_b W_{be}, \tag{10}$$

$$\frac{P_e}{\rho_e} = \frac{1}{\alpha_e} \sum_{b \in \mathscr{F} \setminus \mathscr{E}} V_b \left(\frac{P_b}{\rho_b} - \mathbf{F} \cdot \mathbf{r}_{be} + \frac{u_b^2 - u_e^2}{2} \right) W_{be}, \tag{11}$$

where $\alpha_e \equiv \sum_{b \in \mathscr{F} \setminus \mathscr{E}} V_b W_{be}$ is the renormalization factor of the SPH interpolation. We remark that the summations here are extended to neighboring *fluid* particles only. For boundary elements, the pressure and density are obtained by averaging the three adjacent vertex particles.

Finally, since P_e, ρ_e computed as above are inconsistent with the EOS, an iterative procedure relying on inverting the EOS should be used to correct their values. In our tests, a single iteration $\rho'_e = \rho_0(P/B+1)^{1/\xi}$ has shown to be sufficient.

3 Homogeneous Accuracy

Naive implementations of SPH store particle positions with respect to a global reference frame, even though global positions are never actually used in SPH formulas. As a result of this choice, the accuracy of the evaluation of the distance vector between particles ($\mathbf{r}_{ab}$) and all the related quantities is higher closer to the origin, and becomes lower as the particles move away from the origin.

On the other hand, to make neighbor search computationally efficient, many SPH implementations rely on an auxiliary domain grid with spacing not smaller than the influence radius of the smoothing kernel, which ensures that if particle a is located in cell $\mathbf{C}_a = (X_a, Y_a, Z_a)$, its neighbors are located in the cells $(X_a \pm 1, Y_a \pm 1, Z_a \pm 1)$.

While this auxiliary grid has only been used until recently to speed up the neighbor search, it can be used to provide homogeneous accuracy throughout the domain [4]. To achieve this, in GPUSPH the position of each particle is stored in terms of the distance $\mathbf{f}_a$ to the center of the cell it belongs to. The global position of particle a is never computed as such (except when setting up the problem initially, and when retrieving the particle positions for storage or visualization), and $\mathbf{r}_{ab}$ is computed as $\mathbf{f}_{ab} + \mathbf{K}$ where $\mathbf{K} = \mathbf{C}_{ab}s$ is the distance between the centers of the cells, and s the grid spacing.

This approach has several benefits. It's numerically more stable, particularly in the case of simulations over large domains with a very high resolution, and it's

computationally more efficient, since it allows storing the neighbor list purely in terms of their offset to the neighboring cell, reducing the memory consumption of the neighbor list.

4 Fluid/Solid Interaction

GPUSPH includes a fully coupled fluid/solid interaction model based on the thin shell model for rigid bodies.

Objects are described by their inertia tensor with respect to the principal axes, the location of the center of mass and the rotation of the local reference system with respect to the global reference system. The latter is described using unit quaternions, which simplifies the treatment of rotations and avoids some of the well-known problems associated with the use of Euler angles [10].

For the fluid/solid interaction, objects are discretized as a thin shell of boundary particles, which interact with the fluid using the standard fluid/boundary interaction model. The interaction is used to compute the forces acting on each boundary particle, which are used to derive the total force and torque acting on the object.

The object motion is then integrated, taking interaction with other objects and with the topography into account, using the Open Dynamics Engine [9]. The object motion is then used to derive the new position and velocities for the object particles for the new integration step.

5 Goulours Spillway Debris Flow Test Case

Our sample application is a dam break on a natural topography with floating objects. The dam break is simulated on the topography of the Goulours dam spillway, provided by Électricté de France (EDF). Five trees are placed in the reservoir before the water spill. The trees are modeled as cylinders 8 m tall and with a 1 m diameter. Their density is $\rho_t = 0.7\rho_w$, where $\rho_w = 1000\,\mathrm{kg/m^3}$ is the water density. Boundaries are described using the semi-analytical boundary conditions.

Lacking validation data, the main purpose of this simulation is provide a proof of concept, to show the possibility to simulate fluid/solid interaction with GPUSPH, and its application to the modeling of hydraulic waterworks. We can only provide a qualitative evaluation of the simulation, highlighting some crucial aspects in the interaction between the fluids, the floating objects and the irregular topography.

The first highlight is given by the interaction between the first tree that enters the spillway, and the high velocity flow at the beginning of the dam break. The simulation (Fig. 1) shows the sinking/floating of the tree dragged by the flow.

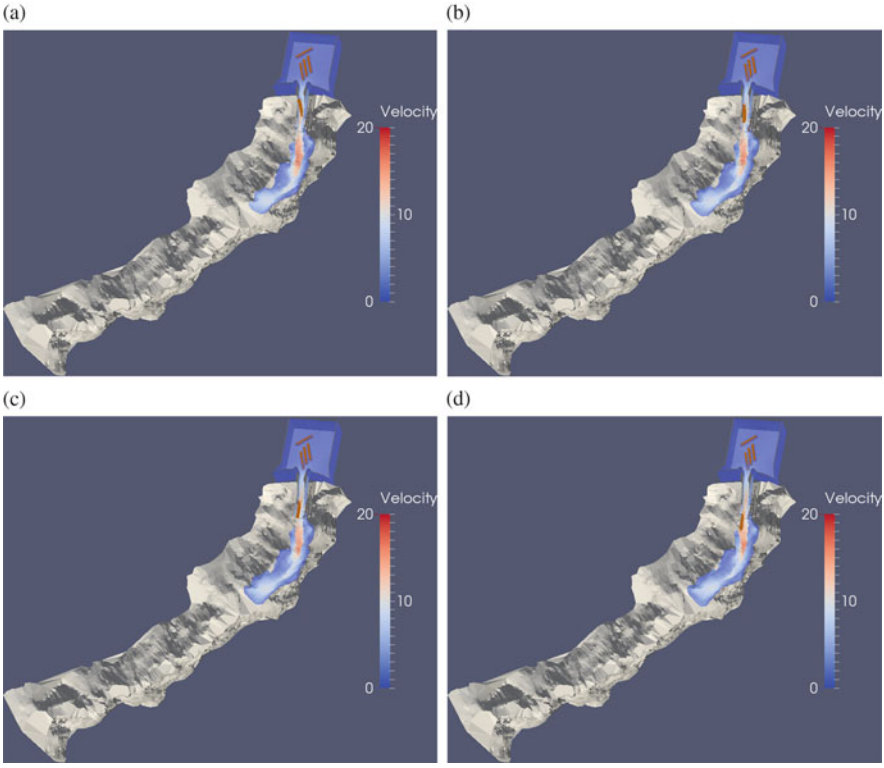


Fig. 1 Zoom-in of the first part of the Goulours spillway showing the interaction of the first tree with the initial water spill. (**a**) $t = 8.5$ s. (**b**) $t = 9$ s. (**c**) $t = 9.5$ s. (**d**) $t = 10$ s

The second highlight is given by a later stage of the simulation, where the first tree has settled down around a bend further down the spillway, and two more trees are able to catch up with it (Fig. 2), potentially causing an obstruction in the spillway.

6 Conclusions

We have shown a possible applications of SPH in the modeling and simulation of hydraulic waterworks. The GPUSPH implementation of SPH in CUDA is used in the examples presented. The results obtained with the use of the unified semi-analytical boundary conditions, in conjunction with the model for fluid/solid interaction present in GPUSPH, show qualitatively good results in the simulation of fully coupled fluid/solid interaction with the flow, in a proof-of-concept application to the simulation of debris flow during a dam-break on a natural topography. The application illustrates the use of numerical modeling with SPH to highlight potential

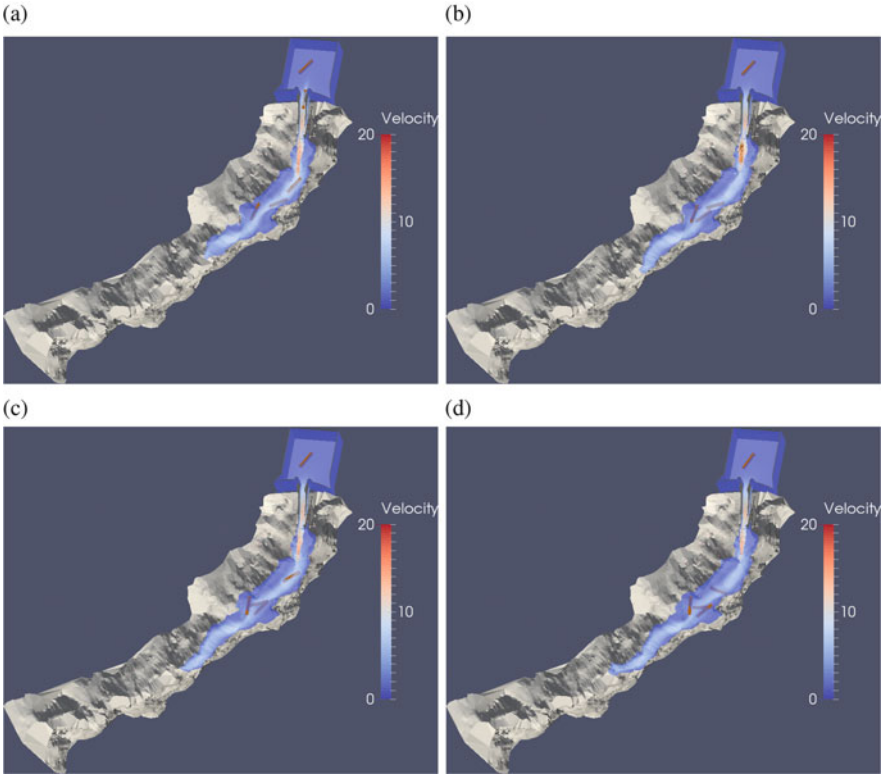


Fig. 2 Two trees reach up with the first tree causing a potential obstruction in the Goulours spillway. (**a**) $t = 20$s (**b**) $t = 22.5$ s (**c**) $t = 25$ s. (**d**) $t = 27.5$ s

obstructions to the waterway and related situations. Future work in this regard will be aimed at the validation of the fluid/solid interaction, and the refinement of the unified semi-analytical boundary conditions by improving the computation of γ.

References

1. Ferrari, A., Dumbser M., Toro E.F., Armanini A.: A new 3D parallel SPH scheme for free surface flows. Comput. Fluids **38**, 1203–1217 (2009)
2. Ferrand, M., Laurence D., Rogers, B.D., Violeau, D., Kassiotis, C.: Unified semi-analytical wall boundary conditions for inviscid, laminar or turbulent flows in the meshless SPH method. Int. J. Numer. Methods Fluids **71**(4), 446–472 (2012)
3. Hérault, A., Bilotta G., Dalrymple R.A.: SPH on GPU with CUDA. J. Hydraul. Res. **48**, 74–79 (2010)
4. Hérault, A., Bilotta, G., Dalrymple, R.A.: Achieving the best accuracy in an SPH implementation. In: Proceedings of the 9th SPHERIC International Workshop (2014)

5. Kulasegaram, S., Bonet, J., Lewis, R.W., Profit, M.: A variational formulation based contact algorithm for rigid boundaries in two-dimensional SPH applications. Comput. Mech. **33**, 316–325 (2004). doi:10.1007/s00466-003-0534-0
6. Leroy, A., Violeau, D., Ferrand, M., Kassiotis, C.: Unified semi-analytical wall boundary conditions applied to 2-D incompressible SPH. J. Comput. Phys. **261**:106–129 (2014)
7. Mayrhofer, A., Rogers B.D., Violeau D., Ferrand, M.: Investigation of wall bounded flows using SPH and the unified semi-analytical wall boundary conditions. Comput. Phys. Commun. **184**(11), 2515–2527 (2013)
8. Mayrhofer, A., Ferrand, M., Kassiotis, C., Violeau, D., Morel, F.-X.: Unified semi-analytical wall boundary conditions in SPH: analytical extension to 3-D. Num. Alg. (2014). doi:10.1007/s11075-014-9835-y
9. Smith, R.: Open dynamics engine. IGC'02 (2002)
10. Vicci, L.: Quaternions and rotations in 3-Space: the algebra and its geometric interpretation. TR01-014, Department of Computer Science, UNC Chapel Hill (2001)

MS 36
MINISYMPOSIUM: SPACETIME MODELS OF GRAVITY IN GEOLOCATION AND ACOUSTICS

Organizers

Jose M. Gambi[1], Michael M. Tung[2] and Manuel Carretero[3]

Speakers

Michael M. Tung[2], Jose M. Gambi[1] and Maria L. Garcia del Pino[4]
Maxwell's Fish-Eye in (2+1)D Spacetime Acoustics

Jose M. Gambi[1], Michael M. Tung[2], Javier Clares[5] and Maria L. Garcia del Pino[4]
Post-Newtonian Effects in Geolocation by FDOA

Jose M. Gambi[1], Javier Clares[5] and Maria C. Rodriguez-Teijeiro[6]
Post-Newtonian Geolocation of Passive Radio Transmitters by TDOA and FDOA

Jose M. Gambi[1], Maria L. Garcia del Pino[4] and Michael M. Tung[2]
Post-Newtonian Orbital Equations for Fermi Frames in the Vicinity of the Earth

[1]Jose M. Gambi, Universidad Carlos III de Madrid, Leganes, Spain.
[2]Michael M. Tung, Universidad Politecnica de Valencia, Valencia, Spain.
[3]Manuel Carretero, Universidad Carlos III de Madrid, Leganes, Spain.
[4]Maria L. Garcia del Pino, Universidad Carlos III de Madrid, Leganes, Spain.
[5]Javier Clares, Universidad Carlos III de Madrid, Leganes, Spain.
[6]Maria C. Rodriguez-Teijeiro, Universidad Carlos III de Madrid, Leganes, Spain.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_125

Keywords

Acoustics
Geolocation

Short Description

The geometrization of gravity forms one of the cornerstones of modern science having an impact on the industrial progress connected to many activities of daily life. In the past decade, substantial research has been invested into post-Newtonian corrections for high-precision navigation, geolocation, and tracking devices, as well as into the design of analogue models of gravity by making use of advanced optical and acoustic metamaterials. Present industrial needs demand innovative development of computationally efficient spacetime models in this interdisciplinary field. Such mathematical models become important for applications requiring very accurate timing as achieved in today's engineering systems which rely on modern atomic clocks. Moreover, acoustic metamaterials—artificially produced with properties not found in nature—challenge the engineer to fabricate acoustic devices with highly unusual features.

Maxwell's Fish-Eye in (2+1)D Spacetime Acoustics

M.M. Tung, J.M. Gambi, and M.L. García del Pino

Abstract In the past few years Maxwell's fish-eye lens has been subject to intense investigation in the context of transformation optics, mainly spurred by the possibility to create perfect imaging without the need to resort to negative refraction, one of the outstanding—but difficult to implement—properties of metamaterials. Here we extend this discussion to an acoustical fish-eye constructed in (2+1)D spacetime. The underlying acoustic wave is governed by a homogeneous spherical Helmholtz equation, which is shown to emerge from a variational principle in inherently covariant manner. The formal analytical solutions of the acoustic potential are derived.

Keywords Acoustic • Helmholtz equation • Maxwell's fish-eye

1 Introduction

One of the central objectives of metamaterial research is first the theoretical conception and then the industrial engineering of artificial materials with remarkable properties which are not found in nature. The clever design of so-far unknown advanced metamaterial devices with useful applications in all sectors of living is its prime purpose. The engineering of acoustic metamaterial devices falls in this category [9] and also serves well to demonstrate the power of the underlying differential-geometric framework in this analogue model of gravity [16]. Apart from perfect acoustic lenses, it is worthwhile mentioning that other industrial applications in this field cover the acoustical improvement of concert halls, the construction

M.M. Tung (✉)
Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera, s/n, E-46022 Valencia, Spain
e-mail: mtung@mat.upv.es

J.M. Gambi • M.L. García del Pino
Gregorio Millán Institute, Universidad Carlos III de Madrid, Avenida de la Universidad 30, E-28911 Leganés, Madrid, Spain
e-mail: gambi@math.uc3m.es; lgarciadelpino@educa.madrid.org

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_126

of ships and submarines invisible to sonar detection, and much more. Interesting applications of acoustic cloaking may be found in [2, 3, 11].

Maxwell's fish-eye lens is the special case of a so-called Lüneburg lens [7, pp. 172–182], which is a spherically symmetric lens characterized by a variable refractive index in its interior. More precisely, the fish-eye is a positively refracting metamedium which is implemented by the stereographic projection of a hypersphere to a plane passing through its origin. In the remainder of this section we will explain what this means in the 2D example, referring to purely spatial dimensions first.

Hence, we shall consider the conformal mapping of the 2-sphere with radius $a > 0$ to the xy-plane, namely $S^2 \to \mathbb{R}^2$ conveniently defined by the following stereographic coordinate transformation with parametrization:

$$\mathbf{r}(x,y) = \begin{pmatrix} \xi \\ \eta \\ \zeta \end{pmatrix} = a \begin{pmatrix} \dfrac{2x/a}{(x/a)^2+(y/a)^2+1} \\ \dfrac{2y/a}{(x/a)^2+(y/a)^2+1} \\ \dfrac{(x/a)^2+(y/a)^2-1}{(x/a)^2+(y/a)^2+1} \end{pmatrix}. \tag{1}$$

Here (ξ, η, ζ) are the auxiliary coordinates on S^2, and (x, y) are the physical coordinates in the projected plane. It is easy to check that $\|\mathbf{r}\| = a$ for all $(\xi, \eta, \zeta) \in S^2$ and that the equator lies in the xy-plane, also with radius a. A straightforward calculation yields for the 2D-metric with coordinates $(x^1, x^2) \equiv (x, y)$ the isotropic result

$$g_{ij} = \left(\frac{2}{(\rho/a)^2+1}\right)^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{with} \quad \rho^2 = x^2 + y^2. \tag{2}$$

Note that Eq. (2) represents a curved space with the non-vanishing Ricci tensor components $R_{11} = R_{22} = 4a^2/(\rho^2 + a^2)^2$ giving the scalar curvature $R = 2/a^2$, which is identical to the 2D-sphere. Comparing this with the acoustic metric of an isotropic metamedium,

$$g_{ij} = n^2(\rho) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{3}$$

immediately provides the positive refractive index

$$n(\rho) = \frac{2}{(\rho/a)^2+1}. \tag{4}$$

See also [5, 6] for different derivations of Eq. (4)

For further derivations we will also require the inverse transformation of Eq. (1) given by

$$x = \frac{\xi}{a-\zeta}\, a, \qquad y = \frac{\eta}{a-\zeta}\, a. \tag{5}$$

It thus becomes obvious that the azimuthal angle, φ, for the xy-plane and for the sphere S^2 is the same:

$$\tan\varphi = \frac{y}{x} = \frac{\eta}{\xi}. \tag{6}$$

Furthermore, by using the standard representation in spherical coordinates with the polar angle ϑ, the radial coordinate is calculated by

$$\rho = \frac{\sqrt{\xi^2+\eta^2}}{a-\zeta}\, a = a\, \cot\left(\vartheta/2\right). \tag{7}$$

We are now in the position to extend the previous results to the covariant formalism in curved spacetime and will proceed in the next section with an exploration of Maxwell's fish-eye in this full spacetime framework.

Section 2.1 first introduces the variational principle in spacetime for the scalar acoustic potential, ϕ, and its associated wave propagation. The proposed fundamental Lagrangian density function on the spacetime manifold allows to derive the equations of motion—related to the homogeneous spherical Helmholtz equation—and thus completely predicts the evolution of non-dissipative acoustic phenomena.

Next, in Sect. 2.2 we are able to move on to the problem at hand, namely the fish-eye lens in (2+1)D spacetime and the general relationship between the constitutive parameters of the virtual and physical acoustic metafluid. These are needed to engineer the metamaterial with fish-eye properties. Finally, we derive formal analytic solutions for the acoustic potential with a fish-eye spacetime metric.

2 Results and Discussion

2.1 Variational Principle for Spacetime Acoustics

Previously, we have introduced a Lagrangian framework to describe macroscopic electrodynamic phenomena within transformation optics [4] and also tackled diffusion on curved manifolds [10, 12, 13]. In [11, 14] we have focussed on the Lagrangian framework which describes acoustic phenomena. Since the advent of special relativity it is known that Maxwell's equations for electrodynamics are inherently covariant, an advantage which is absent in acoustics. Even so, the acoustic theory only requires a scalar potential $\phi : M \to \mathbb{R}$ (in contrast to the vector

potential of electrodynamics), which simplifies matters a bit. As usual, M denotes a smooth spacetime manifold endowed with a Lorentzian metric $\mathbf{g}$, where we assume a positive signature.

For the reformulation of Hamilton's variational principle in acoustics within fully covariant spacetime and for finding the extremal solution of the corresponding action functional, we require a Lagrangian density function of the type

$$\mathscr{L} : M \times TP \to \mathbb{R}, \tag{8}$$

where P the ambient space defined by the acoustic potential. If $N = M \times P$ is the Riemannian manifold representing configuration space, then the jet bundle $J^1N = M \times TP$ indicates that the Lagrangian $\mathscr{L}$ generally is a scalar function of x^μ, ϕ, and $\phi_{,\mu}$ and may also contain the metrics of M and N, see e.g. [1].[1] In other words, partial derivatives of configuration space with respect to all spacetime coordinates are admitted.

Observe also that in the acoustic metamaterial the local fluid velocity, $\mathbf{v}$, the acoustic pressure, p, and the density, ϱ_0, are directly obtained from the scalar potential via [8, 11]

$$v_\mu = -\phi_{;\mu} = -\phi_{,\mu} = \begin{pmatrix} -p/c\varrho_0 \\ \mathbf{v} \end{pmatrix}, \tag{9}$$

where $c > 0$ is the time-independent acoustic wave speed.

Physical constraints (energy-momentum conservation, locality) limit the acoustic Lagrangian density to take the simplest possible form [11], which only contains a kinetic term (in the spacetime sense): $\mathscr{L}(\phi_{,\mu}) = \frac{1}{2}\sqrt{-g}\, g^{\mu\nu}\phi_{,\mu}\phi_{,\nu}$. Therefore, for transformation acoustics the functional derivative of the following associated action integral must vanish [11]:

$$\frac{\delta}{\delta\phi}\int d\mathrm{vol}_g\, g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} = 0. \tag{10}$$

Integration is carried out over a bounded, closed set of spacetime ($n = 4$ for (3+1)D spacetime) and the invariant volume element is defined by $d\mathrm{vol}_g = \sqrt{-g}\, dx^0 \wedge \ldots \wedge dx^{n-1}$ with $g = \det \mathbf{g}$. The solutions of Eq. (10) are just the Euler-Lagrange equations of motion governing the dynamics of the acoustic system.

[1] It is customary to denote by Greek indices the usage of the full range of spacetime values for tensors, whereas Latin indices only run over the spatial values. Comma and semicolon are standard notation for partial and covariant derivatives, respectively.

2.2 Analytic Solutions for the Acoustic Fish-Eye in (2+1)D Spacetime

As shown before, the fish-eye metric in *physical space*, Eq. (2), represents isotropic characteristics similar to those of a 2-sphere, especially with the same scalar curvature $R = 2/a^2$. It is therefore more convenient for the study of the fish-eye metric to work in a spherical *virtual space* [6].

Rewriting the Cartesian metric, Eq. (2), in spherical coordinates for S^2 by considering Eq. (5) and including the time component, readily yields

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & a^2 & 0 \\ 0 & 0 & a^2 \sin^2\theta \end{pmatrix} \tag{11}$$

for the full (2+1)D spacetime metric of the fish-eye with coordinates $(x^0, x^1, x^2) = (ct, \vartheta, \varphi)$. It is $\sqrt{-g} = a^2 \sin\vartheta$. Note that in this simple metric no spacetime mixing occurs and time dilation is absent. For the simplest metric representing time-dilation in uniformly accelerated frames or within a uniform gravitational field we refer to [14, 15].

The metric Eq. (11) is static and its non-vanishing Christoffel symbols are thus identical to those of S^2. Consequently, the geodesics in virtual space are great circles on S^2 with its projections conformally mapped onto physical space, i.e. the xy-plane, also being circles.

Following the outline of [11], we identify the *virtual space* with flat Minkowski space. Moreover, we associate *physical space* with the space having the desirable acoustic properties of the fish-eye, realized by the metric Eq. (11). The so-called constitutive relations establish the connection between physical and virtual acoustic space and their material properties. With Eq. (19) of [11], for the fish-eye the result is

$$\kappa = \kappa_0 \sin\vartheta, \qquad \varrho_{ij} = \varrho_0 \frac{\kappa}{\kappa_0} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{12}$$

where as usual κ is the bulk modulus with a fixed scale κ_0, and ϱ_{ij} is the isotropic mass-density tensor. These relations fully determine the metamaterial properties for engineering the acoustic fish-eye.

On the other hand, the variational principle, Eq. (10), provides the following wave equation for the fish-eye metric

$$\Delta_M \phi = g^{\mu\nu}\phi_{;\mu\nu} = -\frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} + \frac{1}{a^2}\Delta_{S^2}\phi = 0, \tag{13}$$

which is a homogeneous spherical Helmholtz equation combined with a harmonic temporal contribution. For finding its solutions it is therefore natural to try the

following *ansatz* by applying the method of separation of variables:

$$\phi(t,\vartheta,\varphi) = \phi_0(t)Y(\vartheta,\varphi). \tag{14}$$

It is expected that $Y(\vartheta,\varphi)$ are the eigenfunctions of the spherical Laplacian Δ_{S^2}, i.e. spherical harmonics with eigenvalues λ, necessarily of the form $\lambda = l(l+1), l \in \mathbb{N}$. Hence, an explicit calculation yields formally the following solution for the acoustic fish-eye potential

$$\phi(t,\vartheta,\varphi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left[A_{lm}\cos\left(\frac{\sqrt{l(l+1)}}{a}ct\right) + B_{lm}\sin\left(\frac{\sqrt{l(l+1)}}{a}ct\right)\right]Y_{lm}(\vartheta,\varphi), \tag{15}$$

where

$$Y_{lm}(\vartheta,\varphi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}\,P_{lm}\big(\cos\vartheta\big)e^{im\varphi} \tag{16}$$

are the orthonormalized spherical harmonics and P_{lm} the associated Legendre polynomials. The coefficients A_{lm} and B_{lm} will depend on the precise boundary conditions under consideration and can be determined by employing the standard orthogonality relation for $Y_{lm}(\vartheta,\varphi)$.

The final transformation of $\phi(t,\vartheta,\varphi)$ back to physical space with Eqs. (6) and (7) finally yields $\phi(t,x,y)$. This completes the calculation of the 2D fish-eye potential.

3 Conclusions

The acoustic fish-eye is a particularly interesting subject of transformation physics, since it might facilitate the construction of perfect acoustic focussing devices.

In this work, we have sketched the application of a covariant variational principle to determine the acoustic potential for a metamaterial with an underlying (2+1)D fish-eye spacetime metric. The fish-eye metric is the result of the choice of physical and virtual spaces linked by a stereographic mapping of the 2-sphere S^2 to the xy-plane. It is shown that the wave equation for the acoustic potential can be solved analytically in terms of spherical harmonics.

We hope that the variational spacetime approach to transformation acoustics may assist in the design and implementation of other acoustic metadevices and may provide new avenues to research in this field.

References

1. Calin, O., Chang, D.C.: Geometric Mechanics on Riemannian Manifolds: Applications to Partial Differential Equations. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2005)
2. Chen, H.Y., Chan, C.T.: Acoustic cloaking and transformation acoustics. J. Phys. D **43**(11), 113001 (2010)
3. Cummer, S.A., Schurig, D.: One path to acoustic cloaking. New J. Phys. **9**(3), 45–52 (2007)
4. García-Meca, C., Tung, M.M.: The variational principle in transformation optics engineering and some applications. Math. Comput. Model. **57**(7–8), 1773–1779 (2013)
5. Horsley, S.A.R.: Transformation optics, isotropic chiral media and non-Riemannian geometry. New J. Phys. **13**(5), 053053 (2011)
6. Leonhardt, U., Philbin, T.: Geometry and Light—The Science of Invisibility. Dover Publications, Inc., New York (2010)
7. Lüneburg, R.K.: Mathematical Theory of Optics. University of California Press, Berkeley and Los Angeles (1966)
8. Mechel, F.P.: Formulas of Acoustics. Springer, Berlin (2002)
9. Norris, A.N.: Acoustic metafluids. J. Acoust. Soc. Am. **125**(2), 839–849 (2009)
10. Tung, M.M.: Basics of a differential-geometric approach to diffusion: Uniting Lagrangian and Eulerian models on a manifold. In: Bonilla, L.L., Moscoso, M.A., Platero, G., Vega, J.M. (eds.) Progress in Industrial Mathematics at ECMI 2006. Mathematics in Industry, vol. 12, pp. 897–901. Springer, Berlin (2007)
11. Tung, M.M.: A fundamental Lagrangian approach to transformation acoustics and spherical spacetime cloaking. Europhys. Lett. **98**, 34002–34006 (2012)
12. Tung, M.M.: Diffusion on surfaces of revolution. In: Günther, M., Bartel, A., Brunk, M., Schöps, S., Striebel, M. (eds.) Progress in Industrial Mathematics at ECMI 2010, Mathematics in Industry, vol. 17, pp. 643–650. Springer, Berlin (2012)
13. Tung, M.M., Hervás, A.: A differential-geometric approach to model isotropic diffusion on circular conic surfaces in uniform rotation. In: Fitt, A.D., Norbury, J., Ockendon, H., Wilson, E. (eds.) Progress in Industrial Mathematics at ECMI 2008. Mathematics in Industry, vol. 15, pp. 1053–1058. Springer, Berlin (2010)
14. Tung, M.M., Peinado, J.: A covariant spacetime approach to transformation acoustics. In: Fontes, M., Günther, M., Marheineke, N. (eds.) Progress in Industrial Mathematics at ECMI 2012. Mathematics in Industry, vol. XX. Springer, Berlin (2014)
15. Tung, M.M., Weinmüller, E.B.: Gravitational frequency shifts in transformation acoustics. Europhys. Lett. **101**, 54006–54011 (2013). doi:10.1209/0295-5075/101/54006
16. Visser, M., Barcelo, C., Liberati, S.: Analogue models of and for gravity. Gen. Relat. Gravit. **34**, 1719–1734 (2002)

Post-Newtonian Effects in Geolocation by FDOA

J.M. Gambi, M.M. Tung, J. Clares, and M.L. García del Pino

Abstract The post-Newtonian terms included in the Frequency Difference of Arrival equation derived here by means of Synge's world-function are considered to estimate their contribution in the precise Geolocation of passive radio transmitters at rest on the earth surface. Four of these terms are kinematical and the other two are gravitational. The kinematical terms account for the velocities of the radio transmitter and the receivers with respect to the Earth Centered Inertial reference frame, as well as for the relative velocities of the transmitter with respect to the receivers. The other two account for the gravitational attraction of an spherical earth on the receivers. The gravitational time delay has been taken into account to derive these terms.

Keywords Frequency difference of arrival • Geolocation • Post-newtonian

1 Introduction

The Frequency Difference of Arrival (FDOA) equation used in this paper has been derived by means of Synge's world-function [1] and the frequency shift formula [2]. The equation involves two satellites, S_i and S_j, that are in orbit about the earth. The equation reads

$$\begin{aligned} f_{S_i} - f_{S_j} = f_E \Big\{ & \big[\mathbf{n}_j \cdot (\mathbf{v}_{S_j} - \mathbf{v}_E)\big]\big[1 + (\mathbf{n}_j \cdot \mathbf{v}_E)\big] \\ & - \big[\mathbf{n}_i \cdot (\mathbf{v}_{S_i} - \mathbf{v}_E)\big]\big[1 + (\mathbf{n}_i \cdot \mathbf{v}_E)\big] \\ & + \frac{1}{2}\big[|\mathbf{v}_{S_i}|^2 - |\mathbf{v}_{S_j}|^2\big] + \big[\frac{m}{|\mathbf{r}_{S_i}|} - \frac{m}{|\mathbf{r}_{S_j}|}\big] \Big\} + \mathcal{O}(\varepsilon^3), \end{aligned} \tag{1}$$

J.M. Gambi (✉) • J. Clares • M.L. García del Pino
Gregorio Millán Institute, Univ. Carlos III de Madrid, 28911 Madrid, Spain
e-mail: gambi@math.uc3m.es; fclares@fis.uc3m.es; lgarciadelpino@educa.madrid.org

M.M. Tung
Instituto de Matemática Multidisciplinar, Univ. Politècnica de València, 46022 Valencia, Spain
e-mail: mtung@mat.upv.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_127

where f_E is the frequency of emission of the signal emitted by a radio transmitter at t_E, and f_{S_i}, f_{S_j} are the frequencies received by S_i and S_j at the reception instants t_{S_i} and t_{S_j} respectively.

The equation contains all the second order post-Newtonian terms that can be considered in the Earth Centered Inertial (ECI) reference frame. Thus, it contains the kinematical terms $[\mathbf{n}_i\cdot(\mathbf{v}_{S_i}-\mathbf{v}_E)][\mathbf{n}_i\cdot\mathbf{v}_E]$, $[\mathbf{n}_j\cdot(\mathbf{v}_{S_j}-\mathbf{v}_E)][\mathbf{n}_j\cdot\mathbf{v}_E]$, as well as $|\mathbf{v}_{S_i}|^2$ and $|\mathbf{v}_{S_j}|^2$, where $\mathbf{v}_E$ is the velocity of the transmitter at the emission event of the signal, $(\mathbf{r}_E, t_E)$, and $\mathbf{v}_{S_i}$ and $\mathbf{v}_{S_j}$ are the velocities of S_i and S_j at the reception events, $(\mathbf{r}_{S_i}, t_{S_i})$, $(\mathbf{r}_{S_j}, t_{S_j})$, so that both positions and velocities are referred to the ECI reference frame; $\mathbf{n}_i$, $\mathbf{n}_j$ are the directions given by $\mathbf{r}_{S_i}(t_{S_i}) - \mathbf{r}_E(t_E)$ and $\mathbf{r}_{S_j}(t_{S_j}) - \mathbf{r}_E(t_E)$ respectively. The gravitational terms take the form $m/|\mathbf{r}_{S_i}|$ and $m/|\mathbf{r}_{S_j}|$, since this form fulfils the present needs in Geolocation, m being the mass of the earth [3, 4] (Note that in Eq. (1), and in what follows, $c = G = 1$. Therefore, all the basic magnitudes in the simulations below will be given in seconds. In particular, the values of the mass and mean radius of the earth adopted are 1.479×10^{-11} and 2.125×10^{-2} s respectively).

The time delay formula used to derive (1) when $\mathbf{r}_S(t_S)$ is not aligned with $\mathbf{r}_E(t_E)$ and the ECI center is

$$\Delta t = m\Big[2\log\frac{r_S + d_S}{r_E + d_E} + \frac{d_E}{r_E} - \frac{d_S}{r_S}\Big], \tag{2}$$

where $r_E = |\mathbf{r}_E(t_E)|$, $r_S = |\mathbf{r}_S(t_S)|$, $(d_E)^2 = r_E^2 - d^2$, $(d_S)^2 = r_S^2 - d^2$, and d is the Euclidean distance from the ECI center to the straight line joining $\mathbf{r}_E(t_E)$ and $\mathbf{r}_S(t_S)$. S stands for S_i and S_j in each case. The formula used when $\mathbf{r}_S(t_S)$ is aligned with $\mathbf{r}_E(t_E)$ and the ECI center is

$$\Delta t = 2m\log\frac{r_S}{r_E}, \tag{3}$$

which is the limit of (2) when $\mathbf{r}_S(t_S)$ tends to be aligned with $\mathbf{r}_E(t_E)$ and the ECI center. Other frequency shift formulae that may give rise to alternative forms to (1) can be found in [5–7], and a different version of the time delay formulae in (2) and (3) can be found in [8]. In this context see also [10–12].

Since the aim of this work is to estimate the magnitude of the contributions in Geolocation of all the terms described above, let us mention, first, as a matter of reference, that the contribution of $(1/2)[(v_S(t_S))^2 - (v_E(t_E))^2]$ in the frequency shift formula introduced in [7] for the Navigation problem by GPS, and for receivers at rest on the earth surface, is of the order of 10^{-11} (a rough estimation of 8.35×10^{-11}, which is equivalent to a correction of 2.50 cm/s in range rate, can be found in [9]); and second, that to resemble actual geolocations, very highly inclined Low Earth Orbit (LEO) satellites in orbit with zero eccentricity will be considered as receivers. Finally, let us note that the term just mentioned does not appear in (1).

2 Impact of the Post-Newtonian Terms

Before estimating the impact of the post-Newtonian terms on the standard determination of v_E and f_E, let us note that, unlike the terms $A_i \equiv [\mathbf{n}_i \cdot (\mathbf{v}_{S_i} - \mathbf{v}_E)][\mathbf{n}_i \cdot \mathbf{v}_E]$ and $B_j \equiv [\mathbf{n}_j \cdot (\mathbf{v}_{S_j} - \mathbf{v}_E)][\mathbf{n}_j \cdot \mathbf{v}_E]$, the last two terms between brackets in (1), that is, $C_{ij} \equiv (1/2)[|\mathbf{v}_{S_i}|^2 - |\mathbf{v}_{S_j}|^2]$ and $D_{ij} \equiv [m/|\mathbf{r}_{S_i}| - m/|\mathbf{r}_{S_j}|]$, are universal, i.e. do not depend on any characteristic of the radio transmitter. In fact, the only terms that depend on the radio transmitter, through its position and velocity, are A_i and B_j. Furthermore, for orbits with zero eccentricity, $2C_{ij} = D_{ij}$. Therefore, $C_{ij} + D_{ij}$ remains constant along the orbits of S_i and S_j, if they are of zero eccentricity. The value of the constant is $(3/2)[m/a_i - m/a_j]$, where a_i and a_j are the semi-major axis of S_i and S_j respectively.

Let us now assume that S_j is a family of LEO satellites with orbital inclination $i = 85°$ and zero eccentricity, so that the longitudes of the ascending nodes, Ω_j, range from $\Omega_i = 30°\text{W}$ to $\Omega_K = 30°\text{E}$, Ω_i, Ω_K being the longitude of the ascending nodes of S_i and S_K. Let us also assume that a_j range from 2.292×10^{-2} to 2.659×10^{-2} s (that is, from about 500–1600 km). The area of interest for all the couples (S_i, S_j) is bounded by the interval (10°W, 10°E) in longitude and by the interval (40°N, 60°N) in latitude. Then we have that the contribution of $C_{ij} + D_{ij}$ to the standard determination of f_E ranges from 0 to 1.25×10^{-10}, regardless the position of E (Fig. 1). Now, the corrections due to $A_i + B_j$ for any B_j are of the same order of magnitude for any potential emitter within the area of interest. Therefore, we chose, as representative, an emitter located at coordinates (0°E, 50°N).

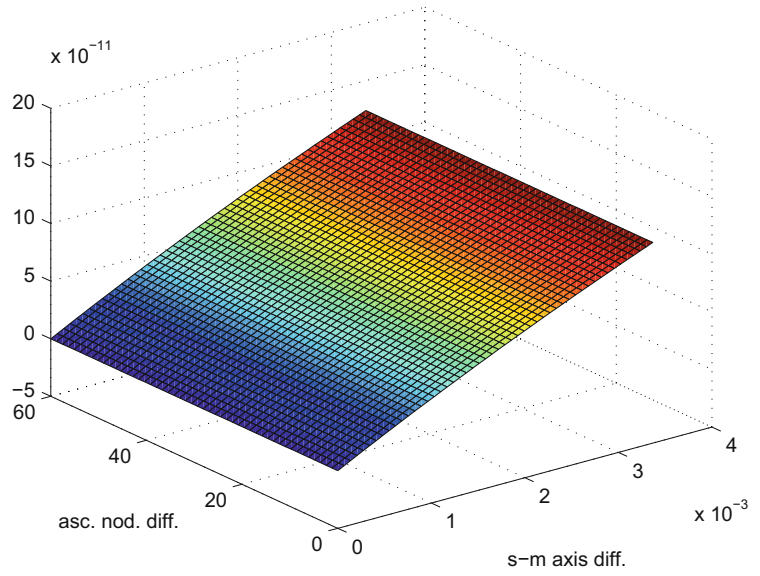


Fig. 1 Contributions of $C_{ij} + D_{ij}$ up to $a_K = 2.659 \times 10^{-2}$ s

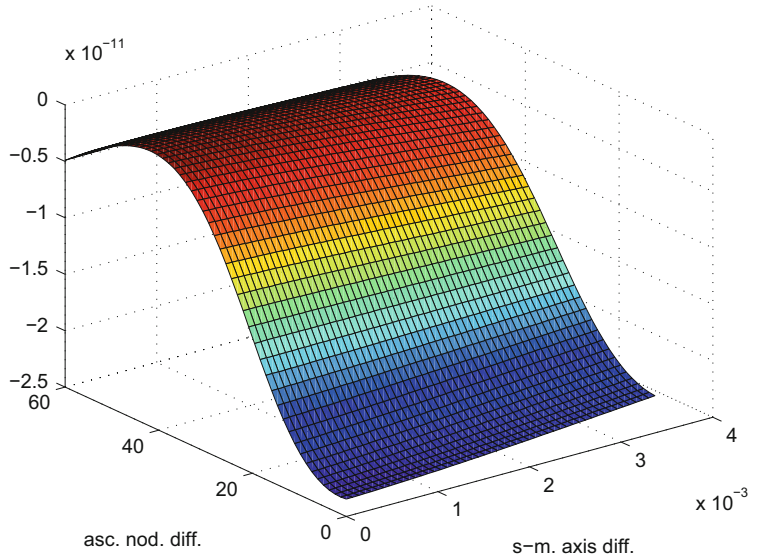


Fig. 2 Contributions of $A_i + B_j$ up to $a_K = 2.659 \times 10^{-2}$ s

Figure 2 shows the contribution of $A_i + B_j$ for each couple (S_i, S_j), so that the last Ω_j is Ω_K. Hence we can conclude that, among $A_i + B_j$ and $C_{ij} + D_{ij}$, the dominant contributor is $C_{iK} + D_{iK}$ (Figs. 1 and 2).

Next, let us assume that a_j range from 2.292×10^{-2} to 2.351×10^{-2} s. Then, keeping Ω_j between Ω_i and Ω_K, we find that the contribution of $C_{ij} + D_{ij}$ for any S_j becomes much smaller than in the previous case (Fig. 3). But the contribution of $A_i + B_j$ practically remains the same for any B_j (Fig. 4). In this case both contributions balance, although never cancel all over the area of interest (Fig. 5).

As a consequence, it is not possible to neglect, even in this case, the total contribution of the post-Newtonian terms. In fact, this contribution is in general of the same order of the contribution of the post-Newtonian term $(1/2)[(v_S(t_S))^2 - (v_E(t_E))^2]$, which is characteristic of the Navigation problem by GPS (it ranges, as was said above, in a neighborhood of 10^{-11}).

The same consequence is valid for semi-major axis intervals much smaller, such as $(2.292 \times 10^{-2}, 2.295 \times 10^{-2})$ (note that in this case $a_K - a_i \cong 4$ km). In fact, despite the contribution of $C_{ij} + D_{ij}$ becomes very small (Fig. 6), there still remains the contribution of $A_i + B_j$, which newly again is similar to those shown in Figs. 2, 4 and 7.

The last step is to check whether or not the order of the contribution of $A_i + B_j$ remains the same, so as to show its size in detail when $a_j = a_i$ for any j. The result is shown in Fig. 8. (Note that the scale of the y-axis is logarithmic.)

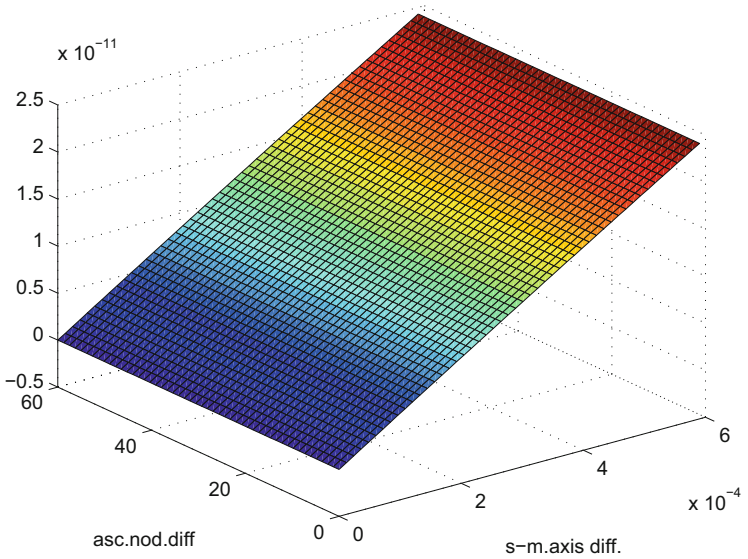


Fig. 3 Contributions of $C_{ij} + D_{ij}$ up to $a_K = 2.351 \times 10^{-2}$ s

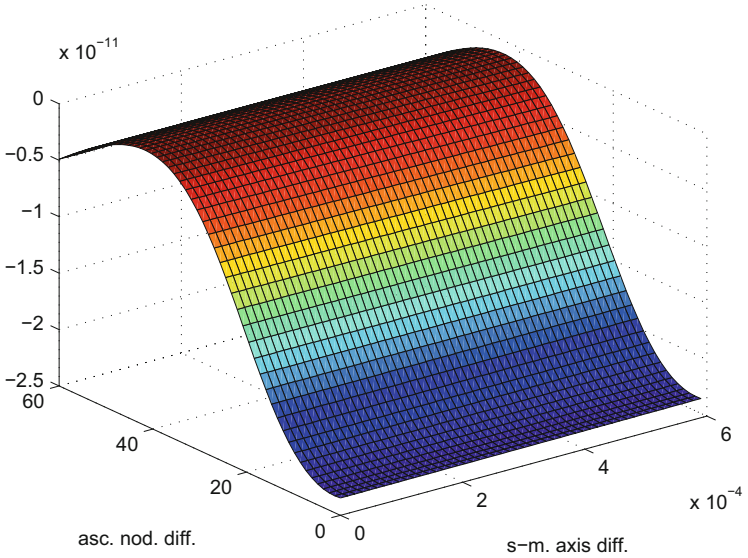


Fig. 4 Contributions of $A_i + B_j$ up to $a_K = 2.351 \times 10^{-2}$ s

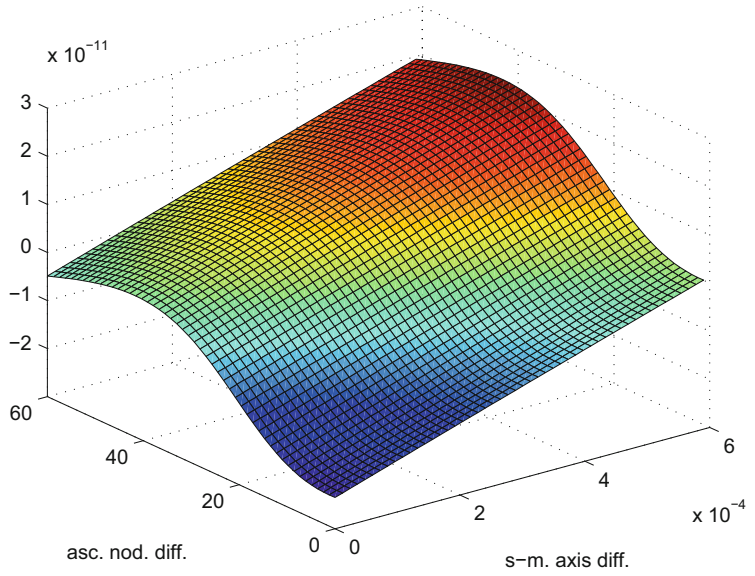


Fig. 5 Total contributions up to $a_K = 2.351 \times 10^{-2}$ s

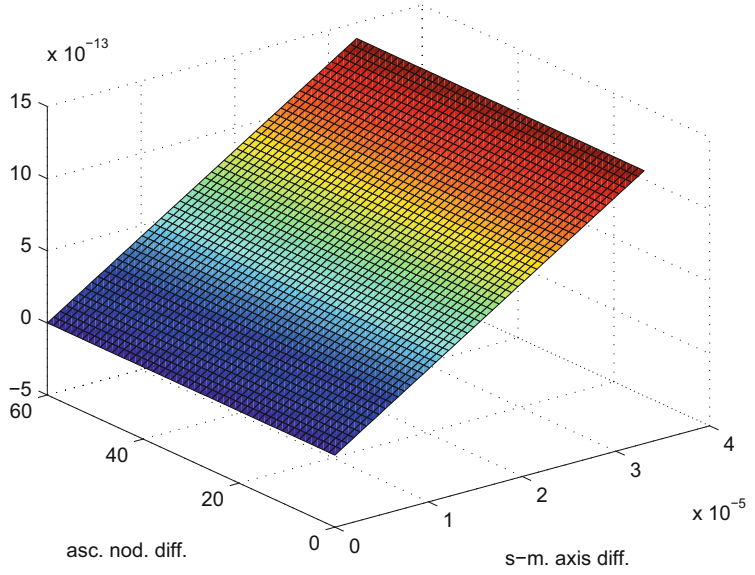


Fig. 6 Contributions of $C_{ij} + D_{ij}$ up to $a_K = 2.295 \times 10^{-2}$ s

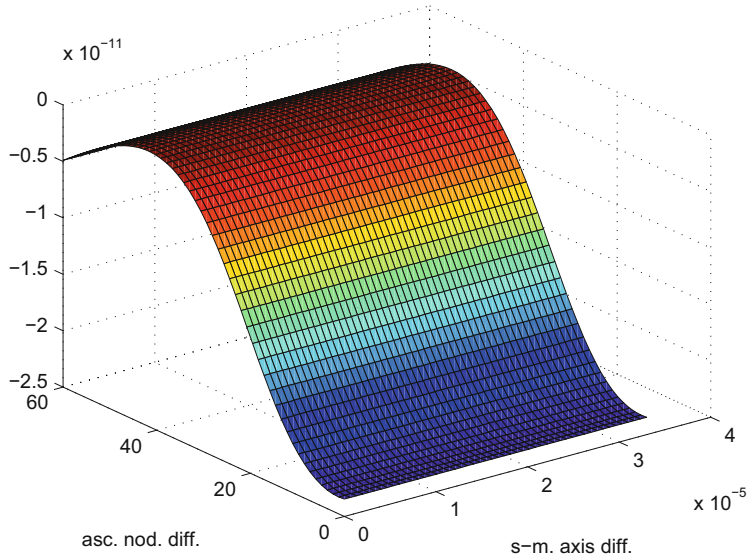


Fig. 7 Contributions of $A_i + B_j$ up to $a_K = 2.295 \times 10^{-2}$ s

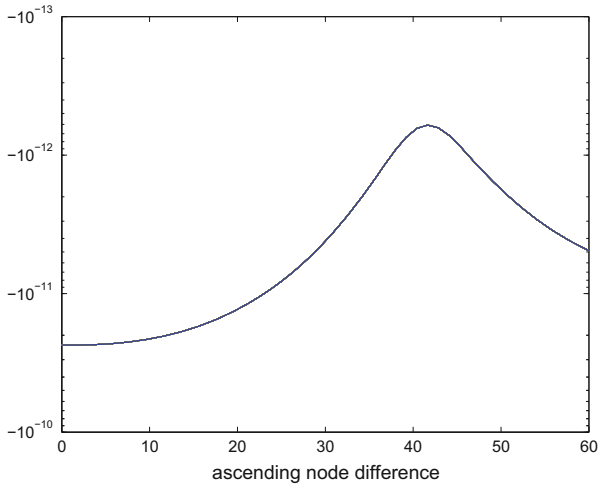


Fig. 8 Contributions of $A_i + B_j$ when $a_i = a_K$

3 Conclusion

According to the previous discussion, the post-Newtonian terms in FDOA equations, such as the one in (1), should be taken into consideration for any couple of LEO satellites. This is so because the total contribution of A_i, B_j, C_{ij}, and D_{ij}

reach values of the order of 10^{-11}, at least within the specified area of interest. Furthermore, even when C_{ij} and D_{ij} could be neglected for some couple of satellites, the corrections due to A_i and B_j always remain in the four equations of this kind needed to locate a radio transmitter.

References

1. Synge, J.L.: Relativity: The General Theory, Chap. 2. North-Holland, New York (1960)
2. Synge, J.L.: Relativity: The General Theory, Chap. 7, p. 304. North-Holland, New York (1960)
3. Ho, K.C., Chan, Y.T.: Solution and performance analysis of geolocation by TDOA. IEEE Trans. Aerosp. Electron. Syst. **29**(4), 1311–1322 (1993)
4. Adamy, D.L.: EW 102: A Second Course in Electronic Warfare. Artech House Radar Library, Chap. 6, p. 169. Horizon House Publications Inc., Boston (2004)
5. Soffel, M.H.: Relativity in Astrometry, Celestial Mechanics and Geodesy, Chap. 5. Springer, Berlin/Heidelberg (1989)
6. Montenbruck, O., Gill, E.S.: Satellite Orbits, Chap. 6, pp. 202–203. Springer, Berlin/Heidelberg (2000)
7. Ashby, N.: Relativistic effects in the global positioning system. http://www.aapt.org/doorway/tgru/articles/Ashbyarticle.pdf (2006)
8. Gambi, J.M., Rodriguez-Teijeiro, M.C., Garcia del Pino, M.L., Salas, M.: Shapiro time-delay within the Geolocation problem by TDOA. IEEE Trans. Aerosp. Electron. Syst. **47**(3), 1948–1962 (2011)
9. Combrinck, L.: General Relativity and Space Geodesy. In: Xu, G. (ed.) Sciences of Geodesy - II. Springer, Berlin/Heidelberg (2013)
10. Gambi, J.M., Rodriguez-Teijeiro, M.C., Garcia del Pino, M.L.: The post-Newtonian Geolocation problem by TDOA. In: Progress in Industrial Mathematics at ECMI 2010, p. 489. Springer, Berlin/Heidelberg (2012)
11. Seeber, G.: Satellite Geodesy, 3rd edn. Walter de Gruyter, Berlin (2003)
12. Combrinck, L.: General Relativity and Space Geodesy. In: Xu, G. (ed.) Sciences of Geodesy - II. Springer, Berlin/Heidelberg (2013)

Post-Newtonian Geolocation of Passive Radio Transmitters by TDOA and FDOA

J.M. Gambi, J. Clares, and M.C. Rodríguez Teijeiro

Abstract Different satellite configurations are considered to show by numerical simulations the influence of the post-Newtonian corrections for the standard locations of radio transmitters by the Time Difference of Arrival method to the solutions of the Newtonian Frequency Difference of Arrival equations. The satellites considered are Low, Mid and Geostationary Earth Orbit satellites in a number never smaller than five. The radio transmitters are supposed to be passive and are placed either on the earth surface or in space.

Keywords Frequency difference of arrival • Geolocation • Post-newtonian • Time difference of arrival

1 Introduction

Time Difference of Arrival (TDOA) and Frequency Difference of Arrival (FDOA) are the most accurate location systems among all the systems presently used to locate passive (i.e. noncooperative) radio transmitters placed on the Earth surface or in space. The standard TDOA methods involve receivers on board three satellites. For this reason they do not yield unique locations; further, they provide location accuracy as large as tens of meters [1].

A new TDOA method that generalizes one method for emitters on the earth surface by Ho and Chan was introduced in a recent paper [2, 3]. The method involves five satellites and provides unique locations both for emitters on the earth surface and in space. The method was also formulated within the post-Newtonian model of the Earth surrounding space in order to increase the accuracy so far reached.

Unlike the standard equations, the FDOA equations used in this paper are aimed to directly find the velocities of the emitters and the frequencies of emission. They also involve five satellites, which may not be the same used to locate the emitters

J.M. Gambi (✉) • J. Clares • M.C. Rodríguez Teijeiro
Gregorio Millán Institute, Universidad Carlos III de Madrid, 28911 Madrid, Spain
e-mail: gambi@math.uc3m.es; fclares@fis.uc3m.es; mdcrodriguez@madrid.uned.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_128

by TDOA. To solve the FDOA equations, the position of the emitters must be previously know. For this reason we use the TDOA method introduced in [3].

The aim of this work is to show by means of numerical simulations the corrections for the velocities of the emitters and for the frequencies of emission that are due to the post-Newtonian corrections to the Classical TDOA equations.

2 TDOA and FDOA Equations

Let us assume that the position of an emitter E at the emission instant of a signal is $\mathbf{r}_E$, so that $r_E = |\mathbf{r}_E|$. Let us also assume that (x, y, z) are the Earth Centered Inertial (ECI) coordinates of E at the emission instant of the signal. Let us assume that $\mathbf{v}_E$ is the velocity of E at the emission instant, and that (v_x, v_y, v_z) are the ECI components of $\mathbf{v}_E$ at that instant. Let $\mathbf{r}_{S_i}$ be the positions of the receivers S_i at the arrival instants of the signal, so that (x_i, y_i, z_i) are the ECI coordinates of S_i at those instants, $r_{S_i} = |\mathbf{r}_{S_i}|$, and $(v_{x_i}, v_{y_i}, v_{z_i})$ are the ECI components of $\mathbf{v}_{S_i}$ $(i = 1, \ldots, 5)$. Finally, let us assume that the frequency of emission of the signal is f_E, and that f_i are the reception frequencies at the reception instants at $\mathbf{r}_{S_i}$.

If $r_i = |\mathbf{r}_i|$, where $\mathbf{r}_i = \mathbf{r}_{S_i} - \mathbf{r}_E$, we have that the time difference of arrival to S_i and S_j, say $r_{i,j}$, is given by $r_i - r_j$ $(i, j = 1, \ldots, 5; i \neq j)$. Then we have (Fig. 1)

$$
\begin{aligned}
r_{3,2} + r_{2,1} - r_{3,1} &= 0,\\
r_{4,3} + r_{3,1} - r_{4,1} &= 0,\\
r_{5,4} + r_{4,1} - r_{5,1} &= 0,
\end{aligned}
$$

so that

$$
\begin{aligned}
r_{3,2}r_{2,1}r_{3,1} &= l_1 + m_1 x + n_1 y + v_1 z,\\
r_{4,3}r_{3,1}r_{4,1} &= l_2 + m_2 x + n_2 y + v_2 z,\\
r_{5,4}r_{4,1}r_{5,1} &= l_5 + m_5 x + n_5 y + v_5 z,
\end{aligned}
\tag{1}
$$

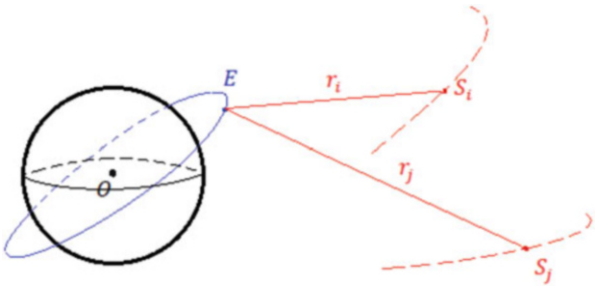


Fig. 1 Magnitudes involved in the Newtonian TDOA equations

where

$$l_1 = r_{3,2}K_1 + r_{2,1}K_3 - r_{3,1}K_2,$$
$$l_2 = r_{4,3}K_1 + r_{3,1}K_4 - r_{4,1}K_3,$$
$$l_5 = r_{5,4}K_1 + r_{4,1}K_5 - r_{5,1}K_4,$$

$$m_1 = -2(r_{3,2}x_1 + r_{2,1}x_3 - r_{3,1}x_2),$$
$$m_2 = -2(r_{4,3}x_1 + r_{3,1}x_4 - r_{4,1}x_3),$$
$$m_5 = -2(r_{5,4}x_1 + r_{4,1}x_5 - r_{5,1}x_4),$$

$$n_1 = -2(r_{3,2}y_1 + r_{2,1}y_3 - r_{3,1}y_2),$$
$$n_2 = -2(r_{4,3}y_1 + r_{3,1}y_4 - r_{4,1}y_3),$$
$$n_5 = -2(r_{5,4}y_1 + r_{4,1}y_5 - r_{5,1}y_4),$$

and

$$v_1 = -2(r_{3,2}z_1 + r_{2,1}z_3 - r_{3,1}z_2),$$
$$v_2 = -2(r_{4,3}z_1 + r_{3,1}z_4 - r_{4,1}z_2),$$
$$v_5 = -2(r_{5,4}z_1 + r_{4,1}z_5 - r_{5,1}z_4),$$

with $K_i = x_i^2 + y_i^2 + z_i^2$.

The Classical TDOA equations are the equations in (1).[1] For the post-Newtonian equations we have $r_{i,j} = (r_i - r_j)(1 - \eta_{ij})$, where $\eta_{ij} = -p_{ij}/(r_i - r_j)$ $(r_i \neq r_j)$ and

$$p_{ij} = m\left\{2\log\left[\frac{\tan(\frac{\theta_{0_i}}{2})\tan(\frac{\theta_j}{2})}{\tan(\frac{\theta_{0_j}}{2})\tan(\frac{\theta_j}{2})}\right] + (\cos\theta_{0_i} - \cos\theta_{0_j}) + (\cos\theta_j - \cos\theta_i)\right\},$$

so that θ_{0_i} (θ_{0_j}) and θ_i (θ_j) are the angles that $\mathbf{r}_E$ and $\mathbf{r}_{S_i}$ $(\mathbf{r}_{S_j})$ make with $\mathbf{r}_i$ $(\mathbf{r}_j)$ and m is the mass of the earth (Fig. 2). The equations are

$$r_{3,2}r_{2,1}r_{3,1} = \bar{l}_1 + \overline{m}_1 x + \overline{n}_1 y + \overline{v}_1 z,$$
$$r_{4,3}r_{3,1}r_{4,1} = \bar{l}_2 + \overline{m}_2 x + \overline{n}_2 y + \overline{v}_2 z,$$
$$r_{5,4}r_{4,1}r_{5,1} = \bar{l}_5 + \overline{m}_5 x + \overline{n}_5 y + \overline{v}_5 z, \tag{2}$$

[1]Note that to derive these equations we take $c = 1$. Note also that to derive the post-Newtonian equations we have also made $G = 1$. For this reason, the initial data in the numerical simulations below are given in seconds.

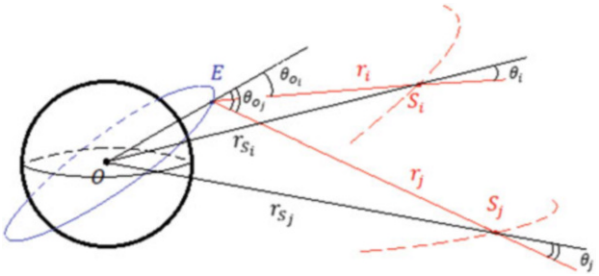


Fig. 2 Magnitudes involved in the post-Newtonian TDOA equations

where

$$\begin{aligned}
\bar{l}_1 &= (r_{3,2}K_1 + r_{2,1}K_3 - r_{3,1}K_2)(1 + \eta_{32} + \eta_{21} + \eta_{31}),\\
\bar{l}_2 &= (r_{4,3}K_1 + r_{3,1}K_4 - r_{4,1}K_3)(1 + \eta_{43} + \eta_{31} + \eta_{41}),\\
\bar{l}_5 &= (r_{5,4}K_1 + r_{4,1}K_5 - r_{5,1}K_4)(1 + \eta_{54} + \eta_{41} + \eta_{51}),
\end{aligned}$$

$$\begin{aligned}
\overline{m}_1 &= (-2)(r_{3,2}x_1 + r_{2,1}x_3 - r_{3,1}x_2)(1 + \eta_{32} + \eta_{21} + \eta_{31}),\\
\overline{m}_2 &= (-2)(r_{4,3}x_1 + r_{3,1}x_4 - r_{4,1}x_3)(1 + \eta_{43} + \eta_{31} + \eta_{41}),\\
\overline{m}_5 &= (-2)(r_{5,4}x_1 + r_{4,1}x_5 - r_{5,1}x_4)(1 + \eta_{54} + \eta_{41} + \eta_{51}),
\end{aligned}$$

$$\begin{aligned}
\overline{n}_1 &= (-2)(r_{3,2}y_1 + r_{2,1}y_3 - r_{3,1}y_2)(1 + \eta_{32} + \eta_{21} + \eta_{31}),\\
\overline{n}_2 &= (-2)(r_{4,3}y_1 + r_{3,1}y_4 - r_{4,1}y_3)(1 + \eta_{43} + \eta_{31} + \eta_{41}),\\
\overline{n}_5 &= (-2)(r_{5,4}y_1 + r_{4,1}y_5 - r_{5,1}y_4)(1 + \eta_{54} + \eta_{41} + \eta_{51}),
\end{aligned}$$

and

$$\begin{aligned}
\overline{v}_1 &= (-2)(r_{3,2}z_1 + r_{2,1}z_3 - r_{3,1}z_2)(1 + \eta_{32} + \eta_{21} + \eta_{31}),\\
\overline{v}_2 &= (-2)(r_{4,3}z_1 + r_{3,1}z_4 - r_{4,1}z_3)(1 + \eta_{43} + \eta_{31} + \eta_{41}),\\
\overline{v}_5 &= (-2)(r_{5,4}z_1 + r_{4,1}z_5 - r_{5,1}z_4)(1 + \eta_{54} + \eta_{41} + \eta_{51}).
\end{aligned}$$

The FDOA equations are based on the standard formula

$$f_i = f_E\big[1 - (\mathbf{r}_i/|\mathbf{r}_i|) \cdot (\mathbf{v}_{S_i} - \mathbf{v}_E)\big],$$

so that $f_{i,j} = f_i - f_j$. The equations are

$$
\begin{aligned}
\left(\frac{\mathbf{r}_1}{|\mathbf{r}_1|} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|}\right) \cdot \mathbf{v}_E - f_{1,2}(f_E)^{-1} &= \frac{\mathbf{r}_1}{|\mathbf{r}_1|} \cdot \mathbf{v}_{S_1} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|} \cdot \mathbf{v}_{S_2}, \\
\left(\frac{\mathbf{r}_3}{|\mathbf{r}_3|} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|}\right) \cdot \mathbf{v}_E - f_{3,2}(f_E)^{-1} &= \frac{\mathbf{r}_3}{|\mathbf{r}_3|} \cdot \mathbf{v}_{S_3} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|} \cdot \mathbf{v}_{S_2}, \\
\left(\frac{\mathbf{r}_4}{|\mathbf{r}_4|} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|}\right) \cdot \mathbf{v}_E - f_{4,2}(f_E)^{-1} &= \frac{\mathbf{r}_4}{|\mathbf{r}_4|} \cdot \mathbf{v}_{S_4} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|} \cdot \mathbf{v}_{S_2}, \\
\left(\frac{\mathbf{r}_5}{|\mathbf{r}_5|} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|}\right) \cdot \mathbf{v}_E - f_{5,2}(f_E)^{-1} &= \frac{\mathbf{r}_5}{|\mathbf{r}_5|} \cdot \mathbf{v}_{S_5} - \frac{\mathbf{r}_2}{|\mathbf{r}_2|} \cdot \mathbf{v}_{S_2}.
\end{aligned} \tag{3}
$$

3 Numerical Simulations

The satellite data used to solve (1) and (2) for emitters in circular orbit about the earth, with $r_E = 2.140 \times 10^{-2}$ s, $\Omega_E = 0°$, and inclinations ranging from 30° to 80° are: $S_1(a_1 = 6.7 \times 10^{-2}\,\text{s}, e_1 = 0, \Omega_1 = -70°, i_1 = 55°, f_1 = 0.12°)$; $S_2(a_2 = 14.002 \times 10^{-2}\,\text{s}, e_2 = 0, \Omega_2 = -99°, i_2 = 0.1°, f_2 = 57°)$; $S_3(a_3 = 14.002 \times 10^{-2}\,\text{s}, e_3 = 0, \Omega_3 = -57°, i_3 = 0.1°, f_3 = 57°)$; $S_4(a_4 = 6.7 \times 10^{-2}\,\text{s}, e_4 = 0, \Omega_4 = 42°, i_4 = 55°, f_4 = 0.12°)$, and $S_5(a_5 = 6.7 \times 10^{-2}, \text{s}, e_5 = 0, \Omega_5 = 70°, i_5 = 55°, f_5 = 0.12°)$. Figure 3 shows the distance between the classical and post-Newtonian locations. The orbital inclinations of the emitters are represented on the X-axis, and the true anomalies at the emission instants on the Y-axis.

The satellite data used to solve (3), with $f_E = 1.025 \times 10^6$ Hz, are: $S_1(a_1 = 3.0 \times 10^{-2}\,\text{s}, e_1 = 0, \Omega_1 = 30°, i_1 = 80°, f_1 = 5°)$; $S_2(a_2 = 2.3 \times 10^{-2}\,\text{s}, e_2 =$

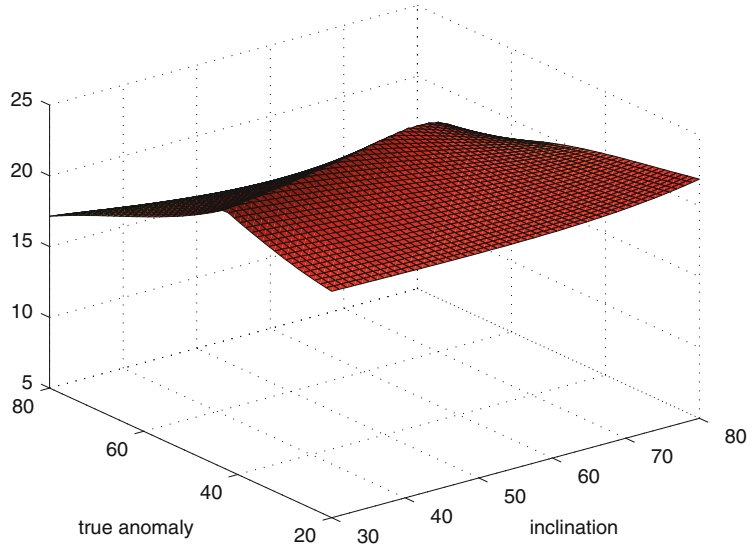


Fig. 3 Corrections to the distances (in m)

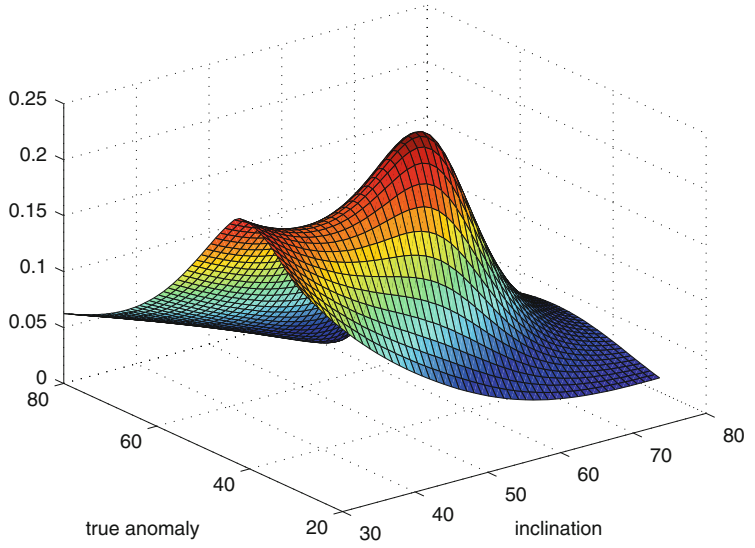


Fig. 4 Corrections to the speeds (in m/s)

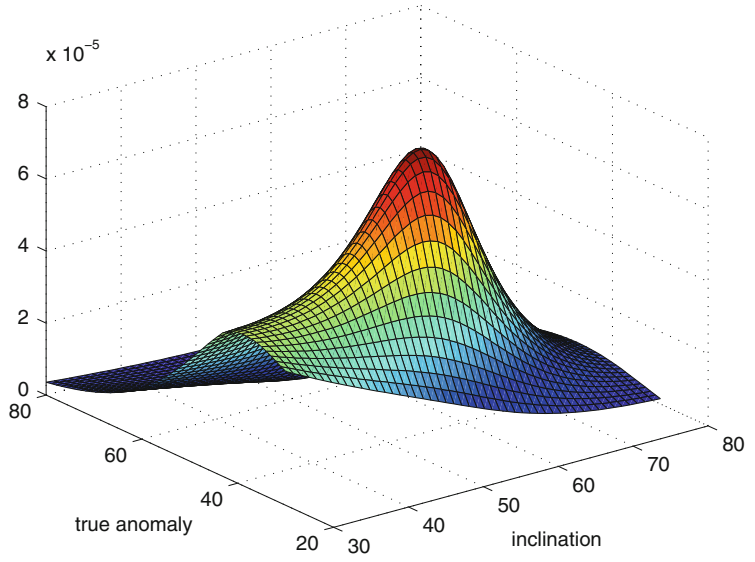


Fig. 5 Corrections to the velocity directions (in radians)

$0, \Omega_2 = 30°, i_2 = 80°, f_2 = 60°$); $S_3(a_3 = 14.002 \times 10^{-2}\,\mathrm{s}, e_3 = 0, \Omega_3 = 0°, i_3 = 0.1°, f_3 = -10°)$; $S_4(a_4 = 14.002 \times 10^{-2}\,\mathrm{s}, e_4 = 0, \Omega_4 = 20°, i_4 = 0.1°, f_4 = 10°)$, and $S_5(a_5 = 14.002 \times 10^{-2}\,\mathrm{s}, e_5 = 0, \Omega_5 = 10°, i_5 = 0.1°, f_5 = 10°)$. Figure 4 shows the influence of the post-Newtonian corrections for the location of the emitters on their speeds; Fig. 5 on the velocity directions, and Fig. 6 on f_E.

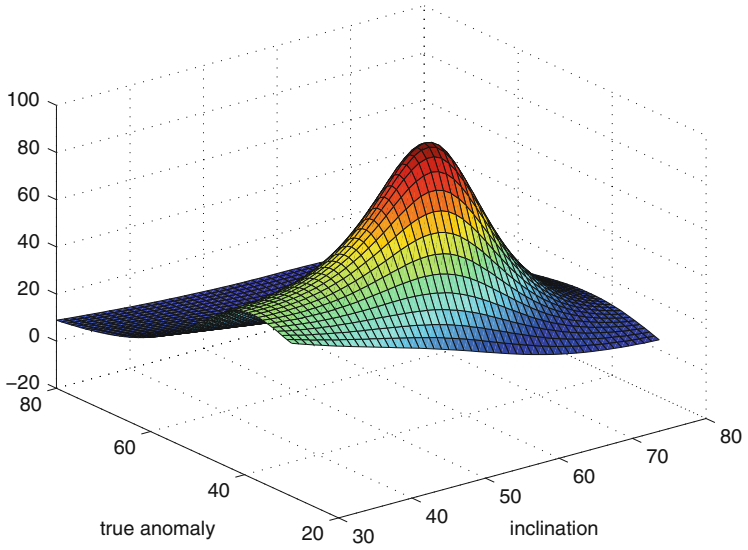


Fig. 6 Corrections to the emitted frequency (Hz)

4 Conclusion

The simulations shown in this work are rather representatives of the size of the corrections for a wide variety of satellite configurations. In fact, the three types of satellites, i.e. Low, Mid and Geostationary Earth Orbit satellites are involved in these simulations. Therefore, the post-Newtonian corrections to the standard locations can be expected to be at least of the order of the meter for any configuration. Hence they can produce significative variations on many standard estimations of the velocities. Better estimations of the velocities must be obtained by using the post-Newtonian equations corresponding to (3).

References

1. Adamy, D.L.: EW 102: A Second Course in Electronic Warfare. Artech House Radar Library, Chap. 6, p. 169. Horizon House Pub. Inc., Boston (2004)
2. Ho, K.C., Chan, Y.T.: Solution and performance analysis of Geolocation by TDOA. IEEE Trans. Aerosp. Electron. Syst. **29**(4), 1311–1322 (1993)
3. Gambi, J.M., Rodriguez-Teijeiro, M.C., Garcia del Pino, M.L.: The post-Newtonian Geolocation problem by TDOA. In: Progress in Industrial Mathematics at ECMI 2010, p. 489, Springer, Berlin/Heidelberg (2012)

Post-Newtonian Orbital Equations for Fermi Frames in the Vicinity the Earth

J.M. Gambi, M.L. García del Pino, and M.M. Tung

Abstract Synge's equations for time-like geodesics in terms of Fermi coordinates are used to derive post-Newtonian equations for the relative motion of satellites in coplanar circular near orbits about the earth. The reference frame, co-moving with the base satellite, is assumed to be a Fermi frame, that is, inertial guided. The resulting system is autonomous, linear, and reduces to the equation of the geodesic deviation for nearby satellites. Hence, it can be used by some Acquisition, Pointing, and Tracking systems to increase the accuracy presently reached in locating passive radio-transmitters.

Keywords Frequency difference of arrival • Post-newtonian • Radio transmission control and surveillance

1 Introduction

The emerging importance of radio transmission control and surveillance is making the implementation of accurate space-based Acquisition, Pointing, and Tracking (APT) systems a relevant issue. In particular, Satellite-to-Satellite laser technology attracts more and more attention due to the fact that this technology has matured substantially in the recent years (see e.g. [1, 2]).

The equations introduced below correspond to the post-Newtonian model of the Earth surrounding space. They can help the satellites equipped with those systems to increase the accuracy in computing the relative position of other satellites in order to determine by the Frequency Difference of Arrival (FDOA) method the velocities of passive radio transmitters placed on the Earth surface or in space [3]. In fact, the equations reduce to the equation of the geodesic deviation for nearby satellites orbiting the earth. To show their benefits, the satellites are assumed to be in coplanar

J.M. Gambi (✉) • M.L. García del Pino
Gregorio Millán Institute, Univ. Carlos III de Madrid, 28911 Madrid, Spain
e-mail: gambi@math.uc3m.es; lgarciadelpino@educa.madrid.org

M.M. Tung
Instituto de Matemática Multidisciplinar, Univ. Politècnica de València, 46022 Valencia, Spain
e-mail: mtung@mat.upv.es

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_129

circular orbits with similar radii. Certainly, among these assumptions the only essential to autonomously set up, at the right instants, initial conditions producing solutions with the accuracy presently required, is the last. But, in summary, here we add the other two for the sake of simplicity.

To derive the equations, Synge's equations for time-like geodesics in terms of Fermi coordinates are used. For this reason these equations are introduced in Sect. 2. Then the equations are derived in Sect. 3. The generalization of these equations for orbits with small eccentricity and/or different orbital planes is briefly discussed in the Conclusions.

2 Synge's Equations

Let E be a space-time with metric $g_{ij}(x^k)$ and world-function $\Omega(x^{k_1}, x^{k_2})$.[1] From now on Latin indices range from 1 to 4, and Greek, from 1 to 3. We also adopt $c = G = 1$, so that $x^4 = t$ and the basic magnitudes are to be given in seconds.

Let $(\lambda^{k_1}_{(\alpha)}(s_1), \lambda^{k_1}_{(4)}(s_1))$ be an orthonormal tetrad Fermi transported along a time-like base world line $C_1(x^{k_1}(s_1))$ with $\lambda^{k_1}_{(4)}(s_1) = A^{k_1}(s_1) = dx^{k_1}/ds_1$, so that $A^{4_1}(s_1) = dt/ds_1$; let $P_2(x^{k_2})$ be an arbitrary event in a time-like geodesic $C_2(x^{k_2}(s_2))$ and let $(X^{(\alpha)}, s_1) = (X_{(\alpha)}, s_1)$ be the Fermi coordinates of $P_2(x^{k_2})$ with respect to C_1 where s_1 and s_2 are the proper times of C_1 and C_2 respectively. If $b_1(s_1)$, the first curvature of C_1, is null for all s_1; if $\Omega_{i_1j_1l_1}$ and $\Omega_{i_1j_1l_2}$ are the third-order covariant derivatives of $\Omega(x^{k_1}, x^{k_2})$ taken as indicated by the indices, that is, with respect to x^{i_1}, x^{j_1} and x^{l_1}, in the first case, and with respect to x^{l_2} for the third derivative in the second case; if, furthermore, $H^{k_2} = A^{k_2}(ds_2/ds_1)$ with $A^{k_2} = dx^{k_2}/ds_2$, and if, finally,

$$dL_{(\alpha)}/ds_1 = \chi L_{(\alpha)} + \Omega_{i_1j_1l_2}\lambda^{i_1}_{(\alpha)}A^{j_1}H^{l_2} + \Omega_{i_1j_2l_2}\lambda^{i_1}_{(\alpha)}H^{j_2}H^{l_2}, \tag{1}$$

with $L_{(\alpha)} = \Omega_{i_1j_2}\lambda^{i_1}_{(\alpha)}H^{j_2}$, where $\Omega_{i_1j_2}$ are the second-order covariant derivatives of $\Omega(x^{k_1}, x^{k_2})$, first with respect to x^{i_1}, and then with respect to x^{j_2}; $\chi = (d^2s_2/ds_1^2)/(ds_2/ds_1)$ and $\Omega_{i_1j_1l_2}$, $\Omega_{i_1j_2l_2}$ are the third-order covariant derivatives whose interpretation is similar to those of the previous derivatives, then Synge's equations read [4]

$$\frac{d^2X_{(\alpha)}}{ds_1^2} = -\Omega_{i_1j_1l_1}\lambda^{i_1}_{(\alpha)}A^{j_1}A^{l_1} - \Omega_{i_1j_1l_2}\lambda^{i_1}_{(\alpha)}A^{j_1}H^{l_2} - \frac{dL_{(\alpha)}}{ds_1}. \tag{2}$$

[1] For any two events in E, $P_1(x^{k_1})$, $P_2(x^{k_2})$, for which there is a unique geodesic $\Gamma_{P_1P_2}$ joining them with equations $x^k = \xi^k(u)$ where u is an affine parameter ranging from 0 to 1, the world-function $\Omega(P_1, P_2)$ is defined by the line integral

$$\Omega(P_1, P_2) = \Omega(x^{k_1}, x^{k_2}) = \frac{1}{2}\int_0^1 g_{ij}U^iU^j du$$

taken along $\Gamma_{P_1P_2}$ where $x^{k_1} \equiv \xi^k(0)$, $x^{k_2} \equiv \xi^k(1)$ and $U^k = d\xi^k/du$.

3 The Post-Newtonian Equations

According to Synge the calculations to integrate Eq. (2) become unmanageable. However, they become much simpler by introducing the following approximations, which are valid for all time-like world lines in the vicinity of the earth: (1) $ds_2/ds_1 = 1$ approx.; (2) C_2 is nearly parallel to C_1; and (3) for any two events P_2, $P_{2'}$ in C_2, with Fermi coordinates $(X^{(\alpha)}, s_1)$ and $(X^{(\alpha)} + dX^{(\alpha)}, s_1 + ds_1)$,

$$2\Omega(P_2, P_{2'}) = g_{(rs)}dX^{(r)}dX^{(s)},$$

with

$$\begin{aligned} g_{(\alpha\beta)} &= \delta_{\alpha\beta} + 2h_{(\alpha_1\beta_2)}, \\ g_{(\alpha 4)} &= 0, \\ g_{(44)} &= -1 + 2h_{(4_1 4_1)} + 2h_{(4_1 4_2)}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} h_{(\alpha_1\beta_2)} &= \frac{3}{2}\sigma^{-3}X^{(\mu)}X^{(\nu)}\int_0^\sigma (\sigma - u)uS_{(\alpha\beta\mu\nu)}du, \\ h_{(4_1 4_1)} &= \frac{3}{2}\sigma^{-3}X^{(\mu)}X^{(\nu)}\int_0^\sigma (\sigma - u)^2 S_{(44\mu\nu)}du, \\ h_{(4_1 4_2)} &= \frac{3}{2}\sigma^{-3}X^{(\mu)}X^{(\nu)}\int_0^\sigma (\sigma - u)uS_{(44\mu\nu)}du, \end{aligned} \tag{4}$$

the integrals being taken along the straight line $x^k(u) = x^{k_1}(1 - u/\sigma) + x^{k_2}u/\sigma$ $(0 \le u \le \sigma)$ where, according to the Fermi coordinates assigned to P_2, $x^{k_1}(s_1)$ are the coordinates of the foot P_1 at C_1 of the geodesic $\Gamma_{P_1P_2}$ drawn from $P_2(x^{k_2})$ to cut orthogonally C_1, so that $\sigma = X_{(\alpha)}X^{(\alpha)}$ and

$$S_{(abcd)} = S_{(abcd)}(x^k(u)) = \left[S_{ijlm}(x^k(u))\right]\left[\lambda^i_{(a)}\lambda^j_{(b)}\lambda^l_{(c)}\lambda^m_{(d)}(x^k(u))\right], \tag{5}$$

where $\lambda^k_{(\alpha)}(x^k(u)), \lambda^k_{(4)}(x^k(u))$ are respectively parallel to $\lambda^{k_1}_{(\alpha)}(x^{k_1}(s_1)), A^{k_1}(x^{k_1}(s_1))$ with respect to the metric $g_{ij}(x^k)$, and

$$S_{ijlm}(x^k(u)) = -\frac{1}{3}(R_{iljm} + R_{imjl})\big(x^k(u)\big), \tag{6}$$

where $R_{abcd}(x^k(u))$ is the Riemann tensor at $x^k(u)$.

In fact, under these hypothesis Eq. (2) become for the post-Newtonian approximation of the earth exterior Schwarzschild field with Earth Centered Inertial (ECI) coordinates

$$\frac{d^2X_{(\alpha)}}{ds_1^2} = -\Omega_{(\alpha_1 4_1 4_1)} - 2\Omega_{(\alpha_1 4_1 4_2)} - \Omega_{(\alpha_1 4_2 4_2)}, \tag{7}$$

where

$$\Omega_{(\alpha_1 4_1 4_1)} = -\Omega_{(\alpha_1 4_1 4_2)} = -\sigma^{-3} X^{(\gamma)} \int_0^\sigma (\sigma - u)^2 R_{(\alpha 4 \gamma 4)} du,$$

$$\Omega_{(\alpha_1 4_2 4_2)} = 2\sigma^{-3} X^{(\gamma)} \int_0^\sigma u^2 R_{(\alpha 4 \gamma 4)} du, \tag{8}$$

and

$$R_{(\alpha 4 \gamma 4)} = R_{(\alpha 4 \gamma 4)}(x^k(u)) = -m \left(\frac{3x^\alpha(u) x^\gamma(u)}{r(u)^5} - \frac{\delta_{\alpha\gamma}}{r(u)^3} \right), \tag{9}$$

where m is the mass of the earth; $x^\delta(u) = x^{\delta_1}(1 - u/\sigma) + x^{\delta_2} u/\sigma$, x^{δ_1} and x^{δ_2} being the ECI coordinates of P_1 and P_2, and $r(u)^2 = x^\delta(u) x^\delta(u)$.

The calculations to integrate Eq. (7) are now as manageable as the calculations to integrate the equations of the geodesic deviation, since these equations are

$$\frac{d^2X_{(\alpha)}}{ds_1^2} = -R_{(\alpha 4 \beta 4)} X^{(\beta)}, \tag{10}$$

with $R_{(\alpha 4 \beta 4)}$ evaluated at $x^{k_1}(s_1)$.

In fact, both systems of equations can be integrated in parallel in order to compute the increment of accuracy while determining the relative motion of any satellite S_2 (whose world line is C_2) with respect to a base satellite S_1 (whose world line is C_1).

In particular, the benefits of (7) compared with the benefits of (10) for a system of satellites designed to locate passive radio transmitters by means of FDOA can easily be shown by merely considering S_1 and S_2 to be under the assumptions mentioned in the Introduction. (A detailed description of the nature of the Geolocation problem by FDOA can be found in [5].)

Thus, after computing the integrals in (8) under those assumptions using the plane orbital coordinates $\overline{x}^1, \overline{x}^2$ for S_1 (the $\overline{x}^1$-axis taken towards the ascending node

of the orbit of S_1) and assuming that $\lambda^{\delta_1}_{(1)}(s_1)$ and $\lambda^{\delta_1}_{(2)}(s_1)$ remain parallel to the $\overline{x}^1$- and $\overline{x}^2$-axis respectively, we have that Eq. (7) become

$$\begin{aligned}\frac{d^2X_{(1)}}{ds_1^2} &= \frac{m}{r_1^3}\left(3\cos^2 M_1 - 1\right)\left(1 - \frac{7}{4}\eta\right)X^{(1)} + \frac{3m}{r_1^3}\cos M_1 \sin M_1\left(1 - \frac{7}{4}\eta\right)X^{(2)},\\ \frac{d^2X_{(2)}}{ds_1^2} &= \frac{3m}{r_1^3}\cos M_1 \sin M_1\left(1 - \frac{7}{4}\eta\right)X^{(1)} + \frac{m}{r_1^3}\left(3\sin^2 M_1 - 1\right)\left(1 - \frac{7}{4}\eta\right)X^{(2)},\end{aligned} \tag{11}$$

with $X_{(3)} = 0$, where $M_1 = M_1(s_1)$ is the mean anomaly of S_1 at s_1; $r_1^2 = x^{\delta_1}x^{\delta_1}$, and $\eta = ((r_2 - r_1)/r_1) \ll 1$, r_2^2 being $x^{\delta_2}x^{\delta_2}$.

We now note with respect to Eq. (11) that η does not depend on s_1. It can also be verified, as a matter of check, that if $\eta = 0$ and the initial condition are $X_{(1)_0} = X_{(2)_0} = X'_{(1)_0} = X'_{(2)_0} = 0$, then we have the expected solution $X_{(1)}(s_1) = X_{(2)}(s_1) = 0$. Therefore, one can derive useful results when $\eta = 0$ and the initial conditions correspond to configurations of S_1 and S_2 for which $M_2 = M_1 + M$, where M_2 is the mean anomaly of S_2 and M takes several constant values properly chosen. In fact, for $\eta = 0$, Eq. (11) become

$$\begin{aligned}\frac{d^2X_{(1)}}{ds_1^2} &= \frac{m}{r_1^3}\left(3\cos^2 M_1 - 1\right)X^{(1)} + \frac{3m}{r_1^3}\cos M_1 \sin M_1 X^{(2)},\\ \frac{d^2X_{(2)}}{ds_1^2} &= \frac{3m}{r_1^3}\cos M_1 \sin M_1 X^{(1)} + \frac{m}{r_1^3}\left(3\sin^2 M_1 - 1\right)X^{(2)},\end{aligned} \tag{12}$$

which coincide with Eq. (10) for nearby satellites.

4 Conclusion

Synge's equations for geodesics in terms of Fermi coordinates can be simplified up to obtain equations more general than the equations of the geodesic deviation, yet useful to increase the accuracy provided by these equations in locating by means of FDOA passive radio transmitters placed on the earth surface or in the vicinity of the earth. In fact, by comparing Eqs. (7) and (10), it is straightforward to see why Eq. (7) reduce to Eq. (10). In particular, it can be deduce from these equations that their difference for the particular case considered to derive Eq. (11) is the parameter η, which in this case is constant, as it is in any other case. In fact, the essential characteristic of η is that it only involves the semi-major axes of S_1 and S_2, so that the initial and subsequent Fermi distances from S_1 to S_2 can be kept under control by two-way laser ranging according to the post-Newtonian framework. Furthermore,

despite Eq. (7) are more complete than Eq. (10), they are equally affordable, since they are still linear and the system remains autonomous.

The generalization of Eq. (11) to equations valid for satellites with different inclination is straightforward, particularly if the orbits are circular and have the same radius. The reason is that the integrals in (8) that correspond to these configurations are very similar to those computed to derive Eq. (11). But the computation of the integrals for orbits with (small) eccentricities, e_1, e_2, is not so simple, since previous computations involving series expansions in e_1 and e_2 are required.

References

1. Guelma, M., Kogan, A., Kazarian, A., Livne, A., Orenstain, M., Michalik, H., Arnold, S.: Acquisition and pointing control for inter-satellite laser communications. IEEE Trans. Aerosp. Electron. Syst. **40**(4), 1239–1248 (2004)
2. Norton, T., Conner, K., Covington, R., Ngo, H., Rink, C.: Development of reprogrammable high frame-rate detector devises for laser communication pointing, Acquisition and Tracking. IEEE Aerospace Conference Paper No. 1414 (2008). doi:10.1109/AERO.2008.4526342
3. Gambi, J.M., Rodriguez-Teijeiro, M.C., Garcia del Pino, M.L.: The post-Newtonian Geolocation problem by TDOA. In: Progress in Industrial Mathematics at ECMI 2010, p. 489. Springer, Berlin/Heidelberg (2012)
4. Synge, J.L.: Relativity: The General Theory, Chap. 2. North-Holland, New York (1960)
5. Adamy, D.L.: EW 102: A Second Course in Electronic Warfare. Artech House Radar Library, Chap. 6, p. 169. Horizon House Pub. Inc., Boston (2004)

MS 37
MINISYMPOSIUM: STRUCTURED NUMERICAL LINEAR ALGEBRA IN IMAGING AND MONUMENT CONSERVATION

Organizers

Marco Donatelli[1] and Stefano Serra Capizzano[2]

Speakers

Alessandro Buccini[3] and Marco Donatelli[1]
Multigrid Regularization Method for Image Deblurring with Arbitrary Boundary Conditions

Daniele Bertaccini[4]
Updating Preconditioners in Digital Image Restoration

Thomas Huckle[5]
Numeric Evaluation of Geometric Continuity in CAD Systems

[1]Marco Donatelli, Università degli Studi dell'Insubria, Como, Italy.
[2]Stefano Serra Capizzano, Università degli Studi dell'Insubria, Como, Italy.
[3]Alessandro Buccini, Università degli Studi dell'Insubria, Como, Italy.
[4]Daniele Bertaccini, Università di Roma "Tor Vergata", Roma, Italy.
[5]Thomas Huckle, Techinische Universitat Munchen, Munchen, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_130

Marco Donatelli[1], Matteo Semplice[6] and Stefano Serra Capizzano[2]
Multigrid Preconditioning for Nonlinear (Degenerate) Parabolic Equations with Application to Monument Degradation

Fabrizio Clarelli[7], Barbara De Filippo[8] and Roberto Natalini[9]
A Free-Boundary Model of Corrosion

Armando Coco[10], Marco Donatelli[1] and Matteo Semplice[6]
Numerical Methods for Nonlinear PDEs Modeling Monument Preservation

Keywords

Image processing
Monument degradation
Structured linear algebra

Short Description

Numerical methods in structured linear algebra/matrix theory are relevant in several applications in the restoration of blurred and noisy images (2D, 3D) and in the modeling of the monument degradation under the action of pollutants (e.g. of biochemical type). In reality, several tools from structured matrix-theory are ready to be used in this context, with the effect of substantial improvements in the computational efficiency and in the precision of the results. Conversely, new specific examples of applications pose new challenging mathematical problems to people working on numerical methods and in structured matrix theory.

The researches in the field require multi-level techniques in order to deal with edges/details present in images and monuments and often the convergence of these techniques is well understood in terms of the spectral properties (eigenvalue localization, eigenvalue distribution, asymptotic behavior, eigenvector character in terms

[6]Matteo Semplice, Università degli Studi di Torino, Torino, Italy.

[7]Fabrizio Clarelli, Istituto per le Applicazioni del Calcolo “M. Picone”, Consiglio Nazionale delle Ricerche, Roma, Italy.

[8]Barbara De Filippo, Istituto per le Applicazioni del Calcolo “M. Picone”, Consiglio Nazionale delle Ricerche, Roma, Italy.

[9]Roberto Natalini, Istituto per le Applicazioni del Calcolo “M. Picone”, Consiglio Nazionale delle Ricerche, Roma, Italy.

[10]Armando Coco, Université de Bordeaux, Bordeaux, France.

of frequencies etc.) of the underlying structures, arising from the approximation of the involved integral and differential operators. We recall that the approximation of integral equations in imaging leads to matrix-structures which are (severely) ill conditioned in a high portion of the high frequency domain, while elliptic/parabolic partial differential equations, also of nonlinear and degenerate type, arise in both problems.

A Free-Boundary Model of Corrosion

F. Clarelli, B. De Filippo, and R. Natalini

Abstract Deterioration of copper and bronze artifacts is one of the main concerns for people working in cultural heritage. In particular a significant effort has been devoted to study the corrosion due to environmental conditions, such as temperature, moisture and the concentration of pollutants. We introduce a mathematical model able to describe the corrosion effects on a copper layer, which is subject to deposition of SO_2. The present model is based on a partial differential equation system with a double free boundary for monitoring and detecting copper corrosion products (mainly brochantite and cuprite).We assume to have a copper sample on which is formed a non protective oxide layer (Cu_2O), and, over this layer, a corrosion product (brochantite) grows. We aim to create a new approach to forecasting corrosion behavior without the necessity of an extensive use of laboratory testing using chemical-physical technologies, while taking into account the main chemical reactions. Although the model was kept simple, just describing the main reaction and transport processes involved, the mathematical simulations and the related model calibration are in agreement with the laboratory experiments.

Keywords Brochantite • Copper • Corrosion • Cultural heritage • Finite difference methods • Free boundary model • Parabolic problems

1 Introduction

The process of deterioration of copper alloys is a major concern for conservators, scientist, art historians and collectors [8], where environmental factors, such as temperature, moisture, concentration of pollutants play a key role in corrosion

F. Clarelli (✉)
CNR Istituto per le Applicazioni del Calcolo "M. Picone", Via Madonna del Piano 10, I-50019 Sesto Fiorentino, FI, Italy
e-mail: f.clarelli@iac.rm.cnr.it

B. De Filippo • R. Natalini
CNR Istituto per le Applicazioni del Calcolo "M. Picone", Via dei Taurini 19, 00185 Roma, Italy
e-mail: b.defilippo@iac.cnr.it; roberto.natalini@cnr.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_131

[1, 5, 7]. In this study we focus on pollutants such as sulfur dioxide (SO_2), being one of the most important factors in the deterioration of bronze. Indeed SO_2 reacts (in presence of water) to produce sulfate acid (H_2SO_4), which causes corrosion phenomena on copper surfaces and produces several corrosion products (such as antlerite, posnjakite, brochantite [6]).

The complexity of corrosion processes involved need a quantitative model approach to develop predictive tools. These methods, similar to those introduced in [3] and [2], can be used not only for the monitoring and detection of artefact alterations, but also for determining optimal intervention strategies.

This mathematical model is able to describe the evolution of copper corrosion and is based on fluid dynamical and chemical relations and characterized by a double free boundary. Its calibration has been elaborated according to the experimental results in [4].

2 The Model

The mathematical model describes the development of a corrosion patina on a copper sample, taking into account the formation of cuprite and brochantite as the principal corrosion products, as in Fig. 1.

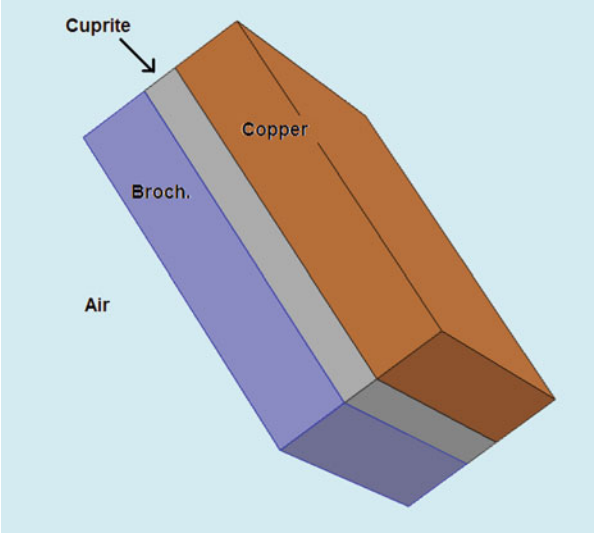


Fig. 1 Example of cuprite and brochantite deposition on a copper sample

We consider the following reaction to describe the cuprite Cu_2O production (see Eq. (1))

$$2Cu + \frac{1}{2}O_2 \rightarrow Cu_2O. \tag{1}$$

While, we consider a simplified reaction for brochantite production in Eq. (2), details regard this approximation are in [3].

$$2Cu_2O + SO_2 + 3H_2O + \frac{3}{2}O_2 \rightarrow Cu_4(OH)_6SO_4, \tag{2}$$

Since we assume these two reactions as instantaneous, we have two free boundaries, and the effective time of reaction is implicitly included in the diffusivity coefficients.

2.1 Equations of the Model

In our model we consider the swelling effect due to the volume change caused by the chemical reactions. We indicate by $\alpha(t)$ the boundary between copper and cuprite, $\beta(t)$ is the boundary between cuprite and brochantite, and $\gamma(t)$ is the boundary between brochantite and external air. Thus, we have four regions:

1. Copper (inner region).
2. Cuprite Cu_2O, between $\alpha(t)$ and $\beta(t)$.
3. Brochantite $Cu_4(OH)_6SO_4$, between $\beta(t)$ and $\gamma(t)$.
4. Laboratory controlled atmosphere on the external side of $\gamma(t)$.

Equations of our model describes mass balances in the brochantite layer and in the cuprite layer, a detailed description of the model and the mass balances on the boundaries are in [3].

2.2 Brochantite Layer Equations

The brochantite formation (Eq. (2)) has been assumed to develop on the cuprite layer, due to the SO_2, H_2O and O_2, which move through the brochantite layer ($\gamma(t) \leq x \leq \beta(t)$) and react with Cu_2O. This reaction implies the formation of a new layer of brochantite.

In the following, the concentration of SO_2 is indicated by S, the water concentration by W and the oxygen concentration by O.

Mass balance of SO_2 in the brochantite layer $\gamma(t) \leq x \leq \beta(t)$ is

$$\frac{\partial S}{\partial t} - D_s \frac{\partial^2 S}{\partial x^2} + \dot{\gamma} \frac{\partial S}{\partial x} = 0, \tag{3}$$

on the external boundary we have the environmental SO_2 concentration (Eq. (4))

$$S(\gamma(t)) = S_{air}(t), \tag{4}$$

and on the boundary $\beta(t)$ we assumed that SO_2 reacts totally with Cu_2O, thus we have

$$S(\beta(t)) = 0. \tag{5}$$

Since the flux of SO_2 at the boundary $\beta(t)$ is proportional to Cu_2O moles consumption, we have a further condition (first free boundary)

$$-n_b \frac{D_s}{M_s} \frac{\partial S}{\partial x} = \frac{1}{2} \frac{\rho_p}{M_p} \dot{\beta}; \tag{6}$$

where M_s, M_p are the molar weight of SO_2 and Cu_2O respectively, ρ_p is the mass density of cuprite and D_s is the diffusivity.

The mass balance of water in the brochantite layer is

$$\frac{\partial W}{\partial t} - D_w \frac{\partial^2 W}{\partial x^2} + \dot{\gamma} \frac{\partial W}{\partial x} = 0. \tag{7}$$

On the boundary $\gamma(t)$ we have

$$W(\gamma(t)) = W_{air}(t). \tag{8}$$

For $x = \beta(t)$, we have that some moles of water are wasted by the reaction (2), and the boundary condition is

$$\frac{J_w}{M_w} = \frac{3}{2} \frac{\rho_p}{M_p} \dot{\beta} + n_b \frac{W}{M_w} \dot{\beta}. \tag{9}$$

Where D_w is the water diffusivity and

$$J_w = n_b \left(-D_w \frac{\partial W}{\partial x} - W \omega_p \dot{\alpha} - W \omega_b \dot{\beta} \right) = n_b \left(-D_w \frac{\partial W}{\partial x} + W \dot{\gamma} \right). \tag{10}$$

Finally the oxygen mass balance in the brochantite layer is

$$\frac{\partial O}{\partial t} - D_o \frac{\partial^2 O}{\partial x^2} + \dot{\gamma} \frac{\partial O}{\partial x} = 0, \tag{11}$$

on the boundary $x = \gamma(t)$ we have

$$O(\gamma(t)) = O_{air}(t), \tag{12}$$

and on the boundary $x = \beta(t)$ we have the condition (13), since the oxygen is wasted by the reaction (2).

$$\frac{J_o}{M_o} = \frac{3}{4}\frac{\rho_p}{M_p}\dot{\beta} + n_b\frac{O}{M_o}\dot{\beta}. \tag{13}$$

where D_o is the oxygen diffusivity and the flux J_o is in Eq. (14)

$$J_o = n_b\left(-D_o\frac{\partial O}{\partial x} + O\dot{\gamma}\right). \tag{14}$$

2.3 Cuprite Layer Equations

In this domain, to avoid confusion, we indicate oxygen by G.

We assume that all oxygen moles, that arrive on the inner boundary $\alpha(t)$, react with copper. Although this assumption is not always valid, it is justified in our model since water plays a key role in the speed of reaction and in our experiments the parameters of Relative Humidity (RH) are near to 100 %. Also, the diffusivity D_g includes implicitly the finite time of reaction.

Oxygen mass balance equation is

$$\frac{\partial G}{\partial t} - D_g\frac{\partial^2 G}{\partial x^2} - \omega_p\dot{\alpha}\frac{\partial G}{\partial x} = 0. \tag{15}$$

The value of $G(\beta(t))$ is given by oxygen on the boundary $\beta(t)$ obtained by Eq. (13).

$$G(\beta(t)) = O(\beta(t)); \tag{16}$$

on the other boundary $\alpha(t)$, oxygen reacts totally with copper (Eq. (17))

$$G(\alpha(t)) = 0. \tag{17}$$

The last condition in $x = \alpha(t)$ is given by the oxygen which reacts totally with the copper moles

$$-n_p\frac{D_g}{M_g}\frac{\partial G}{\partial x} = \frac{1}{4}\frac{\rho_c}{M_c}\dot{\alpha}. \tag{18}$$

Where D_g is the oxygen diffusivity and J_g is

$$J_g = n_p\left(-D_g\frac{\partial G}{\partial x} - G\omega_p\dot{\alpha}\right). \tag{19}$$

3 Calibration

We calibrated the diffusivity coefficients with our experimental tests. To do that, we used the thickness of corrosion products at different time points. Each thickness value has been obtained by a sample average and standard deviation, computed at each time-point.

We used the least square method to find the best parameters. We obtained (in cm^2/s): $D_g = 9.9 \cdot 10^{-9}$, $D_s = 3.96 \cdot 10^{-5}$, $D_o = 9.9 \cdot 10^{-6}$ and $D_w = 3.96 \cdot 10^{-5}$. The evolution in time of the corrosion products thickness, using the best parameters, is in Fig. 2. We can see the difference between the experimental points and the best simulation in Fig. 3.

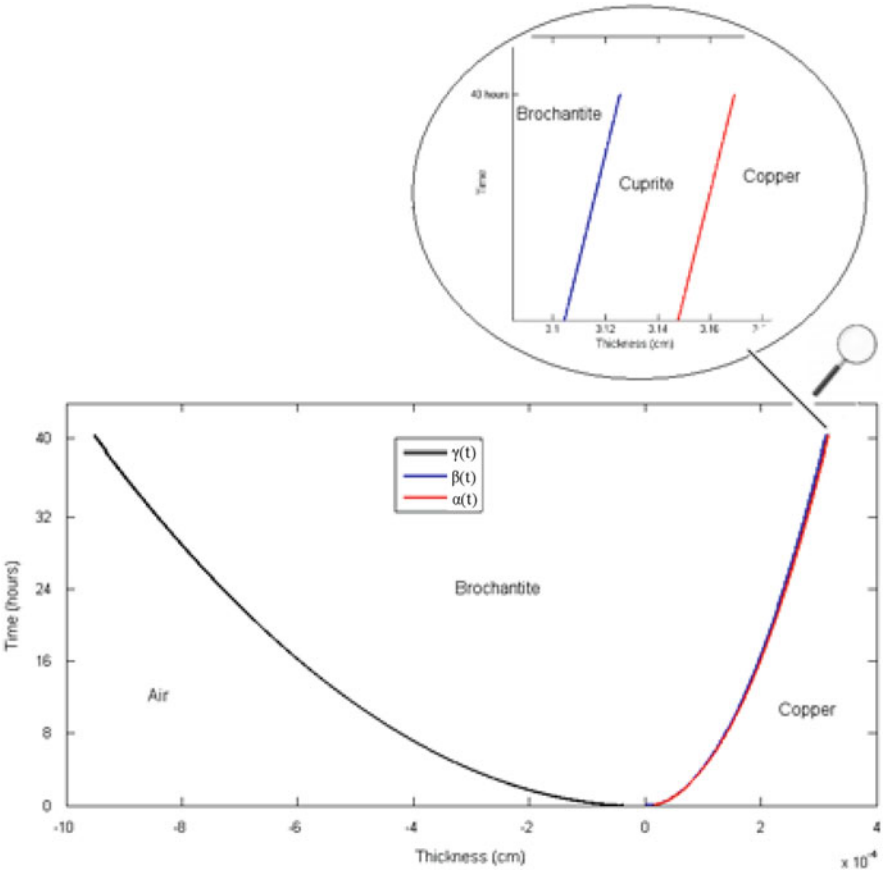


Fig. 2 Evolution of $\gamma(t)$ (*black line*), $\beta(t)$ (*blue line*) and $\alpha(t)$ (*red line*)

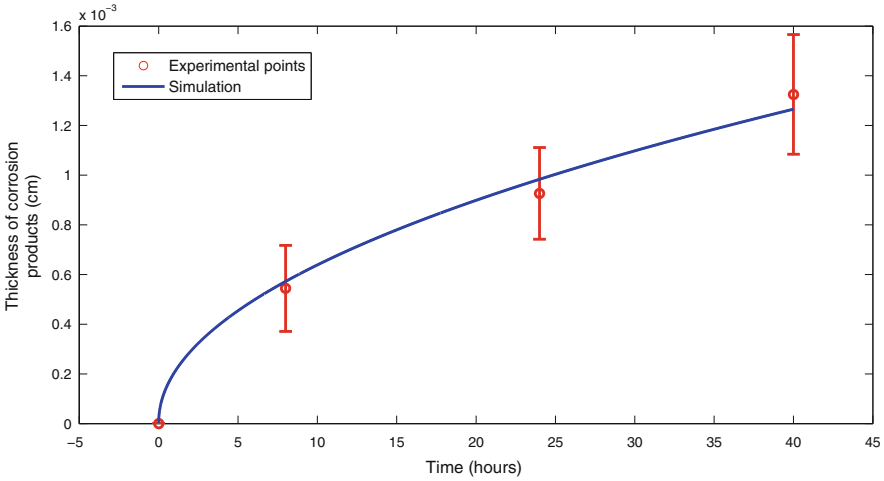


Fig. 3 Simulation (*continuous line*) and experimental points (*in red*)

Details on finite differences numerical schemes can be found in [3]. Simulations, after 40 h, give us the following values: $\gamma = -9.505 \cdot 10^{-4}$, $\alpha = 3.1693 \cdot 10^{-4}$, $\beta = 7.9916 \cdot 10^{-4}$ (cm).

Although the model was kept simple, just describing the main reaction and transport processes involved, the mathematical simulations and the related model calibration are in agreement with the laboratory experiments.

References

1. Chawla, S.K., Payer, J.H.: The early stage of atmospheric corrosion of copper by sulfur dioxide. J. Electrochem. Soc. **137**(1), 60–64 (1990)
2. Clarelli, F., Fasano, A., Natalini, R.: Mathematics and monument conservation: free boundary models of marble sulfation. SIAM J. Appl. Math. **69**(1), 149–168 (2008)
3. Clarelli, F., De Filippo, B., Natalini, R.: Mathematical model of copper corrosion. Appl. Math. Model. **38**(19–20), 4804–4816 (2014)
4. De Filippo, B., Campanella, L., Brotzu, A., Natali, S., Ferro, D.: Characterization of bronze corrosion products on exposition to sulphur dioxide. Adv. Mater. Res. **138**, 21–28 (2010)
5. Fitzgerald, K.P., Nairn, J., Skennerton, G.A.: Atmospheric corrosion of copper and the colour, structure and composition of natural patina on copper. Corros. Sci. **48**, 2480–2509 (2006)
6. Graedel, T.E., Nassau, K., Franey, J.P.: Copper patinas formed in the atmosphere-I. Introduction. Corros. Sci. **27**(7), 639–657 (1987)
7. Nassau, K., Miller, A.E., Graedel, T.E.: The reaction of simulated rain with copper, copper patina, and some copper compound. Corros. Sci. **27**, 703–719 (1987)
8. Scott, D.A.: Copper and Bronze in art: Corrosion, Colorants, Conservation. The Getty Conservation Institute, Los Angeles (2002). ISBN:0-89236-638-9

MS 38
MINISYMPOSIUM: TAILORED MATHEMATICS FOR THE TECHNICAL TEXTILE INDUSTRY

Organizers

Nicole Marheineke[1]

Speakers

Joachim Binnig[2]
The Airlay Process: Basic Principles and Challenges for Tomorrow

Simone Gramsch[3]
Simulation of Fiber Dynamics and Fiber Wall Contacts for Airlay Processes

Christian H. Neßler[4]
Construction of Virtual Nonwovens

Christoph Strohmeyer[5]
Effective Mechanical Properties of Nonwovens Produced by Airlay Processes

Andrew Sageman-Furnas[6]
Exploring Woven Structures Through the Geometry of Chebyshev Nets

[1]Nicole Marheineke, FAU Erlangen-Nürnberg, Germany.
[2]Joachim Binning, AUTEFA Solutions, Germany.
[3]Simone Gramsch, Fraunhofer ITWM, Kaiserslautern, Germany.
[4]Christian H. Neßler, TU Kaiserslautern, Germany.
[5]Christoph Strohmeyer, FAU Erlangen-Nürnberg, Germany.
[6]Andrew Sageman-Furnas, University Göttingen, Germany.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_132

Stephan Martin[7]
Higher-Order Averaging of Linear Fokker-Planck Equations for Fiber Dynamics

Thomas M. Cibis[8] and Nicole Marheineke[1]
Homogenization Strategies for Fiber Curtains and Bundles in Air Flows

Javier Rivero-Rodriguez[9]
Setup of Viscous Cosserat Rod Model Describing Electrospinning

Walter Arne[10]
Homotopy Method for Viscous Cosserat Rod Model Describing Electrospinning

Stefan Schiessl[11]
A Moving Mesh Framework Based on Three Parametrization Layers for 1d PDEs

Keywords

Fiber spinning
Nonwoven manufacturing

Short Description

Spunbond, meltblowing, airlaying, fluid-dynamical sewing as well as electrospinning are just a number of the various manufacturing processes for technical textiles. In the focus of all these processes stand slender objects such as oriented particles, elastic threads or viscous/viscoelastic jets that move due to mechanical, electromagnetic or aerodynamical forces and interact with each other, outer walls and/or surrounding turbulent flows. The application spectrum for the final fabrics, the technical textiles, is extremely broad and ranges from everyday products like diapers and vacuum cleaner bags to high-tech goods like battery separators and medical products. Optimization and design of the production processes with respect to the desired material properties requires tailored mathematical models and methods that allow for accurate and efficient simulations.

[7]Stephan Martin, Imperial College London, UK.
[8]Thomas M. Cibis, FAU Erlangen-Nürnberg, Germany.
[9]Javier Rivero-Rodriguez, University Sevilla, Spain.
[10]Walter Arne, Fraunhofer ITWM, Kaiserslautern, Germany.
[11]Stefan Schiessl, FAU Erlangen-Nürnberg, Germany.

A Moving Mesh Framework Based on Three Parameterization Layers for 1d PDEs

Stefan Schiessl, Nicole Marheineke, and Raimund Wegener

Abstract Solutions of partial differential equations (PDEs) arising in science and industrial applications often undergo large variations occurring over small parts of the domain. Resolving steep gradients and oscillations properly is a numerical challenge. The idea of the *r*-refinement (moving mesh) is to improve the approximation quality—while keeping the computational effort—by redistributing a fixed number of grid points in areas of the domain where they are needed. In this work we develop a general moving mesh framework for 1d PDEs that is based on three parameterization layers representing referential, computational and desired parameters. Numerical results are shown for two different strategies that are applied to a fiber spinning process.

Keywords Fiber spinning • Moving mesh • Three parameterization layers

1 Moving Mesh

In the solution of PDEs steep gradients and oscillations can occur and cause numerical difficulties. In the classical *h*-refinement additional grid points are inserted, the improved approximation quality comes hence with the price of increased computational costs. This work considers the so-called *r*-refinement or moving mesh (see e.g. [4] and references within): the general idea is to improve the approximation quality by redistributing a fixed number of grid points within the domain while keeping the computational effort. For this purpose the existing approaches in literature use two grids: the referential grid in which the model equations are originally formulated and the transformed grid. Restricting to 1d

S. Schiessl (✉) • N. Marheineke
FAU Erlangen-Nürnberg, Lehrstuhl Angewandte Mathematik 1, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: stefan.schiessl@math.fau.de; marheineke@math.fau.de

R. Wegener
Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
e-mail: raimund.wegener@itwm.fraunhofer.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_133

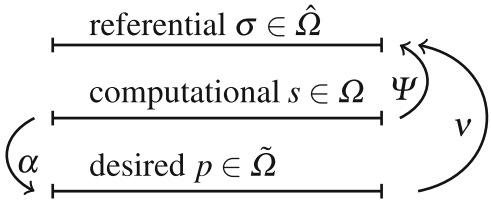


Fig. 1 Illustration of the three parameterization layers

problems we develop a general moving mesh framework that is based on three parameterization layers and discuss different adaption strategies for a suitable reparameterization. In particular, we embed an established MMPDE approach [5] into our framework and compare the results in an example of fiber spinning.

Framework of Three Parameterization Layers We consider three parameterization layers: *referential* $\sigma \in \hat{\Omega}$, *computational* $s \in \Omega$ and *desired* $p \in \tilde{\Omega}$ parameterizations, see Fig. 1. The model equations for an arbitrary application are originally formulated in a depicted parameterization (e.g. Lagrangian or Eulerian description) to which we refer to as the referential parameters. The desired parameters should now reflect some kind of optimal parameterization for the given problem, e.g. the absolute value of a gradient of a solution component becomes constant. As the direct use of the desired parameterization is not numerically beneficial, the computational parameters are additionally introduced. The core of the framework are the time-dependent parameter transformations $\Xi \in \{\Psi, \alpha, \nu\}$ with $\Xi(\cdot, t)$ one-to-one mapping for time $t \in [0, T]$ that are listed in Fig. 1.

To allow for moving grids, the original equations are transformed into the computational parameters by using $\Psi : \Omega \times [0, T] \to \hat{\Omega}$. The computational parameters are not identical to the desired ones, but should approach them, i.e. $\alpha : \Omega \times [0, T] \to \tilde{\Omega}$ should be pulled towards identity id. Consequently, the *r*-refinement aims at an adaption strategy (Part (a) equation for the unknown Ψ) and a description of a desired reparameterization (Part (b) choice for α). The existing moving mesh approaches regard only two parameterizations as the desired layer is implicitly incorporated in the computational layer. The advantage of our proposed framework is a clear separation of all three layers. This provides more flexibility in the modeling as we will see in the following.

In literature there exists various approaches with theoretical statements for fixed domains, e.g. [3, 5]. Following that confinement, we assume here that $\hat{\Omega} = \Omega = \tilde{\Omega} = [0, 1]$. Then, the transformations Ξ can be interpreted as *distribution functions for the parameters*. Supposing sufficient regularity $\Xi \in \mathscr{C}^1([0, 1] \times [0, T], [0, 1])$, the derivatives $f_\Xi = \partial_x \Xi$ describe the *parameter densities*, i.e. $f_\Xi > 0$ and $\int_0^1 f_\Xi(x, t)\, \mathrm{d}x = 1$.

(a) Adaption Strategy: Equation for Ψ, f_Ψ Proceeding from a desired parameter distribution in terms of ν, f_ν, the idea behind the adaption of the computational parameters is a temporal relaxation. In that sense Ψ, f_Ψ fulfills at a later time what is currently deemed optimal with ν, f_ν. In the following we present two strategies: distribution relaxation (DELAX) and moving mesh partial differential equations

(MMPDE). Thereby, the second one was originally proposed and explored in [5] as the MMPDE4 approach and is here embedded in our general framework.

Strategy 1.1 (Distribution Relaxation (DELAX)) Let a parameter density f_α be given. Then Ψ is determined by the evolution equation

$$\partial_t \Psi = -\frac{1}{\tau}\left(\Psi - \Psi \circ \alpha^{-1}\big|_{p=s}\right), \qquad t \in (0,T),\ s \in \Omega \tag{1}$$

with initial condition $\Psi(s,0) = s$ and temporal relaxation parameter $\tau > 0$. Note that $\alpha(s,t) = \int_0^s f_\alpha(s',t)\,\mathrm{d}s'$ holds.

DELAX proceeds from the relaxation ansatz $\Psi(s,t+\tau) = \nu(s,t) = \Psi(\alpha^{-1}(s,t),t)$ for the distribution. Since $f_\alpha > 0$, the order of the grid points is preserved under α (i.e. *no node crossing* property [3]). This property is directly handed over to Ψ. Moreover, in the limit $\tau \to 0$, (1) enforces that $\alpha \equiv \mathrm{id}$. The strategy requires the computation of the inverse of α and the interpolation between the parameterizations. However, the costs are relatively cheap as the domains are one-dimensional. Alternatively, one might also think of an respective evolution equation for f_Ψ

$$\partial_t f_\Psi = -\frac{1}{\tau}\left(f_\Psi - (f_\Psi \circ \alpha^{-1}) f_{\alpha^{-1}}\big|_{p=s}\right).$$

In MMPDE [5] a *monitor function* is introduced to describe the desired reparameterization.

Definition 1.2 (Monitor Function) Let a function $\hat{M} : \hat{\Omega} \times [0,T] \mapsto \mathbb{R}^+$, $t \in [0,T]$ be given and assume that $\hat{M}(\cdot,t) \in \mathscr{C}^0(\hat{\Omega},\mathbb{R}^+)$ is strictly positive and bounded. Then $\hat{M}$ is called a *monitor function*. On the computational domain Ω the monitor function is denoted by $M(s,t) := \hat{M}(\Psi(s,t),t)$.

Strategy 1.3 (Moving Mesh PDE (MMPDE)) Let a monitor function M be given, satisfying $M(\cdot,t) \in \mathscr{C}^1(\Omega,\mathbb{R}^+)$. Assume that $\Psi(\cdot,t) \in \mathscr{C}^2(\Omega,\hat{\Omega})$, then Ψ is determined by the PDE

$$\partial_s(M\,\partial_t f_\Psi) = -\frac{1}{\tau}\partial_s(M f_\Psi), \qquad t \in (0,T]\ s \in (0,1), \tag{2}$$

with the initial and boundary conditions $\Psi(s,0) = s$ and $\Psi(0,t) = 0$, $\Psi(1,t) = 1$ and temporal relaxation parameter $\tau > 0$.

MMPDE proceeds from the relaxation ansatz $f_\Psi(s,t+\tau) = f_\nu(\alpha(s,t),t)$ for the density. By the chain rule $f_\nu \circ \alpha = (f_{\nu^{-1}} \circ \Psi)^{-1} = f_\Psi / f_\alpha$ particularly holds. The desired parameterization is modeled in terms of M with $\mathscr{M}(t) = \int_0^1 \hat{M}(\sigma,t)\,\mathrm{d}\sigma$, i.e.

$$f_{\nu^{-1}} = \frac{\hat{M}}{\mathscr{M}}, \qquad \text{implying} \quad f_\nu \circ \alpha = \frac{\mathscr{M}}{M}, \quad f_\alpha = \frac{f_\Psi M}{\mathscr{M}}.$$

Multiplying $\partial_t f_\Psi = -(f_\Psi - \mathscr{M}/M)/\tau$ with M and taking the space derivative, the integral of the monitor function disappears and the second order MMPDE (2) with mixed partial derivatives is obtained. The property of no node crossing is fulfilled (see proof in [3]). Moreover, in the limit $\tau \to 0, f_\alpha \equiv 1$ is enforced which implies $\alpha \equiv \mathrm{id}$. By dealing with the parameter density f_Ψ, this strategy brings diffusion to the problem. In contrast to DELAX, the MMPDE strategy cannot be formulated directly on the level of the distribution function Ψ.

(b) Desired Reparameterization: Choice of α, f_α The redistribution of the parameters is performed with respect to the chosen density function f_α (or $\hat{M}$, M in MMPDE, respectively). In general, f_α is a model-dependent arbitrarily complicated functional on the solution that should approach $f_\alpha = 1$ by moving the mesh. Hence, it is often set up as gradient or higher derivatives of solution components. Consider the solution $\hat{y} : \hat{\Omega} \times [0, T] \to \mathbb{R}$ with large, strictly positive derivative in the referential parameterization. To obtain a moderate (constant) derivative, we impose $f_\alpha = \partial_s y / \int_0^1 \partial_s y(s', t)\, \mathrm{d}s'$ with $y(s, t) = \hat{y}(\Psi(s, t), t)$. The MMPDE strategy yields the same parameter density for $\hat{M} = \partial_\sigma \hat{y}$.

2 Application

The driving application behind this research are fiber spinning processes where boundary layers occur, for example due to large elongations or mass lumping.

Fiber Spinning Model The spinning of a slender viscous jet under the influence of gravity is characterized by the dimensionless Reynolds number Re (ratio between inertia and viscosity) and Froude number Fr (ratio between inertia and gravity). Consider the jet attached to a wall on one side and with a stress-free end at the other side (Fig. 2a). Its uni-axial dynamics can be described by an initial-boundary value problem for the unknown jet position r, cross-sectional area A, momentum-associated velocity v, inner force (stress) n, elongation e and parameter speed u in $(0, 1) \times [0, T]$

$$\partial_t r = v - ue, \qquad \partial_s r = e, \tag{3a}$$

$$\partial_t A + \partial_s(uA) = 0, \qquad \partial_t(Av) + \partial_s(uAv) = \frac{1}{Re}\partial_s n + \frac{1}{Fr^2}A, \tag{3b}$$

$$\partial_s v = \frac{1}{3}\frac{ne^2}{A}, \tag{3c}$$

supplemented with

$$(r, A, v)(s, 0) = (s, 1, 0), \quad (r, v, u)(0, t) = (0, 0, 0), \quad (n, u)(1, t) = (0, 0). \tag{3d}$$

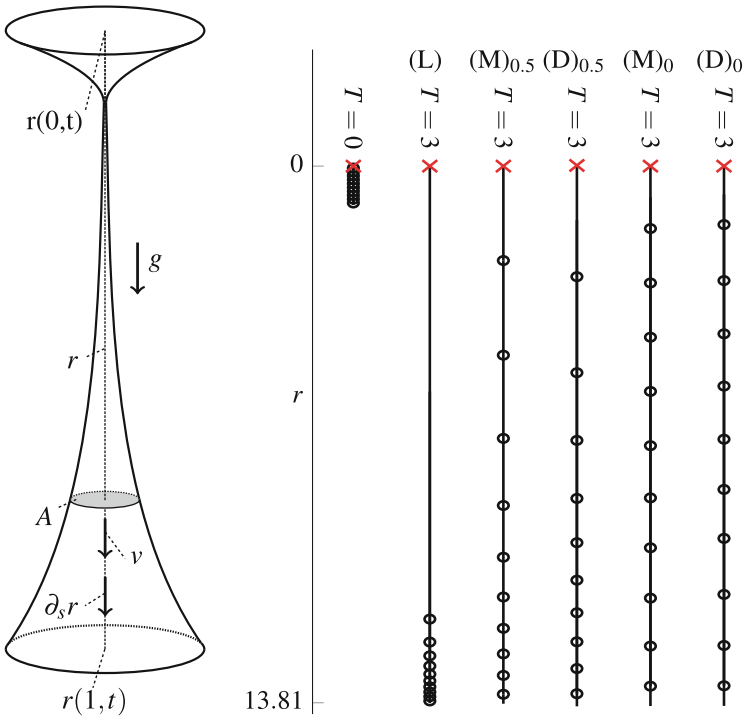


Fig. 2 (**a**) *left:* Jet illustration; (**b**) *right:* Simulation of r and its grid points at time $T = 0$ and $T = 3$ in Lagrangian description and in the moving mesh strategies for different relaxations $\tau = 0.5, 0$ (simulation set-up: $(\mathrm{Re}, \mathrm{Fr}) = (1, 0.5)$ with step sizes $\Delta s = \Delta t = 0.1$)

Equations (3a) state the kinematics, the dynamics is given by the balance Eq. (3b) with the material law in (3c). For fiber spinning the model equations are typically formulated in Lagrangian or Eulerian parameterization (see e.g. [1, 2] for the more sophisticated 2d or 3d rotational spinning). However, to allow for a moving mesh, we consider the dynamics in a computational parameterization $\Omega = [0, 1]$: by containing the unknown parameter speed u, system (3) is not complete yet. There are different possibilities to fix the remaining degree of freedom:

(L) $u \equiv 0$: Lagrangian (material) description (cf. [1])
(E) $e(s,t) = r(1,t)$: scaled Eulerian (spatial) description
$(\mathrm{D})_\tau$ u is modeled as parameter speed using (4) with Ψ determined by DELAX
$(\mathrm{M})_\tau$ u is modeled as parameter speed using (4) with Ψ determined by MMPDE.

$$u(s,t) = -\frac{\partial_t \Psi(s,t)}{\partial_s \Psi(s,t)} \tag{4}$$

Numerical Results The underlying numerical method of choice is a first order finite volume scheme in space (leading to a DAE system) with an implicit Euler method in time [1, Sect. 3] for (3). The resulting nonlinear system is solved

with a Newton method. The flux discretization is performed down-winded for the convective terms and the stress in the balance equations, the other spatial derivatives in (3) are treated in a up-winded manner. In case of moving mesh, the two boundary conditions for u keep the parameterization domain time-independent. The finite volume method is applied to MMPDE with central differences according to [7]. Since the considered parameter density f_α (or $\hat{M}$ respectively) are in general complicated functionals on the solution, analytical derivatives are not available. Hence, a fully implicit time integration may slow down the Newton method due to the need for numerical gradients. We use a semi-implicit time scheme: solely f_α ($\hat{M}$, respectively) is evaluated explicitly. This treatment is reasonable since it only creates an additional temporal delay in the grid similar to the relaxation effect of the τ.

Simulating the jet dynamics in the Lagrangian (material) description (L) shows already for moderate Reynolds and Froude numbers a lumping of the grid points towards the jet end $s = 1$. Moreover, large elongations (large derivatives $\partial_s r = e$) arise at $s = 0$, see Fig. 2b for the example $(\mathrm{Re}, \mathrm{Fr}) = (1, 0.5)$. In the inviscid case $\mathrm{Re} \to \infty$, the model Eq. (3) describe the free fall where the jet position r becomes even discontinuous: $r(s,t) = \mathrm{Fr}^{-2}t^2 + s$ has a jump because of the imposed boundary condition $r(0,t) = 0$. Whereas (L) resolves the mass lumping at the jet end optimally (arising boundary layer since $A(1,t) = A(1,0)$), it is not suitable for the refinement of the large elongations. Here, the Eulerian description (E) yielding $e(\cdot,t) = const.$ is obviously preferable. We apply the moving mesh strategies to increase the approximation quality by rearranging the grid points. To approximate a scaled Eulerian description we choose the parameter density f_α as

$$f_\alpha(s,t) = \frac{\partial_s r}{\int_0^1 \partial_s r(s',t)\,\mathrm{d}s'}$$

(or $\hat{M}(\sigma,t) = \partial_\sigma \hat{r}$ with $\hat{r}(\sigma,t) = r(\Psi^{-1}(s,t),t)$, respectively). The positivity of f_α is ensured due to the initial condition for r, so no node crossing is suspected. In addition to the results of (L), Fig. 2 also shows the redistribution of the grid points for $(\mathrm{D})_{0.5}$ and $(\mathrm{M})_{0.5}$ at time $T = 3$. The effect of the moving mesh strategies over time can be seen in Fig. 3.

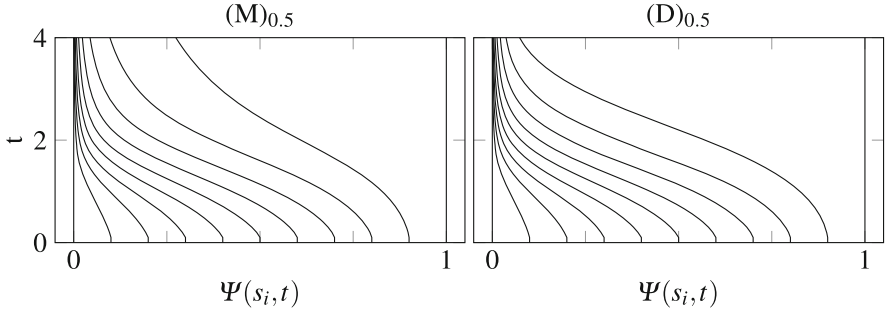


Fig. 3 Equidistant computational grid points $s_i = i/N$, $i = 0, \ldots, N$ with mesh movement $\sigma_i = \Psi(s_i,t)$ in the referential parameterization over time for $(\mathrm{M})_{0.5}$ and $(\mathrm{D})_{0.5}$ (cf. Fig. 2b)

The grid points move to the beginning of the referential parameter domain. In the limit $\tau \to 0$, both strategies yield an almost constant derivative $\partial_s r = e$ as desired (Fig. 2b). Note that the small observed variations in the elongation come from the semi-implicit time scheme which causes a temporal delay. Using a fully implicit time integration the elongation stays constant over the domain for all times.

3 Conclusion

For 1d PDEs we proposed a general moving mesh framework based on three parameterization layers and studied two adaption strategies in the application of fiber spinning. In comparison to the existing sophisticated MMPDE strategy [5], a relatively easy evolution equation for the mesh movement is employed in our DELAX approach. Our numerical studies show that both strategies yield similar results. The possibilities to control the mesh are very promising. The framework provides further ideas on how to model adaption strategies. An extension to time-dependent parameterizations is in work, see e.g. [6].

Acknowledgements This work has been supported by Deutsche Forschungsgemeinschaft (DFG): MA 4526/2-1, WE 2003/4-1.

References

1. Arne, W., Marheineke, N., Meister, A., Wegener, R.: Finite volume approach for the instationary Cosserat rod model describing the spinning of viscous jets. arXiv:1207.0731 (2012)
2. Arne, W., Marheineke, N., Meister, A., Wegener, R.: Numerical treatment of non-stationary viscous Cosserat rod in a two-dimensional Eulerian framework. In: Progress in Industrial Mathematics at ECMI 2012. Springer, Heidelberg (2014)
3. Huang, W., Russell, R.: Analysis of moving mesh partial differential equations with spatial smoothing. SIAM J. Numer. Anal. **34**(3), 1106–1126 (1997)
4. Huang, W., Russell, R.D.: Adaptive Moving Mesh Methods. Applied Mathematical Sciences, vol. 17. Springer, Heidelberg (2010)
5. Huang, W., Ren, Y., Russell, R.D.: Moving mesh partial differential equations (MMPDES) based on the equidistribution principle. SIAM J. Numer. Anal. **31**(3), 709–730 (1994)
6. Schiessl, S., Arne, W., Marheineke, N., Wegener, R.: An adaptive moving mesh approach for hyperbolic conservation laws on time-dependent domains. Proc. Appl. Math. Mech. (PAMM) **14**, 957–958 (2014)
7. Stockie, J.M., Mackenzie, J.A., Russell, R.D.: A moving mesh method for one-dimensional hyperbolic conservation laws. SIAM J. Sci. Comput. **22**(5), 1791–1813 (2001)

Construction of Virtual Non-wovens

Axel Klar, Christian H. Neßler, and Christoph Strohmeyer

Abstract We present a method for the computational construction of virtual non-woven materials in the textile industry. The underlying model is a surrogate model for the lay-down process of a single fibre described by stochastic differential equations. In particular, we illustrate a computational method of constructing a virtual non-woven material from thousands of single fibres. Furthermore, we show a way of identifying contact points between the fibres. These contact points play an essential role in the corresponding fibre network, which is the basis for virtual material testing.

Keywords Airlay process • Fiber lay-down • Fiber network • Nonwoven manufacturing

1 Introduction

Many products like filters are non-woven materials. We consider the case where the material consists of many thousands of relatively short fibres, so called *staple fibres*. Usually the product is a mixture of several different materials, for example, cotton or polymers. The fibres lay down on a conveyor belt where they form the material structure. To fix this structure, a post processing step is needed, which in our case is *thermobonding*. The material is heated so that the polymer fibres melt and stick together at *contact points*. The production process is highly involved and so further improvements can both save money and increase the quality of the products. Due to the complexity of the production process, it is not possible to describe all aspects by a model of first principles. Therefore, in the project *OPAL*, the research on this topic

A. Klar • C.H. Neßler (✉)
Department of Mathematics, Technische Universität Kaiserslautern, Erwin-Schrödinger-Straße, 67663 Kaiserslautern, Germany
e-mail: klar@mathematik.uni-kl.de; nessler@mathematik.uni-kl.de

C. Strohmeyer
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Chair of Applied Mathematics 2, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: strohmeyer@math.fau.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_134

has been separated into three main parts: the fibre flow in a turbulent air stream, the lay-down process on the belt and virtual material testing.

In this paper we concentrate on a computational method of generating virtual fibre webs based on a stochastic differential equation model. The aim is to describe the fibre behaviour in the lay-down area in such a way that it is accurate enough, but also possible to run simulations in a reasonable time frame. In addition, we describe an efficient method for the contact point identification.

2 Fibre Lay-Down Model

We begin by outlining a stochastic differential equation model for the lay-down of a single fibre which includes anisotropic behaviour as well as transport with respect to a reference curve. The underlying model for our fibre simulation was developed for the lay-down process in melt-spinning. We direct the reader to [1] for an overview of this research area and existing two dimensional models, and to [2, 3] for a detailed analysis of the three dimensional model considered here.

Let us assume that we want to simulate a non-woven material consisting of thousands of fibres with identical material properties, for example, length, diameter and the number of crimps of each fibre. Furthermore, we assume that there is no direct interaction between the fibres. Thus, we consider the evolution of the lay-down scenario for each fibre individually. The fibre position is described by a parametrised curve $\xi : [0, T) \to \mathbb{R}^3$ and its orientation by $\tau : [0, T) \to \mathbb{S}^2$, where $\mathbb{S}^2$ is the unit sphere, i.e. $\|\tau\| = 1$. This implies that T is the fibre length. External forces, generated by a potential $V : \mathbb{R}^3 \to \mathbb{R}$, are allowed to act on the fibre. We include a Brownian motion W_t and a noise amplitude $A > 0$ which allows us to control the entanglement of the fibre. The larger the noise, the larger is the entanglement. Moreover, we introduce the reference curve $\gamma : \mathbb{R} \to \mathbb{R}^3$ which is usually chosen as $\gamma(t) = -\psi e_1$, where e_i, $i = 1, 2, 3$ is the canonical basis and ψ is the ratio of production speed to the transport speed of the belt.

With the above, the governing equations for a single fibre with initial conditions $(\xi_0, \tau_0) \in \mathbb{R}^3 \times \mathbb{S}^2$ are

$$\begin{aligned} \mathrm{d}\xi &= \tau \, \mathrm{d}t \,, \\ \mathrm{d}\tau &= -\frac{1}{B+1} \left(n_1 \otimes n_1 + B \, n_2 \otimes n_2\right) \nabla_\xi V(\xi - \gamma) \mathrm{d}t \\ &\quad + A \left(n_1 \otimes n_1 + \sqrt{B} n_2 \otimes n_2\right) \circ \mathrm{d}W_t, \end{aligned} \tag{1}$$

where $\otimes$ denotes the tensor product and $\circ$ Stratonovich integration [1–3]. Here n_1 and n_2 are the normal and binormal to the fibre curve respectively and $B \in [0, 1]$ a

parameter of anisotropy. We note that, due to construction, $n_1 \in \text{span}\{e_1, e_2\}$. This implies that for $B = 0$ we have a process which lives in the $e_1 - e_2$ plane, assuming appropriate initial conditions, whereas for $B = 1$ we obtain the fully isotropic model.

3 Layer Building

Non-woven materials can have a rather complex structure due to the production process and the requirements from industry. Therefore it is important to consider the building of these structures, for example layers, within the virtual generation of fibre webs.

The evolution equation (1) requires appropriate initial conditions $[(x_0, y_0, z_0), \tau_0]$. The initial values x_0 and y_0 are chosen from a distribution provided by our project partners from the transport group of Fraunhofer ITWM, who simulate the turbulent air stream and the fibre transport within this air stream until the fibres are close to the conveyor belt. For related work in melt-spinning, see, for example, [4] and [5].

What remains is to determine z_0. This strongly depends on the material already laid down on the belt. Of course, not all fibres already simulated are relevant, but only those close to the incoming new fibre we want to describe.

Let us assume that we already have simulated N fibres, whose positions are described by ξ^j, $j = 1, \ldots, N$. Then the task is to compute the initial point ξ_0^{N+1} for the $(N+1)$th fibre.

In the single fibre simulations based on (1) we assume a constant discretisation step size Δ for all fibres, which results in a uniform grid $t_1, \ldots, t_M$, $M \in \mathbb{N}$. For $\epsilon > 0$ we define

$$Z_\epsilon(x_0, y_0) := \{(x, y, z) \in \mathbb{R}^3 \big| (x - x_0)^2 + (y - y_0)^2 < \epsilon^2,\ z \in \mathbb{R}\}, \tag{2}$$

which we call the *local (ϵ-) cylinder* for $(x_0, y_0) \in \mathbb{R}^2$. As ϵ increases, the domain describing the layer building also gets larger. Then we define

$$K := \sum_{j=1}^{M} \sum_{i=1}^{N} \left| \{\xi^i(t_j) \in Z_\epsilon(x_0, y_0)\} \right|, \tag{3}$$

the number of already simulated fibre points in the local cylinder. This is illustrated in Fig. 1. We make the assumption that the fibre volume at each discretisation point $\xi^i(t_j) \in Z_\epsilon(x_0, y_0)$ can be approximated by a ball of radius r, where r is the actual fibre radius. This assumption is meaningful as long as Δ is sufficiently small and

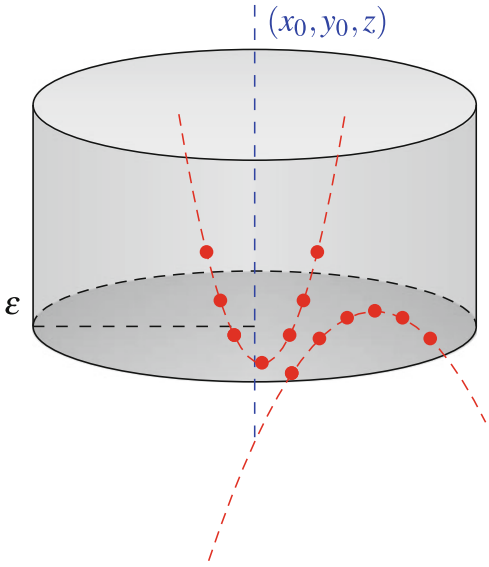


Fig. 1 Local cylinder around (x_0, y_0) with radius ϵ where z_0 is chosen as described by (4). The fibre curves with discretisation points in the cylinder are indicated in *red*

the fibre radius is much smaller than the fibre length. Then we define

$$z_0 = \beta \frac{4/3\pi r^3 K \Delta}{\pi \epsilon^2} = \beta \frac{4r^3 K \Delta}{3\epsilon^2}, \tag{4}$$

where we scale by the cross-section area of $Z_\epsilon(x_0, y_0)$ and take the discretisation step size Δ into account. This approximation describes the situation of compactly distributed balls within $Z_\epsilon(x_0, y_0)$ and close to the belt quite well. In general such a compact distribution is not the case. Real materials can have a much more complicated structure and can have a higher relative volume. This fact is considered by the parameter β in (4), which varies for different materials.

4 Contact Point Identification

Identifying material properties from the simulations described in the previous sections plays an important role in real-life applications. So far we have considered the material as a number of individual fibres rather than a connected fibre net. The idea is to identify the contact points of the fibres and to generate a graph, where the contact points are interpreted as nodes. This graph, corresponding to the virtual non-woven simulated, will then form the basis for virtual material tests, see [6] and [7].

Assuming a constant discretisation step size Δ and M discretisation points for each individual fibre, with a total of N fibres, comparing all data points with another has the complexity $\mathcal{O}((N \cdot M)^2)$. As we consider in general a large number

of fibres and a small step size, a direct point wise comparison would be by far computationally too expensive. Therefore one should consider the simulation data for one fibre as one object and take advantage of this consideration.

For our purposes the so called *bounding box* method is very promising. It is well-known in computer graphics and has multiple applications, for instance in ray-tracing. For more sophisticated methods and an overview on this topic, we refer to [8] and references therein.

The basic idea of the *bounding box* method is to use a simple geometry which contains the object of interest. In our case this is just a box aligned to the coordinate axes for each single fibre. Then we check pairwise if those N boxes intersect or not. If the box-intersection is empty, then we know also that the fibres contained in the boxes must be disjoint and so there are no contact points. However, if the intersection of the boxes is not empty this does not mean that the fibres do intersect. Then we have to consider the parts of the fibres within the intersection box and iterate.

We illustrate our approach by a simple two dimensional example in Fig. 2. In the first step (on the left hand side) we see that the bounding boxes for the red and blue curves intersect, resulting in the black intersection box. In the second step (on the right hand side) the fibre data we need to consider is reduced to the black dashed box. As soon as the fibre data is reduced enough, one can perform a point wise comparison, which is relatively cheap in computational costs, and identify the contact points.

An example of contact point identification considering simulated fibre data is illustrated in Fig. 3. The image shows the situation within a reference volume: only the fibre parts lying in the reference volume are considered. Then the contact point identification with bounding boxes is performed. Red circles indicate both the contact points and intersections with the edges of the reference volume.

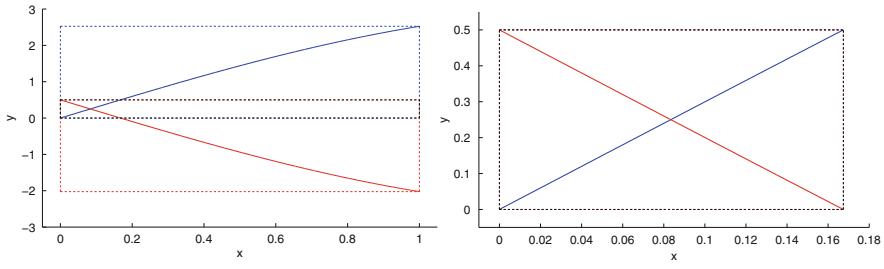


Fig. 2 A simple example for iterated bounding boxes. Note the change in the domain from the *left* to the *right* image

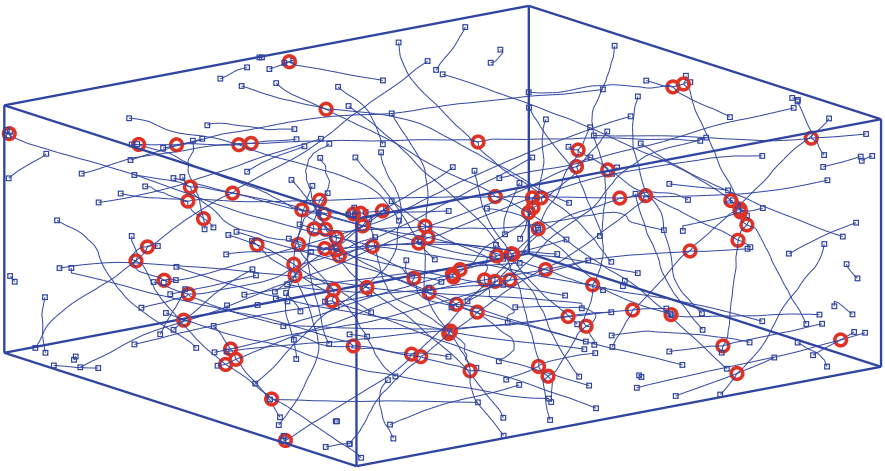


Fig. 3 Fibres, shown as *blue lines*, and their contact points, indicated as *red circles*, within a reference volume

5 Conclusion

We introduced a computational method for the simulation of virtual non-woven materials. This can incorporate the information from turbulent air stream simulations into the fibre web generation. Furthermore, we described a technique for contact point identification based on a bounding box approach. These ideas build the basis for the simulation of a production process of non-woven materials that aims for the optimisation of material properties with respect to process parameters.

Acknowledgements We thank all our partners in the project OPAL. This work has been supported by the German BMBF, Project OPAL 05M2013.

References

1. Klar, A., Marheineke, N., Wegener, R.: Hierarchy of mathematical models for production processes of technical textiles. Z. Angew. Math. Mech. **89**(12), 941–961 (2009)
2. Klar, A., Maringer, J., Wegener, R.: A 3D model for fiber lay-down in nonwoven production processes. Math. Models Methods Appl. Sci. **22**, 9 (2012)
3. Klar, A., Maringer, J., Wegener, R.: A smooth 3D model for fiber lay-down in nonwoven production processes. Kinet. Relat. Models **5** 1, 97–112 (2012)
4. Marheineke, N., Wegener, R.: Modeling and application of a stochastic drag for fiber dynamics in turbulent flows. Int. J. Multiphase Flow **37**, 136–148 (2011)
5. Marheineke, N., Wegener, R.: Fiber dynamics in turbulent flows: general modeling framework. SIAM J. Appl. Math. **66**, 1703–1726 (2006)
6. Lebée, A., Sab, K.: Homogenization of a space frame as a thick plate: application of the Bending-Gradient theory to a beam lattice. Comput. Struct. **127**, 88–101 (2013)

7. Langnese, J., Leugering, G., Schmidt, E.: Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures. Springer Science + Business Media, LLC, New York (1994)
8. Chang, C., Gorissen, B., Melchior, S.: Fast oriented bounding box optimization on the rotation group $SO(3, \mathbb{R})$. ACM Trans. Graph. (TOG) **30**(5), 122:1–6 (2011)

Effective Mechanical Properties of Nonwovens Produced by Airlay Processes

Christoph Strohmeyer and Günter Leugering

Abstract The mechanical properties of nonwoven materials are investigated and optimized. On the micro scale these fabrics are modeled as network of beams. Here linear Timoshenko beams and geometrically exact beams are compared in simple tension tests. Then a homogenization scheme is used to calculate effective material tensors of Kirchhoff-Love plates, where periodically repeatable representative volume elements containing fiber networks define the micro structure. Finally, these effective properties are optimized by changing the shape of the underlying network.

Keywords Airlay process • Beam models • Cosserat rod model • Effective mechanical properties • Fiber network • Nonwoven manufacturing

1 Introduction

In modern textile industry it is of great importance to fabricate custom-tailored products which have to meet specific standards and should fulfill special requirements such as high toughness. The project OPAL (**OP**timization of **A**ir**L**ay processes) investigates the industrial process including the single fiber traveling in turbulent air flow, laying down onto a ramp of loosely connected fibers and finally forming networks being glued together by thermo bonding. The resulting inner structure determines effective material properties of the nonwoven textile. This work is concerned with the latter part of the process. On the micro scale stochastic fiber networks have to be tackled, both from a modeling and a numerical point of view. As first approaches direct (non-)linear simulation as well as energetic homogenization, suggested by [8], are used to obtain information about the macroscopic behavior of fiber networks. Subsequently the homogenization procedure and optimization are combined, which is to be seen as a first step towards controlling industrial relevant scenarios, the ultimate goal of the OPAL project.

C. Strohmeyer (✉) • G. Leugering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Chair of Applied Mathematics 2, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: strohmeyer@math.fau.de; leugering@math.fau.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_135

2 Networks of Beams

Beam models can be derived by inserting an ansatz on how beam-like objects can deform into the 3D equations of continua and subsequently averaging these equations over the cross section. This so-called constrained motion (describing the displacement of a material point x of the beam), derived due to assumptions because of their special geometric properties, has for Timoshenko-like beams in general the form

$$u^{\text{full}}(x) \;=\; u(x_1) + \big(\Lambda(x_1) - I\big)x_A.$$

Here the vector x_A lies in the cross section of the beam, x_1 parameterizes its centerline and $\Lambda \in \mathbb{R}^{3\times 3}$ characterizes the rotation of the cross section. Now there are two extreme cases of beam models:

- The geometrically exact (GE) beam, obtained by taking $\Lambda \in SO(3)$, e.g. parameterized by its rotation vector $\phi \in \mathbb{R}^3$ (see [6]), and nonlinear continuum equations, see e.g. [9].
- The linear Timoshenko (TS) beam, obtained by linearizing $\Lambda(\phi) \in SO(3)$, so that $u^{\text{full}}(x) = u(x_1) + \phi(x_1) \times x_A$, together with the linear continuum equations.

For initially straight beams of length L, oriented along the x_1-axis, we obtain the following similar looking but in terms of complexity very different systems of differential equations, both holding for $x_1 \in (0, L)$:

TS beam	GE beam
$N' + \bar{N} = 0$	$(\Lambda N)' + \bar{N} = 0$
$M' + E_1 \times N + \bar{M} = 0$	$(\Lambda M)' + (E_1 + u') \times (\Lambda N) + \bar{M} = 0$

(1)

All quantities are given in the local basis of the beam $\{E_1, E_2, E_3\}$, $(\cdot)'$ denotes differentiation w.r.t. x_1 and $\times$ is the cross product. If the beams are linear elastic and have a doubly symmetric cross section the constitutive law reads as:

TS beam	GE beam	material data
$N = C_N(u' + E_1 \times \phi)$	$N = C_N \Lambda^T\big(u' - (\Lambda - I)E_1\big)$	$C_N = \operatorname{diag}(EA_c, GA_s, GA_s)$
$M = C_M \phi'$	$M = C_M \operatorname{vec}(\Lambda^T \Lambda')$	$C_M = \operatorname{diag}(GI_t, EI, EI)$

(2)

The operator $\operatorname{vec}(\cdot)$ extracts the axial vector from a skew symmetric matrix and $\operatorname{diag}(\cdot)$ indicates the diagonal elements of a square matrix. The material and geometric properties of the beam are elastic and shear modulus E and G, area of cross section A_c and corrected shear area A_s. Fibers considered in this paper have circular cross sections, so that one has just one second moment of inertia $I = \int x_2^2 \mathrm{d}x_A = \int x_3^2 \mathrm{d}x_A$ and torsional constant $I_t = 2I$.

With these equations at hand a network is built up: given a global basis $\{e_1, e_2, e_3\}$, vertices $\mathscr{V} := \{v_1, \ldots, v_n\} \subset \mathbb{R}^3$and edges $\mathscr{E} := \{b_1, \ldots, b_m\}, b_i := (v_{i_1}, v_{i_2})$, defining the graph $\mathscr{G} := (\mathscr{V}, \mathscr{E})$, the i-th beam connects vertices v_{i_1} and v_{i_2} and has length $L_i = \|v_{i_2} - v_{i_1}\|$. It fulfills the differential systems (1) and (2) with corresponding sets of unknowns (u_i, ϕ_i) (TS), (u_i, Λ_i) (GE) and material matrices C_{N_i}, C_{M_i}. The local basis is given as:

$$E_{1i} := \frac{v_{i_2} - v_{i_1}}{L_i}, \quad E_{j_i} \cdot E_{1i} = 0,\ j = 2, 3 \quad \wedge \quad T_i := \left(E_{1i}, E_{2i}, E_{3i}\right) \in SO(3). \tag{3}$$

Using the transformation matrix T_i we can express local unknowns in the global basis and formulate proper transmission conditions for junctions. Prior to this we define $M_k := \{i : v_k \in b_i\}$ the index set of beams connected to node v_k and

$$x_i^k := \begin{cases} 0, & \text{if } i\text{-th beam starts in node } v_k \\ L_i, & \text{if } i\text{-th beam ends in node } v_k \end{cases} \qquad n_i^k := \begin{cases} -1, & \text{if } x_i^k = 0 \\ 1, & \text{if } x_i^k = L_i \end{cases} \tag{4}$$

It is assumed that fibers are connected rigidly, since they have been connected by thermo bonding during the industrial process, so that we have to ensure continuity of displacement and rotation over joints:

$$\left.\begin{array}{lll} \text{TS:} & T_i u_i(x_i^k) = T_j u_j(x_j^k) \quad \wedge & T_i \phi_i(x_i^k) = T_j \phi_j(x_j^k) \\ \text{GE:} & T_i u_i(x_i^k) = T_j u_j(x_j^k) \quad \wedge & T_i \Lambda_i(x_i^k) T_i^T = T_j \Lambda_j(x_j^k) T_j^T \end{array}\right\} \quad \forall i, j \in M_k \tag{5}$$

as well as balance of forces and moments at nodes that are not fixed by Dirichlet boundary conditions, see [7]:

$$\sum_{i \in M_k} n_i^k T_i N_i(x_i^k) = n_k \quad \wedge \quad \sum_{i \in M_k} n_i^k T_i M_i(x_i^k) = m_k, \tag{6}$$

with n_k, m_k being nodal loads at v_k. Note that conditions (6) are the same for TS and GE beams, just the respective constitutive law (2) has to be used. Furthermore they contain Neumann boundary conditions as special case if only a single beam is connected to node v_k. Remaining nodes are of Dirichlet type:

$$\left.\begin{array}{lll} \text{TS:} & T_i u_i(x_i^k) = u_{Dk} \quad \wedge & T_i \phi_i(x_i^k) = \phi_{Dk} \\ \text{GE:} & T_i u_i(x_i^k) = u_{Dk} \quad \wedge & T_i \Lambda_i(x_i^k) T_i^T = \Lambda_{Dk} \end{array}\right\} \quad \forall i \in M_k. \tag{7}$$

3 An Optimization Problem with Effective Material Data

In the following section an optimization problem, whose cost functional depends on material data of homogenized linear beam networks, is introduced. First we are going to describe the process of homogenization of x_1/x_2-periodic cells (representative volume element, RVE) as Kirchhoff-Love plates. Subsequently a shape optimization problem is proposed where the position of inner nodes of the RVE are design dependent and therefore can be moved freely in space.

3.1 Energetic Homogenization

At this point we are following [8] to compute homogenized material tensors with corresponding cell problems on the RVE, defined as $\Omega_\varepsilon := [0, \varepsilon_1]\times[0, \varepsilon_2]\times[-\varepsilon_3, \varepsilon_3]$. In this approach no strict homogenization in the sense of two-scale convergence, see [1], is carried out, but the Hill-Mandel principle, where the equivalence of inner energies of the micro (beams) and macro scale (plate) is required, see e.g. [5]. This has the advantage that one can simply choose the model as which the fiber net is going to be homogenized—in this case a Kirchhoff-Love (KL) plate. Starting point is the constitutive law of the plate

$$N^{\mathrm{KL}} = A : e + B : \chi, \qquad M^{\mathrm{KL}} = B^T\!: e + D : \chi,$$

that relates the stress resultants $N \in \mathbb{R}^{2\times 2}$ (collecting normal and shear forces) and $M \in \mathbb{R}^{2\times 2}$ (collecting bending and torsional moments) to its strain and curvature, $e, \chi \in \mathbb{R}^{2\times 2}$. The fourth order tensors $A, B, D \in \mathbb{R}^{2\times 2\times 2\times 2}$ are to be calculated from the base cell containing a fiber net. To use the Hill-Mandel principle we have to equate the energy density of the beams (inner energy of the fiber net in the RVE averaged over its in-plane area $\mathscr{A} = \varepsilon_1\varepsilon_2$) to the energy density of a KL-plate

$$\frac{1}{\mathscr{A}} \sum_{i=1}^{m} \frac{1}{2} \int_0^{L_i} N_i^T C_{N_i}^{-1} N_i + M_i^T C_{M_i}^{-1} M_i \mathrm{d}s_i = \frac{1}{2}(e\!:\!A\!:\!e + 2e\!:\!B\!:\!\chi + \chi\!:\!D\!:\!\chi)\,.$$

If the inner forces/moments of the network in the RVE can be constructed as

$$N_i(e, \chi) = N_i^e : e + N_i^\chi : \chi, \quad M_i(e, \chi) = M_i^e : e + M_i^\chi : \chi, \tag{8}$$

with e and χ (constant in Ω_ε) applied as a special loading scenario to the network, it's obvious that the homogenized material data of the plate can be written as

$$\frac{1}{\mathscr{A}} \sum_{i=1}^{m} \int_0^{L_i} (N_i^{\omega_1})^T \cdot C_{N_i}^{-1} \cdot N_i^{\omega_2} + (M_i^{\omega_1})^T \cdot C_{M_i}^{-1} \cdot M_i^{\omega_2} \mathrm{d}s_i, \tag{9}$$

where (9) yields tensor A for $\omega_{1/2} = e$, D for $\omega_{1/2} = \chi$ and B for $\omega_1 = e, \omega_2 = \chi$. In [8] the cell problem is constructed by applying the constrained motion

$$u^{\mathrm{KL}}(x) = \hat{e}x + x_3\hat{\chi}x - \tfrac{1}{2}(x^T\hat{\chi}x)e_3, \quad \phi^{\mathrm{KL}}(x) = \tfrac{1}{2}\nabla \times u^{\mathrm{KL}}(x) = e_3 \times \hat{\chi}x$$

of a KL plate, which can be evaluated in particular at the center line of the i-th beam $\bar{x}_i(x_1) := v_{i_1} + x_1 E_{1i} \in \Omega_\varepsilon$ ($x_3 = 0$ being the neutral plane of the plate) and

$$\hat{e} = \begin{pmatrix} e & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{\chi} = \begin{pmatrix} \chi & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{3\times 3},$$

in average to the RVE allowing for periodic deviations $u_i^{\mathrm{per}}, \phi_i^{\mathrm{per}}$:

$$\begin{aligned} C_{N_i}(u_i' + E_{1i} \times \phi_i)' = 0, \qquad & u_i(e,\chi) = T_i^T u^{\mathrm{KL}}(\bar{x}_i(x_1)) + u_i^{\mathrm{per}} \\ C_{M_i}\phi_i'' + E_{1i} \times C_{N_i}(u_i' + E_{1i} \times \phi_i) = 0, \qquad & \phi_i(e,\chi) = T_i^T \phi^{\mathrm{KL}}(\bar{x}_i(x_1)) + \phi_i^{\mathrm{per}} \end{aligned} \tag{10}$$

for $x_1 \in (0, L_i)$, together with continuity (5) of the **periodic part** $u_i^{\mathrm{per}}, \phi_i^{\mathrm{per}}$ and balance conditions (6). These conditions are enriched by enforcing continuity and balance also for periodic sets of points (periodicity in x_1- and x_2-direction) lying on faces of $\partial\Omega_\varepsilon$. To be physically meaningful (i.e. fiber net is physically connected at periodic points), all points belonging to the same periodic set have to fulfill:

$$\exists k_1, k_2 \in \{-1, 0, 1\}: \quad v_i = v_j + \left(k_1\varepsilon_1,\, k_2\varepsilon_2,\, 0\right)^T. \tag{11}$$

One node of the neutral plane has to be fixed in space via (7) to prevent rigid motions of the beams. Now it is clear that the third order tensors $N_i^{e/\chi}, M_i^{e/\chi}$ in (8) can be constructed by solving (10) where only a single component of e or χ is set to 1:

$$\left(N_i^{e/\chi}\right)_{k\beta\gamma} = (N_i)_k(e,\chi), \quad \left(M_i^{e/\chi}\right)_{k\beta\gamma} = (M_i)_k(e,\chi), \qquad \text{if } (e/\chi)_{\beta\gamma} = 1. \tag{12}$$

3.2 A Shape Optimization Problem

As indicated in the introductory part of this chapter the cost functional of the optimization problem contains entries of the homogenized material tensors (9). Of course there is also the possibility to penalize changes of shape or e.g. stress of the network in the RVE, so that also the states and the design itself may be argument of

the objective functional

$$J(A,B,D,u,\phi,\alpha) \quad \Big(\rightsquigarrow \quad A = A(u,\phi) \quad \rightsquigarrow \quad u = u(\alpha) \Big). \tag{13}$$

As implied by (13), A, B and D depend on solutions of cell problems (10). These solutions again depend on the design variables α coming into the problem as displacements of inner nodes v_i, see e.g. [2], of the underlying graph in the RVE [points belonging to periodic sets still need to fulfill (11)]:

$$v_i(\alpha) = v_i^0 + \alpha_i \qquad \rightsquigarrow \qquad \text{graph } \mathscr{G}(\alpha) := (\mathscr{V}(\alpha), \mathscr{E}). \tag{14}$$

This way the topology of the network remains unchanged but its shape can change completely. To avoid singularities and physically undesirable situations, constraints on the length of all beams as well as box constraints to the design α are imposed:

$$0 < \underline{L}_j \le L_j(\alpha) \le \overline{L}_j, \quad \begin{pmatrix} 0 \\ 0 \\ -\varepsilon_3 \end{pmatrix} - v_i^0 \le \underline{\alpha}_i \le \alpha_i \le \overline{\alpha}_i \le \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} - v_i^0. \tag{15}$$

Additionally, resource constraints are introduced to the problem:

$$\underline{V} \le \sum_{i=1}^{m} \lambda_i L_i(\alpha) \le \overline{V}, \qquad \text{physical volume for } \lambda_i = A_{c_i}. \tag{16}$$

Finally, we state the shape optimization problem:

$$\left.\begin{aligned} &\min_{\alpha} j(\alpha) := J\Big(A(\alpha), B(\alpha), D(\alpha), u(\alpha), \phi(\alpha), \alpha\Big) \\ &\text{s.t.} \quad A, B, D \text{ from (9) with (12) in which} \\ &\qquad (u,\phi)(e,\chi) \text{ solves (10) on graph } \mathscr{G}(\alpha) \text{ (14)} \\ &\qquad \alpha \text{ satisfies (15), (16)} \end{aligned}\right\} \tag{17}$$

Gradient information is calculated via adjoint calculus, see e.g. [4].

4 Results

Numerical results are obtained by Finite Element discretization and `snopt`, an optimization software for constrained problems, see [3]. Material data of the beams is approximately taken from polyester micro fibers: $E = 3.0 \times 10^9\ \mathrm{N/m^2}$, Poisson's

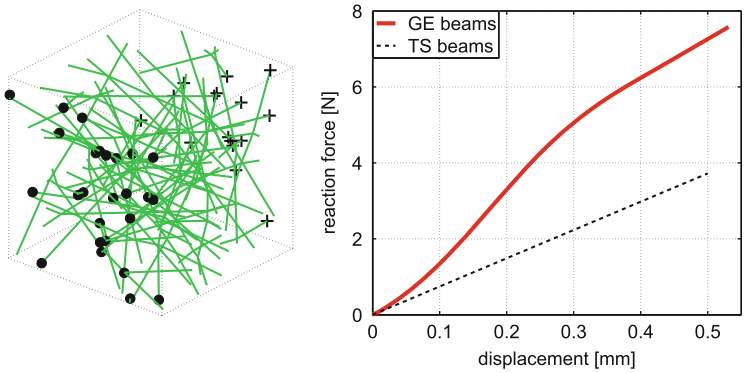


Fig. 1 Numerical setup (*left*) and displacement-force-diagram (*right*). At the face with *circle*—markers homogeneous and at *plus*—markers inhomogeneous Dirichlet conditions are applied and reacting (normal) force is calculated (in the nonlinear case for every load increment)

ratio $\nu = 0.45$ and diameter $D = 25 \times 10^{-6}$ m. With these quantities all entries of C_N and C_M, see (2), can be calculated.

4.1 Tension Tests with Stochastic Fiber Networks

Here a tension experiment is mimicked: On two opposing faces of the unit cell $\Omega_\varepsilon = [0, 0.5\,\text{mm}]^3$ a constant Dirichlet displacement in normal direction is applied and the reacting force in normal direction on one of these faces calculated. In Fig. 1 TS and GE beam models are compared on a stochastic net with 251 nodes and 262 beams, discretized by 1908 four-noded 3D-elements (34, 026 DoFs). The asymptotic nature for small strains is clearly evident, as well as the stiffer behavior of the GE model due to the aligning of the fibers in pulling direction. The influence of crimping and different densities of contact points is yet to be investigated.

4.2 Optimization

In Fig. 2 we see a RVE ($\Omega_\varepsilon = [0, 1\,\text{mm}]^3$) consisting of a cube connected to the corners (initial configuration) and the optimized structure (maximal stiffness in x_1-direction). Figure 3 shows the result if a structure is optimized regarding transversal contraction of the plate. For the results it was set $\overline{V} = L_j = \infty$, see (16), (15).

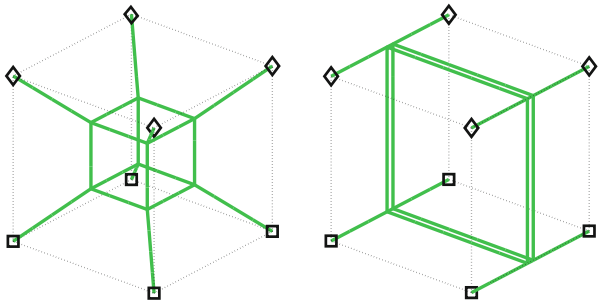


Fig. 2 3D scenario—initial configuration (*left*) and optimized cell structure (*right*) for $J = -A_{1111}$: $j(\alpha^0) = -13.01, j(\alpha^{\text{opt}}) = -5.88 \times 10^3 = -4EA_c \times 10^3$ (four fibers per RVE aligned exactly in x_1-direction). Different markers (*square*/*diamond* on *top*/*bottom*) indicate the two different periodic sets. The minimal length constraint can be seen in the result

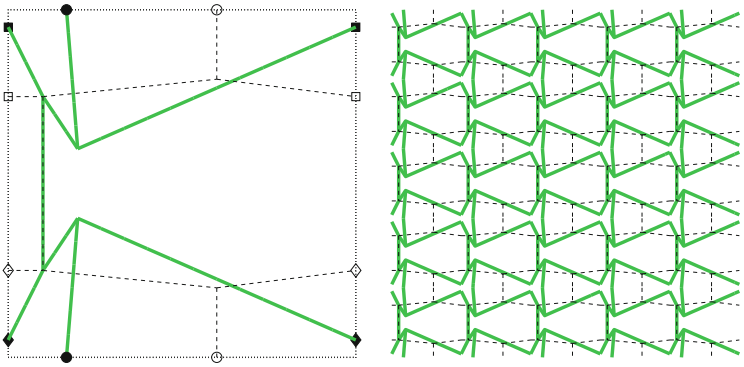


Fig. 3 Auxetic 2D structure—initial (*dashed*) and optimized (*solid*) RVE (*left*) and periodic layer (*right*) for $J = \frac{A_{1122}}{A_{1111}}$: $j(\alpha^0) = 0.11, j(\alpha^{\text{opt}}) = -1.47$

5 Conclusion

The tension tests show the robustness of the GE beam model even for very unstructured and relatively large 3D networks. Since these situations can be handled very well they are going to be used to identify relevant parameters of stochastic nets which then are to be optimized including more aspects of the industrial process.

The presented optimization problem is a very good starting point for further investigations since homogenization and optimization are combined, although it is not exactly in focus of manufacturing nonwovens by airlay processes. The scenarios are chosen such that there are somewhat expectable solutions which are indeed found by the optimization, like the aligning of the fibers in the direction of maximal stiffness and the well-known auxetic honeycomb structure.

Acknowledgements This work has been supported by the German BMBF, Project OPAL 05M2013.

References

1. Allaire, G.: Homogenization and two-scale convergence. SIAM J. Math. Anal. **23**(6), 1482–1518 (1992)
2. Bendsøe, M., Ben-Tal, A., Zowe, J.: Optimization methods for Truss geometry and topology design. Struct. Optim. **7**, 141–159 (1994)
3. Gill, P., Murray, W., Saunders, M.: SNOPT: an SQP algorithm for large-scale constrained optimization. SIAM Rev. **47**(I), 99–131 (2005)
4. Haslinger, J., Mäkinen, R.A.E.: Introduction to Shape Optimization: Theory, Approximation, and Computation. SIAM, Philadelphia (2003)
5. Hohe, J., Becker, W.: Determination of the elasticity tensor of non-orthotropic cellular sandwich cores. Technische Mechanik **19**(4), 259–268 (1999)
6. Ibrahimbegovic, A.: On the choice of finite rotation parameters. Comput. Methods Appl. Mech. Eng. **149**, 49–71 (1997)
7. Langnese, J., Leugering, G., Schmidt, E.: Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures. Springer Science + Business Media, LLC, New York (1994)
8. Lebée, A., Sab, K.: Homogenization of a space frame as a thick plate: application of the Bending-Gradient theory to a beam lattice. Comput. Struct. **127**, 88–101 (2013)
9. Muñoz Romero, J.: Finite-element analysis of flexible mechanisms using the master-slave approach with emphasis on the modelling of joints. Ph.D. thesis, IC London (2004)

Homogenization Strategies for Fiber Curtains and Bundles in Air Flows

Thomas M. Cibis, Christian Leithäuser, Nicole Marheineke, and Raimund Wegener

Abstract In non-woven manufacturing thousands of slender fibers are swirled by air flows before they lay down to form a web. The fiber-fluid interactions have a crucial influence on the quality of the final product. For the purpose of an efficient and fast computation of the multi-scale, two-way coupled interaction problem, we investigate classical homogenization strategies and a new continuum approach for very long fibers suspended in a fluid flow. We compare the results with Direct Numerical Simulation (DNS) and Immersed Boundary Methods for academic examples.

Keywords Fiber-fluid interaction • Fiber dynamics • Homogenization • Immersed boundary method

1 Introduction

Fiber-fluid interactions play a crucial role in many applications, e.g. non-woven manufacturing [14], fiber spinning [5], fiber suspension flows in paper making or dry forming of pulp mats [3, 12, 24]. Their simulation based on the model of first principles suffers from the computational complexity, hence appropriate surrogate models are required. In this work we focus on the effect of a curtain or bundle of long fibers on a surrounding flow field and investigate different homogenization strategies. We present a continuum approach that is based on the description of the fibers as special Cosserat rods and results from a homogenization on the fiber length density. It is compared to classical approximations, such as Darcy's Law and Brinkman's Law for porous media, as well as to numerical results from Immersed

T.M. Cibis • N. Marheineke (✉)
Department Mathematik, Friedrich-Alexander-Universität Nürnberg-Erlangen, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: cibis@math.fau.de; marheineke@math.fau.de

C. Leithäuser • R. Wegener
Fraunhofer Institut für Techno- und Wirtschaftsmathematik, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
e-mail: leithaeuser@itwm.fraunhofer.de; wegener@itwm.fraunhofer.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_136

Boundary Methods and DNS in the academic example "forest of cylinders". We comment on the methods' utility to coupled fiber-fluid simulations in industrial problems.

2 Effect of a Fiber Curtain/Bundle on a Surrounding Flow Field

For modeling the effect of an immersed fiber curtain or bundle on a flow field we present different asymptotic and numerical strategies, proceeding from DNS for the Navier-Stokes equations (NSE). For simplicity we restrict here to a stationary incompressible flow with mass density $\rho_\star$, dynamic viscosity $\mu_\star$, velocity $\boldsymbol{v}_\star$ and pressure $p_\star$. Appropriate boundary conditions need to be supplemented. A generalization of the strategies to a dynamic set-up is possible.

DNS and Comparison Quantity (Curtain Force) $\boldsymbol{F}$ The model of first principles is an interface problem in terms of NSE where the fibers are considered as extended 3D objects Ω_i with velocity $\boldsymbol{v}_i$, $i = 1, 2, \ldots$ in the domain $\Omega_\star \subset \mathbb{R}^3$,

$$\begin{aligned} &\nabla \cdot \boldsymbol{v}_\star = 0, \qquad \rho_\star \left(\boldsymbol{v}_\star \cdot \nabla\right) \boldsymbol{v}_\star = -\nabla p_\star + \mu_\star \Delta \boldsymbol{v}_\star && \text{in } \Omega_\star^\circ := \Omega_\star \backslash \bigcup \Omega_i \\ &\boldsymbol{v}_\star = \boldsymbol{v}_i && \text{at } \Gamma_i := \partial \Omega_i \end{aligned}$$

Due to the actio-reactio principle the fibers' impact on the flow equals the aerodynamic force on the fibers $\boldsymbol{f} = \int_{\bigcup \Gamma_i} \boldsymbol{S}_\star \cdot \boldsymbol{n} \, \mathrm{d}A$ with outer normal $\boldsymbol{n}$ and Newtonian stress tensor $\boldsymbol{S}_\star = -p_\star + \mu_\star (\nabla \boldsymbol{v}_\star + \nabla \boldsymbol{v}_\star{}^T)$. However, because of the required very fine resolution, DNS is in general too memory-intensive and time-consuming and thus limited to academic problems. In view of the desired surrogate models, we hence introduce the force $\boldsymbol{F}$ on the fiber curtain $\overline{\Omega} \subseteq \Omega_\star$ with boundary $\overline{\Gamma}$, i.e. domain containing all fibers $\overline{\Omega} \supseteq \bigcup \Omega_i$, as quantity for further observations. According to the Gaussian integral theorem we have $\boldsymbol{F} := \int_{\overline{\Gamma}} \boldsymbol{S}_\star \cdot \boldsymbol{n} \, \mathrm{d}A = \int_{\Omega_\star^\circ \cap \overline{\Omega}} \nabla \cdot \boldsymbol{S}_\star \, \mathrm{d}V - \boldsymbol{f}$.

Immersed Boundary Methods Immersed Boundary Methods [18, 21] provide a simplification of the interface problem, but are still computationally demanding as they also require a fine resolution of the flow. The slender fibers are represented as 1D objects (curves) $\boldsymbol{r}_i \colon \mathscr{I}_i \subseteq \mathbb{R} \to \mathbb{R}^3$ with velocity $\boldsymbol{v}_i$, tangent $\boldsymbol{\tau}_i$ and diameter d_i, for example using the special Cosserat rod theory [4, 22]. Their impact on the flow is modeled as an external source term in the momentum equation of the flow. Proceeding from an aerodynamic drag model $\boldsymbol{f}_i^{air}$ for the force on a single fiber

(cf. e.g. [8, 16]), the fiber's force is given by a Dirac delta-shaped line source. The force of the entire fiber curtain results then from the superposition of the individual forces,

$$\nabla \cdot \boldsymbol{v}_\star = 0, \qquad \rho_\star \left(\boldsymbol{v}_\star \cdot \nabla\right) \boldsymbol{v}_\star = -\nabla p_\star + \mu_\star \Delta \boldsymbol{v}_\star + \boldsymbol{f}^{jets}, \quad \boldsymbol{f}^{jets} = \sum_i \boldsymbol{f}_i^{jet} \quad \text{in } \Omega_\star$$

$$\boldsymbol{f}_i^{jet}(\boldsymbol{z}) = -\int_{\mathscr{I}_i} \boldsymbol{f}_i^{air}\left(\boldsymbol{\tau}_i(s), \boldsymbol{v}_\star - \boldsymbol{v}_i(s), \rho_\star, \mu_\star, d_i(s)\right) \delta\left(\boldsymbol{z} - \boldsymbol{r}_i(s)\right) \mathrm{d}s$$

In the following, we consider two approaches that differ in the choice of the drag model $\boldsymbol{f}_i^{air}$: in the Numerical Strategy (NUM) the model of [16] is evaluated with flow velocities averaged over the numerical grid (see also [5]), in the Modified Strategy (MOD) a more complex variant [8] is used that requires the solving of integro-differential equations.

Homogenization Strategies Considering a fiber curtain or bundle $\overline{\Omega}$ as porous medium suggests the use of classical homogenization strategies [13, 20]. In the special case of the "forest of cylinders" [1, 2, 9], the Darcy, Darcy-like, Brinkman and "Navier-Stokes" Laws presented in (a)–(d) are derived in a homogenization procedure for an increasing number of fibers with decreasing diameter. The laws differ in the relation of fiber diameter d and neighboring distance a between the fibers. In general, the applicability range of the laws is characterized by the porosity ϕ or the solid fraction $\epsilon = 1 - \phi$ of the medium. The core of the laws is the permeability tensor $\overline{\boldsymbol{K}}$ being a measure of the porousness, it is often modeled as a function of porosity ϕ and Reynolds number Re, see e.g. [10, 15, 23]. Offside of the classical laws stands the new continuum approach that we discuss in (e). Note that the quantities associated to the fiber curtain (fiber continuum) are indicated by $\overline{\cdot}$ throughout this paper, e.g. fiber curtain's velocity $\overline{\boldsymbol{v}}$ and tangent $\overline{\boldsymbol{\tau}}$.

(a) Darcy's Law (DL). The classical Darcy's Law is probably the most well-known homogenization law, it is particularly suitable for dense porous media with $\phi \leq 0.6$ [20]. Here, NSE are solved outside the fiber curtain and the Darcy equations with flow velocity $\boldsymbol{w}_\star$ and pressure $q_\star$ in the curtain. At the curtain surface $\overline{\Gamma}$ we apply the interface conditions of Beavers and Joseph [19], whose main feature is a jump in the tangential velocity component. For the normal component it holds $\boldsymbol{v}_\star \cdot \boldsymbol{n} = \boldsymbol{w}_\star \cdot \boldsymbol{n}$,

$$\begin{aligned} &\nabla \cdot \boldsymbol{v}_\star = 0, && \rho_\star \left(\boldsymbol{v}_\star \cdot \nabla\right) \boldsymbol{v}_\star = -\nabla p_\star + \mu_\star \Delta \boldsymbol{v}_\star && \text{in } \Omega_\star \backslash \overline{\Omega} \\ &\nabla \cdot \boldsymbol{w}_\star = 0, && \boldsymbol{w}_\star - \overline{\boldsymbol{v}} = -\mu_\star^{-1} \overline{\boldsymbol{K}} \nabla q_\star && \text{in } \overline{\Omega} \end{aligned}$$

(b) Darcy-like Law (DlL). The Darcy-like Law has a similar structure to DL, but differs in scaling and modeling of the permeability [1, 2]. The inverse of a constant tensor $\boldsymbol{M}$ takes the place of $\overline{\boldsymbol{K}}$. Moreover, the flow velocity in the curtain is scaled with δ^2 where $\delta = a|\log(d/a)|^{1/2}$ depends on the ratio of fiber diameter and neighboring distance. At the curtain surface $\overline{\Gamma}$ the transition conditions of

Beavers and Joseph are used for the scaled velocity, such as $\boldsymbol{v}_\star \cdot \boldsymbol{n} = \delta^2 \boldsymbol{w}_\star \cdot \boldsymbol{n}$,

$$\begin{aligned} \nabla \cdot \boldsymbol{v}_\star &= 0, \qquad \rho_\star \left(\boldsymbol{v}_\star \cdot \nabla\right) \boldsymbol{v}_\star = -\nabla p_\star + \mu_\star \Delta \boldsymbol{v}_\star \qquad \text{in } \Omega_\star \backslash \overline{\Omega} \\ \nabla \cdot \boldsymbol{w}_\star &= 0, \qquad \boldsymbol{w}_\star - \delta^{-2} \overline{\boldsymbol{v}} = -\mu_\star^{-1} a^2 \boldsymbol{M}^{-1} \nabla q_\star \qquad \text{in } \overline{\Omega} \end{aligned}$$

(c) Brinkman's Law (BL). In Brinkman's Law, NSE are solved in the entire domain for the flow field where the curtain's impact is incorporated as an additional force term. This force depends on the curtain's permeability and is linear in the relative velocity between flow and curtain. It exclusively acts in the curtain area, using the characteristic function $\chi_{\overline{\Omega}}$. Brinkman [6] showed that a porosity $\phi \geq 0.6$ is necessary for the validity of his law. Later, the validity of the law was proven for porous media with $\phi \geq 0.95$, see [11]. In non-woven manufacturing particular attention is paid to BL since the arising fiber curtains have a high porosity due to the fibers' slenderness.

$$\begin{aligned} &\nabla \cdot \boldsymbol{v}_\star = 0, \qquad \rho_\star \left(\boldsymbol{v}_\star \cdot \nabla\right) \boldsymbol{v}_\star = -\nabla p_\star + \mu_\star \Delta \boldsymbol{v}_\star + \boldsymbol{f}_{\boldsymbol{BL}} \qquad \text{in } \Omega_\star \\ &\boldsymbol{f}_{\boldsymbol{BL}} = -\mu_\star \overline{\boldsymbol{K}}^{-1} \left(\boldsymbol{v}_\star - \overline{\boldsymbol{v}}\right) \chi_{\overline{\Omega}} \end{aligned}$$

(d) "Navier-Stokes Law" (NSL)/One-way coupling. The solution of the "undisturbed" NSE can also be seen as a homogenization law. This is particularly appropriate for highly porous media, when the fibers' effect on the flow is negligible small. Thus, NSL corresponds to a one-way coupling [8].

$$\nabla \cdot \boldsymbol{v}_\star = 0, \qquad \rho_\star \left(\boldsymbol{v}_\star \cdot \nabla\right) \boldsymbol{v}_\star = -\nabla p_\star + \mu_\star \Delta \boldsymbol{v}_\star \qquad \text{in } \Omega_\star$$

(e) Continuum approach (CA). The continuum approach [5] results from a homogenization procedure with increasing number of fibers and an ever-decreasing fiber length density. Similarly as in BL, the curtain's impact on the flow is incorporated as an force term in NSE for the flow field. In accordance to the actio-reactio principle this force equals the product of the aerodynamic force on the fiber continuum $\boldsymbol{f}^{\boldsymbol{air}}$ (force per fiber length) and the fiber length density φ (fiber length per volume). The underlying aerodynamic drag model originates from [16] (cf. NUM).

$$\begin{aligned} &\nabla \cdot \boldsymbol{v}_\star = 0, \qquad \rho_\star \left(\boldsymbol{v}_\star \cdot \nabla\right) \boldsymbol{v}_\star = -\nabla p_\star + \mu_\star \Delta \boldsymbol{v}_\star + \boldsymbol{f}_{\boldsymbol{CA}} \qquad \text{in } \Omega_\star \\ &\boldsymbol{f}_{\boldsymbol{CA}}(z) = -\varphi(z) \boldsymbol{f}^{\boldsymbol{air}} \left(\overline{\boldsymbol{\tau}}(z), \boldsymbol{v}_\star(z) - \overline{\boldsymbol{v}}(z), \rho_\star, \mu_\star, \overline{d}(z)\right) \chi_{\overline{\Omega}} \end{aligned}$$

3 Comparison of Strategies in Scenario "Forest of Cylinders"

The "forest of cylinders" is a benchmark scenario [9]. Although it is obviously academic, it gives insight in the applicability of the strategies to fiber curtains or bundles arising in industrial problems. Consider MN infinitely long fixed cylinders of diameter d that are arranged uniformly in a square array with M columns side by side and N rows behind each other with neighboring distance $2a$. For the permeability $\overline{\boldsymbol{K}}$ we use the formulas in [10, 23]; in these cases $\boldsymbol{M} = \pi \mathrm{diag}(1, 1, 1/2)$ holds [2, 9]. We investigate the acting force $\boldsymbol{F}$ in dependence on the Reynolds number Re and the solid fraction ϵ. DNS results are used as reference.

Infinitely Extended Fiber Curtain As first example we consider an infinitely extended periodic curtain ($M = \infty$, $N = 10$). Due to the high symmetry of the set-up the homogenization approaches can even be solved analytically. We have $\mathrm{Re} = d\rho_\star v^{in}_{\star n}/\mu_\star$ and $\epsilon = \pi d^2/\left(16a^2\right)$ where $v^{in}_{\star n}$ is the orthogonal (normal) inflow velocity. Figure 1 shows the typical flow behavior. The inflow velocity is slowed down by the curtain, the non-orthogonal velocity components vanish. The pressure drops in the curtain continuously.

Comparing the different strategies, the orthogonal (normal) force component onto the fiber curtain is visualized in dependence on Re and ϵ in Fig. 2. In the dimensionless consideration, the classical homogenization strategies reflect the dependence on the solid fraction well, but their results turn out to be completely independent of the Reynolds number. In the continuum approach, it is exactly the opposite. One explanation might be that the flow can not duck the curtain and that there remains a constant, not decelerated velocity due to the incompressibility in the orthogonal (normal) direction. Concerning the immersed boundary methods MOD gives satisfying results as long as ϵ is not too large. NUM, in contrast, yields the worst results. The study of the tangential directions is relatively unspectacular: MOD and NUM yield very good agreements to DNS, BL and CA behave also well.

MxN-Fiber Bundle As second example we consider a bundle of $25 \times 10 = 250$ endless fibers. This bundle behaves qualitatively like a single thick fiber (rope) of diameter $\overline{d}$, see Fig. 3. The fluid flow avoids the bundle and circulates around it. As expected, it is slowed down by the bundle. We use here the Reynolds number wrt. to the bundle diameter $\mathrm{Re}_{\overline{d}} = \overline{d}\rho_\star v^{in}_{\star n}/\mu_\star$.

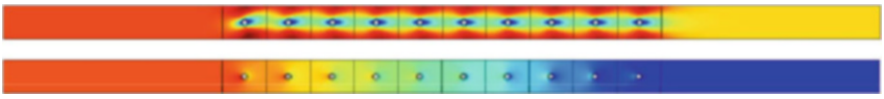


Fig. 1 Flow behavior for an infinitely extended fiber curtain. Cross-sectional view on a curtain column ($N = 10$ rows) being continued periodically up- and downwardly. *Top:* High velocity magnitude in front of the curtain (*red area*), behind the curtain only the orthogonal component remains (*yellow area*). *Bottom:* Continuous pressure decrease (from the *higher red* to the *lower blue* level)

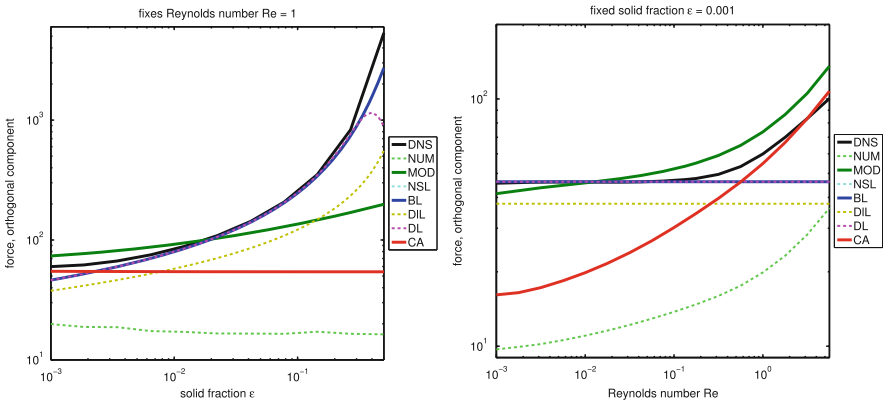


Fig. 2 Comparison of the strategies for an infinitely extended fiber curtain ($N = 10$). Normal force component onto the curtain surface, *left:* for varying ϵ and fixed $\mathrm{Re} = 1$, *right:* for varying Re and fixed $\epsilon = 0.001$

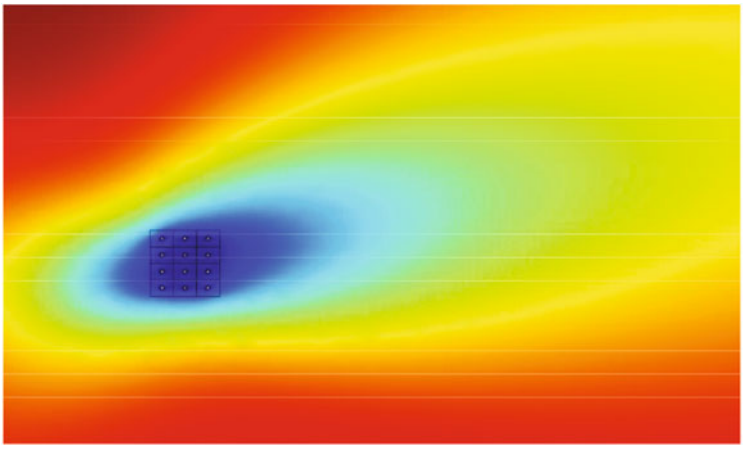


Fig. 3 Flow behavior for an immersed fiber bundle ($M = 4$ columns, $N = 3$ rows). Cross-sectional view: velocity (magnitude) is slowed down near the bundle

Apart from NSL, all strategies yield a relative deviation from the DNS reference by less than 10 % in all force components (see for example the component in direction of the rows in Fig. 4). Obviously, the results of MOD agree best, followed by BL. The reason for the strong deviation of NSL can be found in the fact that the curtain's impact on the flow is relevant and can definitely not be neglected.

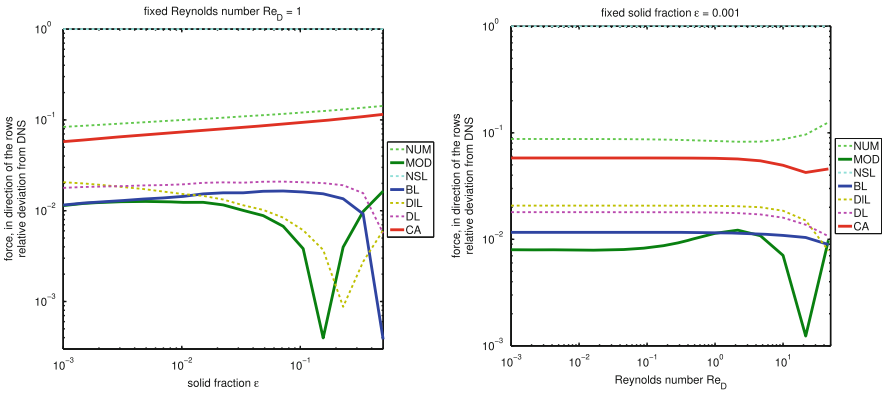


Fig. 4 Comparison of the strategies for a fiber bundle ($M = 25$, $N = 10$). Relative error of the normal force component onto the curtain surface wrt. DNS results, *left:* for varying ϵ and fixed $\mathrm{Re} = 1$, *right:* for varying Re and fixed $\epsilon = 0.001$

4 Discussion and Summary

Aiming the efficient simulation of curtains of long fibers in flow fields, we studied different asymptotic and numerical surrogate models and strategies. The most accurate results are obtained by an immersed boundary method. However, the variant MOD requires the solution of integro-differential equations and is hence very costly. The much cheaper strategy NUM in contrast suffers from the disadvantages of the averaging over the grid cells. Its outcome is grid-dependent. Brinkman's Law and our continuum approach turn out to be satisfying compromises between accuracy and effort. This is not surprising as the strategies are quite similar, dealing with an additional force term in the flow's momentum equation. The force $\boldsymbol{f}_{BL}$ is linear in the relative velocity, which is also true for $\boldsymbol{f}_{CA}$ for small relative velocities. In view of coupled dynamic fiber-fluid simulations the satisfaction of the actio-reactio principle is important. CA fulfills this requirement due to the underlying aerodynamic drag model. For its application to simulating a rotational spinning process in glass wool production we refer to [17]; for further details see [7].

Acknowledgements This work has been supported by German BMBF, 05M2010 and 05M2013.

References

1. Allaire, G.: Continuity of the Darcy's law in the low-volume fraction limit. Ann. Scuola Norm. Sup. Pisa **18**(4), 475–499 (1991)
2. Allaire, G.: Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes. Part I + II. Arch. Rational Mech. Anal. **113**, 209–298 (1991)
3. Andric, J.: Numerical modeling of air-fiber flows. Ph.D. thesis, Chalmers University of Technology, Göteborg (2014)

4. Arne, W., Marheineke, N., Meister, A., Wegener, R.: Numerical analysis of Cosserat rod and string models for viscous jets in rotational spinning processes. Math. Models Methods Appl. Sci. **20**(10), 1941–1965 (2010)
5. Arne, W., Marheineke, N., Schnebele, J., Wegener, R.: Fluid-fiber-interactions in rotational spinning process of glass wool manufacturing. J. Math. Ind. **1**(2), 1–26 (2011)
6. Brinkman, H.: A calculation of the viscous force exerted by a flowing fluid on a dense swarm of particles. Appl. Sci. Res. **A1**, 27–34 (1947)
7. Cibis, T.: Homogenisierungsstrategien für Filament-Strömungs-Wechselwirkungen. Ph.D. thesis, FAU Erlangen-Nürnberg, Erlangen (2015)
8. Cibis, T., Marheineke, N., Wegener, R.: Asymptotic modeling framework for fiber-flow interactions in a two-way coupling. In: Fontes, M., et al. (eds.) Progress in Industrial Mathematics at ECMI 2012, pp. 109–117. Springer, Heidelberg (2014)
9. Cioranescu, D., Murat, F.: A strange term coming from nowhere. In: Cherkaev, A., Kohn, R. (eds.) Topics in the Mathematical Modelling of Composite Materials, pp. 45–93. Birkhäuser, Boston (1997)
10. Drummond, J., Tahir, M.: Laminar viscous flow through regular arrays of parallel solid cylinders. Int. J. Multiphase Flow **10**, 515–540 (1983)
11. Durlofsky, L., Brady, J.: Analysis of the Brinkman equation as a model for flow in porous media. Phys. Fluids **30**(11), 3329–3341 (1987)
12. Hämäläinen, J., Lindström, S.B., Hämäläinen, T., Niskanen, H.: Papermaking fibre-suspension flow simulations at multiple scales. J. Eng. Math. **71**(1), 55–79 (2011)
13. Hornung, U.: Homogenization and Porous Media. Springer, New York (1997)
14. Klar, A., Marheineke, N., Wegener, R.: Hierarchy of mathematical models for production processes of technical textiles. Z. Ang. Math. Mech. **89**(12), 941–961 (2009)
15. Koch, D., Ladd, A.: Moderate Reynolds number flows through periodic and random arrays of aligned cylinders. J. Fluid Mech. **349**, 31–66 (1997)
16. Marheineke, N., Wegener, R.: Modeling and application of a stochastic drag for fibers in turbulent flows. Int. J. Multiphase Flow **37**, 136–148 (2011)
17. Marheineke, N., Liljo, J., Mohring, J., Schnebele, J., Wegener, R.: Multiphysics and multimethods problem of rotational glass fiber melt-spinning. Int. J. Num. Anal. Mod. B **3**(3), 330–344 (2012)
18. Mark, A.: A novel immersed-boundary method for multiple moving and interacting bodies. Ph.D. thesis, Chalmers University of Technology, Göteborg (2007)
19. Nield, D.: The Beavers-Joseph boundary condition and related matters: a historical and critical note. Transp. Porous Media **78**(3), 537–540 (2009)
20. Nield, D., Bejan, A.: Convection in Porous Media. Springer, New York (1992)
21. Peskin, C.: The immersed boundary method. Acta Numer. **11**, 1–39 (2002)
22. Rubin, M.: Cosserat Theories: Shells, Rods and Points. Solid Mechanics and Its Applications. Springer, Dordrecht (2000)
23. Sangani, A., Acrivos, A.: Slow flow past periodic arrays of cylinders with application to heat transfer. Int. J. Multiphase Flow **8**, 193–206 (1982)
24. Svenning, E., Mark, A., Edelvik, F., Glatt, E., Rief, S., Wiegmann, A., Martinsson, L., Lai, R., Fredlund, M., Nyman, U.: Multiphase simulation of fiber suspension flows using immersed boundary methods. Nordic Pulp Paper Res. J. **27**(2), 184–191 (2012)

Homotopy Method for Viscous Cosserat Rod Model Describing Electrospinning

Walter Arne, Javier Rivero-Rodriguez, Miguel Pérez-Saborid, Nicole Marheineke, and Raimund Wegener

Abstract The dynamics of viscous jets in electrospinning processes varies from drop forming, whipping to coiling depending on the parameter regime. To investigate the practically relevant whipping regime more closely we use an asymptotic Cosserat rod model that is given by a stiff boundary value problem of ordinary differential equations. For the efficient simulation of the six-parametric problem we present a numerical approach that is based on a continuation-collocation method. On top of an implicit Runge-Kutta discretization of fourth order, suitable initial guesses and global convergence of the applied Newton method are achieved by a recursive continuation strategy. The numerical results are very convincing, they show the jet characteristics observed in the experiments.

Keywords Cosserat rod model • Electrospinning • Fiber spinning • Homotopy method • Viscous jets • Whipping instability

1 Introduction

Electrospinning processes allow the production of very thin polymer fibers with diameters ranging from less than 3 nm to over 1 μm. By applying an electric charge, a molten polymer extrudes from the spinneret and forms a slender jet due to the high voltage between nozzle and collector, see Fig. 1. Depending on the parameter regime, the observed jet behavior varies from drop forming, whipping instability to coiling [6, 7]. The whipping is of practical interest and can be described by a

W. Arne (✉) • R. Wegener
Fraunhofer ITWM, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
e-mail: walter.arne@itwm.fraunhofer.de; raimund.wegener@itwm.fraunhofer.de

J. Rivero-Rodriguez • M. Pérez-Saborid
Area de Mecanica de Fluidos, Departamento de Ingeneria Aeroespacial y Mecanica de Fluidos, Avenida de los Descubrimientos s/n, 41092 Sevilla, Spain
e-mail: jrivrod@us.es; psaborid@us.es

N. Marheineke
Department Mathematik, FAU Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: marheineke@math.fau.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_137

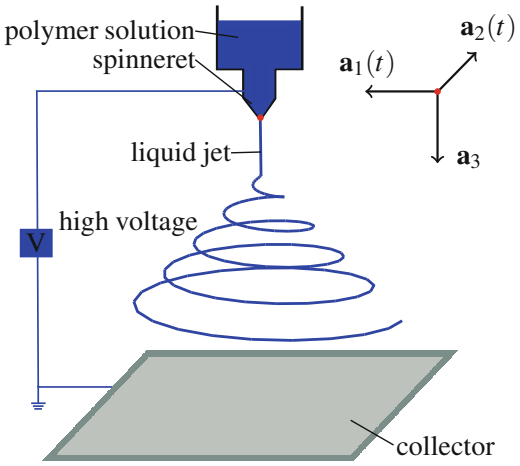


Fig. 1 Sketch of electrospinning set-up with acting electric field $\mathbf{E} = E\mathbf{a_3}$ and gravity $\mathbf{g} = \rho A g \mathbf{a_3}$—consideration of a time-dependent outer basis $\{\mathbf{a_1}(t), \mathbf{a_2}(t), \mathbf{a_3}\}$ that is originated in the spinneret (tip) and rotates wrt. the jet's whipping frequency Ω, i.e. $\boldsymbol{\Omega} = \Omega \mathbf{a_3}$, $\partial_t \mathbf{a_i} = \boldsymbol{\Omega} \times \mathbf{a_i}$

stationary viscous Cosserat rod model.[1] In this paper we focus on the numerical treatment of the six-parametric stiff boundary value problem (BVP) of ordinary differential equations. We use a Lobatto IIIa formula (implicit Runge-Kutta scheme of fourth order) for collocation. The resulting nonlinear system is solved with a Newton method for which global convergence is achieved by a continuation strategy (homotopy method). The initial guess is adapted from an idea of Ribe [4] for viscous rope coiling onto a plane. We conclude with numerical results to a parameter study.

2 Electrospinning Model

In electrospinning a liquid jet leaves the spinneret and moves due to viscous friction, surface tension, gravity and applied electric forces (Fig. 1). In the special Cosserat theory it is described by a curve for the position and a director triad for the cross-sectional orientation. In this work we consider a spun fiber jet of certain length L with stress-free end. To study the whipping instability, we choose a time-dependent outer basis $\{\mathbf{a_1}(t), \mathbf{a_2}(t), \mathbf{a_3}\}$ rotating with the—a priori unknown—jet's whipping frequency Ω, $\boldsymbol{\Omega} = \Omega \mathbf{a_3}$ and introduce the respective spin to the directors, cf. [4, 5] on viscous rope coiling. This makes a 'lay-down' position and the directors time-independent, but introduces fictitious rotational forces and moments due to inertia.

Proceeding from the incompressible viscous Cosserat rod equations [2] derived for rotational spinning, we incorporate electric and capillary force models [7]. The resulting dimensionless BVP is formulated in the director basis $\{\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}\}$ for the

[1] For details on experiments, physical effects and modeling of the electrospinning process we refer to the proceeding article *Setup of viscous Cosserat rod model describing electrospinning* by J. Rivero-Rodriguez et al.

jet curve r, rotational group $\mathsf{R} = (R_{ij}) = (\mathbf{d_i} \cdot \mathbf{a_j}) \in SO(3)$, curvature κ, velocity u, forces n and moments m. The unknown jet length L and whipping frequency Ω are expressed in the dimensionless numbers—length ratio ℓ between jet length and tip–counterelectrode (spinneret–collector) distance and Rossby number Rb as ratio between inertia and rotation—that are determined by two additionally imposed geometric boundary conditions at the jet end. The model is given by

$$
\begin{aligned}
\ell^{-1}\mathsf{R}\cdot\partial_s\check{\mathsf{r}} &= \mathsf{e}_3 & \ell^{-1}\partial_s\mathsf{R} &= -\text{»}\times\mathsf{R}\\
\ell^{-1}\partial_s\text{»} &= -\frac{1}{3}\text{»}n_3 + \frac{4}{3}u\mathsf{P}_{3/2}\cdot\mathsf{m} + \frac{1}{\mathrm{Rb}}\frac{1}{u}\kappa\times\mathsf{e}_3 & \ell^{-1}\partial_s u &= \frac{1}{3}un_3\\
\ell^{-1}\partial_s\mathsf{n} &= -\text{»}\times\mathsf{n} + \mathrm{Re}\,u\left(\text{»}\times\mathsf{e}_3 + \frac{1}{3}n_3\mathsf{e}_3\right) + \frac{2\mathrm{Re}}{\mathrm{Rb}}(\mathsf{R}\cdot\mathsf{e}_3)\times\mathsf{e}_3\\
&\quad + \frac{\mathrm{Re}}{\mathrm{Rb}^2}\frac{1}{u}\mathsf{R}\cdot\left(\mathsf{e}_3\times(\mathsf{e}_3\times\check{\mathsf{r}})\right) - \frac{\mathrm{Re}}{\mathrm{Fr}^2}\frac{1}{u}\mathsf{R}\cdot\mathsf{e}_3 - \mathsf{f}_{ca} - \mathsf{f}_{el}\\
\ell^{-1}\partial_s\mathsf{m} &= -\text{»}\times\mathsf{m} + \frac{4}{\epsilon_\star^2}\mathsf{n}\times\mathsf{e}_3 + \frac{\mathrm{Re}}{3}\left(u\mathsf{P}_3\cdot\mathsf{m} - \frac{1}{4}n_3\mathsf{P}_2\cdot\text{»}\right)\\
&\quad - \frac{\mathrm{Re}}{4\mathrm{Rb}}\frac{1}{u}\mathsf{P}_2\cdot\left(\frac{1}{3}\mathsf{R}\cdot\mathsf{e}_3 n_3 - \frac{1}{3}\mathsf{e}_3 n_3 + \left(\text{»} - \frac{1}{\mathrm{Rb}}\frac{1}{u}\mathsf{e}_3\right)\times\mathsf{R}\cdot\mathsf{e}_3\right)\\
&\quad - \frac{\mathrm{Re}}{4}\left(\frac{1}{u^2}\mathsf{P}_2\cdot(u\text{»} - \frac{1}{\mathrm{Rb}}\mathsf{e}_3 + \frac{1}{\mathrm{Rb}}\mathsf{R}\cdot\mathsf{e}_3)\right)\times\left(u\text{»} - \frac{1}{\mathrm{Rb}}\mathsf{e}_3 + \frac{1}{\mathrm{Rb}}\mathsf{R}\cdot\mathsf{e}_3\right)
\end{aligned}
$$

with capillary and electric forces

$$
\mathsf{f}_{ca} = \frac{1}{\epsilon\mathrm{Ca}}\frac{1}{\sqrt{u}}\left(2\kappa\times\mathsf{e}_3 - \frac{1}{3}n_3\mathsf{e}_3\right), \quad \mathsf{f}_{el} = \frac{4\xi}{\epsilon}\frac{1}{u}\mathsf{R}\cdot\mathsf{e}_3 - \theta\frac{1}{u^2}\log\left(\frac{2}{\epsilon}\sqrt{u}\right)\kappa\times\mathsf{e}_3
$$

and geometric and kinematic boundary conditions at the nozzle $s = 0$ as well as geometric and stress-free dynamic boundary conditions at $s = 1$

$$
\begin{aligned}
\check{\mathsf{r}}(0) &= \mathsf{0} & \mathsf{R}(0) &= \mathsf{P}_1 & \kappa(0) &= 0 & u(0) &= 1\\
\check{r}_1(1) &= 0 & \check{r}_3(1) &= 1 & \mathsf{n}(1) &= \mathsf{0} & \mathsf{m}(1) &= \mathsf{0},
\end{aligned}
$$

$\mathsf{P}_k = \mathrm{diag}(1, 1, k)$, $k \in \mathbb{R}$ and e_i, $i = 1, 2, 3$ canonical basis vectors in $\mathbb{R}^3$. Note that $\check{\mathsf{r}} = (\check{r}_i)$ represents the curve in the outer basis, i.e. $\mathsf{r} = \mathsf{R}\cdot\check{\mathsf{r}}$. For the parametrization of the rotational group R we use unit quaternions [3].

The electrospinning model and hence the jet's solution are characterized by six dimensionless parameters $(\mathrm{Re}, \mathrm{Ca}, \mathrm{Fr}, \xi, \theta, \epsilon) \in \mathbb{R}_0^6$ with Reynolds number Re as ratio between inertia and viscosity, Capillary number Ca as ratio between viscosity and surface tension, Froude number Fr as ratio between inertia and gravity, ξ as ratio between electric field and viscosity, θ as ratio between Coulomb repulsion and viscosity, and ϵ as slenderness ratio between jet diameter and tip–counterelectrode

distance. Note that ϵ plays a special role in the equations since it has two meanings. Whereas it represents the actual (physical) slenderness ratio in the electric f_{el} and capillary forces f_{ca}, it can be considered as a regularization parameter $\epsilon = \epsilon_\star$ in the momentum equation. Interpreting the rod as ϵ-regularized string model in the context of asymptotic analysis [2], we can thus modify the regularization parameter $\epsilon_\star$ to take a moderate value and stabilize the numerics (e.g. $\epsilon_\star = 0.1$).

3 Numerical Approach

The numerical challenge lies in solving the BVP for arbitrary parameter settings. Following [1] we apply an implicit Runge-Kutta scheme (Lobatto IIa formula) of fourth order as collocation method. The resulting non-linear system of equations is solved using a Newton method. The core of the numerics is the applied continuation procedure. As the convergence of the Newton method depends crucially on the initial guess, we adapt the initial guess iteratively by solving a sequence of BVPs with slightly changed parameters, improving so the computational performance and globalizing the convergence. The typical questions in such a continuation (homotopy) method deal with the continuation step size control, an appropriate starting solution and the choice of a continuation path through the parameter space. Thereby, the first one is obviously general and technical, whereas the last two are model-dependent.

Continuation Step Size Control Given a starting solution to the parameter tuple p_s, we seek for a sequence of parameters $\mathsf{p}_s = \mathsf{p}_0, \mathsf{p}_1, \ldots, \mathsf{p}_n = \mathsf{p}_d$ with the desired parameter tuple p_d such that the solution to the respective predecessor BVP provides a good initial guess for the successor. The choice of the continuation path decides about failure or success since there are not always existing solutions and several meaningful ways. We use an adaptive step size control with the number of Newton iterations as quality criterion for the chosen p_i. An interim solution is always computed twice by using one full step and two half steps. If the full step requires more Newton iterations or 10 % more collocation points than both half steps together, the continuation step is reduced by a factor k_1, otherwise it is increased by k_1 for the further computation. If the Newton method fails, the step size is reduced by a factor k_2 and the computation is repeated. As first step we try $\mathsf{p}_1 = \mathsf{p}_s + k(\mathsf{p}_d - \mathsf{p}_s)$ with $k = 10^{-2}$, moreover $k_1 = 1.5, k_2 = 10$.

Starting Solution The initialization is taken from [4], it is the analytical solution

$$\check{\mathsf{r}}(s) = (1 - \cos(\pi s/2), 0, \sin(\pi s/2)), \quad \mathsf{q}(s) = (\cos(\pi s/4), 0, -\sin(\pi s/4), 0)$$
$$\kappa = (0, 1, 0), \quad u = 1, \quad \mathsf{n} = \mathsf{m} = 0$$

with $\ell = \pi/2$ and $\mathrm{Rb}^{-1} = 0$ for a (non-coiling) jet having the form of a quarter circle in the absence of inertia, surface tension and outer forces

$(\mathrm{Re}, \mathrm{Ca}^{-1}, \mathrm{Fr}^{-1}, \xi, \theta) = 0$ and arbitrary $\epsilon \neq 0$. Using continuation on the boundary conditions we obtain a solution associated to viscous rope coiling with vanishing linear and angular velocity at the jet end, i.e. continuation parameters c_i are turned from 0 to 1 in

$$\begin{aligned}
\kappa(0) - (1-c_1)\mathsf{e}_2 &= \mathsf{0} \\
c_2 \check{r}_1(1) - (1-c_2)\check{r}_2(1) &= 0 \\
(-1 - c_3 + c_3/\mathrm{Rb})\mathsf{R}(1) \cdot (\mathsf{e}_3 \times \check{\mathsf{r}}(1)) + u(1)\mathsf{e}_3 - (1-c_4)(\mathsf{e}_2 + \mathsf{e}_3) &= 0 \\
(-1 - c_5 + c_5/\mathrm{Rb})(\mathsf{R}(1) - \mathsf{P}_1) \cdot \mathsf{e}_3 + u(1)\text{»}(1) + (1-c_6)(\mathsf{e}_1 - \mathsf{e}_2 + \mathsf{e}_3) &= 0
\end{aligned}$$

Continuation Path in Parameter Space In electrospinning the whipping is caused by the applied electric forces according to observations. The following continuation strategy turns out to be successful, as it regards the change of the whipping frequency by taking into account the interplay of Re, ξ and θ.

1. Proceed from the solution to viscous rope coiling for desired ϵ and increase Re, ξ, θ up to moderate values [e.g. $(\mathrm{Re}, \xi, \theta) = (1, 1, 2.3)$].
2. Perform continuation on the boundary conditions to obtain a stress-free jet end, i.e. parameter c is turned from 0 to 1 in $\mathsf{n}(1) = (1-c)\tilde{\mathsf{n}}$ and $\mathsf{m}(1) = (1-c)\tilde{\mathsf{m}}$.
3. Change continuously to the desired characteristic parameters of the BVP, keep thereby ξ and θ in balance.

Note that the continuation method is robust in Ca, Fr, ϵ as the respective solutions only slightly change. The forthcoming numerical simulations are performed on a Intel Xeon 2.67 GHz using MATLAB, in particular the routine bvp4c.m is used for the collocation. In the algorithm the preparatory step to obtain the solution associated with viscous rope coiling takes a CPU time of 156 s. Each step (including rejection) requires about 3–5 s. This is a spectacularly good performance in view of the fact that the path through the six-parameteric c_i-space is by no means the diagonal. The actual navigation in the space of the electrospinning parameters is sensitive and takes in general several minutes. With a step size varying over several orders of magnitude and only same hundred steps in total, it clearly stresses the efficiency of the adaptive step size control.

4 Results and Discussion

The numerical results of the electrospinning model are very promising, they show qualitatively the characteristic whipping behavior observed in the experiments [6]. The whipping radius increases for a stronger repulsion (larger θ) or a smaller electric field (smaller ξ), see jet dynamics in Fig. 2. Also the cross-sectional stretching depends strongly on the electrostatic effects. The impact of gravity and surface

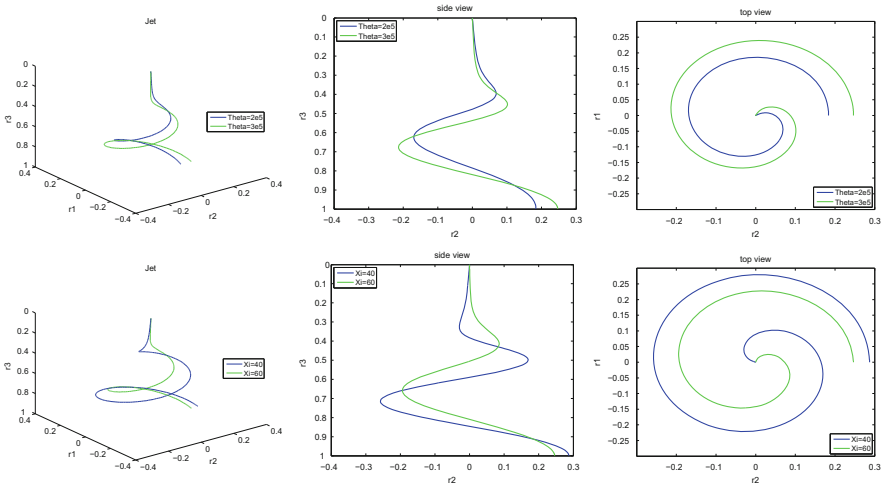


Fig. 2 Jet curve in 3d with 2d projections *side view* and *top view* (from *left to right*) for $(\mathrm{Fr}^{-1}, \mathrm{Ca}, \epsilon, \epsilon_\star) = (0, 1, 8 \cdot 10^{-3}, 10^{-1})$. *Top:* $(\mathrm{Re}, \xi) = (0.17, 40)$ and varying θ (repulsion $\theta = 2 \cdot 10^5, 3 \cdot 10^5$), *bottom:* $(\mathrm{Re}, \theta) = (0.25, 3 \cdot 10^5)$ and varying ξ (electric field $\xi = 40, 60$)

tension on the jet's solution is comparatively small. To improve the quantitative agreement with measurements, model extensions are in work.

Acknowledgement The support of the Ministry of Science and Innovation of Spain (Project DPI 2010-20450-C03-02) is acknowledged.

References

1. Arne, W., Marheineke, N., Meister, A., Wegener, R.: Numerical analysis of Cosserat rod and string models for viscous jets in rotational spinning processes. Math. Models Methods Appl. Sci. **20**(10), 1941–1965 (2010)
2. Arne, W., Marheineke, N., Wegener, R.: Asymptotic transition from Cosserat rod to string models for curved viscous inertial jets. Math. Models Methods Appl. Sci. **21**(10), 1987–2018 (2011)
3. Mahadevan, L., Keller, J.: Coiling of flexible ropes. Proc. R. Soc. Lond. A **452**, 1679–1694 (1996)
4. Ribe, N.: Coiling of viscous jets. Proc. R. Soc. Lond. A **2051**, 3223–3239 (2004)
5. Ribe, N., Habibi, M., Bonn, D.: Stability of liquid rope coiling. Phys. Fluids **18**, 084102 (2006)
6. Riboux, G., Marin, A., Loscertales, I., Barrero, A.: Whipping instability characterization of an electrified visco-capillary jet. J. Fluid Mech. **671**, 226–253 (2011)
7. Yarin, A., Koombhongse, S., Reneker, D.: Bending instability in electrospinning of nanofibers. J. Appl. Phys. **89**(5), 3018–3026 (2001)

Setup of Viscous Cosserat Rod Model Describing Electrospinning

Javier Rivero-Rodríguez, Walter Arne, Nicole Marheineke, Raimund Wegener, and Miguel Pérez-Saborid

Abstract Electrospinning is commonly used to produce very fine polymeric fibers. In this technique, a conducting liquid is pumped from an electrified needle into a surrounding dielectric media and the meniscus formed exhibits a conical shape, known as Taylor cone, due to the balance of electrical and surface tension forces. If the needle electrical potential is sufficiently high, the very strong electric field generated at the cone apex cannot be balanced by surface tension and a very thin jet is issued which eventually develops lateral instabilities that are responsible of additional stretching. In this work, we use a theoretical model that describes the kinematic of the midline of the jet, its radius and convective velocity from an Eulerian framework. Balances of mass, linear and angular momentum applied to a slice of the jet, as well as viscous law for stretching, bending and torsion describe the dynamics (nonlinear PDE in time and arclength of the midline). Capillary and electric forces are included in the momentum balance. If periodic orbits are explored, the time dependence of the PDE disappears when the motion is considered with respect to a frame rotating with the jet. One obtains a boundary value problem of ODEs with the frequency as a free parameter. This model is also suitable for describing other kinds of instabilities, such as the axisymmetric one which takes place in drop formation (dripping regime, electrospray).

Keywords Cosserat rod model • Electrospinning • Fiber spinning • Viscous jets • Whipping instability

J. Rivero-Rodríguez (✉) • M. Pérez-Saborid
Area de Mecanica de Fluidos, Departamento de Ingeneria Aeroespacial y Mecanica de Fluidos, Avenida de los Descubrimientos s/n, 41092 Sevilla, Spain
e-mail: jrivrod@us.es; psaborid@us.es

W. Arne • R. Wegener
Fraunhofer ITWM, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
e-mail: walter.arne@itwm.fraunhofer.de; raimund.wegener@itwm.fraunhofer.de

N. Marheineke
Department Mathematik, FAU Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: marheineke@math.fau.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_138

1 Introduction

The interaction of an intense electrical field with the interface between a conducting liquid and a dielectric medium has been known to exist since Gilbert [4] reported in 1600 the formation of a conical meniscus when an electrified piece of ambar was brought close enough to a water drop. The deformation of the interface is caused by the force that the electric field exerts on the net surface charge induced by the field itself. This phenomenon is at the base of modern devices for the production of micro and nano-structures of interest in several technological fields. As schematized in Fig. 1a, these devices consist essentially of a high-voltage power supply, a metallic needle (spinneret) and a grounded collector (counterelectrode). The metallic needle is connected to a syringe pump through which a conducting liquid can be fed at a constant and controllable rate. When a high voltage (usually in the range of 1–30 kV) is applied, the electric field induces an electric current in the liquid that accumulates electric charge at the surface and causes an electric stress that elongates the pendent drop at the needle's exit in the direction of the field. It is observed that if the field strength is below a certain threshold value the balance of electrostatic and surface tension stresses gives rise to a motionless conical shape commonly known as the Taylor cone [12]. However, above the threshold, the large electrostatic stresses concentrated near the cone tip overcome the surface tension stresses and force the ejection of an electrified liquid jet from the cone tip.

For certain values of the applied voltage and imposed liquid flow rate, the jet emanating from the cone tip is stationary and breaks into spherical droplets at some distance downstream due to axisymmetric Rayleigh-Plateau (varicose) instabilities

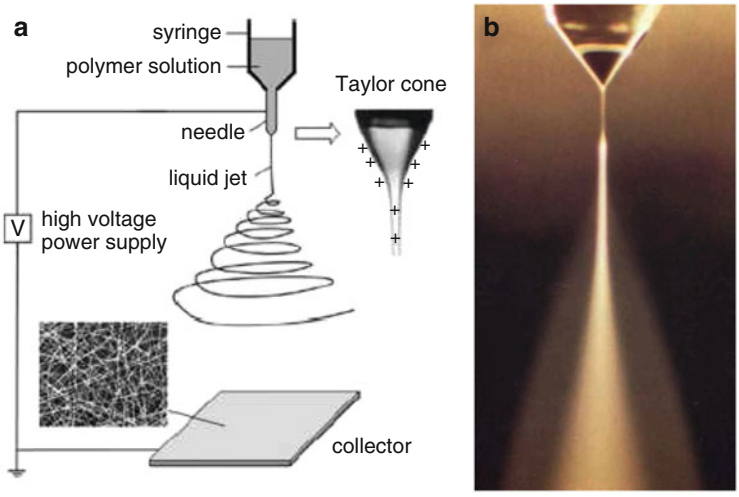


Fig. 1 (**a**) Sketch of an electrospinning or electrospray device (taken from [7]). (**b**) Cone-jet mode of an electrospray (taken from [8])

corrected to account for the presence of surface charge. This so-called cone-jet mode (see Fig. 1b) forms the basis of the electrospray technique [3] for generating small monodisperse drops with great applications in fine coatings, synthesis of powders, micro and nanocapsules, etc.

However, nonsymmetric perturbation modes can also grow due to the net charge carried by the jet. Indeed, if a small portion of the charged jet moves slightly off axis, the charge distributed along the rest of the jet will push that portion farther away from the axis according to Earnshaw's theorem, thus leading to a lateral instability known as whipping or bending instability. If the growth rate associated to this whipping instability is larger than that associated to varicose jet break-up—as may happen, for example, for sufficiently high values of the applied voltage or of the liquid viscosity—, the off-axis movement of the jet becomes the most significant aspect of its evolution (see Fig. 2). The whipping mode manifests itself in the form of chaotic, fast and violent slashes which give rise to very large tensile stresses and to a dramatic jet thinning. This is of fundamental importance in the process known as electrospinning [7, 9], where micro or nanofibers of a polymeric fluid are produced by solidification of the jet issuing from the Taylor cone before it breaks up into droplets. The evaporation of the solvent is greatly enhanced by a reduction of the jet diameter that is typically several orders of magnitude, which makes the electrospinning technique very competitive with other existing ones such phase separation or self-assembly. It must be realized that the chaotic nature of the whipping regime makes very difficult to unravel its detailed structure. However, there are some circumstances which greatly enhance the parametric range for which

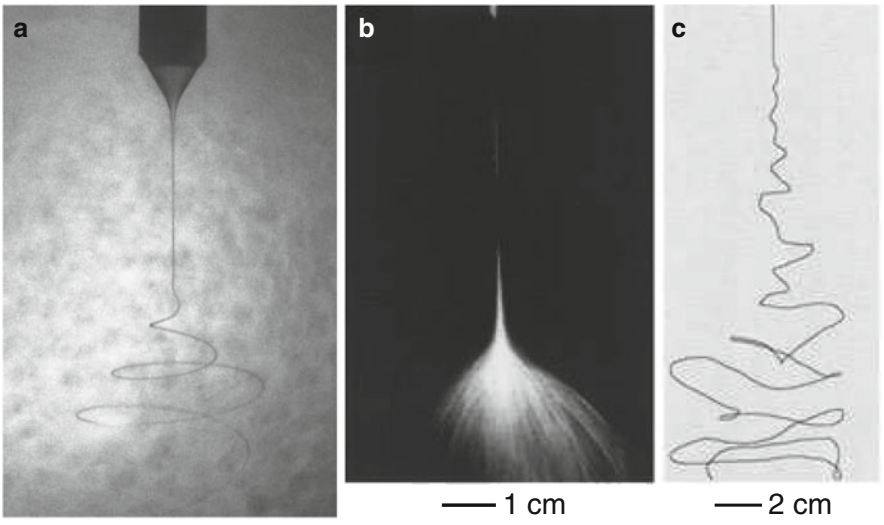


Fig. 2 Whipping instability (**a**) in an electrified jet of glycerine in a bath of hexane (courtesy of Dr. A. Gomez-Marin). (**b**–**c**) Photograph illustrating the chaotic behavior of the whipping mode: (**b**) capture time 1/250 s and (**c**) capture time 18 ns (taken from [7])

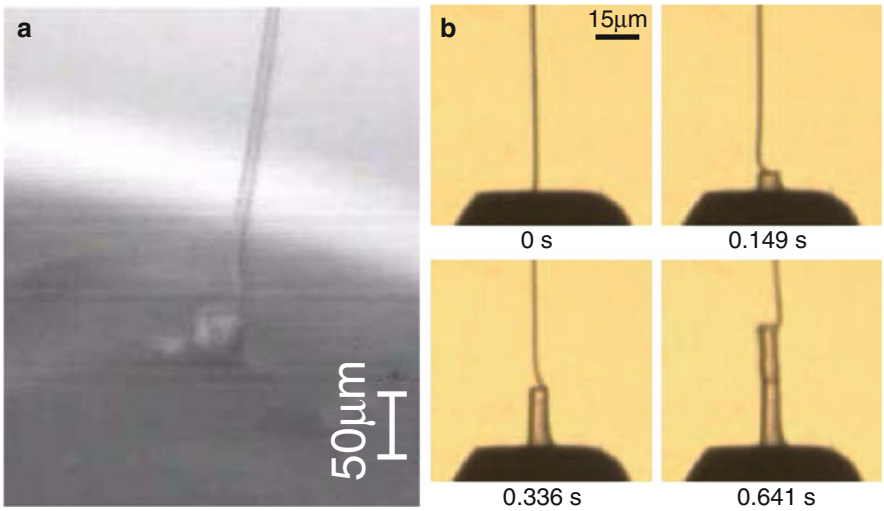


Fig. 3 (**a**) Coiling of an electrified liquid jet (Courtesy of Dr. G. Riboux). (**b**) High speed sequential images of nanocoiling process that yields a free-standing hollow cylinder (taken from [6])

the bending stability leads to a stable helicoidal structure as, for example, when the conducting liquid is surrounded by a dielectric bath [11] or by another coflowing liquid [5].

Yet another mode of operation has been observed for the device sketched in Fig. 1, the so-called coiling mode (see Fig. 3). This mode occurs when the ground electrode is located sufficiently close to the needle's exit so that the liquid jet reaches the plate before being set into chaotic motion by the bending instability. The situation is then the same as if a thin stream viscous fluid such as honey is poured onto a surface from a certain height [10]. Rather than approaching the surface vertically the jet builds on it a helical structure which resembles a pile of coiled rope. The origin of liquid rope coiling is a buckling instability in which an initially vertical fluid stream subject to an axial compressive stress becomes unstable to deformation by bending. Sufficiently far from the counterelectrode, the electrified jet has an helicoidal structure which, under some circumstances, strongly resembles that of the whipping regime in the more stable cases referred to above [5, 11]. The coiling mode has also been observed [6] when polymer nanofibers are electrospun using a grounded pin collector, where the strongly focused electrical field at the ground causes a stable jet which evaporates after exiting the spinneret and gives rise to a dry fiber which impinges on the pin's tip buckling and coiling as shown in Fig. 3b.

As a first step towards the understanding of the physical processes involved in electrospinning, we deal with the modeling and simulation of the whipping regime of an electrified liquid jet. We are particularly interested in the influence of viscosity,

surface tension, gravity, imposed electric field and self-repulsion of the induced charges on the jet dynamics. In the framework of viscous Cosserat rods, we set-up an one-dimensional model that allows for the description of the whipping as stationary process in a rotating frame.

2 Viscous Cosserat Rod for Electrospinning

A jet is a slender long body whose dynamics can be reduced to an one-dimensional description by averaging the underlying balance laws over its cross-sections. In the special Cosserat rod theory there are two constitutive elements: a curve $\mathbf{r} : Q \to \mathbb{E}^3$ specifying the jet position (e.g. midline) and an orthonormal director triad $\{\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}\} : Q \to \mathbb{E}^3$ characterizing the orientation of the cross-sections in the three-dimensional Euclidian space $\mathbb{E}^3$, see Fig. 4a. In $Q = \{(s,t) \in (\mathbb{R}_0^+)^2\}$, t denotes the time and s the arc-length parameter imposing here an Eulerian description. In the following we proceed from the incompressible viscous Cosserat rod model of [1, 2] that was derived for fiber jets in rotational spinning processes on the basis of the work [10] on viscous rope coiling. It allows for stretching, bending and torsion. The incorporated electric and capillary forces are modeled according to [13].

The rod system consists of four kinematic and three dynamic equations, i.e. balance laws for mass (cross-section A), linear and angular momentum,

$$\begin{aligned} \partial_s \mathbf{r} = \mathbf{d_3}, \quad \partial_s \mathbf{d_i} = \boldsymbol{\kappa} \times \mathbf{d_i}, \quad \partial_t \mathbf{r} = \mathbf{v} - u\mathbf{d_3}, \quad \partial_t \mathbf{d_i} = (\boldsymbol{\omega} - u\boldsymbol{\kappa}) \times \mathbf{d_i} \\ \partial_t A + \partial_s(uA) = 0, \quad \rho\partial_t(A\mathbf{v}) + \rho\partial_s(uA\mathbf{v}) = \partial_s \mathbf{n} + \mathbf{f}_{gr} + \mathbf{f}_{ca} + \mathbf{f}_{el} \\ \rho\partial_t(\mathbf{J}\cdot\boldsymbol{\omega}) + \rho\partial_s(u\mathbf{J}\cdot\boldsymbol{\omega}) = \partial_s \mathbf{m} + \mathbf{d_3} \times \mathbf{n} \end{aligned}$$

supplemented with geometric model, viscous material laws and external forces

$$\mathbf{J} = J\mathbf{P}_2, \quad J = \frac{\pi}{4}a^4, \quad A = \pi a^2, \quad \mathbf{n}\cdot\mathbf{d_3} = 3\mu A\partial_s u, \quad \mathbf{m} = 3\mu J\mathbf{P}_{2/3}\cdot\partial_t\boldsymbol{\kappa}$$

$$\mathbf{f}_{gr} = \rho A\mathbf{g}, \quad \mathbf{f}_{ca} = \pi\gamma\partial_s(a\mathbf{d_3}), \quad \mathbf{f}_{el} = 2\pi a\sigma\left(\mathbf{E} - \frac{a\sigma}{2\varepsilon_{per}}\log\left(\frac{H}{a}\right)\boldsymbol{\kappa}\times\mathbf{d_3}\right)$$

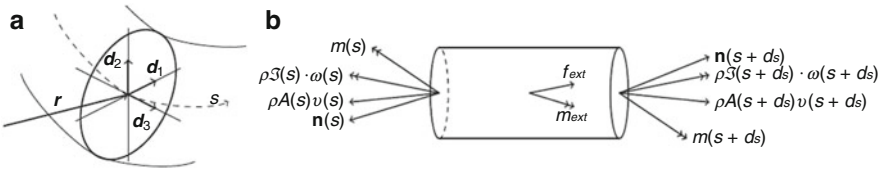


Fig. 4 (**a**) Sketch of jet with midline $\mathbf{r}$ and triad $\mathbf{d_i}$. (**b**) Balance of linear and angular momentum

and $\mathbf{P}_k = \mathbf{d_1} \otimes \mathbf{d_1} + \mathbf{d_2} \otimes \mathbf{d_2} + k\,\mathbf{d_3} \otimes \mathbf{d_3}$, $k \in \mathbb{R}$. The convective speed u can be viewed as one of the Lagrange multipliers to the constraint $\partial_s \mathbf{r} = \mathbf{d_3}$ for the jet tangent (i.e. generalized Kirchhoff constraint with stretching, no shear and arc-length parametrization). The further kinematic equations relate the jet's midline and triad to the curvature $\boldsymbol{\kappa}$, the linear $\mathbf{v}$ and angular $\boldsymbol{\omega}$ velocities. The geometric model for the angular momentum line density with moment of inertia $\mathbf{J}$ preserves the jet's incompressibility. The mass density ρ is considered to be constant and the cross-section circular-shaped of radius a. The tangential contact force $\mathbf{n} \cdot \mathbf{d_3}$ and the couple $\mathbf{m}$ are specified by a linear material law in the strain rate variables with dynamic viscosity μ, whereas the normal force components are the other Lagrange multipliers to the constraint. The external forces are due to gravity $\mathbf{g}$, surface tension γ and the electric field $\mathbf{E}$. The self-repulsion of the induced charges are modeled in terms of the surface charge density σ, the permittivity ε_p and the tip–counterelectrode distance H by help of a local interaction approximation [13]. It is assumed that σ is purely convected with the fluid, i.e. $I = 2\pi a \sigma u$ holds for the electric current.

In the electrospinning process, we have a fixed predominant direction $\mathbf{a_3} = \mathbf{d_3}(0,t)$ (jet tangent at the spinneret $s = 0$) due to the acting electric field $\mathbf{E} = E\mathbf{a_3}$ and gravity $\mathbf{g} = g\mathbf{a_3}$ (Fig. 1a). To investigate the whipping behavior in a stationary manner, we consider a spun jet of certain – a priori unknown – length L with stress-free end. Furthermore, we introduce a time-dependent outer basis $\{\mathbf{a_1}(t), \mathbf{a_2}(t), \mathbf{a_3}\}$, $\partial_t \mathbf{a_i} = \boldsymbol{\Omega} \times \mathbf{a_i}$ rotating with the jet's – a priori unknown – whipping frequency Ω, $\boldsymbol{\Omega} = \Omega \mathbf{a_3}$ and a modified director triad with an additional spin $\partial_t \mathbf{d_i^s} = (\boldsymbol{\omega} - u\boldsymbol{\kappa} + \Omega \mathbf{d_3^s}) \times \mathbf{d_i^s}$, $i = 1,2,3$ [10]. A suitable representation in these bases yields a stationary set-up, but introduces fictitious body forces and couples, such as Coriolis, centrifugal and spin-associated ones, due to inertia in the model equations. The director and outer bases are related by the tensor-valued rotation $\mathbf{R}$, i.e. $\mathbf{R} = \mathbf{a_i} \otimes \mathbf{d_i^s}$. For any quantity we use the following coordinate terminology: $\mathbf{y} = \sum_{i=1}^{3} y_i \mathbf{d_i^s} = \sum_{i=1}^{3} \check{y}_i \mathbf{a_i} \in \mathbb{E}^3$ with $\mathsf{y} = (y_1, y_2, y_3) \in \mathbb{R}^3$ and $\check{\mathsf{y}} = (\check{y}_1, \check{y}_2, \check{y}_3) \in \mathbb{R}^3$ where $\mathsf{y} = \mathsf{R} \cdot \check{\mathsf{y}}$ and $\mathsf{R} = (R_{ij}) = (\mathbf{d_i^s} \cdot \mathbf{a_j}) \in SO(3)$.

Remark 2.1 The periodic rotation of the system around the symmetry axis given by $\mathbf{a_3}$ allows alternatively also the following elegant approach to obtain stationarity. We can express any scalar y and vector-valued $\mathbf{y}$ variables as $y(s,t) = y(s,0)$ and $\mathbf{y}(s,t) = \mathbf{M}(t) \cdot \mathbf{y}^\circ(s)$, where $\mathbf{M}$ represents the rotation tensor with respect to the jet's whipping frequency, $\mathbf{M}(t) = \cos(\Omega t)\mathbf{P}_0(0,0) + \sin(\Omega t)\mathbf{a_3} \times \mathbf{P}_0(0,0) + \mathbf{a_3} \otimes \mathbf{a_3}$. Hence, we get $\partial_t y = 0$ and $\partial_t \mathbf{y}(s,0) = \Omega \mathbf{a_3} \times \mathbf{y}^\circ(s)$ and, in particular, $\mathbf{v}^\circ - u^\circ \mathbf{d_3^\circ} = \Omega \mathbf{a_3} \times \mathbf{r}^\circ$ and $\boldsymbol{\omega}^\circ - u^\circ \boldsymbol{\kappa}^\circ = \Omega(\mathbf{a_3} - \mathbf{d_3^\circ})$ for the linear and angular velocities. In this consideration the relevant two frames are the director triad $\mathbf{d_i^\circ}(s)$ and the reference triad at the nozzle $\mathbf{d_i^\circ}(0)$ with $\mathbf{d_3^\circ}(0) = \mathbf{a_3}$.

In the stationary set-up the mass flux becomes constant, i.e. $Au = const.$ Moreover, the linear and angular velocities can be expressed in terms of the other variables. Incorporating the viscous material laws leads to a boundary value problem of ordinary differential equations for jet curve, triad (rotational group), curvature,

convective speed, contact forces and couples (analogously to the derivation in [2]). To the geometric and kinematic boundary conditions at the nozzle $s = 0$ and stress-free conditions at the end, we impose two additional geometric conditions on the curve end point (height and phase) to determine the free unknown parameters of jet length L and whipping frequency Ω in the problem. In the director basis the dimensionless model equations are then given by

$$\ell^{-1}\mathsf{R}\cdot\partial_s\check{\mathsf{r}} = \mathsf{e}_3, \qquad \ell^{-1}\partial_s\mathsf{R} = -\kappa\times\mathsf{R}$$

$$\ell^{-1}\partial_s\kappa = -\frac{1}{3}\kappa n_3 + \frac{4}{3}u\mathsf{P}_{3/2}\cdot\mathsf{m} + \frac{1}{\mathrm{Rb}}\frac{1}{u}\kappa\times\mathsf{e}_3, \qquad \ell^{-1}\partial_s u = \frac{1}{3}un_3$$

$$\ell^{-1}\partial_s\mathsf{n} = -\kappa\times\mathsf{n} + \mathrm{Re}\,u\left(\kappa\times\mathsf{e}_3 + \frac{1}{3}n_3\mathsf{e}_3\right) + \frac{2\mathrm{Re}}{\mathrm{Rb}}(\mathsf{R}\cdot\mathsf{e}_3)\times\mathsf{e}_3$$
$$+ \frac{\mathrm{Re}}{\mathrm{Rb}^2}\frac{1}{u}\mathsf{R}\cdot\left(\mathsf{e}_3\times(\mathsf{e}_3\times\check{\mathsf{r}})\right) - \frac{\mathrm{Re}}{\mathrm{Fr}^2}\frac{1}{u}\mathsf{R}\cdot\mathsf{e}_3 - \mathsf{f}_{ca} - \mathsf{f}_{el}$$

$$\ell^{-1}\partial_s\mathsf{m} = -\kappa\times\mathsf{m} + \frac{4}{\epsilon^2}\mathsf{n}\times\mathsf{e}_3 + \frac{\mathrm{Re}}{3}\left(u\mathsf{P}_3\cdot\mathsf{m} - \frac{1}{4}n_3\mathsf{P}_2\cdot\kappa\right)$$
$$- \frac{\mathrm{Re}}{4\mathrm{Rb}}\frac{1}{u}\mathsf{P}_2\cdot\left(\frac{1}{3}\mathsf{R}\cdot\mathsf{e}_3 n_3 - \frac{1}{3}\mathsf{e}_3 n_3 + \left(\kappa - \frac{1}{\mathrm{Rb}}\frac{1}{u}\mathsf{e}_3\right)\times\mathsf{R}\cdot\mathsf{e}_3\right)$$
$$- \frac{\mathrm{Re}}{4}\left(\frac{1}{u^2}\mathsf{P}_2\cdot(u\kappa - \frac{1}{\mathrm{Rb}}\mathsf{e}_3 + \frac{1}{\mathrm{Rb}}\mathsf{R}\cdot\mathsf{e}_3)\right)\times\left(u\kappa - \frac{1}{\mathrm{Rb}}\mathsf{e}_3 + \frac{1}{\mathrm{Rb}}\mathsf{R}\cdot\mathsf{e}_3\right)$$

with capillary and electric forces

$$\mathsf{f}_{ca} = \frac{1}{\epsilon\mathrm{Ca}}\frac{1}{\sqrt{u}}\left(2\kappa\times\mathsf{e}_3 - \frac{1}{3}n_3\mathsf{e}_3\right), \quad \mathsf{f}_{el} = \frac{4\xi}{\epsilon}\frac{1}{u}\mathsf{R}\cdot\mathsf{e}_3 - \theta\frac{1}{u^2}\log\left(\frac{2}{\epsilon}\sqrt{u}\right)\kappa\times\mathsf{e}_3$$

and boundary conditions

$$\check{\mathsf{r}}(0) = \mathsf{0}, \qquad \mathsf{R}(0) = \mathsf{P}_1, \qquad \kappa(0) = \mathsf{0}, \qquad u(0) = 1$$
$$\check{r}_1(1) = 0, \qquad \check{r}_3(1) = 1, \qquad \mathsf{n}(1) = \mathsf{0}, \qquad \mathsf{m}(1) = \mathsf{0}$$

with $\mathsf{P}_k = \mathrm{diag}(1,1,k)$, $k\in\mathbb{R}$ and canonical basis $\mathsf{e}_\mathsf{i}\in\mathbb{R}^3$. The system is made dimensionless using the three problem-relevant lengths (jet length L, tip–counterelectrode distance H, nozzle diameter D) and the jet velocity at the nozzle U. The reference values are $\bar{s} = L$, $\bar{r} = H$, $\bar{\kappa} = 1/H$, $\bar{u} = U$, $\bar{n} = \pi\mu UD^2/(4H)$ and $\bar{m} = \pi\mu UD^4/(16H^2)$. Apart from the length ratio $\ell = L/H$ and the Rossby number $\mathrm{Rb} = U/(\Omega H)$ that are induced by the free unknown parameters, the electrospinning model[1] is characterized by six dimensionless numbers: Reynolds

[1]The numerical treatment of the problem is discussed in the proceeding article *Homotopy method for viscous Cosserat rod model describing electrospinning* by Arne et al., the simulations show the experimentally observed whipping behavior.

number Re, Froude number Fr, Capillary number Ca, ratio between electric field and viscosity ξ, ratio between Coulomb repulsion and viscosity θ and slenderness ratio ϵ, i.e.

$$\mathrm{Re} = \frac{\rho U H}{\mu}, \quad \mathrm{Fr} = \frac{U}{\sqrt{gH}}, \quad \mathrm{Ca} = \frac{\mu U}{\gamma}, \quad \xi = \frac{IEH}{\pi \mu D U^2}, \quad \theta = \frac{I^2 H}{\pi^2 \varepsilon_p \mu D^2 U^3}, \quad \epsilon = \frac{D}{H}.$$

Acknowledgement The support of the Ministry of Science and Innovation of Spain (Project DPI 2010-20450-C03-02) is acknowledged.

References

1. Arne, W., Marheineke, N., Meister, A., Wegener, R.: Numerical analysis of Cosserat rod and string models for viscous jets in rotational spinning processes. Math. Models Methods Appl. Sci. **20**(10), 1941–1965 (2010)
2. Arne, W., Marheineke, N., Wegener, R.: Asymptotic transition from Cosserat rod to string models for curved viscous inertial jets. Math. Models Methods Appl. Sci. **21**(10), 1987–2018 (2011)
3. Fernandez de la Mora, J., Loscertales, I.: The current emitted by highly conducting Taylor cones. J. Fluid Mech. **260**, 155–184 (1994)
4. Gilbert, W.: De Magnete. Wiley, New York (1893). Translated by P.F. Mottelay, original work from 1600
5. Guerrero, J., Rivero, J., Gundabala, V., Perez-Saborid, M., Fernandez-Nieves, A.: Whipping of electrified jets. Proc. Natl. Acad. Sci. USA **111**(38), 13763–13767 (2014)
6. Kim, H.Y., Lee, M., Park, K., Kim, S., Mahadevan, L.: Nanopottery: coiling of electrospun polymer nanofibers. Nano Lett. **10**(6), 2138–2140 (2010)
7. Li, D., Xia, Y.: Electrospinning of nanofibers: reinventing the wheel. Adv. Mater. **16**(14), 1151–1170 (2004)
8. Pantano, C., Ganán-Calvo, A.M., Barrero, A.: Zeroth-order, electrohydrostatic solution for electrospraying in cone-jet mode. J. Aerosol Sci. **25**(6), 1065–1077 (1994)
9. Reneker, D., Yarin, A., Fong, H., Koombhongse, S.: Bending instability of electrically charged liquid jets of polymer solutions in electrospinning. J. Appl. Phys. **87**(9), 4531–4547 (2000)
10. Ribe, N.: Coiling of viscous jets. Proc. R. Soc. Lond. A **2051**, 3223–3239 (2004)
11. Riboux, G., Marin, A., Loscertales, I., Barrero, A.: Whipping instability characterization of an electrified visco-capillary jet. J. Fluid Mech. **671**, 226–253 (2011)
12. Taylor, G.: Disintegration of water drops in an electric field. Proc. R. Soc. **280**(1382), 383–397 (1964)
13. Yarin, A., Koombhongse, S., Reneker, D.: Bending instability in electrospinning of nanofibers. J. Appl. Phys. **89**(5), 3018–3026 (2001)

Simulation of Fiber Dynamics and Fiber-Wall Contacts for Airlay Processes

Simone Gramsch, Andre Schmeißer, and Raimund Wegener

Abstract In an airlay process thousands of fibers are distributed by a turbulent air stream to produce a nonwoven. We present models and numerical strategies in order to simulate the dynamics of the fibers until they are laid down to a conveyor belt. In particular, we focus on the effect of the turbulent air flow onto the fibers and their contact with walls. The simulation results of the laydown can be used further, e.g., as input for fiber laydown models in nonwoven production processes.

Keywords Airlay process • Fiber dynamics • Fiber-wall contact • Nonwoven manufacturing • Turbulent airflow

1 Introduction

Nonwoven fabrics are defined as sheet or web structures bonded together by entangling fibers or filaments mechanically, thermally, or chemically. The emphasis is on the prefix 'non'. Nonwovens are not woven. There are three main nonwoven manufacturing processes: dry-lay processes, wet-lay processes, and extrusion processes.

Independent of the nonwoven manufacturing process there are usually two major challenges engineers deal with in textile industry. On the one hand, the nonwoven's market demands lower and lower prices for the end products, hence the production processes must be as economical as possible. On the other hand, the demand for better quality increases due to the fact that nonwovens make more an entrance in consumer goods. Both goals are inconsistent with one another. Simulating nonwoven production processes is a mathematical key technology that enables engineers to optimize the processes with respect to economics as well as quality.

S. Gramsch (✉) • A. Schmeißer • R. Wegener
Fraunhofer ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
e-mail: simone.gramsch@itwm.fraunhofer.de; andre.schmeisser@itwm.fraunhofer.de; raimund.wegener@itwm.fraunhofer.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_139

2 Airlay Processes

An airlay process is a typical example of a dry-lay process. First, the raw material is opened by a carding roller system. The carding system consists of a main rotating card cylinder and several smaller disentangling roller cards (see Fig. 1). Typically, the raw material are natural fibers, but also man-made fibers can be processed. In contrast to extrusion processes an airlay process works with short- or long-staple fibers to form the web. After their disentangling the fibers are transported in an air flow. The fibers leave the rotating card cylinder due to centrifugal forces. Then air and fibers move towards a conveyor belt that is continuously sliding into machine direction (abbreviated in the following with MD). The conveyor belt is under suction such that the fibers lay down on it and the deposited material is condensed. An additional web forming roll is placed near to the deposition zone in order to condense the laid down material further. More details about airlay processes and the next steps in the process chain like bonding or finishing can be found in [1].

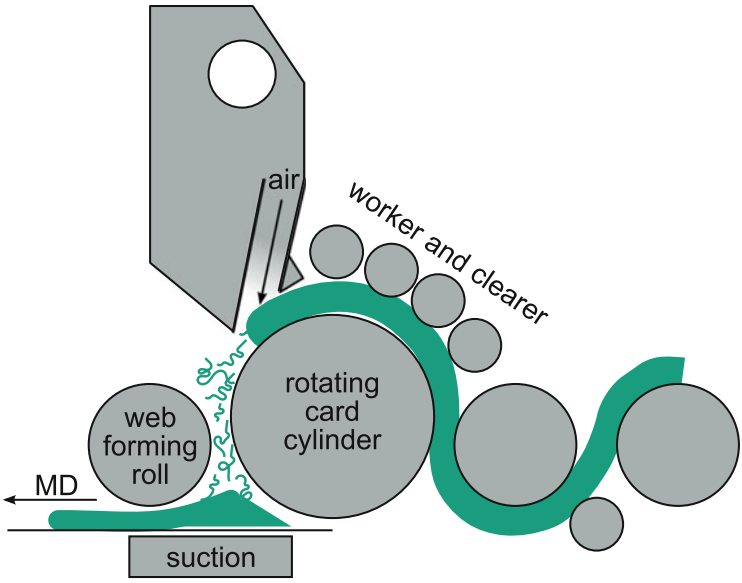


Fig. 1 Principle design of an airlay production process

3 Model for Fiber Dynamics

The basis for the modeling of the fiber dynamics is the Cosserat-Rod theory. A good overview about Cosserat-Rods can be found in [2]. The basic idea of fiber models is to describe a fiber as a curve with oriented cross-sections. By averaging over the cross-sectional area we obtain one-dimensional balance laws of linear and angular momentum. We close the system of equations by modeling the angular momentum dependent of the angular velocity and specifying the material laws. A detailed description of the fiber models is derived in [5]. We focus in the following on a model for elastic, inextensible fibers.

Let $\mathbf{r} : (s_a, s_b) \times \mathbb{R}_0^+ \to \mathbb{R}^3$ and $T : (s_a, s_b) \times \mathbb{R}_0^+ \to \mathbb{R}^3$ be the center-line and the tangential contact force of the fiber. With $s \in (s_a, s_b)$ we denote a material point of the fiber. Then the fiber dynamics equation reads for a time $t > 0$ as

$$\begin{aligned} (\rho A)\partial_{tt}\mathbf{r} &= \partial_s \left(T\partial_s\mathbf{r} - \partial_s((EI)\partial_{ss}\mathbf{r})\right) + \mathbf{f}_{ext}, \\ \|\partial_s\mathbf{r}\| &= 1. \end{aligned} \tag{1}$$

The line density and the bending stiffness of the fiber are denoted by $(\rho A)(s)$ and $(EI)(s)$, respectively. In more detail we have density ρ, Young's modulus E, cross-section area A, and moment of inertia I with $I = A^2/(4\pi)$ for a circular cross-section.

A crucial point in the fiber dynamics equation is the modeling of the external forces $\mathbf{f}_{ext}(\mathbf{r}, \partial_s\mathbf{r}, \partial_t\mathbf{r}, s, t)$. For an airlay process the dominating external force is the air drag, hence we have to deal with the effect of aerodynamics in more detail. The air drag model for fibers is based on the construction principle of an infinite cylinder that is circulated by a flow. The air drag coefficients are derived by considering the relative velocity of fiber and flow as well as the angle of attack. For more details we refer the reader to [9]. In general, the air flow in an airlay process is turbulent. Hence, we model the aerodynamic force on the fiber as a stochastic drag force that summarizes these turbulent effects as described in [9]. In [4] a new approach to reconstruct the turbulent velocity fluctuations on basis of turbulence models is developed that will be included in our simulation framework in the near future.

The fiber-wall contact is modeled by an anholonomic constraint that is based on a signed distance function $H \in \mathscr{C}^2$ representing the machinery parts. In case of a contact between a fiber point and a part of the machinery we extend the dynamics equations by the constraint $H = 0$ and an additional force $\lambda\nabla H/\|\nabla H\|$ normal to the geometry. Here, $\lambda(s, t)$ acts as a Lagrangian multiplier with respect to the constraint, i.e., in a formal notation

$$(\rho A)\partial_{tt}\mathbf{r} = \ldots + \lambda\frac{\nabla H}{\|\nabla H\|}, \quad (\lambda = 0 \ \wedge \ H > 0) \quad \vee \quad (\lambda > 0 \ \wedge \ H = 0). \tag{2}$$

In practice, we implemented a predictor-corrector scheme to detect and handle the contacts according to this model. Finally, the fiber dynamics models have to be completed by appropriate boundary and initial conditions.

4 Numerical Strategies and Implementation

Simulations of nonwoven production processes at industrial scale require stable numerics and a highly efficient implementation. First, the fiber model (1) has prototypically been implemented and tested in MATLAB. In the early stage of the software development we designed the numerical algorithms with high performance by applying the following strategy.

Equation (1) is formulated as a system of differential equations of order 1 for the state variables curve $\mathbf{r}$, velocity $\mathbf{v}$, and tangential contact force T. The semidiscretization with respect to $s \in (s_a, s_b)$ is based on a finite volume scheme, i.e., we introduce $\mathbf{r}_i$ and $\mathbf{v}_i$ as mean values in cells with center $s_i = (i-1)\ \Delta s$, $i = 1, \ldots, N$. Using a staggered grid with contact forces $T_{i+1/2}$ at the cell edges leads to

$$\partial_t \mathbf{r}_i = \mathbf{v}_i, \quad (\rho A)\partial_t \mathbf{v}_i = \mathbf{rhs}_{i+1/2} - \mathbf{rhs}_{i-1/2} + \mathbf{f}_{ext}, \quad \|\partial_s \mathbf{r}\|_{i-1/2} = 1$$
$$\mathbf{rhs}_{i+1/2} = T_{i+1/2}(\partial_s \mathbf{r})_{i+1/2} - (\partial_s((EI)\partial_{ss}\mathbf{r}))_{i+1/2}, \quad i = 2, \ldots, N-2.$$

Here, $(\partial_s \mathbf{r})_{i+1/2}$ and $(\partial_s((EI)\partial_{ss}\mathbf{r}))_{i+1/2}$ are approximated by first order finite differences. Additionally, the boundary conditions have to be specified (see [5] for more details). In summary, we gain a DAE-system that is solved by an implicit Euler method. The resulting non-linear system of equations is solved with a Newton method considering the Jacobian analytically and using an Armijo rule for the relaxation control. This strategy is highly efficient, one minor flaw is the treatment of fiber-wall contacts. Due to strong variations of Newton iterations occurring at time steps with contacts adaptive time step strategies do not show any advantages with respect to quality and costs. In practice, geometries in CFD simulations are described as triangular meshes. Particular attention should therefore be paid on the generation of the smooth distance function $H \in \mathscr{C}^2$ introduced in (2). Our approach uses a linear combination of the triangle plane distance functions. They are weighted by radial Gaussian kernels normalized to give a partition of unity. A new smoothing approach based on convolutions is currently being developed (see [3]).

To achieve maximal performance the numerical algorithms are implemented in the C++ simulation tool FIDYST (Fiber Dynamics Simulation Tool) using the Qt framework and OpenGL. The underlying CFD data must be imported in a standard data format called EnSight Gold Case that is also used for further post-processing.

5 Simulation Results

For the first time, we present fiber simulation results of the airlay process K12 from the company Autefa Solutions. Autefa Solutions specified the process parameters of a reference scenario. The simulation of the air flow was performed by the commercial flow solver FLUENT. Then a mixture of fibers was simulated. The mixture consists of 30 % bicomponent fibers (core PES, surface PET) and 70 % solid fibers (PES). Table 1 summarizes the different material properties.

The simulation of the fibers included overall 1000 fibers. The initial conditions were identical for all fibers. The starting point of the fibers leaving the rotating card cylinder was fixed. The initial velocity was assumed to be equal to the effective velocity of the rotating card cylinder. The belt is non-moving and treated in the CFD simulation as a porous medium. The total simulation time was 0.1 s. Figure 2 shows the simulation results at certain timesteps.

In the following we focus on the laydown of the fibers on the belt, since the quality of the web is one of the optimization goals. Figure 3 shows the laydown distribution of the fibers on the conveyor belt. 287 bicomponent and 559 solid fibers have reached the belt after a total simulation time of 0.1 s. The remaining fibers are still moving in the air due to turbulence effects.

Bearing in mind that all fibers have the same initial conditions (besides the variations in density) it is impressive to see their spatial distribution on the non-moving belt due to turbulent air effects. For real industrial applications the spatial distributions must be superposed in both directions: machine direction (MD) and cross machine direction (CD). Hereby, we use stochastic models that mimic the characteristic laydown properties of the fibers (see [5–7]). In order to apply such models for an industrial scale the number ΔN of stochastic fibers that fall during the time Δt in a rectangular section of width w on the moving belt must be adjusted to the real machine throughput. With given total mass rate $\dot{m}$ of the industrial production process regarding a total working width W we obtain

$$\Delta N = \frac{\dot{m}}{(\rho A) l} \cdot \frac{w}{W} \cdot \Delta t.$$

Table 1 Material properties of fibers for the reference scenario

Material property	Bicomponent fiber	Solid fiber
Line density	4.4 e-07 kg/m	6.7 e-07 kg/m
Density	1325 kg/m^3	1380 kg/m^3
Length	60 mm	60 mm
Elasticity modulus	3 kN/mm^2	3 kN/mm^2

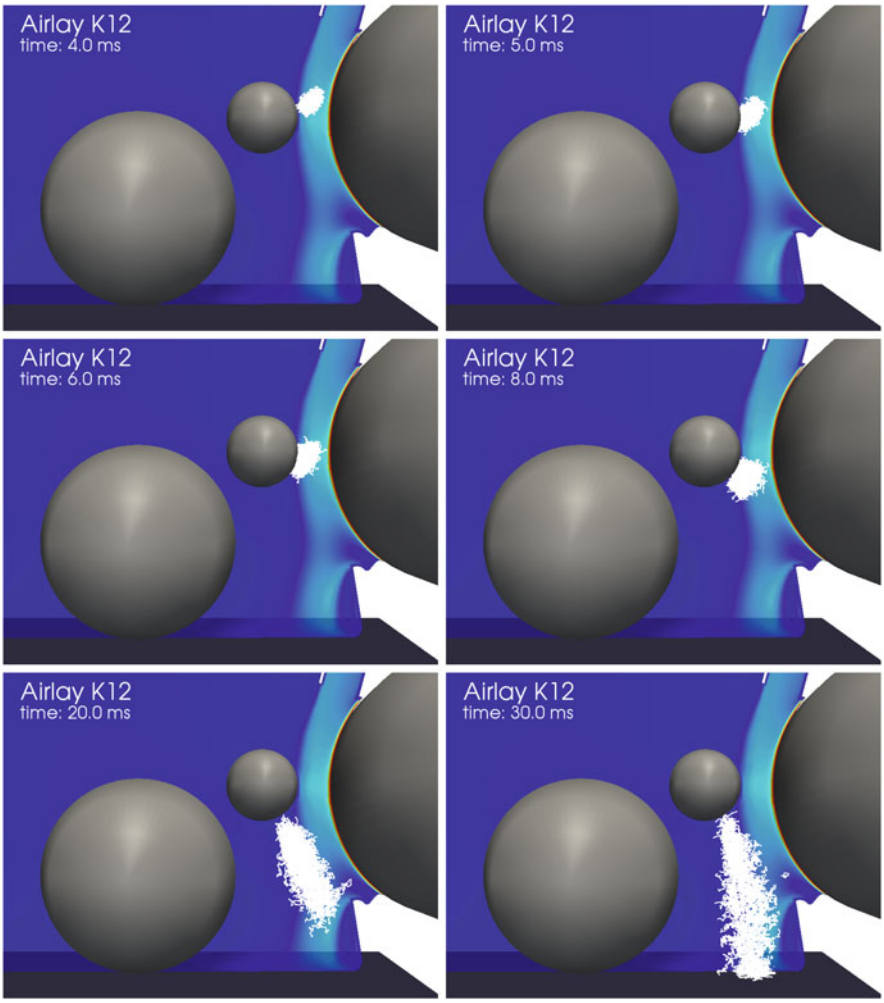


Fig. 2 Simulation results of 1000 fibers in the airlay process K12 at certain timesteps (constant time step Δt = 1e-5 s). A cut through the air flow, the geometry of the web forming roll, the baffle pipe, and the rotating card cylinder is displayed. The fibers collide with the baffle pipe and move towards the belt

Hereby, (ρA) denotes the line density and l the fiber length. Now, we can use the fiber distributions displayed in Fig. 3 as starting point for the surrogate stochastic models and the optimization of the nonwoven quality.

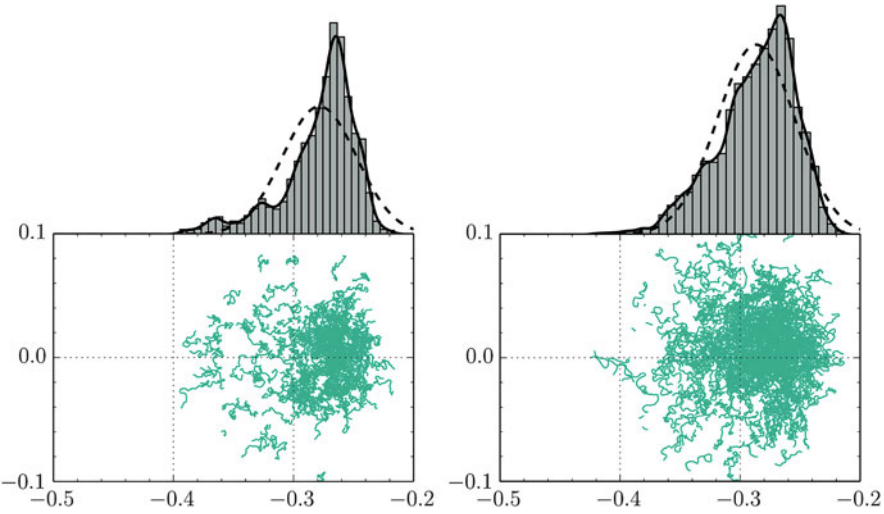


Fig. 3 Fiber laydown distributions. The normalized histogram (*top*) is obtained by summarizing the fiber mass (*bottom*) in cross machine direction and fitted by a normal distribution (*dashed line*) and a Gaussian kernel density estimation (*continuous line*)

6 Conclusion

We have presented a summary of a fiber dynamics model that is appropriate to simulate an airlay process. We have sketched the numerical algorithm that is used to solve the dynamic equations for elastic, inextensible fibers. They form the basis of the software tool FIDYST (Fiber Dynamics Simulation Tool) that is capable of simulating nonwoven production processes at industrial scale. Furthermore, simulations of fiber dynamics in an airlay process have successfully been performed for the first time. A short discussion of the laydown results shows how mathematical techniques can be used in order to optimize the product quality.

Obviously, the fibers on the belt have a big influence on the air flow since they block the suction through the belt. Hence, studying the build up of the nonwoven on the conveyor belt is the next step for a deeper understanding of the airlay process. Additionally, analyzing the effect of the fibers on the air is a challenging issue. First ideas to consider the back-coupling of fibers to air are derived in [8] and will be developed further. We also expect that including the turbulence reconstruction model will significantly improve the validity of simulations for nonwoven production processes.

Acknowledgements The authors would like to acknowledge their industrial partner, the company Autefa Solutions GmbH, for the interesting and challenging problem. This work has been supported by German Bundesministerium für Bildung und Forschung, Schwerpunkt ‘Mathematik für Innovationen in Industrie und Dienstleistungen’, Project OPAL 05M13AMD.

References

1. Albrecht, W., Fuchs, H., Kittelmann, W. (eds.): Nonwoven Fabrics: Raw Materials, Manufacture, Applications, Characteristics, Testing Processes. Wiley-VCH Verlag GmbH, New Jersey (2003)
2. Antman, S.S.: Nonlinear Problems of Elasticity. Springer, New York (2006)
3. Arne, W., Leithäuser, C., Schmeißer, A.: Modeling and simulation along the process chain for filaments and nonwovens. In: Proceedings of the 2nd Young Researcher Symposium (YRS) 2013, pp. 78–83. Fraunhofer Verlag, Kaiserslautern
4. Hübsch, F., Marheineke, N., Ritter, K., Wegener, R.: Random field sampling for a simplified model of melt-blowing considering turbulent velocity fluctuations. J. Stat. Phys. **150**(6), 1115–1137 (2013)
5. Klar, A., Marheineke, N., Wegener, R.: Hierarchy of mathematical models for production processes of technical textiles. ZAMM **89**, 941–961 (2009)
6. Klar, A., Maringer, J., Wegener, R.: A 3d model for fiber lay-down in nonwoven production processes. Math. Models Methods Appl. Sci. **22**(9), 1250020:1–18 (2012)
7. Klar, A., Maringer, J., Wegener, R.: A smooth 3d model for fiber lay-down in nonwoven production processes. Kinet. Relat. Models **5**(1), 57–112 (2012)
8. Marheineke, N., Liljo, J., Mohring, J., Schnebele, J., Wegener, R.: Multiphysics and multimethods problem of rotational glass fiber melt-spinning. Int. J. Num. Anal. Mod. B **3**(3), 330–344 (2012)
9. Marheineke, N., Wegener, R.: Modeling and application of a stochastic drag for fiber dynamics in turbulent flows. Int. J. Multiphase Flow **37**, 136–148 (2011)

MS 39
MINISYMPOSIUM: THE EMERGING DISCIPLINE OF PHARMACOMETRICS: AT THE CROSSROAD OF MATHEMATICS AND MODERN PHARMACEUTICAL SCIENCES

Organizers

Fahima Nekka,[1] Jun Li[2] and Matti Heilio[3]

Speakers

Fahima Nekka[1]
Old Problems, New Solutions: Variability and Nonlinearity in Biopharmaceutical Processes and Mathematical Model-Based Problem Solving

Marco Veneroni[4]
Analysis of a Variational Model for Liquid Crystal Shells

Matylda Jablonska-Sabuka[5]
Drug Dose Optimization in HIV Treatment

[1]Fahima Nekka, Université de Montréal, Canada.
[2]Jun Li, Université de Montréal, Canada.
[3]Matti Heilio Lappeenranta University of Technology, Finland.
[4]Marco Veneroni,, Università degli Studi di Pavia, Italy.
[5]Matylda Jablonska-Sabuka, Lappeenranta University of Technology, Finland.

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_140

Keywords

Drug variability
Modeling and simulation
Optimisation of drug use
Pharmacometrics

Short Description

In the drug development arena, the rapid accumulation of new quantitative methodologies and tools pushed the emergence of systemic and mechanistic studies of pharmacology that drive the drug R&D. Particularly, tools based on modeling and simulation (M&S) gained a large popularity in the milieu considering the increasing number of success stories involving M&S. The efficient use of these tools heavily relies on advanced mathematical methodologies and their appropriateness to the problem at hand. This minisymposium will exemplify this field with mathematical applications to concrete pharmaceutical problems.

A Probabilistic Strategy for Group-Based Dose Adaptation

Guillaume Bonnefois, Olivier Barrière, Jun Li, and Fahima Nekka

Abstract Individualized dose adaptation usually requires Therapeutic Drug Monitoring (TDM) based on patient blood samplings. However this invasive approach, generally accompanied with discomfort and cost, is not always justified since it may occur that the resulting dose adaptation does not significantly differ in a population whose individuals share similar characteristics. Inspired by the principle of maximum likelihood, we propose a probabilistic approach, based on population-pharmacokinetic modeling and simulation, to evaluate the therapeutic performance of a dosing regimen in terms of dose and time. Two types of therapeutic indicators, time-based and concentration-based, are suggested to assess quantitatively different drug regimens with the aim to identify the optimal one. For the population under investigation, our results identified a stable and robust optimal regimen and determined critical times including toxicity. Moreover, for a same therapeutic target, our approach enables to identify more than one corresponding regimen, giving thus a great flexibility in clinical practice.

Keywords Population-pharmacokinetic modeling • Probabilistic approach • Therapeutic drug monitoring

G. Bonnefois • F. Nekka (✉)
Faculté de Pharmacie, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, QC, Canada H3C 3J7
e-mail: guillaume.bonnefois@umontreal.ca; fahima.nekka@umontreal.ca

O. Barrière
Faculté de Pharmacie, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, QC, Canada H3C 3J7

Inventiv Health, Montréal, QC, Canada
e-mail: olivier.barriere@inventivhealth.com

J. Li
Faculté de Pharmacie, Centre de Recherches Mathématiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, QC, Canada H3C 3J7
e-mail: li@crm.umontreal.ca

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_141

1 A Probabilistic Strategy to Assist Dose Adaptation

Dose adaptation enables to tailor a patient's specific dose and time schedule in order to maximise efficacy while minimising toxicity [1]. It aims to target the best therapeutic outcomes for an individual or a specific population and is applied for a variety of drugs, especially those exhibiting narrow therapeutic windows such as immunosuppressants and anticancer drugs [2]. The Population Pharmacokinetics (Pop-PK) approach, taking into account the inter and intra individual variability, has been adopted to model relationships between dose, concentration, and effect/toxicity [4].

Several approaches have been used for dose adaptation using linear relationship principles, nomograms, and individual PK Bayesian estimation methodologies [2]. These methods, either being deterministic, overlooking thus the individual particularity, or involving an invasive blood sampling, motivate the search of alternative methods, with the objective to determine a uniform dosing regimen for a subpopulation, for example, a group of individuals that share similar characteristics. Working in this direction, we here develop a Pop-PK model-based computational strategy for the selection of the optimal drug regimen accounting for dose as well as administration time schedule.

The current work provides a solution for dose adaptation from a probabilistic point of view, which, for reasons of limitations in blood sampling, can be of particular relevance for sensitive populations such as pediatrics and geriatrics.

1.1 Regimen Performance

Our computational strategy aims to determine the best drug regimen that gives rise to a therapeutic target based on a proposed quantitative performance. Within the framework of Pop-PK modeling, the performance can be defined as:

$$\text{Performance}(\text{Regimen}, \text{Pop-PK model})\Big|_{\text{Drug Pop-PK model}}$$

1.2 Regimen Design

A dosing regimen will be defined on a daily basis with the following notations:

$$\text{Regimen} = (\mathbf{D}, \boldsymbol{\tau}), \text{ where}$$

$$\mathbf{D} = (D_1, D_2, D_3, \cdots, D_k)$$
$$\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3, \cdots, \tau_k), \qquad \tau_1 < \tau_2 < \cdots < \tau_k$$

Each pair (D_i, τ_i), $i = 1, \cdots, k$, represents a dose and its corresponding dosing time.

Thus, the total daily dose (TDD) is:

$$\text{TDD} = \sum_{i=1}^{k} D_i.$$

When performing a regimen selection, all possible dose regimens will be evaluated and compared.

1.3 Criteria for Screening Dosing Regimens at Steady-State

In order to evaluate the dosing regimens, we refer to the Therapeutic Window (TW) of a drug for which PK and effects are linked. TW is defined by a minimum effective concentration ($TW_{\min}$) and a minimum toxic concentration ($TW_{\max}$).

The performance of the PK profile of a particular dosing regimen will be evaluated in terms of TW. For this, we define four therapeutic indicators (TI) for the selection of the best drug regimen. Two types of TI are proposed as follows.

1.3.1 Time-Based Therapeutic Indicators

One time-based TI is defined as the daily time spent by a steady-state PK profile within the TW. This is named the effective time TI_{eff} of a PK profile and given by:

$$TI_{Eff} = Length\{t : TW_{\min} \leq C(t) \leq TW_{\max}\} = \int_{1\,\text{day}} \chi_{TW}(C(t))dt$$

where $\chi_{TW}(C(t)) = 1$ if the value of $C(t)$ is within TW, and $\chi_{TW}(C(t)) = 0$ otherwise.

Another time-based TI refers to the toxicity of the PK profile and is defined as the daily time where the concentrations are beyond $TW_{\max}$. This is named the toxic time TI_{tox} of a PK profile and given by:

$$TI_{Tox} = Length\{t : C(t) > TW_{\max}\} = \int_{1\,\text{day}} \chi_{[TW_{\max}, +\infty)}(C(t))dt$$

where $\chi_{[TW_{\max}, +\infty)}(C(t)) = 1$ if the value of $C(t)$ is beyond $TW_{\max}$, and 0 otherwise.

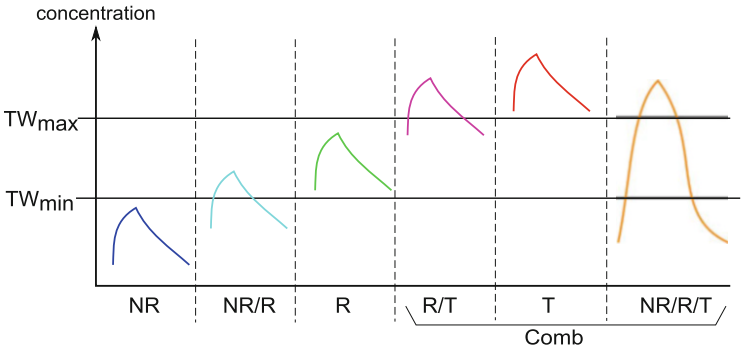


Fig. 1 PK profiles corresponding to the six therapeutic categories. Non responders (NR), responders (R), and toxic (T) are simple categories whereas hybrid ones are comprised of NR/R, R/T, and NR/R/T. Comb groups categories where toxicity arises (Comb = R/T+T+NR/R/T)

1.3.2 Concentration-Based Therapeutic Indicators

The performance of dosing regimens based on different concentration levels that the associated PK profile may reach in reference to TW can be evaluated. For this, three therapeutic zones that lie below, within, and beyond TW are referred to as non-effective, effective, and toxic zones, respectively. We define six mutually exclusive therapeutic categories which are further divided into simple (the whole daily PK profile remains within a unique zone) and hybrid classes (the PK profile moves between zones) as illustrated in Fig. 1. In the current paper, we only consider the use of concentration-based TI: TI_R and $\mathrm{TI}_{Comb} = \mathrm{TI}_T + \mathrm{TI}_{R/T} + \mathrm{TI}_{NR/R/T}$.

1.3.3 Evaluation of Dosing Regimens

Based on the TIs defined above, the performance of dosing regimens is evaluated using Monte-Carlo simulations (N=1000). Indeed, each dosing regimen generates a large variety of PK profiles which may belong to different categories. This is induced by the variability present in the Pop-PK model. To use the time-based TI, the average toxic time (TI_{Tox}) or the average effective time (TI_{Eff}) can be evaluated as follows:

$$TI = \frac{1}{N}\sum_{i=1}^{N} TI_i$$

where N is the total number of simulated PK profiles and TI refers to either TI_{eff} or TI_{tox}.

Similarly, the concentration-based TI (TI_R and TI_{Comb}), can be determined for a given dosing regimen. Its probability to produce PK profiles belonging to a CAT

(here R and Comb) is defined as:

$$TI_{CAT}(Regimen) = Prob(\text{CAT}) = \frac{1}{N} \times \#(\text{PK} \in \text{CAT})$$

where # indicate the number of PK profiles belonging to the specific CAT.

1.3.4 Selection of the Best Regimen

Using the previously described quantitative evaluation of the dosing regimen, the best regimen is selected as follows. A multi-objective approach is considered with a combination of TI with associated weights in order to determine the Regimen $= (\mathbf{D}, \boldsymbol{\tau})$ that maximizes (or minimizes) this combination. For this, a pool of regimens was set up, for which the TIs values, including their maximum TI_{max} and minimum TI_{min}, are calculated. Then the performance of each dosing regimen is evaluated by:

$$\text{Performance}(Regimen) = \sum_{i=1}^{I} w_i \begin{cases} \dfrac{TI_{\max} - TI_{Regimen}}{TI_{\max} - TI_{\min}} & \text{for minimization} \\ \dfrac{TI_{Regimen} - TI_{\min}}{TI_{\max} - TI_{\min}} & \text{for maximization,} \end{cases}$$

where w_i are the weights suggested or supported by the clinical experience and chosen by the user in order to favor or penalize $\text{TI} = \text{TI}_i$ on the right of bracket and $\sum_{i=1}^{I} w_i = 1$, I is the number of TIs considered. The normalization is necessary for the uniformity of units so that comparison can be made.

1.3.5 Software and Implementation

Two major software platforms were used for the development and implementation of the algorithm. NONMEM (version VII, Icon Development Solutions, Ellicott City, MD) was used to simulate the steady-state PK profiles. Implementation in MATLAB (R2008, MathWorks, Inc.) allowed for the simulation of dosing regimen, data analysis and the production of graphical outputs as illustrated in Fig. 2.

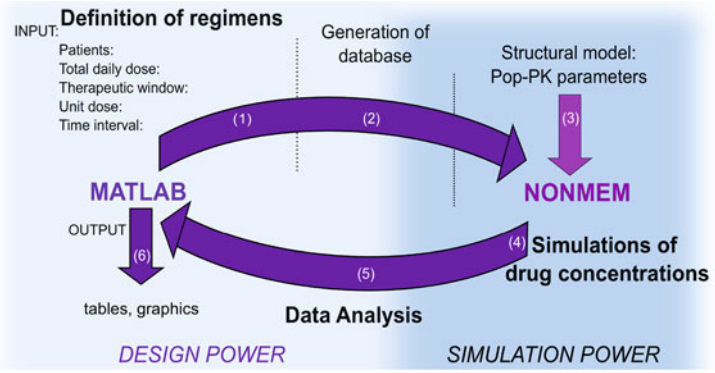


Fig. 2 Overview of the algorithm. In a first step, the number of patients, total daily dose and therapeutic index were defined. Then giving a dose unit and time interval enables to define regimens. Drug concentrations were simulated and, for each regimen, TIs are calculated to obtain the performance

2 Results

Figures 3 and 4 illustrate numerically and graphically regimen performance for the QD regimen using carbamazepine as a drug illustrative case with its associated Pop-PK model [3], where weights for TIs emphasising toxicity rather than efficacy have been considered.

In Fig. 3, the left panel shows the simulated steady state concentration distributions, from the lower 10 % to the upper 90 %, which have been calculated to highlight time-based TIs. In the right panel, a pie chart represents the probability partition of the six therapeutic categories for all generated PK profiles for the concentration-based TIs.

Moreover, for the same therapeutic target (ex: regimens for which $TI_R > 50\,\%$), we are able to identify more than one regimen that satisfy this criterion.

In Fig. 4, an additional characterization of the PK profiles is presented. In the upper panel, the time evolution of percentages of concentrations compared to TW are illustrated for each simple or hybrid category. These conditional probabilities allow the identification of critical therapeutic times including toxicity.

In the lower panel of Fig. 4, the distribution of effective times of PK profiles in each therapeutic category are reported. While the results are trivial for simple categories (NR, R, T), different patterns can be observed for hybrid categories and allow a better investigation and evaluation of the benefit and risk for a given regimen.

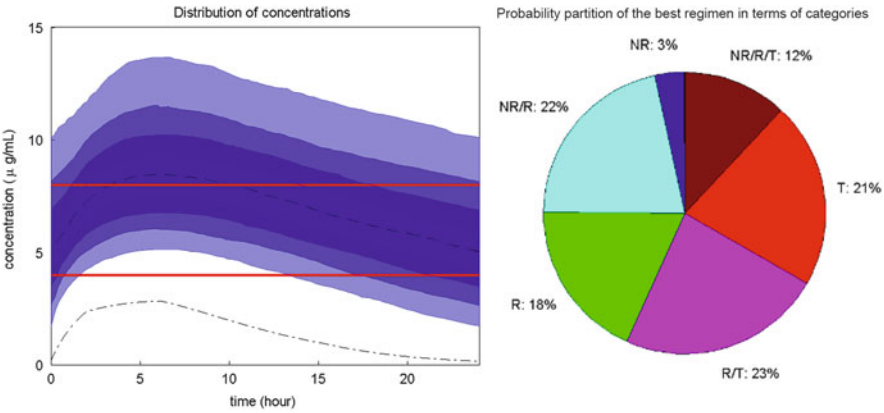


Fig. 3 *Left panel*: Distribution of concentrations at steady state during a 24 h time interval for a given regimen. The *middle dotted line* represents the median of concentrations at any time. The TW is indicated with *two thick horizontal red lines*; *Right panel*: Probabilities of PK profiles generated from this given regimen belonging to six therapeutic categories

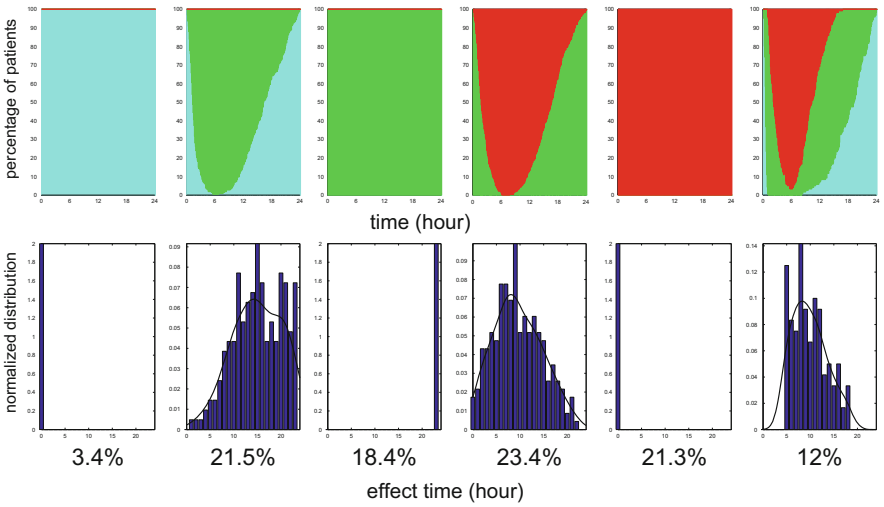


Fig. 4 *Upper panel*: Time evolution of percentages of concentrations below (*cyan*), within (*green*) and beyond (*red*) the therapeutic window; *Lower panel*: the distribution of effective times of PK profiles. From *left to right*: NR, NR/R, R, R/T,T and NR/R/T with their corresponding percentages. Percentage indicated below the figure are the probabilities of PK profile generated from the given regimen belonging to six CAT as illustrated in Fig. 3, *right*

3 Conclusion

In this paper, a new computational methodology for dose adaptation that integrates a Pop-PK modeling and simulation has been proposed and developed. Based on the concept of TW, several TIs have been revisited in the framework of the Pop-PK approach to evaluate the performance of dosing regimens. This allows us to determine the optimal regimens in terms of doses and dosing times. In the context of drug research and development continuum, our approach can be iteratively applied through clinical phases to refine the search for dosing and schedule.

Acknowledgements The authors acknowledge the financial support of NSERC-Industrial Chair in Pharmacometrics, FRQNT, NSERC, Mprime, Novartis, Pfizer and InVentiv Health.

References

1. Canal, P., Gamelin, E., Vassal, G., Robert, J.: Benefits of pharmacological knowledge in the design and monitoring of cancer chemotherapy. Pathol. Oncol. Res. **4**(3), 171–178 (1998)
2. de Jonge, M.E., Huitema, A.D.R., Schellens, J.H.M., Rodenhuis, S., Beijnen, J.H.: Individualised cancer chemotherapy: strategies and performance of prospective studies on therapeutic drug monitoring with dose adaptation: a review. Clin. Pharmacokinet. **44**(2), 147–173 (2005)
3. Punyawudho, B., Ramsay, E.R., Brundage, R.C., Macias, F.M., Collins, J.F., Birnbaum, A.K.: Population pharmacokinetics of carbamazepine in elderly patients. Ther. Drug Monit. **34**(2), 176–181 (2012). doi: 10.1097/FTD.0b013e31824d6a4e. URL http://dx.doi.org/10.1097/FTD.0b013e31824d6a4e
4. Sheiner, L.B., Beal, S., Rosenberg, B., Marathe, V.V.: Forecasting individual pharmacokinetics. Clin. Pharmacol. Ther. **26**(3), 294–305 (1979)

Part II
Contributed Sessions

αAMG Based on Weighted Matching for Systems of Elliptic PDEs Arising from Displacement and Mixed Methods

Pasqua D'Ambra and Panayot S. Vassilevski

Abstract Adaptive Algebraic Multigrid (or Multilevel) Methods (αAMG) are introduced to improve robustness and efficiency of classical algebraic multigrid methods in dealing with problems where no a priori knowledge or assumptions on the near-null kernel of the underlined matrix are available. Recently we proposed an adaptive (bootstrap) AMG method, αAMG, aimed to obtain a composite solver with a desired convergence rate. Each new multigrid component relies on a current (general) *smooth vector* and exploits pairwise aggregation based on weighted matching in a matrix graph to define a new automatic, general-purpose coarsening process, which we refer to as "*the compatible weighted matching*". In this work, we present results that broaden the applicability of our method to different finite element discretizations of elliptic PDEs. In particular, we consider systems arising from displacement methods in linear elasticity problems and saddle-point systems that appear in the application of the mixed method to Darcy problems.

Keywords Adaptive algebraic multigrid method • Weighted matching

1 Introduction

Algebraic Multigrid Methods (AMG) were introduced in the mid-1980s as *plug-in* solvers for large and sparse linear systems of equations $A\mathbf{x} = \mathbf{b}$, with the final aim to define automatic coarsening process only depending on the coefficient matrix [4, 15]. These methods are particularly efficient for systems arising from scalar second-order elliptic partial differential equations (PDEs), where a characterization of the *algebraically smooth error*, which is the error component not reduced by a simple relaxation scheme (such as Gauss-Seidel relaxation), is available [9]. This

P. D'Ambra (✉)
ICAR-CNR, Via P. Castellino, 111, 80131 Napoli, Italy
e-mail: pasqua.dambra@cnr.it

P.S. Vassilevski
CASC-LLNL, P.O. Box 808, L-561, Livermore, CA 94551, USA
e-mail: panayot@llnl.gov

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_142

error, corresponding to the eigenvectors of A with small associated eigenvalues (*near-null kernel of A*), must be nearly exactly represented in the coarse-space in order to be eliminated by the coarse-grid correction process. Therefore, a main focus in the current state-of-the-art AMG methods is to define strategies for building coarse variables and intergrid operators which are able to adapt themselves to the properties of the near-null kernel of the problem at hand in order to preserve efficiency and robustness for dealing with more general classes of problems than the traditional scalar elliptic PDEs, including systems of elliptic PDEs, convection-diffusion equations and also more general non-PDE problems. In this direction, Adaptive Algebraic Multigrids (αAMG) have been proposed [6, 8], where main idea is to use appropriate adaptive steps aimed to "identify" smooth error components which the current solver is not able to efficiently handle so that they can be used to improve the solver by modifying the coarsening scheme without using any specific a priori knowledge about these error components. In [11] we proposed a new αAMG method which relies on a bootstrap strategy aimed to compute a composite solver with a desired convergence rate. We demonstrated its effectiveness when applied to symmetric positive definite (s.p.d.) systems arising from finite element discretization of highly anisotropic scalar elliptic PDEs on structured and unstructured meshes. Here, we extend the application of the method to systems of elliptic PDEs coming from linear elasticity and Darcy flow in porous media in mixed setting. In Sect. 2 we outline the αAMG based on the compatible weighted matching; in Sect. 3 we describe the model problems and introduce the Bramble-Pasciak transformation used for extending our method in dealing with symmetric indefinite systems stemming from the mixed finite element discretization of Darcy problems; finally, in Sect. 4 we present results obtained by a prototype Matlab version of our αAMG solver.

2 Main Features of αAMG Based on Compatible Weighted Matching

In [11], we proposed a bootstrap process aimed to build a composite solver of the following form:

$$\mathbf{x}_k = \prod_{r=0}^{2m+1} (I - B_r^{-1}A)\mathbf{x}_{k-1}, \qquad k = 1, 2, \ldots, \tag{1}$$

with $B_{m+r} = B_{m+1-r}$, $r = 1, \ldots, m+1$. Each B_r is an AMG-cycle built with its own aggregation procedure of unknowns driven by a weighted matching for the original matrix graph with weights depending on the most recently computed algebraically smooth vector $\mathbf{x}_k$ with respect to the current composite solver. In more details, starting from a general (random) given vector, we build an initial AMG-cycle represented by the operator B_0 and apply it to the homogeneous system

$A\mathbf{x} = 0$, starting with a nonzero random initial iterate $\mathbf{x}_0$, and successively computing $\mathbf{x}_k := (I - B_0^{-1}A)\mathbf{x}_{k-1}$ for a fixed number of iterations. The iterative process provides an approximation to the eigenvector of $B_0^{-1}A$ corresponding to the minimal eigenvalue of $B_0^{-1}A$, i.e., of the algebraically smooth vector corresponding to the current solver. This last vector is then used to build a new AMG-cycle represented by the operator B_1 to be composed as in (1) and tested on the homogeneous system. The bootstrap process is stopped when the process represented by (1) reaches a desired convergence rate.

Each new AMG operator B_r is built by using pairwise aggregation of unknowns driven by weighted matching algorithms for the matrix graph. Such matching algorithms are widely exploited in sparse matrix computations to enhance matrix diagonal dominance [12]. More aggressive coarsening (than pairwise aggregation) can be obtained by combining multiple steps of the pairwise aggregation. Our main idea was to exploit the concept of compatible relaxation introduced in [5] for selecting the coarse-vector space. Since for the coarse space, we choose piecewise constant interpolant (that interpolates exactly the current smooth vector), we choose a complementary space such that on each aggregate (of pair of vertices) it is spanned by a vector orthogonal to the restriction of the smooth vector to that (pairwise) aggregate. To actually choose the aggregates, we use weights based on these orthogonal vectors so that the resulting A_f matrix corresponding to the space complementary to the coarse space have maximal product of its diagonal entries. For the actual details on the respective algorithms and results on scalar PDEs, we refer to [11]. Here, we investigate the use of more accurate interpolation operators obtained by weighted-Jacobi smoothing of the piecewise constant interpolation operators coupled with aggressive coarsening. This leads to smoothed aggregation type adaptive AMG method [7], which exhibits improved convergence and scalability properties with general reduction of setup costs.

Our coarsening process, which we referred to as *compatible weighted matching*, has the advantage to be independent of user-defined parameters; furthermore, it overcomes the limitations of the characterization of strength of connectivity between pairs of unknowns, well motivated only for algebraic systems with M-matrices. The latter concept is generally used in both the coarse space selection and in the interpolation scheme for classical AMG schemes. We stress that computing optimal matching has a super-linear computational complexity, whereas we are interested in (optimal) AMG with linear complexity, that is why we apply an approximate algorithm to find sub-optimal weighted matchings in a graph [13]; this approach was demonstrated to be effective in computing suitable compatible weighted matchings in the difficult case of highly non-grid aligned anisotropic scalar elliptic PDEs.

3 Case Studies: Linear Elasticity and Darcy Problems

We focus on two types of elliptic PDEs particularly relevant for many engineering applications, such as Lamé equations for linear elasticity and Darcy equations for flow in porous media in mixed system setting. Of main interest is to demonstrate

the feasibility of our method on general s.p.d linear systems, where the coefficient matrix is not an M-matrix, as well as, on some symmetric but indefinite systems of saddle-point form.

The most widely used mathematical model for studying deformation of materials due to the application of external forces are the following Lamé equations, which are equilibrium equations written in terms of the displacement field $\mathbf{u}$:

$$\mu \Delta \mathbf{u} + (\lambda + \mu)\,\mathrm{grad}(\mathrm{div}\mathbf{u}) = \mathbf{f} \qquad \mathbf{x} \in \Omega \tag{2}$$

where $\mathbf{u} = \mathbf{u}(\mathbf{x})$ is the displacement vector, Ω is the 3D spatial domain, and λ and μ are the Lamé constants. A mix of Dirichlet boundary conditions and so-called traction conditions are usually applied to have a unique solution. Discretization of (2) by finite element method, if each scalar component of the displacement vector $\mathbf{u} = (u, v, w)$ is considered separately (*unknown-based* [14] discretization), leads to s.p.d. systems of equations whose coefficient matrix can be written in the following block form:

$$A = \begin{bmatrix} A_{uu} & A_{uv} & A_{uw} \\ A_{vu} & A_{vv} & A_{vw} \\ A_{wu} & A_{wv} & A_{ww} \end{bmatrix}$$

We note that if $\mu >> \lambda$, the above matrix is spectrally equivalent to its block diagonal, corresponding to the matrix coming from discretization of Laplace equation per each unknown component. In this case, block-wise version of the classical AMG are efficient solver. In general, A is not strongly block-diagonally dominant and problem-dependent multigrid operators have to be considered to improve convergence of AMG [1]. In the present work we demonstrate that our αAMG is able to obtain a solver with a desired convergence rate for general elasticity problems, without any a priori information on the problem neither on the discretization scheme.

The second type of systems of PDEs we considered in this work comes from the Darcy problem of flows in porous media. It is a boundary value problem associated to the following second order elliptic equation:

$$-\,\mathrm{div}k(\mathbf{x})\,\mathrm{grad}\,p = f(\mathbf{x}) \qquad \mathbf{x} \in \Omega, \tag{3}$$

where $p = p(\mathbf{x})$ is the flow pressure, Ω is the spatial domain, and $k(\mathbf{x})$ is the permeability coefficient. In a mixed finite-element formulation, the flow velocity field $\mathbf{u} = -k\nabla p$ is introduced and Eq. (3) becomes $\mathrm{div}\mathbf{u} = f$. The resulting problem is a system of two first order vector equations which can be discretized by using a pair of finite element spaces leading to the following indefinite system of saddle-point form:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ f \end{bmatrix},$$

where A is an s.p.d. matrix. Such linear systems, especially for highly variable or discontinuous permeability coefficient, are very challenging for general iterative solvers, and more specifically for algebraic multigrid (see [2, 16]). Here we propose to use an approach based on the Bramble-Pasciak preconditioner [3] which transforms the saddle-point matrix into a s.p.d. matrix. They utilize a preconditioner matrix M for the A block, such that $A - M$ is s.p.d., and transform the saddle-point matrix into the following s.p.d. one:

$$\widehat{\mathscr{A}} = \begin{bmatrix} AM^{-1} - I & 0 \\ BM^{-1} & -I \end{bmatrix} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} = \begin{bmatrix} AM^{-1}A - A & (AM^{-1} - I)B^T \\ B(M^{-1}A - I) & BM^{-1}B^T \end{bmatrix}. \tag{4}$$

A good choice in practice is a diagonal matrix M assembled from the local element-based diagonal matrices $diag(M_{fem})$, where $M_{fem} = 1/2\lambda_{min}D_{fem}$ and $D_{fem} = diag(A_{fem})$. Here, A_{fem} is the local element mass matrix for each finite element and λ_{min} is the minimal eigenvalue of the generalized local eigenvalue problem $A_{fem}\mathbf{q} = \lambda D_{fem}\mathbf{q}$. In this case the transformed matrix (4) can be explicitly computed at a cost of a moderate increase in the total number of nonzero elements. In the following Section we report some numerical results related to the application of our adaptive AMG on Darcy problems discretized by the mixed finite-element method, in the above transformed s.p.d. form.

4 Numerical Results

In this Section we report some preliminary results which illustrate the ability of our method to solve the systems of equations introduced in Sect. 3 both in 2D and in 3D domains. We investigate the convergence behavior and the setup cost for increasing mesh size of the discretization. The setup cost is measured in terms of AMG components *nstages* built by the bootstrap process to reach a desired convergence rate set to 0.7. The obtained convergence rate ρ was estimated by applying the solver in (1) for 15 iterations at each new built. We also report, per each test case and per each mesh with n nodes, the average number of levels *nlev* of all solver components and the average of their operator complexity *cmpx*, which gives information on the cost of the application of one cycle; *cmpx* is defined as the ratio between the sum of nonzero entries of the matrices of all levels and the number of nonzero entries of the fine-grid matrix. Each AMG component, built on the base of the *compatible weighted matching* coarsening method, is a general v-fold cycle [16], where one-sweep is alternated with three sweeps in the next level; In this way, we ensure linear cost per cycle since our coarsening is based on pairwise aggregation. Symmetric Gauss-Seidel relaxation (one iteration) is employed as pre/post smoothing while direct solver (based on LU factorization) is used at the coarsest level. In order to achieve aggressive coarsening we combine four steps of pairwise aggregation based on compatible matching, which allows us to define coarse matrices with a

Table 1 Linear elasticity problems: setup cost when unsmoothed (on the left) and smoothed aggregation (on the right) are used

Composite αAMG setup						Composite αAMG setup					
	n	nstages	ρ	nlev	cmpx		n	nstages	ρ	nlev	cmpx
beam 2D	4386	8	0.61	3	1.12	beam 2D	4386	5	0.69	3	1.25
	16,962	10	0.69	3	1.12		16,962	7	0.63	3	1.20
	66,690	12	0.69	4	1.12		66,690	9	0.68	4	1.23
	264,450	20	0.68	5	1.10		264,450	12	0.70	5	1.19
beam 3D	2475	9	0.61	3	1.20	beam 3D	2475	8	0.53	3	1.53
	15,795	11	0.67	3	1.20		15,795	10	0.60	3	1.78
	111,843	16	0.67	4	1.23		111,843	12	0.64	4	2.40
	839,619	25	0.57	5	1.09		839,619	17	0.61	5	1.34

coarsening ratio of at most 16 at each level; the process is stopped when the size of the coarsest matrix is at most 100.

As test case for linear elasticity, we consider Eq. (2) on a beam characterized by $\mu = 0.42$ and $\lambda = 1.7$; one side of the beam is considered fixed and the opposite end is pushed downward. The problem is discretized using linear finite elements on triangular (2D) and tetrahedral meshes (3D) on different mesh sizes, obtained by uniform refinement, with the software package MFEM (http://mfem.googlecode.com). In Table 1, we summarize our results obtained both in the case of constant piecewise interpolation, i.e., unsmoothed aggregation, (on the left) and with smoothed aggregation (on the right). We observe that our method is able to achieve convergence factors less than the desired one for all the cases, although no a priori information on the spectral properties of the matrices neither on the particular features of the system of PDEs and of its discretization were used. We notice that the number of the necessary bootstrap steps generally increases with increasing the mesh size, especially for 3D problems; the largest size mesh requires five more bootstrap steps with respect to the medium size mesh. The total number of bootstrap steps, as expected, is reduced if the smoothed aggregation is applied; furthermore, smoothed aggregation coupled with our aggressive coarsening based on a combination of more steps of pairwise aggregation produces a moderate increase in the operator complexity, leading to a general reduction both in the setup and the application cost of the method.

For the Darcy problems, we consider saddle-point systems stemming from a realistic problem with highly variable permeability coefficients, describing a 3D petroleum reservoir obtained from the 10th Society of Petroleum Engineers (SPE) Comparative Solution Project [10]. We present results for Dirichlet problems (i.e. pressure given on the boundary) discretized by using MFEM with structured hexahedral meshes. For discretization, we used first-order Raviart-Thomas spaces [16] for velocity and piecewise-constant functions for pressure. We apply the Bramble-Pasciak transformation described in Sect. 3 to obtain the corresponding s.p.d. matrix (4). We observe that for the considered test case and the employed

Table 2 Darcy problems: setup cost when unsmoothed (on the left) and smoothed aggregation (on the right) are used

Composite αAMG Setup						Composite αAMG Setup					
	n	nstages	ρ	nlev	cmpx		n	nstages	ρ	nlev	cmpx
SPE10	1403	2	0.50	2	1.07	SPE10	1403	2	0.57	2	1.12
	10,652	3	0.68	3	1.12		10,652	3	0.69	3	1.34
	33,645	5	0.65	4	1.13		33,645	5	0.66	4	1.46
	88,800	7	0.65	4	1.14		88,800	6	0.69	4	1.54

mesh sizes, the number of nonzeros in the transformed matrix has an increase of about 80 % with respect to the original saddle-point matrix. In Table 2 we report results for different mesh sizes (note that here n is the size of the saddle-point matrix) for both unsmoothed and smoothed aggregation, when the algorithmic choices were the same as in the elasticity problems. We observe that the adaptive solver is able to obtain the required convergence rate with a moderate number of setup steps, demonstrating the potential of the coupling between Bramble-Pasciak transformation and the adaptive solver to handle well indefinite systems of saddle-point type coming from realistic flow problems. The increase in the number of bootstrap steps needed to obtain the desired convergence rate for increasing mesh size is moderate, showing good scalability properties also in the case of unsmoothed aggregation. We also observe that in this case the impact of smoothed aggregation based on a weighted Jacobi smoother on the convergence behaviour and scalability is not as significant.

Acknowledgements The work of the author "P. D'Ambra" was partially supported by GNCS-INdAM. The work of the author "P.S. Vassilevski" was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

1. Baker, A.H., Kolev, Tz.V., Yang, U.M.: Improving algebraic multigrid interpolation operators for linear elasticity. Numer. Linear Algebra Appl. **17**, 495–517 (2010)
2. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005)
3. Bramble, J.H., Pasciak, J.E.: A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. Math. Comput. **50**, 1–17 (1988)
4. Brandt, A., McCormick, S.F., Ruge, J.W.: Algebraic multigrid (AMG) for sparse matrix equations. In: Evans, D.J. (ed.) Sparsity and Its Applications, pp. 257–284. Cambridge University Press, Cambridge, (1984)
5. Brandt, A.: General highly accurate algebraic coarsening. Electron. Trans. Numer. Anal. **10**, 1–20 (2000)
6. Brandt, A., Brannick, J., Kahl, K., Livshitz, I., Bootstrap AMG. SIAM J. Sci. Comput. **33**, 612–632 (2011)

7. Brezina, M., Falgout, R.D., MacLachlan, S., Manteuffel, T., McCormick, S., Ruge, J.: Adaptive smoothed aggregation αSA multigrid. SIAM Rev. **47**, 317–346 (2005)
8. Brezina, M., Falgout, R.D., MacLachlan, S., Manteuffel, T., McCormick, S.F., Ruge J.: Adaptive algebraic multigrid. SIAM J. Sci. Comput. **27**, 1261–1286 (2006)
9. Briggs W.L., Henson V.E., McCormick S.F.: A Multigrid Tutorial, 2nd edn. SIAM, Philadelphia (2000)
10. Christie, M.A., Blunt, M.J.: Tenth SPE comparative solution project: a comparison of upscaling techniques. SPE Reserv. Eng. Eval. **4**, 308–317 (2001)
11. D'Ambra, P., Vassilevski, P.S.: Adaptive AMG with coarsening based on compatible weighted matching. Comput. Vis. Sci. **16**, 59–76 (2013)
12. Duff, I.S., Koster, J.: On algorithms for permuting large entries to the diagonal of a sparse matrix. SIAM J. Matrix Anal. Appl. **22**, 973–996 (2001)
13. Preis, R.: Linear time 1/2-approximation algorithm for maximum weighted matching in general graphs. In: STACS'99, LNCS, vol. 1563, pp. 259–269. Springer, Berlin (1999)
14. Ruge, J.W.: AMG for problems of elasticity. Applied Mathematics and Computation. Elsevier, New York (1986)
15. Ruge, J.W., Stüben, K.: An introduction to algebraic multigrid (AMG). In: McCormick, S.F. (ed.) Multigrid Methods. Frontiers in Applied Mathematics, vol. 3, pp. 73–130. SIAM, Philadelphia (1987)
16. Vassilevski, P.S.: Multilevel block factorization preconditioners. Matrix-based Analysis and Algorithms for Solving Finite Element Equations. Springer, New York (2008)

A Mathematical Model of the Ripening of Cheddar Cheese

Winston L. Sweatman, Steven Psaltis, Steven Dargaville, and Alistair Fitt

Abstract Cheddar cheese undergoes a number of biochemical changes during ripening. These processes were modelled with differential equations in a project at MISG2013 (the 2013 mathematics-in-industry study group) at Queensland University of Technology, Australia. Models could aid in the prediction of cheese quality from initial measurements. The model is presented and the effect of small changes in initial conditions is explored.

Keywords Biochemical process • Cheese ripening model

1 Introduction

The mathematical model for cheese ripening described here was developed during the 2013 Mathematics-in-Industry Study Group at QUT in Australia. A full description is given in the report [4]. The project was brought by the Fonterra Co-operative Group. A predictive model would be helpful for adjusting factory processes following pre-ripening measurements. The current model is summarised here and its behaviour is illustrated as some initial conditions are varied.

The particular focus is on cheddar-type cheeses. These are produced as 20 kg blocks in a process lasting a few hours. The blocks are then ripened in temperature controlled (cool) storage for a period of months. The three principal milk constituents of cheese (casein, milk fat and lactose) are responsible for generating a

W.L. Sweatman (✉)
Centre for Mathematics in Industry, Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand
e-mail: w.sweatman@massey.ac.nz

S. Psaltis • S. Dargaville
School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD, Australia
e-mail: steven.psaltis@qut.edu.au; dargaville.steven@gmail.com

A. Fitt
Oxford Brookes University, Oxford, UK
e-mail: afitt@brookes.ac.uk

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_143

number of flavour compounds during cheese ripening [3]. These compounds include peptides, amino acids and fatty acids that are accounted for in the current model.

During the ripening there is an initial phase of about 2–4 weeks in which bacteria are active. This is followed by a longer phase in which chemical reactions take place catalysed by enzymes with minimal bacterial activity. These phases have been modelled separately in an earlier study [1]. The present model instead combines the phases into a coherent whole with consistent parameters between the different phases.

2 The Cheese-Ripening Model

Three key processes in the production of cheese are included in the model [4]. These are the breakdown of sugar (lactose), protein (casein) and fat (milk fat). The variables involved are listed in Table 1 together with illustrative initial values. Other symbols represent constant parameters (Table 2). It should be remembered that the reactions, variables and parameters in this model are representative of the whole process which is rather more complex.

Bacteria are crucial to the process. As a part of their life cycle they consume lactose and produce lactic acid, as described in (1)–(3).

$$\frac{dX}{dt} = \frac{\mu_m L X}{K_\ell + L} - k_\ell X, \tag{1}$$

$$\frac{dL}{dt} = \left(-\frac{\mu_m}{Y_x} - \mu_{LA} \right) \frac{LX}{K_\ell + L}, \tag{2}$$

$$\frac{d\alpha}{dt} = \kappa_1 \frac{\mu_{LA} L X}{K_\ell + L}, \tag{3}$$

Table 1 Variables and initial conditions

Symbol	Description	Value	Units
X	Bacterial cells	3.692×10^9	cfu g^{-1}
L	Lactose	15.366	mg g^{-1}
α	Lactic acid	1	mg g^{-1}
E_1	Proteinase	$3.692 \times 10^9 \times e_{1,0}$	U g^{-1}
E_2^o	Extracellular dipeptidase	0.023	U g^{-1}
A	Casein	258	mg g^{-1}
C	Amino acids	2.07059	mg g^{-1}
B	Dipeptides	24.7013	mg g^{-1}
E_L	Extracellular lipase	1	U g^{-1}
T	Triglycerides	9.5	mg g^{-1}
F	Fatty acids	0.5	mg g^{-1}
t	Time	0	days

Bacteria also produce enzymes for the breakdown of protein and fat. The protein breakdown is modelled through (4)–(8). The enzymes proteinase E_1 and extracellular dipeptidase E_2^o are produced by the bacterial action.

$$\frac{dE_1}{dt} = \frac{\rho_1 \mu_m L X}{K_\ell + L} - k_1 E_1, \tag{4}$$

$$\frac{dE_2^o}{dt} = \mu k_\ell X - k_2 E_2^o, \tag{5}$$

$$\frac{dA}{dt} = -\frac{V_f E_1 A}{K_A + A}, \tag{6}$$

$$\frac{dC}{dt} = \frac{V_b E_2^o B}{K_B + B}, \tag{7}$$

$$\frac{dB}{dt} = -\zeta \frac{dA}{dt} - \frac{1}{\zeta}\frac{dC}{dt}. \tag{8}$$

The differential equations representing the breakdown of fat are similar to those for the breakdown of protein. Lipase E_L is produced by the bacteria and converts triglycerides T into fatty acids F.

$$\frac{dE_L}{dt} = \mu_L k_\ell X - k_3 E_L, \tag{9}$$

$$\frac{dT}{dt} = -\frac{V_T E_L T}{K_T + T}, \tag{10}$$

$$\frac{dF}{dt} = -\kappa_2 \frac{dT}{dt}. \tag{11}$$

The parameters in the equations, and listed in Table 2, were obtained in various ways. Several come from values garnered from the literature by Kim et al. [1]. Others were numerically fitted to experimental data presented by Kim et al. [1]. These experimental data are divided into two phases: that when bacteria are active and that when they are not. The parameters for the fat evolution were fitted to data given by Marsili [2]. The constant of proportionality κ_1 was not fitted, and the value of initial proteinase activity $e_{1,0}$ (U cfu^{-1}) is not required.

Table 2 Parameter values

Symbol	Description	Value	Units
K_ℓ	Reaction constant	4.2322×10^4	mg g^{-1}
k_ℓ	Cell death rate	0.2388	day^{-1}
μ_m	Cell reaction constant	4.7295×10^3	day^{-1}
Y_x	Lactose yield constant	1.04×10^9	cfu mg^{-1}
μ_{LA}	Lactic acid reaction constant	3.692×10^{-7}	mg cfu^{-1} day^{-1}
ρ_1	Proteinase reaction constant	$1.2972 \times e_{1,0}$	U cfu^{-1}
k_1	Extracellular dipeptidase destruction rate	0.005	day^{-1}
μ	Dipeptidase reaction constant	0.5151	U cfu^{-1}
k_2	Proteinase destruction rate	0.0235	day^{-1}
V_f	Casein reaction constant	$1.9752 \times 10^{-11} \times (e_{1,0})^{-1}$	mg U^{-1} day^{-1}
K_A	Casein reaction rate constant	0.207	mg g^{-1}
V_b	Amino acids reaction constant	9.4449×10^{-12}	mg U^{-1} day^{-1}
K_B	Amino acid reaction rate constant	1.15	mg g^{-1}
ζ	Scaling constant for proteolysis reactions	1.08	Dimensionless
μ_L	Lipase reaction constant	2.2119×10^{-4}	U cfu^{-1}
k_3	Lipid destruction rate	0.00256	day^{-1}
V_T	Triglyceride reaction constant	2.864×10^{-9}	mg U^{-1} day^{-1}
K_T	Triglyceride reaction constant	1.5537	mg g^{-1}
κ_2	Constant of proportionality relating triglycerides and fatty acids	1	Dimensionless

3 Coupling and Variation with Initial Conditions

The evolution of the bacterial cells and lactose influences that of the other variables (3)–(11). However, there is no back-coupling and so these quantities may be found independently once the values for bacterial cells and lactose are known.

The equations for bacterial cells and lactose (1)–(2) are coupled. From (2), we note that, assuming lactose is present, the quantity of bacterial cells can be expressed in terms of the lactose present and its derivative (which is negative),

$$X = -\frac{Y_x(K_\ell + L)}{(\mu_m + \mu_{LA}Y_x)L}\frac{dL}{dt}, \tag{12}$$

and, further, the evolution of lactose can be described by a second-order non-linear equation

$$\frac{d^2L}{dt^2} + \left(k_l - \frac{\mu_m L}{K_l + L} - \frac{K_l}{(K_l + L)\,L}\frac{dL}{dt}\right)\frac{dL}{dt} = 0. \tag{13}$$

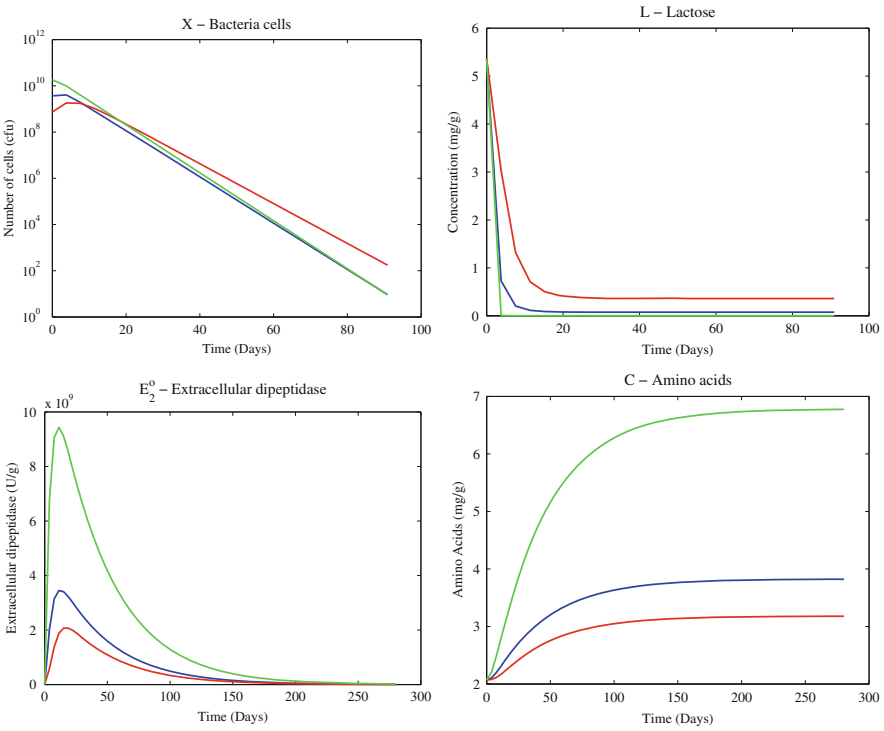


Fig. 1 Effect of varying initial bacteria population

We now explore the effects of small changes in the initial conditions. Figure 1 shows the original results (blue) and those for which the initial quantity of bacteria was increased by a factor of five (green) and decreased by a factor of five (red). The quantities of bacteria are quite similar after 15 days. Thereafter the decline in bacteria is marginally slower for the lower initial quantity as it does not exhaust the initial supply of lactose. Increasing the quantity of bacteria produces more enzymes to break down protein and fat, E_1, E_2^o and E_L, with consequent increases in these reaction rates.

The amount of casein (A) varies naturally in cow milk during the year. However, varying the initial quantity by plus or minus 10 mg/g the effect is only noticeable in casein levels (Fig. 2). Most quantities' evolution is not affected by casein levels and further A is so much larger than K_A that its rate of change is essentially constant.

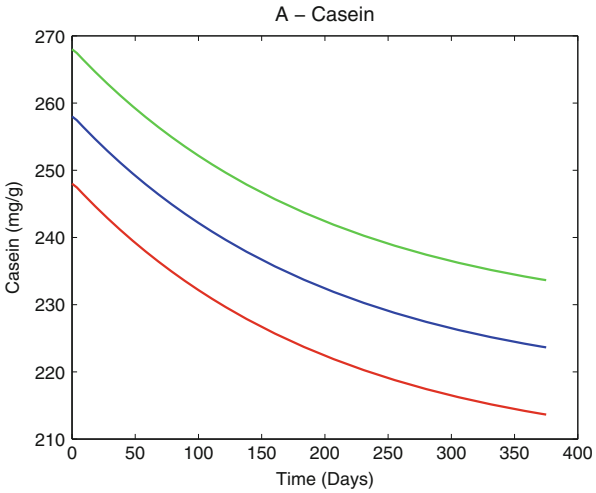


Fig. 2 Evolution of casein for different initial values. The three graphs shown are for $A = 248, 258$ and 268 mg/g

4 Discussion and Conclusions

We have presented a model for processes of biochemical change during the ripening of cheese. This was developed at the Australian study group MISG2013 and a fuller account of this modelling process together with further investigations is given elsewhere [4]. Here we have further commented on the coupling of the equations and investigated the effect of some changes in initial conditions. Additional data will enable more accurate fitting of model parameters and further verification of the model.

Acknowledgements We are grateful to Fonterra Cooperative Group for providing this project for MISG2013 and to the Industry representative Joanne Simpson for her valuable input. We thank the director for the MISG Troy Farrell and his team for the hospitality at QUT. We also acknowledge the other team members for their contribution to the MISG project: Matthew Adams, Pamela Burrage, Elliot Carr, Eamon Conway, Mark Flegg, Tony Gibb, Peter van Heijster, Michael Jackson, Brodie Lawson Sama Low Choy, Nurul Syaza Abdul Latif, Andrew Macfarlane, Louise Manitzky, Kaye Marion, Phil Watson, Bill Whiten and Andy Wilkins.

References

1. Kim, J.K., Starzak, K., Preckshot, G.W., Marshall, R., Bajpai, R.K.: Critical reactions in ripening of cheeses - a kinetic analysis. Appl. Biochem. Biotechnol. **45**, pp. 51–68 (1994)
2. Marsili, R.: Monitoring chemical changes in cheddar cheese during ageing by high performance liquid chromatography and gas chromatography techniques. J. Dairy Sci. **68**, 3155–3161 (1985)

3. Singh, T.K., Drake, M.A., Cadwallader, K.R.: Flavor of cheddar cheese: a chemical and sensory perspective. Compr. Rev. Food Sci. Food Saf. **2**, 166–189 (2003)
4. Sweatman, W.L., Psaltis, S., Dargaville, S., Fitt, A., Gibb, T., Lawson, B., Marion, K.: The mathematical modelling of cheese ripening. ANZIAM J. **55**, M1–M38 (2014)

An Alternative Stochastic Volatility Model

Youssef El-Khatib and Abdulnasser Hatemi-J

Abstract Stochastic volatility modelling is of fundamental importance in financial risk management. Among the most popular existing models in the literature are the Heston and the CEV stochastic models. Each of these models has some advantages that the other one lacks. For example, the CEV model and the Heston model have different relative properties concerning the leverage as well as the smile effects. In this work we deal with the hybrid stochastic volatility model that is based on the CEV and the Heston models combined. This alternative model is expected to perform better than any of the two previously mentioned models in terms of dealing with both the leverage and the smile effects. We deal with the pricing and hedging problems for European options. We first find the set of equivalent martingale measures (E.M.M.). The market is found to be incomplete within this framework since there are infinitely many of E.M.M. We then find the targeted E.M.M. by minimizing the entropy. Using Ito calculus and risk-neutral method enable us to find the partial differential equation (P.D.E.) corresponding to the option price. Moreover, we use Clark-Ocone formula to obtain a hedging strategy that minimizes the distance between the payoff and the value of the hedged portfolio at the maturity. This hedging strategy is among the most efficient available strategies.

Keywords Financial risk management • Option pricing and hedging • Stochastic volatility model

Y. El-Khatib (✉)
Department of Mathematical Sciences, UAE University, P.O. Box 15551, Al-Ain, United Arab Emirates
e-mail: Youssef_Elkhatib@uaeu.ac.ae

A. Hatemi-J
Department of Economics and Finance, UAE University, P.O. Box 17555, Al-Ain, United Arab Emirates
e-mail: Ahatemi@uaeu.ac.ae

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_144

1 Introduction

The classical Black and Scholes model (see [2]) is used regularly for the evaluation of options. However, this model suffers from several deficiencies among other the so called smile effect as well as the leverage effect. A well-known approach that for improving the Black Scholes model is to incorporate jumps in the stochastic process. The literature contains quite large number of research work on this issue, we can cite for instance [4, 5, 12]. In [6, 7] we can find new types of stochastic volatility models where the main objective is to try to capture the impact of the financial crises. In [1] the author suggests a model that combines stochastic volatility and jumps. In addition, the stochastic volatility models are considered useful tools for taking into account the smile phenomena and to some extent the leverage effect. One of the most popular stochastic volatility model is the Heston model [11]. Another useful model within this context is the Constant Elasticity Variance (CEV) model developed by Cox [3], which is also widely used by practitioners to capture the leverage effect. This paper suggests a combined[1] Heston-CEV model, which is expected to sustain the advantages of each model while reducing their weaknesses.

The remaining part of the paper is organized as follows. Section 2 presents the model. Section 3 deals with the pricing of European options within this new context together with the underlying hedging strategy. The last section concludes the paper.

2 The Model

Assume that the probability space is $(\Omega, \mathscr{F}, P)$. Assume also that $(W_t)_{t\in[0,T]}$ and $(B_t)_{t\in[0,T]}$ are two Brownian motion processes such that $d\langle W_t, B_t\rangle = \varrho dt$ and $|\varrho| < 1$. We also consider the filtration $(\mathscr{F}_t)_{t\in[0,T]}$ to be the natural filtration generated by W and B. The market is consisting of two assets: a risky asset $S = (S_t)_{t\in[0,T]}$ to which is related an European call option and a riskless one given by

$$dA_t = r_t A_t dt, \quad t \in [0, T], \quad A_0 = 1, \tag{1}$$

where r_t is a deterministic measure of time varying interest rate. Assume that the data generating process for the stock price at time t, denoted by S_t, is the following stochastic differential equation:

$$dS_t = \mu_t S_t dt + \sigma S_t^{\alpha} \sqrt{Y_t} dW_t, \tag{2}$$

$$dY_t = \nu(\theta - Y_t)dt + b\sqrt{Y_t} dB_t \tag{3}$$

[1] The combined Heston-CEV model has independently been investigated by others see for example [9].

where $t \in [0, T]$ and $S_0 = x > 0$. The parameters σ, α, ν, θ and b are all constant numbers and μ_t is a deterministic function. Note that σ is related to the volatility of the underlying asset, α is the elasticity of the underlying asset variance.

2.1 Change of Probability and Equivalent Martingale Measures

In order to insure the no arbitrage condition and according to the first fundamental theorem of asset pricing we need to move to a new probabilistic environment where the probability is a P-Equivalent Martingale Measure (P-EMM). It is well-known that if Q is a P-equivalent probability then by the Radon-Nikodym theorem there exists a $\mathscr{F}_T$-measurable random variable, ρ_T such that $Q(A) = E_P[\rho_T 1_A]$, $A \in \mathscr{P}(\Omega)$. Notice that ρ_T is strictly positive P-a.s, since Q is equivalent to P and $E_P[\rho_T] = E_P[\rho_T 1_\Omega] = 1$. It is common to use the notation $\rho_T := \frac{dQ}{dP}$. Consider now the P-martingale $\rho = (\rho_t)_{t\in[0,T]}$ defined by

$$\rho_t := E_P[\rho_T \mid \mathscr{F}_t] = E_P\left[\frac{dQ}{dP} \mid \mathscr{F}_t\right].$$

The next proposition gives the Radon-Nikodym density of an EMM with respect to P.

Proposition 1 *Let Q be a P-EMM. The Radon-Nikodym density of Q with respect to P is given by*

$$\rho_T = \exp\left(\int_0^T (\beta_t dW_t + \gamma_t dB_t) - \frac{1}{2}\int_0^T (\beta_t^2 + \gamma_t^2 + 2\varrho\beta_t\gamma_t)dt\right) \tag{4}$$

where $(\beta_t)_{t\in[0,T]}$ and $(\gamma_t)_{t\in[0,T]}$ are two predictable processes. Moreover β_t and γ_t are related by

$$\mu_t - r_t + \sigma S_t^{\alpha-1}\sqrt{Y_t}(\beta_t + \varrho\gamma_t) = 0. \tag{5}$$

Proof A complete proof is available on request. □

The previous proposition leads to the following corollary.

Corollary 1 *The market of the model (1)–(3) is incomplete.*

Proof A complete proof is available on request. □

We just saw from the previous proposition that there is an infinite number of P-EMM. We find the P-EMM that minimizes the relative entropy because this will

minimize the Kullback-Leibler distance within these settings (see for instance [10, 13]). Our aim is to minimize

$$I(Q^\gamma, P) = E_P\left[\frac{dQ^\gamma}{dP}\ln\frac{dQ^\gamma}{dP}\right] = E_P\left[\rho_T^\gamma \ln \rho_T^\gamma\right], \tag{6}$$

over all the P-EMM. The following proposition gives the P-EMM that minimizes the relative entropy.

Proposition 2 *Let* $\hat{\gamma} = 0$ *and* $\hat{\beta} = \frac{r_t - \mu_t}{\sigma S_t^{\alpha-1}\sqrt{Y_t}}$. *The P-EMM* $\hat{Q}$ *defined by its Radon-Nikodym density*

$$e_T = \exp\left(\int_0^T \frac{r_t - \mu_t}{\sigma S_t^{\alpha-1}\sqrt{Y_t}} dW_t - \frac{1}{2}\int_0^T \left(\frac{r_t - \mu_t}{\sigma S_t^{\alpha-1}\sqrt{Y_t}}\right)^2 dt\right),$$

minimizes the relative entropy.

Proof Since we deal with continuous stochastic processes we can apply Theorem 1 of [14] which shows that the reverse relative entropy $I(P, Q^\gamma) = E_{Q^\gamma}\left[\frac{dP}{dQ^\gamma}\ln\frac{dP}{dQ^\gamma}\right]$ can be used instead of the relative entropy given by (6). We have

$$\beta_t^2 + 2\varrho\beta\gamma_t = (\beta_t + \varrho\gamma_t)^2 - \varrho\gamma_t^2 = \left(\frac{\mu_t - r_t}{\sigma S_t^{\alpha-1}\sqrt{Y_t}}\right)^2 - \varrho\gamma_t^2.$$

Thus,

$$I(P, Q^\gamma) = E_P\left[\frac{1}{2}\int_0^T \left(\gamma_t^2(1-\varrho) + \left(\frac{\mu_t - r_t}{\sigma S_t^{\alpha-1}\sqrt{Y_t}}\right)^2\right) dt\right].$$

Therefore, we need to minimize the following function:

$$f(x) = \frac{1}{2}\left((1-\varrho)x^2 + \left(\frac{\mu_t - r_t}{\sigma S_t^{\alpha-1}\sqrt{Y_t}}\right)^2\right).$$

Note that since $|\varrho| < 1$, $f(x)$ has an absolute minimum at $x = 0$. This ends the proof. □

3 Pricing and Hedging

In this section we find the PDE of the option price as well as a hedging strategy that minimizes the variance. In a complete market, one is interested in finding a strategy that leads to a portfolio value that is equal to the payoff at maturity. In

an incomplete model, this type of strategies are not available, thus the question is which one is the best. The answer will depend on in which sense the strategy is better. Here we define the best strategy to be the one that minimizes the distance between the payoff and the value of underlying portfolio (for more details about this approach one can refer to [8]). From now on, we work with $\hat{Q}$ i.e. the P-EMM that is minimizing the entropy given by $\hat{\beta}$ from Proposition 2. In the previous section, we found the probability measures that insure the market is arbitrage free. Thus, we need to express our model under the new probability space. For this purpose, we define the following:

$$\hat{W}_t = W_t - \int_0^t \hat{\beta}_s ds = W_t - \int_0^t \frac{r_s - \mu_s}{\sigma S_s^{\alpha-1}\sqrt{Y_s}} ds, \quad t \in [0,T].$$

By using the Girsanov theorem $\hat{W}$ is a $\hat{Q}$-Brownian motion. Moreover, under $\hat{Q}$, $(S_t)_{t\in[0,T]}$ satisfies

$$\begin{aligned} dS_t &= r_t S_t dt + \sigma S_t^{\alpha}\sqrt{Y_t} d\hat{W}_t, \quad t \in [0,T], \\ dY_t &= \nu(\theta - Y_t)dt + b\sqrt{Y_t} dB_t, \quad t \in [0,T]. \end{aligned}$$

Next we find the price of the option using the PDE approach.

3.1 Option Price PDE

The following proposition gives the PDE of the option price for our model.

Proposition 3 *The price of an European call option with maturity T on a stock with price $(S_t)_{t\in[0,T]}$ defined by the model (1), (2) and (3) and with strike K can be written at maturity as $(C := C(t, S_t, Y_t))_{t\in[0,T]}$ and it satisfies the following PDE.*

$$C_t + r_t x C_x + \nu(\theta - y)C_y + \frac{1}{2}\sigma^2 x^{2\alpha} y C_{xx} + \frac{1}{2} b^2 y C_{yy} + b\sigma\varrho x^{\alpha} y C_{xy} - r_t C = 0, \qquad (7)$$

with the terminal condition $C(T, S_T, Y_T) = h(S_T) := (S_T - K)^+$.

Proof By Itô Lemma, we obtain $dC = L_t dt + \sigma S_t^{\alpha}\sqrt{Y_t} C_x d\hat{W}_t + b\sqrt{Y_t} C_y dB_t$, where

$$L_t := C_t + r_t S_t C_x + \nu(\theta - Y_t)C_y + \frac{1}{2}\sigma^2 S_t^{2\alpha} Y_t C_{xx}. + \frac{1}{2} b^2 Y_t C_{yy} + b\sigma\varrho S_t^{\alpha} Y_t C_{xy}.$$

Since, $\left(e^{-\int_0^t r_s ds} C\right)_{t\in[0,T]}$ is a $\hat{Q}$-martingale then $L_t = r_t C$ which gives (7). Complete proof is available on request. □

In the next section we deal with the hedging problem.

3.2 Hedging

Let η_t and ζ_t denote the number of units invested at time t in the risky and risk-less assets respectively. Thus the value V_t of the portfolio at time t is given by

$$V_t = \zeta_t A_t + \eta_t S_t, \quad t \in [0, T].$$

Assuming that the portfolio is self-financing, we can state the following.

Proposition 4 *The payoff $h(S_T) = (S_T - K)^+$ is not marketable (attainable). However if $D^{\hat{W}}$ and D^B are the Malliavin derivatives with respect to $\hat{W}$ and B respectively. Then, we have*

$$\begin{aligned} C_0 &= E_{\hat{Q}}[(S_T - K)^+] e^{-\int_0^T r_s ds}, \\ C_x &= \sigma^{-1} S_t^{-\alpha} Y_t^{-\frac{1}{2}} E_{\hat{Q}}[D_t^{\hat{W}} (S_T - K)^+ \mid \mathscr{F}_t] e^{-\int_t^T r_s ds}, \\ C_y &= b^{-1} Y_t^{-\frac{1}{2}} E_{\hat{Q}}[D_t^B (S_T - K)^+ \mid \mathscr{F}_t] e^{-\int_t^T r_s ds}. \end{aligned} \tag{8}$$

Proof We use the expansion of $d(e^{-\int_0^t r_s ds} C)$ and the following equalities

$$V_T = V_0 e^{\int_0^T r_t dt} + \int_0^T e^{\int_t^T r_s ds} \eta_t \sigma S_t^{\alpha} \sqrt{Y_t} d\hat{W}_t, \tag{9}$$

$$h(S_T) = E_{\hat{Q}}[h(S_T)] + \int_0^T E_{\hat{Q}}[D_t^{\hat{W}} h(S_T) \mid \mathscr{F}_t] d\hat{W}_t + E_{\hat{Q}}[D_t^B h(S_T) \mid \mathscr{F}_t] dB_t. \tag{10}$$

A more detailed proof is available on request. □

The previous proposition is in alignment with the market incompleteness. Since the payoff is not attainable, we search in this case for a portfolio that leads to a value that is the closest to $h(S_T)$. We need to determine in which sense the closeness should be defined. In this paper, we choose to find the hedging strategy that leads to a portfolio that minimizes the distance between the value of the portfolio at maturity V_T and the payoff $h(S_T)$. The next proposition gives the strategy that minimizes the variance $E_{\hat{Q}}\left[(h(S_T) - V_T)^2\right]$.

Proposition 5 *The strategy minimizing $E_{\hat{Q}}\left[(h(S_T) - V_T)^2\right]$ is given by*

$$\hat{\eta}_t = \sigma^{-1} S_t^{-\alpha} Y_t^{-\frac{1}{2}} E[D_t^{\hat{W}} (S_T - K)^+ \mid \mathscr{F}_t] e^{-\int_t^T r_s ds} = C_x. \tag{11}$$

Moreover the distance between the payoff and value of the portfolio at maturity is in this case given by

$$E_{\hat{Q}}\left[(h(S_T) - \hat{V}_T)^2\right] = \int_0^T (E_{\hat{Q}}[D_t^B h(S_T) \mid \hat{\mathscr{F}}_t])^2 dt.$$

Proof By comparing (9) and (10) we obtain

$$E_{\hat{Q}}\left[(h(S_T)-\hat{V}_T)^2\right] = E_{\hat{Q}}\left[\left(\int_0^T E_{\hat{Q}}[D_t^B f(S_T) \mid \hat{\mathscr{F}}_t]dB_t\right)^2\right]$$
$$+E_{\hat{Q}}\left[\left(\int_0^T \left(E[D_t^{\hat{W}} f(S_T) \mid \hat{\mathscr{F}}_t] - e^{\int_t^T r_s ds}\hat{\eta}_t \sigma S_t^{\alpha}\sqrt{Y_t}\right) d\hat{W}_t\right)^2\right]$$
$$= E_{\hat{Q}}\left[\int_0^T g(\hat{\eta}_t)dt\right],$$

where

$$g(x) = (E_{\hat{Q}}[D_t^B f(S_T) \mid \hat{\mathscr{F}}_t])^2 + \left(E_{\hat{Q}}[D_t^{\hat{W}} f(S_T) \mid \hat{\mathscr{F}}_t] - e^{\int_t^T r_s ds} x\sigma S_t^{\alpha}\sqrt{Y_t}\right)^2.$$

The minimum is reached at $g'(x) = 0$. Therefore, the strategy that minimizes the variance is given by (11). The second part of the equality is obtained from (8). This ends the proof. □

4 Conclusions

In this work, an alternative stochastic volatility model has been considered. It combines the CEV and heston models. The combined model is more consistent with the reality than the CEV or the Heston model separatively. The pricing and hedging problems for the considered model have been investigated. After providing the Radon-Nikodym density for an arbitrary equivalent martingale measure, we show that the market is incomplete. Within this situation, the PDE of the option price for a European call option was derived under the minimal entropy martingale measure. Using the Malliavin calculus and the Clark-Ocone formula, the strategy that minimizes the variance was also obtained.

Acknowledgements The authors are indebted to an anonymous referee for constructive comments that resulted in improving the paper. However, usual disclaimer applies. The second author would like to acknowledge the funding support from the UAE University via the UPAR program.

References

1. Bates, D.S.: Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. Rev. Financ. Stud. **9**, 69–107 (1996)
2. Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–654 (1973)

3. Cox, J.: Notes on option pricing I: constant elasticity of diffusions. Unpublished Draft, Stanford University (1975)
4. El-Khatib, Y., Al-Mdallal, Q.M.: Numerical simulations for the pricing of options in jump diffusion markets. Arab J. Math. Sci. **18**(2), 199–208 (2012)
5. El-Khatib, Y., Hatemi-J, A.: On the calculation of price sensitivities with a jump-diffusion structure. J. Stat. Appl. Probab. **1**(3), 171–182 (2012)
6. El-Khatib, Y., Hatemi-J, A.: Computations of price sensitivities after a financial market crash. Electr. Eng. Intell. Syst. **6**(2), 239–248 (2013)
7. El-Khatib, Y., Hajji, M.A., Al-Refai, M.: Options pricing in jump diffusion markets during financial crisis. Appl. Math. Inf. Sci. **7**(6) (2013)
8. Föllmer, H.; Sondermann, D.: Hedging of non-redundant contingent claims. In: Hildenbrand, W., Mas-Colell, A. (eds.) Contributions to Mathematical Economics, pp. 205–223. North-Holland, Amsterdam (1986)
9. Forde, M., Pogudin, A.: The large maturity smile for the SABR and CEV-Heston model. Int. J. Theor. Appl. Finan. **16**(8), 1350047 (2014)
10. Frittelli, M.: The minimal entropy martingale measure and the valuation problem in incomplete markets. Math. Financ. **10**(1), 39–52 (2000)
11. Heston, S.: A closed-form solutions for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)
12. Merton, R.C.: Option pricing when underlying stock returns are discontinuous. J. Financ. Econ. **3**, 125–44 (1976)
13. Schweizer M.: On the minimal martingale measure and the Föllmer-Schweizer decomposition, Stoch. Anal. Appl. **13**, 573–599 (1995)
14. Schweizer, M. : A minimality property of the minimal martingale measure, Stat. Probab. Lett. **42**, 27–31 (1999)

A Nonlinear CVFE Scheme for an Anisotropic Degenerate Nonlinear Keller-Segel Model

Clément Cancès, Moustafa Ibrahim, and Mazen Saad

Abstract In this paper, we consider a nonlinear Control Volume Finite Element (CVFE) scheme to solve an anisotropic degenerate Keller-Segel model over general meshes. This scheme, whose construction is based on the Godunov scheme to approximate the degenerate diffusion fluxes provided by the conforming finite element reconstruction on a primal triangular mesh and on a nonclassical upwind finite volume mesh to approximate the other terms over a dual mesh, ensures the discrete maximum principle whatever the anisotropy of the problem and without any restriction on the transmissibility coefficients. Numerical experiment is provided with full anisotropic and heterogeneous diffusion tensors over general mesh.

Keywords Control volume finite element scheme • Keller-Segel model

1 The Anisotropic Degenerate Keller-Segel Model

Let Ω be an open bounded polygonal and connected subset of $\mathbb{R}^2$, and let $t_{\mathrm{f}} > 0$ be a fixed finite time. We are interested in the modified degenerate Keller-Segel system [6] modeling the chemotaxis process given by the set of parabolic equations

$$\begin{cases} \partial_t u - \operatorname{div}\left(\Lambda(\mathbf{x})\, a(u)\, \nabla u - \Lambda(\mathbf{x})\, \chi(u)\, \nabla v\right) = f(u) & \text{in } Q_{t_{\mathrm{f}}} = \Omega \times (0, t_{\mathrm{f}}), \\ \partial_t v - \operatorname{div}\left(D(\mathbf{x})\, \nabla v\right) = g(u, v) & \text{in } Q_{t_{\mathrm{f}}} = \Omega \times (0, t_{\mathrm{f}}), \\ \left(\Lambda(\mathbf{x})\, a(u)\, \nabla u - \Lambda(\mathbf{x})\, \chi(u)\, \nabla v\right) \cdot \mathbf{n} = 0 & \text{on } \Sigma_{t_{\mathrm{f}}} = \partial\Omega \times (0, t_{\mathrm{f}}), \\ D(\mathbf{x})\, \nabla v \cdot \mathbf{n} = 0 & \text{on } \Sigma_{t_{\mathrm{f}}} = \partial\Omega \times (0, t_{\mathrm{f}}), \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad v(\mathbf{x}, 0) = v_0(\mathbf{x}) & \text{in } \Omega. \end{cases} \tag{1}$$

C. Cancès
Laboratoire Jacques-Louis, UMR 7598, CNRS, UPMC Univ. Paris 6, F-75005 Paris, France
e-mail: cances@ljll.math.upmc.fr

M. Ibrahim (✉) • M. Saad
Laboratoire de mathématiques Jean Leray, UMR 6629 CNRS, École Centrale de Nantes, F-44321 Nantes, France
e-mail: moustafa.ibrahim@ec-nantes.fr; mazen.saad@ec-nantes.fr

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_145

In the above model, the density of the cell-population and the chemoattractant concentration are represented by $u = u(\mathbf{x}, t)$ and $v = v(\mathbf{x}, t)$ respectively. Next, $a(u)$ is a density-dependent diffusion coefficient, and $\Lambda(\mathbf{x})$ is the diffusion tensor in a heterogeneous medium. Furthermore, the function $\chi(u)$ is the chemoattractant sensitivity, and $D(\mathbf{x})$ is the diffusion tensor for v. The function $f(u)$ describes cell proliferation and cell death. The function $g(u, v)$ describes the rates of production and degradation of the chemoattractant; here, we assume it is the linear function given by

$$g(u, v) = \alpha u - \beta v, \qquad \alpha, \beta \geq 0. \tag{2}$$

We give the main assumptions made about the system:

(A1) The cell-density diffusion $a : [0, 1] \longrightarrow \mathbb{R}^+$ is a continuous function such that, $a(0) = a(1) = 0$, and $a(u) > 0$ for $0 < u < 1$.

(A2) The chemosensitivity $\chi : [0, 1] \longrightarrow \mathbb{R}^+$ is a continuous function such that, $\chi(0) = \chi(1) = 0$. Furthermore, we assume that there exists a function $\mu \in C\left([0, 1]; \mathbb{R}^+\right)$, such that $\mu(u) = \dfrac{\chi(u)}{a(u)}$ and $\mu(0) = \mu(1) = 0$.

(A3) The diffusion tensors Λ and D are two bounded, uniformly positive symmetric tensors on Ω, that is: $\forall \mathbf{w} \neq 0, T_- |\mathbf{w}|^2 \leq \langle T(\mathbf{x})\mathbf{w}, \mathbf{w}\rangle \leq T_+ |\mathbf{w}|^2$, $T = \Lambda$ or D.

(A4) The cell proliferation function $f : [0, 1] \longrightarrow \mathbb{R}$ is a continuous function such that, $f(0) \geq 0$ and $f(1) \leq 0$.

(A5) The initial function u_0 and v_0 are two functions in $L^2(\Omega)$ such that, $0 \leq u_0 \leq 1$ and $v_0 \geq 0$.

In the sequel, we use the Lipschitz continuous nondecreasing function $\xi : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$\xi(u) := \int_0^u \sqrt{a(s)}\, \mathrm{d}s, \qquad \forall u \in \mathbb{R}. \tag{3}$$

and the following set of functions: $\eta(v)$, $p(v)$, $\Gamma(v)$ and $\Phi(v)$ defined in $\mathbb{R}$ by

$$\eta(v) = \max(0, \min(v, 1)), \qquad p(v) = \int_1^v \frac{1}{\eta(s)}\, \mathrm{d}s,$$

$$\Gamma(v) = \int_1^v p(s)\, \mathrm{d}s, \qquad \Phi(v) = \int_0^v \frac{1}{\sqrt{\eta(s)}}\, \mathrm{d}s.$$

The Keller-Segel model (1) has been numerically investigated by many authors. For instance, Andreianov et al. studied in [1] the finite volume scheme for system (1) with isotropic diffusion tensors. Recently, Ibrahim and Saad proposed and analyzed in [5] a CVFE scheme for the case of anisotropic diffusion tensors and under the

assumption that all the transmissibility coefficients are nonnegative. Here we extend the idea given by Cancès and Guichard [2] to the degenerate Keller-Segel system (1) over general meshes. The theoretical results presented in this contribution are detailed in the article version [3].

2 The Nonlinear CVFE Scheme for System (1)

The discretization of system (1) requires the definition of two type of approximations: the finite element approximation over a primal triangular mesh and the finite volume approximation over a dual barycentric mesh (Fig. 1).

Let $\mathscr{T}$ be a conforming triangulation of the domain Ω. We define $h_{\mathscr{T}}$ and $\theta_{\mathscr{T}}$ to be the size and the regularity of the mesh Ω defined by: $h_{\mathscr{T}} := \max_{T\in\mathscr{T}} h_T$ and $\theta_{\mathscr{T}} = \max_{T\in\mathscr{T}} \frac{h_T}{\rho_T}$, where h_T is the diameter of the triangle T and ρ_T is the diameter of the incircle of the triangle T. We denote by $\mathscr{V}$ the set of vertices of the triangulation $\mathscr{T}$ and by $\mathscr{E}$ the set of edges of $\mathscr{T}$. For every vertex $K \in \mathscr{V}$ (located at position $\mathbf{x}_K$), we denote by $\mathscr{E}_K$ the subset of $\mathscr{E}$ consisting of the edges having $\mathbf{x}_K$ as an extremity. An edge joining two vertices K and L is denoted by σ_{KL}. For the construction of the dual barycentric mesh, we denote by T_K the set of all triangles having K as a vertex. There exists a unique dual element ω_K constructed around a vertex $K \in \mathscr{V}$ by connecting the barycenters $\mathbf{x}_T$ of the triangles $T \in \mathscr{T}_K$ with the barycenters $\mathbf{x}_\sigma$ of the edges $\sigma \in \mathscr{E}_K$. We denote by $\mathscr{H}_{\mathscr{T}}$ the usual $\mathbb{P}_1$-finite element space defined by

$$\mathscr{H}_{\mathscr{T}} = \{\phi \in \mathscr{C}^0\left(\overline{\Omega}\right); \phi_{|_T} \in \mathbb{P}_1\left(\mathbb{R}\right), \forall T \in \mathscr{T}\}$$

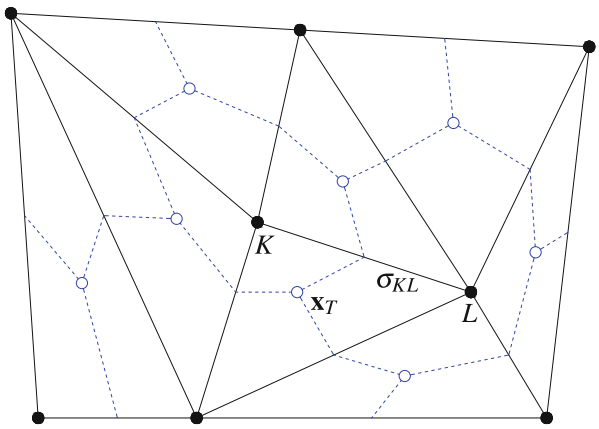


Fig. 1 Triangular mesh $\mathscr{T}$ and Donald dual mesh $\mathscr{M}$: dual volumes, vertices, interfaces

and by $(\varphi_K)_{K\in\mathscr{V}}$ its canonical basis. Furthermore, we consider the discrete control volumes space $\mathscr{X}_{\mathscr{M}}$ defined by

$$\mathscr{X}_{\mathscr{M}} = \{\phi : \Omega \longrightarrow \overline{\mathbb{R}}, \phi_{|\omega_K} \text{ is constant }, \forall K \in \mathscr{V}\}.$$

In this paper, we restrict our study to the case of uniform time discretization with the time step given by $\Delta t = t_{\mathrm{f}}/\,(N+1)$ for a given nonnegative integer N.
We set $t^n = n\Delta t$ for $0 \le n \le N+1$, and introduce the space and time discrete spaces

$$\mathscr{H}_{\mathscr{T},\Delta t} = \{\phi \in L^\infty(Q_{t_{\mathrm{f}}})\,; \phi(\mathbf{x},t) = \phi\left(\mathbf{x}, t^{n+1}\right) \in \mathscr{H}_{\mathscr{T}}, \forall t \in (t^n, t^{n+1}], 0 \le n \le N\},$$

$$\mathscr{X}_{\mathscr{M},\Delta t} = \{\phi \in L^\infty(Q_{t_{\mathrm{f}}})\,; \phi(\mathbf{x},t) = \phi\left(\mathbf{x}, t^{n+1}\right) \in \mathscr{X}_{\mathscr{M}}, \forall t \in (t^n, t^{n+1}], 0 \le n \le N\}.$$

For a given $\left(u_K^n\right)_{n\in\{0,\cdots,N+1\},K\in\mathscr{V}}$ (resp. $\left(v_K^n\right)_{n\in\{0,\cdots,N+1\},K\in\mathscr{V}}$), there exists a unique function $u_{\mathscr{T},\Delta t} \in \mathscr{H}_{\mathscr{T},\Delta t}$ (resp. $v_{\mathscr{T},\Delta t} \in \mathscr{H}_{\mathscr{T},\Delta t}$) and a unique $u_{\mathscr{M},\Delta t} \in \mathscr{X}_{\mathscr{M},\Delta t}$ (resp. $v_{\mathscr{M},\Delta t} \in \mathscr{X}_{\mathscr{M},\Delta t}$) such that

$$u_{\mathscr{T},\Delta t}\left(\mathbf{x}_K, t^{n+1}\right) = u_{\mathscr{M},\Delta t}\left(\mathbf{x}_K, t^{n+1}\right) = u_K^{n+1}, \qquad \forall K \in \mathscr{V},\ \forall n \in \{0,\ldots,N\},$$
$$v_{\mathscr{T},\Delta t}\left(\mathbf{x}_K, t^{n+1}\right) = v_{\mathscr{M},\Delta t}\left(\mathbf{x}_K, t^{n+1}\right) = v_K^{n+1}, \qquad \forall K \in \mathscr{V},\ \forall n \in \{0,\ldots,N\}.$$

Let m_K be the 2-dimensional Lebesgue measure of ω_K for every $K \in \mathscr{V}$. The nonlinear CVFE scheme for the discretization of system (1), is given by the following set of equations: for all $K \in \mathscr{V}$

$$u_{\mathscr{M}}^0(\mathbf{x}) = u_K^0 = \frac{1}{m_K}\int_{\omega_K} u_0(\mathbf{y})\,\mathrm{d}\mathbf{y}, \quad v_{\mathscr{M}}^0(\mathbf{x}) = v_K^0 = \frac{1}{m_K}\int_{\omega_K} v_0(\mathbf{y})\,\mathrm{d}\mathbf{y}. \tag{4}$$

and for all $n \in \{0,\ldots,N\}$

$$\begin{aligned} m_K\frac{u_K^{n+1}-u_K^n}{\Delta t} &+ \sum_{\sigma_{KL}\in\mathscr{E}_K} \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right) \\ &- \sum_{\sigma_{KL}\in\mathscr{E}_K} \Lambda_{KL}\mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right) = m_K f\left(u_K^{n+1}\right), \\ m_K\frac{v_K^{n+1}-v_K^n}{\Delta t} &+ \sum_{\sigma_{KL}\in\mathscr{E}_K} D_{KL}\eta_{KL}^{n+1}\left(p\left(v_K^{n+1}\right) - p\left(v_L^{n+1}\right)\right) = m_K\left(\alpha u_K^n - \beta v_K^{n+1}\right). \end{aligned} \tag{5}$$

In the above system, we have set $T_{KL} = -\int_\Omega T(\mathbf{x})\nabla\varphi_K(\mathbf{x})\cdot\nabla\varphi_L(\mathbf{x})\,\mathrm{d}\mathbf{x} = T_{LK}$, with $T \equiv \Lambda$ or D. Denoting by $I_{KL}^{n+1} = [\min(u_K^{n+1}, u_L^{n+1}), \max(u_K^{n+1}, u_L^{n+1})]$, and by

$J_{KL}^{n+1} = [\min(v_K^{n+1}, v_L^{n+1}), \max(v_K^{n+1}, v_L^{n+1})]$, then a_{KL}^{n+1} and η_{KL}^{n+1} are given by

$$a_{KL}^{n+1} = \begin{cases} \max_{s \in I_{KL}^{n+1}} a(s) & \text{if } \Lambda_{KL} \geq 0, \\ \min_{s \in I_{KL}^{n+1}} a(s) & \text{if } \Lambda_{KL} < 0, \end{cases} \quad \eta_{KL}^{n+1} = \begin{cases} \max_{s \in J_{KL}^{n+1}} \eta(s) & \text{if } D_{KL} \geq 0, \\ \min_{s \in J_{KL}^{n+1}} \eta(s) & \text{if } D_{KL} < 0. \end{cases}$$

Finally, μ_{KL}^{n+1} is set to be equals to

$$\mu_{KL}^{n+1} = \begin{cases} \mu_{\downarrow}\left(u_K^{n+1}\right) + \mu_{\uparrow}\left(u_L^{n+1}\right), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0, \\ \mu_{\uparrow}\left(u_K^{n+1}\right) + \mu_{\downarrow}\left(u_L^{n+1}\right), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) < 0. \end{cases}$$

The functions $\mu_{\uparrow}$ and $\mu_{\downarrow}$ are deduced from the function μ introduced in (A2) and given by

$$\mu_{\uparrow}(z) = \int_0^z \left(\mu'(s)\right)^+ \mathrm{d}s, \qquad \mu_{\downarrow}(z) = -\int_0^z \left(\mu'(s)\right)^- \mathrm{d}s.$$

Note that the scheme (5) is locally conservative on the dual median mesh $\mathscr{M}$. Indeed, denoting by

$$F_{KL}^{n+1} = \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right) - \Lambda_{KL}\mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right)$$

and

$$\Phi_{KL}^{n+1} = D_{KL}\eta_{KL}^{n+1}\left(p\left(v_K^{n+1}\right) - p\left(v_L^{n+1}\right)\right),$$

it follows from the symmetry properties $T_{KL} = T_{LK}$ for $T \equiv \Lambda$ or D, $a_{KL}^{n+1} = a_{LK}^{n+1}$, $\eta_{KL}^{n+1} = \eta_{LK}^{n+1}$, and $\mu_{KL}^{n+1} = \mu_{LK}^{n+1}$ that the following conservation relation holds:

$$F_{KL}^{n+1} + F_{LK}^{n+1} = 0 = \Phi_{KL}^{n+1} + \Phi_{LK}^{n+1}, \qquad \forall \sigma_{KL} \in \mathscr{E}.$$

3 Discrete Estimates, Existence and Convergence of the Scheme

As a major feature of the method we propose, the uniform bounds are preserved at the discrete level. We give the discrete maximum principle

Proposition 1 *For all $K \in \mathscr{V}$, and all $n \in \{0, \ldots, N+1\}$, we have $0 \leq u_K^n \leq 1$ and $v_K^n \geq 0$.*

Sketch of the proof We take a dual element ω_K such that $u_K^{n+1} = \min_{L\in\mathscr{V}}\{u_L^{n+1}\}$, then we multiply the first equation of system (5) by $-u_K^{n+1}$ and sum over $K \in \mathscr{V}$. One obtains that $u_K^{n+1} \geq 0$ by noting that $a_{KL}^{n+1} = 0$ when $\Lambda_{KL} \geq 0$, and that $a_{KL}^{n+1}(\Lambda_{KL})^+ \left(u_K^{n+1} - u_L^{n+1}\right)\left(u_K^{n+1}\right)^- \geq 0$, and $\mu_{KL}^{n+1}\Lambda_{KL}^+\left(v_K^{n+1} - v_L^{n+1}\right)^-\left(u_K^{n+1}\right)^- = 0$. Similarly, we get $u_K^{n+1} \leq 1$ by taking ω_K such that $u_K^{n+1} = \max_{L\in\mathscr{V}}\{u_L^{n+1}\}$.
The proof of the last claim $v_K^n \geq 0$ is a direct consequence of the nonnegativity of u_K^n.

Now we give some discrete properties on the CVFE scheme (4)–(5).

Proposition 2 *For all $n \geq 0$, there exists a constant C independent of h such that*

$$\begin{aligned}&\sum_{K\in\mathscr{V}} m_K \Gamma\left(v_K^{n+1}\right) + \Delta t \sum_{\sigma_{KL}\in\mathscr{E}} D_{KL}\left(\Phi(v_K^{n+1}) - \Phi(v_L^{n+1})\right)^2 \\ &\leq \sum_{K\in\mathscr{V}} m_K \Gamma(v_K^{n+1}) + \Delta t \sum_{\sigma_{KL}\in\mathscr{E}} D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1}) - p(v_L^{n+1})\right)^2 \leq C.\end{aligned} \tag{6}$$

This estimate is obtained by multiplying the second equation of system (5) by $\Delta t p(v_K^{n+1})$, by using the convexity of the function Γ, and by using the definition of the function Φ.

The Proposition 2 as well as the assumption (A3) on Λ and D lead to the following proposition

Proposition 3 *There exists a constant $C > 0$ independent of h such that*

$$\iint_{Q_{t_f}} \Lambda \nabla v_{\mathscr{T},\Delta t} \cdot \nabla v_{\mathscr{T},\Delta t} \mathrm{d}\mathbf{x}\,\mathrm{d}t = \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathscr{E}} \Lambda_{KL}(v_K^{n+1} - v_L^{n+1})^2 \leq C. \tag{7}$$

Using Estimates (6) and (7), we obtain an analogous estimate to the one in Proposition 2; it is given in the following proposition.

Proposition 4 *For all $n \geq 0$, there exists a constant C independent of h such that*

$$\begin{aligned}&\sum_{K\in\mathscr{V}} m_K \left(u_K^{n+1}\right)^2 + \Delta t \sum_{\sigma_{KL}\in\mathscr{E}} \Lambda_{KL}\left(\xi(u_K^{n+1}) - \xi(u_L^{n+1})\right)^2 \\ &\leq \sum_{K\in\mathscr{V}} m_K \left(u_K^{n+1}\right)^2 + \Delta t \sum_{\sigma_{KL}\in\mathscr{E}} \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right)^2 \leq C.\end{aligned} \tag{8}$$

In order to prove the existence of a solution $\left(u_K^{n+1}, v_K^{n+1}\right)_{K\in\mathscr{V}, n\in\{0,\ldots,N\}}$ to the CVFE scheme (4)–(5), we use the previous a priori estimates and the following result ensuring that no component v_K^{n+1} of the discrete solution can go to zero. In what

follows, we denote by $u_{\mathscr{M},\Delta t}$ and $v_{\mathscr{M},\Delta t}$ the unique elements of $\mathscr{X}_{\mathscr{M},\Delta t}$ such that

$$u_{\mathscr{M},\Delta t}(\mathbf{x}_K, t) = u_K^{n+1}, \quad v_{\mathscr{M},\Delta t}(\mathbf{x}_K, t) = v_K^{n+1}, \quad \forall K \in \mathscr{V}, \forall n \geq 0.$$

Proposition 5 *Assume that* $\int_\Omega u_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$ *or* $\int_\Omega v_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$*, then there exists* $r_h > 0$ *depending on the data as well as on the mesh* $\mathscr{T}$ *and* Δt *such that*

$$v_K^{n+1} \geq r_h, \quad \forall K \in \mathscr{V}, \forall n \in \{0, \ldots, N\}.$$

Moreover, there exists a solution $\left(u_K^{n+1}, v_K^{n+1}\right)_{K\in\mathscr{V}}$ *to the scheme* (4)–(5)*.*

The proof of Proposition 5 is similar to the one given in [2], we rely on the topological degree argument to get the existence results.

We are in a position to state the main result on the convergence of the CVFE scheme (4)–(5) towards the weak solution of the continuous system (1) as the time and space discretization steps go to zero. Specifically, we have the following theorem

Theorem 1 *Let* $(\mathscr{T}_m)_{m\geq 1}$ *be a sequence of conforming triangulations of* Ω *such that* $h_{\mathscr{T}_m} \to 0$ *as* $m \to \infty$*, and let* $(\Delta t_m)_{m\geq 1}$ *be a sequence of time steps such that* $\Delta t_m \to 0$ *as* $m \to \infty$*. For all* $q \in [1, \infty)$*, the discrete solution* $(u_{\mathscr{M}_m,\Delta t_m}, v_{\mathscr{M}_m,\Delta t_m})_m$ *converges (up to an unlabeled subsequence) in* $L^q(Q_{t_f})$ *towards the weak solution of the continuous system* (1) *as* $m \to \infty$*.*

Sketch of the proof In order to obtain the convergence result, we follow the technique employed in [4]. Thus, we use the Kolmogorov compactness criterion relying on some estimates on differences of time and space translates of the discrete solutions $(u_{\mathscr{M}_m,\Delta t_m}, v_{\mathscr{M}_m,\Delta t_m})_{m\geq 1}$. Then, we identify the limit solution as a weak solution to the continuous problem (1).

4 Numerical Experiment

In this section, we simulate the chemotaxis process in a 2-D domain. To implement the CVFE scheme (4)–(5), we use the Newton's algorithm coupled with a biconjugate method to solve the nonlinear systems arising from the use of a fully non-linear implicit Euler which is adopted in order to obtain a stable numerical scheme. In the numerical test, we consider a general unstructured mesh $\Omega = (0, 1)^2$ made of 5 193 triangles that contains obtuse angles. Here, we consider a general triangular mesh and not a cartesian mesh in order to show the efficiency of the numerical scheme to tackle anisotropic convection-diffusion problem over a general mesh. We fix: $\Delta t = 0.002$, $\alpha = 0.01$, $\beta = 0.05$, $a(u) = d_u u(1-u)$, $d_u = 0.0005$, $\chi(u) = \zeta \times (u(1-u))^2$, $\zeta = 0.05$, and $f \equiv 0$. By definition, we have $\mu(u) = \frac{\zeta}{d_u} u(1-u)$

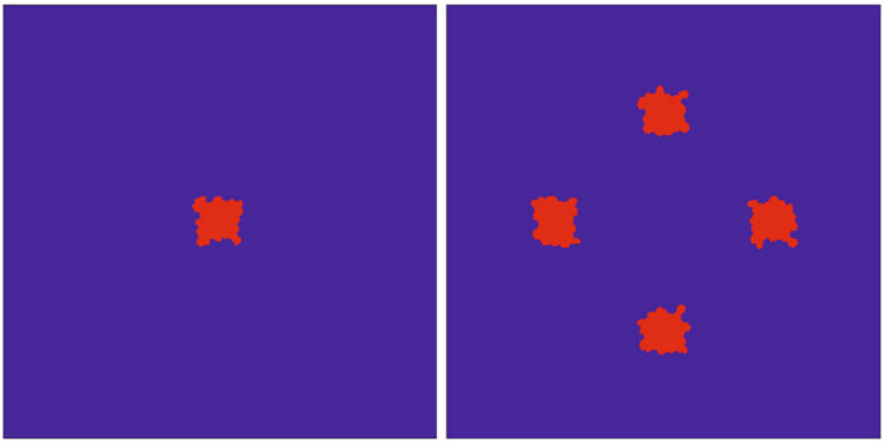


Fig. 2 Initial condition for the cell density u (*left*) with $0 \leq u \leq 1$ and for the chemoattractant concentration v (*right*) with $0 \leq v \leq 5$

then, the numerical flux function μ_{KL}^{n+1} is given using the following functions:

$$\mu_{\uparrow}(z) = \mu\left(\min\{z, \frac{1}{2}\}\right), \text{ and } \mu_{\downarrow}(z) = \mu\left(\max\{z, \frac{1}{2}\}\right) - \mu\left(\frac{1}{2}\right), \quad \forall z \in (0,1)^2.$$

Furthermore, we assume that the initial conditions are defined by regions, and we assume zero-flux boundary conditions. For instance, the cell density is initially defined by $u_0(\mathbf{x}, \mathbf{y}) = 1$ in the square region given by $(\mathbf{x}, \mathbf{y}) \in [0.45, 0.55]$ and 0 otherwise (see Fig. 2). The initial chemoattractant concentration is defined by $v_0(\mathbf{x}, \mathbf{y}) = 5$ in the space region given by $(\mathbf{x}, \mathbf{y}) \in ([0.2, 0.3] \times [0.45, 0.55]) \cup ([0.45, 0.55] \times [0.2, 0.3]) \cup ([0.45, 0.55] \times [0.7, 0.8]) \cup ([0.7, 0.8] \times [0.45, 0.55])$.

The diffusion tensors are defined, for all $\mathbf{x} \in (0,1) \times (0,1)$, by

$$\Lambda(\mathbf{x}) = \begin{pmatrix} 7 & 2 \\ 2 & 10 \end{pmatrix}, \quad D(\mathbf{x}) = d\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \quad d = 0.0001.$$

Figures 3 and 4 represent the evolution of the cell density at time $t = 0.8$, $t = 1.4$, $t = 3$ and the distribution of the chemoattractant at $t = 3$ over the dual barycentric mesh. We observe in practice that the bounds on the solution stated in Propositions 1 and 5 are well respected by the proposed CVFE scheme.

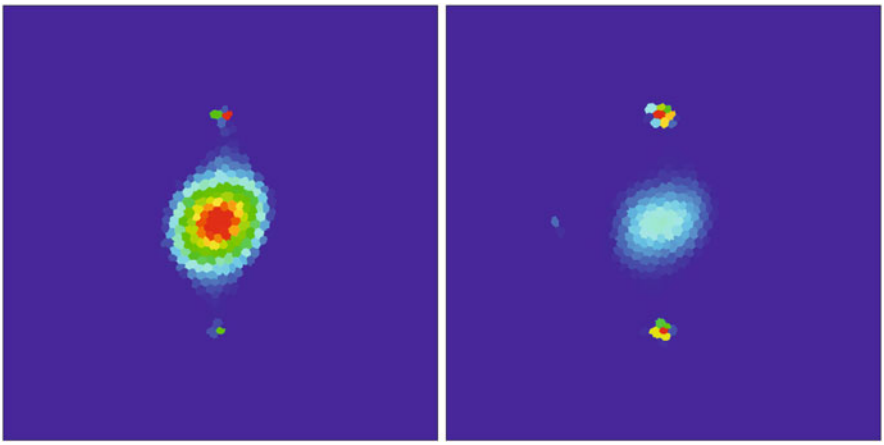


Fig. 3 Evolution of the cell density u at time $t = 0.8$ with $0 \leq u \leq 0.44$ (*left*), and at time $t = 1.4$ with $0 \leq u \leq 0.954$ (*right*)

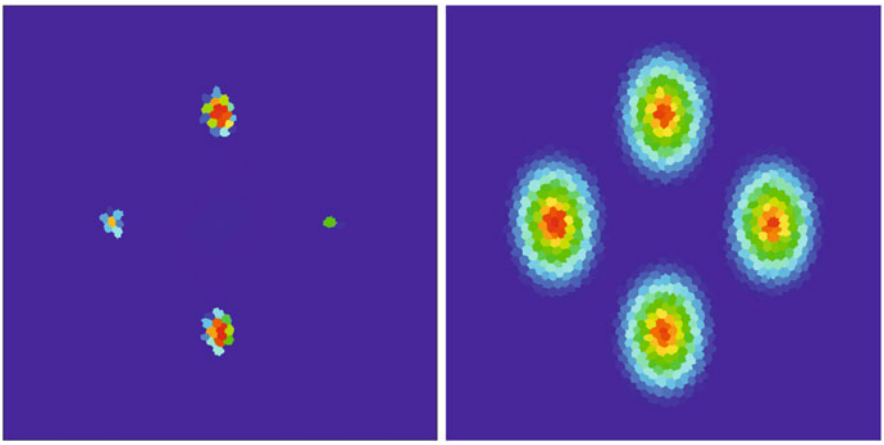


Fig. 4 Evolution of the cell density u at time $t = 3$ with $0 \leq u \leq 0.985$ (*left*), and of the chemoattractant at the same time with $2.37 \times 10^{-10} \leq v \leq 2.192$ (*right*)

References

1. Andreianov, B., Bendahmane, M., Saad, M.: Finite volume methods for degenerate chemotaxis model. J. Comput. Appl. Math. **235**(14), 4015–4031 (2011)
2. Cancès, C., Guichard, C.: Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. Math. Comput. **85**, 549–580 (2016)
3. Cancès, C., Ibrahim, M., Saad, M.: Positive nonlinear CVFE scheme for degenerate anisotropic Keller-Segel system (2015). Preprint, HAL: hal-01119210

4. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handbook of Numerical Analysis, vol. 7, pp. 713–1018. Elsevier, Amsterdam (2000)
5. Ibrahim, M., Saad, M.: On the efficacy of a control volume finite element method for the capture of patterns for a volume-filling chemotaxis model. Comput. Math. Appl. **68**(9), 1032–1051 (2014)
6. Keller, E.F., Segel, L.A.: Model for chemotaxis. J. Theor. Biol. **30**(2), 225–234 (1971)

Combining Traditional Optimization and Modern Machine Learning: A Case in ATM Replenishment Optimization

Harry Raymond Joseph

Abstract ATM Replenishment has become a widely studied popular problem in the modern age of human-machine interaction, due to several reasons. This paper presents a solution that is a two-part system. The first part or the analytics section is capable of providing very highly accurate, hourly forecasts of withdrawals from an ATM, for which past data is available. The Machine Learning algorithm used for obtaining forecasts, M5P, is based on a decision tree approach that reinforces various characteristics empirically found on withdrawal patterns in ATMs. The second and more important section formulates a simple mixed binary, goal programming problem. The weights are decided by the bank at the beginning of each period, and is particularly advantageous in decision flexibility terms. This is done specially keeping in mind the ever-changing operating budgets and customer service goals. In terms of hard numbers, this work describes a system which generates daily schedules with an error in withdrawal forecast per month (non-absolute addition) as low as 0.7 % at a correlation coefficient of 0.92.

Keywords ATM replenishment optimization • Human-machine interaction • Machine learning

1 ATM Replenishment Optimization

ATMs have replaced serpentine queues of earlier days in large banks as customers lined up to withdraw money from their accounts. The first ATMs appeared in the early 90s and have since seen rapid indictment, with almost 90 % of withdrawals being done through ATMs. ATM service has become such an important factor that many banks have spent millions in optimizing and enriching customer experience in this aspect. The importance is reflected in the customer-side as well, with many customers considering ATM service as an important factor before opening an

H.R. Joseph (✉)
Indian Institute of Technology Madras, Chennai, India

TU Munich, Germany
e-mail: raymond.harry@tum.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_146

account with a particular bank. The most studied aspect with regard to ATMs is that of cash-outs [1–3].

Cash-outs are defined as an event occurring at an ATM such that the ATM does not have sufficient cash repository to meet the demand for the next 4 h of operation. This demand for the next 4-h window is calculated through several statistical procedures ranging from simple interpolation to advanced correlation techniques that consider several parameters. Particularly, the availability of data in abundance that provides an excellent avenue of opportunity for forecasting, the increasing spotlight on customer satisfaction—which many banks value highly in the competitive sector that is banking and the mere cost-cutting incentive are aspects which have made this problem both interesting and challenging. This work looks at an all-round solution in terms of tractability, simplicity and flexibility that lies at the marriage of modern machine learning and conventional operations research based optimization [4–6].

2 The Case: Industrial Environment

The following section describes the operating conditions under which the system proposed in this solution is expected to operate and introduces the basic problem statement.

2.1 Operating Environment and Assumptions

Given blow are a set of assumptions that are to be considered in providing the solution.

1. Transport is charged for a 'to-fro' trip from central cash facility to ATM and back to cash facility from ATM to return cartridge.
2. Given any reasonable refilling schedule, the transporter is willing to oblige, with services available at all times—no constraints imposed by the transporter on the scheduling of refills.
3. Some new ATMs do not have a large transaction record database as they are newly inaugurated or due to several other possible changes. Some forms of interpolation or meta-inspired tools have to be developed for these cases in making forecasts and providing solutions.
4. Some ATMs are co-located with others. For this purpose, we would like to define co-location as: two ATMs are co-located if a customer at one ATM can see the other or if the ATMs are located within a 300-m distance from each other.
5. Location data for each ATM is available. Data may also be obtained using a simple web-crawling application on Google Maps.
6. All ATMs don't have similar characteristics—different ATMs have varying full-capacities.

7. ATMs are declared to be cashed-out if the cash in the inventory is forecast to not be able to meet the next 4 h of demand. This is an expected safety buffer for uncertainty in the forecasts of the demand. On visiting an ATM, the truck always fills the ATM to capacity.

2.2 Challenges

A list of challenges that make the problem slightly more difficult are considered below.

1. The banks are unable to provide a quantitative preference between the two conflicting goals to be pursued—customer service goals and transportation austerity measures.
2. As stated earlier, some new ATMs do not have a large transaction record database as they have been recently inaugurated.

2.3 Operating Goals

The goals that must be pursued are:

1. The banks require the solution to make accurate forecasts of refilling schedules.
2. The schedule needs to be optimized to minimize transportation costs charged by the transporter.

3 Solution Description

This section describes in detail the work to be completed, and the theoretical basis behind the various procedures, that are likely to be followed.

3.1 Information Flow: Theoretical Basis: Information Theory and Machine Learning

After a comprehensive survey academic literature on ATM networks, the following qualitative claims have been recorded regarding withdrawal patterns—these claims are important as they provide the theoretical basis for the machine learning process—choice of algorithm and input schema in the relational algebra sense [7, 9].

1. Withdrawals may depend on seasonality—example month of the year, date and day of the month, time of the day and so on.
2. Withdrawals depend on macroeconomic parameters such as Inflation percentage and price level.
3. Withdrawals at an ATM depend on the location of the ATM—crowded with high population density, less sparse population coverage, ATMs on highways, etc.

Almost certainly, one may model withdrawals at any ATM as belonging to one of the two components: (a) Random/Floating Population and (b) Fixed or Periodic Withdrawals. In which case, the withdrawal over a 4-h window, $W_{L,T}$ may be modeled as the sum of two stochastic variables: one the Poisson counter and two the Gaussian Normal distribution, each with mean and variance modified appropriately.

$$W_{L,T} = P(\lambda) + N(\mu, \sigma) \tag{1}$$

In such a model, Machine Learning and Data Analytics methods such as decision trees have proven effective. Since our task is that of predicting a floating point number—the withdrawal, we will use the well-known *M5P* algorithm for predicting the same. The results of which, for predicted withdrawals from the test data set are presented in a later section. Increasing the granularity of the output forecast, collecting more data and data pre-processing are some of the major aspects that need to be worked on.

3.2 The M5P Algorithm

The *M5P* [10] algorithm is a modification over Quinlan's *M5* algorithm for inducing trees for regression models. Given the stochastic splitting presented above in (1) *M5P* appears to be the most optimal algorithm for forecasting ATM withdrawals with reasonable computational expense. The *M5P* algorithm, is based on the well-known decision trees which ensure each splitting reduces the entropy of the split groups through an apparent renormalization. For instance if at the split G there are C classes such that $p(C_i)$ is the probability of rows being split at the split G belonging to the class C_i then, entropy of the data being split at G is defined as:

$$H(G) = \sum p(C_i) Log(1/p(C_i)) \tag{2}$$

Similarly, the information gain I_G at the split G on the attribute A_G is defined as:

$$H(A_G, G) = H(G) - \sum p(i) H(i) \tag{3}$$

M5P combines however several novelties over traditional decision tree based machine learning methods. For instance, the decision-tree induction algorithm used to build a tree, minimizes the intra-subset variation in the class values down each

branch. The splitting procedure in M5P stops if the class values of all instances that reach a node vary very slightly, or only a few instances remain. Second, the tree is pruned back from each leaf. When pruning an inner node is turned into a leaf with a regression plane. Third, to avoid sharp discontinuities between the subtrees a smoothing procedure is applied that combines the leaf model prediction with each node along the path back to the root, smoothing it at each of these nodes by combining it with the value predicted by the linear model for that node [8].

3.3 Traditional Analytics: Goal Programming Based Optimization

The following section describes the analytical components describing the operation research problem: Given that $W_{L,T}$ is the withdrawal forecast for the next 4 h at the Lth ATM at time **T** of the day, then, the fund equation representing the cash dynamics may be modeled as:

$$F_{L,T-1} + R_{L,T} - W_{L,T} - L_{L,T} = F_{L,T} \tag{4}$$

$R_{L,T}$ is the replenishment decision for ATM **L** at time **T** of the day. $F_{L,T-1}$ represents cash remaining in the cartridge after withdrawal in the previous hour, $\mathbf{T-1}$. $W_{L,T}$ represents the demand forecast for the next 4 h, obtained from the above, data analytics section employing the M5P algorithm. $L_{L,T}$ is the left over money in the cartridge in case of replenishment. This money is returned to the bank. $L_{L,T}$ is significant only if the replenishment decision $R_{L,T}$ is non-zero. Now:

$$R_{L,T} = T_L S_{L,T} \tag{5}$$

In the above equation,$\mathbf{T_L}$ is the capacity of the **L**th ATM. This ensures flexibility in modeling and factors the aspect of non-uniform ATM full-capacities. $S_{L,T}$ is one of the variables that appear in the objective function. It is binary and is 1 if the **L**th ATM is replenished at time **T**, else it takes the value 0. SL, T is hence a record of the number of replenishments. Also if replenished, $L_{L,T}$ becomes active—whatever is the cash remaining after withdrawal in the previous hour, it is taken away when the cartridge is replaced and the ATM is brought to full capacity. Hence, $L_{L,T}$ may be defined as:

$$L_{L,T} = S_{L,T} F_{L,T-1} \tag{6}$$

The goal for this mixed binary goal programming problem is:

Minimize: $\sum_{L,T}(|F_{L,T}| + MS_{L,T})$

M is the weight of the second objective: transport austerity. As can be seen from above, the problem is a simple linear programming problem, that can be solved to

resolve values of variables $S_{L,T}$ and $F_{L,T}$ that satisfy the objective function measures. Please note that the system runs the analysis at a granularity of 1 h, and determines the signs and magnitude of the variables for the next 4 h—this ensures the 4-h safety buffer for determining cash-outs.

3.4 Input Data Structure

Input data is sent to the data analytics engine that applies the M5P algorithm and outputs forecasts for hourly demand. The exact schema **S** that will be required to be input for a particular ATM at a given location is as below:

S: (Day, Date, Month, Time, Amount Withdrawn, Daily Price Level, Inflation Percentage)

Note that the daily price level and inflation percentage are of utmost importance, since they dictate the trends in savings and spending in the economy and hence withdrawal patterns. This data is easily available from the relevant regulatory authority, if unavailable with the bank. This data on hourly forecast withdrawals obtained is now fed by the system into the LP solver component along with the Goal weight M. We anticipate that priorities between customer service goals and transport austerity are expected to keep shifting. The system provides a flexible replenishment schedule, allowing for these changes, to reflect whatever objectives the bank has set for that period.

3.5 Output Data Structure

The output data structure provides us with corresponding replenishment schedules and the flow/error variable values. The system is run with the hourly withdrawal forecasts at say the beginning of the month, to forecast all replenishment schedules and corresponding flow elements for the rest of the month. Output data structure is a schema comprising:

O: (Date of Replenishment, Time of Replenishment)

In addition, the system also outputs: Expected Number of Cash-Outs and Number of Replenishment Events. Given below is a diagrammatic representation of the complete process associated with the system (Fig. 1).

4 Special Considerations

This section addresses two special considerations: New ATMs with insufficient transaction history for analytics and Co-Located ATMs.

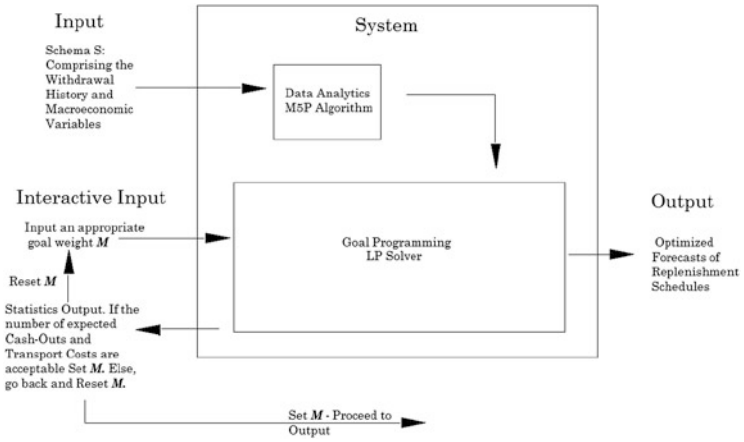


Fig. 1 Top-level representation of the solution

New ATMs The strategy to deal with this challenge is rather straightforward. Analytics learning methods are extended—data analytics methods are employed on a new schema, which consists of the following attributes:

S: (Day, Date, Month, Time, Amount Withdrawn, Daily Price Level, Inflation Percentage, Number of Bank Customers in that Location, Population Density around the ATM)

Note that the above columns have data from all ATM transactions at all Locations. Withdrawals are forecast at the new locations using learning from previous transactions, in addition with two attributes: Number of Bank Customers in that Location and Population Density. It is expected that the former is available with the bank, the latter may be obtained from the competent authority.

Co-Located ATMs These are as defined previously. In these cases, the capacity of the two co-located ATMs together will be added and ensuing analysis is done by adding forecast demands. For this purpose, they maybe visualized as a single ATM with capacity twice as that of a normal ATM. Inclusion of capacity parameter T_C in the modeling facilitates this analysis.

5 Results and Conclusions

In this paper a solution that utilizes traditional optimization methods as well as modern machine learning algorithms to maximize ATM replenishments in banks has been developed. The conventional goal programming mixed binary formulation enables banks to adapt to changing goals in terms of a trade-off between customer satisfaction and operating expenses. The Machine learning forecasting algorithm

Table 1 The following table shows the output obtained from the learning engine for a given month

Date	Number of transactions	Predicted	Observed withdrawals	Error/flow
19	31	3772.09	3441.00	−331.09
13	42	4652.61	4998.00	345.39
21	27	2557.23	2862.00	304.77
7	31	2860.63	3689.00	828.37
23	35	3185.23	3080.00	−105.23
9	31	2860.63	2697.00	−163.63
5	43	4482.33	4859.00	376.67
18	30	2620.96	2700.00	79.04
10	28	2442.20	3052.00	609.80
30	43	3902.50	3655.00	−247.50
25	46	4150.17	3818.00	−332.17
22	34	3159.50	3026.00	−133.50
16	35	3207.12	2975	−232.12
2	45	4020.10	3825	−195.10

seems to perform particularly well, with a correlation coefficient of 0.92 and an error of 0.7 % (Table 1).

Several research avenues seem to open up while considering useful extensions to this problem. For instance, the seasonality in terms of day of the week and month of the year seem to be little emphasized. Perhaps feature-extraction methods in these aspects may considerably improve results. In addition, further research considering a differing stochastic model of withdrawals, might also be an interesting exercise.

References

1. Adams, R., Brevoort, K., Elizabeth, K.K.: Who competes with whom? the case of depository institutions. Working paper 2005–03, Federal Reserve Board (2005)
2. Bernhardt, D., Massoud, N.: Endogenous ATM location and pricing. Working paper (2002a)
3. Calem, P., Carlino, G.: The concentration/conduct retlationship in bank deposit markets. Rev. Econ. Stat. **72**, 268–276 (1991)
4. Gowrisankaran, G., Krainer, J.: The welfare consequences of ATM surcharges: evidence from a structural entry model. Working paper (2004)
5. Hannan, T.: Bank retail fees and multimarket banking. Working paper, Federal Reserve Board (2004)
6. Hannan, T., Prager, R.: The competitive implications of multimarket branching. J. Bank. Finance **28**, 1889–1914 (2004)
7. Isshi, J.: Interconnection pricing and compatibility in network industries: ATM networks in the banking industry. Working paper, Department of Economics, Harvard University (2004)

8. Kramer, S.: In OPenTox documentation. Retrieved from http://www.opentox.org/dev/documentation/components/m5p as of 31 May 2014
9. Matutes, C., Padilla, A.J.: Shared ATM networks and banking competition. Eur. Econ. Rev. **38**, 1113–1138 (1994)
10. Wang, Y., Witten, I.H.: Induction of model trees for predicting continuous classes. In: Poster Papers of the 9th European Conference on Machine Learning (1997)

Detection of Shadow Artifacts in Satellite Imagery Using Digital Elevation Models

Ivan Martynov and Tuomo Kauranne

Abstract There are numerous methods for shadow identification in satellite imagery using variety of means of image processing and pattern recognition. We, however, suggest to detect shadow artifacts using a straight geometrical approach with the help of digital elevation models (DEM). To demonstrate the pipeline of shadow detection process we use Landsat imagery and SRTM DEM database as the sources of images and MATLAB as software to run computational operations. Landsat provides a huge amount of images covering the surface of our planet. The SRTM collection covers nearly all terrain areas, thus excluding oceans, seas and other zero elevation areas. Both databases are widely used for image processing in various projects working with geographical data. The computational tools of MATLAB in their turn help to easily create functions and scripts for image processing in a relatively simple and fast way.

Keywords DEM • Homography • Landsat • Satellite imagery • SRTM

1 Introduction

Shadows appear in nearly every satellite image and, therefore, may hide and obscure important features of a terrain. In various areas of satellite image processing it is important to classify the terrain into segments (forest, mountains, water bodies, etc.) or analyze the terrain for other features. Consequently, detection and removal of shadow artifacts in satellite images is important for image analysis [1, 2].

In this paper we represent an approach for detection of shadow areas using DEM (Digital Elevation Model) as an accessorial tool. Among a number of DEM data we have decided to exploit SRTM (Shuttle Radar Topography Mission) DEMs due to their availability. A DEM image would require latitude and longitude coordinates of image corners.

I. Martynov (✉) • T. Kauranne
Lappeenranta University of Technology, Skinnarilankatu 34, 53850 Lappeenranta, Finland
e-mail: ivan.martynov@lut.fi; tuomo.kauranne@lut.fi

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_147

There are several satellites providing imagery for the Earth. We have chosen Landsat because they provide a great collection of moderate resolution (30x30 m) images for over 40 years and the images are available for everyone due to free access. There are many images for every region of our planet, thus making it possible to choose among them one image or few which would satisfy the needs. For example, choosing an image with less clouds or from a certain date (Landsat provide data from 1972). Otherwise, it is possible to use collections of other satellites. The image would require the Sun position relatively to the Earth (azimuth and elevation angles) and, likewise for a DEM data, latitude and longitude coordinates of image corners. These are the necessary parameters in order to compute where the shadows are cast and to match the satellite image with the DEM data.

Given the solar angles we can use the DEM image to compute coordinates of shadow areas in a specific region. Then homography is applied in order to convert pixel coordinates into latitude and longitude values since they are unique for both Landsat and SRTM images. The latitude and longitude coordinates are then converted into pixel values using homography for the Landsat image in order to locate the shadow artifacts in the image.

We have chosen to use MATLAB as a computational tool since they provide convenient environment for a relatively fast coding of functions and scripts for image processing.

2 Shadow Detection Pipeline

Assume we have chosen a region and acquired Landsat images for this region (7–8 images for Landsat 5–7 or 10–11 images for Landsat 8). In order to begin shadow calculation we need corresponding SRTM files and the solar angles relatively to the Earth. The angles can be extracted from meta data which comes in a separate text file together with the Landsat images. Usually, the file looks like `Landsat_code_MTL.txt` using the `MTL` part as a special identifier for the meta file. From the same file latitude and longitude values for corner points are fetched.

Using the latitude and longitude values we can choose SRTM DEM files which would cover one Landsat image (tile) entirely. The size of one SRTM image is 1° degree latitude by 1° longitude, which makes roughly 110 km by 100 km depending on the location. The latitude value does not change a lot, but the longitude kilometers lessen significantly when coming towards one of the Poles. As can be seen in Fig. 1 an SRTM image covers only a part of the Landsat image (due to the difference in latitude and longitude coordinates). The size of a Landsat image is about 170 km by 185 km and thus nine SRTM images are required to fully cover the Landsat image.

We can acquire the SRTM DEM data using, for example, Earth Resources Observation and Science (EROS) Center. After we have acquired all necessary files we need to compute shadow areas for every DEM image. Apparently, the steps would be the same for every image.

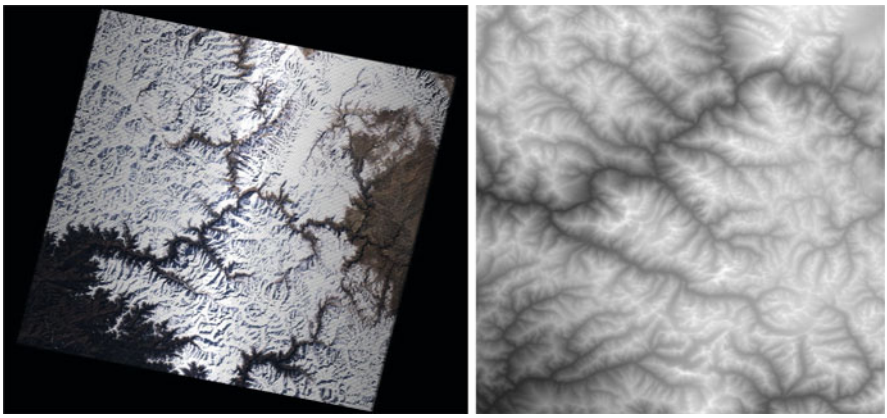


Fig. 1 Landsat and SRTM DEM images

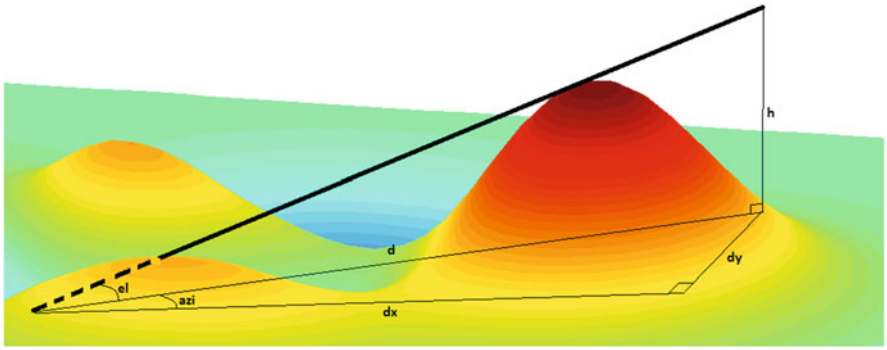


Fig. 2 Solar ray geometry

A DEM image pixels represent heights of a terrain. Firstly, we decide to define where shadows fall by drawing an imaginary solar ray going through one of the peaks until the ray reaches the ground (see Fig. 2). Knowing the height (denoted as h in Fig. 2) at the certain peak point, the solar elevation angle (el) and the point on the ground we can calculate the length of a potential shadow (d). The height value is taken from the DEM file, the elevation angle has been acquired earlier and the ground point is considered to have either the zero height or the lowest meaningful value in the DEM image. In case of a non-zero ground point the height should be accordingly adjusted (height minus the lowest value). Let us assume that the ground point has height zero for simplicity.

The next step is to calculate shifts in x and y directions in order to obtain coordinates of the ground point. For this purpose we use the solar azimuth (azi). Since we know the length of the potential shadow then we can apply simple trigonometry formulae to calculate the shifts (dx and dy). The shifts allow us to parametrize the line with one end in the peak and the other in the ground. Since the

peak point has known coordinates (x_0, y_0, z_0) then the ground point has coordinates $(x_0 + dx, y_0 + dy, 0)$. Therefore, the line would be defined as shown in (1):

$$\frac{x - x_0}{dx} = \frac{y - y_0}{dy} = \frac{z - z_0}{dz} = t. \tag{1}$$

Afterwards, we calculate the coordinates of all pixels which the shadow line is to cross and we name them potential shadow pixels. Sometimes the solar ray would not reach the ground because of an obstacle (a hill or another mountain peak) as it can be seen in Fig. 2. Therefore, we check every potential shadow pixel starting from the closest to the chosen peak but excluding the peak itself (because it is never in shadow). We continue towards the ground pixel until an obstacle is met or the ground is reached. The check is performed using the parameter t from (1) which can be calculated using the x or y coordinate of the current pixel and then used to calculate the z height value of the line. Then the z value is compared with the actual height value in the DEM image to determine whether the solar ray has been blocked or not.

After all true shadow pixels have been identified we can move to the next peak. In order to simplify and speed up the process of shadow detection we ignore peaks (pixels) which are already in shadow. Besides, we sort all peaks in the descending order because the highest peaks assume to cast the longest shadows and, therefore, more pixels could be omitted in the following steps. The algorithm continues until all pixels have been processed and then creates a binary image with true values assigned to shadow areas (see Fig. 3).

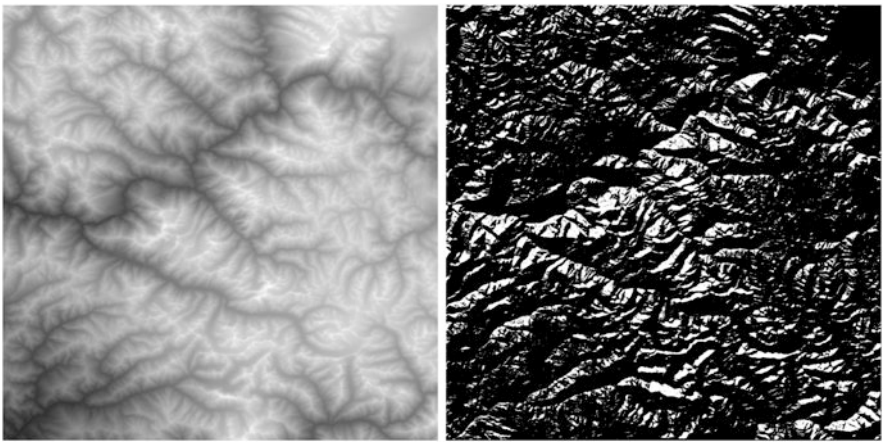


Fig. 3 Shadow mask computed for an SRTM DEM image

The next maneuver would be to convert the pixel coordinates into latitude and longitude coordinates. For this purpose we apply a plane-to-plane homography. Apparently, we may easily acquire the corner coordinates of the DEM image. The question is how to calculate the corresponding geographic coordinates. These values can be extracted from the SRTM file name due to its construction. The file name is formed as follows `N28E086.hgt` where the first and second letters define north, south, west or east directions (`N`, `S`, `E`, `W`). The numbers following the letters show the latitude and longitude values of the southwest corner correspondingly. In this example the image covers latitudes 28–29 North and longitudes 86–87 East.

When the geographic coordinates have been acquired we can compute a homography matrix. The matrix is applied to convert the earlier computed shadow pixel coordinates into latitude and longitude values. Since the geographic coordinates are unique we can then convert the latitude and longitude values into pixel coordinates using the homography technique. In order to compute a homography matrix for the Landsat image we repeat the same steps: acquire the corner pixel coordinates, extract the latitude and longitude values for corners from the meta data file and then compute the homography matrix. Then we use the matrix to calculate the pixel coordinates using the geographic coordinates. Since the resolution of SRTM images is 90×90 m and the resolution of Landsat images is 30×30 m the Landsat shadow mask is morphologically dilated using the square of size 3×3 pixels as a structured element. Exception applies if an SRTM image covers an area in USA: the resolution is the same as for the Landsat image—30×30 m. The final shadow pixel are shown in Fig. 4 as red areas.

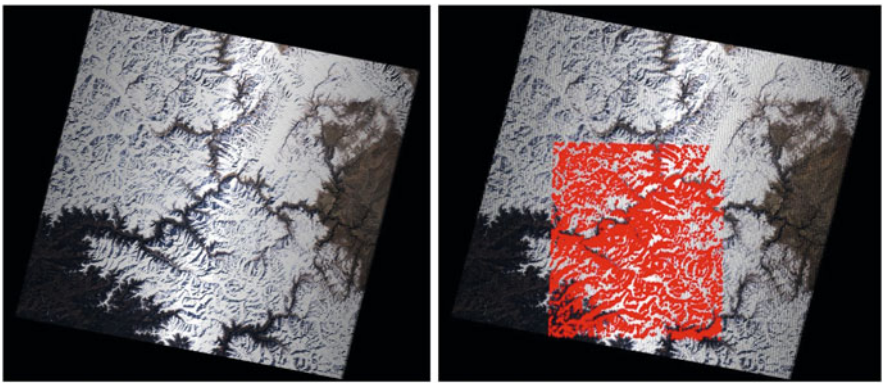


Fig. 4 Shadow mask over the Landsat image

3 Discussions

There are few known problems concerning our method of shadow detection.

1. The resolution of SRTM DEM images is coarser than the resolution of Landsat (as well as several other satellites) images. This means that small shadows as well as the borders of bigger shadow areas might be easily left unidentified. DEM images with a better resolution would definitely improve the accuracy;
2. An SRTM DEM image may have void zones thus there is no meaningful data. Usually the pixel values in these zones are set to be very low (for example, $-2^{15}+1$). One way to deal with this problem is to ignore these zones and consider that there are no shadow artifacts inside void areas. Another way is to interpolate the missing data using an appropriate algorithm;
3. Clouds add shadows which can not be detected in DEM images. Clouds as well can cover areas (mountain slopes, hills) which might be identified as shadow artifacts in the DEM images. An example of a shadow mask covering an entire Landsat image having clouds is shown in Fig. 5. In order to solve the cloud issue, we can use one tool called Fmask which is able to identify clouds and their shadows in a Landsat image rather well [3].

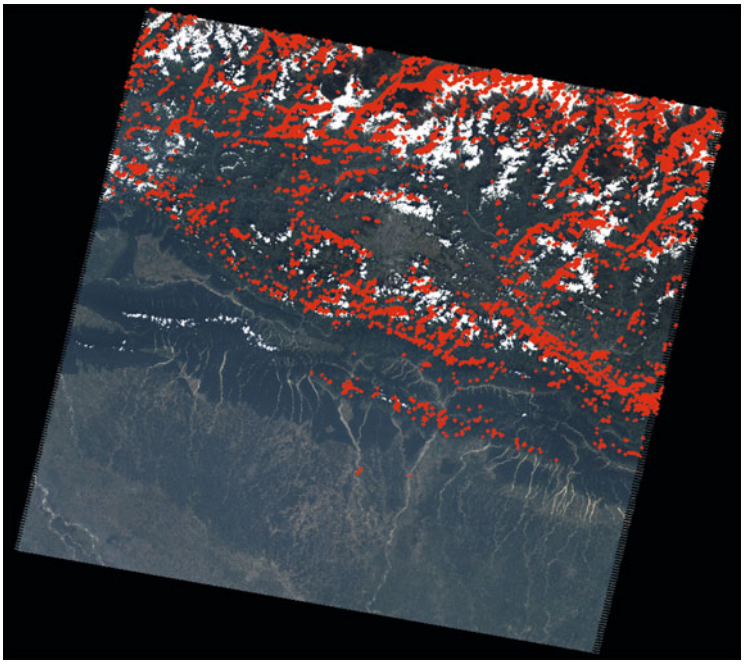


Fig. 5 Landsat image with clouds and its shadow mask

References

1. Makarau, A., Richter, R., Müller, R., Reinartz, P.: Adaptive shadow detection using a blackbody radiator model. IEEE Trans. Geosci. Remote Sens. **49**(6), 2049–2059 (2011)
2. Mazzoni, D., Horváth, Á., Garay, M.J., Tang, B., Davies, R.: A MISR cloud-type classifier using reduced support vector machines. In: Eighth Workshop on Mining Scientific and Engineering Datasets. SIAM International Conference on Data Mining (2005)
3. Zhu, Z., Woodcock, C.E.: Object-based cloud and cloud shadow detection in landsat imagery, remote sensing of environment. Remote Sens. Environ. **118**, 83–94 (2012)

Efficient Numerical Simulation of the Wilson Flow in Lattice QCD

Michèle Wandelt and Michael Günther

Abstract Lattice Quantum Chrome Dynamics (Lattice QCD) is a gauge theory formulated on a highly dimensional grid or lattice of points in space and time. It aims at determining observables such as the mass of elementary particles as accurate as possible, with computational costs as low as possible at the same time. Thus high performance computing tools are inevitable, as well as the construction of HPSC hardware tailored to the needs of Lattice QCD. In the Hybrid Monte Carlo (HMC) approach (Duane et al., Phys. Lett. B, 195:216, 1987 http://dx.doi.org/10.1016/0370-2693(87)91197-X), Monte Carlo simulations involving a molecular dynamics step in its core are performed, which yield physical values provided with their statistical errors.

In this talk we concentrate on the Wilson Flow, a system of differential equations defined on the Lie group $SU(3)$. The Wilson Flow can be used, e.g., to determine the physical lattice spacing which influences the result of the HMC simulations. We focus on tailored Runge-Kutta Lie group integration methods with step size prediction. The numerical results confirm that our strategy is able to reduce the statistical errors of the simulation.

Keywords Lattice quantum chrome dynamics • Wilson flow

1 Introduction

Quantum Chromo Dynamics (QCD) is a quantum field theory that describes the strong interaction between fundamental constituents of matter inside subatomic particles. The discretized version of QCD is formulated on a 4-dimensional grid—or lattice—in space and time and called Lattice QCD. It aims at the computation of observables like the mass of elementary particles which is theoretically done via the computation of path integrals. Due to the fact that these integrals are very high dimensional, this calculation is done using Monte Carlo simulations [1]. During

M. Wandelt (✉) • M. Günther
Lehrstuhl für Angewandte Mathematik/Numerische Analysis, Fachbereich C - Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany
e-mail: wandelt@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_148

these simulations, a sequence $[U]^0 \rightarrow [U]^1 \rightarrow [U]^2 \rightarrow \ldots$ of field configurations is computed which consists of a set of matrices being elements of the special unitary matrix Lie group $SU(3)$. Moreover, the observables are determined as expectation values of certain operators of the different configurations. As a byproduct of the Monte Carlo simulation, the so-called Wilson flow can be computed, see [2, 3]. It is a flow in the field space and can be used to investigate certain physical properties of the lattice as, for example, the physical lattice spacing. The Wilson flow is defined by a system of differential equations of the kind

$$\dot{V}(t) = Z([V(t)]) \cdot V(t) \; . \tag{1}$$

Since the variables $V(t)$ are elements of the matrix Lie group $SU(3)$ and the variables $Z([V(t)])$ elements of the appropriate matrix Lie algebra $\mathfrak{su}(3)$ we have a differential equation on the manifold $SU(3)$. This means, the solution has to be also in the Lie group. Thus, we have to choose a numerical method that ensures a solution in the Lie group like, for example, Munthe-Kaas Runge-Kutta (RK-MK) methods. Usually, the Wilson Flow is computed via Runge-Kutta methods for Lie groups of fixed convergence order.

In this paper, we concentrate on the numerical integration of the Wilson flow using step size prediction. In Sect. 2, we start with a brief explanation of RK-MK schemes for differential equations of type (1). Then, we focus on step size prediction for RK-MK schemes in Sect. 3. Afterwards, we show the numerical results for a RK-MK scheme of convergence order (2)3 in Sect. 4. Here, we compute the Wilson flow and investigate the so-called Wilson energy as observable. Then, we adapt the step size prediction for the whole set of variables of a field configuration. Finally, we show some simulation results.

2 Runge Kutta Methods for Lie Groups

In the Wilson flow, a differential equation on a Lie group which is a differentiable manifold has to be solved. This differential equation has a special structure:

$$\dot{V} = Z \cdot V \tag{2}$$

with V being an element of a Lie group G and Z an element of the Lie algebra $\mathfrak{g}$. This kind of differential equation can be solved using the theorem of Magnus [4]. That means, the unknown Lie group element V can be replaced by a mapping

$$V = \exp(\Omega) V_0 \tag{3}$$

with unknown Lie algebra element Ω. Then, Ω is the solution of the differential equation

$$\dot{\Omega} = d \exp_{\Omega}^{-1}(Z) \tag{4}$$

in the Lie algebra with $\dot{\Omega}$ being the derivative of the inverse exponential mapping and initial value $\Omega(0) = 0$. At the end, the solution of (2) is given as the mapping (3) of the solution Ω of (4). Thereby, the derivative of the inverse exponential map can be rewritten as infinite series

$$d\exp_{\Omega}^{-1}(Z) = \sum_{k=0}^{\infty} \frac{B_k}{k!} ad_{\Omega}^{k}(Z)$$

with B_k being the kth Bernoulli number and adjoint operator ad_{Ω}^{k} which is a mapping

$$ad_{\Omega}(A) := [\Omega, A] = \Omega \cdot A - A \cdot \Omega$$

in the Lie algebra $\mathfrak{g}$. For the numerical simulation, the infinite series of the derivative of the inverse exponential map has to be truncated. This truncation induces a model error which should be smaller or equal than the convergence order of the numerical method used for the detection of the solution Ω.

Munthe-Kaas describes a suitable truncation for Runge-Kutta methods [5, 6] as follows: for a Runge-Kutta method of convergence order p, the truncation index q has to be larger than $p - 2$:

$$\dot{\Omega} = \sum_{k=0}^{q} \frac{B_k}{k!} ad_{\Omega}^{k}(Z) \quad , q \geq p - 2 \tag{5}$$

A Runge-Kutta method for the differential equation (2) can be computed in three steps: Start with the mapping $V = \exp(\Omega)V_0$. Then, use an appropriate numerical integration scheme to solve the differential equation

$$\dot{\Omega} = d\exp_{\Omega}^{-1}(Z) = \sum_{k=0}^{q} \frac{B_k}{k!} ad_{\Omega}^{k}(Z)$$

with initial value $\Omega(0) = 0$, e.g. the Munthe-Kaas Runge Kutta scheme. Finally, map the solution Ω via Eq. (3) from the Lie algebra to the Lie group.

The RK-MK for the computation of the solution of the differential equation (4) is given as:

$$\Omega_1 = h \sum_i b_i K_i \quad \text{with} \quad K_i = f_q(Y_i, Z_i)$$

$$Y_i = h \sum_k a_{ik} K_k, \qquad Z_i = Z(V_i), \qquad V_i = \exp(Y_i) \cdot V_0$$

Here, $f_q(Y_i, Z_i)$ is described by (5) as

$$f_q(Y_i, Z_i) = B_0 \cdot Z + B_1 \cdot [\Omega, Z] + \frac{B_2}{2} \cdot \big[\Omega, [\Omega, Z]\big] + \ldots + \frac{B_q}{q!} ad_{\Omega}^{q}(Z)$$

This means, the function f_q has to be suitably truncated according to the desired convergence order of the method. For example, a convergence order $p = 2$ can be achieved using $f_0 = B_0 \cdot Z$, and for $p = 3$ we would need $f_1 = B_0 \cdot Z + B_1 \cdot [\Omega, Z]$.

3 Step Size Control

Our aim is to solve the equation $\dot{V} = Z \cdot V$ using a step size control. Here, we use a common step size control as, for example, described in [7] and combine it with a Munthe-Kaas Runge-Kutta method. We proceed as follows: Start from initial values V_0 with a given step size and compute the solutions $\hat{V}_1$ of convergence order p and V_1 of convergence order $p+1$. Here, the RK-MK method is adapted for the step size control: start with $V_1 = \exp(\Omega_1) \cdot V_0$ and $\hat{V}_1 = \exp(\hat{\Omega}_1) \cdot V_0$. To reach the desired convergence order p of $\hat{V}_1$ and $p+1$ of V_1, the RK-MK algorithm is given as

$$\begin{aligned}
\Omega_1 &= h \sum_i b_i K_i \quad \text{with} \quad K_i = f_{p-2}(Y_i, Z_i) \\
Y_i &= h \sum_j a_{ij} K_j, \qquad Z_i = Z(V_i), \qquad V_i = \exp(Y_i) \cdot V_0 \\
\hat{\Omega}_1 &= h \sum_i \hat{b}_i \hat{K}_i \quad \text{with} \quad \hat{K}_i = f_{p-1}(\hat{Y}_i, \hat{Z}_i) \\
\hat{Y}_i &= h \sum_j a_{ij} \hat{K}_j, \qquad \hat{Z}_i = Z(\hat{V}_i), \qquad \hat{V}_i = \exp(\hat{Y}_i) \cdot V_0 \ .
\end{aligned}$$

The measure for the error is calculated as

$$\text{err} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \Big(\frac{||\hat{\Omega}_1 - \Omega_1||_j}{ATOL + RTOL \cdot ||\hat{\Omega}_1||_j} \Big)^2} \ .$$

As we work on a set of matrix Lie algebra elements, the computation of the error measure has to be adapted to a set of Lie algebra elements: The norms $||\hat{\Omega}_1 - \Omega_1||_j$ and $||\hat{\Omega}_1||_j$ have to be chosen as matrix norms like, for example, the Frobenius norm, row sum norm or the spectral norm. Afterwards, the optimal step size h_{opt} is computed as

$$h_{\text{opt}} = h \cdot \sqrt[p+1]{\frac{1}{\text{err}}} \cdot \rho$$

with safety factor ρ. Additionally, the step size should not increase or decrease too fast which is prevented by

$$h_{\text{opt}} = \min\Big(\alpha \cdot h, \max\big(\beta \cdot h, h_{opt} \big) \Big) \ .$$

If the error measure is small enough, i.e. err ≤ 1, the step is accepted and V_1 taken as initial value for the new step, otherwise the step is repeated. In any case, the new step size is set to h_{opt}.

Remark 1 (Step Size Control for the Wilson Flow) The Wilson Flow is a flow in the field space, i.e. for a lattice of n variables there are n differential equations

$$\dot{V}_j(t) = Z([V(t)]) \cdot V_j(t) \quad j = 1, \ldots, n$$

to be solved. The calculation for one Wilson Flow starts at one of the given configurations, e.g. $[U]^i$ which serves as initial values $[V]^0$. Here, we have to refresh our mind with the fact that the variables $V_j(t)$, $j = 1, \ldots, n$ are elements of the special unitary Lie group $SU(3)$. The function $Z([V(t)])$ maps an element $V_j \in SU(3)$ to its appropriate special unitary Lie algebra $\mathfrak{su}(3)$:

$$V_j \to Z_j = Z([V]) \quad , \qquad SU(3) \to \mathfrak{su}(3).$$

Thereby, the function Z does not just depend on V_j itself but of several adjacent variables V_k (considered to be constants at this moment). This dependence is induced by the notation $Z([V])$. Considering the elements $V_j \in SU(3), Z_j \in \mathfrak{su}(3)$ we have a system of the aforementioned differential equations on Lie groups with solution being as well in the Lie group.

4 Numerical Results

We compute the Wilson flow for one single configuration that consists of n lattice points. Then, we measure the Wilson energy

$$E = \sum_p \text{Real Trace}(\mathbf{1} - U(p)) \tag{6}$$

whose formula is, for example, described in [2]. We have implemented a Munthe-Kaas Runge-Kutta method of convergence order (2)3 with Bogacki-Shampine coefficients given in Table 1. This means, we have a Runge-Kutta method of four stages and have to compute

$$\Omega_1 = h \sum_{i=1}^{k=3} b_i K_i \quad \text{with} \quad K_i = f_1(Y_i, Z_i) = f_0(Y_i, Z_i) + B_1[Y_i, Z_i]$$

$$\hat{\Omega}_1 = h \sum_{i=1}^{k=4} \hat{b}_i \hat{K}_i \quad \text{with} \quad \hat{K}_i = f_0(Y_i, Z_i)$$

Table 1 Bogacki-Shampine coefficients

0					
1/2	1/2				
3/4	0	3/4			
1	2/9	1/3	4/9		
	2/9	1/ 3	4/9	0	$\leftarrow b$
	7/24	1/4	1/3	1/8	$\leftarrow \hat{b}$

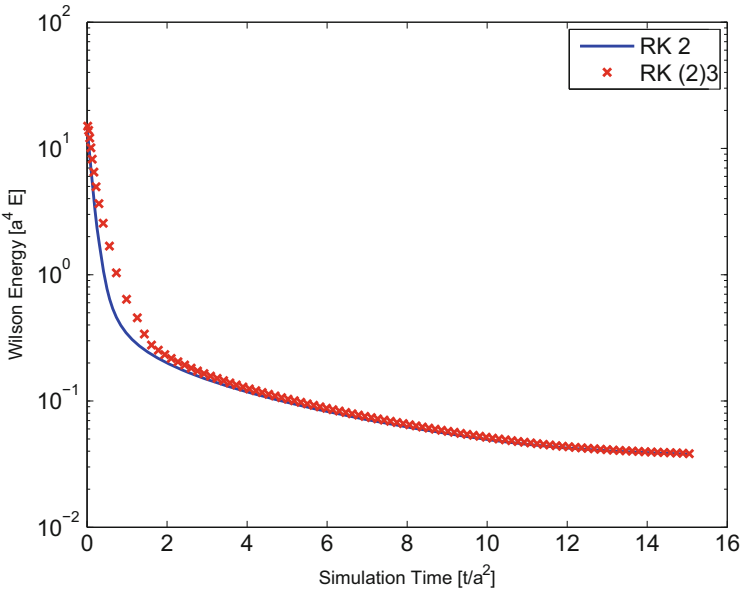


Fig. 1 Wilson energy computed with a Runge-Kutta method of convergence order 2 (*blue*) and with step size control (*red*)

for all n points in the configuration. Then, the solutions V_1 of convergence order three and $\hat{V}$ of convergence order two are reached for all lattice points via

$$V_1 = \exp(\Omega_1)V_0 \quad \text{and} \quad \hat{V}_1 = \exp(\Omega_1)V_0 \,.$$

Since the model error of $\hat{K}_i$ is larger than the one of K_i, we use the better approximation K_i instead of $\hat{K}_i$ if it is already available (this is the case in the first three stages). In Fig. 1, we compare the Wilson energy (6) computed (via the Wilson flow) with a RK method of order 2 with one computed with the aforementioned step size prediction. Here, the parameters for the step size control are set to ATOL=$1e{-}3$, RTOL=0, $\rho = 0.8$, facmin=0.5 and facmax=2. We see that a step size prediction works for the Wilson flow which consists of a set of matrices being Lie group elements.

5 Conclusion and Outlook

Usually, the Wilson flow is computed via a Runge-Kutta method with fixed step size and physicists are interested in the mean values of the observables computed from many configurations including their statistical errors. There are two advantages of the step size prediction explained here: first of all, the computational effort is reduced exploiting the dynamics of the system. Here, the parameters controlling the step size prediction have to be approved in a next step. Then, the step size prediction controls the numerical error such that the statistical errors can be reduced in a suitable manner. This has to be investigated in a next step.

Acknowledgements This work was supported by the Deutsche Forschungsgemeinschaft through the Collaborative Research Center SFB-TRR 55 Hadron Physics from Lattice QCD. I would like to acknowledge M. Lüscher for the deployment of his software package SU3YM.

References

1. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195**, 216 (1987). http://dx.doi.org/10.1016/0370-2693(87)91197-X
2. Lüscher, M.: Trivializing maps, the Wilson flow and the HMC algorithm (2009). http://arxiv.org/abs/0907.5491
3. Lüscher, M.: Properties and uses of the Wilson flow in lattice QCD (2010). http://arxiv.org/abs/1006.4518
4. Magnus, W.: On the exponential solution of differential equations for a linear operator. Commun. Pure Appl. Math. **7**(4), 649–673 (1954). http:/dx.doi.org/10.1002/cpa.3160070404
5. Munthe-Kaas, H.: Runge-Kutta methods on Lie groups. BIT **38**, 92–111 (1998)
6. Munthe-Kaas, H.: High order Runge-Kutta methods on manifolds. Appl. Numer. Math. **29**, 115–127 (1999)
7. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I – Nonstiff Problems. Springer, Berlin (2000)

Electro-Manipulation of Droplets for Microfluidic Applications

L.T. Corson, C. Tsakonas, B.R. Duffy, N.J. Mottram, C.V. Brown, and S.K. Wilson

Abstract There is a growing technology-driven interest in using external influences to move or shape small quantities of liquids, a process that is referred to as microfluidic actuation. The use of electrical, rather than mechanical, forces to achieve this actuation is convenient, because the resultant devices contain no moving parts. In this work we consider a sessile drop of an incompressible liquid with a high conductivity resting on the lower substrate inside a parallel-plate capacitor subjected to a relatively low frequency A.C. field. With the application of an electric field the drop deforms into a new static shape where the apex of the drop rises towards the upper electrode in order to balance the Maxwell electric stresses, surface tension and hydrostatic pressure on the interface. From experimental, numerical and asymptotic approaches we determine a predictive equation for the deformation as a function of initial contact angle and drop width, surface tension and applied voltage.

Keywords Electro-manipulated droplets • Microfluidic actuation

1 Introduction

There is a growing technology-driven interest in using external influences to move or shape small quantities of liquids, referred to as microfluidic actuation. The use of electrical, rather than mechanical, forces to achieve this actuation is convenient, because the resultant devices contain no moving parts. Existing non-mechanical microfluidic actuation techniques that are driven by the application of a voltage include electrowetting and liquid dielectrophoresis [10], with many applications including lab-on-a-chip [11], polymer surface patterning [17], as well

L.T. Corson • B.R. Duffy • N.J. Mottram (✉) • S.K. Wilson
Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, UK
e-mail: nigel.mottram@strath.ac.uk

C. Tsakonas • C.V. Brown
School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*, Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_149

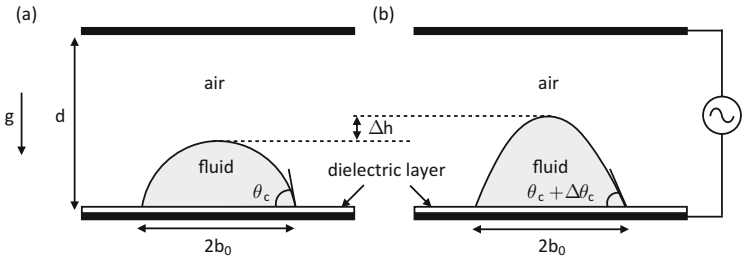


Fig. 1 Sketch of the geometry of a sessile drop resting on the lower substrate inside a parallel-plate capacitor. This substrate consists of an electrode coated with a *thin solid* dielectric layer. (**a**) No electric field applied. (**b**) An electric field applied across the capacitor deforms the drop

as optimisation of optical properties for polymer microlenses [13, 21], and droplet driven displays [7].

When an ionic, conducting liquid drop is subjected to a uniform electric field, the drop deforms as a result of the electric stresses on the interface, and it elongates in the direction of the electric field [18]. In this work we consider a sessile drop of an incompressible liquid with a high conductivity resting on the lower substrate inside a parallel-plate capacitor (Fig. 1) subjected to a relatively low frequency A.C. field. This situation is of particular interest to display device applications where the deformation of the drop can be used to change the optical properties of an image pixel [7]. With the application of an electric field the drop deforms into a new static shape where the apex of the drop rises towards the upper electrode in order to balance the Maxwell electric stresses, surface tension and hydrostatic pressure due to gravity on the interface. The lower electrode is coated with a thin solid dielectric layer, so the liquid drop is shielded from both electrodes. In this situation the mobile ions will reconfigure to reduce the electric field inside the drop to zero, so that the electric potential of the drop is a constant.

Previous experimental work on the deformation of sessile conductive drops in this geometry has included work on soap bubbles [2], polymer drops [14], water drops in air [3, 4, 15], water drops immersed in dielectric oil [16], and various alcohols in air [5, 6, 19]. As well as different liquids, these experiments also considered different substrate treatments (untreated, hydrophilic and hydrophobic), and therefore the initial contact angles of the drop varied greatly (specifically from 15° to 160°) [20]. Theoretical work in this geometry has tended to employ numerical techniques to solve the coupled electrostatic and augmented Young–Laplace equations for the electric field and drop profile. To simplify the process, many authors consider small drops where the assumption of negligible gravity is valid (see e.g. [1, 2, 14]).

In this paper we consider, experimentally and theoretically, the situation of pinned conductive liquid drops with contact angles that are close to $\pi/2$. Using both numerical and asymptotic approaches we find solutions to the coupled electrostatic and augmented Young–Laplace equations which agree very well with

the experimental results. Our asymptotic solution for the drop profile extends that of Basaran and Scriven [2] to drops that have initial contact angles close to $\pi/2$ and higher values of the electric field, and provides a predictive equation for the changes in the height as a function of the zero-field contact angle, drop radius, surface tension and applied voltage.

2 Experimental Setup

Figure 1 shows the experimental setup. A sessile drop of the liquid trimethylolpropane triglycidyl ether (TMPGE) rests on the lower substrate inside a parallel-plate capacitor with gap d between the electrodes. TMPGE is often considered a non-conducting dielectric material. However, dielectric studies show that, at the frequencies and voltages used in our experiments, this is a lossy material with a high conductivity that masks the dielectric polarisability so that the liquid is more accurately considered to be a conductive liquid. The electrodes were formed from a continuous layer of transparent conductor, indium tin oxide, on borosilicate glass slides. On the lower substrate the electrode is coated with a 1 μm thick layer of the dielectric material SU8 as well as a commercial hydrophobic coating to give contact angles close to $\pi/2$. The surface tension γ of the liquid was found to be $40.5\,\mathrm{mN\,m^{-1}}$ and the value of the density ρ was measured as $1157\,\mathrm{kg\,m^{-3}}$. In this study AC voltages at 1 kHz were used, and accurate values for the small height changes in the range 1–40 μm were obtained using a 20× microscope objective. Experiments were conducted for eight drops of various sizes with zero-field contact angles ranging from 86.1° to 93.1° (1.50–1.62 rad) and a range of cell gap to drop radius ratios from 2.45 to 4.21. In all experiments, the drop contact line was observed to be pinned with no appreciable movement even at the highest voltages used. Experimental results for the change in the height of the drop apex Δh will be shown in Sect. 4 when comparisons with numerical solutions of the theoretical model are made.

3 Theoretical Model

In the theoretical model of the experiment described in Sect. 2, an axisymmetric drop of an incompressible, perfectly conductive liquid rests on the lower substrate inside a parallel-plate capacitor surrounded by air, as shown in Fig. 1. Consistent with the experimental results, it is assumed that the drop is static and the contact line is pinned. We denote the constant drop base radius by b_0; the zero-field contact angle by θ_c; and the contact angle with an electric field applied by $\theta_c + \Delta\theta_c$, where $\Delta\theta_c$ is the electric-field-induced change in the contact angle. The electrodes are separated by a constant distance d, and we assume that the thickness of the dielectric layer on top of the lower electrode is negligible, so that the electric potential at the top of this

layer can be assumed to be zero. This is a reasonable approximation given that the thickness of the dielectric layer (1 μm) is small compared to the other dimensions of our system: b_0 and d are of the order of millimetres. At the top electrode the electric potential is equal to V.

We use spherical polar coordinates with their origin at the centre of the base of the drop with r denoting the distance from the origin and θ the angle the radial vector makes with the axis of symmetry. The drop interface is then defined as the zero level of the function $\eta = r - R(\theta)$, so that at any particular angle θ, the distance of the drop interface from the origin is $r = R(\theta)$.

The electric field $\mathbf{E} = -\nabla U$, where $U(r, \theta)$ is the electric potential, and the drop interface $r = R(\theta)$ are governed by Laplace's equation in the bulk and the normal stress balance, often termed the augmented Young–Laplace equation, on the drop interface. Since the drop is assumed to be a perfectly conducting liquid, the electric potential inside the drop is constant, and is determined by the close proximity of the lower electrode which is fixed at $U = 0$. The upper substrate is held at a potential $U = V$. The boundary conditions for the interface are those of axisymmetry and that the contact line is fixed at $r = b_0$. In addition, the volume of the drop $\mathscr{V}$ is assumed constant.

The governing equations and boundary conditions are made dimensionless by scaling distance by b_0, so that $r = b_0 r^*$ and $R = b_0 R^*$, and by writing

$$\mathscr{V} = \frac{2\pi b_0^3}{3}\mathscr{V}^*, \quad \mathbf{E} = \frac{V}{d}\mathbf{E}^*, \quad U = \frac{Vb_0}{d}U^*, \quad p - p_a = \frac{\gamma}{b_0}p^*. \tag{1}$$

We define a non-dimensional electric Bond number, gravitational Bond number, and scaled cell gap as

$$\delta^2 = \frac{\epsilon_0 \epsilon_2 V^2 b_0}{\gamma d^2}, \quad G = \frac{\rho g b_0^2}{\gamma}, \quad D = \frac{d}{b_0}, \tag{2}$$

respectively. Here p_a is the constant air pressure, ρ is the constant fluid density, γ is the constant surface tension, ϵ_0 is the permittivity of free space and $\epsilon_2 = \epsilon_{\text{air}}/\epsilon_0$ is the relative permittivity of the surrounding air; ϵ_2 is sufficiently close to one that we take it to equal unity.

Then, with the stars dropped for clarity, the electric potential U and the drop interface R must satisfy

$$\nabla^2 U = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial U}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial U}{\partial \theta}\right) = 0, \tag{3}$$

$$p - GR\cos\theta + \delta^2\left((\mathbf{E}\cdot\mathbf{n})^2 - \frac{1}{2}|\mathbf{E}|^2\right) = \nabla\cdot\mathbf{n}, \tag{4}$$

where $\mathbf{n} = \nabla\eta/|\nabla\eta|$ is the drop interface outward unit normal, so that $\nabla\cdot\mathbf{n}$ is twice the mean curvature. The electric potential and the drop interface must also satisfy

the boundary conditions

$$U(r, \pi/2) = 0, \quad U(R, \theta) = 0, \quad U(r, \theta) = D \text{ on } r\cos\theta = D, \tag{5}$$

$$R(\pi/2) = 1, \quad R'(0) = 0, \tag{6}$$

and the volume constraint

$$\mathscr{V} = \int_0^{\pi/2} R^3 \sin\theta \, \mathrm{d}\theta. \tag{7}$$

In order to compare to experimental measurements we consider the change in the height of the drop apex, $\Delta h = R(0) - R(0)|_{\delta^2=0}$, where $R(0)|_{\delta^2=0}$ is the zero-field height of the drop apex.

Using this theoretical model of the experimental system we will carry out numerical simulations and compare with the experimental results. Using evidence from these numerical simulations, we can find asymptotic solutions in appropriate limits, although these calculations are only summarised here; further details can be found in Corson et al. [9].

4 Numerical Results and Comparison with Experimental Results

The theoretical model described above was solved numerically using COMSOL [8] and MATLAB [12], where solutions to Laplace's equation (3), subject to (5), and solutions to the normal stress balance (4), subject to (6), were found iteratively until convergence was achieved. Figure 2 compares the experimentally measured change in the height of the drop apex (stars) to the full numerical solution (solid line) for the eight experimental drops. Figure 2 illustrates the validity of the numerical solutions over a range of parameter values: the gravitational Bond number G which increases from panel (a)–(h), the cell gap to drop radius D which decreases from panel (a)–(h), and the initial contact angle θ_c. We see that there is very good agreement between the experimental results and the numerical solution.

Figure 2 also shows numerical solutions using two additional simplifying assumptions: with gravity neglected $G = 0$ (dashed line) and with the upper electrode very far from the drop $D \to \infty$ (dashed-dotted line). We see that, for all values of D we consider, the simplifying assumption of large cell gap is valid. Unsurprisingly, since $G > 0.1$ for all drops, the numerical solutions with $G = 0$ overestimate the deformation, although the numerical solutions with gravity included do reproduce the experimental results.

From Fig. 2 we see that the deformation may be approximated by $\Delta h = \alpha_{0,2}\delta^2 + \alpha_{1,2}\epsilon\delta^2 + \alpha_{0,4}\delta^4$, and fitting to the experimental results we find coefficients $\alpha_{0,2} = 0.366 \pm 0.012$, $\alpha_{1,2} = -1.059 \pm 0.419$ and $\alpha_{0,4} = 0.090 \pm 0.096$. For the theoretical

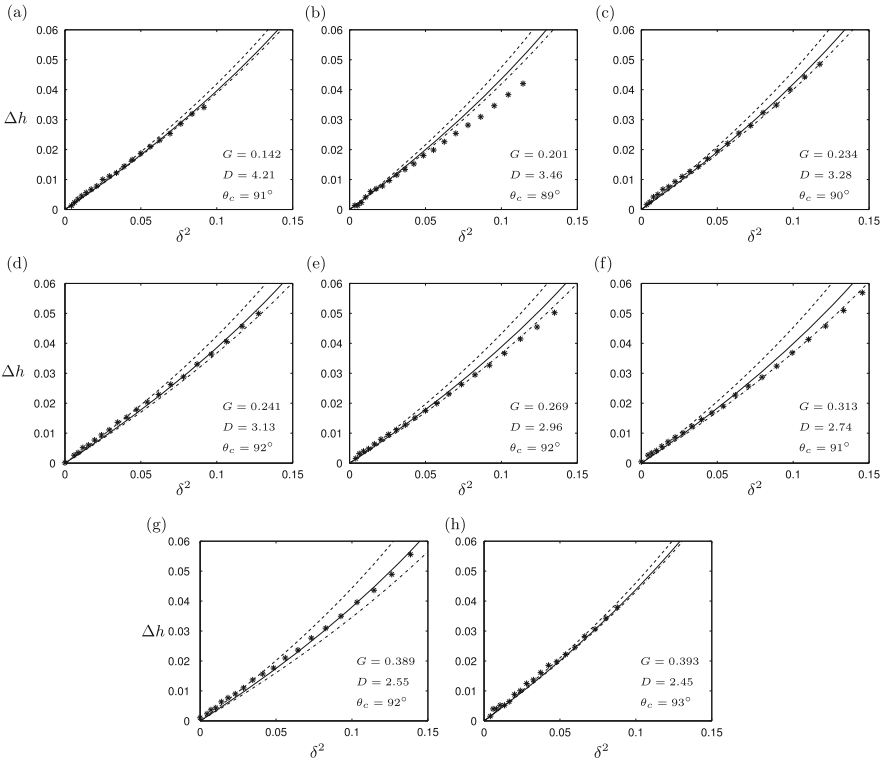


Fig. 2 The change in the height of the drop apex Δh plotted as a function of electric Bond number δ^2 for each experiment (*stars*) along with the full numerical solution (*solid lines*). Also shown are numerical solutions using two additional simplifying assumptions: with gravity neglected, $G = 0$ (*dashed lines*), and with the upper electrode very far from the drop, $D \to \infty$ (*dashed-dotted line*)

model we find the numerically determined coefficients $\alpha_{0,2} = 0.375, \alpha_{1,2} = -0.966$ and $\alpha_{0,4} = 0.541$. The experimental coefficients for $\alpha_{0,2}$ and $\alpha_{1,2}$ agree well with the theoretical coefficients, although the large amount of scatter in experimental values for the $\alpha_{0,4}$ coefficient suggests that the level of error in the experimental values is of the same size as δ^4.

5 Summary and Discussion

In this short paper we have considered, both experimentally and theoretically, the deformation due to an electric field of a pinned nearly hemispherical static sessile drop of a liquid with a high conductivity. Numerical solutions of the theoretical model were found to agree very well with experimental results for eight drops with

contact angles ranging from 86.1° to 93.1° and cell gap to drop radius ratios from 2.45 to 4.21.

For these experiments it was also noted that the assumption of an infinite cell gap $D \to \infty$ is a good approximation to the experimental situation. Also, although a model with the further simplifying assumption of zero-gravity $G = 0$ did not accurately reproduce the experimental results, the fit to experiments was sufficiently close to consider a simplified model in order to make analytical progress. Therefore, an approximate analysis of the theoretical model, with $G \to 0$ and $D \to \infty$, was undertaken (further details of which can be found in Corson et al. [9]). For this analysis we obtain

$$\Delta h = \frac{3}{8}\delta^2 - \left(\frac{1}{4} + \ln 2\right)\epsilon\delta^2 + \left(\frac{69}{64} - \frac{3}{4}\ln 2\right)\delta^4 \simeq 0.375\delta^2 - 0.943\epsilon\delta^2 + 0.558\delta^4,$$

which is a very good approximation to the numerically obtained results from the full model, and can readily be extended to all orders. The expressions for Δh that are described in this paper predict a reduction in the leading order deformation at the apex of the drop as the contact angle decreases (ϵ increases). The numerical implementation of the theoretical model, as well as the approximate analytical solution, therefore provide accurate solutions for the drop profile $R(\theta)$ and electric potential $U(r, \theta)$, and form a useful predictive tool for the electro-manipulation of a sessile conductive drop in a parallel-plate capacitor.

References

1. Adamiak, K.: Numerical investigation of shape of liquid droplets in an electric field. In: Brebbia, C.A., Kim, S., Oswald, T.A., Power, H. (eds.) Boundary Element XVII, pp. 494–511. Computational Mechanics Publications, Southampton (1995)
2. Basaran, O.A., Scriven, L.E.: Axisymmetric shapes and stability of pendant and sessile drops in an electric field. J. Colloid Interface Sci. **140**, 10–30 (1990)
3. Bateni, A., Susnar, S., Amirfazli, A., Neumann, A.: Development of a new methodology to study drop shape and surface tension in electric fields. Langmuir **20**, 7589–7597 (2004)
4. Bateni, A., Ababneh, A., Elliott, J.A.W., Neumann, A., Amirfazli, A.: Effect of gravity and electric field on shape and surface tension of drops. Adv. Space. Res. **36**, 64–69 (2005)
5. Bateni, A., Laughton, S., Tavana, H., Susnar, S., Amirfazli, A., Neumann, A.: Effect of electric fields on contact angle and surface tension of drops. J. Colloid Interface Sci. **283**, 215–222 (2005)
6. Bateni, A., Amirfazli, A., Neumann, A.: Effects of an electric field on the surface tension of conducting drops. Colloids Surf. A **289**, 25–38 (2006)
7. Blankenbach, K., Rawert, J.: Bistable electrowetting displays. Proc. SPIE **7956**, 795609 (2011)
8. COMSOL Multiphysics: Version 4.3b. COMSOL Inc. (2013)
9. Corson, L.T., Tsakonas, C., Duffy, B.R., Mottram, N.J., Sage, I.C., Brown, C.V., Wilson, S.K.: Deformation of a nearly hemispherical conducting drop due to an electric field: theory and experiment. Phys. Fluids **26**, 122106 (2014)
10. Jones, T.B.: Microfluidic schemes using electrical and capillary forces. J. Phys. Conf. Ser. **142**, 012054 (2008)

11. Kaler, K.V.I.S., Prakas, R., Chugh, D.: Liquid dielectrophoresis and surface microfluidics. Biomicrofluidics **4**, 022805 (2010)
12. MATLAB: Version 8.1 (R2013a). The MathWorks Inc., Natick, MA (2013)
13. O'Neill, F., Owen, G., Sheridan, J.: Alteration of the profile of ink-jet-deposited UV-cured lenses using applied electric fields. Optik **6**, 158–164 (2005)
14. Reznik, S.N., Yarin, A.L., Theron, A., Zussman, E.: Transient and steady shapes of droplets attached to a surface in a strong electric field. J. Fluid Mech. **516**, 349–377 (2004)
15. Roero, C.: Contact angle measurements of sessile drops deformed by DC electric field. In: Mittal, K.L. (ed.) Contact Angle, Wettability and Adhesion, vol. 4, pp. 165–176. Federal Institute of Technology Zurich, Zurich (2006)
16. Roux, J.M., Achard, J.L., Fouillet, Y.: Forces and charges on an undeformable droplet in the DC field of a plate condenser. J. Electrost. **66**, 283–293 (2008)
17. Schaffer, E., Thurn-Albrecht, T., Russell, T.P., Steiner, U.: Electrohydrodynamic instabilities in polymer films. Europhys. Lett. **53**, 518–524 (2010)
18. Taylor, G.: Disintegration of water drops in an electric field. Proc. R. Soc. Lond. A **280**(1382), 383–397 (1964)
19. Tsakonas, C., Corson, L., Sage, I.C., Brown, C.V.: Electric field induced deformation of hemispherical sessile drops of ionic liquid. J. Electrost. **76**(6), 437–440 (2014)
20. Vancauwenberghe, V., Di Marco, P., Brutin, D.: Wetting and evaporation of a sessile drop under an external electric field: a review. Colloids Surf. A **432**, 50–56 (2013)
21. Zhan, Z., Wang, K., Yao, H., Cao, Z.: Fabrication and characterization of aspherical lens manipulated by electrostatic field. Appl. Opt. **48**(22), 4375–4380 (2009)

Fiber Suspension Flows: Simulations and Existence Results

Uldis Strautins

Abstract Main result of this article is demonstrating the weak global in time well posedness result for the equations governing fiber suspension flows for sufficiently small initial data under mild assumptions about the nonlinear equation for fiber orientation dynamics and the constitutive law, thus extending the previous local in time results. The required estimate of growth of the H^2 norm is granted if the L^∞ norm of fiber orientation state variables remains bounded. This is the case for fiber orientation tensors.

Keywords Existence result • Fiber orientation dynamics • Fiber suspension flow

1 Introduction

Technical fiber suspension flows are commonly modelled as flows of a non-Newtonian fluid with a material law characterizing the relation between stress and strain dependent on the local microstructure of the fluid. The microstructure is primarily characterized by fiber orientation and concentration although it also depends on a multitude of other parameters such as shape and interfacial phenomena on the fiber-matrix surface. Fiber orientation is commonly characterized by orientation tensors which are low order momenta of the distribution function. This description is concise (fiber orientation can be described using just five independent scalar fields) and allows modelling the microstructure dynamics in a closed way once a closure approximation has been chosen. As a result, one ends with a system of incompressible generalized Navier-Stokes equation and a transport equation governing the fiber orientation.

The FO equations depend on the model chosen, but also on the closure approximation, see e.g. [1]. Assuming that higher order coefficients in expansion of the ODF vanish, the linear closure approximation is obtained. Powerful well-posedness results have been obtained by Munganga et al. for this linear case [4, 5].

U. Strautins (✉)
Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulvāris 29, Rīga LV-1459, Latvia
e-mail: uldis.strautins@lu.lv

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_150

Otherwise the equations are non-linear, and in this case much less is known. A fixed point argument used by Guilope and Saut [3] to prove existence for a viscoelastic fluid was the point of departure. From there Galdi et al. [2] have demonstrated the well posedness in a weak sense for the Folgar-Tucker equation with the quadratic closure approximation (we note, however, that the necessary estimates fail if the orientational distribution is exactly proportional to the scalar shear rate). Strautins has extended that approach to a wide class of nonlinear equations under mild assumptions; these are not restricted to fiber suspensions alone. This paper aims to extend these results to global in time results by exploiting the fact that the supremum norm of the components of FO tensor are bounded (e.g., for the second order moment $a^{(2)}$ we have $Tr(a^{(2)}) = 1$ and $a^{(2)} \geq 0$).

2 Equations

We formulate the equations in a general setting. Linear momentum of the suspension is governed by a Navier-Stokes equation:

$$Re\frac{Dv}{Dt} - \triangle v + \nabla p - \mathrm{div} T = \rho b, \tag{1}$$

$$\nabla \cdot v = 0,$$

where v is velocity, p is pressure and T is extra stress tensor due to the fibers. Typically T depends on the fourth order fiber orientation tensor, e.g., $T = N_p a^{(4)} : D + N_s(D \cdot a^{(2)} + a^{(2)} \cdot D)$ for constants N_p and N_s, where D is the symmetrical part of ∇v, but we only require the linearity of this constitutive law wrt. components of ∇v.

Fiber orientation is described by a finite set of scalars governed by a system of transport-reaction equations, for example, the components of $a^{(2)}$ and Folgar-Tucker equation. We allow arbitrary finite set of parameters, which we group in a vector field s. This field satisfies the transport-reaction equations

$$s_t + (v \cdot \nabla)s = F(s, \nabla v).$$

We assume that $s = 0$ is a stationary state when the suspension does not flow $\nabla v = 0$, i.e., $F(0, 0) = 0$. Some fiber orientation diffusion models having isotropic orientation diffusion have $a^{(2)} = \frac{1}{3}I$ as stationary point; then we define the components of s as the components of

$$a^{(2)} - \frac{1}{3}I$$

We define the space of solenoidal vector fields with vanishing trace $V := H^1_{0,\sigma}(\Omega)$. Let $\Pi : H^1_0 \to V$ be the standard projection to incompressible fields, and denote $\mathscr{L} = -\Pi\triangle$.

Next we formulate an initial-boundary value problem for (v, s) and state our main assumptions.

Problem $\mathscr{P}$: Find

$$v(\cdot, t) \in V \text{ and } s(\cdot, t) \in (H^2(\Omega))^{d_2},$$

such that for almost all $t \in (0, T)$ the following equations are satisfied

$$Re[v_t + (v \cdot \nabla)v] + \eta\mathscr{L}v - \mathrm{div}T = b,$$

$$T = \mathscr{T}(s, \nabla v),$$

$$s_t + (v \cdot \nabla)s = \mathscr{F}(s, \nabla v),$$

and the given initial conditions $v(\cdot, 0) = v_0$ and $s(\cdot, 0) = s_0$ hold.

Here $\mathscr{T}$, $\mathscr{F}$ satisfy the following assumptions.

$\mathscr{T} \in \left[C^2(\mathbb{R}^{d_2+d^2})\right]^{d^2}$, the function $\mathscr{T}(s, \kappa)$ is linear in κ, and together with its first and second order derivatives has a polynomial growth with respect to $|s|$. Furthermore, whenever $|s| = 0$ or $|\kappa| = 0$, the gradient $\nabla_{(s,\kappa)}\mathscr{T} = 0$.

$\mathscr{F} \in \left[C^2(\mathbb{R}^{d_2+d^2})\right]^{d_2}$, the function $\mathscr{F}(s, \kappa)$ is linear in κ, and together with its first and second order derivatives has a polynomial growth with respect to $|s|$. Moreover, $\mathscr{F}(0, 0) = 0$. □

An important assumption is that $\mathscr{F}$ must be C^2 w.r.t. velocity gradient. This condition is not fulfilled in the classical Folgar-Tucker model where the fiber orientation diffusion is assumed to be proportional to the local scalar shear rate, which is not differentiable in the origin.

3 Main Result

In the dissertation [6] the following result is proven:

Theorem 1 *Let Ω have C^3 boundary, $b \in L^2_{\mathrm{loc}}(\mathbb{R}^+; H^1)$, $b' \in L^2_{\mathrm{loc}}(\mathbb{R}^+; H^{-1})$, $v_0 \in D(\mathscr{L})$ and $s_0 \in H^2(\Omega)^{d_2}$. Then there exist positive constants K and T such that whenever*

$$\|b'\|_{L^2(0,T;H^1)\cap L^2(0,T;H^{-1})} + \|v_0\|_{D(\mathscr{L})} + \|s_0\|_{H^2} \le K,$$

then Problem $\mathscr{P}$ admits at least one solution (v, P, s) in $\Omega \times (0, T)$ in the spaces

$$v \in L^2(0,T;H^3),\ \ v' \in L^2(0,T;V), \quad P \in L^2(0,T;H^2),$$
$$s \in L^\infty(0,T;H^2),\ s' \in L^\infty(0,T;H^1).$$

The constant K depends only on the domain Ω and on the material constants, while T also depends on the data. Finally, such solutions satisfy the estimate

$$\|v\|_{L^2(0,T;H^3)\cap L^\infty(0,T;D(\mathscr{L}))} + \|v'\|_{L^2(0,T;V)\cap L^\infty(0,T;H)}$$
$$+\|P\|_{L^2(0,T;H^2)} + \|s\|_{L^\infty(0,T;H^2)} + \|s'\|_{L^\infty(0,T;H^1)} \le k,$$

where k depends on the data in such a way that

$$k \to 0 \ \text{ as } \ \|b\|_{L^2(0,T;H^1)} + \|b'\|_{L^2(0,T;H^{-1})} + \|v_0\|_{H^2} + \|s_0\|_{H^2} \to 0.$$

We extend it to the following.

Theorem 2 *Let the ordinary differential equation $\dot{s} = \mathscr{F}(s, \kappa(t))$ have a compact invariant manifold, whatever the matrix valued function $\kappa(t)$ be, and the initial data s_0 are from this invariant manifold, then the weak solution of Theorem 1 can be continued uniquely for all time points $t \in \mathbb{R}_+$.*

We proceed the proof as for Theorem 1 in [6]. The assumptions about boundedness of s allows us to estimate $\|\nabla s \otimes \nabla s\|_{L^2} \le c\|s\|_{L^\infty}\|s\|_{H^2}$ by Gagliardo-Nirenberg, hence from (3.18) in [6] we deduce the linear bound $\frac{d}{dt}\|s\|^2_{H^2} \le c_1\|s\|^2_{H^2}$.

4 Conclusions

Interpretation of simulation results should rely on qualitative theory of the underlying equations. For example, in fiber suspension flows in channel domain there may be $\nabla v = 0$ on the centerline at all times. Thus, the initial fiber orientation state is retained while in arbitrary close neighbourhood the solution tends to the stationary solution. Does the analytical solution break down at certain time? The result of this paper guarantees that the solutions of fiber orientation models do not develop singularities under the assumptions of the Theorem.

However, we have seen that the Folgar-Tucker model does not satisfy the conditions. This equation can be modified by using a mollified version of the local shear rate as suggested in [6].

Acknowledgements This work was partially supported by the grant 623/2014 of the Latvian Council of Science.

References

1. Chung, D.H., Kwon, T.H: Fiber orientation in the processing of polymer composites. Korea-Aust. Rheol. J. **14**, 175–188 (2002)
2. Galdi, G.P., Reddy, D.B.: Well-posedness of the problem of fiber suspension flows. J. Non-Newtonian Fluid Mech. **83**, 205–230 (1999)
3. Guillope, C., Saut, J.C.: Existence results for the flow of viscoelastic fluids with a differential constitutive law. Nonlinear Anal. Theory Methods Appl. **15**, 849–869 (1990)
4. Munganga, J.M.W.: Existence and stability of solutions for steady flows of fiber suspension flows. J. Math. Fluid Mech. **15**, 197–214 (2013)
5. Munganga, J.M.W., Reddy, D.B.: Local and global existence of solutions to the equations for fibre suspension flows. Math. Models Methods Appl. Sci. **12**, 1177–1203 (2002)
6. Strautins, U.: Flow-driven orientation dynamics in two classes of fibre suspensions. Ph.D. thesis, TU Kaiserslautern (2008)

Global Existence of Weak Solutions to an Angiogenesis Model

N. Aïssa and R. Alexandre

Abstract We prove global existence of a weak solution to the angiogenesis model proposed by A. Tosin, D. Ambrosi, L. Preziosi in Bull. Math. Biol. (2006) 7, 1819-1836. The model consists of compressible Navier-Stokes equations coupled with a reaction-diffusion equation describing the concentration of a chemical solution responsible of endothelial cells migration and blood vessels formation.

Proofs are based on the control of the entropy associated to the hyperbolic equation of conservation mass and the adaptation of the results of P.L. Lions dealing with compressible fluids which are inevitable for all models dealing with compressible Navier-Stokes equations.

We use the vanishing artificial viscosity method to prove existence of solutions, the main difficulty for passing to the limit is the lack of compactness due to hyperbolic equation which usually induces resonance phenomenon. This is overcome by using the concept of the compactness of effective viscous pressure combined with suitable renormalized solutions to the hyperbolic equation of mass conservation.

Keywords Angiogenesis model • Existence result • Vanishing artificial viscosity method

1 Introduction

In this paper, we discuss a mathematical model introduced by Ambrosi et al. [1], Tosin et al. [7] describing the formation of blood vessels in the organism. The self-assembly of endothelial cells into a vascular labyrinth is called vasculogenesis and is responsible for the early formation of blood vessels in the embryo.

We refer to the papers [1, 7] for the physical and biological relevances of the model considered herein. We set $Q_T = (0, T) \times \Omega$ where $T > 0$ is a fixed time and

N. Aïssa (✉)
Laboratoire AMNEDP, USTHB, BP 32 El Alia 16111, Algeria
e-mail: aissa.naima@gmail.com

R. Alexandre
Paris Tech, 12 rue Edouard Manet 75013 Paris, France
e-mail: radjesvarane.alexandre@paristech.fr

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_151

Ω is a bounded and regular subset of $\mathbb{R}^2$ or $\mathbb{R}^3$. The model problem settled in Q_T is given by

$$\begin{cases} \partial_t n + \nabla \cdot (n\, v) = 0, \\ \partial_t (nv) + \nabla \cdot (nv \otimes v) - \mu \Delta v - \xi \nabla (\nabla \cdot v) + \gamma nv + \nabla(\varphi(n)) = an \nabla c, \\ \partial_t c - d\, \Delta c + \dfrac{1}{\tau}\, c = \alpha(n)\, n, \\ n(0) = n_0,\ (nv)(0) = m_0,\ c(0) = c_0 \text{ in } \Omega, \\ v = 0,\ \nabla c \cdot N = 0 \text{ on } (0,T) \times \partial\Omega. \end{cases} \tag{1}$$

State variable n represents the density of cellular matter and v it's velocity field. The chemotaxis force is $f_{chem} = a\, \nabla c$ where c represents concentration of the chemical soluble factor and a measures the intensity of cell response per unit mass. It is assumed that the chemotactic interaction has an attractive character, that is $a > 0$. The dissipative interaction with the substratum is taken into account by the term $f_{diss} = -\gamma nv$, $\gamma > 0$. Parameter $d > 0$ is the diffusion factor, $\alpha(n) > 0$ is the rate of release and $\tau > 0$ is the half-life of the soluble mediator.

Moreover, N is the outward unit normal to the boundary, $\mu > 0$ and $\xi = \mu + \lambda > 0$ are the Lamé parameters appearing in the usual Navier-Stokes theory. Finally, $\varphi(n)$ represents the repulsive force felt by the cells when they crowd, according to the papers [1, 7], we shall consider the following explicit example

$$\varphi(n) = \mathbf{1}_{\{n > n_c\}} (n - n_c)^k, \tag{2}$$

where $n_c > 0$ is a constant called the close-packing density and $k > \frac{3}{2}$. In order to simplify the proofs and to highlight the main ideas, we will consider the case when $k = 4$. The case of powers less than 4 will be discussed at the end of the paper. Our main theorem for model (1) is given by

Theorem 1 *Let the pressure function φ be given by (2) and the entropy function $\psi(n)$ be defined by $\psi(n) = n \int_0^n \frac{\varphi(s)}{s^2} ds$. Assume*

(i) $n_0 \geq 0$, $n_0 \in L^4(\Omega)$, $\psi(n_0) \in L^1(\Omega)$,
(ii) $v_0 \in L^2(\Omega)$, $n_0 v_0 \in L^2(\Omega)$,
(iii) $c_0 \in H^1(\Omega)$, $\alpha(n)$ *is such that* $|\alpha|_\infty \leq \alpha_0$, $|\alpha'(n)| \leq \frac{\alpha_1}{n}$ *for some positive constants* α_0, α_1.

Then, there exists a global weak solution (n, v, c) to (1) satisfying the energy estimate

$$\mathscr{E}(t) + \int_0^t \mu |\nabla v|^2 + \xi |\nabla \cdot v|^2 + n|v|^2 dx ds \leq C_1 \int_0^t (\mathscr{E}(s) + 1) ds + C_2 \mathscr{E}_0, \tag{3}$$

with some constants $C_k > 0$ depending only on the data for $k = 1, 2$. The energy of the system is defined by

$$\mathscr{E}(t) = \int_\Omega \{\psi(n) + \frac{1}{2}n|v|^2 + b|c|^2\}dx.$$

for some appropriate positive constant b.

We note that the convex function ψ satisfies the differential equation

$$s\psi'(s) - \psi(s) = \varphi, \quad \psi(0) = 0, \tag{4}$$

and is related to the definition of renormalized solution of the continuity equation that will be recalled later. Through the paper, we shall use the standard Hilbert spaces $L^2(\Omega), H^m(\Omega)$ and $L^p(0,T;H^m(\Omega))$ equipped with the norms denotes $\|.\|, \|.\|_{H^m(\Omega)}$ and $\|.\|_{L^p(0,T;H^m(\Omega))}$ respectively.

Weak solutions alluded to in the above statement are finite energy weak solutions of problem (1). In particular, (n, v, c) satisfy

1. $n \geq 0$, $n \in L^\infty(0,T;L^4(\Omega))$, $v \in (L^2(0,T;H_0^1(\Omega)))^2$, $\sqrt{n}v \in L^2(0,T;L^2(\Omega))$, $c \in L^2(0,T;H^2(\Omega))$;
2. The energy is locally integrable on $(0,T)$ and energy inequality (3) holds in $\mathscr{D}'(0,T)$;
3. Equations from (1) are satisfied in $\mathscr{D}'((0,T)\times\Omega)$ and in addition, mass conservation law holds in the sense of renormalized solutions.

Plan of the Paper In Sect. 2 we use Galerkin method to solve an approximate system with artificial viscosity; in Sect. 3, we give the main tools for the vanishing viscosity method. Finally, in Sect. 3.3, we prove the strong convergence of the mass density and prove the main theorem.

2 Approximate System

It is natural to introduce an artificial viscosity to transform the hyperbolic equation of mass conservation into a parabolic one. The new problem can be treated by standard methods.

2.1 Approximate System with Artificial Viscosity

$$\begin{cases} \partial_t n + \nabla\cdot(nv) - \varepsilon\Delta n = 0 \text{ on } (0,T)\times\Omega, \\ \nabla n\cdot N = 0 \text{ on } \partial\Omega \text{ and } n(0,x) = n_0 \text{ on } \Omega. \end{cases} \tag{5}$$

$$\begin{cases} \partial_t(nv) + \nabla \cdot (nv \otimes v) - \mu \Delta v - \xi \nabla(\nabla \cdot v) + \gamma nv + \nabla(\varphi(n)) + \varepsilon \nabla v . \nabla n = a\, n . \nabla c, \\ v = 0 \text{ on } \partial\Omega, (nv)(0,x) = m_0 \text{ on } \Omega. \end{cases} \tag{6}$$

$$\begin{cases} \partial_t c - d\Delta_x c + \frac{1}{\tau} c = \alpha(n) n \text{ on } (0,T) \times \Omega, \\ \nabla c \cdot N = 0 \text{ on } \partial\Omega, c(0,x) = c_0. \end{cases} \tag{7}$$

The artificial viscosity is balanced in the momentum equation by the term $\varepsilon \nabla v . \nabla n$ in order to control energy estimates.

2.2 Second Level Approximation: Galerkin Approximation

Let $\{\eta_i\}_{i=0}^{\infty}$ be an orthonormal basis in $L^2(\Omega)$ and an orthogonal basis in $H_0^1(\Omega)$. We denote $X_m =$ span $\{\eta_j\}_{1 \le j \le m}$ and P_m the orthogonal projection of $L^2(\Omega)$ onto X_m.

Adapting proofs of Novotny and Straskhaba [6], p. 352 by combining Galerkin's method and Schauder's fixed point theorem, we get existence and uniqueness of an approximate solution $(n^m, v^m, c^m) \in \mathscr{C}^0(0,T; X_m)$ to (5)–(6)–(7) satisfying the following uniform estimates with respect to m which are deduced from energy estimates and interpolation argument

$$\|v_m\|_{L^2(0,T;W^{1,2}(\Omega))} + \|n_m\|_{L^\infty(0,T;L^4(\Omega))} + \|n_m |v_m|^2\|_{L^\infty(0,T;L^1(\Omega))} \le L(\mathscr{E}_0, T), \tag{8}$$

$$\|c_m\|_{L^2(0,T;W^{2,2}(\Omega))} \le L(\mathscr{E}_0, T), \tag{9}$$

$$\sqrt{\varepsilon} \|\nabla n_m\|_{L^2(\Omega_T)} + \varepsilon \|\Delta n_m\|_{L^2(\Omega_T)} \le L(\mathscr{E}_0, T, \Omega), \tag{10}$$

$$\|n_m v_m\|_{L^\infty(0,T;L^{\frac{8}{5}}(\Omega))} + \|\partial_t n_m\|_{L^2(0,T;(W^{1,\frac{8}{3}}(\Omega))^\star)} \le L(\mathscr{E}_0, T, \Omega), \tag{11}$$

$$\|n_m^{\frac{16}{3}}\| \le L(\mathscr{E}_0, T, \varepsilon, \Omega), \tag{12}$$

Here L is a positive constant which is, in particular, independent of m. Moreover, if ε is not explicitly written in the argument of L, then L is independent of ε. We may pass to the limit as $m \to \infty$. Indeed, we deduce from the previous estimates that for subsequences,

$$v_m \rightharpoonup v_\varepsilon \ \text{in } L^2(0,T;H_0^1(\Omega)),\ n_m \rightharpoonup n_\varepsilon \ \text{weakly}-\star \ \text{in } \ L^\infty(0,T;L^4(\Omega)), \tag{13}$$

$$\nabla n_m \rightharpoonup \nabla n_\varepsilon \ \text{weakly in } \ L^2(\Omega_T),\ \ c_m \to c_\varepsilon \ \text{strongly in } \ L^2(0,T;H^1(\Omega)). \tag{14}$$

Moreover, using Aubin's compactness result, $n_m \to n_\varepsilon$ in $L^2(\Omega_T)$. Consequently, using (12) and the Vitali's theorem; we get $n_m \to n_\varepsilon$ in $L^p(\Omega_T)$, $1 \leq p < \frac{16}{3}$. Then $\varphi(n_m) \to \varphi(n_\varepsilon)$ strongly in $L^{\frac{4}{3}}(\Omega_T)$ as $m \to \infty$. Finally we check that $(n_\varepsilon, v_\varepsilon, c_\varepsilon)$ is a solution to (5)–(6)–(7).

We have thus proven the existence of solutions of the approximate problem.

3 Main Ideas for Letting $\varepsilon \to 0$ and Mathematical Tools

Using estimates (8) and (9) which are also valid for $(n_\varepsilon, v_\varepsilon, c_\varepsilon)$, we get for subsequences,

$$v_\varepsilon \rightharpoonup v \text{ in } L^2(0,T;H_0^1(\Omega)), \quad n_\varepsilon \rightharpoonup n \text{ in } L^2(\Omega_T) \text{ weak}, c_\varepsilon \to c \text{ in } L^2(0,T;H^1(\Omega)) \text{ strong}.$$

At this stage, we loose the strong convergence of n_ε because the constant occurring in (10) depends on ε. Consequently $\varphi(n_\varepsilon)$ should converge to a mere measure. The limiting problem would, a priori, be different from the initial one. The procedure we will use is not standard and is an adaptation of Lions [4]. The ideas are the following and will be developed later:

- The real pressure exerted on a unit volume is not only $\varphi(n)$ but $\varphi(n) - (\nu + \xi)\nabla \cdot v$ which is called the effective viscous pressure. This term behaves as though it was a strongly convergent quantity in the following sense [4]

$$(\varphi(n^\varepsilon) - (\nu + \xi)\nabla \cdot v^\varepsilon \rightharpoonup F, \quad n^\varepsilon \rightharpoonup n) \Rightarrow (\varphi(n^\varepsilon) - (\nu + \xi)\nabla \cdot v^\varepsilon)n^\varepsilon \to Fn \quad \text{a.e.}$$

- Renormalized solutions of the continuity equation with suitable test functions allow to control density oscillations.
- Strong convergence under strict convexity theorem leads to the strong convergence of the density n_ε in at least $L^1(Q_T)$.

Next, we recall some mathematical tools which will be useful in the sequel.

3.1 Renormalized Solutions

It is well known that hyperbolic systems of conservation laws are not well posed in the class of distributional solutions. The appearance of discontinuities represents a source of energy dissipation which is not captured by the weak formulation. Multiplying the continuity equation

$$\partial_t n + \nabla \cdot (nv) = f, \quad n(0) = n_0, \tag{15}$$

by expression $B'(n)$ where B is a smooth function, we formally deduce the equation

$$\partial_t B(n) + \nabla \cdot (B(n)v) + b(n)\nabla \cdot v = B'(n)f, \quad b(s) = B'(s)s - B(s), \tag{16}$$

Definition 1 We shall say that n (and v) is a renormalized solution of the continuity equation (15) on $(0,T)\times\Omega$ if (16) holds in $\mathscr{D}'((0,T)\times\Omega)$ for any functions $B \in \mathscr{C}[0,\infty)\cap\mathscr{C}^1(0,\infty), b \in \mathscr{C}[0,\infty), B(0) = 0$ (Including that the continuity equation holds in $\mathbb{R}^2$ by extending the solution by 0 outside Ω).

By well known results due to DiPerna-Lions transport theory [2], if $n \in L^2(0,T;L^2(\Omega))$ is a weak solution associated to $v \in L^2(0,T;H^1(\Omega))$ then n is a renormalized solution of the continuity equation on $(0,T)\times\Omega$.

3.2 *Compactness of the Effective Viscous Pressure*

We check that Proposition 7.36 in [6] applies in our case since we have uniform bounds of n_ε in $L^5(Q_T)$ [5] which allows us to conclude that $\varphi(n_\varepsilon) \rightharpoonup \overline{\varphi(n)}$ in at least $L^{\frac{5}{4}}(Q_T)$. Then the compactness of the effective viscous pressure writes

$$\overline{\varphi(n)n} - (\xi+\mu)\overline{n\nabla\cdot v} = \overline{\varphi(n)}n - (\xi+\mu)n\nabla\cdot v \ \text{ a.e.} \tag{17}$$

where n, $\overline{\varphi(n)n}$, $\overline{n\nabla\cdot v}$ and $\overline{\varphi(n)}$ are respectively the weak limits of n_ε, $\varphi(n_\varepsilon)n_\varepsilon$, $n_\varepsilon\nabla\cdot v_\varepsilon$ and $\varphi(n_\varepsilon)$.

3.3 *Strong Convergence of the Density and Proof of the Main Theorem*

As $n \in L^2(\Omega_T)$ and $v \in L^2(0,T,H^1(\Omega))$ then n is a renormalized solution to the continuity equation $\partial_t n + \nabla\cdot(nv) = 0$ so it satisfies

$$\partial_t b(n) + \nabla\cdot(b(n)v) + (nb'(n) - b(n))\nabla\cdot v = 0 \tag{18}$$

with the test function $b(s) = s\,ln(s)$. Next, n_ε satisfies renormalized inequality

$$\partial_t b(n_\varepsilon) + \nabla\cdot(b(n_\varepsilon)v_\varepsilon) + (n_\varepsilon b'(n_\varepsilon) - b(n_\varepsilon))\nabla\cdot v_\varepsilon - \varepsilon\Delta b(n_\varepsilon) \le 0 \tag{19}$$

with the test function $b_h(s) = s\,ln(s+h)$. Integrating over $(0,T)\times\Omega$ and letting $h\to 0$ and $\varepsilon\to 0$ we get

$$\int_\Omega \overline{n(T)log(n(T))} - n(T)log(n(T))dx \le \int_{\Omega_T} n\nabla\cdot v - \overline{n\nabla\cdot v}dxdt$$

Hence using the compactness of the effective viscous flux (17)

$$\int_\Omega \overline{n(T)log(n(T))} - n(T)log(n(T))dx \leq \frac{1}{\mu+\xi}\int_{\Omega_T}(\overline{\varphi(n)}n - \overline{\varphi(n)n})dxdt \tag{20}$$

Next we use results on monotone and convex operators to control density oscillations. Since log and φ are nondecreasing, using Lemmas 3.33 and 3.35 in [6], we get $\overline{n(T)log(n(T))} - n(T)log(n(T)) \geq 0$ while $(\overline{\varphi(n)}n - \overline{\varphi(n)n}) \leq 0$. Hence

$$\overline{\varphi(n)}n = \overline{\varphi(n)n} \ \text{ a.}e. \tag{21}$$

Next, as φ is nondecreasing, thanks to the results on monotone operators (Cf. Lemma 3.39 in [6]), we deduce from (21) that

$$\overline{\varphi(n)} = \varphi(n), \ \text{ a.}e. \ in \ \ \Omega_T. \tag{22}$$

Finally, as φ is strictly convex, we can use Theorem 2.11 in [3] to conclude that $n_\varepsilon \to n$ a.e. in Ω_T. Then by the density improved uniform estimates of Lions [4], Lions [5], Novotny and Straskhaba [6] in $L^5(\Omega_T)$ and Vitali's theorem

$$n_\varepsilon \to n \ strongly \ \ in \ L^p(\Omega_T), 1 \leq p < 5. \tag{23}$$

Consequently $\varphi(n_\varepsilon) \to \varphi(n)$ in $L^1(\Omega)$ and we check that (n, v, c) is a weak solution of (1).

Remark 1 In the case when the pressure function is in the form

$$\varphi(n) = \mathbf{1}_{\{n>n_c\}}(n-n_c)^k, \ \ 2 \leq k < 4,$$

Galerkin approximation convergence fails because (10) relies on L^4 uniform estimates of the density n. To overcome this difficulty, we modify the pressure term $\nabla\varphi(n)$ by adding the artificial pressure $\delta\nabla n^\beta$ for some constants $\beta \geq 4$, $\delta > 0$. As $n \in L^2(Q_T)$, the DiPerna-Lions's transport theory applies to the continuity equation and the previous arguments remain valid for letting $\delta \to 0$ to get a global weak solution to (1). We refer to [6] for more details.

If $\frac{9}{5} \leq k < 2$, using the Lions's approach [4] and an improved estimate on the density we deduce from [5] that $n \in L^2(Q_T)$. Then we can proceed like in the previous case.

If $\frac{3}{2} < k < \frac{9}{5}$, n may not belong to $L^2(Q_T)$ and the DiPerna-Lions theory fails. By Fereisl's approach [3, 6], we can overcome this by using generalized renormalized solutions to control density oscillations. We refer to [3, 6] for more details. Finally the case when $k < \frac{3}{2}$ is an open problem.

References

1. Ambrosi, D., Gamba, A., Serini, G.: Cell directional and chemotaxis in vascular morphogenesis. Bull. Math. Biol. **66**, 1851–1873 (2004)
2. DiPerna, R.J., Lions, P.-L.: Ordinary differential equations, transport theory and Sobolev spaces. Invent. Math. **98**, 511–547 (1989)
3. Feireisl, E.: Dynamics of Viscous Compressible Fluids. Oxford University Press, Oxford (2003)
4. Lions, P.L.: Mathematical Topics in Fluid Mechanics, vol. 2. Compressible Models. Oxford Lecture Series in Mathematics and its Applications, vol. 10. Oxford University Press, Oxford (1998)
5. Lions, P.L.: Bornes sur la densité pour les équations de Navier-Stokes compressibles isentropiques avec conditions aux limites de Dirichlet . C. R. Acad. Sci. Paris Série I **328**, 659–662 (1999)
6. Novotny, A., Straskhaba, I.: Introduction to the Theory of Compressible Flow. Oxford University Press, Oxford (2004)
7. Tosin, A., Ambrosi, D., Preziosi, L.: Mechanics and chemotaxis in the morphogenesis of vascular networks. Bull. Math. Biol. **68**(7), 1819–1836 (2006)

High-Order Compact Schemes for Black-Scholes Basket Options

Bertram Düring and Christof Heuer

Abstract We present a new high-order compact scheme for the multi-dimensional Black-Scholes model with application to European Put options on a basket of two underlying assets. The scheme is second-order accurate in time and fourth-order accurate in space. Numerical examples confirm that a standard second-order finite difference scheme is significantly outperformed.

Keywords Black-Scholes model • Computational finance • Option pricing

1 Introduction

The multidimensional Black-Scholes model for option pricing (e.g. [8]) considers $n \in \mathbb{N}_{\geq 2}$ underlying assets $S_i \in [0, \infty[$ for $i = 1, \ldots, n$, where each asset follows a geometric Brownian motion,

$$\mathrm{d}S_i(t) = \mu_i S_i(t)\mathrm{d}t + \sigma_i S_i(t)\mathrm{d}W^{(i)}(t), \tag{1}$$

where $\mu_i \in \mathbb{R}$ is the drift and $\sigma_i \geq 0$ is the volatility of the asset S_i, respectively, for $i = 1, \ldots, n$ and $\mathrm{d}W^{(i)}(t)$ denotes a Wiener Process at time $t \in [0, T]$ for some $T > 0$. The correlation between the assets is given by $\mathrm{d}W^{(i)}(t)\mathrm{d}W^{(j)}(t) = \rho_{ij}\mathrm{d}t$. The Lemma of Itô and standard no-arbitrage arguments lead to the following (backward in time) parabolic partial differential equation with mixed second-order derivative

B. Düring
Department of Mathematics, University of Sussex, Pevensey II, Brighton, BN1 9QH, UK
e-mail: b.during@sussex.ac.uk

C. Heuer (✉)
Angewandte Mathematik und Numerische Analysis, Fachbereich C – Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, Gaußstrasse 20, 42119 Wuppertal, Germany
e-mail: c.heuer@sussex.ac.uk

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_152

terms for the option price $V = V(S_1, S_2, \ldots, S_n, t)$ (see, e.g. [8]),

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sum_{i=1}^{n} \sigma_i^2 S_i^2 \frac{\partial^2 V}{\partial S_i^2} + \sum_{\substack{i,j=1 \\ i<j}}^{n} \rho_{ij}\sigma_i\sigma_j S_i S_j \frac{\partial^2 V}{\partial S_i \partial S_j} + \sum_{i=1}^{n} r S_i \frac{\partial V}{\partial S_i} - rV = 0,$$

with $S_i > 0$, $t \in [0, T[$ and $r \geq 0$ denoting the riskless interest rate. When examining a European Put basket option, the final condition is given by

$$V(S_1, \ldots, S_n, T) = \max\left(K - \sum_{i=1}^{n} \omega_i S_i, 0\right),$$

where the asset weights satisfy $\sum_{i=1}^{n} \omega_i = 1$ and additionally $\omega_i > 0$ for $i = 1, \ldots, n$ if we have short-selling restrictions. Suitable boundary conditions are discussed later.

The transformations

$$x_i = \frac{\gamma}{\sigma_i} \ln\left(\frac{S_i}{K}\right), \quad \tau = T - t \quad \text{and} \quad u = e^{r\tau}\frac{V}{K}, \tag{2}$$

where $\gamma > 0$ is a constant scaling parameter, yield the (forward in time) parabolic partial differential equation

$$u_\tau - \frac{\gamma^2}{2}\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2} - \gamma^2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \rho_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \gamma \sum_{i=1}^{n} \left[\frac{\sigma_i}{2} - \frac{r}{\sigma_i}\right] \frac{\partial u}{\partial x_i} = 0, \tag{3}$$

where $x \in \mathbb{R}^n$ and $\tau \in \Omega_\tau =]0, T]$. Under the same transformations the initial condition for a European Put basket is given by

$$u(x_1, \ldots, x_n, 0) = \max\left(1 - \sum_{i=1}^{n} \omega_i e^{\frac{\sigma_i x_i}{\gamma}}, 0\right). \tag{4}$$

When looking for numerical methods to approximate solutions to problem (3), (4), subject to suitable boundary conditions, finite difference schemes can be employed, at least for space dimensions up to three. Standard discretisations, however, only yield second-order convergence in terms of the spatial discretisation parameter. Alternatively, high-order compact schemes can be used which only use points on a compact computational stencil, while having fourth-order consistency in space, see for example [1, 3, 4, 6, 7] and the references therein. A drawback is that the derivation of high-order compact schemes (and their numerical stability analysis) is algebraically demanding, hence most works in this area restrict themselves to the one-dimensional case. An additional complication is present in (3) in form of the mixed second-order derivative terms.

In a forthcoming paper [2] we derive new high-order compact schemes for a rather general class of linear parabolic partial differential equations with mixed second-order derivative terms and time- and space-dependent coefficients in arbitrary space dimension $n \in \mathbb{N}$. In the present paper we focus on the multi-dimensional Black-Scholes model (3), (4). We present a new high-order compact scheme which is second-order accurate in time and fourth-order accurate in space. To ensure high-order convergence in the presence of the initial condition (4) with low regularity we employ the smoothing operators of Kreiss et al. [5]. Numerical examples for pricing European Put options on a basket of two underlying assets confirm that a standard second-order finite difference scheme is significantly outperformed.

2 Discrete Two-Dimensional Black-Scholes Equation

For the discretisation of (3) with $n = 2$ we replace the spatial domain by the rectangle $\Omega = [x_{\min}^{(1)}, x_{\min}^{(1)}] \times [x_{\min}^{(2)}, x_{\min}^{(2)}]$ with $-\infty < x_{\min}^{(i)} < x_{\min}^{(i)} < \infty$ for $i = 1, 2$. On Ω, we define the grid

$$G_h^{(2)} = \left\{ (x_{i_1}^{(1)}, x_{i_2}^{(2)}) \in \Omega \mid x_{i_k}^{(k)} = x_{\min}^{(k)} + i_k h,\ 1 \le i_k \le N_k,\ k = 1, 2 \right\}, \tag{5}$$

where $h > 0$, $N_k \in \mathbb{N}$ and $x_{\max}^{(k)} = x_{\min}^{(k)} + N_k h$ for $k = 1, 2$. By $\overset{\circ}{G}{}_h^{(2)}$ we denote the interior of $G_h^{(2)}$. We present the coefficients of a semi-discrete scheme of the form

$$\sum_{j_1=i_1-1}^{i_1+1} \sum_{j_2=i_2-1}^{i_2+1} \left[\hat{M}_{j_1,j_2} \partial_\tau U_{j_1,j_2}(\tau) + \hat{K}_{j_1,j_2} U_{j_1,j_2}(\tau) \right] = \tilde{g}(x, \tau),$$

at time τ for each point $x \in \overset{\circ}{G}{}_h^{(2)}$ for the two-dimensional Black-Scholes equation using $n = 2$ in (3). By $U_{j_1,j_2}(\tau)$ we denote the approximation of $u(x_{i_1}^{(1)}, x_{i_2}^{(2)}, \tau)$ after semi-discretisation in space with $\left(x_{i_1}^{(1)}, x_{i_2}^{(2)}\right) \in G_h^{(2)}$.

The general idea underlying the derivation of the high-order compact scheme is to operate on the differential equation (3) as an additional relation to obtain finite difference approximations for high-order derivatives in the truncation error. Inclusion of these expressions in a central difference method for Eq. (3) increases the order of accuracy to fourth order while retaining a compact stencil. A detailed derivation of this scheme and a thorough von Neumann stability analysis are presented in a forthcoming paper [2]. In the two-dimensional case we obtain the

following coefficients

$$\begin{aligned}
\hat{K}_{i_1,i_2} =& -\frac{2\gamma^2\rho_{12}^2}{3h^2} + \frac{5\gamma^2}{3h^2} + \frac{\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)^2}{3} + \frac{\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)^2}{3},\\
\hat{K}_{i_1\pm1,i_2} =& \frac{\gamma^2\rho_{12}^2}{3h^2} \pm \frac{\gamma\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)}{3h} \mp \frac{\gamma\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)\rho_{12}}{3h} - \frac{\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)^2}{6} - \frac{\gamma^2}{3h^2},\\
\hat{K}_{i_1,i_2\pm1} =& \frac{\gamma^2\rho_{12}^2}{3h^2} \pm \frac{\gamma\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)}{3h} \mp \frac{\gamma\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)\rho_{12}}{3h} - \frac{\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)^2}{6} - \frac{\gamma^2}{3h^2},\\
\hat{K}_{i_1\pm1,i_2-1} =& \pm\frac{\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)}{12} - \frac{\gamma\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)}{12h} \pm \frac{\gamma\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)}{12h}\\
&- \frac{\gamma\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)\rho_{12}}{6h} \pm \frac{\gamma\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)\rho_{12}}{6h} - \frac{\gamma^2}{12h^2} \pm \frac{\gamma^2\rho_{12}}{4h^2} - \frac{\gamma^2\rho_{12}^2}{6h^2},\\
\hat{K}_{i_1\pm1,i_2+1} =& \frac{\gamma\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)}{12h} \mp \frac{\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)}{12} \pm \frac{\gamma\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)}{12h}\\
&+ \frac{\gamma\,\rho_{12}\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)}{6h} \pm \frac{\gamma\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)\rho_{12}}{6h} - \frac{\gamma^2}{12h^2} \mp \frac{\gamma^2\rho_{12}}{4h^2} - \frac{\gamma^2\rho_{12}^2}{6h^2},
\end{aligned}$$

as well as

$$\begin{aligned}
M_{i_1+1,i_2\pm1} =& M_{i_1-1,i_2\mp1} = \pm\frac{\rho_{12}}{24}, \qquad & M_{i_1,i_2} =& \frac{2}{3},\\
M_{i_1\pm1,i_2} =& \frac{1}{12} \mp \frac{h\left(\frac{\sigma_1}{2} - \frac{r}{\sigma_1}\right)}{12\gamma}, \qquad & M_{i_1,i_2\pm1} =& \frac{1}{12} \mp \frac{h\left(\frac{\sigma_2}{2} - \frac{r}{\sigma_2}\right)}{12\gamma}.
\end{aligned}$$

Additionally, $\tilde{g}(x,\tau) = 0$ for $x \in \overset{\circ}{G}{}_h^{(2)}$ and $\tau \in \Omega_\tau$. After presenting the high-order compact discretisation for the spatial interior we now discuss the boundary conditions.

3 Discretisation of the Boundary Conditions

The first boundary we discuss is $S_i = 0$ for some $i \in \{1,2\}$ at time $t \in [0,T[$. Once the value of the asset is zero, it stays constant over time, see (1). If only one asset reaches its minimum value, using $S_i = 0$ for $i \in \{1,2\}$ in the multi-dimensional Black-Scholes equation with $n = 2$ leads to the one-dimensional Black-Scholes

equation for the asset S_j with $j = \{1, 2\} \setminus i$. One can either transform the solution of the one-dimensional Black-Scholes partial differential equation using (2) or derive a fourth-order compact scheme for these boundaries similarly to the space interior. If both asset values are minimal, we have

$$u(x_{\min}^{(1)}, x_{\min}^{(2)}, \tau) = u(x_{\min}^{(1)}, x_{\min}^{(2)}, 0)$$

for $\tau \in]0, \tau_{\max}]$ after transforming with (2).

Upper boundaries are boundaries with $S_i = S_i^{\max} > 0$ with $i \in \{1, 2\}$ at time $t \in [0, T[$. For a sufficiently large $S_i^{\max}$, we can approximate

$$\frac{\partial V(S_1, S_2, t)}{\partial S_i}\Big|_{S_i = S_i^{\max}} \equiv 0, \tag{6}$$

with $S_k \in \left[S_k^{\min}, S_k^{\max}\right]$ for $k = \{1, 2\} \setminus \{i\}$. If only one underlying asset S_i reaches its maximum value, using (6) in the two-dimensional Black-Scholes differential equation leads to the one-dimensional Black-Scholes differential equation for the underlying asset S_j with $j = \{1, 2\} \setminus \{i\}$. One can either transform the solution of this equation using (2) or transform the one-dimensional Black-Scholes differential equation using (2) and derive a fourth-order compact scheme for these boundaries. When both underlying assets reach their maximum value, we have

$$u(x_1^{\max}, x_2^{\max}, \tau) = u(x_1^{\max}, x_2^{\max}, 0)$$

for $\tau \in]0, \tau_{\max}]$ after using the transformations (2). Since the boundaries behave similar, we have

$$u(x_1^{\min}, x_2^{\max}, \tau) = u(x_1^{\min}, x_2^{\max}, 0), \qquad u(x_1^{\max}, x_2^{\min}, \tau) = u(x_1^{\max}, x_2^{\min}, 0),$$

for $\tau \in]0, \tau_{\max}]$.

4 Time Discretisation

We use an equidistant time grid of the form $\tau = k\,\Delta\tau$ for $k = 0, \ldots, N_\tau$ with $N_\tau \in \mathbb{N}$. Using a Crank-Nicolson-type time discretisation with step size $\Delta\tau$ leads to

$$\sum_{j_1=i_1-1}^{i_1+1} \sum_{j_2=i_2-1}^{i_2+1} \left[\hat{M}_{j_1,j_2} + \frac{\Delta\tau}{2}\hat{K}_{j_1,j_2}\right] U_{j_1,j_2}^{k+1}$$

$$= \sum_{j_1=i_1-1}^{i_1+1} \sum_{j_2=i_2-1}^{i_2+1} \left[\hat{M}_{j_1,j_2} - \frac{\Delta\tau}{2}\hat{K}_{j_1,j_2}\right] U_{j_1,j_2}^{k} + (\Delta\tau) g(x)$$

at each point $\left(x_{i_1}^{(1)}, x_{i_1}^{(2)}\right) \in G_h^{(2)}$, where only points of the compact stencil are used. By U_{i_1,i_2}^k we denote the approximation of $u(x_{i_1}^{(1)}, x_{i_2}^{(2)}, \tau_k)$. For the Crank-Nicolson type time discretisation this compact scheme has consistency order two in time and four in space. Thus, using $\Delta\tau \in \mathcal{O}\left(h^2\right)$, leads to fourth-order consistency in terms of the spatial stepsize $h > 0$.

5 Numerical Experiments

In this section we present numerical experiments for the Black-Scholes European Puts basket option in space dimension $n = 2$. According to [5], we cannot expect fourth-order convergence if the initial condition u_0 is only in $C^0(\Omega)$. In [5] suitable smoothing operators are identified in Fourier space. Since the order of consistency of our high-order compact schemes is four, we use the smoothing operator Φ_4 (see [5]), given by its Fourier transformation

$$\hat{\Phi}_4(\omega) = \left(\frac{\sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}}\right)^4 \left[1 + \frac{2}{3}\sin^2\left(\frac{\omega}{2}\right)\right].$$

This leads to the smoothed initial condition given by

$$\tilde{u}_0(x_1, x_2) = \frac{1}{h^2} \int_{-3h}^{3h} \int_{-3h}^{3h} \Phi_4\left(\frac{x}{h}\right) \Phi_4\left(\frac{y}{h}\right) u_0(x_1 - x, x_2 - y)\, \mathrm{d}x\mathrm{d}y,$$

for any stepsize $h > 0$, where $\Phi_4(x)$ denotes the Fourier inverse of $\hat{\Phi}_4(\omega)$. If u_0 is smooth enough in the integrated region around $(x_1, x_2) \in \Omega$, we have $\tilde{u}_0(x_1, x_2) = u_0(x_1, x_2)$. Thus it is possible to identify the points where smoothing is necessary for a given initial condition. This approach reduces the necessary computations significantly. Note that as $h \to 0$, the smoothed initial condition $\tilde{u}_0$ converges to the original initial condition u_0 given in (4). Hence the approximation of the smoothed problem tends towards the true solution of (3).

For examining the numerical convergence rate we use the relative l^2-error $\|U_{\text{ref}} - U\|_{l^2} / \|U_{\text{ref}}\|_{l^2}$, as well as the l^∞-error $\|U_{\text{ref}} - U\|_{l^\infty}$, where U_{ref} denotes a reference solution on a fine grid and U is the approximation. We determine the numerical convergence order of the schemes as the slope of the linear least square fit of the individual error points in the loglog-plots of error versus number of discretisation points per spatial direction. We compare the high-order compact scheme to a standard second-order scheme, which results from applying the standard central difference operators directly in (3) with $n = 2$. We use the following

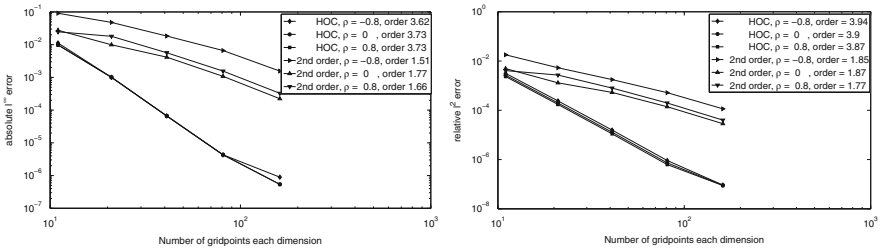


Fig. 1 Absolute l^∞-error and relative l^2-error for two-dimensional Black-Scholes Basket Put with smoothed initial condition

parameters,

$$\sigma_1 = 0.25,\ \sigma_2 = 0.35,\ \gamma = .25, \quad r = \log(1.05),\ \omega_1 = 0.35 = 1 - \omega_2.$$

and $K = 10$. We set the parabolic mesh ratio $\Delta\tau/h^2 = 0.4$, but emphasise that neither the von Neumann stability analysis presented in [2] nor additional numerical experiments reveal any restrictions on this relation, indicating unconditional stability of the scheme. We use different values $\rho_{12} = -0.8$, $\rho_{12} = 0$ and $\rho_{12} = 0.8$ for the correlation. In Fig. 1 we show plots of the l^∞-error and the relative l^2-error. The high-order compact scheme performs highly similar for the three different correlation values, the points are almost identical. The numerical convergence orders for the high-order compact scheme range between 3.62 and 3.73 for the l^∞-error, and between 3.87 and 3.94 for the relative l^2-error. The high-order compact scheme significantly outperforms the standard second-order discretisation in all cases.

Acknowledgements The second author was partially supported by the European Union in the FP7-PEOPLE-2012-ITN Program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN *STRIKE—Novel Methods in Computational Finance*).

References

1. Düring, B., Fournié, M.: High-order compact finite difference scheme for option pricing in stochastic volatility models. J. Comput. Appl. Math. **236**(17), 4462–4473 (2012)
2. Düring, B., Heuer, C.: High-order compact schemes for parabolic problems with mixed derivatives in multiple space dimensions. SIAM J. Numer. Anal. **53**(5), 2113–2134 (2015)
3. Düring, B., Fournié, M., Heuer, C.: High-order compact finite difference schemes for option pricing in stochastic volatility models on non-uniform grids. J. Comput. Appl. Math. **271**(18), 247–266 (2014)
4. Karaa, S., Zhang, J.: Convergence and performance of iterative methods for solving variable coefficient convection-diffusion equation with a fourth-order compact difference scheme. Comput. Math. Appl. **44**(3–4), 457–479 (2002)

5. Kreiss, H., Thomee, V., Widlund, O.: Smoothing of initial data and rates of convergence for parabolic difference equations. Commun. Pure Appl. Math. **23**, 241–259 (1970)
6. Spotz, W., Carey, G.: Extension of high-order compact schemes to time-dependent problems. Numer. Methods Partial Differ. Equ. **17**(6), 657–672 (2001)
7. Tangman, D., Gopaul, A., Bhuruth, M.: Numerical pricing of options using high-order compact finite difference schemes. J. Comp. Appl. Math. **218**(2), 270–280 (2008)
8. Wilmott, P.: Derivatives: The Theory and Practice of Financial Engineering. Wiley, Chichester (1998)

Mathematical Formulation of Bioventing Optimal Design Strategies

Filippo Notarnicola

Abstract Bioventing is a technology used to abate the presence of pollutants in the subsoil. Microorganisms biodegrade the pollutant but the biochemical reaction requires oxygen and so an airflow is induced in the subsoil by means of injection and/or extraction wells.

Costs, final result and decontamination time are reliant on contaminant type, soil permeability and several other factors, but oxygen subsoil concentration plays a very important role. For this reason a rational choice of well locations and flow rates is required.

The mathematical definition of the optimal design problem will be set-up starting from a simplified mathematical model describing the bioventing system.

A formal definition of decontaminated subsoil will be given and the set of system control variables will be identified. Optimization strategies such as cost minimization and time optimization will be mathematically described.

Keywords Bioventing • Optimal design

1 Introduction and the Simplified Mathematical Model

In bioventing, the pollutant is removed from the subsoil by bacterial activity and the process can be enhanced by adding oxygen into the soil using air injection wells. The literature regarding bioventing is very vast and involves many aspects: microbial, biochemical, geophysical and so on [3, 4, 6, 8].

This publication concerns the problem of designing a subsoil bioventing intervention that is, starting from the knowledge of soil characteristics and pollutant concentration, to identify appropriate well locations and well air flow rates useful to achieve predetermined aims. Different goals can be pursued in the design phase. Starting from the mathematical model describing the physical phenomenon, two optimization strategies will be proposed: the first one reduces the remediation cost

F. Notarnicola (✉)
Istituto per le Applicazioni del Calcolo - Consiglio Nazionale delle Ricerche, Via Amendola 122-d, 70123 Bari, Italy
e-mail: f.notarnicola@ba.iac.cnr.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_153

and the second one reduces the remediation time. Here, as a first step, we do not analyze the analytical or computational procedures to solve the optimal problem, but we describe the two design objectives giving a formal mathematical definition of them.

The mathematical model outlined in this paper is based on the continuum approach for the fluid flow in porous media [1, 5] and it is non stationary, multi phase and multi component. It is a simplified version of the general model in [7] and it is based on the following simplifying assumptions.

Three different phases are present: the *air* gas phase, the *water* liquid phase and the *pollutant* liquid phase. The water and the pollutant phases are considered to be immiscible and therefore, although both are liquid, they are treated as distinct phases. Only air is in the gas phase, composed of two components, oxygen and non-oxygen. The micro-organisms which biodegrade the pollutant grow where oxygen and hydrocarbons are available and they are subject to *Fick*ian diffusion.

There are nine unknowns of the system and they depend on space and time:

- H, C, G the water, pollutant and gas *phase saturations*,
- X_O and X_N, the gas (air) relative oxygen and non oxygen *fractions*,
- p_H, p_C, p_G the water, pollutant and gas *phase pressures*,
- B, the *bacteria concentration*

Using mass conservation and the expression of the generalized Darcy law for multiphase flow it is possible to write the following continuity equations. For the water, the pollutant, the oxygen and the non-oxygen fractions of the air we have:

$$\frac{\partial}{\partial t}(\Phi H \rho_H) = \mathbf{div}\left(\rho_H \frac{k_H}{\mu_H}\mathbf{K}\,(\mathbf{grad} p_H - \rho_H\,\mathbf{g})\right) + \mathbf{div}\,(\Phi\,\mathbf{D}_H\ \mathbf{grad}\ H\rho_H) + \frac{1}{Y_H}\ \Phi\rho_O X_O G\ \ \Phi\rho_C\, C\ \ B \tag{1}$$

$$\frac{\partial}{\partial t}(\Phi C \rho_C) = \mathbf{div}\left(\rho_C \frac{k_C}{\mu_C}\mathbf{K}\,(\mathbf{grad} p_C - \rho_C\,\mathbf{g})\right) + \mathbf{div}\,(\Phi\,\mathbf{D}_C\ \mathbf{grad}\ C\rho_C) - \left(\frac{1}{Y_C} g(C,O)B\ +\ M_C\ \ \Phi\rho_O X_O G\ \ \Phi\rho_C\, C\ \ B\right) \tag{2}$$

$$\frac{\partial}{\partial t}(\Phi G X_O \rho_O) = \mathbf{div}\left(X_O\rho_O \frac{k_G}{\mu_G}\mathbf{K}\,(\mathbf{grad} p_G - \rho_G\,\mathbf{g})\right) + \mathbf{div}\,(\Phi\,\mathbf{D}_G\ \mathbf{grad}\ GX_O\rho_O) - \left(\frac{1}{Y_O}\{g(C,O) - d\}B + M_O\ \ \Phi\rho_O X_O G\ \ \Phi\rho_C\, C\ \ B\right) + r_O \tag{3}$$

$$\frac{\partial}{\partial t}(\Phi G X_N \rho_N) = \mathbf{div}\left(X_N\rho_N \frac{k_G}{\mu_G}\mathbf{K}\,(\mathbf{grad} p_G - \rho_G\,\mathbf{g})\right) + \mathbf{div}\,(\Phi\,\mathbf{D}_G\ \mathbf{grad}\ GX_N\rho_N) + r_N \tag{4}$$

Each of the above equations contains: the Darcy term, the diffusive/dispersive term, the source term describing the biochemical reaction due to the bacteria activity and, finally, the source terms representing injection or extraction wells, that is: r_O, r_N.

In each point of the space domain for water, pollutant, gas saturations and for the oxygen and non oxygen gas relative fractions, the following two equalities hold:

$$H + C + G = 1 \tag{5}$$

$$X_O + X_N = 1 \tag{6}$$

The bacteria spatial concentration satisfies the following diffusion and reaction continuity equation:

$$\frac{\partial}{\partial t} B = D_B \ \textbf{div grad}\ B + g(C, O)B - dB \tag{7}$$

where D_B is the bacteria diffusion coefficient and $g(C, O)$ (a positive function usually based on Monod type terms, [7]) and d (a positive constant) represent the microorganism growth and decay rate, respectively.

In a porous media system at the interface between different phases there is a difference of pressure (see [1, pp. 441–449], [5, pp. 50–60]) called capillary pressure which depends on the phase saturations and porous media characteristics. If we assume that the water phase H wets the pollutant phase C and that C wets the gas phase G then for the capillary pressures pc_{CH} and pc_{GC}, we have:

$$pc_{CH} = p_C - p_H = f_1(H, C, G) \tag{8}$$

$$pc_{GC} = p_G - p_C = f_2(H, C, G) \tag{9}$$

The two functions $f_1(H, C, G)$ and $f_2(H, C, G)$ are considered known and several expressions of them have been proposed (see [5, pp. 54–60]).

The nine Eqs. (1)–(9) describe the simplified bioventing system and the following parameters appear in the equations:

- $\mathbf{K}$ the intrinsic permeability tensor;
- k_H, k_C, k_G the relative permeability of water, pollutant and gas phases;
- μ_H, μ_C, μ_G the dynamic viscosity of water, pollutant and gas phases;
- ρ_H, ρ_C, ρ_G, ρ_O, ρ_N the density of water, pollutant, gas, oxygen and non-oxygen part of the gas phase, respectively. The gas phase density, $\rho_G = \rho_O X_O + \rho_N X_N$;
- $\mathbf{D}_H$, $\mathbf{D}_C$, $\mathbf{D}_G$ the dispersion tensor of water, pollutant and gas phases;
- $\mathbf{g} = (0, 0, -g)^T$ the gravitational acceleration vector;
- M_C, M_O the hydrocarbon and oxygen *metabolic consumption constants*;
- Y_C, Y_O the hydrocarbon and oxygen *yield coefficients*.

Air is injected or extracted by N active wells and we suppose that the bioventing model is written in a spatial three dimensional domain. Each one of the N wells, identified by the index $i = 1, \ldots, N$, has horizontal position (x_i, y_i) and consists of

a vertical segment of length l_i (the air permeable zone) located at a depth z_i from the ground level. Therefore each well is spatially described by (x_i, y_i, z_i, l_i). Moreover, for the well i, q_i is the total volumetric air flow rate injected or extracted and it is uniformly distributed along the segment l_i.

For reasons of optimization, the flow rates of the wells can change during the decontamination intervention: then we suppose that the terms q_i depend on time. Let us denote, at time t, the vector of the flow rates of the N wells with $\mathbf{q}(t)$ and the vector—constant in time—of the geographical descriptions of the N wells with $\mathbf{w}$:

$$\mathbf{q}(t) = (q_1(t), \ldots, q_N(t)) \qquad \mathbf{w} = [(x_1, y_1, z_1, l_1), \ldots, (x_N, y_N, z_N, l_N)] \tag{10}$$

If $s = (x, y, z)$ is a point in the three dimensional spatial domain, we denote the function describing, in space and time, the air source term with $\Sigma(s, t; \mathbf{q}(t), \mathbf{w})$. The presence of $\mathbf{q}(t)$ and $\mathbf{w}$ in $\Sigma(s, t; \mathbf{q}(t), \mathbf{w})$ shows that its definition depends on the N wells. Now, by denoting the oxygen fraction of the air in the atmosphere with p , we can finally define the external source terms in Eqs. (3) and (4):

$$r_O(s, t) = p\,\Sigma(s, t; \mathbf{q}(t), \mathbf{w}) \qquad r_N(s, t) = (1 - p)\,\Sigma(s, t; \mathbf{q}(t), \mathbf{w}) \tag{11}$$

2 Design Optimization and Control

The final goal of the decontamination is to lead the pollutant under a level fixed by laws and regulations, in all the points of the polluted regions; the biochemical decontamination reaction requires oxygen and, therefore, air is injected in the subsoil by means of wells.

2.1 Control Variables and Space State Definitions

The bioventing process will be controlled by ruling the air velocity flow field, that is choosing convenient geographical positions and air flow rates of the wells: it follows that $\mathbf{w}$ and $\mathbf{q}(t)$, used to define $\Sigma(s, t; \mathbf{q}(t), \mathbf{w})$, are the control variables of the system and all the unknowns depend on them.

In theory N, the number of the wells, could change during the decontamination intervention but, in this paper, we consider it N as fixed. The value of N should be sufficiently high and the optimal control procedure will select the strictly necessary active wells imposing a null flow rate to the others.

For physical reasons the control variables $\mathbf{q}(t)$ and $\mathbf{w}$ are subject to constraints. The flow rates of the wells are limited by the pump equipment power. If M_t is the maximum total flow rate available for the bioventing intervention and M is the maximum flow rate of each single well then, for $\mathbf{q}(t)$, we have the following set

of the admissible time dependent vector functions:

$$\mathcal{Q} = \left\{\mathbf{q}(t) : \mathbb{R}_+ \to \mathbb{R}^N \;\middle|\; \sum_{i=1}^{N} |q_i(t)| \le M_t, \quad |q_i(t)| \le M \;\; (i = 1, \ldots, N), \quad t > 0\right\} \tag{12}$$

We explicitly observe that the flow rates of the wells can assume positive or negative values for injection or extraction wells, respectively.

Moreover, the set of the admissible control values for the well positions is:

$$W_c = \left\{\mathbf{w} = [(x_1, y_1, z_1, l_1), \ldots, (x_N, y_N, z_N, l_N)] \in \mathbb{R}^{4N} \;\middle|\; 0 \le l_i \le L \quad \text{and} \quad Z_{min} \le z_i \le Z_{max} \;\; \text{for} \;\; i = 1, \ldots, N\right\} \tag{13}$$

where the constraints on l_i limit the vertical active portion of the wells; limitations are also imposed to the depths z_i, due to subsoil characteristics.

To solve the system (1)–(9) an initial condition is required. Denoting a point in the spatial domain with $s = (x, y, z)$, let $S_0 = \left(H^0(s), C^0(s), G(s), X_O^0(s), X_N^0(s), p_H^0(s), p_C^0(s), p_G^0(s), B^0(s)\right)$ a spatial initial configuration of all the nine unknowns of the model at the time $t = 0$; in the framework of the control theory S_0 is the initial state of the system. In the subsequent definitions we shall consider the initial state S_0 as assigned and fixed. Now, for the initial condition S_0, suppose that the functions $H(s,t)$, $C(s,t)$, $G(s,t)$, $X_O(s,t)$, $X_N(s,t)$, $p_H(s,t)$, $p_C(s,t)$, $p_G(s,t)$, $B(s,t)$ are the solution of the differential problem (1)–(9) in correspondence to the well source terms r_O and r_N which, in turn, depend on the vector function $\mathbf{q}(t) \in \mathcal{Q}$ and on the constant vector $\mathbf{w} \in W_c$.

The goal of a decontamination intervention is the removal of the pollutant inside the contaminated domain and, therefore, we will focus our attention on the pollutant spatial concentration. Then, if $C(s,t)$ is the solution of the system (1)–(9) in correspondence to the initial condition S_0 and wells description $\mathbf{q}$ and $\mathbf{w}$, and $C(s,T)$ is the space distribution of the pollutant saturation at a fixed time T, we define the *space state of the pollutant at time T* as the following space dependent function:

$$\widetilde{C}(s; T, S_0, \mathbf{q}, \mathbf{w}) = C(s,T) \tag{14}$$

We observe that in $\widetilde{C}(s; T, S_0, \mathbf{q}, \mathbf{w})$ the values $T, S_0, \mathbf{q}, \mathbf{w}$ are explicitly mentioned to show the dependence of the pollutant space state from those values. From this definition it follows that $\widetilde{C}(s; 0, S_0, \mathbf{q}, \mathbf{w}) = C^0(s)$.

For a given initial condition S_0, we would like to define the set of the *pollutant reachable space states at time T* whose elements are obtained in correspondence to an admissible system control:

$$R(T) = \left\{\widetilde{C}(s; T, S_0, \mathbf{q}, \mathbf{w}) \text{ is a } \textit{space state of the pollutant at time } T \;\middle|\; (\mathbf{q}, \mathbf{w}) \in \mathcal{Q} \times W_c \text{ and } T > 0\right\} \tag{15}$$

The set $R(T)$ contains several possible spatial contaminant distributions: but only some of them can be associated to a decontaminated state of the soil. Therefore the set of the *acceptable contaminant states at time T* will be defined, that is the subset $A(T) \subset R(T)$, containing the contaminant reachable space states which, at time T, fulfills laws and regulations. If the law requires the pollutant level to be below a fixed constant P_L, we have:

$$A(T) = \left\{ \widetilde{C}(s; T, S_0, \mathbf{q}, \mathbf{w}) \in R(T) \quad \middle| \quad \| C(s,t) \|_\infty \leq P_L \text{ for } \ t \geq T \right\} \tag{16}$$

where $C(s,t)$ is the pollutant solution of the system (1)–(9) which is connected with $\widetilde{C}(s; T, S_0, \mathbf{q}, \mathbf{w})$ and, therefore, such that $C(s,T) = \widetilde{C}(s; T, S_0, \mathbf{q}, \mathbf{w})$. Observe that if $\widetilde{C}(s; T, S_0, \mathbf{q}, \mathbf{w}) \in A(T)$ and $C(s,t)$ is its related contaminant solution function then, from the definition (16), it follows that T is not necessarily the exact time when $C(s,t)$ achieves a decontaminate state. In other words, if $T' < T$ and $A(T')$ is the set of the acceptable contaminant states at time T', it is also possible that $C(s,T') = \widetilde{C}(s; T', S_0, \mathbf{q}, \mathbf{w}) \in A(T')$. On the other hand, at time T, the set $A(T)$ could be empty if, for example, the time T is excessively short.

Now, we define the set whose elements are all the *acceptable contaminant states up to time T*:

$$\mathbf{A}(T) = \left\{ \bigcup_{0 \leq \tau \leq T} A(\tau) \right\} \tag{17}$$

In what follows we shortly denote the elements of $\mathbf{A}(T)$ as $C_A(s)$. Finally, for each initial condition of the system S_0, we can define the set of the control variables which lead the system to an acceptable contaminant state within a time T:

$$\Gamma(S_0, T) = \left\{ (\mathbf{q}, \mathbf{w}) \in \mathscr{Q} \times W_c \quad \middle| \quad \text{exists} \quad C_A(s) \in \mathbf{A}(T) \right\} \tag{18}$$

If the set $\mathbf{A}(T)$ is empty the same is for $\Gamma(S_0, T)$.

2.2 Cost and Time Optimization

We suppose that the total cost of the decontamination intervention consists of the costs of realizing wells, of the cost of the air pumping equipment and of the cost of the total volume of the used pumped air.

If we consider that the control variables of the system are $\mathbf{q}$ and $\mathbf{w}$ we have that, starting from a fixed initial state S_0, for each decontaminate state $C_A(s) \in \mathbf{A}(T)$ the cost function depends on the values $(\mathbf{q}, \mathbf{w}) \in \Gamma(S_0, T)$ associated with $C_A(s)$.

Then, if N is the maximum number of the active wells, the cost function is:

$$J(\mathbf{q}, \mathbf{w}) = m_e + Nm_w + m_u \sum_{i=1}^{N} \int_0^T |q_i(t)| dt \tag{19}$$

where m_e is the expense for the pumping equipment, m_w is the cost to realize a single well and m_u is the cost to pump a unitary air volume. The integral in (19) involves the absolute value of q_i since we suppose that m_u is the same for positive or negative values of $q_i(t)$. Finally, if we assume that the goal of the control problem is to minimize the total decontamination cost, we have to find:

$$\min_{(\mathbf{q},\mathbf{w}) \in \Gamma(S_0,T)} J(\mathbf{q}, \mathbf{w}) \tag{20}$$

Now the problem of seeking the minimal intervention time will be defined. Consider a value of time T and suppose that the set $\mathbf{A}(T)$ is not empty. From definition (18) it follows that for each pair of control variables $(\mathbf{q}, \mathbf{w}) \in \Gamma(S_0, T)$ there is at least one element $C_A(s) \in \mathbf{A}(T)$. Moreover, $C_A(s)$ comes from a solution $C(s, t)$ of the system (1)–(9) and, as shown at about the end of Sect. 2.1, the element $C_A(s)$ may not be the only element in $\mathbf{A}(T)$ related with the same function $C(s, t)$.

Then, for each $(\mathbf{q}, \mathbf{w}) \in \Gamma(S_0, T)$ and $T > 0$, it is possible to define the following set of decontamination times:

$$\theta(\mathbf{q}, \mathbf{w}, T) = \left\{ 0 \leq \tau \leq T \,\middle|\, \text{exists } C_A(s) \in \mathbf{A}(T) \text{ such that } C_A(s) = \widetilde{C}(s; \tau, S_0, \mathbf{q}, \mathbf{w}) \right\} \tag{21}$$

Moreover, for each $(\mathbf{q}, \mathbf{w}) \in \Gamma(S_0, T)$, the function $\Theta(\mathbf{q}, \mathbf{w}) = \inf \theta(\mathbf{q}, \mathbf{w}, T)$ can be defined. Therefore, the minimal intervention time can be found by seeking:

$$\min_{(\mathbf{q},\mathbf{w}) \in \Gamma(S_0,T)} \Theta(\mathbf{q}, \mathbf{w}) \tag{22}$$

3 Conclusion

In this paper a mathematical model describing a bioremediation system has been reported in order to consider the optimal design problem.

For the bioventing, the control variables have been identified, a cost function and the related cost optimal control problem have been mathematically defined.

Moreover, another optimization criterion has been described, that is the minimization of the remediation intervention time having available some fixed limited technical resources.

The solutions of the two optimization problems described in this paper appear to be difficult from a mathematical and computational point of view. Further developments may consist of easier to manage optimization procedures or of dividing the two procedures into sub steps which are easier to approach.

References

1. Bear, J.: Dynamics of Fluids in Porous Media. Elsevier, New York (1972)
2. Bressan, A., Piccoli, B.: Introduction to the Mathematical Theory of Control. American Institute of Mathematical Sciences, Springfield (2007)
3. Cookson, J.J.: Bioremediation Engineering: Design and Application. Mc GrawHill, New York (1995)
4. EPA: Bioventing Principles and Practice. Vol 1: Bioventing Principles (EPA/540/R-95/534a). United States Environmental Protection Agency, Office of Research and Development, Washington (1995)
5. Helmig, R.: Multi phase Flow and Transport Processes in the Subsurface. Springer, Berlin (1997)
6. Khan, F., Husain, T., Hejazi, R.: An overview and analysis of site remediation technologies. J. Environ. Manage. **71**(2), 95–122 (2004)
7. Notarnicola, F.: Subsoil decontamination with biological techniques: a bio-fluid dynamics problem. In: Aletti, G., Burger, M., Micheletti, A., Morale, D. (eds.) Math Everywhere: Deterministic and Stochastic Modelling in Biomedicine, Economics and Industry, pp. 265–276. Springer, Berlin (2007)
8. Rathfelder, K., Lang, J., Abriola, L.: A numerical model (MISER) for the simulation of coupled physical, chemical and biological processes in soil vapor extraction and bioventing systems. J. Contam. Hydrol. **43**(3–4), 239–270 (2000)

Numerical Simulation of Heat Transfer in Underground Electrical Cables

R. Čiegis, G. Jankevičiūtė, A. Bugajev, and N. Tumanova

Abstract The aim of this project is to develop a virtual modelling tool which can be used to construct optimal design of power transmission lines and cables. They should meet the latest power transmission network technical and economical requirements. The mathematical model is based on a general heat conduction equation describing the diffusion, convection and radiation processes. We take into account a linear dependence of the resistance on temperature. The velocity of convective transport of the heat in air regions is obtained by solving a coupled thermoconvection problem including the heat conduction problem and a standard Navier-Stokes model of the flow in air. The changes of material coefficients in soil due to influence of heating are taken by solving a simplified mass balance equation for flows in porous media. The FVM is used to solve the obtained system of differential equations. Discretization of the domain is done by applying "aCute" mesh generator, which is a modification of the well-known Triangle mesh generator. The discrete schemes are implemented by using the OpenFOAM tool. Parallel versions of basic algorithms are also investigated. Results of computational experiments of simulation of real industrial underground cables are presented.

Keywords Heat transfer • Optimal design • Power transmission lines and cables

1 Introduction

This research is aimed to develop design rules for power transmission lines and cables, which have to meet the latest power transmission network technical and economical requirements. At present the power lines are over-dimensioned by up to 60 % in terms of transmitted power. However, today, as the new distributed generating capacities are installed e.g. large wind farms, bio-gas plants or waist-to-energy plants, the infrastructure of power grid must be re-designed or new optimization strategies for the available grid developed. Power cables for power

R. Čiegis (✉) • G. Jankevičiūtė • A. Bugajev • N. Tumanova
Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
e-mail: rc@vgtu.lt; gerda.jankeviciute@vgtu.lt; andrej.bugajev@vgtu.lt; natalija.tumanova@vgtu.lt

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_154

distribution applications are still rated according to IEC 287 and IEC 853 standards, which use the Neher and McGrath methods from 1957 [1, 10], relaying on the thermal equivalent circuit technique. However, today there are many applications where analytical and heuristic formulas cannot describe precisely enough the conditions under which the cables are installed. An example could be an underground cable route, where the installation conditions for a cable are different only for a short distance including only a short crossing of the road. The present standards require that the cable's current-carrying capacity must be reduced due the worst case conditions. Today the cost effective designing of cable installations comes us an urgent need, since the copper price level has reached its maximum since decades. At the same time the safety of the design is also must be guaranteed. Thus a direct simulation of cables including all specific conditions under which the cables are operating is needed. Some approaches of direct simulation are presented in [7, 8], where finite element and finite volume methods are applied for numerical simulation of underground cables.

The knowledge of dynamics (in time) of heat distribution in/around electrical cables is necessary to optimize the usage of electricity transferring infrastructure. It is important to determine various parameters of the cable networks: maximal electric current for the cable; optimal cable parameters in certain circumstances; cable life expectancy and many other engineering factors.

2 Mathematical Models

A good review on mathematical models for simulation of heat transfer in underground and overhead electrical cables is presented in [8]. Here we will describe the most important details on high voltage cables and their installation conditions. Figure 1 illustrates a typical high voltage cable, consisting of conductor (most often copper or aluminum), conductor screen, insulation layer (cross-linked polyethylene), insulation screen, metallic shielding (copper tape/wire, aluminum/lead sheath) and outer covering (polyvinyl chloride, polyethylene, nylon).

Various installation environments are considered in applications. Our main interest is to simulate the following environments: (a) cables in air, (b) cables directly buried, (c) cables in pipe and the pipe is directly buried, (d) cable groups (duct banks).

There are two possible arrangements of cables in the installation: (a) single core arrangement and (b) three-core arrangement. In each case a group of cables can be arranged horizontally or vertically (see, Fig. 2).

In recent years a new interesting technology is applied to improve the cooling of high-voltage cables [12]. The forced cooling systems are developed employing general principle: solid body or other medium is placed near cable or it's conductor resulting the abstraction of excess heat from the conductor. This can be implemented

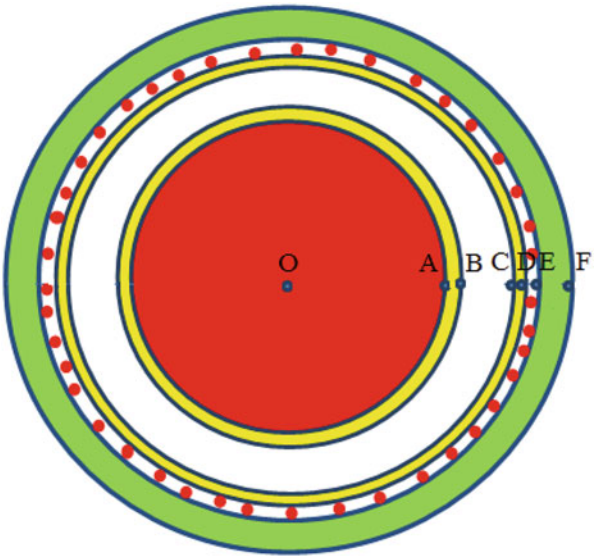


Fig. 1 A typical structure of high-voltage (110 kV) cables consisting of six main layers

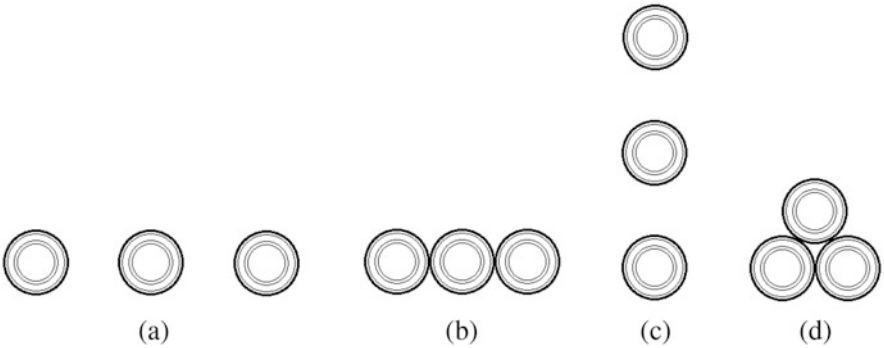


Fig. 2 Examples of cables layout topologies: (**a**, **b**) three single core cables are arranged horizontally; (**c**) three single core cables are arranged vertically; (**d**) three-core cable

by the following means:

1. Copper or aluminum slab is put under or over the cables. These metals are characterized by high thermal conductivity coefficient, thus help to avoid overheating, especially in case of transient current increase.
2. Pipes with water are arranged around the cables. Water is pumped and it's velocity is adjustable.
3. Special duct with fluid (oil) flowing along the cable is installed in the cable itself. The fluid serves for the abstraction of excess heat from the conductor.

Now we will describe basic mathematical models which are used in our tool for simulation of heat transfer in different environments.

2.1 Heat Transfer in Solids

Electric current flowing through cable generates heat. It's distribution in time and space is described by heat equation in heterogeneous medium. Since the bungle of cables consisting of metal cores and various insulation layers is placed in a medium, we assume that the coefficients of thermal conductivity, material density and specific heat capacity are discontinuous piece-constant functions depending on spatial coordinates.

Since the heat transfer mechanism in the underground electrical cables is quite complicated and many processes can be described only approximately, the structure of cables also can be simplified. For most cases it is enough to take one slice of isolation material. The length of a cable (or many cables) is much bigger than its diameter, thus effects along the cable's length can be neglected. A sand or soil area is much bigger than cables' area, so two-dimensional models are sufficient for the analysis. However, properties of different filler should be taken into account, we consider sand, wet and dry soil and other materials.

For heat transfer in underground cables we assume the diffusion to be the main transfer mechanism. A mathematical model of the heat source is described by the Joule–Lenz law. Then the mathematical model of non-stationary heat-transfer is given by the parabolic differential equation [2, 3]:

$$\begin{cases} c\rho\dfrac{\partial T}{\partial t} = \nabla\cdot(\lambda\nabla T) + q_0(1+\alpha(T-T^*))I^2, & t\in[0,t_{max}], x\in\Omega, \\ T(x,0) = T_0, & \text{when } x\in\Omega, \\ T \text{ and } \lambda\nabla T \text{ are continuous}, & \text{when } x\in\Omega, \end{cases} \tag{1}$$

here $x = (x_1, x_2)$, $T(x,t)$ is temperature in the Kelvin scale, $\lambda(x) > 0$ is the heat conductivity coefficient, $q(x,t,T)$—the source function. $\rho(x) > 0$ is the mass density, $c(x) > 0$ is the specific heat capacity, $\nabla\cdot(\lambda\nabla T)$ defines the diffusion operator, T_0 is the initial temperature. We take into account a linear dependence of the resistance on temperature, T^* is the reference temperature and I is the electrical current. Due to different properties of materials included into the model, coefficients λ, c, ρ are discontinuous and their values may vary 1000 times. That makes the simulation task very challenging.

Various boundary conditions are applied to describe the heat flow through boundaries of the domain:

- $T(x,t) = T_{b1}$ for the upper boundary, $T(x,t) = T_{b2}$ for the lower boundary, $\frac{\partial T}{\partial x_1} = 0$ for the left and right boundaries are applied for modelling boundary conditions in winter ($T_{b1} = T_{b2}$) or summer ($T_{b1} > T_{b2}$).

- Boundary conditions of the third type $\lambda \frac{\partial T}{\partial x_2} = \alpha(T(x,t) - T_{air})$ are applied on the upper boundary to evaluate the cooling effects of wind in air.

Numerical approximation of various boundary conditions is considered in [4].

2.2 Cable in Pipe

Figure 3 shows a single cable placed at the bottom of a plastic pipe buried in the soil. The plastic pipe is placed in the center of the soil domain. Due to its relatively low heat conduction coefficient a plastic pipe represents a significant thermal resistance for the cooling of the cable. The main heat transfer mechanism in air is described by air circulation inside the plastic pipe. Velocities of the free convection process are computed by solving the Navier-Stokes equations (the conservation of continuity, momentum and energy equations) in the air area. The air flow is usually modelled as compressible non-reacting fluid. In order to the mathematical model simpler, we assume the air to be incompressible for small velocities (less than 80–100 m/s). Assuming the velocities are not large, the model of laminar incompressible flow is given by the following system of equations [2]:

$$\nabla \cdot \mathbf{u} = 0, \tag{2}$$

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho \mathbf{u} \nabla \cdot \mathbf{u} - \nabla \cdot (\eta \nabla \mathbf{u}) = -\nabla p - \rho \alpha \mathbf{g}(T - T_0), \tag{3}$$

$$\rho c \Big(\frac{\partial T}{\partial t} + \nabla \cdot (\mathbf{u} T) \Big) - \nabla \cdot (\lambda \nabla T) = q, \tag{4}$$

where $\rho(x) > 0$ is the density of material in particular area, $\mathbf{u}(x,t)$ is velocity of the flow, p is the pressure, η is the dynamic viscosity, α is the thermal expansion

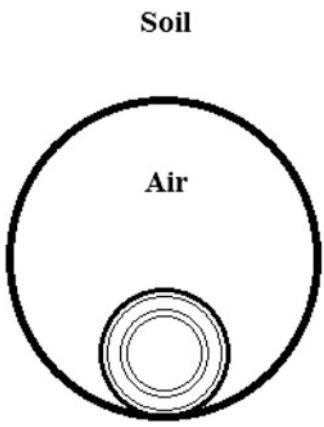


Fig. 3 A single cable in directly buried pipe

coefficient. The heat conduction equation (4) is extended to the whole domain. In the soil, pipe or cable area term $\nabla \cdot (\mathbf{u}T)$ is zero, resulting to the non-stationary problem (1).

2.3 Heat Transfer in Soil

In most papers the soil is considered as solid material. Here we simulate a full model of heat and water transfer in variable saturated soil. The porous medium is considered to be rigid and unsaturated, hence two phases are present: liquid (water) and gas (water vapor and air). The temperature of the solid, liquid and gas phases are considered to be in local thermal equilibrium. The Richards equation is obtained from the mass conservation equation for liquid phase and Darcy law [6]:

$$\varepsilon \frac{\partial \rho_w S^l}{\partial t} + \nabla \cdot \left(-\rho_w \mathbf{K} \frac{k^l_{rel}(S^l)}{\mu_w} \big(-\nabla p^c(S^l) - \rho_w \mathbf{g} \big) \right) = 0, \tag{5}$$

where ε is porosity of the porous medium, ρ_w is density of the water, S^l is saturation of the liquid phase, p^c is capillary pressure, $\mathbf{K}$ is intrinsic permeability tensor of the porous medium, μ_w is viscosity of the water, $\mathbf{g}$ is vector of gravitational acceleration. In general, the capillary pressure also depends on the temperature $p^c = p^c(S^l, T)$.

For the description of the heat transfer, the energy conservation is used

$$(\rho c)_{eff} \frac{\partial T}{\partial t} + \left(\rho_w \mathbf{K} \frac{k^l_{rel}(S^l)}{\mu_w} \big(\nabla p^c(S^l) + \rho_w \mathbf{g} \big) \right) \cdot \nabla T = \nabla \cdot \big(\lambda_{eff} \nabla T \big) + q, \tag{6}$$

where $(\rho c)_{eff}$ is the effective heat capacity, λ_{eff} is the effective heat conductivity. The soil-atmosphere interface is an important boundary condition affecting subsurface movement of liquid water and heat under field conditions.

3 Numerical Approximations

The obtained systems of PDEs are solved by using Finite Volume Method. For implementation of constructed discrete schemes we have used OpenFOAM (Open source Field Operation And Manipulation) tool [11]. It is a C++ toolbox (library) targeted for the development of customized numerical solvers for partial differential equations. Since mathematical models (1)–(6) consists of different models in different regions, application of OpenFOAM functionality requires some non-trivial modifications of basic models presented in the library of cases. For example in order to approximate heat conduction equation in solid medium (1), two sub-tasks should be solved accurately. First, the numerical fluxes of the discrete solution must be orthogonal to the boundary of finite volumes. In our solver this problem is solved by

using a proper Delaunay triangulation of the domain. Discretization of the domain is done by applying "aCute" mesh generation tool, which is modification of the well-known Triangle mesh generator. Second, a special interpolation should be used for definition of discontinuous coefficients λ in diffusion term, namely *harmonic* version of the interpolation formula.

The given model requires to solve multi-physics problems in different subregions. The robust, efficient and accurate numerical simulation of such processes makes big challenges for selection of appropriate mathematical and numerical methods, and for software implementation part of the project. Since we are mostly interested in solving non-stationary problems, the so-called loosely coupled schemes are used for approximation systems of multi-physics problems. Alternatives can be to use solvers based on fixed-point iterations or to implement monolithic solvers based on implicit time-approximation schemes. A good review on monolithically-coupled numerical algorithms is given in [9].

Parallelization in OpenFOAM is robust and implemented at a low level using MPI library. Solvers are built using high level objects and, in general, don't require any parallel-specific coding. They will run in parallel automatically. Thus there is no need for users to implement standard steps of any parallel code: decomposition of the problem into subproblems, distribution of these tasks among different processes, implementation of data communication methods. OpenFOAM employs a common approach for parallelization of numerical algorithms—domain decomposition. The mesh and its associated fields are split into sub-domains, which are allocated to different processes. Results of computational experiments with parallel version of the solver are presented in [5]. There was tested the parallel performance of the conjugate gradient solver with diagonal incomplete Cholesky preconditioner (DIC/CG) and generalized geometric-algebraic multigrid (GAMG) solver. GAMG showed better times, however DIC/CG linear solver is to be recommended for the parallel computations on parallel computing systems with large number of processors and cores. The weighting factors (supported by OpenFOAM) in mesh partitioning algorithm allow efficient utilization of heterogeneous computing nodes for our parallel application.

4 Computational Experiments

Here we present results of simulation of heat transfer in one electrical cable which is placed at the bottom of a plastic pipe buried in the soil. We have simulated the heat transfer during three summer months, when the soil is assumed to be semi-dry, boundary conditions are taken $T = 293\,\mathrm{K}$, and the electrical current is equal to 470 A. Thus the solution practically reached a stationary phase. Figure 4 shows a distribution of temperature and a velocity field for a single cable placed at the bottom of a plastic pipe buried in the soil.

Figure 5 shows a distribution of temperature for different arrangements of trefoil cable under the same conditions.

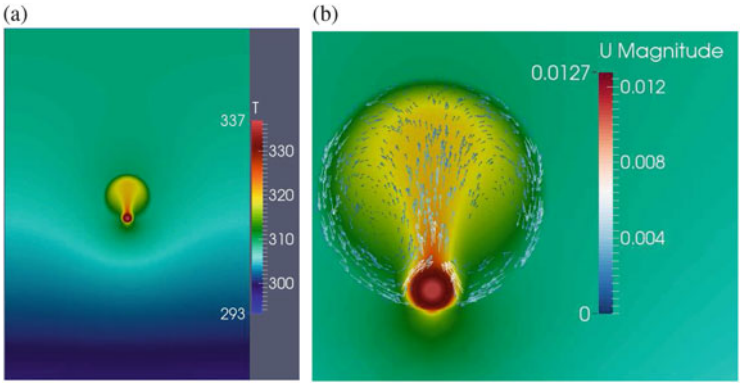


Fig. 4 Simulation results for a single cable in directly buried pipe: (**a**) distribution of temperature, (**b**) velocity field

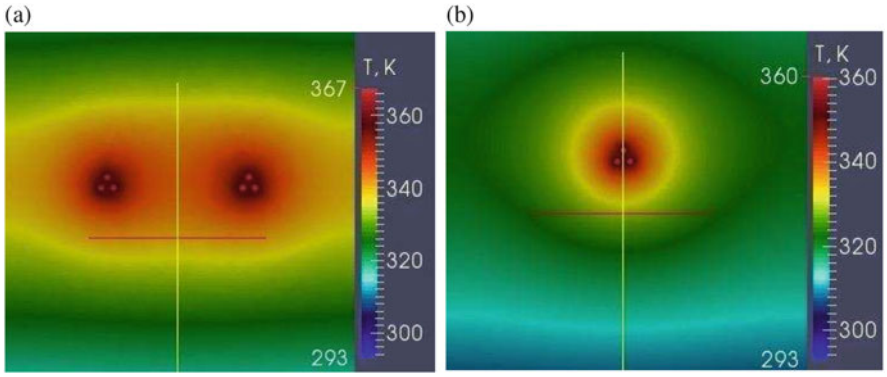


Fig. 5 Simulation results for trefoil cables: (**a**) electrical current $I = 609$ A, (**b**) $I = 675$ A

References

1. Anders, G.: Rating of Electric Power Cables in Unfavorable Thermal Environment. IEEE Press Series on Power Engineering. Wiley, New York (2005)
2. Bergman, T., Incropera, F.: Fundamentals of Heat and Mass Transfer. Wiley, New York (2011)
3. Čiegis, R., Ilgevičius, A., Liess, H., Meilūnas, M., Suboč, O.: Numerical simulation of the heat conduction in electrical cables. Math. Model. Anal. **12**(4), 425–439 (2007)
4. Čiegis, R., Jankevičiūtė, G., Suboč, O.: Numerical simulation of the heat conduction in composite materials. Math. Model. Anal. **15**(1), 9–22 (2010)
5. Čiegis, R., Starikovičius, V., Bugajev, A.: On efficiency of the OpenFOAM-based parallel solver for the heat transfer in electrical power cables. In: Proceedings of the First International Workshop on Sustainable Ultrascale Computing Systems (NESUS 2014), August 27–28, Porto, Portugal, 2014, pp. 43–46. Computer Architecture, Communications, and Systems Group (ARCOS), Madrid (2014)
6. Izumi, T.: Inverse modeling of variably saturated subsurfaces water flow in isotermal/non-isothermal soil. Mem. Fac. Agr., Ehime Univ. **57**, 1–36 (2012)

7. Karahan, M., Kalenderli, Ö.: Coupled electrical and thermal analysis of power cables using finite element method. In: Vikhrenko, P.V. (ed.) Heat Transfer - Engineering Applications. InTech, Rijeka (2011). doi: 10.5772/27350
8. Makhkamova, I.: Numerical investigations of the thermal state of overhead lines and underground cables in distribution networks. Ph.D. Thesis, School of Engineering and Computing Sciences of Durham University (2010)
9. Muddle, R., Mihajlovic, M., Heil, M.: An efficient preconditioner for monolithically-coupled large-displacement fluid-structure interaction problems with pseudo-solid mesh updates. J. Comput. Phys. **231**, 7315–7334 (2012)
10. Neher, J., McGrath, M.: The calculation of the temperature rise and load capability of cable systems. AIEE Trans. Part III **76**, 752–772 (1957)
11. Openfoam. URL http://www.openfoam.org
12. Rizk, F., Trinh, G.: High Voltage Engineering. Taylor & Francis, London (2014)

Numerical Study of Forced MHD Convection Flow and Temperature Around Periodically Placed Cylinders

Harijs Kalis and Maksims Marinaki

Abstract In this paper we consider 2D stationary boundary value problems for the system of magnetohydrodynamic (MHD) equations and the heat transfer equation. The viscous electrically conducting incompressible liquid moves between infinite cylinders with square or round sections placed periodically. We also consider similar 2D MHD channel flow with periodically placed obstacles on the channel walls. We analyse the 2D forced and free MHD convection flow and temperature around cylinders and obstacles in homogeneous external magnetic field. The cylinders, obstacles and walls of the channel with constant temperature are heated. The distributions of electromagnetic fields, forces, velocity and temperature fields have been calculated using the method of finite differences.

The goal of such investigation is to obtain the distributions of stream function, temperature, velocity and the vortex formation in the plane of the cross-section between the cylinders and obstacles as function of the external magnetic field and of the direction of the gravitation.

Keywords Electroconductive liquid • Heat transfer • Magnetohydrodynamic convection flow

1 Introduction

In many physical experiments and technological applications it is important to mix and heat an electroconductive liquid: liquid-metals (steel, mercury, lithium), liquid magnetic materials, electrolyte, water, air. Liquid metals are considered to be the most promising coolants for high temperature applications, like nuclear fusion reactors, because of the inherent high thermal diffusivity, thermal conductivity and hence, excellent heat transfer characteristics.

In the developed mathematical models vortex-type structures appear in liquid flows, as well as in problems related to energy conversion in new technological

H. Kalis (✉) • M. Marinaki
Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia
e-mail: kalis@lanet.lv; maksims.marinaki@lu.lv

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_155

devices. MHD convection flow of a viscous incompressible fluid around cylinder with combined effects of heat and mass transfer is an important problem prevalent in many engineering applications. These types of problems find their application in nuclear reactor cooling systems and energy transport systems.

Heat exchanger systems are employed in numerous industries. Steam generation in boiler, air cooling within the coil of the conditioner and automotive radiators represent just some of the conventional applications of this mechanical system. For the in-line arrangements of tube banks (cylinders) a fluid at prescribed mass flow rate of velocity U_0 and an inlet ambient temperature T_0 much lower than the wall temperature T_w enters the cylinders from the left and exits at the right. By taking advantage of special geometrical features, such as the inherent repetitive nature of the flow behaviour, the computational fluid domain allows the possible exploitation of symmetric and periodic boundary conditions in speeding up the computations and in turn enhancing the computational accuracy of the simplified geometries. Using the conditions of symmetry and periodicity we can consider only two cylinders. The heat transfer significant influence on the fluid flow behaviour without the magnetic field is investigated in [3]. We consider the viscous electrically conducting incompressible liquid. The liquid moves on the plane (x, y) in the Ox-axis direction between infinite cylinders (tube banks) placed periodically in the (x, y) plane. The cross-section of the cylinders are square or circle. We consider 2D stationary boundary value problems for the system of magnetohydrodynamic (MHD) equation. We analyse the 2D flow and temperature around these cylinders in homogeneous external magnetics field, depending on direction of the gravitation.

This process of the magnetohydrodynamics (MHD) is considered with the so-called inductionless approximation. This would mean that the action of a moving liquid on the external magnetic field can be neglected [2].

The external magnetic field, Lorentz force, dimensionless stationary Navier-Stokes equations, numerical domain with two cylinders and the system of three equations for calculating the stream function, vorticity and temperature are defined.

The solution of the problem is obtained using the method of finite differences, Gauss-Seidel iterations and specific boundary conditions for vorticity function. Some numerical results are analysed.

2 Mathematical Model

The magnetic field creates the $F_x(t, x, y), F_y(t, x, y)$ components of the Lorentz force $\mathbf{F}$.

From the vector of Lorentz force $\mathbf{F} = \mathbf{J} \times \mathbf{B}, \mathbf{J} = \boldsymbol{\sigma}(\mathbf{E} + \mathbf{V} \times \mathbf{B})$ for the 2D magnetic field we can obtain

$$J_z = \sigma(V_x B_y - V_y B_x + E_z), \quad F_x = -B_y J_z, \quad F_y = B_x J_z,$$

where $E_z = const$, J_z are the azimuthal components of the electric field vector $\mathbf{E}$ and the density vector of the electric current $\mathbf{J}$, B_x, B_y are the components of the

magnetic induction vector $\mathbf{B}$ for the homogeneous magnetic field, σ is the electric conductivity, V_x, V_y are the components of the velocity vector $\mathbf{V}$.

The external homogeneous 2D magnetic field has two components of the induction in the following dimensionless form
$B_x = B_0\cos(\alpha), B_y = B_0\sin(\alpha)$, where α is the angle between the Ox-axis and the direction of the induction vector, B_0 is the magnitude of magnetic field. Then $J_z = \sigma(E_z + B_0(V_x\sin(\alpha) - V_y\cos(\alpha)))$.

We analyse the flow depending on two types of homogeneous magnetic field: the field parallel to Ox-axis ($\alpha = 0$) and transverse field ($\alpha = \frac{\pi}{2}$).

Using the vorticity function $\zeta = \frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y}$, one obtains

$$\begin{cases} \frac{\partial V_x}{\partial t} - \zeta V_y = -\frac{\partial \bar{p}}{\partial x} - \frac{1}{Re}\frac{\partial \zeta}{\partial y} + \frac{Gr}{Re^2}T\sin(\beta) - S\sin(\alpha)j_z, \\ \frac{\partial V_y}{\partial t} + \zeta V_x = -\frac{\partial \bar{p}}{\partial y} + \frac{1}{Re}\frac{\partial \zeta}{\partial x} - \frac{Gr}{Re^2}T\cos(\beta) + S\cos(\alpha)j_z, \\ \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} = 0, \\ \frac{\partial T}{\partial t} + V_x\frac{\partial T}{\partial x} + V_y\frac{\partial T}{\partial y} = \frac{1}{Pe}\nabla^2 T + \frac{K_T}{Pe}j_z^2, \end{cases} \tag{1}$$

where $j_z = e_z + V_x\sin(\alpha) - V_y\cos(\alpha)$, e_z are the dimensionless forms of the azimuthal components for the electric current density and the electric field, $\bar{p} = p + 0.5\mathbf{V}^2$, $Re = \frac{U_0 L_0}{\nu}$, $S = \frac{\sigma B_0^2 L_0}{\rho U_0}$, $Gr = \frac{\beta_t g(T_w - T_0)L_0^3}{\nu^2}$ are Reynolds, Stewart and Grashof numbers, $Pe = PrRe, Pr = \frac{\nu\rho C_p}{k}$, $K_T = \frac{\sigma B_0^2 L_0^2 U_0^2}{kT_w - T_0}$ are Prandtl number and heat source parameters. The parameter K_T is considered as negligible.

The hydrodynamical stream function ψ can be determined via formulas
$V_x = \frac{\partial \psi}{\partial y}, V_y = -\frac{\partial \psi}{\partial x}$. By eliminating the pressure $\bar{p}$ from the system (1) one obtains

$$\begin{cases} \frac{\partial \zeta}{\partial t} - J(\psi, \zeta) = \frac{1}{Re}\nabla^2\zeta - \frac{Gr}{Re^2}(\frac{\partial T}{\partial x}\cos(\beta) + \frac{\partial T}{\partial y}\sin(\beta)) + Sf, \\ \zeta = -\nabla^2\psi, \\ \frac{\partial T}{\partial t} - J(\psi, T) = \frac{1}{Pe}\nabla^2 T, \end{cases} \tag{2}$$

where $f = \sin(2\alpha)\frac{\partial^2\psi}{\partial x\partial y} + \cos^2(\alpha)\frac{\partial^2\psi}{\partial x^2} + \sin^2(\alpha)\frac{\partial^2\psi}{\partial y^2}$ is the z-component of the vector $curl\mathbf{F}$,

$J(\psi, v) = \frac{\partial \psi}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial \psi}{\partial y}\frac{\partial v}{\partial x}$ is the Jacobian of the functions ψ, v, $v = \zeta; T$, ∇^2 is the Laplace operator.

Using the boundary conditions (BCs) of symmetry and periodicity we can consider the domain that contains only quarters of two or four picked out cylinders.

The periodically placed cylinders (PC) are **arranged in the parallel series**. We consider the domain $\Omega = \Omega_1 \bigcup \Omega_2$ (see Figs. 1 and 2), where
$\Omega_1 = \{(x, y) : l_1 \le x \le l_2, 0 \le y \le L_1\}$, $\Omega_2 = \{(x, y) : 0 \le x \le l, L_1 \le y \le L\}$,
$0 < l_1 < l_2 < l, 0 < L_1 < L$.
Here $C_1 = \{(x, y) : 0 < x < l_1, 0 < y < L_1\}$ and $C_2 = \{(x, y) : l_2 < x < l, 0 < y < L_1\}$ are quarters of cylinders,
$L^1 = \{(x, L) : 0 \le x \le l\}, L^2 = \{(x, 0) : l_1 \le x \le l_2\}$ are the plane of symmetry with BCs $V_y = 0, \frac{\partial T}{\partial y} = \zeta = 0, \psi = \psi_0$ on L^1, $\psi = 0$ on L^2,

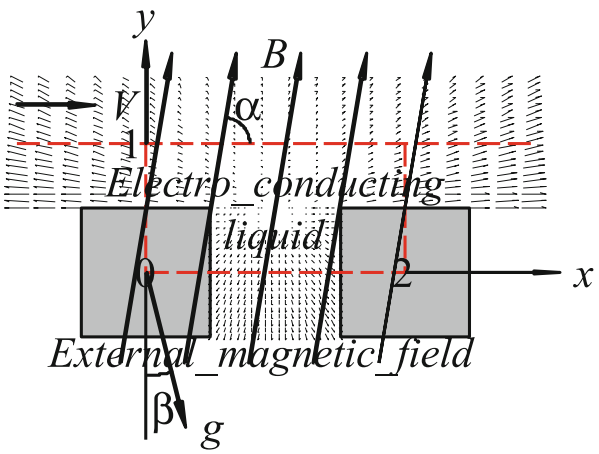


Fig. 1 Domain for parallel placed cylinders (two cylinders, $L_1 = 0.5, L = 1, l_1 = 0.5, l_2 = 1.5, l = 2$)

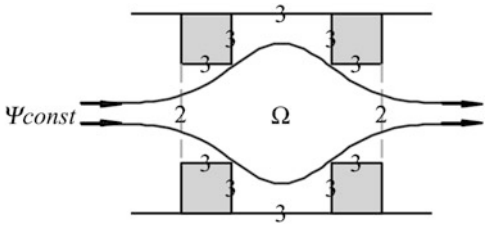


Fig. 2 Domain for channel flow (quarters of four cylinders, $L_1 = 0.5, L_2 = 1.5, L = 2, l_1 = 0.5, l_2 = 1.5, l = 2$)

$W^1 = \{(x, L_1) : 0 < x \le l_1\}, W^2 = \{(x, L_1) : l_2 \le x < l\}, W^3 = \{(l_1, y) : 0 < y \le L_1\}$ and $W^4 = \{(l_2, y) : 0 < y \le L_1\}$ are the walls of the cylinders with the non-slip BCs $T = 1, V_x = V_y = \psi = 0$,
$I_n = \{(0, y) : L_1 < y \le L\}$ is the inlet and $O_t = \{(l, y) : L_1 < y \le L\}$ is the outlet with the periodical BCs for ψ, ζ, T, U_x, U_y.
In case of free convection $\psi_0 = 0$.
For the **additional channel flow with symmetry** (CFS),
$L^2 = W^5 = \{(x, 0) : l_1 \le x \le l_2\}$ is the wall of the half-channel Ω with the non-slip BCs $T = 1, V_x = V_y = \psi = 0$.

3 Numerical Algorithm for Solving the Problem

We consider a uniform square grid $((N + 1) \times M)$:

(1) $\Omega_1^h = \{(x_i, y_j), x_i = (i-1)h, y_j = (j-1)h, \}, i = \overline{N_1, N_2}, j = \overline{1, M_1}$, $(N_1 - 1)h = l_1, (M_1 - 1)h = L_1$.
(2) $\Omega_2^h = \{(x_i, y_j), x_i = (i-1)h, y_j = (j-1)h, \}, i = \overline{1, N+1}, j = \overline{M_1, M}$, $(N - 1)h = l_1, (M - 1)h = L$,

where $h = \frac{l_1}{N_1-1} = \frac{l_2}{N_2-1} = \frac{l}{N-1} = \frac{L}{M-1} = \frac{L_1}{M_1-1}$. In the following any unknown function $q(x, y)$ is approximated by the grid function taking values $q_{i,j} \approx q(x_i, y_j)$.

The Eq. (2) are made stationary and in the uniform grid (x_i, y_j) are replaced with difference equations of second order approximation in 5-point stencil and the numerical calculations are carried out by using Gauss-Seidel iterations with under-relaxation for vorticity and temperature.

The difference equations of second order approximation in 5-point stencil are in the following form:

$$\begin{cases} 4\Psi_{i,j} = \bar{S}\Psi_{i,j} + h^2\zeta_{i,j}, \\ \frac{Re}{4}J_{i,j}(\Psi,\zeta) - 4\zeta_{i,j} + \bar{S}\zeta_{i,j} = \\ = Ha^2((\sin(\alpha))^2 d_y^2\Psi_{i,j} + (\cos(\alpha))^2 d_x^2\Psi_{i,j} - 0.25\sin(2\alpha)d_{x,y}^2\Psi_{i,j}) + \\ + \frac{Grh}{Re}(\cos(\beta)d_xT_{i,j} + \sin(\beta)d_yT_{i,j}), \\ \frac{Pe}{4}J_{i,j}(\Psi,T) - 4T_{i,j} + \bar{S}T_{i,j} = -K_Th^2(e_z + h^{-1}(\cos(\alpha)d_x\Psi_{i,j} + \sin(\alpha)d_y\Psi_{i,j})^2, \end{cases} \tag{3}$$

where $\bar{S}q_{i,j} = q_{i,j-1} + q_{i,j+1} + q_{i-1,j} + q_{i+1,j},\ q = \Psi; T; \zeta,$
$d_x^2\Psi_{i,j} = 2\Psi_{i,j} - \Psi_{i+1,j} - \Psi_{i-1,j},\ d_y^2\Psi_{i,j} = 2\Psi_{i,j} - \Psi_{i,j+1} - \Psi_{i,j-1},$
$d_{xy}^2\Psi_{i,j} = \Psi_{i+1,j+1} + \Psi_{i-1,j-1} - \Psi_{i-1,j+1} - \Psi_{i+1,j-1},$
$d_xq_{i,j} = 0.5(q_{i+1,j} - q_{i-1,j}),\ d_yq_{i,j} = 0.5(q_{i,j+1} - q_{i,j-1}),\ q = \Psi; T,$
$J_{i,j}(\Psi, q) = (\Psi_{i+1,j} - \Psi_{i-1,j})(q_{i,j+1} - q_{i,j-1}) - (q_{i+1,j} - q_{i-1,j})(\Psi_{i,j+1} - \Psi_{i,j-1}),$
$q = \zeta; T,\ Ha = \sqrt{Re \cdot S}$. is the Hartman number.

The numerical calculations for (3) are carried out by Gauss-Seidel iterations with under-relaxation for ζ, T functions :

$$\zeta_{i,j}^m = \omega_1\zeta_{i,j}^z + (1-\omega_1)\zeta_{i,j}^{m-1}, T_{i,j}^m = \omega_2T_{i,j}^z + (1-\omega_2)T_{i,j}^{m-1}, m = 1, 2, \cdots,$$

where $\zeta_{i,j}^z, T_{i,j}^z$ are the grid functions value in central mesh points, obtained in the m-th iteration, $\omega_1, \omega_2 \in (0, 1)$ are the relaxation coefficients.

The discrete BCs [1] with $O(h^2)$ on the walls w are computed in the following form:

$$\zeta_w^m = \frac{\gamma}{2h}(-4\Psi_{w-1}^m + \Psi_{w-2}^m + 3\Psi_w) + \zeta_w^{m-1},$$

where $\Psi_{w-1}^m, \Psi_{w-2}^m$ are the values of $\Psi_{i,j}$ for one or two step h distance from the wall in the interior normal direction. On the corner of the wall the value of ζ is equal to the average value of two nearest ζ values of the wall.

The velocity components are obtained in the following way:

$$Vx_{i,j} = \frac{1}{h}(\Psi_{i,j+1} - \Psi_{i,j}), Vy_{i,j} = -\frac{1}{h}(\Psi_{i+1,j} - \Psi_{i,j}), V = \sqrt{(V_x)^2 + (V_y)^2}.$$

The dimensionless fluid volumes between the two sections $x = 0$,

$$x = l_* = 0.5(l_1 + l_2) \text{ are } Q_1 = \int_{L_1}^{L} V_x(0, y)dy = Q_2 = \int_{0}^{L} V_x(l_*, y)dy = 1.$$

4 Some Numerical Results

Numerical results were obtained for $l_1 = 0.5, l_2 = 1.5, l = 2, L = 1, L_1 = 0.5$ for flow with symmetry, $Re = 40; 100,\ S = 0; 2.5; 20,\ Pr = 4$ (for electrolyte), $Gr = 0;\ 25{,}000,\ \beta = 0; \pm\frac{\pi}{2}, \alpha = 0; \frac{\pi}{2}, \omega_1, \omega_2 \in [0.1, 0.4]$.
Calculations and their graphical visualization were made by means of the computer program MATLAB for regular grid:

$$h = 0.0125, N_1 = 41, N_2 = 121, N = 161, M_1 = 41, M = 81.$$

We apply the iterations with maximal errors $\leq 10^{-7}$ for Ψ and $\leq 10^{-4}$ for ζ and T (the number of iterations $\in$ [10,000, 100,000], running time ~5–10 min). The convergence might be further improved by considering ADI-type methods. In Figs. 3, 4, 5, 6, 7 and 8 we show the obtained levels of stream function and temperature for different values of parameters.

One can conclude, that for large values of the *Ha* number on the walls the Hartman boundary layers develop and the flow then becomes vortex-free (Fig. 4).

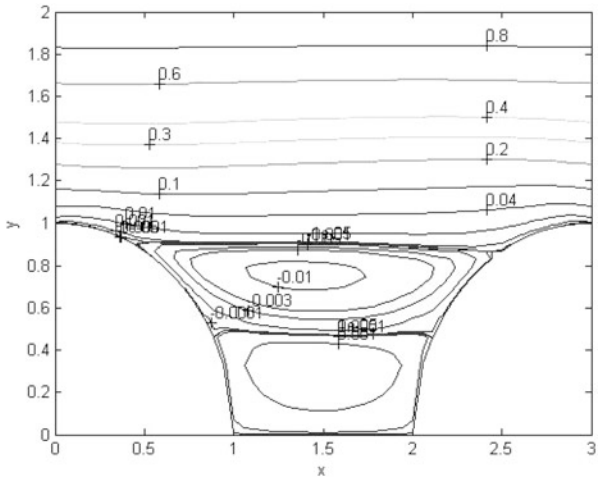


Fig. 3 Levels of stream function in PC at $Re = 100,\ S = 20,\ \alpha = 0,\ Gr = 0$

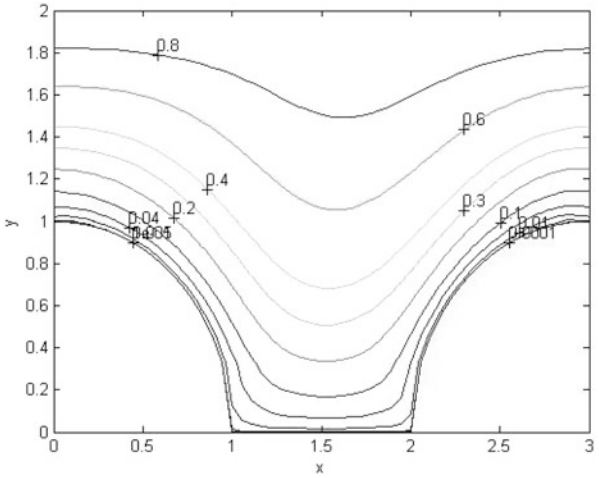


Fig. 4 Levels of stream function in PC at $Re = 100,\ S = 20,\ \alpha = \frac{\pi}{2},\ Gr = 0$

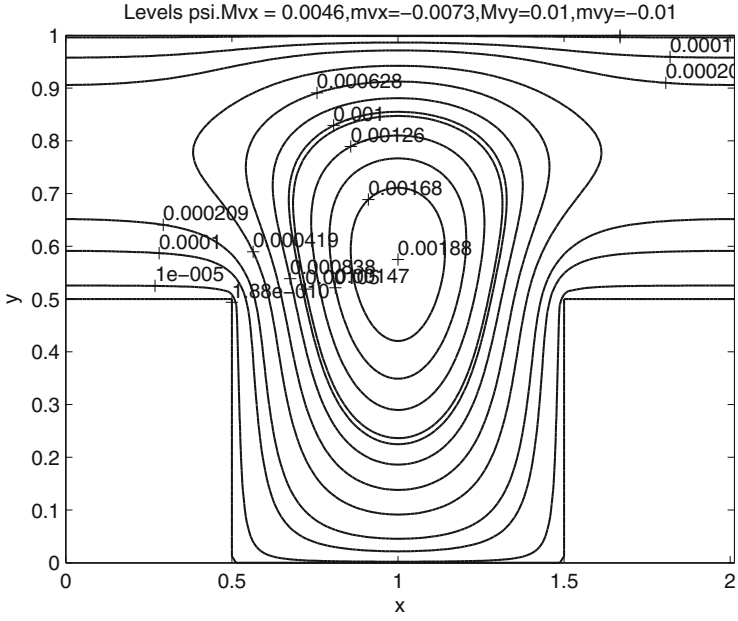


Fig. 5 Levels of stream function in PC for free convection at $Re = 40,\ S = 0,\ Gr = 25{,}000,\ \beta = \frac{\pi}{2}$

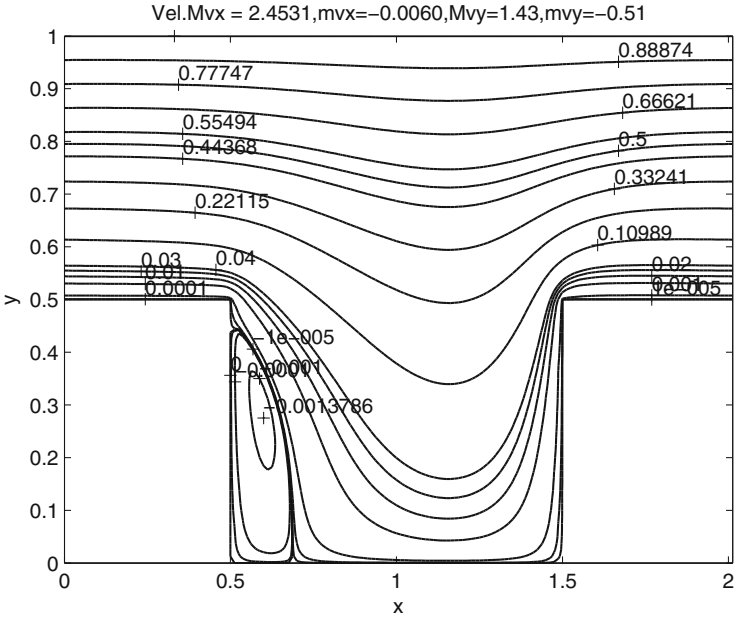


Fig. 6 Levels of stream function in PC at $Re = 40,\ S = 2.5,\ \alpha = \frac{\pi}{2},\ Gr = 0$

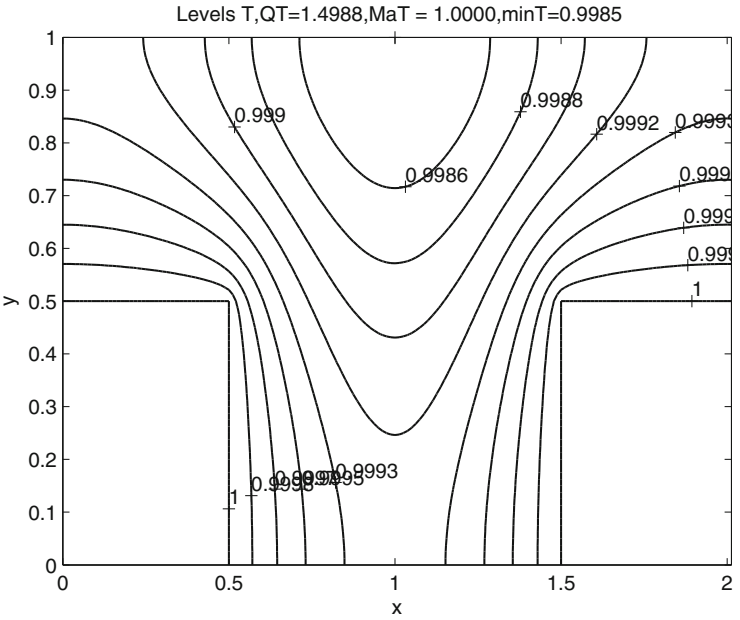


Fig. 7 Levels of temperature in PC for free convection at $Re = 40,\ S = 0,\ Gr = 25{,}000,\ \beta = \frac{\pi}{2}$

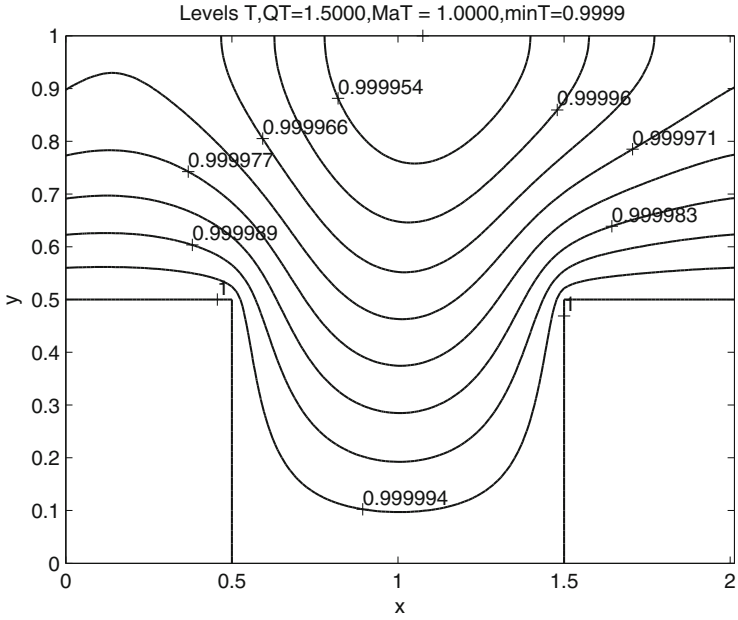


Fig. 8 Levels of temperature in CFS for free convection at $Re = 40$, $\alpha = 0$, $S = 25$, $Gr = 25{,}000$, $\beta = \frac{\pi}{2}$

If $\alpha = \frac{\pi}{2}$, vorticity and integral heat quantity decrease in the magnetic field (Fig. 6); for $\alpha = 0$ the vorticity increases while the integral heat quantity decreases in the magnetic field (Figs. 3 and 8). For $\beta = \frac{\pi}{2}$ and free convection the central flow moves in x-direction and the vortex between cylinders rotates clockwise; similar behaviour is observed for temperature (Figs. 5 and 7).

5 Conclusions

- The 2D MHD free convection flow and temperature fields have been calculated in case of periodically placed cylinders (PC) and obstacles in the channel (CFS).
- It is noticed that in the transverse magnetic field the vorticity and the integral heat quantity both decrease whereas in the longitudinal magnetic field the decrease is seen only in regard to heat quantity (the vorticity then increases noticeably).
- In the strong transverse magnetic field the vortexes were not observed.
- The integral heat quantity for the CFS convection is greater compared to the PC convection.

Acknowledgements This work was partially supported by the grant 623/2014 of the Latvian Council of Science.

References

1. Dorodnicyn, A., Meller, N.A.: On some methods for solving the equations Navier–Stokes (in Russian). In: Abstract of 3-rd Congress of Theoretical and Applied Mechanics, Moscow (1968)
2. Geljfgad, J., Lielausis, O., Cherbinin, E.: Liquid Metal on the Influence of Electromagnetic Forces (in Russian). Zinatne, Riga (1976)
3. Tu, J., Yeoh, G.H., Liu, C.: Computational Fluid Dynamics. A Practical Approach. Elsevier/BH, Amsterdam/Boston (2008)

On Detecting the Shape of an Unknown Object in an Electric Field

Jukka-Pekka Humaloja, Timo Hämäläinen, and Seppo Pohjolainen

Abstract The problem discussed in this paper is detecting the shape of an unknown object in a 2-dimensional static electric field. For simplicity, the problem is defined in a partially rectangular domain, where on a part of the boundary the potential and/or its normal derivative are known. On the other part of the boundary the boundary curve is unknown, and this curve is to be determined. The unknown part of the boundary curve describes the shape of the unknown object.

The problem is defined in the complex plane by an analytic function $w = f(z) = u(x, y) + iv(x, y)$ with the potential u as its real part. Then the inverse function is given as $f^{-1}(w) = x(u, v) + iy(u, v)$, where the functions x and y are harmonic in a rectangle with an unknown boundary condition on one boundary. The alternating-field technique is used to solve the unknown boundary condition.

Keywords Alternating-field technique • Boundary reconstruction problem • Shape detection

1 Introduction

The problem discussed in this paper is detecting the shape of an unknown object in a 2-dimensional static electric field. For simplicity, the problem is defined in a partially rectangular domain, where on a part of the boundary the potential and its normal derivative are known, on the second part of the boundary homogeneous Neumann boundary conditions are used. On the third part of the boundary the potential is known, but the boundary curve is unknown, and this curve is to be reconstructed. The unknown part of the boundary curve describes the shape of the unknown object.

J.-P. Humaloja (✉) • T. Hämäläinen • S. Pohjolainen
Department of Mathematics, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland
e-mail: jukka-pekka.humaloja@tut.fi; timo.t.hamalainen@tut.fi; seppo.pohjolainen@tut.fi

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_156

There are quite a number of different approaches for solving such boundary reconstruction problems, e.g., the method of fundamental solutions [1], the boundary element method [5] or using an indicator function derived from Green's identities [3]. However, our efforts to apply these methods to the problem at hand have been rather unsuccessful. A more suitable method for our problem was found out to be the alternating-field technique on the inverted plane [4], where the region of the problem is conformally mapped to a rectangle in the inverted plane. In the inverted plane all the boundaries of the region are fixed, instead we have an unknown boundary condition on the boundary corresponding the free boundary in the original problem. The missing boundary condition is determined using the iterative alternating-field technique. We will adjust the technique presented in [4] to our problem and demonstrate its functionality on a few test cases.

2 Problem Formulation

Let $a, b \in \mathbb{R}$ such that $a < b$ and let $h : [a, b] \to \mathbb{R}$ such that $h \in C([a, b])$. Now, define domain R by

$$R = \{(x, y) \in \mathbb{R}^2 \,|\, x \in [a,\, b],\, y \in [0,\, h(x)]\}. \tag{1}$$

Let the lines $x = a$ and $x = b$ be perfectly insulated and the line $y = 0$ be perfectly conducting. If a constant voltage potential $u_0 = 1$ is applied to the curve $y = h(x)$, then u_0 generates the electric field $e = -\nabla u$, where the electrical potential u satisfies the following mixed boundary value problem:

$$\begin{aligned} \nabla^2 u &= 0 \quad \text{in } R, \\ \partial_x u &= 0 \quad \text{on } x = a \text{ and } x = b, \\ u &= 0 \quad \text{on } y = 0, \\ u &= 1 \quad \text{on } y = h(x). \end{aligned} \tag{2}$$

When h is known and sufficiently regular, it is well-known that the mixed boundary value problem given in Eq. (2) has a unique solution. However, if h is unknown, but instead we are given an additional boundary condition $-\partial u_y = g(x)$ on the line $y = 0$ with $g : [a, b] \to \mathbb{R}$, the inverse problem of finding h is nonlinear and ill-posed. Additionally, in practice we do not actually know the entire function g but only its values at some discrete points $x_i \in [a, b]$. The geometry for the problem is displayed in Fig. 1.

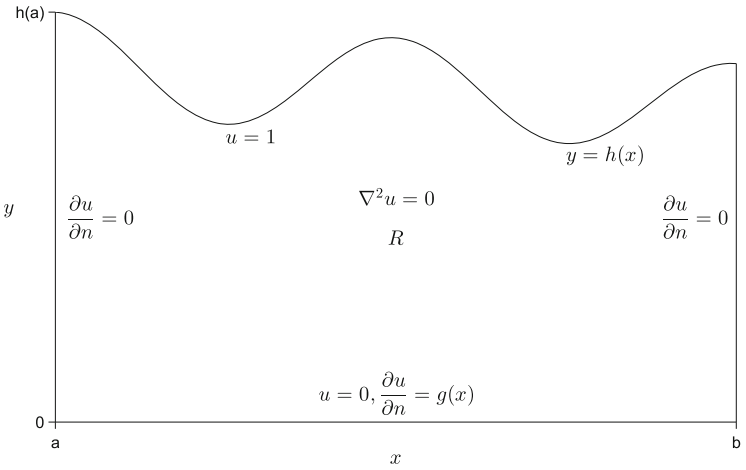


Fig. 1 The problem on the region R

3 Problem on the Inverted Plane

Since the region R is simply connected, the harmonic potential u has a harmonic conjugate v in R such that the complex potential $w = u + iv$ is analytic there. The component functions u and v are known to be connected by the Cauchy-Riemann equations $\partial_x u = \partial_y v$ and $\partial_y u = -\partial_x v$, and thus, it is possible to determine the boundary conditions for v from the boundary conditions of u [2].

It can be seen directly from the Cauchy-Riemann equations that the equipotential lines of u are the lines where $\partial_n v = 0$ and conversely, the lines where $\partial_n u = 0$ are the equipotential lines of v. It yet remains to determine the values of v on the equipotential lines $x = a$ and $x = b$. From the boundary condition $\partial_n u = g(x)$ on the line $y = 0$ we obtain $g(x) = -\partial_y u = \partial_x v$ and thus, the change in the value of v between the lines $x = a$ and $x = b$ is given by $\int_a^b g(x)\mathrm{d}x$ which can be evaluated, e.g., using Simpson's rule. Since the values of v can be determined up to an additive constant, we may assign $v(x = a) = -\int_a^b g(x)\mathrm{d}x = V$ and $v(x = b) = 0$. Furthermore, we may obtain the value of v anywhere on the line $y = 0$ from $v(x) = -\int_x^b g(s)\mathrm{d}s$, which becomes necessary when we determine the boundary conditions for the inverted problem.

Now we have harmonic conjugates u and v which are real and imaginary parts of the analytic function $w = f(z) = u + iv$. If the function f is invertible in R and if $f'(z_0) \neq 0$ at each point $z_0 \in R$, then f has an analytic inverse $f^{-1}(w) = x(u, v) + iy(u, v)$ such that $f^{-1}[f(z)] = z$ [2]. Since the inverse of f is analytic in $f(R)$, its component functions x and y are harmonic conjugates there, i.e.,

$$\partial_{uu}x + \partial_{vv}x = 0 \quad \text{and} \quad \partial_{uu}y + \partial_{vv}y = 0 \tag{3}$$

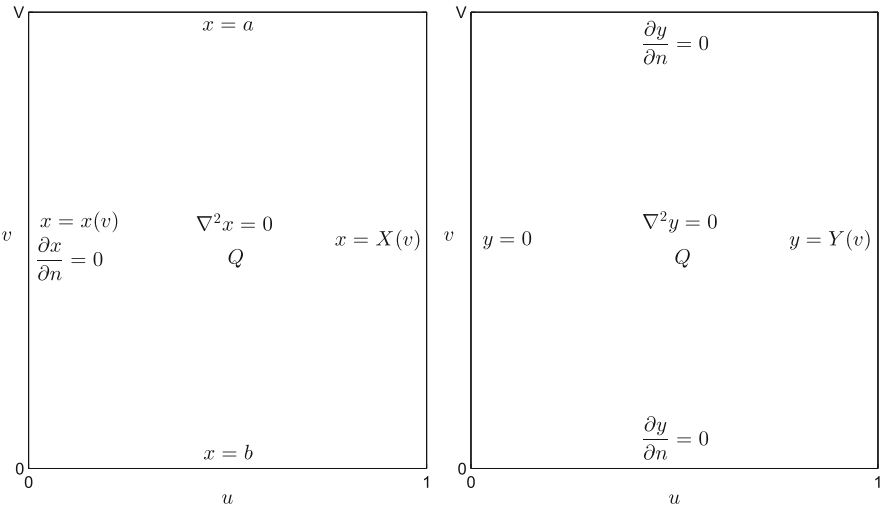


Fig. 2 The inverted problems for x and y on the region Q

and

$$\partial_u x = \partial_v y \quad \text{and} \quad \partial_v x = -\partial_u y \tag{4}$$

The region in the inverted plane is a rectangle $u \in [0, 1], v \in [0, V]$ as shown in Fig. 2, where also the boundary conditions for the inverted problems are given. The boundary conditions $y = 0$, $x = a$ and $x = b$ are obtained directly from the geometry of the original problem, and the corresponding homogeneous Neumann boundary conditions are obtained from the Cauchy-Riemann equations (4). Furthermore, there is an additional boundary condition $x = x(v)$ on the line $u = 0$, which is the inverse of $v = v(x) = -\int_x^b g(s)\mathrm{d}s$. In practice the values of $x(v)$ are only required at some discrete points v_i, which can be interpolated from $v = v(x)$, e.g., by using splines.

The unknown boundary conditions $x = X(v)$ and $y = Y(v)$ on the line $u = 1$ represent the unknown boundary curve which is now mapped to a fixed line. With different values of v we will obtain points (x, y) on the z-plane, which construct the curve $y = h(x)$. The unknown boundary conditions are to be determined using the alternating-field technique which is described next.

4 Alternating-Field Technique

The alternating-field technique is described in [4] by Nilson and Tsuei. The procedure given in the following is schematically similar to the one in [4], but some steps are altered due to differences in the geometries of the problems. In outline,

the alternating-field technique is an iterative procedure, where we find convergent estimates for $X(v)$ and $Y(v)$ by solving Laplace's equation, by turns, for x and y. The convergence of $X(v)$ and $Y(v)$ is measured by the change in the arc length parameter s given by

$$s_i = \sum_{k=1}^{i} \sqrt{\Delta X(v_k)^2 + \Delta Y(v_k)^2}, \tag{5}$$

i.e., s is the arc length parameter of the unknown boundary curve $y = h(x)$.

For the procedure, the region Q is covered by an $M \times N$ rectangular mesh of size $\Delta u = 1/(N-1)$ and $\Delta v = V/(M-1)$. The mesh points are denoted by (u_j, v_i) such that $u_1 = 0$, $u_N = 1$, $v_1 = V$ and $v_M = 0$, and the value of x (resp. y) at a point (u_j, v_i) is denoted by x_{ij} (resp. y_{ij}). Laplace's equation is solved as a system of linear equations, where the coefficient matrix is in $\mathbb{R}^{MN \times MN}$, but it has only five nonzero diagonals. Solutions can be computed effectively by using sparse LU decomposition which needs to be computed only once for the coefficient matrices of x and y.

The steps for the iterative procedure are as follows:

0. Make an initial guess for $X(v)$. Note that $X(v_i) \in [a, b]$ for every $i \in \{1, 2, \ldots, M\}$ and that $X(v_1) = a$ and $X(v_M) = b$. Then perform steps 1–6 to obtain the first iterates for $X(v)$ and $Y(v)$ and an initial estimate for the arc length parameter s.
1. Assign boundary conditions for $x(u, v)$ field, i.e., set

$$\begin{aligned} &x_{1j} = a, \forall j \in \{1, 2, \ldots, N\}, && x_{Mj} = b, \forall j \in \{1, 2, \ldots, N\}, \\ &x_{i1} = x(v_i), \forall i \in \{1, 2, \ldots, M\} && x_{i2} = x_{i1}, \forall i \in \{1, 2, \ldots, M\}, \\ &x_{iN} = X(v_i), \forall i \in \{1, 2, \ldots, M\}. \end{aligned} \tag{6}$$

2. Solve Laplace's equation for x in Q.
3. Calculate new $Y(v)$ by the formula

$$\begin{aligned} &Y(v_i) = -\int_0^1 \partial_v x_{ij} \mathrm{d}u, \; i \in \{2, 3, \ldots, M-1\}, \\ &Y(v_1) = Y(v_2), \qquad Y(v_M) = Y(v_{M-1}), \end{aligned} \tag{7}$$

where

$$\partial_v x_{ij} \approx \frac{x_{(i-1)j} - x_{(i+1)j}}{2\Delta v} \tag{8}$$

and the integral can be evaluated, e.g., using Simpson's rule.

4. Assign boundary conditions for $y(u, v)$ field, i.e., set

$$\begin{aligned} &y_{1j} = y_{2j}, \forall j \in \{1, 2, \ldots, N\}, \; y_{Mj} = y_{(M-1)j}, \forall j \in \{1, 2, \ldots, N\}, \\ &y_{i1} = 0, \forall i \in \{1, 2, \ldots, M\}, \quad y_{iN} = Y(v_i), \forall i \in \{1, 2, \ldots, M\}, \end{aligned} \tag{9}$$

   where $Y(v_i)$ is given by Eq. (7).
5. Solve Laplace's equation for y in Q.
6. Calculate new $X(v)$ by the formula

$$\begin{aligned} &X(v_i) = x_{i1} + \int_0^1 \partial_v y_{ij} \mathrm{d}u, \; i \in \{2, 3, \ldots, M-1\}, \\ &X(v_1) = a, \qquad\qquad X(v_M) = b, \end{aligned} \tag{10}$$

   where

$$\partial_v y_{ij} \approx \frac{y_{(i-1)j} - y_{(i+1)j}}{2\Delta v} \tag{11}$$

   and the integral can be evaluated, e.g., using Simpson's rule. Then calculate a new arc length parameter s^* from the newly obtained $X(v)$ and $Y(v)$ using Eq. (5).
7. Check convergence for s, i.e., calculate $||s^* - s||^2$. If necessary, set $s = s^*$ and return to step 1. A new estimate for $X(v)$ is given by Eq. (10).

The usual criterion for the procedure to stop is to see, whether $||s^* - s||^2$ is sufficiently small. Other possibility would be, e.g., to inspect the convergence rate of s and determine a suitable stopping criterion based on its changes.

5 Numerical Test Cases

The procedure described in the previous section is tested on four different boundary curves $y = h(x)$. The curves, as well as the approximations obtained from the procedure, are presented in Fig. 3. In each of the cases, we have $a = 0$ and $b = 1$, and the value of $g(x)$ is computed at 21 evenly spaced points on the interval $[0, 1]$. Thus, the uv-plane is covered by a 21×21 rectangular mesh. Initial guess for $X(v)$ in all the cases is $x(v)$ and the stopping criterion for the procedure is $||s^* - s||^2 < 10^{-10}$. Data on error norms, absolute and relative maximum errors and the number of iterations required for $||s^* - s||^2 < 10^{-10}$ is displayed in Table 1 for each test case. The numbering of the cases corresponds to the order of the images in Fig. 3.

Based on Table 1 and Fig. 3 we see that for the most part the approximated boundary points Y agree with the actual boundary curve $h(X)$. However, a few noticeable errors occur as well. These errors are mostly caused by the conformal mapping from the xy-plane to the uv-plane. Namely, if $h'(0) \neq 0$, $h'(1) \neq 0$ or the curve $y = h(x)$ contains non-smooth points, those points are non-analytical points for the function $f(z) = u + iv$ and thus, the mapping $f(z)$ is not conformal

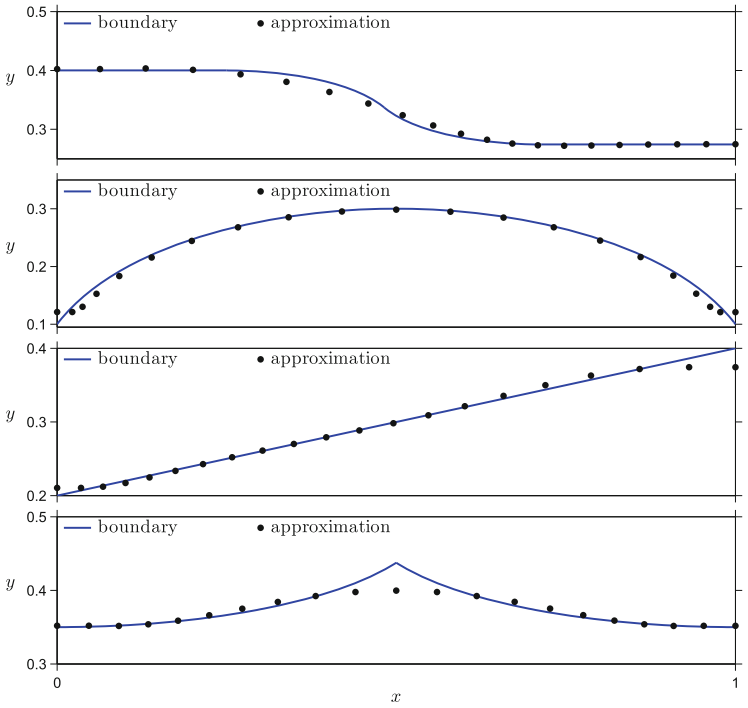


Fig. 3 The boundary curves $y = h(x)$ for the test cases 1–4 and the approximated boundary points obtained using the alternating-field technique

Table 1 Results for the numerical test cases

Case	$\|\|Y - h(X)\|\|_2$	$\|\|Y - h(X)\|\|_\infty$	$\max\left\{\frac{\|Y-h(X)\|}{h(X)}\right\}$	Iterations
1	0.0285	0.0161	0.0424	29
2	0.0458	0.0211	0.2110	18
3	0.0320	0.0256	0.0641	21
4	0.0434	0.0379	0.0866	33

at those points, which will cause errors. Most probably some errors arise from the alternating-field technique as well. However, there are no existing stability or error estimates for the technique, so these errors are virtually unknown. Regardless, it would seem that the errors caused by sources other than the conformal mapping are rather insignificant in comparison.

References

1. Borman, B., Ingham, D.B., Johansson, B.T., Lesnic, D.: The method of fundamental solutions for detection of cavities in EIT. J. Integr. Equ. Appl. **21**(3), 383–406 (2009)
2. Brown, J.W., Churchill, R.V.: Complex Variables and Applications, 7th edn. McGraw-Hill, New York (2004)
3. Karamehmedovic, A., Knudsen, K.: Inclusion estimation from a single electrostatic boundary measurement. Inverse Prob. **29**(2), 18 (2013)
4. Nilson, R.H., Tsuei Y.G.: Free boundary problem of ECM by alternating-field technique on inverted plane. Comput. Methods Appl. Mech. Eng. **6**(3), 265–282 (1975)
5. Quinn, D.W., Oxley, M.E.: The boundary element method applied to moving boundary problems. Math. Comput. Model. **14**, 145–150 (1990)

Tracking of Reference Robot Trajectory Using SDRE Control Method

Elvira Rafikova, Luiz Henrique de Vitro Gomez, and Marat Rafikov

Abstract The application of the SDRE-State Dependent Ricatti Equation method for the tracking control of a nonholonomic mobile robot is presented in this work. The proposed control law minimize the quadratic cost functional consisting of tracking errors and control efforts. The numerical simulations demonstrate the efficacy of the control method applied to track the linear and circular trajectory reference robot.

Keywords Control problem • Robot motion • Tracking

1 Introduction

In this paper it is considered a tracking control problem applied to a differential steering nonholonomic robot. This problem received attention in [1] in which a locally exponentially stabilizing control was proposed. A dynamic feedback linearization technique for wheeled mobile robot was presented in [2]. Global tracking control laws were proposed in [3]. Model-based predictive control for the differential steering mobile robot is presented in [4]. In [5] the switched control is proposed for nonholonomic mobile robot. The control method considered in this paper is SDRE-State Dependent Riccati Equation control. This strategy has become very popular within the control community over the last decade, providing a very effective algorithm for synthesizing nonlinear feedback controls by allowing nonlinearities in the system states while additionally offering great design flexibility through state dependent weighting matrices. For this method Mracek and Cloutier [6] proved local asymptotic stability for a multivariable case system with SDRE feedback controller. A review of SDRE method can be found in [7].

E. Rafikova (✉) • L.H. de Vitro Gomez • M. Rafikov
Federal University of ABC, Av. dos Estados, 5001 Santo Andre, Brazil
e-mail: elvira.rafikova@ufabc.edu.br; luizhenriquesamurai@hotmail.com; marat.rafikov@ufabc.edu.br

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_157

2 Robot Model

The problem under consideration is the tracking of a reference robot trajectory for a wheeled mobile robot of differential steering type, shown in Fig. 1. The vehicle is equipped with two identical, parallel, nondeformable, standard wheels that are actuated separately. Moreover, pure rolling and nonslipping contact with the ground causes restrictions on the degree mobility of the robot which limits the initial velocity vector set.

The kinematics of mobile robot is given by the model:

$$\begin{aligned} \dot{x} &= v\cos\theta \\ \dot{y} &= v\cos\theta \\ \dot{\theta} &= \omega \end{aligned} \tag{1}$$

Where $\left[\, x\ y\ \theta \,\right]^T$ are the position and orientation coordinates of the robot and $\left[\, V\ \omega \,\right]^T$ are the linear and angular velocities of the robot, respectively. The linear and angular velocities of the robot is given by the composition of the velocities of the right V_R and left V_L :

$$V = \frac{(V_R + V_L)}{2}, \omega = \frac{(V_R + V_L)}{2r} \tag{2}$$

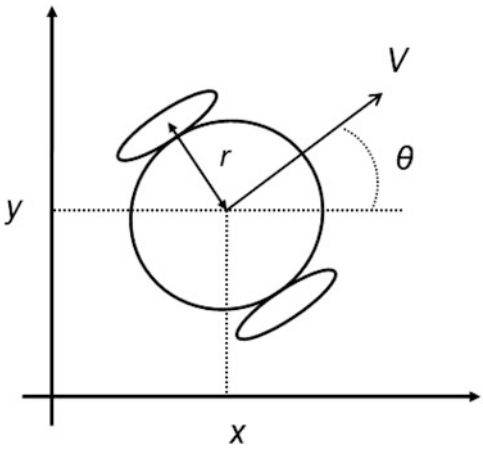


Fig. 1 The differential steering mobile robot schematic model showing the coordinates system

3 Control Tracking Problem of the Differential Steering Robot

When the robot is controlled to follow the reference robot, it usually has some state error. This error can be expressed by vector $e(t) = \begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix}^T$, defined as:

$$\mathbf{e} = \begin{bmatrix} x - x_r \\ y - y_r \\ \theta - \theta_r \end{bmatrix} \tag{3}$$

where $\begin{bmatrix} x_r & y_r & \theta_r \end{bmatrix}^T$ are the position and orientation coordinates of the reference robot. It is considered that the reference system is the same robotic system with desired position and velocity then it can be written in the same form as (1) :

$$\begin{aligned} \dot{x_r} &= v_r \cos(\theta_r) \\ \dot{y_r} &= v_r \sin(\theta_r) \\ \dot{\theta_r} &= \omega_r \end{aligned} \tag{4}$$

Taking into account (1) and (4) the following error system is obtained:

$$\begin{aligned} &\dot{e_1} = v_r \cos(\theta_r)\cos(e_3) - v_r \sin(\theta_r)\sin(e_3) + u_1 \cos(\theta_r)\cos(e_3) - \\ &-u_1 \sin(\theta_r)\sin(e_3) - v_r \cos(\theta_r) \\ &\dot{e_2} = v_r \sin(\theta_r)\cos(e_3) + v_r \sin(e_3)\cos(\theta_r) + u_1 \sin(\theta_r)\cos(e_3) + \\ &+u_1 \sin(e_3)\cos(\theta_r) - v_r \sin(\theta_r) \\ &\dot{e_3} = u_2 \end{aligned} \tag{5}$$

where a control vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 \end{bmatrix}^T$ is defined as:

$$\mathbf{u} = \begin{bmatrix} v - v_r \\ \omega - \omega_r \end{bmatrix} \tag{6}$$

Supposing that e_3 is small, $\cos(e_3) = 1$ and $\sin(e_3) = e_3$, (5) can be presented in following form:

$$\begin{bmatrix} \dot{e_1} \\ \dot{e_2} \\ \dot{e_3} \end{bmatrix} = \mathbf{A} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} + \mathbf{B} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{7}$$

where the matrices A, B are :

$$\mathbf{A} = \begin{bmatrix} 0\ 0 & -v_r \sin(\theta_r) \\ 0\ 0 & v_r \cos(\theta_r) \\ 0\ 0 & 0 \end{bmatrix} \tag{8}$$

and

$$\mathbf{B} = \begin{bmatrix} \cos(\theta_r) - e_3 \sin(\theta_r) & 0 \\ \sin(\theta_r) + e_3 \cos(\theta_r) & 0 \\ 0 & 1 \end{bmatrix} \tag{9}$$

4 SDRE Control Method Applied to Trajectory Control of the Robot

The explanation of the main idea of the method follows ahead. Consider the general infinite-horizon, input-affine, autonomous, nonlinear regulator problem of the form. The nonlinear system is presented in linear form with state dependent matrices:

$$\dot{\mathbf{e}} = A(e)\,\mathbf{e} + B(e)\,\mathbf{u} \tag{10}$$

The minimized functional is:

$$J = \frac{1}{2} \int_0^\infty \left[\mathbf{e}^T Q(e) \mathbf{e} + \mathbf{u}^T R(e) \mathbf{u} \right] dt \tag{11}$$

where $\mathbf{e} \in \mathbb{R}^{\kappa}$ is a state vector and matrices $Q(e)$ and $R(e)$ are positive definite for all e .

The control $\mathbf{u}$ is given by equation:

$$\mathbf{u} = -\mathbf{R}^{-1}(e)\mathbf{B}^T(e)\mathbf{P}(e)\mathbf{e} \tag{12}$$

where the matrix $\mathbf{P}(e)$ is obtained from :

$$\mathbf{P}(e)\mathbf{A}(e) + \mathbf{A}^T(e)\mathbf{P}(e) - \mathbf{P}(e)\mathbf{B}(e)\mathbf{R}^{-1}(e)\mathbf{B}^T(e)\mathbf{P}(e) + \mathbf{Q}(e) = 0 \tag{13}$$

Equation (13) is a state dependent Riccati equation. According to [6] the regulator (12) with $\mathbf{P}(e)$ obtained from (13), is suboptimal.

5 Numerical Results

The goal of this section is to obtain the suboptimal control strategy when the reference robot trajectories are linear and circular. The numerical results are presented in this section in order to demonstrate the efficacy of control method.

5.1 Linear Reference

The reference robot trajectory is linear when $\theta_r = const$. In this case, the robot control functions are determined in following form:

$$\begin{aligned} v &= v_r + u_1 \\ \omega &= u_2 \end{aligned} \tag{14}$$

where u_1 and u_2 are determined by (12) and (13).

The error system and robot tracking trajectories are presented in Figs. 2 and 3, respectively.

In the Fig. 2 the error coordinates are driven to zero by the effort of the control in 5 s.

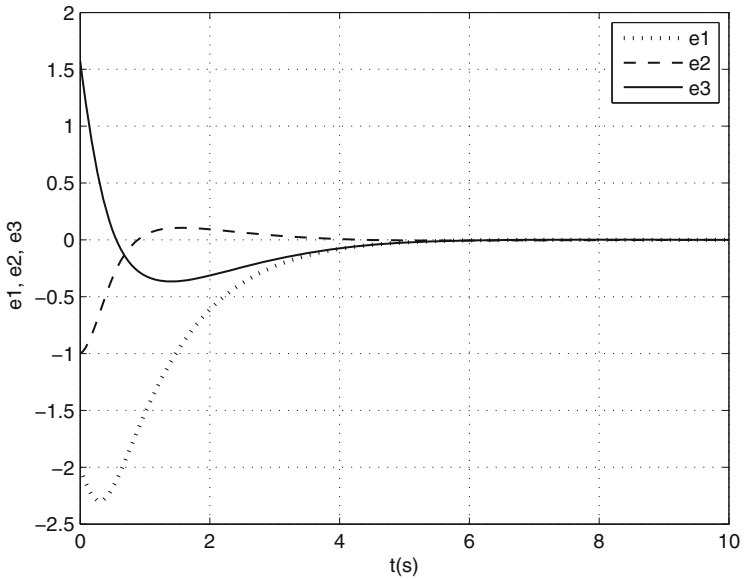


Fig. 2 Error system trajectory for $\theta_r = \pi/4$ and $v_r = 1$

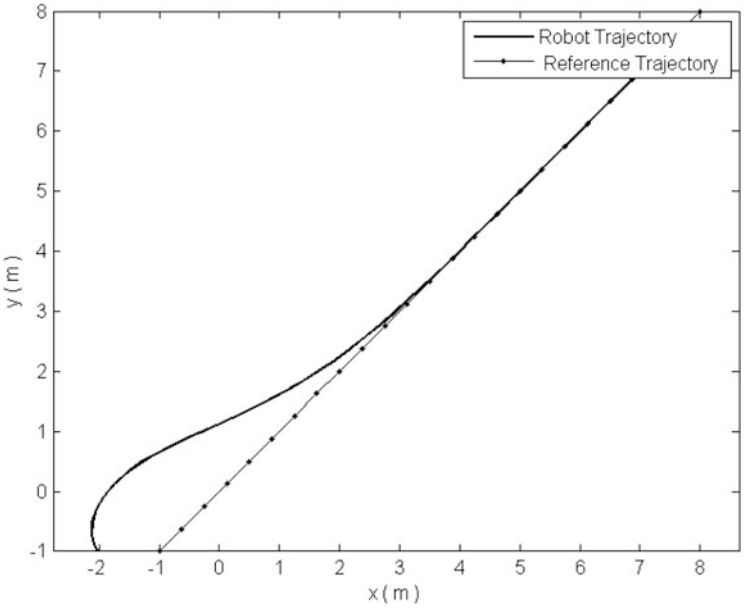


Fig. 3 Robot tracking trajectory for $\theta_r = \pi/4$ and $v_r = 1$

The Fig. 3 represents the robot trajectory (solid line) successfully converging to the reference trajectory represented by solid line with dots. The initial condition of the robot system are $x_0 = -2, y_0 = -1$ and $\theta_0 = 3\pi/4$.

5.2 Circular Reference

The reference robot trajectory is circular when $\omega_r = const$. In this case, the robot control functions are determined from the following form:

$$\begin{aligned} v &= v_r + u_1 \\ \omega &= \omega_r + u_2 \end{aligned} \tag{15}$$

where u_1 and u_2 are determined by (12) and (13). The error system and robot tracking trajectories are presented in Figs. 3 and 4, respectively.

In the Fig. 4 error coordinate trajectories are driven to zero in less than 3 s by the control effort.

In Fig. 5 the line with triangular shape represent the robot trajectory converging to the circular reference (represented by dashed line).

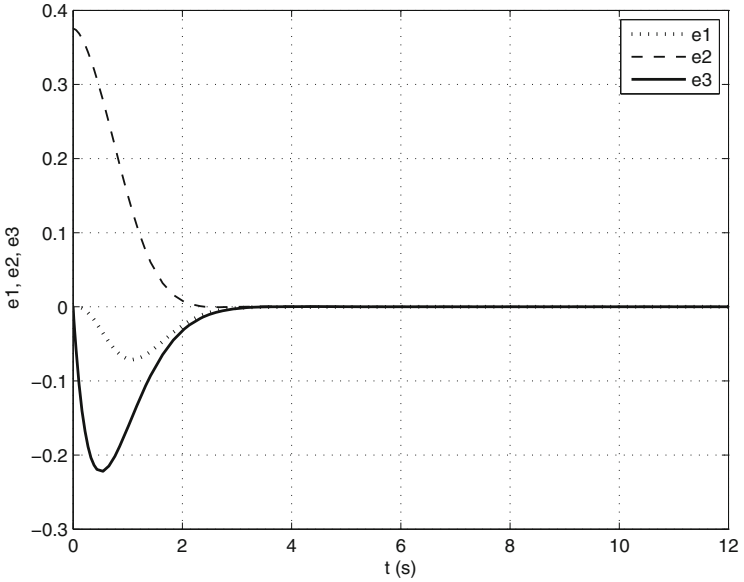


Fig. 4 Error system trajectory for $\theta_r = \pi/4$ and $v_r = 1$

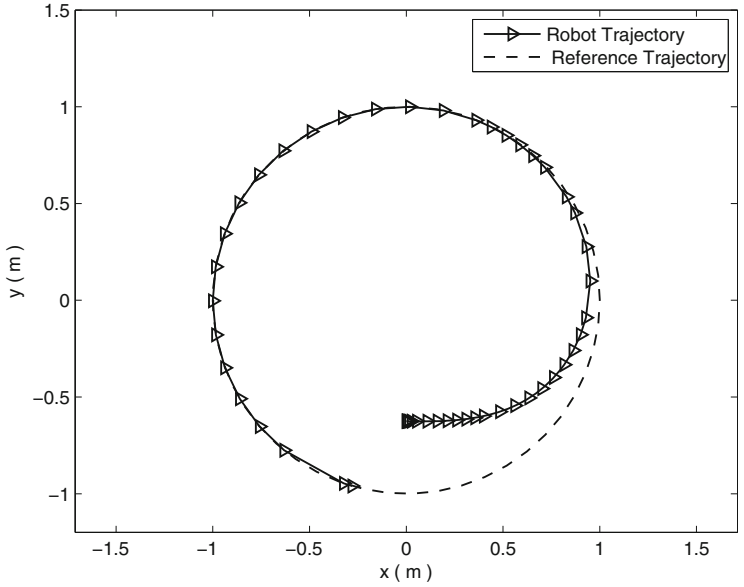


Fig. 5 Robot tracking trajectory for $\theta_r = \pi/4$ and $v_r = 1$

6 Concluding Remarks

The application of the SDRE method for the tracking control of a mobile robot is presented in this work. The proposed control law minimize the quadratic cost functional consisting of tracking errors and control efforts. Assuming the hypothesis that orientation coordinate errors are small, the error system is transformed in the form which is adequate for SDRE method. The numerical simulations show that the proposed algorithm can be applied even when the orientation coordinate errors are not small. The next steps of this work include the real-time implementation of the proposed control strategy on experimental mobile robot.

References

1. Walsh, G., Tilbury, D., Sastry, S., Murray, R., Laumond, J.: Stabilization of trajectories for systems with nonholonomic constraints. IEEE Trans. Autom. Control **39**(1), 216–222 (1994)
2. Canudas de Wit, C., Sordalen, O.J.: Exponential stabilization of mobile robots with nonholonomic constraints. IEEE Trans. Autom. Control **37**(11), 1791–1797 (1992)
3. Jiang, Z.P., Nijmeijer, H.: A recursive technique for tracking control of nonholonomic systems in chained form. IEEE Trans. Autom. Control **44**(2), 265–279 (1999)
4. Klancar, G., Skrjanc, I.: Tracking error model-based predictive control for mobile robots in real time. Robot. Auton. Syst. **55**, 460–469 (2007)
5. Sankaranarayanan, V., Mahindrakar, A.D.: Switched control of a nonholonomic mobile robot. Commun. Nonlinear Sci. Numer. Simul. **14**, 2319–2327 (2009)
6. Mracek, C.P., Cloutier, J.R.: Control designs for the nonlinear benchmark problem via the state-dependent Riccati equation method. Int. J. Robust Nonlinear Control **8**, 401–433 (1998)
7. Cimen, T. :Systematic and effective design of nonlinear feedback controllers via the state-dependent Riccati equation (SDRE) method. Annu. Rev. Control. **34**(1), 32–51 (2010)

Part III
Anile Prize Lecture

Design of Silicon Based Integrated Optical Devices Using the Finite Element Method

Paolo Pintus

Abstract Among the components needed in photonic integrated circuits, dielectric waveguides and small footprint ring resonators play a key role for many applications and require sophisticated electromagnetic analysis and design. In this work, we present an accurate vectorial mode solver based on the finite element method. Considering a general nonreciprocal permittivity tensor, the proposed method allows us to investigate important cases of practical interest. To compute the electromagnetic modes, the Rayleigh-Ritz functional is derived for the non-self adjoint case, it is discretized using the node elements and the penalty function is added to remove the spurious solutions. Although the use of the penalty function is well known for the waveguide problem, it has been introduced for the first time (to the best of our knowledge) in the ring resonator modal analysis. The resulting quadratic eigenvalue problem is linearized and solved in terms of the propagation constant for a given frequency (i.e., γ-formulation). Unlike the earlier developed mode solvers, our approach allows us to precisely compute both forward and backward propagating modes in the nonreciprocal case. Moreover, it avoids time-consuming iterations and preserves matrix sparsity, ensuring high accuracy and computational efficiency.

Keywords Electromagnetic modes • Finite element method • Modal analysis • Optical device • Photonic integrated circuit

1 Introduction

The study of the electromagnetic field propagation in optical devices is the key starting point to investigate photonic components before their manufacturing. Therefore, rigorous mathematical models are very important tools to perform precise simulations and accurate design. In this work, we present a mode solver for

P. Pintus (✉)
Scuola Sant'Anna, via Moruzzi 1, 56124, Pisa, Italy

CNIT National Laboratory of Photonic Networks, via Moruzzi 1, 56124 Pisa, Italy
e-mail: paolo.pintus@sssup.it

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7_158

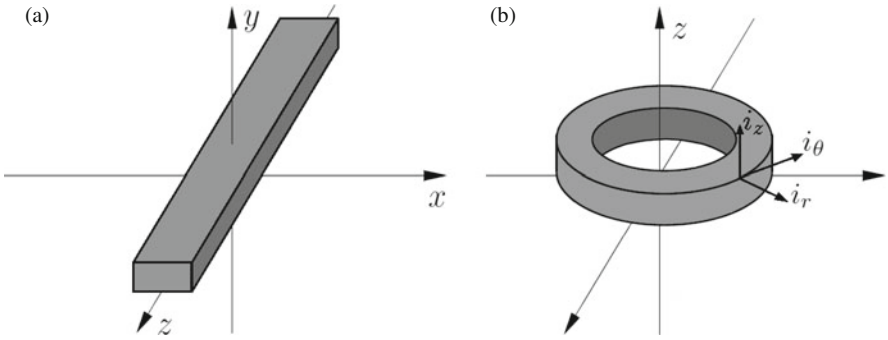


Fig. 1 Device under investigation. (**a**) Straight waveguide. (**b**) Ring resonator

dielectric waveguides and micro ring resonators that are the building blocks of all photonic integrated circuits. Those components are schematically shown in Fig. 1.

To perform their modal analysis, several numerical methods are commonly used such as the finite element method (FEM) [6, 8], the finite difference method (FDM) [3, 9], the method of lines (MoL) [1], and the film mode matching (FMM) [17, 21]. Due to the possibility of using adaptive meshing, FEM shows several advantages. It usually provides better approximation and requires less memory to store the stiffness matrix with respect to FDM, while it is more appropriate than MoL and FMM for modal analysis of graded index waveguides [20], and in general for waveguides with complex cross-section geometry and refractive index profiles [14]. In addition, it is the most suited to solve deformation and stress problems in solids, like for the case of stress-induced effects in optical waveguides [23].

With the FEM, the solution is numerically computed as a linear combination of basis functions which, in electromagnetism, are usually of two kinds: *node elements* (also called *Lagrangian elements*) and *edge elements* (also called *Nédélec elements*) [6]. Edge elements have mainly three important advantages: (1) the spurious solutions can be effectively removed in several electromagnetic problem formulations, (2) the boundary conditions at material interface and conducting surface can be easily imposed, (3) there are no difficulties in treating conducting and dielectric edges and corners related to the field singularities [6, 13]. On the other hand, node elements are more efficient as regards to the storage requirements and the number of floating point operations (FLOPs) [13]. Moreover, the solutions computed using node elements provide higher accuracy when extremely flat or elongated elements are used in the mesh [13]. It is worth noting that in order to completely get rid of any spurious solutions introduced by the node elements, the penalty function can be added to the functional [18, 19]. In this work, the node elements are used, since we do not consider waveguides or ring resonators with field singularities (e.g., by using magnetic-field formulation) and we assume a zero-field condition on the border. However, the method can be implemented also with edge elements.

2 Mathematical Model

An electromagnetic mode is a solution of the Maxwell's equation which propagates in a waveguide or in a ring resonator without sources. More formally, a mode is an eigenfunction of one of the differential operators $\mathscr{L}_H$ and $\mathscr{L}_E$ that are defined as

$$[\mathscr{L}_H \mathbf{H}](\mathbf{r}) = \nabla \times \left[\boldsymbol{\varepsilon}_r^{-1}(\mathbf{r}) \nabla \times \mathbf{H}(\mathbf{r})\right] - \left(\frac{\omega}{c}\right)^2 \boldsymbol{\mu}_r(\mathbf{r}) \mathbf{H}(\mathbf{r}), \tag{1a}$$

$$[\mathscr{L}_E \mathbf{E}](\mathbf{r}) = \nabla \times \left[\boldsymbol{\mu}_r^{-1}(\mathbf{r}) \nabla \times \mathbf{E}(\mathbf{r})\right] - \left(\frac{\omega}{c}\right)^2 \boldsymbol{\varepsilon}_r(\mathbf{r}) \mathbf{E}(\mathbf{r}), \tag{1b}$$

where $\mathbf{r}$ is the position vector, $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic field, $\boldsymbol{\varepsilon}_r$ and $\boldsymbol{\mu}_r$ are the relative permittivity and relative permeability tensor, while ω and c are the angular frequency and the speed of light in the vacuum, respectively. Those differential operators are derived from the Maxwell's equations, assuming linear, instantaneous and time invariant media [7]. Because at optical frequency $\boldsymbol{\mu}_r = 1$, the normal and tangent component of $\mathbf{H}$ are continuous across any boundary separating two different media [5, 11]. For this reason, the formulation of Eq. (1a) in term of the magnetic field is preferred. As can be seen from Fig. 1, straight waveguides and ring resonators are characterized by a continuous translational symmetry along one direction. By considering the coordinate system in Fig. 1 (i.e., Cartesian coordinates for the waveguide and cylindrical coordinates for the ring), the waveguide and the ring resonator are invariant structures with respect to the z-axis and the θ-axis, respectively. Assuming the fields propagate as harmonic waves and taking into account the translational symmetry, the magnetic fields in the waveguide and in the ring are

$$\mathbf{H}^{wg} = \mathbf{H}(x, y) e^{i\omega t - \gamma z}, \qquad \mathbf{H}^{rr} = \mathbf{H}(r, z) e^{i\omega t - \gamma \theta}, \tag{2}$$

where $\gamma \in \mathbb{C}$ is called *propagation constant*. Note that $\mathbf{H}^{rr}$ can be derived from $\mathbf{H}^{wg}$ as follows $x \mapsto r$, $y \mapsto z$, and $z \mapsto \theta$. Here, we assumed that $\boldsymbol{\varepsilon}_r$ is

$$\boldsymbol{\varepsilon}_r^{wg}(x, y) = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} & -\varepsilon_{xz} \\ \varepsilon_{xy} & \varepsilon_{yy} & \varepsilon_{yz} \\ \varepsilon_{xz} & -\varepsilon_{yz} & \varepsilon_{zz} \end{pmatrix}, \qquad \boldsymbol{\varepsilon}_r^{rr}(r, z) = \begin{pmatrix} \varepsilon_{rr} & -\varepsilon_{r\theta} & \varepsilon_{rz} \\ \varepsilon_{r\theta} & \varepsilon_{\theta\theta} & -\varepsilon_{\theta z} \\ \varepsilon_{rz} & \varepsilon_{\theta z} & \varepsilon_{zz} \end{pmatrix}, \tag{3}$$

for the waveguide and the ring resonator problem, respectively. The tensors in Eq. (3) have complex entries and generalize those presented by Konrad in [10], permitting to consider also lossy materials. In addition, they include the case presented by Lu and Fernandez [12] and allow us to investigate nonreciprocal and anisotropic waveguides. The tensor $\boldsymbol{\varepsilon}_r$ is not symmetric neither adjoint, therefore $\mathscr{L}_H$ results non-self-adjoint. To solve the problems, we introduce the adjoint operator

$$\left[\mathscr{L}_H^a \mathbf{H}^a\right](\mathbf{r}) = \nabla \times \left\{\left[\boldsymbol{\varepsilon}_r^a(\mathbf{r})\right]^{-1} \nabla \times \mathbf{H}^a(\mathbf{r})\right\} - \left(\frac{\omega}{c}\right)^2 \mathbf{H}^a(\mathbf{r}), \tag{4}$$

where $\mathbf{H}^a$ is the adjoint field and $\boldsymbol{\varepsilon}_r^a$ is the adjoint permittivity tensor [2]. Because the domain of $\mathscr{L}_H^a$ is unchanged, the adjoint fields can be written as well as in Eq. (2), where γ^a is the adjoint propagation constant. To compute the eigenfunctions of the operator (1a), we look for the stationary points of the Rayleigh-Ritz functional [2]

$$F(\mathbf{H}, \mathbf{H}^a) = \iiint_V \left[\nabla \times \mathbf{H}^a \cdot \boldsymbol{\varepsilon}_r^{-1} \nabla \times \mathbf{H}\right] dV - \left(\frac{\omega}{c}\right)^2 \iiint_V \mathbf{H} \cdot \mathbf{H}^a \, dV, \tag{5}$$

where $\mathbf{H}$ and $\mathbf{H}^a$ belongs to the Sobolev space $H(curl)$. To removed the spurious solutions from the spectrum of interest, the penalty function has been added

$$\widetilde{F}(\mathbf{H}, \mathbf{H}^a) = F(\mathbf{H}, \mathbf{H}^a) + \alpha_p \iiint_V \nabla \cdot \mathbf{H}^a \, \nabla \cdot \mathbf{H} \, dV, \tag{6}$$

where the constant α_p is a free parameter and it is usually chosen equal to 1. Equation (6) implies that $\mathbf{H}, \mathbf{H}^a \in H(curl) \cap H(div)$.

If $\boldsymbol{\varepsilon}_r^a = \boldsymbol{\varepsilon}_r^t$, the relationships between the direct and adjoint fields are

$$\left(H_x^a, H_y^a, H_z^a\right)^t = \left(H_x, H_y, -H_z\right)^t, \qquad \left(H_r^a, H_\theta^a, H_z^a\right)^t = \left(H_r, -H_\theta, \, H_z\right), \tag{7}$$

for the waveguide and the ring resonator, respectively [15]. Because the field amplitudes in Eq. (2) vary exclusively in the cross-section as well as $\boldsymbol{\varepsilon}_r$, only the xy and rz planes are discretized. So the magnetic fields are approximates as

$$\mathbf{H}_n^{wg} = \sum_{k=1}^{n} \left[h_{x,k} \, \phi_k(x, y) \, \mathbf{i}_x + h_{y,k} \, \phi_k(x, y) \, \mathbf{i}_y + h_{z,k} \, \phi_k(x, y) \, \mathbf{i}_z\right] e^{i\omega t - \gamma z}, \tag{8}$$

$$\mathbf{H}_n^{rr} = \sum_{k=1}^{n} \sqrt{r} \left[h_{r,k} \, \phi_k(r, z) \, \mathbf{i}_r + h_{\theta,k} \, \phi_k(r, z) \, \mathbf{i}_\theta + h_{z,k} \, \phi_k(r, z) \, \mathbf{i}_z\right] e^{i\omega t - \gamma \theta}, \tag{9}$$

where n is the number of mesh nodes. The factor $\sqrt{r}$ has been introduced to avoid difficulties in the integration of the singular terms for the ring resonator modes [10]. To simplify the notation, let us introduce the vector of the unknown coefficients

$$\mathbf{h}^{wg} = \left(h_{x,1} \ldots h_{x,n} \; h_{y,1} \ldots h_{y,n} \; h_{z,1} \ldots h_{z,n}\right)^t, \tag{10}$$

$$\mathbf{h}^{rr} = \left(h_{r,1} \ldots h_{r,n} \; h_{\theta,1} \ldots h_{\theta,n} \; h_{z,1} \ldots h_{z,n}\right)^t, \tag{11}$$

for the two cases, respectively. As a result, the discretized functional (6) is

$$\widetilde{F}(\mathbf{H}) \approx \mathbf{h}^t \left[\gamma^2 M + \gamma \, C + K\right] \mathbf{h} - \omega^2 \, \mathbf{h}^t L \mathbf{h}, \tag{12}$$

where the matrices M, C, K and L are symmetric and explicitly reported in the Appendix. Since the modes are the zeros of the Rayleigh-Ritz functional, we derive

Eq. (12) with respect to the unknown vector. According to the known/unknown parameters, an ω-formulation or a γ-formulation can be derived [19]. In the first case, the propagation constant γ is provided as an input parameter and $(\omega, \mathbf{h})$ is the eigenvalue-eigenvector pair of a generalized eigenvalue problem. Vice versa, fixing ω, we have a quadratic eigenvalue problem (QEP) [22], where $\mathbf{h}$ is the eigenvector and γ its eigenvalue. To compute the modes, the QEP is linearized according to [22].

3 Numerical Results

To validate the model, we compare it with the perturbation method, which is generally used to study nonreciprocal waveguides/ring resonators [4]. In those optical component, the forward and backward propagating modes have different γ. Considered the nonreciprocal ring resonator presented in [16] we computed the difference between the two propagation constant $\Delta\gamma = \gamma^+ - \gamma^-$, where $\pm$ refer to the two directions. While the real part of $\Delta\gamma$ is almost negligible, its imaginary part is different for the two directions, providing a different resonant frequency. The plot of the imaginary part of $\Delta\gamma$ is shown in Fig. 2 with respect to the thickness of the ring (Si thickness). As we can see, the results of the two methods are in good agreement.

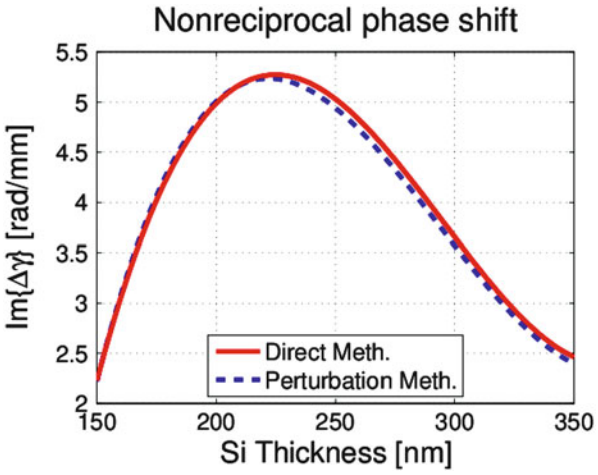


Fig. 2 Comparison between the perturbation method and the proposed one

4 Conclusion

In this work we presented a vectorial mode solver for nonreciprocal and anisotropic waveguides and ring resonators. With the described methods the mode can be accurately computed. Moreover, solving directly the QEP avoids the use of iterative approaches, speeding up simulations and improving the convergence of the eigenvalues at the same time.

Acknowledgements This work has been partially supported by a PhD scholarship from the Scuola Sant'Anna, by the Tuscany Region through the project *ARNO-T3* PAR FAS 2007–2013 and by the European Commission grant *IRIS*, project no. 619194 FP7-ICT.

Appendix

To compute the matrices in Eq. (12) let us define the following matrices

$$
\begin{aligned}
R_{ij} &= \iint_S \phi_i \phi_j \, dxdy, \quad P_{ij} = \iint_S \phi_i \phi_j \, x^2 dxdy, \quad J_{ij} = \iint_S \phi_i \frac{\partial \phi_j}{\partial y} \, dxdy, \\
N_{ij} &= \iint_S \phi_i \frac{\partial \phi_j}{\partial x} \, dxdy, \quad D_{ij} = \iint_S \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial y} \, dxdy, \quad E_{ij} = \iint_S \frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} \, dxdy, \\
Z_{ij} &= \iint_S \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial x} \, dxdy,
\end{aligned}
\tag{13}
$$

where R, P, D and E are symmetric. If $\mathbf{p} = \boldsymbol{\varepsilon}_r^{-1}$, M, L, C and K for the waveguide problem are

$$
\begin{aligned}
M &= \begin{pmatrix} -p_{yy}R & p_{yx}R & 0 \\ p_{xy}R & -p_{xx}R & 0 \\ 0 & 0 & \alpha_p R \end{pmatrix}, \qquad L = \begin{pmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & -R \end{pmatrix}, \\
C &= \begin{pmatrix} p_{zy}J^t - p_{yz}J & p_{yz}N - p_{zx}J^t & p_{yx}J - p_{yy}N - \alpha_p N^t \\ p_{xz}J - p_{zy}N^t & p_{zx}N^t - p_{xz}N & p_{xy}N - p_{xx}J - \alpha_p J^t \\ p_{xy}J^t - p_{yy}N^t - \alpha_p N & p_{yx}N^t - p_{xx}J^t - \alpha_p J & 0 \end{pmatrix}, \\
K &= \begin{pmatrix} p_{zz}D + \alpha_p E & -p_{zz}Z + \alpha_p Z^t & p_{zy}Z - p_{zx}D \\ -p_{zz}Z^t + \alpha_p Z & p_{zz}E + \alpha_p D & p_{zx}Z^t - p_{zy}E \\ p_{xz}D - p_{yz}Z^t & p_{yz}E - p_{xz}Z & p_{xy}Z + p_{yx}Z^t - p_{xx}D - p_{yy}E \end{pmatrix}.
\end{aligned}
\tag{14}
$$

Note that the matrices for the ring resonator are defined as in Eq. (13) where $x \mapsto r$ and $y \mapsto z$. To compact the notation, we also defined the matrices

$$\begin{aligned} &S = \tfrac{3}{2}R + N, \qquad T = \tfrac{1}{2}R + N, \qquad X = \tfrac{3}{2}J^t + Z, \qquad Y = \tfrac{1}{2}J^t + Z, \\ &U = \tfrac{9}{4}R + \tfrac{3}{2}\left(N + N^t\right) + E, \qquad V = \frac{3}{4}R + \tfrac{3}{2}N + \tfrac{1}{2}N^t + E, \\ &W = \tfrac{1}{4}R + \tfrac{1}{2}N + \tfrac{1}{2}N^t + E. \end{aligned} \tag{15}$$

Therefore, M, L, C and K for the ring resonator are

$$\begin{aligned} M &= \begin{pmatrix} -p_{zz}R & 0 & p_{rz}R \\ 0 & \alpha_p R & 0 \\ p_{zr}R & 0 & -p_{rr}R \end{pmatrix}, \qquad L = \begin{pmatrix} P & 0 & 0 \\ 0 & -P & 0 \\ 0 & 0 & P \end{pmatrix}, \\ C &= \begin{pmatrix} p_{\theta z}J^t - p_{z\theta}J & p_{zr}J - p_{zz}S - \alpha_p S^t & p_{z\theta}T - p_{\theta r}J^t \\ p_{rz}J^t - p_{zz}S^t - \alpha_p S & 0 & p_{zr}S^t - \alpha_p J - p_{rr}J^t \\ -p_{\theta z}T^t + p_{r\theta}J^t & p_{rz}S - \alpha_p J^t - p_{rr}J & p_{\theta r}T^t - \alpha_p T \end{pmatrix}, \\ K &= \begin{pmatrix} p_{\theta\theta}D + \alpha_p U & -p_{\theta r}D + p_{\theta z}X & \alpha_p X^t - p_{\theta\theta}Y \\ p_{r\theta}D - p_{z\theta}X^t & p_{rz}X + p_{zr}X^t - p_{rr}D - p_{zz}U & p_{z\theta}V - p_{r\theta}Y \\ \alpha_p X - p_{\theta\theta}Y^t & -p_{\theta z}V^t + p_{\theta r}Y^t & p_{\theta\theta}W + \alpha_p D \end{pmatrix}. \end{aligned} \tag{16}$$

References

1. Berini, P., Wu, K.: Modeling lossy anisotropic dielectric waveguides with the method of lines. IEEE Trans. Microwave Theory Tech. **44**(5), 749–759 (1996)
2. Chew, W.: Waves and Fields in Inhomogeneous Media. Wiley-IEEE Press, New York (1999)
3. Fallahkhair, A.B., Li, K.S., Murphy, T.E.: Vector finite difference modesolver for anisotropic dielectric waveguides. IEEE/OSA J. Lightwave Technol. **26**(11), 1423–1431 (2008)
4. Gabriel, G.J., Brodwin, M.E.: The solution of guided waves in inhomogeneous anisotropic media by perturbation and variational methods. IEEE Trans. Microwave Theory Tech. **13**(3), 364–370 (1965)
5. Jackson, J.D.: Classical Electrodynamics, 3rd edn. Wiley, New York (1999)
6. Jin, J.: The Finite Element Method in Electro-Magnetics, 2nd edn. Wiley, New York (2002)
7. Joannopoulos, J.D., Johnson, S.G., Winn, J.N., Meade, R.D.: Photonic Crystals: Molding the Flow of Light, 2nd edn. Princeton University Press, Princeton (2007)
8. Kakihara, K., Kono, N., Saitoh, K., Koshiba, M.: Full-vectorial finite element method in a cylindrical coordinate system for loss analysis of photonic wire bends. Opt. Express **14**(23), 11128–11141 (2006)
9. Kim, S., Gopinath, A.: Full-vectorial finite element method in a cylindrical coordinate system for loss analysis of photonic wire bends. IEEE/OSA J. Lightwave Technol. **14**(9), 2085–2092 (2006)
10. Konrad, A.: High-order triangular finite elements for electromagnetic waves in anisotropic media. IEEE Trans. Microwave Theory Tech. **25**(5), 353–360 (1977)
11. Landau, L.D., Lifshits, E.M.: Electrodynamics of Continuous Media. A Course of Theoretical Physics, vol. 8. Pergamon, New York (1960)

12. Lu, Y., Fernandez, F.A.: An efficient finite element solution of inhomogeneous anisotropic and lossy dielectric waveguides. IEEE Trans. Microwave Theory Tech. **41**(6), 1215–1223 (1993)
13. Mur, G.: Edge elements, their advantages and their disadvantages. IEEE Trans. Magn. **30**(5), 3552–3557 (1994)
14. Photon Design: Integrated optics software FIMMWAVE 4.1. http://www.photond.com/
15. Pintus, P.: Accurate vectorial finite element mode solver for magneto-optic and anisotropic waveguides. Opt. Express **22**(13), 15737–15756 (2014)
16. Pintus, P., Tien, M.C., Bowers, J.E.: Design of magneto-optical ring isolator on SOI based on the finite element method. IEEE Photon. Technol. Lett. **23**(22), 1670–1672 (2011)
17. Prkna, L., Hubálek, M., Ctyroký, J.: Field modeling of circular microresonators by film mode matching. IEEE J. Sel. Top. Quantum Electron. **11**(1), 217–223 (2005)
18. Rahman, B.M.A., Davies, J.B.: Penalty function improvement of waveguides solution by finite elements. IEEE Trans. Microwave Theory Tech. **32**(8), 922–928 (1984)
19. Selleri, S., Zoboli, M.: Performance comparison of finite-element approaches for electromagnetic waveguides. J. Opt. Soc. Am. A **14**(7), 1460–1466 (1997)
20. Sher, S.M., Pintus, P., Di Pasquale, F., Bianconi, M., Montanari, G.B., De Nicola, P., Sugliani, S., Prati, G.: Design of 980nm-pumped waveguide laser for continuous wave operation in ion implanted $Er : LiNbO_3$. IEEE J. Quantum Electron. **47**(4), 526–533 (2011)
21. Sudbø, A.S.: Film mode matching: a versatile numerical method for vector mode field calculations in dielectric waveguides. Pure Appl. Opt. **2**(3), 211–233 (1993)
22. Tisseur, F., Meerbergen, K.: The quadratic eigenvalue problem. SIAM Rev. **43**(2), 235–286 (2001)
23. Ye, W.N., Xu, D.X., Janz, S., Cheben, P., Picard, M.J., Lamontagne, B., Tarr, N.G.: Birefringence control using stress engineering in silicon-on-insulator (soi) waveguides. IEEE/OSA J. Lightwave Technol. **23**(3), 1308–1318 (2005)

Author Index

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7

Subject Index

G. Russo et al. (eds.), *Progress in Industrial Mathematics at ECMI 2014*,
Mathematics in Industry 22, DOI 10.1007/978-3-319-23413-7

B

Q

R

Y

Z