Mathematics in Industry 26
The European Consortium for Mathematics in Industry

Peregrina Quintela · Patricia Barral
Dolores Gómez · Francisco J. Pena
Jerónimo Rodríguez · Pilar Salgado
Miguel E. Vázquez-Méndez *Editors*

Progress in Industrial Mathematics at ECMI 2016

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Springer

MATHEMATICS IN INDUSTRY **26**

THE EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

SUBSERIES

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

Peregrina Quintela • Patricia Barral •
Dolores Gómez • Francisco J. Pena •
Jerónimo Rodríguez • Pilar Salgado •
Miguel E. Vázquez-Méndez
Editors

Progress in Industrial Mathematics at ECMI 2016

Springer

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Editors
Peregrina Quintela
Department of Applied Mathematics
University of Santiago de Compostela
Santiago de Compostela, Spain

Patricia Barral
Department of Applied Mathematics
University of Santiago de Compostela
Santiago de Compostela, Spain

Dolores Gómez
Department of Applied Mathematics
University of Santiago de Compostela
Santiago de Compostela, Spain

Francisco J. Pena
Department of Applied Mathematics
University of Santiago de Compostela
Santiago de Compostela, Spain

Jerónimo Rodríguez
Department of Applied Mathematics
University of Santiago de Compostela
Santiago de Compostela, Spain

Pilar Salgado
Department of Applied Mathematics
University of Santiago de Compostela
Santiago de Compostela, Spain

Miguel E. Vázquez-Méndez
Department of Applied Mathematics
University of Santiago de Compostela
Lugo, Spain

ISSN 1612-3956 ISSN 2198-3283 (electronic)
Mathematics in Industry
The European Consortium for Mathematics in Industry
ISBN 978-3-319-63081-6 ISBN 978-3-319-63082-3 (eBook)
DOI 10.1007/978-3-319-63082-3

Library of Congress Control Number: 2017961824

Mathematics Subject Classification (2010): 15-xx, 34-xx, 35-xx, 37-xx, 39-xx, 41-xx, 47-xx, 49-xx, 60-xx, 62-xx, 65-xx, 68-xx, 70-xx, 74-xx, 76-xx, 78-xx, 80-xx, 90-xx, 91-xx, 92-xx, 93-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

Preface

This volume contains the proceedings of the *19th European Conference on Mathematics for Industry* (ECMI 2016) held in Santiago de Compostela, Spain, from June 13 to June 17, 2016.

Under the auspices of the European Consortium for Mathematics in Industry (ECMI), the European Conferences on Mathematics for Industry are organized every 2 years with the aim to reinforce the interaction between academy and industry, leading to innovation in both fields. These conferences also encourage industrial sectors to propose challenging problems where mathematicians can provide insight and new ideas. They are one of the main forums where significant advances in industrial mathematics are presented, bringing together prominent figures from business, science, and academia to promote the use of innovative mathematics to industry.

ECMI 2016 was jointly organized by the Department of Applied Mathematics at the Universidade de Santiago de Compostela (USC) and the Spanish Network for Mathematics and Industry (math-in). The conference was a great success, attracting more than 350 participants from about 40 countries, involving the 5 continents. Although the majority came from Europe, representing 26 countries, there was also an important representation from Australia, America (Canada, Mexico, and the USA), Africa (Nigeria, Sudan, Tanzania, and Uganda), and Asia (China, India, Israel, the Philippines, and Japan).

The ECMI 2016 scientific program consisted of 10 plenary talks by some of the leading researchers in industrial mathematics, 39 minisymposia in specific areas covering a wide variety of recent developments, and 19 sessions of contributed talks. In short, a total of 306 presentations was distributed along 5 days, with 8 parallel sessions per day. In these presentations, there were, directly or indirectly involved, about 50 companies that, in one way or another, funded through collaborations the research presented to meet their specific demands. Many success stories from industry collaborations were presented. We would like to thank them all for their support as it contributed to the success of this event and was crucial to allow many young researchers to participate in ECMI 2016. In order to facilitate this

Group photo at the Gala dinner (forecourt of Hostal dos Reis Católicos)

academic-industry interaction, the European Consortium for Mathematics in Industry in collaboration with the European Mathematical Society launched a European one-stop shop, the European Service Network of Mathematics for Industry and Innovation (EU-MATHS-IN), to provide an agile access to the most advanced mathematical techniques to companies.

The conference program also paid special attention to establishing discussion forums in various fields, such as mathematics in the H2020 program, master's programs related to industrial mathematics, or study groups as a tool for dissemination and promotion of mathematical technology.

In this 19th edition, a wide variety of applications were presented covering problems in *electronics* (15%), *energy and environment* (14%), and *mechanics and mechatronics* (12%), among the industrial sectors with the highest number of talks. When classifying the talks according to the societal challenges, the EU Framework Programme for Research and Innovation H2020 observed that 14% of them fell into the *Climate Action, Environment, Resource Efficiency and Raw Materials* challenge and 13% into *Health, Demographic Change and Wellbeing*, while 12% belonged to *Europe in a Changing World—Inclusive, Innovative and Reflective Societies*. These percentages clearly show that mathematics is a cross-cutting technology through all industrial sectors and all societal challenges.

Special mention must be made to the dissemination event *The Mathematical Way to the Oscars* which was held in the Auditorium of Abanca (Pazo de Ramirás) and featured the participation of Professor Joseph M. Teran, from the University of California, as invited speaker from academia, and two representatives from Spanish companies familiar with the use of mathematical technology in movie production.

Three special lectures are included in all ECMI conferences:

- The Alan Tayler Memorial Lecture was established to honor Alan Tayler who was one of the founding members of ECMI. The 2016 Alan Tayler Memorial Lecture, *Mathematical Modelling of Lithium Ion Batteries*, was delivered by S. Jon Chapman, from the University of Oxford.
- The Anile-ECMI Prize for Mathematics in Industry was established to honor Professor Angelo Marcello Anile (1948–2007) of Catania. The prize is dedicated to young researchers with excellent PhD theses in industrial mathematics. The 2016 Anile-ECMI Prize for Mathematics in Industry was awarded to Francesc Font, of the University of Limerick, who participated at ECMI 2016 with the talk *Influence of Substrate Melting on the Laser-Induced Dewetting of Nanothin Films*.
- The Hansjörg Wacker Memorial Prize was established in memory of ECMI founding member Hansjörg Wacker (1939–1991). The prize is awarded for the best mathematical dissertation at the master's level on an industrial project written by a student from an ECMI institution. The 2016 Hansjörg Wacker Memorial Prize was awarded to Elisa Riccietti from the Università degli Studi Firenze, who participated at ECMI 2016 with the talk *Numerical Methods for Optimization Problems: An Application to Energetic Districts*.

A book of abstracts of ECMI 2016 gathering information about all the conference talks was published by the University Press of Santiago de Compostela in its collection *Cursos e Congressos*.

Continuing the tradition of the ECMI conferences, a new honorary member was appointed. This honor was awarded to Alfredo Bermúdez de Castro, professor of applied mathematics at the Universidade de Santiago de Compostela. Professor Bermúdez is a pioneering mathematician who for the past 30 years has been involved in developing new mathematical technologies tailored to solve industrial problems in several fields such as solid mechanics, fluids, acoustics, electromagnetism, and chemical kinetics; he is a reference of industrial mathematics in the world.

ECMI 2016 proceedings compiles more detailed information on a good representative sample of the conference program. This book of proceedings illustrates the breakthrough of industrial mathematics in the world, how the challenges posed from real industrial problems are an engine for advancing mathematical knowledge, and the great versatility of mathematics to address them. The proceedings are classified into four parts: plenary lectures, ECMI awards, minisymposia, and contributed talks.

In Part I, three of the plenary talks are included showing the application of multidimensional semiparametric and predictive models to solve environmental problems, how mathematical modeling can help to understand the way immune system regulates the mechanisms to distinguish friends from foes upon inspection of circulating antigens, or the application of the weighted least action principle to design a specific beam shaping lens.

In Part II of these proceedings, a paper related to the 2016 Hansjörg Wacker Memorial Prize is included.

In order to be as faithful as possible to the ECMI 2016 program, the papers corresponding to minisymposia talks are classified in the same order of intervention as in the program and are listed in Part III. A brief description of the objectives and contents of the corresponding minisymposia is also included.

Finally, Part IV includes the proceedings of contributed talks sorted alphabetically by author.

ECMI 2016 received generous support from ECMI, the Universidade de Santiago de Compostela, the Spanish Network for Mathematics and Industry, the Spanish Ministry of Economy and Competitiveness, the US Naval Research Office, the Thematic Network of Mathematics and Industry, the Technological Institute for Industrial Mathematics (ITMATI), and the Galician Network of Industrial Mathematics (Red TMATI) funded by the Xunta de Galicia. We also had the collaboration of several institutions or companies such as GDI, Iberia, Renfe, Springer, Xacobeo Galicia, and Santiago Turismo. We would like to thank them all for their support as it contributed to the success of this event and was crucial to allow many young researchers to participate in ECMI 2016.

We would like to address our warmest thanks to the invited speakers, T. Abboud (France), S.J. Chapman (UK), L.J. Cummings (USA), L. Formaggia (Italy), W. González (Spain), M.A. Herrero (Spain), B. Kaltenbacher (Austria), P.M. Pardalos (USA), K. Rubinstein (USA), and J.M. Teran (USA), for coming to Santiago de Compostela and contributing to the success of the conference with the high quality of their presentations. We are also greatly indebted to the members of the Scientific Committee (A. Bermúdez, D. Hömberg, S. O'Brien, A. Bátkai, A.K. Belyaev, A.L. Bertozzi, L. Bonilla, E. Carrizosa, L. P. Cook, P. Joly, T. Kauranne, T. Myers, J. Ockendon, A. Quarteroni, G. Russo, O. Scherzer, B. Wagner, and G. Wake) for their efforts to select the excellent invited speakers of ECMI 2016. We wish to thank also all participants, organizers of special sessions, chairs, and attendees of the different sessions, for their contributions and attendance, without whom there would have been no conference. Special thanks go to the contributors of this volume.

Finally, we would like to express our gratitude to Elisa Eiroa, manager of the research group in mathematical engineering (mat+i); Fe Sampayo, technology translator in math-in; Manuel Porto; secretary of the Department of Applied Mathematics; and Carlos Grela, technical assistant of the mat+i group, for their help in organizing this event. In this section of appreciation, we cannot forget the important support given by the conference assistants, Begoña, Cristina, Ernesto, Javier, Juan, Luis, Manuel, Marcos, and Pedro; all of them are students of the master's program in industrial mathematics and therefore representatives of the future generation of ECMI people. Finally, we would like to thank the faculties of biology and mathematics at the Universidade de Santiago de Compostela, which provided the spaces for this conference.

We members of the Organizing Committee, and also editors of this volume, were deeply involved in the preparation of ECMI 2016 and very much enjoyed this

experience especially in working with the numerous participants. We are happy to complete our work by editing this volume and wish that you will find the ECMI 2016 proceedings interesting, stimulating, and a great experience.

Santiago de Compostela and Lugo
March 2017

Peregrina Quintela
Patricia Barral
Dolores Gómez
Francisco J. Pena
Jerónimo Rodríguez
Pilar Salgado
Miguel E. Vázquez-Méndez

Contents

Part I
Plenary Lectures

Semiparametric Prediction Models for Variables Related with Energy Production

Wenceslao González-Manteiga, Manuel Febrero-Bande, and María Piñeiro-Lamas

Abstract In this paper a review of semiparametric models developed throughout the years thanks to extensive collaboration between the Department of Statistics and Operations Research of the University of Santiago de Compostela and a power station located in As Pontes (A Coruña, Spain) property of Endesa Generation, SA, is shown. In particular these models were used to predict the levels of sulfur dioxide in the environment of this power station with half an hour in advance. In this paper also a new multidimensional semiparametric model is considered. This model is a generalization of the previous models and takes into account the correlation structure of errors. Its behaviour is illustrated in the prediction of the levels of two important pollution indicators in the environment of the power station: sulfur dioxide and nitrogen oxides.

1 Introduction: An Environmental Problem

The coal-fired power station in As Pontes is one of the production centers owned by Endesa Generation SA in the Iberian Peninsula. It is located in the town of As Pontes de García Rodríguez, northeast of A Coruña province.

This power station was designed and built to make use of lignite from the mine located in its vicinity. This solid fuel is characterized by its high moisture and sulphur contents and its low calorific value. Throughout the years the plant has undergone several transformation processes in their facilities with the aim of reducing emissions of sulphur dioxide (SO_2). The power station completed its last

W. González-Manteiga (✉) • M. Febrero-Bande
Faculty of Mathematics, University of Santiago de Compostela, Rúa Lope Gomez de Marzoa, s/n, Santiago de Compostela, Spain
e-mail: wenceslao.gonzalez@usc.es; manuel.febrero@usc.es

M. Piñeiro-Lamas
CIBER Epidemiología y Salud Pública, Complexo Hospitalario da Universidade de Santiago, Santiago de Compostela, Spain
e-mail: maria.pineiro@usc.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_1

adaptation in 2008 to consume, as primary fuel, imported subbituminous coal, characterized by its low sulphur and ash contents.

The location of the power plant close to natural sites of high ecological value, such as the Natural Park *As Fragas do Eume* and existing legislation, mean that it has existed since the beginning a great concern for its impact on the environment. Therefore the station has a *Supplementary Control System of Air Quality* that allows it to make changes in operating conditions in order to reduce emissions when the weather conditions are adverse to the spread of the emitted smoke plume, specifically containing SO_2, and there are significant episodes of impaired air quality. Spanish law, by rules and regulations sets maximum concentrations that can be achieved for these gases in a given period of time. In particular, for this plant the only limit that might be exceeded at any time, is one that is established on the 1 h mean from the concentration of SO_2 in the soil, the value of $350\,\mu g/m^3$.

Then the problem is to be able to predict using the information received continuously at sampling stations and the past information, the future values for SO_2 levels. Statistical forecast models are the key to get these predictions and suggest a course of action to the plant operators.

In recent years changes in environmental legislation and in the power station itself, and the construction of a new station combined cycle natural gas require the design models to obtain the simultaneous prediction of two pollution indicators in the environment. The fuels that are going to be used make that the main interest lies in predicting the values of the nitrogen oxides (NO_x) simultaneously with the values of SO_2.

All these changes have created a new problem: predicting 1 h mean concentrations of sulphur dioxide and nitrogen oxides, measured in the environment of the two facilities. Faced with this new approach, the statistical forecast models are again an effective tool. Thus, a multidimensional prediction general model is designed (see Sect. 3).

2 One-Dimensional Predictive Models

2.1 Models Designed to Solve the Environmental Problem

Resulting from the collaboration over the past years between the Department of Statistics and Operations Research at the University of Santiago de Compostela and the Environment Section of the power station, the *Inmission Statistical Forecasting System* (SIPEI, in Spanish) have been created employing statistical models to provide predictions for the levels of SO_2 with a half an hour horizon.

Due to data availability with minutal frequency in real-time and current legislation, is considered the 1 h mean from both of the values of SO_2 and NO_x, for predictions of future values of both pollutants. Thus two time series are constructed, X_{1t} and X_{2t}, for which the subscript t represents a minutal instant, and each value

will be an average of the actual values for the last hour:

$$X_{1t} = \frac{1}{60}\sum_{i=0}^{59} \mathrm{SO}_2(t-i) \text{ and } X_{2t} = \frac{1}{60}\sum_{i=0}^{59} \mathrm{NO}_x(t-i),$$

where $\mathrm{SO}_2(t)$ and $\mathrm{NO}_x(t)$ represent the concentration of SO_2 and NO_x , respectively, at time t, measured in $\mu\mathrm{g}/\mathrm{m}^3$.

The series of 1 h SO_2 means has a characteristic behaviour, highly influenced by weather conditions and local topography. It takes values close to zero for long periods of time, and it can suddenly and sharply increase (episodes) in bad weather for the dispersion of the smoke plume. Nowadays, the serie of 1 h NO_x has a behaviour similar to that of SO_2, but on a smaller scale (see Fig. 1). The main objective of the developed statistical models is to predict the episodes, so our interest is centred on the values that occur less along the time series. Because of this, a kind of memory called *Historical Matrix* was designed [13], which will be essential to the behaviour of all developed models so far. This matrix is composed of a large number of vectors based on $(X_{t-l}, \ldots, X_t, X_{t+k})$: real data of bihourly SO_2 or NO_x means, chosen so as to cover the full range of variable in question and make the role

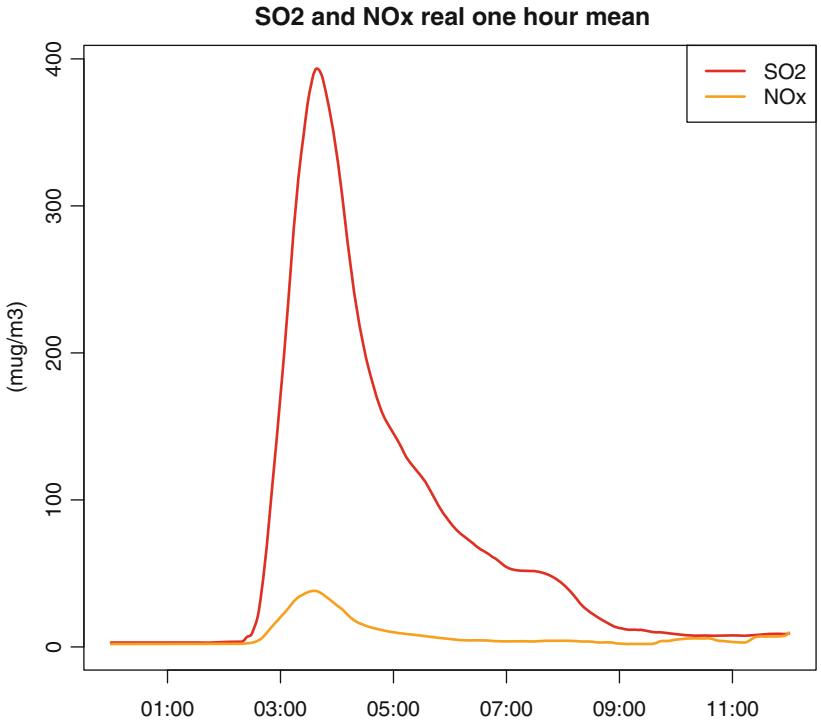


Fig. 1 Episode depicted in one of sampling stations

of historical memory. To ensure that cover the entire range of the variable, the matrix is divided into blocks according to the level of the response variable, X_{t+k}. To update the memory, in every instant, when a new observation is received, the historical matrix is renewed in the following way: the class to which the new observation belongs is found and then the oldest datum in such class leaves the matrix and the new observation enters it. With a sample built this way, makes sure that always have updated information on the full variation range of the interest variable, and over the years this concept has been adapted to the different statistical techniques used.

2.1.1 The First Semiparametric Model

In the early years of development, the data transmission frequency to SIPEI was pentaminutal, and also, the legislation in force at that time sets the limit values for the 2 h mean of the SO_2. For this reason, the prediction models for SO_2 levels initially worked with series of bihourly means. The objective was to obtain the prediction, with a half an hour horizon, for this time series. Therefore, each time it receives a new observation, X_t, it has to predict the value at six times ahead, X_{t+6}.

A *semiparametric approach* was considered [8] which generalizes the Box–Jenkins models as follows:

$$X_{t+\kappa} = \varphi_\kappa(X_t, X_{t-l}) + Z_{t+\kappa},\ \kappa, l \in \mathbb{Z}^+$$

where Z_t has an ARIMA structure independent of X_t [1].

In particular, at each time t the regression function $\varphi_6(X_t, X_{t-1}) = \mathbb{E}(X_{t+6}/X_t, X_{t-1})$ is estimated with the Nadaraya–Watson kernel type estimator using the information provided by the historical matrix. The second step is to calculate the residual time series $\hat{Z}_{t-64}, \ldots, \hat{Z}_t$ relative to the last 6 h, where $\hat{Z}_i = X_i - \hat{\mathbb{E}}(X_i/X_{i-6}, X_{i-7})$ for each i and fits an appropriate ARIMA model for it. Finally we get the Box-Jenkins prediction of $\hat{Z}_{t+6}$. The final point prediction proposed is given by: $\hat{\mathbb{E}}(X_{t+6}/X_t, X_{t-1}) + \hat{Z}_{t+6}$.

2.1.2 Partially Linear Model

The information used by the previous semiparametric models to obtain the predictions is the past of the time series; however it might be useful to introduce additional information in order to improve these predictions. Specifically, meteorological and emission variables have been used with, the so–called *partially linear models* [14] to estimate 2 h mean values of SO_2 with an hour in advance.

Data in the form of (V_t, Z_t, Y_t) is considered, where V_t is a vector of exogenous variables, $Z_t = (X_t, X_{t-s})$ and $Y_t = X_{t+12}$ being X_t the series of bihourly SO_2 means; and is assumed that this series conform to the following partially linear model: $Y_t = V_t'\beta + \varphi(Z_t) + \epsilon_t$, where ϵ_t is an error term of mean equals to zero.

2.1.3 Neural Networks

The change in the interest series established by the European Council Directive 1999/30/CE, from 2 h means to 1 h means, causes the time series to be less smooth. At the beginning was adapted the semiparametric model designed to work on the new series of 1 h means. The results showed a considerable increase in terms of the variability of the given predictions, regarding the results usually obtained for the series of 2 h means.

In an attempt to improve the response given by the SIPEI, and in particular, its point predictions with half an hour horizon, new predictors based on *neural networks models* were developed [5].

The neural network has been designed to provide predictions, with half an hour in advance, 1 h mean values of SO_2. It consists of an input layer, one hidden layer and an output layer. The number of nodes in the output layer is determined by the size of the response to be obtained from the network; in this case interested in a prediction for X_{t+6}. As input to the network it has been taken the bidimensional vector (X_{t-3}, X_t) and the nodes in the hidden layer have been taken as the activation function of a logistic function, and in the output layer the identity function. The predictor given by the neural network has the following expression:

$$\hat{X}_{t+6} = o_1 = \sum_{j=1}^{L} \omega_{1j}^{o} f_j^h(\theta_j^h + \omega_{j1}^h X_{t-3} + \omega_{j2}^h X_t)$$

with $f_j^h(z) = \frac{1}{1+e^{-z}}$.

The weights $\{\omega_{j1}^h, \omega_{j2}^h, \omega_{1j}^o; j = 1, \ldots, L\}$ and the trends $\{\theta_j^h; j = 1, \ldots, L\}$ are determined during the training process, as well as the final L number of hidden layer nodes, that is chosen like the value which neural network provides better results, after having trained networks with identical architecture and different values of L. To design the training set of the neural network it have been considered historical matrices, formerly introduced, suitably adapted.

Figure 2 shows the forecasts given half an hour before by the neural network with 50 nodes in its hidden layer for an episode depicted in one of the measuring stations. The good behavior of the forecast (dotted line) can easily be seen. The procedures based on neural networks accurately predict the real 1 h mean SO_2 air quality values (solid line). These models were optimized later with boosting learning techniques [4].

2.1.4 Functional Data Model

The 1 h mean values of SO_2 can be treated as observations of a stochastic process in continuous time. The interest is, as it was discussed above, to predict a half-hour horizon, so that each of the curves is an interpolated data on half an hour. In this case curves were obtained by considering six pentaminutal consecutive observations,

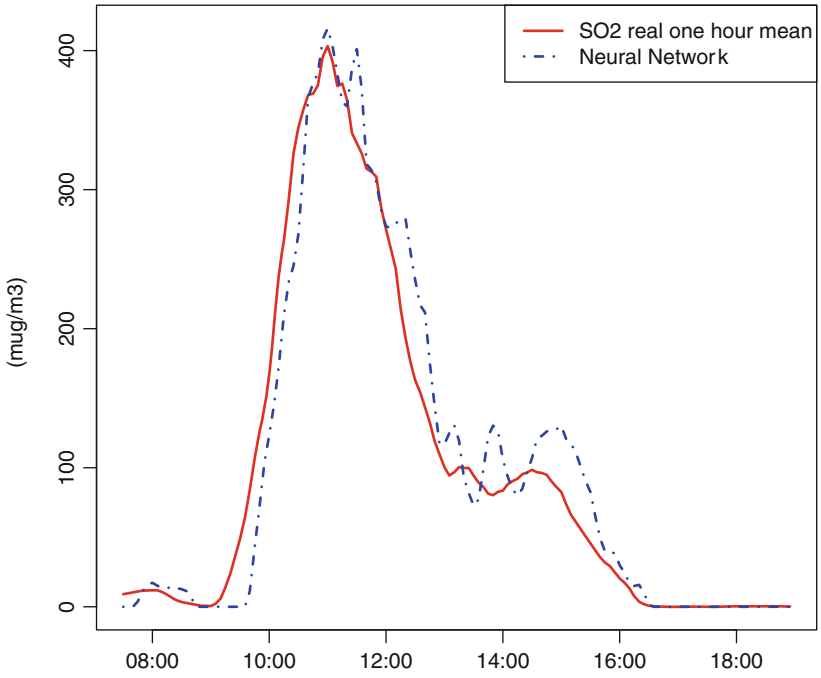


Fig. 2 Episode depicted in one of sampling stations. Prediction given by the neural network [5]

with sampling points for each functional data. Therefore, we use random variables with values in Hilbert space $H = L^2([0, 6])$ with the form $X_t(u) = x(6t + u)$.

The following statistical model is considered $X_t = \rho(X_{t-1}) + \epsilon_t$, where ϵ_t is a Hilbertian strong white noise and $\rho : H \to H$ is the operator to estimate. For the estimation of ρ, a *functional kernel estimator* has been used in the *autoregressive of order-one Hilbertian framework*. Furthermore, it has been conveniently adapted the concept of historical matrix to the case where the data are curves [6].

2.1.5 Other Approaches Designed to Predict Probabilities

The models described, so far, provide point predictions of SO_2, but other techniques have also been developed in order to predict probabilities. The aim of these alternative models is to estimate the probability that the series of bihourly SO_2 measures exceeds a certain level r with an hour anticipation, namely in our case, we predict $\mathbb{P}(Z_t) = \mathbb{P}(X_{t+12} > r|Z_t)$ being $Z_t = (X_t, X_t - X_{t-3})$. To do it *additive models with an unknown link function* [15] have been used.

It has also been considered more complex *generalized additive models* (GAM) with second-order interaction terms [16]. They have shown that the GAM with interactions detects the onset of episodes earlier than it does GAM on its own.

2.2 Alternative One-Dimensional Models: Additive Models

In the statistical literature there is a wide range of one-dimensional models which can be used to predict the levels of SO_2. We will focus on the techniques we will use in the next section to construct our multidimensional model: additive models for continuous response.

There have been a number of proposals for fitting the additive models. Friedman and Stuetzle [7] introduced a backfitting algorithm and [2] studied its properties. Mammen et al. [12] proposed the so called *smooth backfitting* by employing projection arguments. Let $\{(Y_t, Z_t)\}_{t=1}^{T}$ be a random sample of a strictly stationary time series, with Y_t one-dimensional and Z_t q-dimensional following the model:

$$Y_t = m(Z_t) + \epsilon_t \, t \in \mathbb{Z} \tag{1}$$

where $\{\epsilon_t\}$ is a white noise process and $\mathbb{E}[\epsilon_t | Z_t] = 0$.

Typically is assumed that the function m is additive with component functions m_j, for $j = 0, \ldots, q$, thus

$$Y_t = m_0 + m_1(Z_{1t}) + \ldots + m_q(Z_{qt}) + \epsilon_t \tag{2}$$

A generalized kernel nonparametric estimation can be given using *smooth backfitting* for the functions $m_1, \ldots, m_q$ (see again the above mentioned papers).

3 Multidimensional Semiparametric Prediction

The new goal is to incorporate the prediction of NO_x with half an hour in advance, as well as to continue getting the predictions of SO_2, as has already been commented. The idea is to generalize the one-dimensional semiparametric approach proposed by García-Jurado et al. [8] taking into account the structure of correlation between the vectorial series that is intended to predict.

3.1 The Model

In general, the following vectorial model is considered

$$Y_t = \varphi(Z_t) + \epsilon_t, \; t \in \mathbb{Z} \tag{3}$$

where $Y_t = (Y_{1t}, \ldots, Y_{rt})$, $Z_t = \left(Z_{1t}, \ldots, Z_{qt}\right)$ and $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{rt})$, where ϵ_t has a VAR(p) structure of the form

$$\epsilon_t = \sum_{i=1}^{p} \Phi_i \epsilon_{t-i} + \xi_t \quad \text{for all } \ t \in \mathbb{Z},$$

independent of Z_t, where the Φ_i are fixed $(r \times r)$ coefficients matrices and ξ_t is a r-dimensional white noise process, i.e. $\mathbb{E}(\xi_t) = 0$, $\mathbb{E}(\xi_t \xi_t') = \Sigma_\xi$ and $\mathbb{E}(\xi_t \xi_s') = 0$ for $l \neq s$.

Our main objective is to predict the vector Y_t using a sample of size T, κ instants ahead. The prediction $\hat{Y}_t$ of Y_t is defined by

$$\dot{Y}_t = \hat{\varphi}_\kappa(Z_t) + \dot{\epsilon}_{t+\kappa} \tag{4}$$

where $\hat{\varphi}_\kappa$ is a nonparametric estimate of φ and $\dot{\epsilon}_{t+\kappa}$ the prediction given, κ instants ahead, for the residual series constructed as $\hat{\epsilon}_t = Y_t - \hat{\varphi}_\kappa(Z_t)$.

3.2 Estimations

We suppose that the model 3 is verified. The first step is to make a nonparametric estimation of φ independently for each of the r components of Y_t: $\varphi(Z_t) = (\varphi^1(Z_t), \ldots, \varphi^r(Z_t))$. Furthermore, we assume that the functions φ^k are additive with component functions $\varphi^{k,j}$, for $k = 1, \ldots, r$ and $j = 0, \ldots, q$, thus

$$\varphi^k(Z_t) = \varphi^{k,0} + \varphi^{k,1}(Z_{1t}) + \ldots + \varphi^{k,q}(Z_{qt}), \ k = 1, \ldots, r. \tag{5}$$

Therefore, we estimate r additive models with q covariates using the smooth backfitting techniques which have been cited in the previous section (see also, [10] for more details).

3.3 Other Considerations: The Phenomenon of Cointegration

Sometimes the vectorial processes can be cointegrated, so one has to take into account the *structure of correlation* between the series. The notion of cointegration has been one of the most important concepts in time series since [9] and [3] that formally developed it. The issue has broad applications in the analysis of economic data as well as several publications in the economic literature.

Let $Y_t = (Y_{1t}, \ldots, Y_{rt})'$ be a vector of r time series integrated of order 1 ($I(1)$). Y_t is said to be *cointegrated* if a linear combination of them exists that it is stationary

(I(0)), i.e., if there exists a vector $\beta = (\beta_1, \ldots, \beta_r)'$ such as

$$\beta' Y_t = \beta_1 Y_{1t} + \ldots + \beta_r Y_{rt} \sim I(0)$$

The vector β is called the *cointegration vector*. This vector is not unique since for any scalar c the linear combination $c\beta' Y_t = \beta^* Y_t \sim I(0)$. Therefore, normalization is often assumed to identify a unique β. A typical normalization is $\beta = (1, -\beta_2, \ldots, -\beta_r)'$.

Johansen [11] addresses the issue of the cointegration within an error correction model in the framework of vector autoregressive models (VAR). Consider then a general model VAR(p) for the vector of r series Y_t

$$Y_t = \Phi_0 D_t + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + \xi_t, \ t = 1, \ldots, T$$

where D_t contains deterministic terms (constant, trend, ...).

Suppose Y_t is $I(1)$ and possibly cointegrated. Then the VAR representation is not the most suitable representation for analysis because the cointegrating relationships are not explicitly apparent. The cointegrating relationships become apparent if the VAR model is transformed to the *vector error correction model* of order p (VECM(p))

$$\Delta Y_t = \Phi_0 D_t + \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \ldots + \Gamma_{p-1} \Delta Y_{t-p+1} + \xi_t$$

where $\Pi = \Phi_1 + \ldots + \Phi_p - I_r$, $\Gamma_k = -\sum_{j=k+1}^{p} \Phi_j, \ k = 1, \ldots, p-1$ and $\Delta Y_t = Y_t - Y_{t-1}$. The matrix Π is called the *long-run impact matrix* and Γ_k are the *short-run impact matrices*. Moreover the rank of the singular matrix Π provides information on the number of cointegration relations that exist, i.e., *the rank of cointegration*. Johansen proposes a sequential procedure of likelihood ratio tests to estimate this range.

3.4 Prediction Scheme

We present now the prediction scheme step by step:

1. Every instant t, $\varphi(Z_t)$ is estimated with the smooth backfitting technique independently for each of r components. The estimate of $\varphi = \varphi_\kappa$ is done κ instants ahead: $\hat{\varphi}_\kappa$.
2. The residuals series $\hat{\epsilon}_t$ is computed by

$$\hat{\epsilon}_t = Y_t - \hat{\varphi}_\kappa(Z_t)$$

3. The following step is to make an appropriate adjustment on the model error structure (VECM) and to obtain the prediction of $\hat{\epsilon}_t$, κ instants ahead: $\dot{\epsilon}_{t+\kappa}$.

4. The proposed final prediction is given by (4).

This scheme is a natural generalization of the one-dimensional prediction models.

4 Real Data Application

The general model proposed in Sect. 3.1 was implemented for the particular case of the prediction of levels of SO_2 and NO_x in the vicinity of power station and combined cycle.

Let X_t be the bidimensional series formed by the 1 h mean series of SO_2 and NO_x at instant t. In terms of Eq. (4), we consider $Y_t = X_{t+\kappa}$ and $Z_t = (X_t, X_t - X_{t-5})$. If $\hat{X}_i$ denotes the observed values for past instants ($i \leq t$) and the best prediction for future instants ($i > t$), the aim is to predict X_{t+30} following the next algorithm:

- Every instant t, $\varphi(Z_t)$ is estimated with additives models and the information provided by the historical matrix, independently for each component. The estimate of φ is done at 30 instants ahead: $\dot{Y}_t = \dot{X}_{t+30} = \hat{\varphi}_{30}(Z_t) + \dot{e}_{t+30}$.
- The residuals series $\hat{e}_t$ is computed by $\hat{e}_t = Y_t - \hat{\varphi}_{30}(Z_t)$ and a test of model adequacy is performed (for instance, the Ljung-Box test) for each component of the series concerning the last 4 h.
- If any of the components of the residuals series is not white noise, a test is performed to explore if the vectorial residual series is cointegrated. If this is the case, an adequate VECM is adjusted. If the series is not cointegrated, a VAR model is fitted.
- Thus $\dot{e}_{t+30}$ is obtained .
- The proposed final prediction given by the *Semiparametric Bidimensional Model* with the nonparametric part estimated at 30 instants (SPBM) is:

$$\dot{X}_{t+30} = \hat{\varphi}_{30}(Z_t) + \dot{e}_{t+30}.$$

To observe the behavior of the prediction model we have evaluated its performance on two episodes of air quality alteration, whose information has not been included in the historical matrix.

Figure 3 shows the forecasts given half an hour before by the proposed models for an episode depicted in one of sampling stations. The good behavior of the forecasts can easily be seen. They estimate quite well the real 1 h mean of SO_2 and NO_x values. This is confirmed in Table 1. This table contains three measures of accuracy for the pure nonparametric predictors (NPM) and the proposed semiparametric predictor, based on the following criteria:

(a) Squared error: $\mathrm{SE} = (y_t - \hat{y}_t)^2$
(b) Absolute error: $\mathrm{AE} = |y_t - \hat{y}_t|$
(c) Relative absolute error (%): $\mathrm{RAE} = 100\left|\frac{y_t - \hat{y}_t}{y_t}\right|$

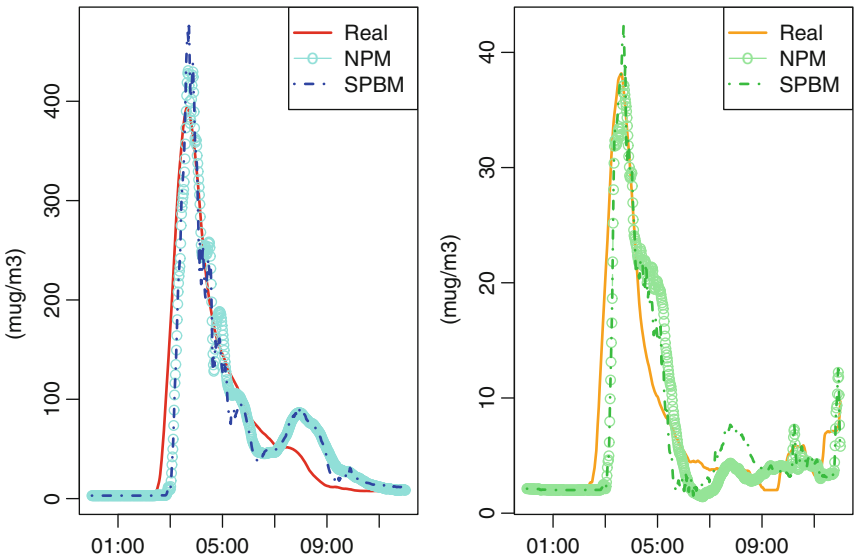


Fig. 3 Episode depicted in one of sampling stations. Predictions given by the bidimensional semiparametic models for the 1 h SO_2 (*left*) and NO_x (*right*) mean

Table 1 SO_2 and NO_x forecast errors

	SO_2						NO_x					
	SE		AE		RAE		SE		AE		RAE	
Model	*M*	*Md*	*M*	*Md*	*M*	*Md*	*M*	*Md*	*M*	*Md*	*M*	*Md*
SPBM	1265.28	635.65	27.91	25.21	27.15	18.33	15.83	8.87	3.25	2.97	21.35	9.28
NPM	1043.23	372.02	24.20	19.29	24.18	15.99	30.35	18.28	4.42	4.28	29.77	12.17

The mean (*M*) and the median (*Md*) of these three measures have been computed for the period covering the pollution incident proper (02.00–10.00 h). The SO_2 nonparametric prediction with the historical matrix captures very well the behavior of the real series (RAE: 24.18%) while the semiparametric prediction is not able to overcome (RAE: 27.15%). However, the NO_x prediction given by SPBM (RAE: 21.35%) notably improves one obtained by the NPM (RAE: 29.77%). Furthermore, the residuals series is cointegrated 123 times (8.37%), mainly when the episode higher values occur.

In another SO_2 episode depicted in one of sampling stations (see Fig. 4) the predictors behavior is some different. The SO_2 prediction given by NPM (RAE: 43.92%) does not entirely capture the behavior of the real series so, the semiparametric prediction (RAE: 38.48%) gets it done better as shown in Table 2. In this episode, the NO_x values are very low (practically zero) and therefore there are no cointegration relationships.

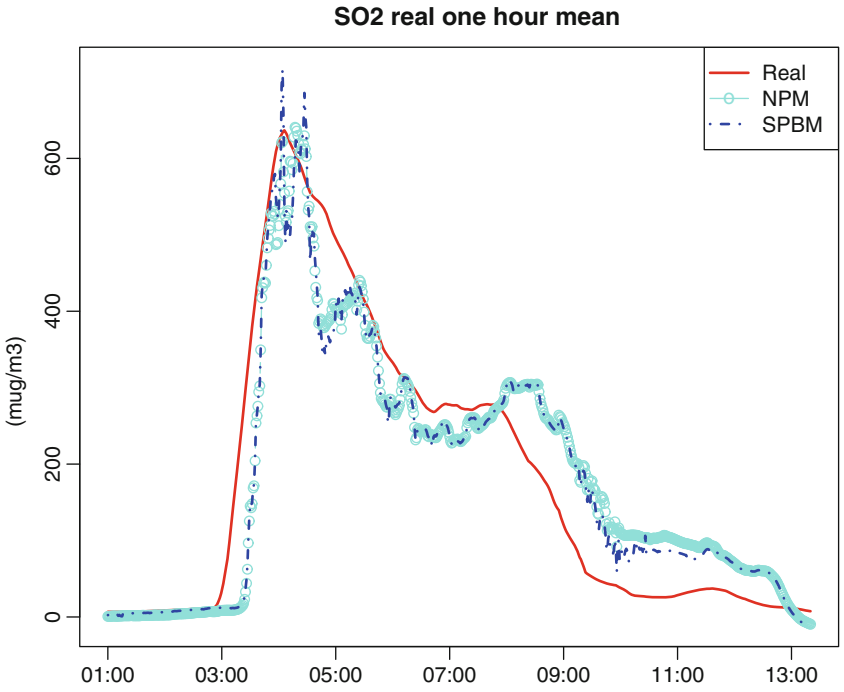


Fig. 4 Episode depicted in one of sampling stations. Predictions given by the bidimensional semiparametic models for the 1 h SO_2 mean

Table 2 SO_2 forecast errors

	SO_2					
	SE		AE		RAE	
Model	*M*	*Md*	*M*	*Md*	*M*	*Md*
SPBM	5782.18	2566.27	62.40	50.66	38.48	16.31
NPM	5833.85	3055.29	64.17	55.27	43.92	16.39

Acknowledgements The work by Wenceslao González-Manteiga and Manuel Febrero-Bande was partially supported by grants MTM2013-41383-P from Ministerio de Economía y Competitividad, Spain.

References

1. Box, G., Jenkins, M., Reinsel, C.: Time Series Analysis: Forecasting and Control. Wiley, Hoboken, NJ (2008)
2. Buja, A., Hastie, T., Tibshirani, R.: Linear smoothers and additive models. Ann. Stat. **17**, 453–510 (1989)
3. Engle, R., Granger, C.: Co-integration and error correction: representation, estimation and testing. Econometrica **57**, 251–276 (1987)

4. Fernández de Castro, B., González-Manteiga, W.: Boosting for real and functional samples: an application to an environmental problem. Stoch. Env. Res. Risk A **22**(1), 27–37 (2008)
5. Fernández de Castro, B., Prada-Sánchez, J., González-Manteiga, W., Febrero-Bande, M., Bermúdez Cela, J., Hernández Fernández, J.: Prediction of SO2 levels using neural networks. J Air Waste Manag Assoc **53**(5), 532–539 (2003)
6. Fernández de Castro, B., Guillas, S., González-Manteiga, W.: Functional samples and bootstrap for predicting sulfur dioxide levels. Technometrics **47**(2), 212–222 (2005)
7. Friedman, J., Stuetzle, W.: Projection pursuit regression. J. Am. Stat. Assoc. **76**(376), 817–823 (1981)
8. García-Jurado, I., González-Manteiga, W., Prada-Sánchez, J., Febrero-Bande, M., Cao, R.: Predicting using Box–Jenkins, Nonparametric, and Bootstrap Techniques. Technometrics **37**(3), 303–310 (1995)
9. Granger, C.: Co-integrated variables and error-correcting models. Ph.D. thesis, Discussion Paper 83-13. Department of Economics, University of California at San Diego (1983)
10. Hastie, T.J., Tibshirani, R.J.: Generalized Additive Models, vol. 43. CRC Press, Boca Raton, FL (1990)
11. Johansen, S.: Statistical analysis of cointegration vectors. J. Econ. Dyn. Control **12**(2), 231–254 (1988)
12. Mammen, E., Linton, O., Nielsen, J.: The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Ann. Stat. **27**(5), 1443–1490 (1999)
13. Prada-Sánchez, J., Febrero-Bande, M.: Parametric, non-parametric and mixed approaches to prediction of sparsely distributed pollution incidents: a case study. J. Chemom. **11**(1), 13–32 (1997)
14. Prada-Sánchez, J., Febrero-Bande, M., Cotos-Yáñez, T., González-Manteiga, W., Bermúdez-Cela, J., Lucas-Domínguez, T.: Prediction of SO2 pollution incidents near a power station using partially linear models and an historical matrix of predictor-response vectors. Environmetrics **11**(2), 209–225 (2000)
15. Roca-Pardiñas, J., González-Manteiga, W., Febrero-Bande, M., Prada-Sánchez, J., Cadarso-Suárez, C.: Predicting binary time series of SO2 using generalized additive models with unknown link function. Environmetrics **15**(7), 729–742 (2004)
16. Roca-Pardiñas, J., Cadarso-Suárez, C., González-Manteiga, W.: Testing for interactions in generalized additive models: application to SO2 pollution data. Stat. Comput. **15**(4), 289–299 (2005)

Emergent Behaviour in T Cell Immune Response

Clemente F. Arias and Miguel A. Herrero

Abstract The ability of our immune system to fight off challenges posed by pathogenic agents (external or internal) is amazing. Indeed, many times during a normal lifespan immune cells have to identify and destroy incoming threats while leaving harmless cell trafficking undisturbed. Most remarkably, this careful regulation of body function is achieved in the absence of any organ in charge of controlling immune response. The latter is just an emergent property resulting from a very limited number of individual actions taken by immune cells, using only local information from their immediate neighbourhood. We shortly review here some striking aspects of this emergent behaviour. In particular, we will focus our attention on two issues, namely the way immune system regulates the number of effector T cells required to wipe out an acute infection and the mechanisms to distinguish friends from foes upon inspection of circulating antigens.

1 Introduction

The immune system (IS) provides efficient response to meet threats coming from unwanted agents proliferating in the organism, whether of external (as in the case of microbial infection) or internal (for instance tumour growth) origin. While a substantial amount of information about the way in which IS operates is now available [5] many crucial questions concerning its function remain largely unanswered.

A key mechanism in the fight against pathogenic agents is antigen recognition, which can be briefly described as follows. Dendritic cells and macrophages, two particular types of immune cells, circulate the body taking samples of peptides from the tissues. If an infection is detected in a particular site in the organism,

C.F. Arias
Grupo Interdisciplinar de Sistemas Complejos, Madrid, Spain
e-mail: tifar@ucm.es

M.A. Herrero (✉)
Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad Complutense de Madrid, Plaza de las Ciencias, 3, 28040 Madrid, Spain
e-mail: herrero@mat.ucm.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_2

dendritic cells and macrophages increase their phagocytic activity. This change in behaviour provides a first mechanism to control the potential threat, since it will entail the direct destruction of many pathogens by dendritic cells and macrophages. Simultaneously, these cells release alert signals that recruit other immune cells into the site of the infection. They also move peptides collected from the attacked tissue to specific membrane structures called Major Histocompatibility Complexes (MHC) and migrate to nearby lymph nodes.

In the lymph nodes, dendritic cells loaded with peptides (called antigens at this stage) interact with T lymphocytes (or T cells for short). T cells are endowed with a membrane receptor (the T Cell Receptor or RCR) that recognizes antigens bound to the MHC of dendritic cells. The spatial structure of the TCRs is highly variable, so that not all TCRs will show the same affinity for a given antigen. Furthermore, T cells are clonal, i.e., they are equipped with multiple copies of only one type of TCR. In the lymph node T cells are in a naïve state, meaning that they are inactivated and unfit for fight. If the TCR of a naïve T cell does not bind any of the antigens presented by a dendritic cell, the T cell remains inactive. However, if the TCR and an antigen fit smoothly (a fact known as immune synapse) a strong response to that particular antigen is readily mounted. In particular, as a result of immune synapse, the naïve T cell activates and becomes an effector T cell. The structural diversity of the TCR implies that only a small fraction of naïve T cells activate in response to the antigens presented by a dendritic cell. Activated T cells undergo a fast replication process that increases their number, which at the beginning may just consist in a few hundreds over the whole body, up to millions in few hours, a process termed as clonal expansion.

The resulting population of effector T cells is ready to fight any cell showing antigens as those already identified by its TCR, and in most cases the carrier pathogens are wiped out. The large population of T cells generated along the way is then mostly disposed of. In fact, the majority of them undergo apoptosis (cell suicide) whereas a small percentage is kept alive to provide a reservoir of memory cells. The latter are already instructed to react to the same type of pathogen in case it dares to show up again. See for instance [7] for details about the process just described.

While the scheme just sketched is widely accepted, a number of details concerning its precise implementation remain to be fully understood. For instance, a large percentage of the antigens displayed for identification by dendritic cells have been generated by the host. This occurs because dendritic cells phagocytize pathogens, but also the remains of host cells that have been destroyed in the course of the infection. For this reason, T cells showing affinity for self-antigens might activate in the course of an infection and trigger an immune response against host cells. In this case, friends are mistaken as foes, which results in autoimmune diseases. On the other hand, antigens coming from potentially dangerous pathogens might be mistaken as innocuous and left unchecked, with potentially devastating consequences. Consequently, an efficient and finely tuned T cell screening capacity is required to distinguish threatening from non-threatening antigens. Evidence shows that such screening capacity actually exists to a considerable degree, although

it may occasionally fail to work during our lifetime. This raises the natural question of understanding what are the precise biological mechanisms that control antigen identification. In spite of significant advances during last century [6] this remains a major issue in Immunology to this day.

We have just noticed that telling friends from foes is a major issue that our IS has to address, but this is by no means the only intriguing fact in immune response that calls for explanation. In some cases it may happen that a pathogenic threat is identified as such, but the resulting T cell action falls short of eliminating it. For instance it could occur that, after perhaps an initial remission of the challenging population, large numbers of T cells would no longer be deemed necessary, and clonal contraction be turned on to trim the T cell population down to a residual value, only to see pathogen figures eventually rebound. Such relapse may result in chronic disease or even in host collapse. As a matter of fact, immune evasion as that described above may be a consequence of insufficient T cell activation or result from a suitable hide-and-seek strategy played by the invaders, or both.

Even when things go as expected and pathogens are successfully eliminated, a number of acknowledged features of T cell operation appear at odds with our intuition. For instance, one may expect that in the case of acute infection clonal expansion should occur immediately after a pathogen is detected, and similarly clonal contraction would take place right after pathogen population had experienced a sharp decrease which forecasts its extinction. However, while the first statement is basically correct, the second one is not. In fact, clonal contraction is known to be a delayed effect, so that T cell population continues to expand well after pathogens have been practically eliminated. This situation is depicted in Fig. 1 below.

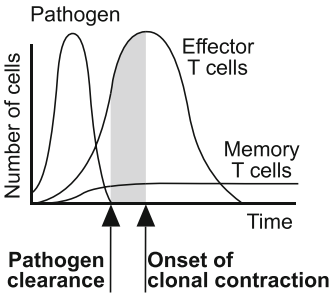


Fig. 1 Clonal contraction is a delayed effect. In case of successful response to acute infection, effector T cells continue to be generated well after the pathogen population has become virtually eliminated. Then the bulk of the T cell population is eventually disposed of by apoptosis. A comparatively small number of activated T cells are saved and become memory cells, ready to quickly recognize the same pathogen in case it appears again

2 T Cell Response as an Emergent Behaviour

From what has been said, it might appear as natural that a specialized organ in the body should be in charge of gathering intelligence about incoming threats (external or internal) and deciding when, where and how an immune response should be provided to meet them. It is quite remarkable that such organ has never been found to exist. Instead, T cell activation seems to be consequence of purely local decisions, taken by individual cells that, while able to scan their immediate neighbourhood for chemical cues, have only a dramatically limited repertoire of possible choices after processing such information. As far as we know, these actions are substantially limited to divide, die or differentiate into a restricted variety of related cell types. Extremely efficient as it turns out to be, immune response appears to be an emergent property that unfolds at the cell population level out of a very reduced set of possible individual cell actions.

These considerations immediately lead to further questions. For instance, we have seen that any activated T cell divides and the same will happen to their daughter cells during clonal expansion. For how long will T cells continue to divide? In other words, is the lineage of a single activated T cell a priori specified? In the affirmative, how is this done? Similarly, one may wonder when and why an activated T cell decides that is time to commit suicide. Is this decision a priori specified from the beginning of its lineage or is it made instead after an evaluation of the stage of the infection? In that case, and since any cell can only process information from its local neighbourhood, how comes that clonal contraction seems to be started at a comparatively late stage, when the pathogenic threat has been already overcome?

Providing a comprehensive review of the various models proposed to account for issues as those recalled above is well beyond the scope of this short note. We shall content ourselves herein with a rough description of recent approaches to deal with two of the questions that have been already mentioned, namely the delayed nature of clonal contraction and the discrimination between pathogenic and non pathogenic antigens [1, 2].

Concerning the first question, it has been already observed that activated T cells continue to divide even in the absence of pathogenic stimulation. A way to account for that fact has been forwarded in [1]. That work is mainly concerned with T cell immune response to an acute infection, and is therefore assumed that the corresponding antigen has been identified as threatening, and effector T cells have been generated to eliminate it. A simple, single cell based ODE model is derived in [1] that reproduces the delay observed in clonal contraction under the following assumptions.

1. Upon activation, any naïve T cell undergoes a first asymmetric division which results in the formation of a memory T cell and an effector T cell. At the modelling level, the difference between these two phenotypes lies mainly in the nature of the cross-talk among their membrane receptors, which will be explained in detail in point 3 below. Subsequent divisions are always symmetric, that is, in each case they lead to two daughter cells belonging to the same phenotype as their mothers.

2. During their lifetime, each activated T cell has to decide between two options: to divide or to die. This choice is not random. It depends instead on the relative proportion of two inhibitor proteins: Rb (which prevents the cell to enter its division cycle) and Bcl2 (which inhibits the internal apoptosis cell program) [3, 4]. The first of such proteins that falls below a critical level, that we may set equal to zero for convenience, determines the irreversible fate of the cell. For instance, depletion of Rb commits cell to enter a division process.
3. The dynamics of the amounts of Rb and Bcl2 is mediated by signals processed by membrane receptors, which will be referred to as death (D) and proliferation (P) receptors for convenience. Both types can interact among themselves, and do regulate the internal amount of Rb and Bcl2 by processing information provided by external chemical cues. These last are assumed to be basically at local equilibrium around each cell. Interestingly, D and P receptors exhibit different types of feedback: Proliferation (P) receptors are stimulated by pathogen antigen, and lose activity according to a first order Arrhenius kinetics. They also stimulate death (D) receptors, which are deactivated in a much longer time scale. Importantly, the kinetics of D receptors does not proceed in a symmetrical manner, since they do not activate P receptors.

Once a T cell has divided, the number of D and P receptors of the mother cell is equally distributed among the daughter cells. Memory cells are not equipped with death receptors. In other words, they are immortal in the time scale of the process under consideration. A striking result obtained in [1] is that such assumptions suffice to account for the delay in clonal contraction represented in Fig. 1. They also allow to explain the difference in length of cell cycles among dividing cells and moreover show that immortal cells can undergo a limited number of (symmetric) divisions before entering a quiescent stage. The reader is referred to [1] for a detailed description of these issues.

3 The Growth Threshold Conjecture

We conclude this paper with a short description of a result obtained in [2] concerning the mechanism that allows T cells to safely distinguish, at least most of the times, between friendly and unfriendly antigens. A way to ascertain how this is achieved goes under the name of growth threshold conjecture (GTC) and can be described as follows. Consider a dynamical interaction between a population of T cells ($T(t)$) and that of a pathogen ($P(t)$) given by the following rules:

$$\begin{cases} T''(t) = -kT(t) + \lambda P(t) \\ P'(t) = \alpha P(t) - \beta T(t)P(t) \\ T(0) = 0 \\ T'(0) = 0 \\ P(0) = P_0 \geq P_m \end{cases} \quad \text{, for } T \geq 0, P \geq P_m \tag{1}$$

Simple as it is, the ODE system above encodes some assumptions worth to be stressed. To begin with, the first differential equation postulates that, in the absence of pathogen stimulation, a T cell population would be characterized by an intrinsic elastic response that allow for oscillatory solutions. The presence of pathogens would have the effect of an external forcing term. On its turn, the second equation there states that in the absence of T cells, pathogen population would grow exponentially. However, as soon as T cells enter the game, they start to eliminate pathogens according a second-order kinetics, which accounts for possible encounters between $T(t)$ and $P(t)$, as represented by the last term in the right of that equation. The relative efficiency of each process involved is measured by the four parameters appearing in these two equations, which are completed by suitable initial conditions representing pathogen invasion and absence of activated T cells when time starts to run. As a matter of fact, the previous system can be rewritten in rescaled variables that allow to reduce the number of parameters involved to two instead of four. In this manner one arrives at:

$$\begin{cases} T''(t) = -T(t) + P(t) \\ P'(t) = \alpha^* P(t) - \beta^* T(t) P(t) \\ T(0) = 0 \\ T'(0) = 0 \\ P(0) = 1 \end{cases} \quad \text{, for } T \geq 0, P \geq P_m^*, \tag{2}$$

where $\alpha^* = \dfrac{\alpha}{\sqrt{k}}$, $\beta^* = \dfrac{\beta\lambda P_0}{k\sqrt{k}}$ and $P_m^* = \dfrac{P_m}{P_0}$.

The system thus obtained does not seem to allow for explicit integration. However numerical simulations reveal a crucial dependence of its dynamics in terms of the two rescaled parameters therein retained. This is described in Fig. 2 below.

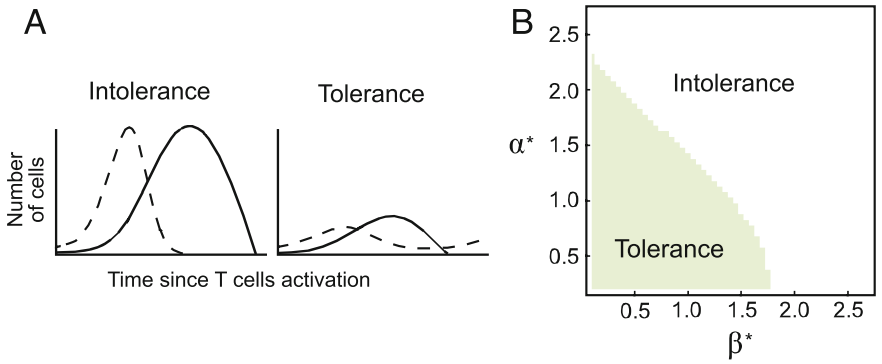


Fig. 2 Immune response is said to result in tolerance (**a**, *right*) when it does not suffice to wipe out a designated pathogen and is instead said to produce intolerance when it does succeed (**a**, *left*). In terms of our rescaled system, the parameter space can be divided in two regions where tolerance (respectively intolerance) occurs separated by a boundary curve (**b**)

It follows from Fig. 2 that for a given value of any of the two parameters involved, one may leave a region where tolerance (or intolerance) prevails to enter its complementary region by just modulating the value of the remaining parameter. In fact, a careful analysis of that figure reveals that, according to our model, pathogens and tumours can escape the action of T cells by either reducing T cells clearance rates (for instance, by immunosuppression or by hiding from the IS) or by reducing their proliferation rates; showing a low profile would thus pay off for pathogens. A further interesting consequence of the proposed model is that actions as inflammation, which increase clearance rates, or fever, which increases proliferation rates in pathogens, are quite effective to reduce intolerance, thus contributing to eventual pathogen suppression.

4 Concluding Remarks

In this note we have shortly reviewed some aspects of immune response where mathematical modelling can help to understand the nuts and bolts of biological mechanisms at work. In particular, models as those briefly recalled allow to select the minimum of acknowledged biological facts required to efficiently perform the tasks described. Important as this parsimony principle may be, a number of significant questions remain to be ascertained as yet. For instance, it would be interesting to know if only the facts selected in each situation can give raise to the observed behaviours. Moreover, when several alternatives are possible, one may wonder if their redundancy could represent an evolutionary advantage rather than a hindrance. We intend to address these issues in future work.

Acknowledgements This work has been partially supported by MINECO Grant MTM2014-53156-P.

References

1. Arias, C.F., Herrero, M.A., Acosta, F.J., Fernandez-Arias, C.: A mathematical model for a T cell fate decision algorithm during immune response. J. Theor. Biol. **349**, 109–120 (2014)
2. Arias, C.F., Herrero, M.A., Cuesta, J.A., Acosta, F.J., Fernandez-Arias, C.: The growth threshold conjecture: a theoretical framework for understanding T-cell tolerance. R. Soc. Open Sci. **2**(7), 150016 (2015)
3. Czabotar, P.E., Lessene, G., Strasser, A., Adams, J.M.: Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. Nat. Rev. Mol. Cell Biol. **15**(1), 49–63 (2014)
4. Giacinti, C., Giordano, A.: RB and cell cycle progression. Oncogene **25**(38), 5220–5227 (2006)
5. Janeway Jr., C.A., Travers, P., Walport, M.: Immunobiology: The Immune System in Health and Disease. Garland Science, New York (2014)
6. Matzinger, P.: Tolerance, danger, and the extended family. Annu. Rev. Immunol. **12**, 991–1045 (1994). doi:10.1146/annurev.iy.12.040194.005015
7. Sompayrac, L.M.: How the Immune System Works. John Wiley & Sons, Hoboken, NJ (2015)

Ray Mappings and the Weighted Least Action Principle

Jacob Rubinstein, Yifat Weinberg, and Gershon Wolansky

Abstract Two basic problems in optics are presented. The solutions to both problems are formulated in terms of the associated ray mappings. An alternative formulation based on a weighted sum of the actions along the rays is derived. Existence of solutions is established via the Weighted Least Action Principle. Numerical methods for computing the ray mappings are discussed. Finally, we demonstrate the theoretical considerations by presenting a complete solution to a specific beam shaping lens design.

1 Introduction

Ray mappings are fundamental objects in geometrical optics. Given an optical element, its Hamilton's point eikonal function includes the information on *all* possible ray mappings induced by it. However, in many applications one looks for specific ray mappings that satisfy certain constraints. In this paper we shall consider two such cases. In the first case we search for all ray mappings associated with a single monochromatic beam whose intensity is known at two planes. This question, known as the phase retrieval problem arises in many applications ranging from astronomy [9, 14] to ophthalmology [8]. The second case concerns the shaping of a collimated beam with arbitrary incident and refracted intensity distributions.

While these two cases involve two completely different problems in optics, we shall show that their *mathematical* formulation is essentially identical. We shall then introduce two characterizations of these problems, provide practical methods to compute the ray mappings, and from the mappings obtain the unknown phase in case one, and the beam shaping lens in case two.

In the next section we derive the ray mapping equations for the phase retrieval problem, while in Sect. 3 we derive the equations for the beam shaping lens. In Sect. 4 we show that both ray mapping problems can be solved by the Weighted Least Action Principle (WLAP) which is a natural extension of the Fermat principle

J. Rubinstein (✉) • Y. Weinberg • G. Wolansky
Department of Mathematics, Technion, Haifa, Israel
e-mail: koby@tx.technion.ac.il; yifatpw@gmail.com; gershonw@math.technion.ac.il

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_3

of least time [10, 11]. Numerical methods for computing the ray mappings, or equivalently finding the critical points of the Weighted Least Action functional, along with an example of lens design are provided in Sect. 5.

2 Phase Retrieval

We denote $u(x, y, z)$ a monochromatic wave propagating in the positive z direction. We express u in the form

$$u(x, y, z) = A(x, y, z)e^{ik\phi(x,y,z)}. \tag{1}$$

The goal is to find the phase (or eikonal) ϕ from measurements of its intensities $I = A^2$ at two planes, say $z = 0$ and $z = h$:

$$I_1(x, y) := I(x, y, 0), \quad I_2(x, y) := I(x, y, h). \tag{2}$$

The wave u is assumed to obey the Helmholtz equation $\Delta u + k^2 u = 0$, where k is the wave number, and we assume that the refraction index is $n = 1$.

In the geometrical optics limit the phase is equivalent to the rays, which are the normals to the wavefronts. A point (x, y) at the aperture Σ on the plane $z = 0$ is mapped into a point $T(x, y) = (x_p, y_p)$ at the imaged aperture Σ_p on the second plane $z = h$. To find the relation between the ray mapping and the phase $\phi(x, y, 0)$ we use the Rayleigh-Sommerfeld diffraction integral:

$$u(x_p, y_p, h) = \frac{-ik}{2\pi} \int_{\Sigma} A(x, y, 0) \frac{\exp(ik(\phi(x, y, 0) + d(x_p, y_p, x, y))}{d} \cos\theta \, dxdy. \tag{3}$$

Here $d = \left((x_p - x)^2 + (y_p - y)^2 + h^2\right)^{1/2}$, and θ is the angle made by the line connecting (x, y) and (x_p, y_p) with the plane $z = 0$. The integral is approximated in the large k limit by the stationary phase method. The stationarity condition is $\nabla(\phi(x, y, 0) + d(x_p, y_p, x, y)) = 0$, or equivalently:

$$\nabla\phi = -\nabla_2 d = \frac{(x_p - x, y_p - y)}{\left((x_p - x)^2 + (y_p - y)^2 + h^2\right)^{1/2}}, \tag{4}$$

where $\nabla_2 d(x_p, y_p, x, y)$ denotes the gradient of d with respect to (x, y). We assume that Eq. (4) implies a one to one correspondence between Σ and Σ_p, which is equivalent to assuming that no caustic surface intersects either planes $z = 0, h$.

Using Eq. (4), we can write down the classical stationary phase approximation of $u(x_p, y_p)$ and then use it to obtain the following relation between the intensities at

the two planes:

$$I_2(T(x,y))|J(T(x,y)| = I_1(x,y). \tag{5}$$

Here J is the Jacobian of the ray mapping $(x_p, y_p) = T(x,y)$, and $|J|$ denotes the absolute value of the determinant of J. Therefore, the phase retrieval problem is equivalent to finding a ray mapping $T : (x,y) \to (x_p, y_p)$ that satisfies both Eqs. (4) and (5).

3 Beam Shaping

Consider an incident collimated beam, propagating in the z direction, whose cross sectional intensity is $I_1(x,y)$. The goal is to design a lens with two freeform surfaces that converts it into a new collimated beam with cross sectional intensity $I_2(x,y)$. Such problems are of great importance in the laser industry [4]. The case where I_1 and I_2 are both radially symmetric is well-known, and can be reduced to simple ODEs. However, we consider the most general case of arbitrary distributions I_1 and I_2. The schematic beams and optical element are depicted in Fig. 1. In Fig. 2 we present an example, where the target is to design a lens that converts an incoming collimated beam with six spots (for instance a beam created by a LED-based device) into a homogeneous collimated beam.

Let $f(x,y)$ and $g(x,y)$ be the front and back surfaces of the lens, respectively. The refraction index of air is 1, and we denote the lens refraction index n. Consider a ray that starts from a point (x,y) on a surface $z = 0$ on the left side of the lens. After two refractions, the ray is mapped into a point $(x_p, y_p) = T(x,y)$ on a plane $z = h$ on the right side of the lens. The ray mapping T is constrained by the energy conservation equation (5). In addition, the mapping T and the lens surfaces f and g are related by Snell's law of refraction and by the fact that both planes $z = 0, h$ are wavefronts, since the incident and refracted beams are collimated. This implies that the optical

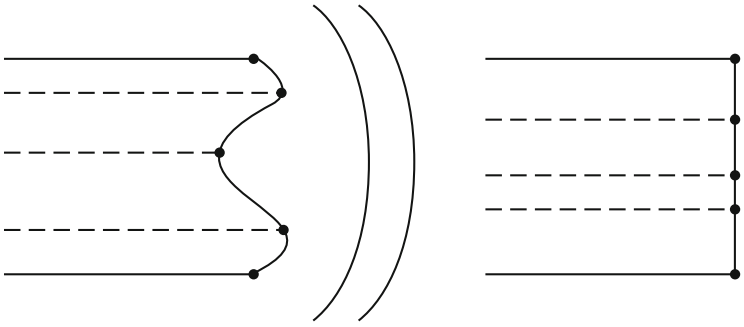


Fig. 1 Shaping arbitrary collimated beams

Fig. 2 An incoming beam consisting of six spots arranged over a circle

path between all points (x, y) and their images $T(x, y)$ is constant. Combining these facts, and after some algebra, the following equation is obtained [12] that relates f and T.

$$\nabla f = \frac{-n((x_p, y_p) - (x, y))}{\sqrt{\chi - bn|(x_p, y_p) - (x, y)|^2}}, \tag{6}$$

where χ and b are two constants that depend upon n. A similar equation holds for the second surface g.

4 The Weighted Least Action Principle

We presented in Sects. 2 and 3 two different optical problems that are fully determined by specific ray mappings. Surprisingly, both problems led to a similar mathematical formulation. Both ray mappings are required to satisfy the energy conversation condition (5). In addition, in both cases the ray mapping T must satisfy a 'symmetry' condition such as (4) or (6). Three questions arise. The first is whether such ray mappings T exist at all. If the answer to this question is positive, the second question is how many solutions exist, and what are their properties. The third question is how to *compute* the ray mappings.

It is very hard to answer these questions directly from Eqs. (4)–(5). Together they constitute a very complicated PDE of the notorious Monge-Ampere type. Instead, we shall answer all three questions using an equivalent formulation of the problem.

For this purpose, we associate each ray connecting (x, y) and $(x_p, y_p) = T(x, y)$ with an *action* $C(T, (x, y))$, and then define the *weighted* total action to be

$$M(T) = \int_{\Sigma} C(T, (x, y)) I_1(x, y)\, dxdy, \tag{7}$$

where Σ denotes the beam's cross sectional domain. The weighted action M is restricted to the set $\mathscr{T}(I_1, I_2)$ of mappings that satisfy condition (5).

Theorem 1 *A ray mapping T is a critical point of M in the class $\mathscr{T}(I_1, I_2)$ if and only if T satisfies the following relation*

$$\nabla_2 C(x_p, y_p, x, y) = \nabla \zeta, \tag{8}$$

for some 'potential' function $\zeta(x, y)$. The notation $\nabla_2 C$ denotes the gradient of C with respect to (x, y).

It is now possible to identify both Eqs. (4) and (6) to be of the form of Eq. (8). Therefore the phase retrieval problem can be solved by associating the unknown phase $\phi(x, y, 0)$ with the critical points of M for the choice $C(x_p, y_p, x, y) = d(x_p, y_p, x, y)$. Similarly the beam shaping lens can be solved by identifying the front lens surface f with the critical points of M for the choice $C(x_p, y_p, x, y) = -\sqrt{\chi - bn|(x_p, y_p) - (x, y)|^2}$.

We proceed to answer the questions we posed at the head of this section. Existence of minimizers of M, at least for convex actions C was proved by Brenier [3]. Existence of maximizers to these functionals is proved in [13]. Furthermore, we conjecture that there always exist also critical points that are neither minimizers nor maximizers. It remains to consider the computation of the critical points.

5 Implementation

The problem of computing minimizers of M received quite a bit of attention. For example we refer to [1, 2, 6] where PDE-based approach is suggested. An alternative combinatorial approach is to sample the two intensities I_1, I_2 faithfully by two sets of points (x^j, y^j) for I_1 and (x_p^j, y_p^j) for I_2, where $j = 1, 2, \ldots, N$. The problem of minimizing M is the same as computing the permutation π that minimizes

$$\sum_{j=1}^{N} C(x_p^{\pi(j)}, y_p^{\pi(j)}; x^j, y^j).$$

The same method can be used to find the maximizers of $M(T)$. While searching over all permutations has of course huge complexity, much more efficient algorithms,

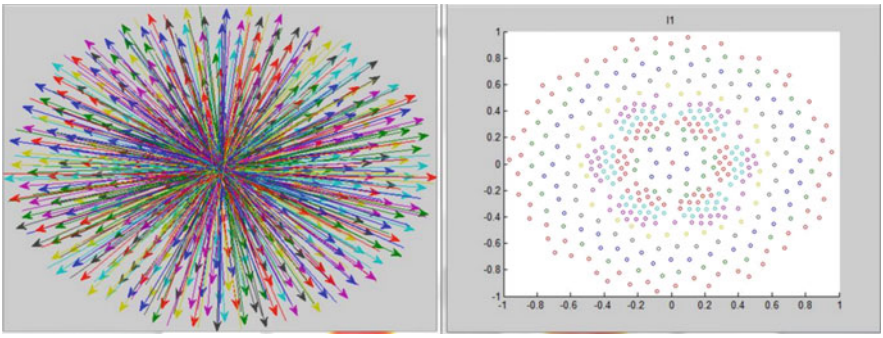


Fig. 3 A discrete sampling of the intensity of the incoming beam (*right*) and the ray mapping (*left*)

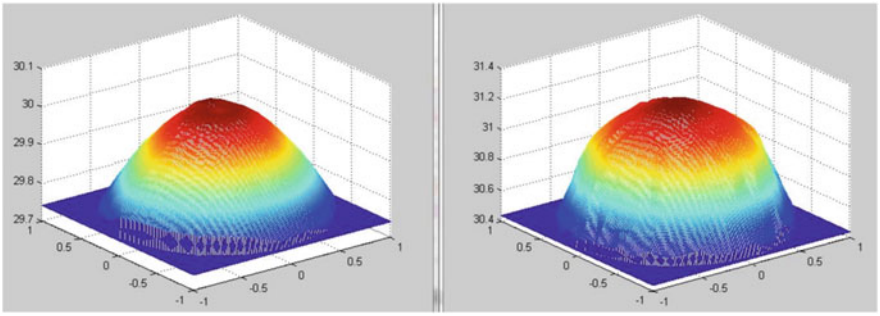


Fig. 4 The surfaces (f on the *left* and g on the *right*) of the beam shaping lens

such as the "Hungarian algorithm" [5] were proposed with complexity of $O(N^3)$. An even more effective multiscale algorithm was proposed recently by Merigot [7], although it is limited to quadratic cost functions. We recently developed a very fast $O(N)$ multiscale algorithm that works for all convex C, at the price of some moderate assumptions on the potential function ζ.

Without spelling out the algorithm's details, we proceed to solve the beam shaping example of Sect. 3. The intensities were sampled with 400 points (right drawing of Fig. 3). In this case we computed the *maximizer* of M since this solution guarantees that both surfaces f and g are convex, which is a very important consideration in the lens manufacturing. The ray mapping is depicted in the left drawing of Fig. 3. Finally, we depict in Fig. 4 the lens surfaces.

Acknowledgements This work is supported by grants from the Israel Science Foundation.

References

1. Angenent, S., Haker, S., Tannenbaum, A.: Minimizing flows for the Monge-Kantorovich problem. SIAM J. Math. Anal. **35**, 61–97 (2003)
2. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math. **84**, 375–393 (2000)
3. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. **64**, 375–417 (1991)
4. Dickey, F.M.: Laser beam shaping. Opt. Photonics News **14**, 31–35 (2003)
5. Kuhn, H.W.: The Hungarian Method for the assignment problem. Naval Res. Logist. Q. **2**, 83–97 (1955)
6. Lee, C., Lowengrub, J., Rubinstein, J., Zheng, X.: Phase reconstruction by the weighted least action principle. J. Opt. A Pure Appl. Opt. **8**, 279–289 (2006)
7. Merigot, Q.: Multiscale approach to optimal transport. Comput. Graph. Forum **30**, 1584–1592 (2011)
8. Nam, J., Rubinstein, J., Thibos, L.: Wavelength adjustment using an eye model from aberrometry data. J. Opt. Soc. Am. A **27**, 1561–1574 (2010)
9. Roddier, F.: Curvature sensing and compensation: a new concept in adaptive optics. Appl. Opt. **27**, 1223–1225 (1998)
10. Rubinstein, J., Wolansky, G.: A variational principle in optics. J. Opt. Soc. Am. A **21**, 2164–2172 (2004)
11. Rubinstein, J., Wolansky, G.: A weighted least action principle for dispersive waves. Ann. Phys. **316**, 271–284 (2005)
12. Rubinstein, J., Wolansky, G.: Intensity control with a free-form lens. J. Opt. Soc. Am. A **24**, 463–469 (2007)
13. Rubinstein, J., Wolansky, G.: Geometrical optics and optimal transport. J. Opt. Soc. Am. A (to appear)
14. Teague, M.R.: Deterministic phase retrieval: a Green's function solution. J. Opt. Soc. Am. **73**, 1434–1441 (1983)

Part II
Hansjörg Wacker Memorial Prize

Numerical Methods for Optimization Problems Arising in Energetic Districts

Elisa Riccietti, Stefania Bellavia, and Stefano Sello

Abstract This paper deals with the optimization of energy resources management of industrial districts, with the aim of minimizing the customer energy bill. Taking into account real time information on energy needs and production and on energy market prices, a cost function is built that should be minimized. Here we focus on the solution of the arising nonlinear constrained optimization problem. We describe the two solvers that have been employed for its solution: a Sequential Linear Programming and a Particle Swarm Optimization.

1 Introduction

This paper is realized in collaboration with the research centre 'Enel Ingegneria e Innovazione - Ricerca' in Pisa, Italy and it is related to an Enel project devoted to the realization of a software tool for the numerical solution of optimization problems arising in energetic districts, [4]. An energetic district is a complex that comprises machines able to generate, absorb or accumulate energy and that is interlaced with the electrical grid with which it can exchange energy. Different machines can be part of a district: generators (electrical, photovoltaic), accumulators, electrical or thermal loads.

Enel research centre aim is to develop a software tool able to find the best asset of the machines in a district, i.e. the one that guarantees the lowest expenses for the district management. The software tool builds the district model, an objective

E. Riccietti (✉)
Dipartimento di Matematica e Informatica "Ulisse Dini", Università di Firenze, viale G.B. Morgagni 67a, 50134 Firenze, Italy
e-mail: elisa.riccietti@unifi.it

S. Bellavia
Dipartimento di Ingegneria Industriale, Università di Firenze, Via di S. Marta, 3, 50139 Firenze, Italy
e-mail: stefania.bellavia@unifi.it

S. Sello
Enel Ingegneria e Innovazione - Ricerca, Pisa, Italy
e-mail: stefano.sello@enel.com

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_4

function that measures the costs related to the district management, and functions modelling the physical constraints on the machines. A nonlinear minimization problem with nonlinear constraints arises from this modelling. In this paper we focus on the solution of this optimization problem.

In finding the optimal solution a compromise between two needs should be found. On one hand one aims at finding the machine asset corresponding to the lowest possible value of the objective function. On the other hand the optimization process should be as quick as possible, as it needs to be performed several times in a day, when the needs of the district or the weather conditions affecting the renewable resources change.

For this reason two different solvers were employed and compared: a Sequential Linear Programming (SLP) and a Particle Swarm Optimization (PSO).

SLP is an iterative solver that generates a sequence of linear approximations of the original nonlinear problem. It is supported by theoretical results and it is ensured to converge to a local minimum. With this method it is not guaranteed to find the global minimum, actually the generated sequence usually converges to the local minimum nearer to the starting guess.

PSO is an heuristic research algorithm designed to find a global minimum of the problem. It is a derivative free algorithm that requires just function values. For this reason it is characterized by a slower convergence than SLP that is a first order method.

PSO solver is then expected to find a better solution than SLP, but the optimization process is expected to be longer. To investigate on this, the solvers were tested on many different realistic examples of energetic districts, and on a real energetic district provided by Enel.

2 Optimization Problem

The variables that need to be optimized are the physical parameters of the machine that affect their functioning (such as the electrical power produced by generators, that stored by batteries, the thermal power provided by boilers). The aim of the optimization process is to build a plan of the machines parameters for the following day to minimize the expenses related to the district management. The objective function f represents the overall daily cost of energy obtained as a result of the difference between purchase costs (fuel and electric energy) and incomings (incentives and sales revenues).

If n is the problem dimension, $f : \mathbb{R}^n \to \mathbb{R}$ is a nonlinear function. The optimization problems is a constrained problem, with both nonlinear constraints and bound constraints. The first ones represent physical constraints on the machines and are modelled by a nonlinear function $g : \mathbb{R}^n \to \mathbb{R}^p$. Bound constraints arise from the fact that the variables are normalized in the interval $[0, 1]$ or $[-1, 1]$.

The resulting optimization problem is the following nonlinear constrained problem:

$$\min_x f(x) \tag{1}$$

$$x_{\min} \leq x \leq x_{\max}, \tag{2}$$

$$g(x) \leq 0, \tag{3}$$

where $x \in \mathbb{R}^n$ is the parameters vector, $x_{\min} \in \mathbb{R}^n$ and $x_{\max} \in \mathbb{R}^n$ denote the related bound constraints.

3 Sequential Linear Programming (SLP)

Sequential Linear Programming (SLP) is an iterative method to find local minima of nonlinear constrained optimization problems [1], by solving a sequence of linear programming problems. In our approach first a penalty function is built treating the constraints as a penalty term, i.e. a new term is added to function f that is positive if the constraints are not satisfied, and is zero otherwise. The resulting function is used as the new objective function. Following [1, 5] the l_1 penalty function was chosen:

$$\Phi(x, \nu) = f(x) + \nu \sum_{i=1}^{p} \max(0, -g_i(x)),$$

where $\nu > 0$ is the penalty parameter. Then, at each iteration k, linear models of function f and of the constraints are computed, approximating them in a neighbourhood of the current solution approximation x_k with first order Taylor series:

$$m_k(d) = f(x_k) + \nabla f(x_k)^T d, \qquad g_i(x_k) + \nabla g_i(x_k)^T d \qquad i = 1, \ldots, p,$$

where $d = x - x_k$ is the step. To obtain a globally convergent method, SLP approach is usually coupled with a trust-region strategy, [2]. Then, a new constraint is added to the bound constraints, that consists of a bound on the step-length:

$$\|d\|_\infty \leq \Delta_k,$$

where Δ_k is called the trust-region radius. The linearized functions replace the nonlinear ones in the penalty function and the following linear minimization

problem is solved:

$$\min_d f(x_k) + \nabla f(x_k)^T d + \nu \sum_{i=1}^{p} \max(0, -g_i(x_k) - \nabla g_i(x_k)^T d) \tag{4}$$

$$\max((x_{\min} - x_k)_j, -\Delta_k) \leq d_j \leq \min((x_{\max} - x_k)_j, \Delta_k), \qquad j = 1, \ldots, n. \tag{5}$$

At each iteration the computed solution d_k is used as a step to define the new solution approximation: $x_{k+1} = x_k + d_k$ and the trust-region radius is updated. The step is accepted if a sufficient decrease in the objective function f is obtained, and in this case the trust-region radius is possibly enlarged, otherwise the step is rejected, the trust-region radius is reduced and the subproblem is solved again.

Following [5], for this approach we proved the following important theoretical results:

1. global convergence of sequence $\{x_k\}$ generated by the algorithm to a KKT point of $\Phi(x, \nu)$ (if ν is big enough a minimizer of Φ is a solution of (1)–(3)),
2. convergence of the Lagrange multipliers estimates of problem (4)–(5) to the multipliers of problem (1)–(3).

This second theoretical results allows us to implement a reliable stopping criterion, as it provides a measure of the distance of the solution approximation found to a KKT point of the original system.

4 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is a stochastic evolutionary method designed to converge to global minimum, [3]. It is inspired to the behaviour of bird swarms. Following the natural metaphor, PSO evolves a population of individuals, referred to as particles, within the search space, that behave according to simple rules but interact to produce a collective behaviour to pursuit a common aim, in this case the localization of a global minimum. The swarm is composed of s particles, each of them represents a solution of the optimization problem and it is represented by a vector $x \in \mathbb{R}^n$. To each particle it is associated a velocity vector v too. The method is based on an iterative algorithm and at each step k vectors x and v are updated:

$$v_{k+1}^i = w v_k^i + c_1 r_1 (p_{best,k}^i - x_k^i) + c_2 r_2 (p_{best,k}^g - x_k^i),$$
$$x_{k+1}^i = x_k^i + v_{k+1}^i,$$

where x_k^i and v_k^i are the position and velocity vector of the i-th particle, $i = 1, \ldots, s$, at the k-th iteration, $p_{best,k}^i$ and $p_{best,k}^g$ are respectively the best position reached by the i-th particle so far and the best position reached by the whole swarm (the best position is the one that corresponds to the lowest value of the objective function),

c_1, c_2 and w are positive weights, r_1 and r_2 are random variables with uniform distribution in [0, 1].

The process is stopped as soon as a maximum number of iterations is reached or when the objective function is not sufficiently decreased within a fixed number of iterations. The computed approximate solution is $p^g_{best,k*}$ where k^* is the last iteration index.

Bound constraints are handled bringing back on the nearest boundary a particle x that has left the search space and changing the sign of the particle velocity, to avoid a new violation on the next iteration.

Originally PSO methods were developed to deal with problems with just bound constraints, and later they were employed to solve also constrained problems [6]. Usually a penalty function approach is used. We employed a quadratic penalty function:

$$\Phi(x, \tau_k) = f(x) + \frac{1}{2\tau_k} \sum_{i=1}^{p} g_i(x)^2,$$

with τ_k penalty parameter that is decreased at each iteration to penalize with increasing severity constraints violations.

The main advantage of PSO methods is that they do not require neither regularity assumptions on the objective function nor to compute the first derivatives and are so suitable when few information on the objective function are available. Clearly the fact that few information on f are used, leads to a really slow method that requires many iterations to converge. Furthermore these methods are heuristic and there are not theoretical results that ensure convergence of the method.

5 Numerical Tests

The software tool equipped with the two solvers was tested on 13 different models of synthetic energetic districts and on an existing district in Pisa, provided by Enel.

The numerical experimentation has been carried out on a PC equipped with a AMD Phenom(tm)II X4 965 Processor 3.40 GHz, 8.00 GB RAM, Windows 7 Professional 64 bit and using Matlab R2012a.

Results shown in the following tables are the average of those obtained over 10 runs, varying the starting point for SLP solver. In the headings of the following tables **f** and **k** are the arithmetic mean of the function values and the number of iterations, $\boldsymbol{\sigma_f}$, $\boldsymbol{\sigma_k}$ are the standard deviations, min **f**, max **f** are the minimum and maximum function values obtained, **time/it**(s) is the time for an iteration in seconds, **time**(m) is the total time in minutes.

As an example we report in Table 1 the detailed results of the optimization process provided by the two algorithms in one of the 13 tests. The model of energetic district includes an accumulator, a boiler, a CHP, a wind generator, a photovoltaic

Table 1 Comparison of solvers on an example of energetic district

Solver	**f**	σ_f	max **f**	min **f**	**k**	σ_k	**time/it(s)**	**time(m)**
PSO	27.2	0.2	27.6	26.9	1004	881	0.5	9
SLP	26.9	0.3	27.5	26.6	51	12	1.4	1.2

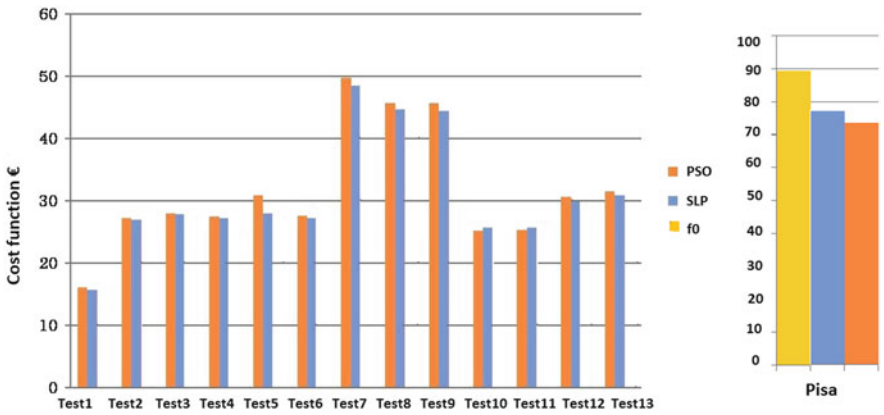


Fig. 1 Comparison of objective function values obtained by SLP and PSO on the 13 tests cases (*left*), comparison of actual (f_0) and optimized management provided by the two solvers, on the test case of Pisa district (*right*)

generator, 5 loads. The optimization problem has 398 variables subject to 213 process constraints and 796 bound constraints.

From this test it is possible to deduce the following remarks.

The convergence rate of SLP algorithm is really higher than that of PSO, as it employs first order informations on f. PSO method requires a very high number of iterations to find the minimum and so the computational time is higher. As expected SLP is more suitable for real time optimization.

The capability of PSO method to find a global minimum does not clearly emerge in these tests, actually in many tests the solution found by SLP is better, as it is highlighted in Fig. 1 (left), in which we report for all the tests the comparison of the objective function values obtained by the two solvers at the end of the optimization process. On the other hand, in unconstrained tests on strongly nonlinear problems taken from the literature, PSO performance results to be much higher than SLP one, as it manages to find the global minimum, while the sequence generated by SLP approaches the local minimum nearer to the starting guess. We can notice that the energetic district test cases are characterized by a high number of nonlinear constraints. On one hand we can think that the PSO constraints handling strategy could be improved, on the other hand the standard deviation on function values is really low, so it is possible that the problem does not have many local minima, and the use of PSO algorithm is not worth it.

Table 2 Results of the optimization of the energetic district of Pisa, $\mathbf{f_0}$ is the cost function for the actual management (without optimization by our procedures)

Solver	$\mathbf{f_0}$	**f**	σ_f	max **f**	min **f**	**k**	σ_k	**time/it(s)**	**time(m)**
PSO	89.3	73.5	0.08	73.6	73.4	377	34	0.4	2.7
SLP	89.3	74.9	2.8	80.6	73.4	23	9	2.6	1

On the other hand in the test performed on the real example of energetic district, that is much less constrained, PSO gains better performance than SLP.

This is a real district in Pisa that comprises: a CHP characterized by rated power 25 kWe and rated thermal power 75 kWt, a gas boiler with rated thermal power 35 kWt, a tank for the storage of hot water with capacity 9400 kJ/ °C, a photovoltaic generator with rated power 14 kWe, a wind farm with rated power 3 kWe, 2 loads. The arising optimization problem has 288 variables, 576 bound constraints and 1 physical constraint.

For this district we have at our disposal data referring to the actual management of local resources, so that we can evaluate savings arising from the optimized management provided by our procedures, as it is depicted in Fig. 1 (right) . In fact in Table 2, that shows the results of the optimization of Pisa district provided by the two methods, we report also the value of the unoptimized objective function f_0, that is computed using those data and represents the cost of the actual management.

In this case PSO algorithm performs better than SLP algorithm, providing a lower value of the objective function and also a really smaller standard deviation. In this case there is just one mild nonlinear constraint so the constraint handling strategy does not have a strong impact on the search process and the capability of the evolutionary algorithm to search for the global minimum appears as in an unconstrained test case. We can also notice that in this case the execution time is reasonable also for PSO algorithm.

Comparing the unoptimized management of local resources to the optimized one, we can see that the employment of the software tool guarantees considerable savings in the energy bill, about 18% daily saving.

Acknowledgements Work partially supported by INdAM-GNCS.

References

1. Byrd, R.H., Gould, N.I.M., Nocedal, J., Waltz, R.A.: An algorithm for nonlinear optimization using linear programming and equality constrained subproblems. Math. Program. **100**, 27–48 (2003)
2. Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust-Region Methods., SMPS/SIAM Series on Optimization, SIAM, Philadelphia (2000)
3. Eberhart, R.C., Kennedy, J.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 1942–1948 (1995)

4. Ferrara, W., Riccardi, J., Sello, S.: Enel customer optimization service expert system development, Technical Report, Enel Ingegneria e Ricerca, Pisa (2014)
5. Fletcher, R., Sainz de la Maza, E.: Nonlinear programming and nonsmooth optimization by successive linear programming. Math. Program. **43**, 235–256 (1989)
6. Parsopoulos K.E., Vrahatis, M.N.: Particle swarm optimization method for constrained optimization problems. In: Intelligent Technologies - Theory and Applications: New Trends in Intelligent Technologies, pp. 214–220. IOS Press, Amsterdam (2002)

Part III
Minisymposia

Minisymposium: Advanced Numerical Methods for Hyperbolic Problems

Giovanni Russo and Sebastiano Boscarino

Description

The purpose of the minisymposium was to gather researchers interested in industrial application of numerical methods for hyperbolic problems. Hyperbolic systems are involved in the description of many physical problems, ranging from classical applications in aerodynamics to shallow water models for the simulation of waves in lakes and rivers, from combustion problems to hydrodynamical models of semiconductors, to mention just a few examples. Because of the actuality and impact of the subject, a recent European Marie Curie ITN project, called ModCompShock (Modelization and computation of shocks and interfaces) has been approved and has just started. The project involves eight main European Universities and several partner research centers. Several participants to the project were invited to contribute to the minisymposium, which became a unique opportunity of scientific exchange between this community and ECMI.

The talks included in the minisymposium were the following:

- P. Mulet. Departament de matemàtica aplicada, Universitat de València (Spain). *Highly accurate weighted extrapolation for finite difference schemes on Cartesian meshes for curved domains.*
- M. Dumbser. University of Trento (Italy). *High order ADER schemes for a unified first order hyperbolic formulation of continuum mechanics.*
- S. Imperiale. Inria Saclay (France). *Space/time domain decomposition methods for wave propagation simulation.*
- C. Parés. Laboratorio de Métodos Numéricos, Universidad de Málaga (Spain). *Well-balanced methods for the Shallow Water equations in spherical coordinates.*

G. Russo (✉) • S. Boscarino
Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: russo@dmi.unict.it; boscarino@dmi.unict.it

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_5

- S. Boscarino. University of Catania (Italy). *Semi-implicit method for all mach number flow for the Euler equations of gas dynamics on staggered grid.*
- J. Schuetz. Vakgroep wiskunde en statistiek, UHasselt (Belgium). *A novel IMEX splitting for the isentropic Navier-Stokes equations.*

High Order Extrapolation Techniques for WENO Finite-Difference Schemes Applied to NACA Airfoil Profiles

Antonio Baeza, Pep Mulet, and David Zorío

Abstract Finite-difference WENO schemes are capable of approximating accurately and efficiently weak solutions of hyperbolic conservation laws. In this context high order numerical boundary conditions have been proven to increase significantly the resolution of the numerical solutions. In this paper a finite-difference WENO scheme is combined with a high order boundary extrapolation technique at ghost cells to solve problems involving NACA airfoil profiles. The results obtained are comparable with those obtained through other techniques involving unstructured meshes.

1 Introduction

Cauchy problems for hyperbolic conservation laws usually involve weak solutions, which yields difficulties to tackle them numerically and has motivated the development of high resolution shock capturing schemes (onwards HRSC). In this paper we will focus on HRSC finite-difference schemes, that use a discretization with a Cartesian mesh, to numerically solve problems involving NACA airfoil profiles, which are of a great interest in the field of aeronautical engineering.

Due to the nature of a Cartesian mesh, if enough care is not taken at the boundary neighborhood the accuracy order will be lost. Therefore, a strategy to extrapolate the information from the computational domain plus the boundary conditions (if any) at the ghost cells should be developed, taking into account as well that singularities may be eventually positioned near the extrapolation zone in order to avoid an oscillatory behaviour of the scheme.

A. Baeza • P. Mulet (✉) • D. Zorío
Departament de Matemàtiques, Universitat de València, C/Dr. Moliner, 50, Burjassot, 46100 València, Spain
e-mail: antonio.baeza@uv.es; mulet@uv.es; david.zorio@uv.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_6

Some authors have tackled this issue with different techniques, such as second order Lagrange interpolation with limiters in [9] and an inverse Lax-Wendroff procedure for inlet boundaries and high order least squares interpolation for outlet boundaries, equipped with scale and dimension dependent weights in order to account for discontinuities, in [10, 11].

The contents of the paper are organized as follows: Section 2 stands for a brief description of some details about the numerical scheme and the procedure to automatically mesh complex domains through a Cartesian grid, as well as the criterion to choose the extrapolation stencil for the boundary conditions; in Sect. 3 an extrapolation procedure with scale and dimension independent weights is described; in Sect. 4 a numerical experiment involving a NACA0012 profile is shown as a matter of illustration; finally, in Sect. 5 some conclusions are drawn.

2 Numerical Scheme and Meshing Procedure

The equations that are considered here are d-dimensional $m \times m$ systems of hyperbolic conservation laws.

$$u_t + \nabla \cdot f(u) = 0, \quad u : \Omega \times \mathbb{R}^+ \subseteq \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^m, \quad f : \mathbb{R}^m \to \mathbb{R}^m, \tag{1}$$

with prescribed initial condition $u(x;0) = u_0(x)$, $x \in \Omega$, and suitable boundary conditions.

2.1 Numerical Scheme

For the sake of simplicity, let us describe the procedure for $d = 1$ and $\Omega = (0,1)$. To approximate the solution of (1), these equations are then discretized in space by means of a Shu-Osher finite-difference approach [8], by taking $x_j = (j + \frac{1}{2})h$, $0 \leq j < N$, $h = \frac{1}{N}$, therefore the spatial scheme can be written in local terms as follows:

$$u_t = -\frac{\hat{f}_{j+\frac{1}{2}}(t) - \hat{f}_{j-\frac{1}{2}}(t)}{h} + \mathcal{O}(h^r), \tag{2}$$

where

$$\hat{f}_{j+\frac{1}{2}}(t) = \hat{f}(u(x_{j-k+1}, t), \ldots, u(x_{j+k}, t))$$

are obtained through r-th order accurate WENO spatial reconstructions [5] with the Donat-Marquina flux-splitting technique [3].

The ODE (2) is then discretized (in time) through a Runge-Kutta procedure [7], yielding a fully discretized numerical scheme. This combination of techniques was proposed by Marquina and Mulet in [6].

2.2 Meshing Procedure

We summarize the meshing procedure for complex (non-rectangular) domains. We focus on the two-dimensional case for simplicity.

The computational domain is given by the intersections of the horizontal and vertical mesh lines that belong to the physical domain:

$$\mathscr{D} := \left((\overline{x} + h_x\mathbb{Z}) \times (\overline{y} + h_y\mathbb{Z})\right) \cap \Omega,$$

where $h_x, h_y > 0$ are the grid sizes for each respective dimension and $\overline{x}, \overline{y}$ are fixed real values.

In order to advance in time using WENO schemes of order $2k-1$, k additional cells are needed at both sides of each horizontal and vertical mesh line. If these additional cells fall outside the domain they are usually named *ghost cells*.

A normal direction associated to each ghost cell P and the domain boundary, $\partial\Omega$, is then computed by finding the point P_0 such that

$$\|P - P_0\|_2 = \min\{\|P - B\|_2 : \quad B \in \partial\Omega\}$$

and then considering the corresponding (interior) normal vector, $\overrightarrow{n}(P) = \overrightarrow{PP_0}$.

To extrapolate the information at each ghost cell, we first interpolate/extrapolate information from the computational domain to points on the normal line, and then extrapolate the obtained data together with the boundary condition, if any, in order to fill the missing value of the corresponding ghost cell.

Figure 1 illustrates graphically the selection procedure of nodes from the computational domain, which can be described through the following summarized algorithm:

1. Extrapolate/interpolate the selected data from the computational domain, $N_{i,j}$, $1 \le i, j \le R+1$, to the corresponding nodes N_i on the normal line, $1 \le i \le R+1$. The axis direction is chosen in terms of the angle of the normal vector.
2. If there are prescribed boundary conditions, the information contained in the normal line involving the recently extrapolated/interpolated values at the points

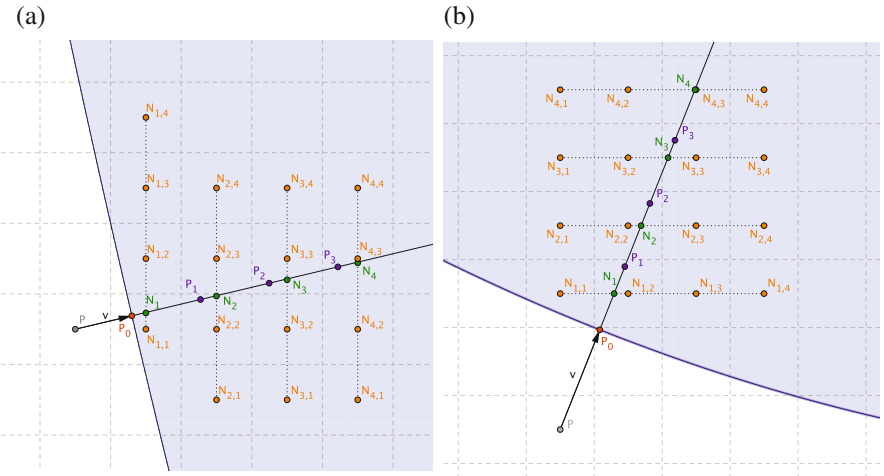


Fig. 1 Examples of choice of stencil ($R = 3$). We use the stencil $\mathscr{S}(P) = \{N_1, N_2, N_3, N_4\}$ in case of outflow boundary and conditions and the stencil $\mathscr{S}(P) = \{P_0, P_1, P_2, P_3\}$ in case of Dirichlet boundary conditions. **(a)** vertical boundary ($|\tan(\theta)| \leq 1$). **(b)** horizontal boundary ($|\tan(\theta)| > 1$), where θ is the angle of v, the normal vector to the boundary at P_0

N_i, $1 \leq i \leq R+1$, is used to interpolate/extrapolate them at new nodes P_i, $1 \leq i \leq R$, in a way such that the stencil including the boundary node P_0 (corresponding to the intersection of the normal line with $\partial\Omega$) are equally spaced.

3. There are two possibilities:
   (a) If boundary conditions are prescribed, we extrapolate the information contained in the stencil P_i, $0 \leq i \leq R$, at the ghost cell P.
   (b) Otherwise, we extrapolate directly the information contained in the points obtained at the first step, N_i, $1 \leq i \leq R+1$, to the ghost cell P.

In case of reflecting boundary conditions for the two dimensional Euler equations, with velocity component $\overrightarrow{v}$, then one defines

$$\overrightarrow{n} = \frac{\overrightarrow{P_0P}}{\left\|\overrightarrow{P_0P}\right\|}, \quad \overrightarrow{t} = \overrightarrow{n}^{\perp},$$

and obtains normal and tangential components of $\overrightarrow{n}$ at each point P_i of the mentioned segment by:

$$v^t(P_i) = \overrightarrow{v}(P_i) \cdot \overrightarrow{t}, \quad v^n(P_i) = \overrightarrow{v}(P_i) \cdot \overrightarrow{n}.$$

The extrapolation procedure is applied to $v^t(P_i)$ to approximate $v^t(P)$ and to $v^n(P_i)$ and $v^n(P_0) = 0$ to approximate $v^n(P)$. Once $v^t(P), v^n(P)$ are approximated, the approximation to $\overrightarrow{v}(P)$ is set to

$$\overrightarrow{v}(P) = v^t(P)\overrightarrow{t} + v^n(P)\overrightarrow{n}.$$

More detailed information about the meshing procedure can be found at [1, 2].

3 Extrapolation Procedure

Likewise the WENO scheme idea, we need to ensure that the extrapolation at the ghost cells is performed smoothly, namely, either there are no discontinuities in the extrapolation stencil or, in such case, a procedure to detect them and take only information from the correct side of the discontinuity is developed.

Let us consider a stencil of equally spaced nodes $x_0 < \cdots < x_R$, their corresponding nodal values $u_j = u(x_j)$ and x_* the extrapolation node. Then we define j_0 as the index corresponding to the closest node in the stencil to the extrapolation node, namely $j_0 = \operatorname{argmin}_{j\in\{0,\ldots,R\}} |x_j - x_*|$. We then define the corresponding smoothness indicator for each stencil $\{x_k, \ldots, x_{k+r_0}\}, 0 \le k \le R-r_0$:

$$I_k = \frac{1}{r}\sum_{\ell=1}^{r_0}\int_{x_0}^{x_r} h^{2\ell-1} p_k^{(\ell)}(x)^2 dx, \quad 0 \le k \le R - r_0,$$

where p_k is the polynomial of degree at most k such that $p_k(x_{i+j}) = u_{i+j}$ for $0 \le j \le r_0$, $0 \le i \le R - r_0$.

If r-th order accuracy is desired, $r \le R$, by stability motivations (see [2]), we now define p to be a polynomial of degree r satisfying $p(x_i) = u_i$, $0 \le i \le R$, by least squares.

Finally, we account for discontinuities by using properly the smoothness indicators computed previously. We thus define the following global average weight

$$\omega := \frac{(R - r_0 + 1)^2}{\left(\sum_{k=0}^{R-r_0} I_k\right)\left(\sum_{k=0}^{R-r_0} \frac{1}{I_k}\right)}.$$

This weight satisfies $0 \le \omega \le 1$ and

$$\omega = \begin{cases} 1 - \mathcal{O}(h^{2c_s}) \text{ if the stencil is } C^{r_0} \text{ with a } s\text{-th order zero derivative,} \\ \mathcal{O}(h^2) \text{ if the stencil contains a discontinuity,} \end{cases}$$

with $c_s := \max\{r_0 - s, 0\}$.

The final result of the extrapolation is thus defined by $u_* = \omega p(x_*) + (1 - \omega)u_{i_0}$. The above expression can be interpreted as a continuous transition between the full order extrapolation in case of smoothness in the stencil ($\omega \approx 1$) and the reduction to a first order constant extrapolation, taking the closest node from the stencil with respect to the extrapolation point as the reference value, in presence of discontinuities on the stencil ($\omega \approx 0$).

4 Numerical Experiment

In this section we show a numerical experiment with the flow given by the 2D Euler equations (with adiabatic constant $\gamma = 1.4$) around a NACA airfoil profile for $R = 8$, $r = 4$, $r_0 = 2$, WENO5 spatial discretization and RK3 time scheme (fifth order spatial accuracy and third order time accuracy). Prior tests in [2] showed that this scheme is indeed fifth order accurate in space in smooth problems, even when complex geometries are considered.

We consider a Mach 2 flow interacting with a NACA0012 profile [4] and we perform the simulation until $T = 5$, whose results are shown in Fig. 2 for the density and pressure fields.

The results are consistent with those obtained in Feistauer et al. [4] using finite volume techniques.

5 Conclusions

In this paper a technique to perform high order boundary extrapolations in finite difference methods has been successfully applied to supersonic inviscid flow around a slender NACA airfoil.

The extrapolation attains the desired accuracy in smooth problems as shown in [2] and behaves properly in presence of discontinuities, which is useful for problems of practical interest like the one presented herein.

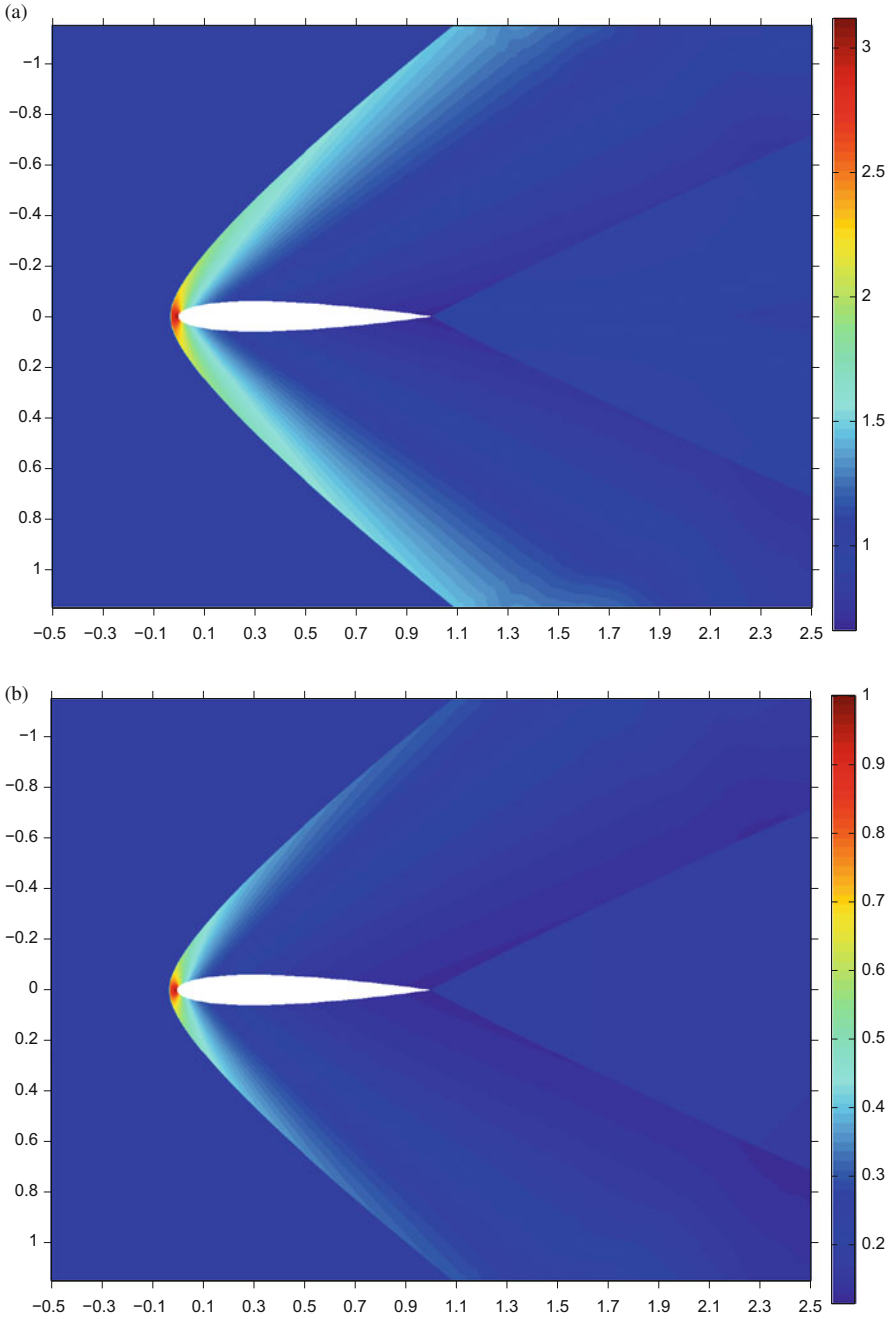


Fig. 2 Density and pressure fields for the supersonic NACA0012 interaction test. (**a**) Density field. (**b**) Pressure field

Acknowledgements This research was partially supported by Spanish MINECO grant MTM2014-54388-P.

References

1. Baeza, A., Mulet, P., Zorío, D.: High order boundary extrapolation technique for finite difference methods on complex domains with Cartesian meshes. J. Sci. Comput. **66**, 761–791 (2016)
2. Baeza, A., Mulet, P., Zorío, D.: High order weighted extrapolation for boundary conditions for finite difference methods on complex domains with Cartesian meshes. J. Sci. Comput. **69**, 170–200 (2016)
3. Donat, R., Marquina, A.: Capturing shock reflections: an improved flux formula. J. Comput. Phys. **125**, 42–58 (1996)
4. Feistauer, M., Felcman, J., Straškraba, I.: Mathematical and Computational Methods for Compressible Flow. Clarendon Press, Oxford (2003)
5. Jiang, G.S., Shu, C.W.: Efficient implementation of Weighted ENO schemes. J. Comput. Phys. **126**, 202–228 (1996)
6. Marquina, A., Mulet, P.: A flux-split algorithm applied to conservative models for multicomponent compressible flows. J. Comput. Phys. **185**, 120–138 (2003)
7. Shu, C.W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)
8. Shu, C.W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes, II. J. Comput. Phys. **83**(1), 32–78 (1989)
9. Sjogreen, B., Petersson, N.: A Cartesian embedded boundary method for hyperbolic conservation laws. Commun. Comput. Phys. **2**, 1199–1219 (2007)
10. Tan, S., Shu, C.W.: Inverse Lax-Wendroff procedure for numerical boundary conditions of conservation laws. J. Comput. Phys. **229**, 8144–8166 (2010)
11. Tan, S., Wang, C., Shu, C.W., Ning, J.: Efficient implementation of high order inverse Lax-Wendroff boundary treatment for conservation laws. J. Comput. Phys. **231**(6), 2510–2527 (2012)

The Influence of the Asymptotic Regime on the RS-IMEX

Klaus Kaiser and Jochen Schütz

Abstract In this work, we investigate the performance and explore the limits of a novel implicit-explicit splitting (Kaiser and Schütz, A high-order method for weakly compressible flows. Commun. Comput. Phys. 22(4): 1150–1174, 2017) for the efficient treatment of singularly perturbed ODEs. We consider a singularly perturbed ODE where, based on the choice of initial conditions, the unperturbed equation does not necessarily describe the behavior of the perturbed one accurately. For the splitting presented in Kaiser and Schütz, (A high-order method for weakly compressible flows. Commun. Comput. Phys. 22(4): 1150–1174, 2017), this has a tremendous influence as it explicitly depends on the solution to the unperturbed equation. That this indeed poses a problem is shown numerically; but also the remedy of using the 'correct' asymptotics is presented. Comparisons with a fully implicit and a standard implicit-explicit splitting are shown.

1 Introduction

Computing the solution to a singularly perturbed differential equation accurately *and* efficiently is a task that is ubiquitous in many scientific applications. One class of methods that is able to deal with such equations are IMEX schemes [1, 2]. At the very core of these schemes is a suitable identification of *stiff* terms (that should be treated in an *IMplicit* manner) and *non-stiff* terms (that should be treated in an *EXplicit* manner). Such a splitting of the equation can have tremendous influence on the accuracy and the efficiency of an IMEX scheme. Over the past few years, a splitting based on the unperturbed solution (the so-called *reference solution*, hence *RS-IMEX* splitting) has been developed, and it has been tested, amongst others, on an ODE and a PDE example [7, 8]. In all the applications, the solution to the perturbed problem was a very close approximation to the unperturbed solution. This

K. Kaiser
IGPM, RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany
e-mail: kaiser@igpm.rwth-aachen.de

J. Schütz (✉)
Faculty of Sciences, University of Hasselt, Agoralaan Gebouw D, 3590 Diepenbeek, Belgium
e-mail: jochen.schuetz@uhasselt.be

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_7

is not necessarily always the case for finite perturbation parameter, even if there is convergence for this parameter to zero. In the present work, we therefore try to give a glimpse on the behavior of the RS-IMEX splitting for a singularly perturbed system that has a richer structure. As a prototype equation, we consider the one suggested and analyzed in [5], namely

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} v \\ z \\ w \end{pmatrix} = \begin{pmatrix} \frac{1}{\varepsilon}\left(-z + f_2 v^2 + f_3 v^3\right) \\ v - w \\ \sqrt{\varepsilon}(\mu - g_1 z) \end{pmatrix}, \tag{1}$$

with, in analogy to [5], $f_2 = \frac{3}{2}, f_3 = -1$, $g_1 = \frac{1}{2}$ and $\mu = \frac{4}{100}$. It is fairly intuitive to see that the $\varepsilon \to 0$ limit (if there is any) fulfills the differential-algebraic equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} 0 \\ z_{(0)} \\ w_{(0)} \end{pmatrix} = \begin{pmatrix} -z_{(0)} + f_2 v_{(0)}^2 + f_3 v_{(0)}^3 \\ v_{(0)} - w_{(0)} \\ 0 \end{pmatrix}. \tag{2}$$

For the initial conditions we consider in this work, which are always of *well-prepared* form

$$z(t=0) = f_2 v(t=0)^2 + f_3 v(t=0)^3,$$

the solution to this last equation has a fairly simple structure, in particular, it exists over the whole real axis with a constant limit state as $t \to \infty$. The solution for $\varepsilon > 0$, on the other hand, has a quite rich structure (see e.g. [5] for a more thorough discussion of the equation). According to [5], we choose $w(t=0) = -\frac{1}{100}$ and the values of $v(t=0)$ are chosen as ± 0.1. All computations are done up to $t_{end} = 10$.

In Fig. 1 the solutions for different values of ε are plotted. We can observe that in all cases, for 'large' values of ε, a peak in the solution is present which disappears for $\varepsilon \to 0$. For the case $v(t=0) = 0.1$ we observe that the reference solution which is computed by (2), is not the limit we observe in the solution. Even more, we observe that there is an initial boundary layer (see Fig. 2 for a zoom of Fig. 1) for the case $v(t=0) = 0.1$. The conclusion that the solution converges towards the reference solution for $v(t=0) = -0.1$ springs to mind.

In what follows, we take different IMEX splittings and compare their behavior. Naturally, our focus is on the behavior of the RS-IMEX splitting and its performance in comparison to the other schemes. In particular, how does it perform if there is no valid asymptotic solution or if the solution converges towards a different reference solution?

This work is guided by the following questions:

- How does the RS-IMEX scheme behave if unperturbed and perturbed solution are not close to each other?
- In particular, is the RS-IMEX scheme then still uniformly stable as observed in [8]?
- If not, are there potential remedies?

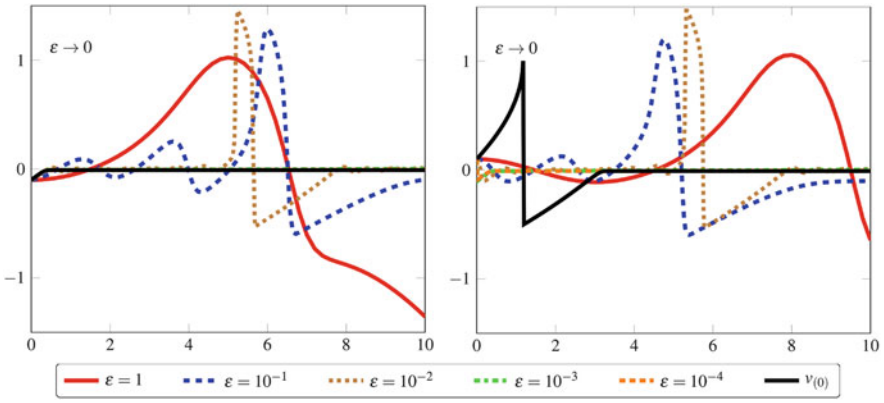


Fig. 1 *Left*: $v(t = 0) = -0.1$, *right*: $v(t = 0) = 0.1$

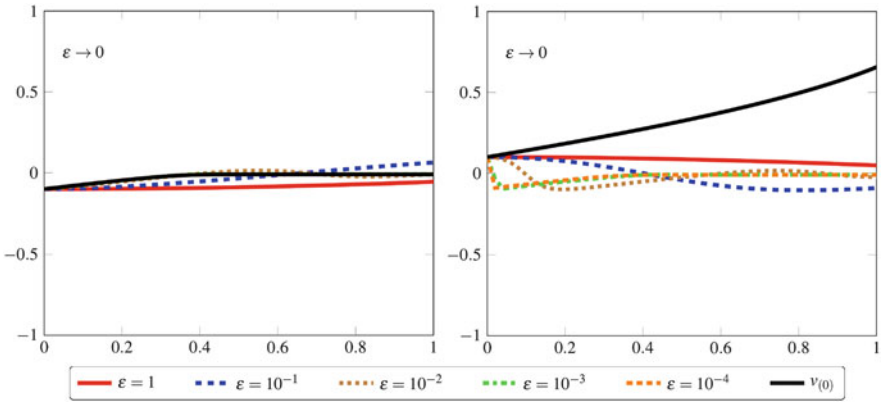


Fig. 2 Zoom into Fig. 1. *Left*: $v(t = 0) = -0.1$, *right*: $v(t = 0) = 0.1$. Note the initial layer for the case on the *right*

2 Comparison of Splittings

In this section we consider three different splittings, see Table 1:

- A 'canonical' splitting (SPLIT), where the $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$-terms are treated implicitly. A similar splitting was used in [3] for van der Pol's equation.
- The RS-IMEX splitting, where the reference solution is approximately computed by a fully implicit method, see [6] for details.
- A fully implicit method.

Numerical computations are run for all these splittings with IMEX Euler [2] and IMEX BPR-353 [4] method.

Table 1 Overview of different splittings used

Splitting	Implicit part	Explicit part
SPLIT	$\begin{pmatrix} \frac{-z+f_2y^2+f_3y^3}{\varepsilon} \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ y-w \\ \sqrt{\varepsilon}(\mu - g_1 z) \end{pmatrix}$
Implicit	$\begin{pmatrix} \frac{-z+f_2y^2+f_3y^3}{\varepsilon} \\ y-w \\ \sqrt{\varepsilon}(\mu - g_1 z) \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$
RS-IMEX	$\begin{pmatrix} \frac{-z+f_2 2y_{(0)}y+3f_3y_{(0)}^2y-f_2y_{(0)}^2-2f_3y_{(0)}^3}{\varepsilon} \\ y-w \\ \sqrt{\varepsilon}(\mu - g_1 z) \end{pmatrix}$	$\begin{pmatrix} \frac{f_2y^2+f_3y^3-f_2 2y_{(0)}y-3f_3y_{(0)}^2y+f_2y_{(0)}^2+2f_3y_{(0)}^3}{\varepsilon} \\ 0 \\ 0 \end{pmatrix}$

Left: implicit, *right*: explicit part

Decent Asymptotics: $\mathbf{v(t = 0) = -0.1}$

First, the easier case is considered, see also Fig. 1. In particular, there is an $\varepsilon \to 0$ asymptotic. Figures 3 and 4 show convergence results for IMEX Euler and IMEX BPR-353, respectively. All methods yield the desired convergence order, but the peak ($\varepsilon \geq 10^{-2}$) makes it somewhat difficult. Furthermore, both SPLIT and RS-IMEX have convergence problems for large values of Δt. For $\Delta t \to 0$, the RS-IMEX splitting produces approximations with a smaller error, and the error curve is generally smoother. Last but not least, there are stability problems for SPLIT, see Fig. 3 and the RS-IMEX splitting, see Fig. 4, for large values of ε. This behavior has also been observed for the RS-IMEX method presented in [6].

Difficult Asymptotics: $\mathbf{v(t = 0) = 0.1}$

This is certainly the most challenging setting for the RS-IMEX splitting, since the reference solution does not coincide with the $\varepsilon \to 0$ limit. The implications can clearly be seen in Figs. 5 and 6. Indeed one can see that the RS-IMEX splitting has huge stability problems for small values of ε and small values of Δt. This could be due to the initial layer, but most probably, it is due to the fact that the reference solution is not asymptotically close to the original solution; an assumption that stands at the beginning of the RS-IMEX approach. The other splittings are stable, but perform quite poorly.

The RS-IMEX approach is quite flexible, it allows tuning the reference solution in a certain sense. In particular, one can take the solution that results from taking $v_{(0)}(t = 0) = -0.1$, which seems to be a better approximation. Corresponding results are presented in Figs. 7 and 8. The results for the SPLIT splitting and the

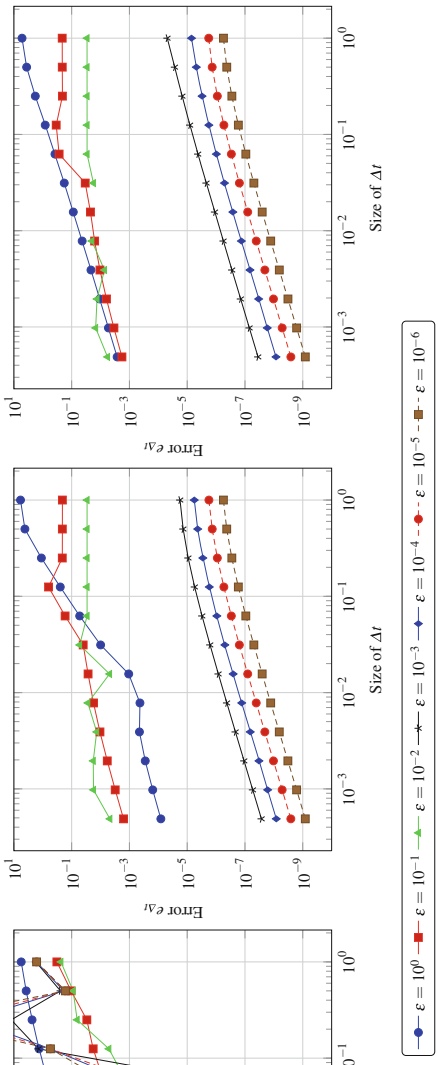


Fig. 3 Numerical results for IMEX-Euler method. Case $v(t = 0) = -0.1$ is considered. *Left*: SPLIT, *middle*: Implicit, *right*: RS-IMEX. Values not plotted are NaN, which show an instability of the method

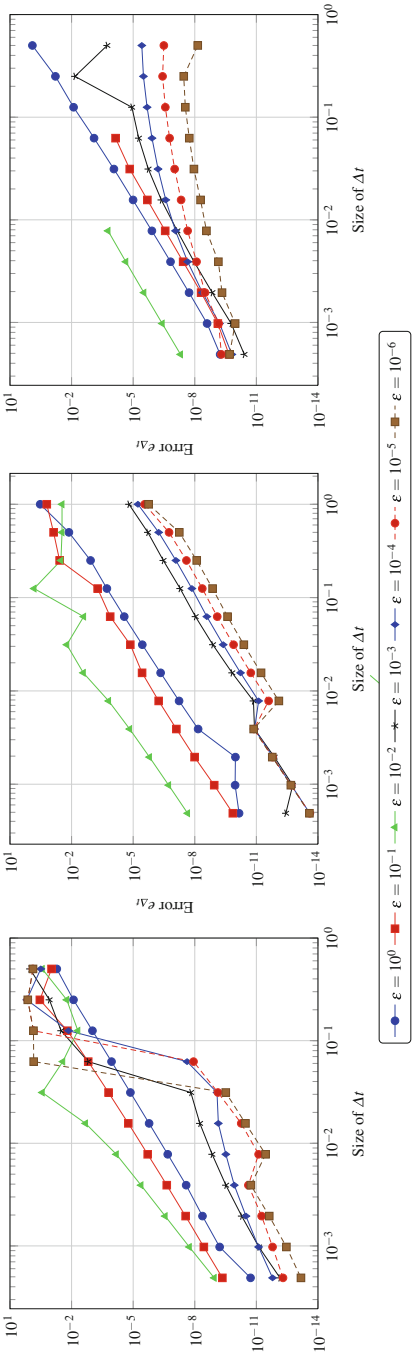


Fig. 4 Numerical results for IMEX BPR-353 method. Case $v(t = 0) = -0.1$ is considered. *Left*: SPLIT, *middle*: Implicit, *right*: RS-IMEX. Values not plotted are NaN, which show an instability of the method

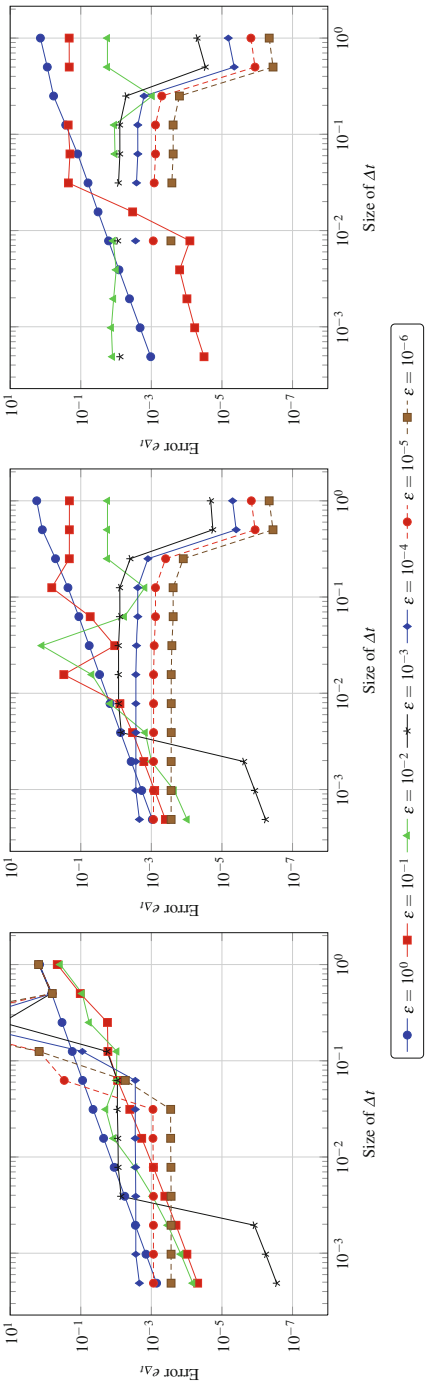


Fig. 5 Numerical results for IMEX-Euler method. Case $v(t = 0) = 0.1$ is considered. *Left*: SPLIT, *middle*: Implicit, *right*: RS-IMEX. Values not plotted are NaN, which show an instability of the method

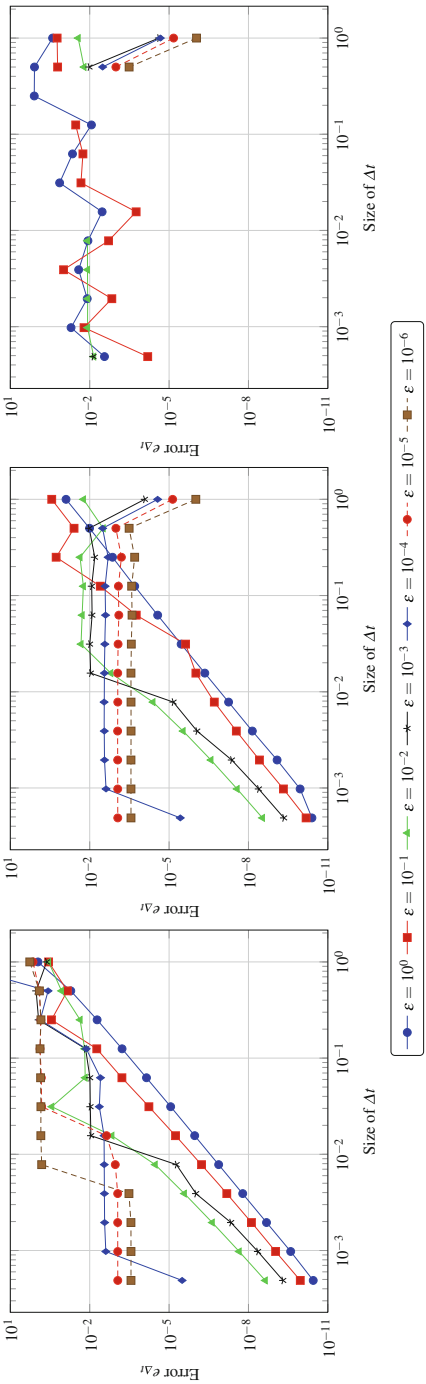


Fig. 6 Numerical results for IMEX BPR-353 method. Case $v(t = 0) = -0.1$ is considered. *Left*: SPLIT, *middle*: Implicit, *right*: RS-IMEX. Values not plotted are NaN, which show an instability of the method

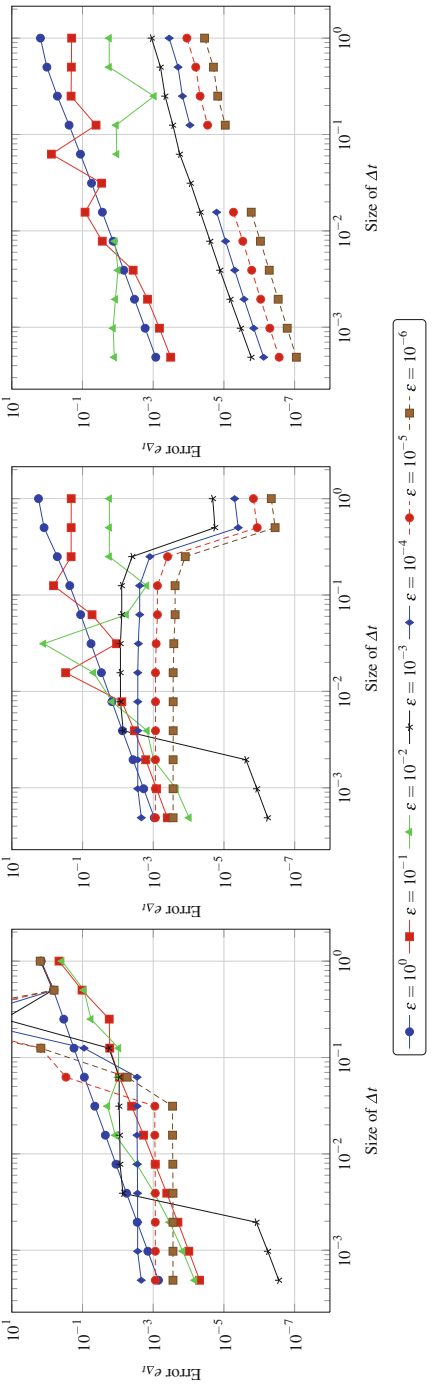


Fig. 7 Numerical results for IMEX-Euler method. Case $v(t = 0) = 0.1$ is considered. *Left*: SPLIT, *middle*: Implicit, *right*: RS-IMEX. The novelty is here that the RS-IMEX method uses the 'correct' limit solution. Values not plotted are NaN, which show an instability of the method

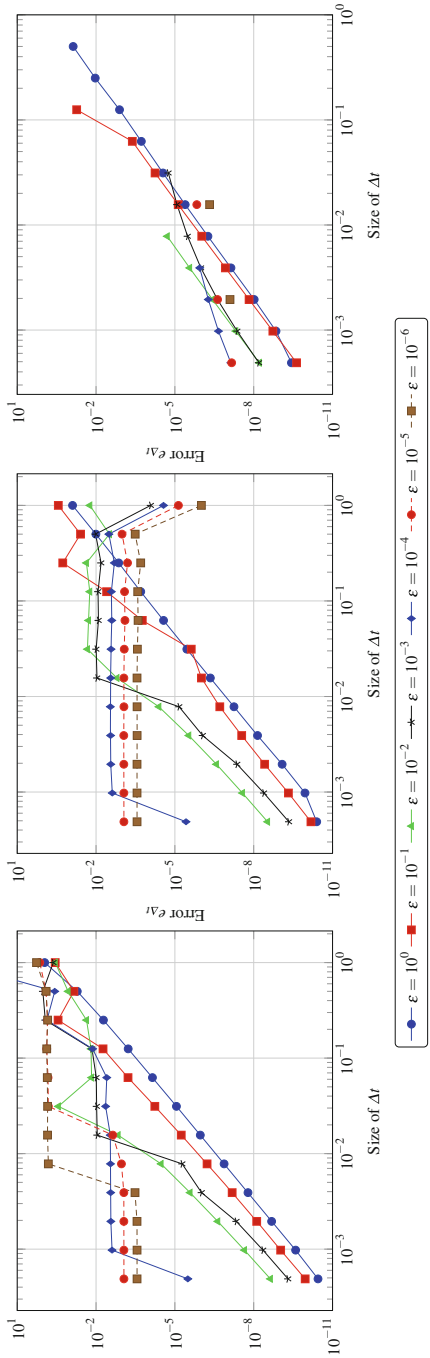


Fig. 8 Numerical results for IMEX BPR-353 method. Case $v(t = 0) = 0.1$ is considered. *Left*: SPLIT, *middle*: Implicit, *right*: RS-IMEX. The novelty is here that the RS-IMEX method uses the 'correct' limit solution. Values not plotted are NaN, which show an instability of the method

implicit method (we didn't change anything there) are the same as in the plots before. The RS-IMEX splitting, on the other hand, now performs much better. It has some stability issues for large values of Δt, which can be explained due to the small step size needed to resolve the initial boundary layer. After crossing this critical point the RS-IMEX splitting behaves similar or—in most cases—even better compared to the other splittings.

3 Conclusion and Outlook

In this work, we have demonstrated the performance of some stiff ODE solvers, with a focus on the RS-IMEX splitting. In particular, one could observe problems with the RS-IMEX method whenever the asymptotics are not the 'straightforward' way. The flexibility of the method, however, could in some cases save the day. Of course all this relies on the fact that one is able to identify the correct asymptotics of the problem. This could be achieved by computing the reference solution 'step-by-step' with the possibility of adapting it to the behavior of the approximate solution. Furthermore, it is well-established that such problems should be computed with some sort of adaptivity. Besides doing it via embedded Runge-Kutta methods—which are hard to get for high-order IMEX schemes—one can also use the difference of the solution and the reference solution as an error indicator. Based on this indicator, one can then directly reduce the time step or switch to another, more suited, method. This is in particular important as a step toward *all-Mach* schemes for the Navier-Stokes equations, where the (local) speed of the flow (i.e., in our setting the value of ε) can be small in one and quite large in another region.

Acknowledgements We thank Peter de Maesschalck for fruitful discussions. The first author has been partially supported by the German Research Foundation (DFG) project NO 361/3-3, and the University of Hasselt in the framework of the BOF 2016.

References

1. Ascher, U.M., Ruuth, S., Wetton, B: Implicit-Explicit methods for time-dependent partial differential equations. SIAM J. Numer. Anal. **32**, 797–823 (1995)
2. Ascher, U.M., Ruuth, S., Spiteri, R: Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. Appl. Numer. Math. **25**, 151–167 (1997)
3. Boscarino, S.: Error analysis of IMEX Runge-Kutta methods derived from differential-algebraic systems. SIAM J. Numer. Anal. **45**, 1600–1621 (2007)
4. Boscarino, S., Pareschi, L., Russo, G.: Implicit-explicit Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit. SIAM J. Sci. Comput. **35**(1), A22–A51 (2013)
5. De Maesschalck, P., Kutafina, E., Popović, N.: Sector-delayed-Hopf-type mixed-mode oscillations in a prototypical three-time-scale model. Appl. Math. Comput. **273**, 337–352 (2016)

6. Kaiser, K., Schütz, J.: A high-order method for weakly compressible flows. Commun. Comput. Phys. **22**(4), 1150–1174 (2017)
7. Kaiser, K., Schütz, J., Schöbel, R., Noelle, S.: A new stable splitting for the isentropic Euler equations. J. Sci. Comput. **70**(3), 1390–1407 (2017)
8. Schütz, J., Kaiser, K.: A new stable splitting for singularly perturbed ODEs. Appl. Numer. Math. **107**, 18–33 (2016)

Minisymposium: Aeroacoustics

Patrick Joly and Jean-François Mercier

Description

Aeroacoustics can be seen as a branch of fluid mechanics that deals with the propagation of acoustic waves in moving fluids (flows). There are obviously many industrial applications of aeroacoustics to begin with aeronautics and car industry (controlling the noise emitted by airplanes and cars is definitely a major issue) but also the control of noise in domestic ducts like air-conditioning for instance.

As a consequence, understanding in depth sound propagation in flows and being able to simulate numerically the corresponding physical phenomena are two major issues that represent scientific challenges for applied mathematicians. Even one could think that computational aeroacoustics is only a part of computational fluid dynamics, as a matter of fact, the modelization of acoustics in flows is a discipline of its own right. Indeed, taking into account the small proportion of mechanical energy involved in the production of sound, one traditionally uses mathematical models issued from various versions of the linearization of full nonlinear models, such as Euler of Navier-Stokes equations, around a reference flow that is most often considered as stationary (time independent) at the time scales of interest.

Dealing with such models from both mathematical and computational points of view raises new interesting and non trivial questions.

The first one is the choice of a mathematical model and the design of a corresponding numerical method, the two aspects being intimately linked. Indeed, many "equivalent" models are available (Linearized Euler Equations, Galbrun's equations, Golstein's equations, Möhring's equations, ...) and lead to various simplifications according to particular properties of the reference flow (laminar flow,

P. Joly (✉) • J.-F. Mercier
POEMS, CNRS-INRIA-ENSTA UMR 7231, 828 Boulevard des Maréchaux, 91762 Palaiseau, France
e-mail: patrick.joly@inria.fr; jean-francois.mercier@ensta-paristech.fr

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_8

potential (or irrotational) flow, ...). These models mainly differ by the choice of the unknowns that are used but also in their mathematical form that is more or less well adapted to a given mathematical (for the analysis) or numerical (for simulation) approach. These choices will also be guided by the fact that one is interested in a transient problem (in this case, one treats an evolution problem in which time is a variable) or in a time harmonic problem (one looks for a solution that oscillates in time at a given frequency). In the recent past years, various numerical methods have been developed around the use of continuous or discontinuous Galerkin methods for space discretization and, when needed, finite differences in time. Such methods are still the object of various improvements and their analysis still raise open questions.

A second very important question is the treatment of boundary conditions. This question can be divided into two parts. The first one concerns physical boundaries that is related to the interaction of acoustic waves with walls. The second one is related to artificial boundaries introduced for bounding the computational domain. In the first case, various effective boundary conditions (whose expression is, by the way, influenced by the choice of the "interior" model) have been recently proposed to take into account thin boundary layers as well as absorbing walls. Their use raises delicate questions related to the well-posedness and stability analysis of the corresponding boundary value problems. In the second case, which also raises difficult stability questions, several existing methods are competing, including Perfectly Matched Layers or time domain integral equations (retarded potentials).

The aim of this minisymposium was to propose to give an overview of the state of the art around these questions and to emphasize the most recent developments in the domain.

The talks included in the minisymposium were the following:

- J.F. Mercier. POEMS, CNRS-INRIA-ENSTA (France) *A brief review on some open questions in aeroacoustics modeling.*
- C. Legendre. Free Field Technologies (Belgium) *Effects of non-uniform mean flows on sound propagation using scalar operators.*
- A. Semin. Technische Universität Berlin (Germany) *Asymptotic modelling of the wave-propagation over microperforated absorbers.*
- A. Bensalah. POEMS, CNRS-INRIA-ENSTA *Introduction to the Goldstein Model in aeroacoustics.*
- J. Chabassier. INRIA Bordeaux (France) *Solution of time-harmonic Galbrun's equation in the context of helioseismology.*
- J. Rodríguez. Universidade de Santiago de Compostela (Spain) *Boundary integral equations and discontinuous Galerkin methods for transient aeroacoustics.*

Simulation of Reflection and Transmission Properties of Multiperforated Acoustic Liners

Adrien Semin, Anastasia Thöns-Zueva, and Kersten Schmidt

Abstract In this paper we study the fundamental and higher-order modes propagation in a cylindrical acoustic duct with a multi-perforated liner section. This study relies on an established approximate model that is mathematically verified by a multiscale analysis and that takes the presence of the liner into account through transmission conditions. We simulate the reflection and transmission behaviour by an *hp*-adaptive finite element method that effectively resolves the solution in presence of strong singularities at the rim of the duct. Moreover, we introduce a new mode matching method based on the complete mode decomposition that depends in the liner section on the Rayleigh conductivity. It turns out that the mode matching method achieves similar accuracies with all propagating and just a number of evanescent modes.

1 Introduction

Nowadays combustion processes create acoustic sources of higher intensity, which in their turn create acoustic instabilities at particular frequencies and may even harm the live time of the gas turbine. It is also a well known fact that perforations or orifices in general can be applied to reduce a noise pollution and widely used, for instance soundproofing perforated wall panels, car mufflers, acoustic liners. The perforated liners in combustion chambers are able to suppress thermoacoustic instabilities and can provide a substantial amount of acoustic damping [5].

A. Semin (✉) • A. Thöns-Zueva
Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: adrien.semin@math.tu-berlin.de; anastasia.thoens@math.tu-berlin.de

K. Schmidt
Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany

Institut für Mathematik, Brandenburgische Technische Universität Cottbus-Senftenberg, Platz der deutschen Einheit 1, 03046 Cottbus, Germany
e-mail: kersten.schmidt@math.tu-berlin.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_9

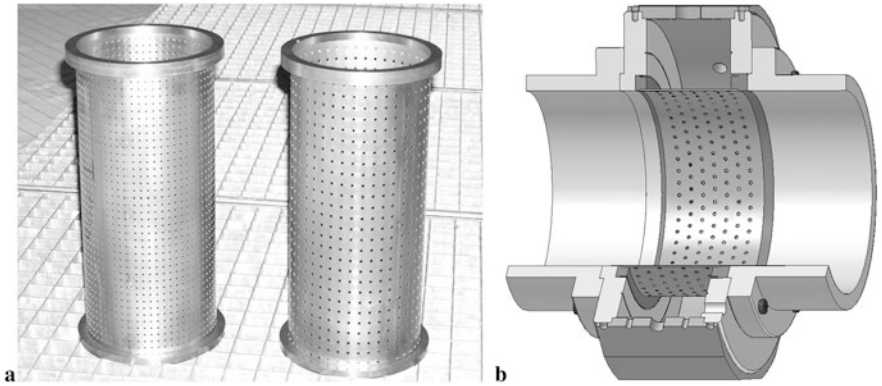


Fig. 1 (**a**) Perforated liners in a duct acoustic test rig (DLR Institute of Propulsion Technology). (**b**) Simplified geometry of a combustion liner for acoustic studies: a cylindrical perforated liner (*Courtesy of F. Bake, DLR Berlin*)

A combustion instability is a discrete frequency phenomena and the frequency characteristics of the liner is one of its most important features. To understand the frequency characteristics the performance of a liner is measured in experiments in acoustic test benches (see Fig. 1b) over a range of frequencies [6]. The major properties are described with the Rayleigh conductivity introduced by J.W.S. Rayleigh [8] that is the quotient of the pressure jump and the velocity through the liner and closely related to the wave length and so to the frequency ω. Usually an approximation of the Rayleigh conductivity is obtained in experiments by comparing the measurements with a solution obtained by a modal decomposition of the sound field in the duct and above the liner. In single-mode wave-guides the incident plane wave is the only propagating mode, i.e. fundamental, and represent the dominant behavior. However, the interaction of the plane wave with higher order propagating modes (at higher frequencies) is of a large interest and studied experimentally and theoretically e.g. in [6, 9, 10]. In this article we show first that the usual impedance transmission conditions [1] are obtained as the limit of the δ-dependent model where hole distances scale like δ and the hole diameters like δ^2, even when considering the singular behavior close to the rim of the duct. This analysis is based on a solution representation close to the perforated wall, close to their edges that are the rims of the duct and away from the perforated wall and using the method of matched asymptotic expansion (cf. [2, 3]) and the solution of a Laplace problem around one hole. Then, we introduce a new mode matching method that uses N propagating and evanescent modes in each of two parts of the duct and $4N$ propagating and evanescent modes in the liner and analyse its accuracy numerically in comparison with a highly accurate finite element solution.

2 Surface Homogenization

We consider an infinite cylindrical duct of radius R_d that is connected to a chamber of total radius $R_\mathrm{c} > R_\mathrm{d}$ and of length L by a multi-perforated circular thin liner $\Omega^\delta_\mathrm{liner}$. The wall of the duct is continued with in the liner section with an array of perforations with distances δ along the direction of the waveguide axis and distances proportional to δ in the angular direction of the cross-section. We also assume that the diameter of each hole is of order δ^2 as the thickness of the perforated wall. As $\delta \to 0$, the domain $\Omega^\delta_\mathrm{liner}$ degenerates to the limit interface $\Gamma_\mathrm{liner} = \{R_\mathrm{d}\} \times (0, 2\pi) \times (0, L)$ (in cylindrical coordinates). We denote then by $\Omega = (0, R_\mathrm{d}) \times (0, 2\pi) \times \mathbb{R} \cup (0, R_\mathrm{c}) \times (0, 2\pi) \times (0, L)$ the infinite duct without the wall and $\Sigma^- = \{R_\mathrm{d}\} \times (0, 2\pi) \times \{0\}$, $\Sigma^+ = \{R_\mathrm{d}\} \times (0, 2\pi) \times \{L\}$ the re-entrant edges of Ω (see Fig. 2). Finally, we denote by Ω^δ the infinite duct with the multi-perforated liner, i.e., $\Omega^\delta := \Omega \setminus \Omega^\delta_\mathrm{liner}$. On this domain we introduce the Helmholtz transmission problem to be considered in this article. For an incident wave p_inc of wave number k_0 coming from the left we seek p^δ as the solution of the total field problem

$$-\Delta p^\delta - k_0^2 p^\delta = 0, \quad \text{in } \Omega^\delta \tag{1a}$$

$$-\nabla p^\delta \cdot \mathbf{n} = 0, \quad \text{on} \partial\Omega^\delta \tag{1b}$$

$$\lim_{z\to\pm\infty} \pm\partial_z(p^\delta - p_\mathrm{inc}) - T^1_\pm(p^\delta - p_\mathrm{inc}) = 0, \tag{1c}$$

where $T^1_\pm$ is the Dirichlet-to-Neumann operator based on the projection on the outgoing propagative modes [4, Eq. (2.7)]. We assume the incident plane wave p_inc to be a guided mode of the duct (see Sect. 3), being of the form $\psi_{j_0}(r)\exp(\mathrm{i}(m\theta + \beta_{j_0} z))$ satisfying (1a) and (1b). As $\delta \to 0$ the solution p^δ of (1) tends to its limit value p_0, i.e., $p^\delta(\mathbf{x}) = p_0(\mathbf{x}) + o(1)$, where p_0 is satisfies

$$-\Delta p_0 - k_0^2 p_0 = 0, \quad \text{in } \Omega \setminus \Gamma_\mathrm{liner}, \tag{2}$$

(1b) on $\partial\Omega$ and (1c). However, this expansion is non-uniform in the whole domain Ω^δ but only away from the limit interface Γ_liner. Close to the perforated liner and away from the two re-entrant edges $\Sigma^\pm$ the solution is represented as

$$p^\delta(\mathbf{x}) = \Pi_0(\mathbf{x}_\Gamma, \mathbf{X}) + \delta\Pi_1(\mathbf{x}_\Gamma, \mathbf{X}) + o(\delta). \tag{3}$$

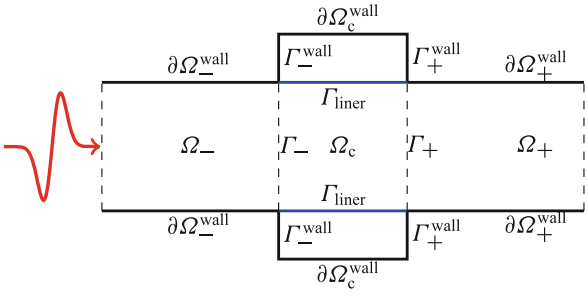


Fig. 2 Split of the domain Ω into the two semi-infinite waveguides $\Omega_\pm$ and the multi-perforated liner section Ω_c. Impedance transmission conditions on the interface Γ_liner approximate the behaviour of the many perforations

In (3) the boundary layer correctors Π_0 and Π_1 are functions of the slow variable $\mathbf{x}_\Gamma := (\theta, z)$ and of the fast variable $\mathbf{X} := (R, \Theta, Z) = ((r - R_\mathrm{d})/\delta, \theta/\delta, z/\delta)$ and are 1-periodic in Z and $2\pi/N_\Theta$-periodic in Θ when assuming that the number of holes in the angular direction is N_Θ/δ. The boundary layer correctors are solution of the Laplace equation in the infinite cylinder $(R, \Theta, Z) \in (-0.5, -0.5) \times (-\pi/N_\Theta, \pi/N_\Theta) \times \mathbb{R}$ with Neumann conditions at $Z = 0$ and with a point source at $(0, 0, 0)$. Solving these problems and writing the matching between the macroscopic part of the solution and its periodic boundary corrector gives equivalently transmission conditions for p_0 on the limit interface. More precisely,

$$\big[\nabla p_0 \cdot \mathbf{n}\big](\mathbf{x}) = 0 \quad \text{and} \quad \big\langle \nabla p_0 \cdot \mathbf{n} \big\rangle(\mathbf{x}) = C_R \big[p_0\big](\mathbf{x}) \quad \text{for } \mathbf{x} \in \Gamma_\mathrm{liner}, \tag{4}$$

where the parameter C_R is obtained from the boundary layer correctors and turns out to be proportional to the inverse of the Rayleigh conductivity K_R [7, 8]. Moreover, $\mathbf{n}$ denotes the normalized normal vector on Γ_liner pointing into the side chamber, $\big[\,\cdot\,\big]$ and $\big\langle\,\cdot\,\big\rangle$ stand for the jump and average over the interface, respectively. Close to the two re-entrant edges $\Sigma^\pm$, the solution p^δ is represented as $p^\delta(\mathbf{x}) = P_0^\pm(\theta, \mathbf{X}^\pm) + o(1)$, the near-field corrector $P_0^\pm$ depends on the slow variable θ and on the fast variable $\mathbf{X}^\pm := (R^\pm, \Theta, Z^\pm)$ with $\mathbf{X}^- = ((r - R_\mathrm{d})/\delta, \theta/\delta, z/\delta)$ and $\mathbf{X}^+ = ((r - R_\mathrm{d})/\delta, \theta/\delta, (z - L)/\delta)$. This near-field corrector is $2\pi/N_\Theta$ periodic with respect to Θ and is bounded towards $|\mathbf{X}^\pm| \to \infty$. Therefore, it is stated that the limit term p_0 remains bounded in a vicinity of the two re-entrant edges $\Sigma^\pm$ and contains no edge singularity.

3 Modes Matching

For the incoming wave $\psi_{j_0}(r)\exp(\mathrm{i}(m\theta + \beta_{j_0} z))$ the solution p_0 of (2)–(4) depends on θ due to the linearity of this problem and the axial symmetry of Ω as $\exp(\mathrm{i}\, m\, \theta)$. Hence, we can represent it in the two semi-infinite waveguides $\Omega_\pm$ (see Fig. 2) as

$$p_0(r, \theta, z) = p_\mathrm{inc}(r, \theta, z) + \exp(\mathrm{i}\, m\, \theta) \sum_{j=0}^{\infty} \alpha_j^- \, \psi_j(r) \, \exp(-\mathrm{i}\, \beta_j\, z) \quad \text{in } \Omega_-, \tag{5a}$$

$$p_0(r, \theta, z) = \exp(\mathrm{i}\, m\, \theta) \sum_{j=0}^{\infty} \alpha_j^+ \, \psi_j(r) \, \exp(\mathrm{i}\, \beta_j\, z) \quad \text{in } \Omega_+, \tag{5b}$$

where $\beta_j = \beta_j(k_0)$ is the wave number for the j^th eigenmode, $\psi_j(r)$ is solution of the generalized 1D eigenvalue problem

$$\frac{1}{r}\partial_r(r\partial_r\psi_j(r)) = \big(\beta_j^2 - k_0^2 - \tfrac{m^2}{r^2}\big)\psi_j(r), \qquad r \in (0, R_\mathrm{d}), \tag{6a}$$

$$\partial_r\psi_j(r) = 0, \qquad r \in \{0, R_\mathrm{d}\}, \tag{6b}$$

and $\alpha_j^{\pm} = \alpha_j^{\pm}(k_0)$ is the modal amplitude of the mode. Using the theory of self-adjoint operators and the properties of the Bessel function J_m of the first kind, the eigenvalues β_j are simple and isolated, $\beta_j^2 \in (-\infty, k_0^2]$ so that we choose $\beta_j \in \mathbb{R}_+ \cup \mathrm{i}\mathbb{R}_+$ to satisfy (1c), and they are ordered by strictly decreasing values of β_j^2. The mode shapes are 2D-orthonormal to each other, i.e. $\int_0^{R_d} \psi_j(r)\,\psi_k(r)\,r\mathrm{d}r = \delta_{jk}$. For j large enough the wave number β_j becomes purely complex and the corresponding modes are not able to propagate. These modes are referred to as evanescent modes. Their amplitude decays exponentially with axial distance from their source.

In the liner section Ω_c, the solution p_0 of (2)–(4) with lateral Neumann boundary condition is similarly written under the form

$$p_0(r,\theta,z) = \exp(\mathrm{i}\,m\,\theta)\sum_{j=0}^{\infty} \psi_j'(r)\big(\alpha_j'^{+}\exp(\mathrm{i}\,\beta_j'\,z) + \alpha_j'^{-}\exp(\mathrm{i}\,\beta_j'\,(L-z))\big). \tag{7}$$

The eigenpairs (β_j', ψ_j') have for any $C_R > 0$ the same above mentioned properties as the eigenpairs (β_j, ψ_j) and the eigenfunctions ψ_j' are now solutions of

$$\frac{1}{r}\partial_r(r\partial_r\psi_j'(r)) = \big(\beta_j'^2 - k_0^2 - \tfrac{m^2}{r^2}\big)\psi_j'(r), \quad r \in (0,R_\mathrm{d}) \cup (R_\mathrm{d}, R_\mathrm{c}), \tag{8a}$$

$$\partial_r\psi_j'(r) = 0, \quad r \in \{0, R_\mathrm{c}\}, \tag{8b}$$

$$\big[\partial_r\psi_j'\big] = 0, \quad \big\langle\partial_r\psi_j'\big\rangle = C_R\big[\psi_j'\big]. \tag{8c}$$

It remains to match the pressure and its normal derivative of (5) and (7) on $\Gamma_\pm$ and to satisfy the Neumann boundary condition on $\Gamma_\pm^{\mathrm{wall}}$. We propose to match in a least L^2-square sense, truncating (5) (respectively (7)) to the first N modes (resp. to the first $2N$ modes to the left and $2N$ modes to the right), and the pressure jump (resp. the jump of its normal derivative for $|r| < R_d$ together with the Neumann boundary condition for $|r| > R_d$) is projected onto each ψ_j, $0 \leqslant j < N$ (resp. ψ_j', $0 \leqslant j < 2N$). In the following we introduce $< u, v >:= \int_0^{R_d} u(r)\,v(r)\,r\mathrm{d}r$.

The projection of the pressure jump at $z = 0$ and $z = L$ on the mode ψ_j gives

$$\alpha_j^- - \sum_{k=0}^{2N-1} < \psi_k', \psi_j > \big(\alpha_k'^{-}\exp(\mathrm{i}\,\beta_j'\,L) - \alpha_k'^{+}\big) + < \psi_{\mathrm{inc}}, \psi_j > = 0, \tag{9}$$

$$\exp(\mathrm{i}\,\beta_j\,L)\alpha_k^+ - \sum_{k=0}^{2N-1} < \psi_k', \psi_j > \big(\alpha_k'^{-} - \exp(\mathrm{i}\,\beta_k'\,L)\alpha_k'^{+}\big) = 0, \tag{10}$$

and the projection of the jump of the normal derivative at $z = 0$ and $z = L$ on the mode ψ'_j (assuming that it is normal derivative from the liner section for $|r| > R_d$)

$$\sum_{k=0}^{N-1} < \psi_k, \psi'_j > \beta_k \alpha_k^- - \beta'_j \exp(\mathrm{i}\,\beta'_j L)\alpha_j'^- + \beta'_j \alpha_j'^+ - \beta_{\mathrm{inc}} < \psi_{\mathrm{inc}}, \psi'_j > = 0, \tag{11}$$

$$\sum_{k=0}^{N-1} < \psi_k, \psi'_j > \beta_k \exp(\mathrm{i}\,\beta_k L)\alpha_k^+ + \beta'_j \alpha_j'^- - \beta'_j \exp(\mathrm{i}\,\beta'_j L)\alpha_j'^+ = 0. \tag{12}$$

Solving (9)–(12) gives the modal amplitudes $\alpha_j^\pm$ and $\alpha_j'^\pm$ of (5) and (7).

4 Numerical Results

We simulate the reflection and transmission properties for the example of duct with $R_\mathrm{d} = 40\,\mathrm{mm}$, a chamber with $R_\mathrm{c} = 64.375\,\mathrm{mm}$ and $L = 72\,\mathrm{mm}$. The wavenumber k_0 is given by $k_0 = 2\pi f_0/c_0$, where $c_0 = 343.2\,\mathrm{m\,s^{-1}}$ is the speed of sound in air and $f_0 = 1100\,\mathrm{Hz}$ is the frequency of the incident wave. As incident wave p_{inc} we opt for a plane wave given by $p_{\mathrm{inc}}(r, \theta, z) = \exp(\mathrm{i}\,k_0\,z)$. This configuration admits only one propagative mode.

First, we solve for the reflection and transmission coefficients related to (2)–(4) with a finite element method, for which we truncate the domains $\Omega_\pm$ at distance $L_\pm = 100\,\mathrm{mm}$, where we set a Robin boundary condition corresponding the Dirichlet-to-Neumann operator $T^1_\pm$ on the artificial boundaries $\{z = \{-L_-, L + L_+\}\}$. For the numerical solution we use the C++ library Concepts (www.tu-berlin.de/?concepts). We start with a coarse mesh generated using the mesh generator Gmsh (www.gmsh.info) containing 752 hexahedra and a polynomial degree 2 in all cells (level 0 with 6319 degrees of freedom) and refine the mesh geometrically towards the interface Γ_{liner} representing the liner, the walls $\partial\Omega_\pm^{\mathrm{wall}}$ of the duct and the interfaces between the duct and the liner sections. We take the best solution after 11 levels of refinement, leading to about 340,000 degrees of freedom, computed in parallel on 16 nodes in 2.3 h, as a reference. We observe in Fig. 3a that the relative error in the first transmission coefficient α_0^+ and the first reflection coefficient α_0^- decay exponentially in the refinement level ℓ where the number of degrees of freedom increase like $\ell^{2.5}$.

For numerical computation of the modes in the duct and the perforated liner section we use a finite element discretization of the two cross-sections with 2D meshes containing 36 curved quadrilaterals in the liner section and 20 curved quadrilaterals in the duct both with a uniform polynomial degree 7. As eigenvalue solver we use Arpack++ (www.caam.rice.edu/software/ARPACK/arpack++.html).

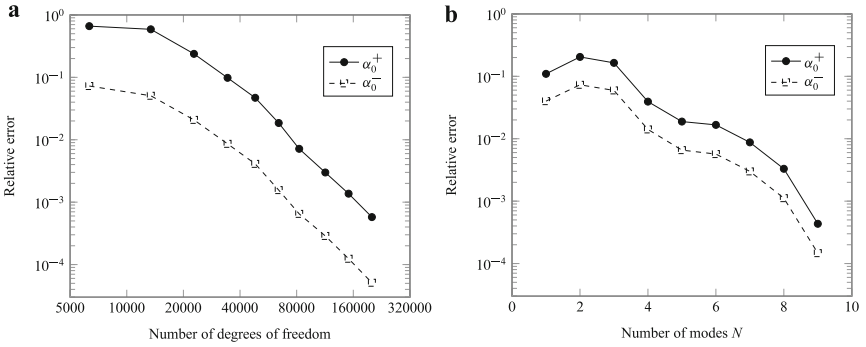


Fig. 3 Relative error in the first transmission coefficient α_0^+ and the first reflection coefficient α_0^- for (**a**) an *hp*-adaptive finite element method and (**b**) the new mode matching method

The relative error in the transmission and reflection coefficients α_0^+ and α_0^- with the new mode matching method, see (9)–(12), is shown in Fig. 3b as function of the modes N in the duct (corresponding to $2N$ modes in the liner section). What is observed is that only with a few evanescent modes the error decreases to much lower levels. Keeping in mind that computational effort increases only quadratically in N the mode matching method is an extremely efficient method despite the occurrence of singularities. The solution obtained with $N = 9$ modes and the solution of the *hp*-adaptive computation with 9 levels of refinement leading to about 200,000 degrees of freedom have the same level of accuracy.

Acknowledgements The authors gratefully acknowledge the financial support from the Einstein Foundation Berlin (grant number IPF-2011-98) and the research center MATHEON through the Einstein Center for Mathematics Berlin (project MI-2) and are thankful to the fruitful exchange with the DLR Berlin.

References

1. Bendali, A., Fares, M., Piot, E., Tordeux, S.: Mathematical justification of the rayleigh conductivity model for perforated plates in acoustics. SIAM J. Appl. Math. **73**(19), 438–459 (2013)
2. Delourme, B., Schmidt, K., Semin, A.: On the homogenization of thin perforated walls of finite length. Asymptot. Anal. **97**(3–4), 211–264 (2016). doi:10.3233/ASY-151350
3. Delourme, B., Schmidt, K., Semin, A.: On the homogenization of the Helmholtz problem with thin perforated walls of finite length. arXiv:1611.06001 (2016)
4. Goldstein, C.: A finite element method for solving Helmholtz type equations in waveguides and other unbounded domains. Math. Comput. **39**(160), 309–324 (1982)
5. Ingard, U., Labate, S.: Acoustic circulation effects and the nonlinear impedance of orifices. J. Acoust. Soc. Am. **22**(2), 211–218 (1950). doi:10.1121/1.1906591
6. Lahiri, C.: Acoustic performance of bias flow liners in gas turbine combustors. PhD thesis (2014). doi:10.14279/depositonce-4270

7. Laurens, S., Tordeux, S., Bendali, A., Fares, M., Kotiuga, P.R.: Lower and upper bounds for the Rayleigh conductivity of a perforated plate. ESAIM: Math. Model. Numer. Anal. **47**(6), 1691–1712 (2013). doi:10.1051/m2an/2013082
8. Rayleigh, J.W.S.: The Theory of Sound. In: Dover Classics of Science and Mathematics, vol. 2. Dover, New York (1945)
9. Weng, C., Bake, F.: An analytical model for boundary layer attenuation of acoustic modes in rigid circular ducts with uniform flow. Acta Acust. united Acust. **102**, 1138–1141 (2016). doi:10.3813/AAA.919025
10. Weng, C., Otto, C., Peerlings, L., Enghardt, L., Bake, F.: Experimental investigation of sound field decomposition with higher order modes in rectangular ducts. In: AIAA/CEAS Aeroacoustics Conference (2016). doi:10.2514/6.2016-3035

Minisymposium: Applied Mathematics in Stent Development

Tuoi T.N. Vo and Sean McGinty

Description

Coronary artery disease is a global problem and devising effective treatments is the subject of intense research activity throughout the world. Over the past decade, stents have emerged as one of the most popular treatments. Acting as a supporting scaffold, these small mesh devices are now routinely inserted into arteries where the blood flow has become dangerously restricted. Stents have evolved from bare metal scaffolds to polymer coated drug-delivery vehicles and, more recently, sophisticated fully biodegradable drug delivery configurations. Despite these advances, significant opportunities to improve on arterial stent design remain. The relative success of coronary artery stenting has led to the emergence of stenting technology for the carotid, neural and peripheral vasculature. In addition, the adaptability of the stent concept has opened horizons beyond the vasculature, with stent technology now being developed for, amongst others, pulmonary, gastro-intestinal and structural heart applications.

This minisymposium covered the development of innovative models to help industry optimise and improve the stent design, to identify the key parameters governing the behaviour of the system, to simulate the flow of plasma around complex stent geometries, to identify the drug release mechanism, and to help decrease the number of experimental studies, thereby saving time and money. As well as providing a platform for contributed talks from members of the newly formed ECMI special interest group '*Advancing the design of medical stents*', this

T.T.N. Vo (✉)
Department of Mathematics and Statistics, MACSI, University of Limerick, Limerick, Ireland
e-mail: tuoi.vo@ul.ie

S. McGinty
School of Engineering, University of Glasgow, Glasgow, Scotland
e-mail: sean.mcginty@glasgow.ac.uk

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_10

minisymposium was opened to researchers with expertise in continuum mechanics, physiological flow modelling, structural and soft tissue mechanics, numerical analysis, mathematical biology and multi-objective optimisation, to name but a few.

The talks included in the minisymposium were the following:

- S. McGinty. University of Glasgow (Scotland). *The role of mathematics in stent development.*
- Abdul I. Barakat. Ecole Polytechnique (France). *Optimizing the performance of drug-eluting stents: simulations and experiments.*
- T.T.N. Vo. University of Limerick (Ireland). *Mathematical models of drug release from polymer-free drug-eluting stents.*
- G. Pontrelli. IAC-CNR, Rome (Italy). *Variable porosity coatings as a means of controlling drug release from stents.*
- J. Escuer. University of Zaragoza (Spain). *Numerical simulation of drug transport in arterial wall under healthy and atherosclerotic conditions.*

Mathematical Modelling of Drug Elution from Drug-Filled Stents

Tuoi T.N. Vo, Amy M.M. Collins, and William T. Lee

Abstract In this paper, we outline the first model to describe drug elution from a drug-filled stent. This novel polymer-free drug-eluting stent stores the therapeutic drug in the inner layer of a tri-layer wire. This inner layer acts a reservoir releasing the drug through laser-drilled holes on the outer surface of the stent struts. We simplify the general model using the assumption of low drug solubility and consider a special case where the dissolution occurs in a uniform downward direction. The main advantage of our simplified model is the ability to achieve analytical solutions. These solutions allow for calculating the drug release profile rapidly and for identifying the dependence of the various parameters of the system. We find that the duration of drug elution is prolonged when the solubility or diffusivity of the drug decreases.

1 Introduction

A Drug-filled Stent (DFS, Medtronic, Santa Rosa, CA) is an innovative polymer-free drug-eluting stent which is designed to obtain controlled and sustained drug elution without a polymer. The polymeric platform, although traditionally crucial in the control and delivery of the therapeutic drug, is widely believed to cause irritation due to degradation leading to chronic inflammation and other long term complications [2]. The new DFS is made from a novel tri-layer wire which consists of the outer cobalt alloy layer for strength, the middle tantalum layer for radiopacity, and the inner core layer. After removal of the core material in this inner layer, the hollowed out centre can be used as a drug reservoir for the therapeutic drug, sirolimus. The drug is released from a single continuous inner core to the artery wall through multiple laser-drilled holes on the outer surface of the stent [5, 6].

T.T.N. Vo (✉) • A.M.M. Collins
MACSI, University of Limerick, Limerick, Ireland
e-mail: tuoi.vo@ul.ie; amy.m.m.collins@gmail.com

W.T. Lee
University of Portsmouth, Portsmouth, UK
e-mail: william.lee@port.ac.uk

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_11

In this paper, we develop a general mathematical model to describe drug elution by diffusion via direct interaction with the arterial wall. The drug transport is described by an advection-diffusion equation and a Navier-Stokes equation is used to describe fluid transport between the arterial wall and the stent. The model also includes a Stefan condition to determine the motion of a moving dissolution front separating the undissolved and dissolved drug in the stent inner core [4]. This model is simplified by nondimensionalising and assuming that the drug solubility is very slow. Then we are able to derive analytical solutions from the simple model to obtain release profiles that can be compared to experimental data. We also aim to identify key parameters associated to the control and delivery of drug release from a DFS.

2 The Model

We consider a small segment of the stent strut of length w and focus on the drug release from a single hole with diameter r_H ($r_H \ll w$). Let L the depth of the hole and R the total depth of hole and DFS core as seen in Fig. 1. Next, let Ω_1 represent the empty region of aqueous liquid inside the stent strut with the drug concentration $c_1(x, y, z, t)$, Ω_2 represent the region of the hole with the drug concentration $c_2(x, y, z, t)$, and $\partial\Omega_{12}$ represent the interface between these two regions. On $\partial\Omega_{12}$, we ensure continuity of both the drug concentration and flux for all time t. We also impose a perfect sink condition $c_2 = 0$ at $y = 0$ to account for the rapid drug lost to the arterial wall. Initially we assume that no drug is present in Ω_2 and solid drug fills the entire DFS core so that Ω_1 does not exist. Therefore, we assume that c_1 and c_2 are zero at $t = 0$.

Now let $\partial\Omega_1$ represent the interface of the solid which is initially stationary but once in contact with the aqueous liquid the drug begins to dissolve leading to the

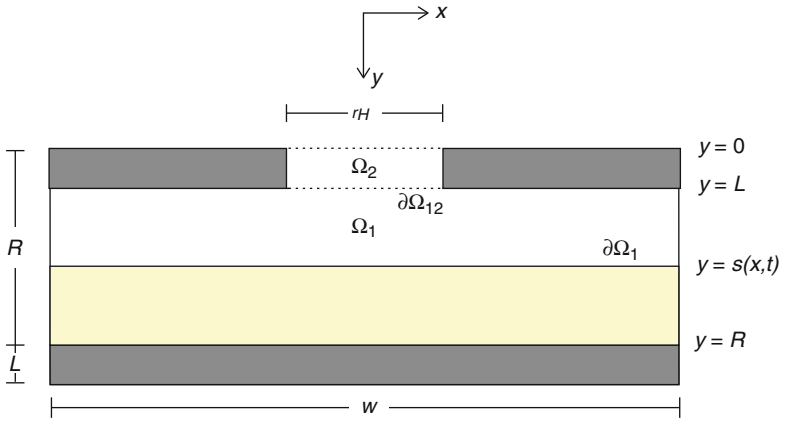


Fig. 1 Cross-section of the stent segment considered in the models in Sect. 2.2

mixture of the two phases. This results in the formation of a new homogeneous phase (the solution) and a moving boundary which we will denote as $y = s(x, z, t)$. This moving boundary is modelled by the Stefan condition [4]:

$$\frac{\partial s}{\partial t} = \frac{D}{c_0 - c_s}(\hat{\mathbf{n}} \cdot \nabla c_1), \tag{1}$$

where c_s is the drug solubility, which is the maximum possible concentration of dissolved drug in the release medium, and $\hat{\mathbf{n}}$ is the unit normal vector pointing into Ω_2. At the interface between the undissolved and dissolved drug, we suppose that the drug concentration c_1 is at solubility, so that $c_2 = c_s$ on $y = s(x, z, t)$. Once dissolution occurs, transport of the drug solute is modelled by the following advection-diffusion and Navier-Stokes equations as well as the associated incompressibility condition:

$$\begin{aligned} \frac{\partial c_i}{\partial t} + \nabla \cdot (\mathbf{u} c_i) &= D\nabla^2 c_i \quad \text{for} \quad i = 1, 2, \\ \rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u}\right) &= -\nabla p + \mu \nabla^2 \mathbf{u}, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \tag{2}$$

where D is the diffusion coefficient, $\mathbf{u}$ is the flow velocity of the solution, p is the pressure, and μ is the dynamic viscosity. Here, we assume that the density of the aqueous liquid ρ is constant.

2.1 Model Reduction

For simplification, we reduce the problem by considering the two-dimensional geometry and rewrite the system (2) in dimensionless form using the following scales:

$$x, y, s \sim R, \quad t \sim \frac{R^2}{Dc^*}, \quad c_1, c_2 \sim c_s, \quad \mathbf{u} \sim \frac{Dc^* w}{Rr_H}, \quad p \sim \frac{\mu D c^* w}{R^2 r_H},$$

where $c^* = \frac{c_s}{c_0 - c_s}$, to obtain:

$$\begin{aligned} \delta c^* \frac{\partial c_i}{\partial t} + c^* \nabla \cdot (\mathbf{u} c_i) &= \delta \nabla^2 c_i \quad \text{for} \quad i = 1, 2, \\ \delta \mathrm{Re} \frac{\partial \mathbf{u}}{\partial t} + \mathrm{Re}(\mathbf{u} \cdot \nabla)\mathbf{u} &= -\nabla p + \nabla^2 \mathbf{u}, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \tag{3}$$

where $\delta = \frac{r_H}{w}$, $\lambda = \frac{L}{R}$, $\alpha = \frac{c_0}{c_s}$, and $\mathrm{Re} = \frac{\rho D c^*}{\mu}$ which is equivalent to the standard definition of the Reynolds number $\mathrm{Re} = \frac{\rho U L^*}{\mu}$, where $U = \frac{Dc^* w}{R r_H}$ and $L^* = \frac{R r_H}{w}$. Hence, non-dimensionalisation may indicate a physical relationships in these four dimensionless parameters which can be altered for optimised drug release.

Drug adhesion and solubility are long thought to be vital to prolonging the duration of elution. Many companies opt for therapeutic drugs which have very low solubility and strong adhesion properties. Therefore, we consider the case where drug solubility is low by imposing the condition that $c_s \ll c_0$ or, equivalently, $c^* \ll 1$. Furthermore, we assume that $c^* \ll \delta \ll 1$. This leads to the uncoupling of Eqs. $(3)_1$ and $(3)_2$. Thus, we can neglect the effects of advection in the system, and obtain the following reduced model:

$$\begin{aligned}
&\nabla^2 c_i = 0 \quad \text{for} \quad i = 1, 2, \quad \text{in } \Omega_1, \Omega_2, \\
&\frac{\partial s}{\partial t} = \hat{\mathbf{n}} \cdot \nabla c_1, \quad c_1 = 1, \quad \text{on } \partial\Omega_1, \\
&c_1 = c_2 \quad \text{and} \quad \hat{\mathbf{n}} \cdot \nabla c_1 = \hat{\mathbf{n}} \cdot \nabla c_2 \quad \text{on } \partial\Omega_{12}, \\
&c_2 = 0 \quad \text{on} \quad y = 0, \\
&s = y^*, \quad c_i = 0 \quad \text{at} \quad t = 0,
\end{aligned} \tag{4}$$

where $y^*(x)$ represents the initial level of the solid drug. Next, we consider a simplified case of uniform dissolution.

2.2 A Simple Model for Uniform Dissolution

In this case, we consider that the hole is large enough for drug dissolution to occur in a uniformly downward direction. The above system of Eq. (4) is converted to be one-dimensional by horizontal averaging in a similar manner to the method used in the reduction of the St. Venant shallow flow equations [1, 3]. To do this, we integrate over the horizontal regions for both Ω_1 and Ω_2:

$$\frac{w}{R}\nabla^2 \tilde{c}_1(y,t) = 0 \quad \text{where} \quad \tilde{c}_1(y,t) = \frac{R}{w}\int_{-\frac{w}{2R}}^{\frac{w}{2R}} c_1(x,y,t)\,dx, \tag{5}$$

$$\frac{r_H}{R}\nabla^2 \tilde{c}_2(y,t) = 0 \quad \text{where} \quad \tilde{c}_2(y,t) = \frac{R}{r_H}\int_{-\frac{r_H}{2R}}^{\frac{r_H}{2R}} c_2(x,y,t)\,dx, \tag{6}$$

and obtain analytical solutions for $\tilde{c}_1$, $\tilde{c}_2$, and $s(t)$:

$$\tilde{c}_1 = \frac{\lambda(\delta-1)-\delta y}{\lambda(\delta-1)-\delta s(t)}, \quad \tilde{c}_2 = -\frac{y}{\lambda(\delta-1)-\delta s(t)}, \tag{7}$$

where

$$s(t) = \frac{\lambda(\delta - 1) + \sqrt{2\delta^2 t + \lambda^2}}{\delta}. \tag{8}$$

The time, t_f, at which all the solid drug has completely dissolved is calculated by solving $s(t_f) = 1$ and is given by:

$$t_f = \frac{\lambda^2(\delta - 2) - 2\lambda(\delta - 1) + \delta}{2\delta}. \tag{9}$$

Denoting by $M(t)$ the amount of drug that has dissolved in the DFS core by time t, we obtain:

$$M(t) = c_0 A_1 (R - L) - A_1 \int_{s(t)}^{R} c_0 dy - A_1 \int_{L}^{s(t)} \tilde{c}_1 dy - A_2 \int_{0}^{L} \tilde{c}_2 dy, \tag{10}$$

where A_1 and A_2 are the surface area of Ω_1 and Ω_2, respectively. The total amount of the drug is given by $M(\infty) = A_1(R-L)c_0$. Therefore, the fraction of drug dissolved in the DFS core by time t is given by $M(t)/M(\infty)$:

$$\frac{M(t)}{M(\infty)} = \begin{cases} 1 - \frac{1-s(t)}{1-\lambda} - \frac{1}{\alpha(1-\lambda)}\left(\int_{\lambda}^{s(t)} \tilde{c}_1 dy + \delta \int_{0}^{\lambda} \tilde{c}_2 dy\right) & \text{for} \quad 0 \le t \le t_f, \\ 1 \quad \text{for} \quad t > t_f. \end{cases} \tag{11}$$

The release profile which is important when comparing with experimental results is found by plotting $M\% = M(t)/M(\infty) \times 100$ versus time.

To investigate the effects of different parameters, we plot the release profile for various values of α, D, λ, and δ. From Fig. 2, we find that increasing α, lowering drug solubility c_s, or decreasing the diffusion coefficient D results in significantly longer elution periods as expected. Note that the values of α and D depend on the properties of a specific drug.

We also consider the importance of parameters associated to the geometry of the stent: δ and λ. We first look at the effect of varying λ which corresponds to the variation of the volume of the drug channel. We initially expected that increasing λ would lead to an increase in the duration of the elution period. However, we found that a stent with a 50% drug channel and a 70% drug channel have very similar endpoints, as seen in Fig. 3a. Therefore, structurally we would recommend a 50% drug channel as opposed to a 70% drug channel as little benefit in terms of the drug elution duration is gained from this increase. Note that larger a drug channel results in weaker a stent strut. Figure 3b shows the release profile with variation of δ to simulate the presence of 1,3, 5, and 7 holes per 1000 μm length of stent strut each with the average diameter of 20 μm. We find that decreasing the hole size or number of holes per unit length of stent leads to an increase in the elution period.

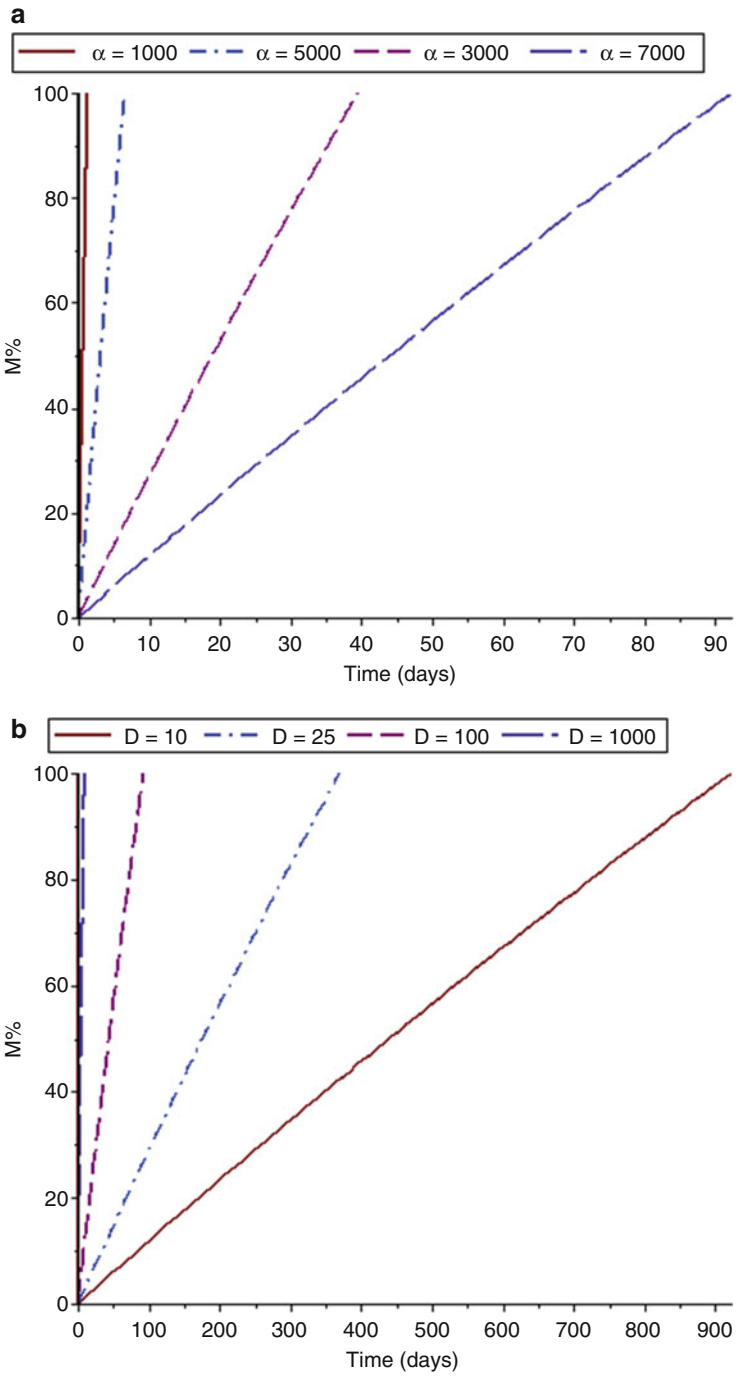


Fig. 2 Release profiles correspond to the variation of α with $D = 100$ $(\mu m)^2/s$ (**a**), and D with $\alpha = 70{,}000$ (**b**). Here $\delta = 0.10$ and $\lambda = 0.33$ (50% drug channel)

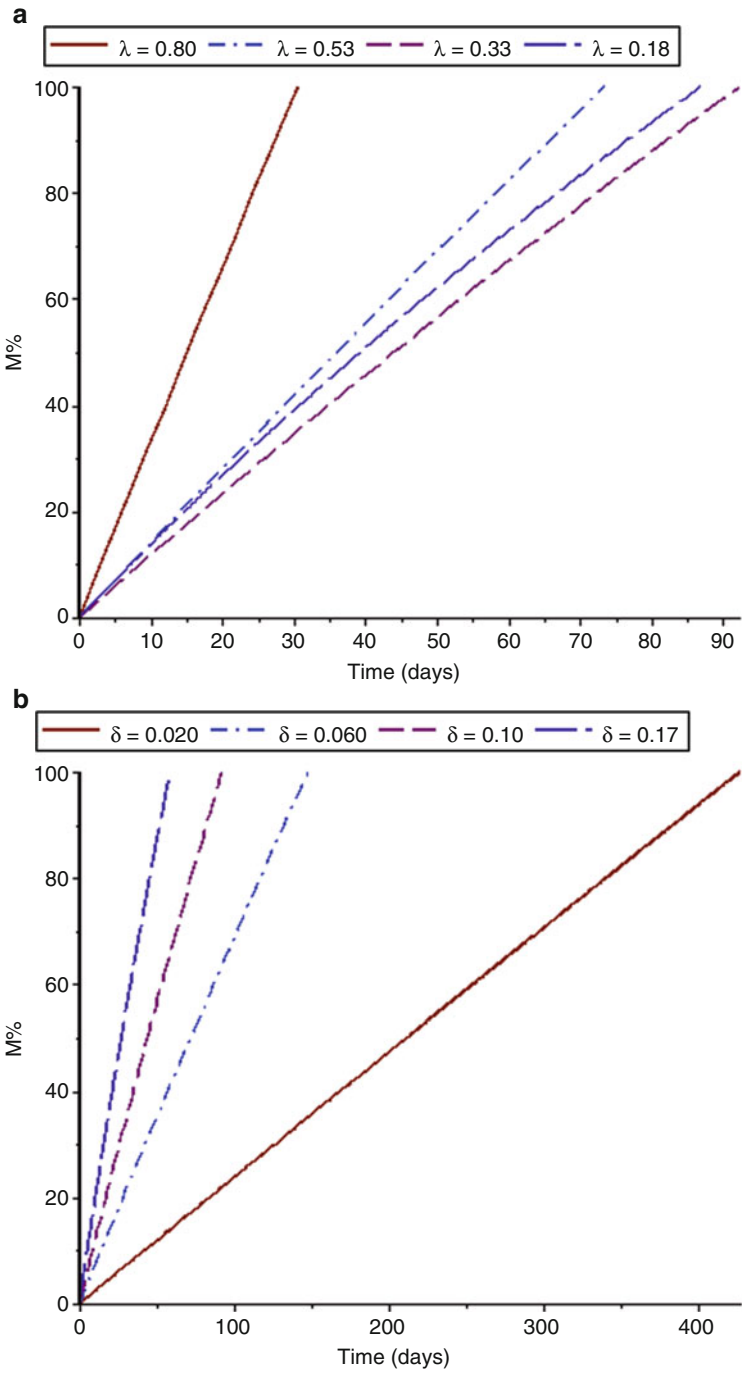


Fig. 3 Release profiles correspond to the variation of λ with $\delta = 0.10$ (**a**), and δ with $\lambda = 0.33$ (50% drug channel) (**b**). Here $D = 100\ (\mu m)^2/s$ and $\alpha = 70{,}000$

3 Conclusion

We developed mathematical models to describe the drug release from a drug-filled stent. The general model includes two governing equations: an advection-diffusion equation to describe solute transport and a Navier-Stokes equation to describe fluid transport from the thin film between the stent and the arterial tissue. This model was simplified using the assumption that the solubility concentration is much smaller than the initial concentration of the therapeutic drug. This assumption allowed us to uncouple the two governing equations and neglect the effects of fluid velocity of the system. We also assumed that an average hole diameter is much smaller than the diameter of the stent segment considered. A perfect sink condition was used on the outer surface of the stent and there was no presence of the drug initially in the hole.

A special case where the dissolution occurs in a uniform downward direction was considered. In this case, we only considered a two dimensional model where the horizontal direction is averaged for simplicity. This simple model can provide some insight into identifying key parameters that control the drug release rate from drug-filled stents. These parameters include the average hole size, volume of the drug channel, drug solubility, and diffusivity. In the optimisation of these parameters, we must also consider the physical restrictions of this alteration on the properties associated to both the drug and stent geometry. When considering the alteration of the drug we examined the diffusivity and the solubility equilibrium. We found that decreasing the solubility or diffusivity of the drug results in a longer elution period. While for the modification of stent geometry, we found that longer elution periods occur when the diameter of the holes is narrowed or the number of holes per unit length of the stent is reduced.

In our results, the shape of the release profile is linear which is not the case in polymer-coated drug eluting stents. This linearity may be due to our assumptions used to simplify the model. However, the main advantage of our simplified model is the ability to achieve analytical solutions. These solutions not only allow for release profiles to rapidly calculated, but they also clearly show the dependence of the release rate on various parameters of the system. This is of great benefit when considering the required design parameters to achieve a particular release profile.

Acknowledgements The authors acknowledge the support of MACSI, the Mathematics Applications Consortium for Science and Industry (http://ulsites.ul.ie/macsi/), funded by the Science Foundation Ireland Investigator Award 12/IA/1683.

References

1. Abdul Hadi, M.S., Mustapa, M.Z., Saad, S.: Complete derivation of 2D shallow-water model from the primitive equations governing geophysical flows. In: 8th International Conference on Marine Technology, 20–22 October 2012, Terengganu, Malaysia (2012)

2. Abizaid, A., Costa, J.R.: New drug-eluting stents an overview on biodegradable and polymer-free next-generation stent systems. Circ. Cardiovasc. Interv. **3**, 384–393 (2010)
3. Cea, L., Puertas, J., Vázquez-Cendón, M.E.: Depth averaged modelling of turbulent shallow water flow with wet-dry fronts. Arch. Comput. Meth. Eng. **14**(3), 303–341 (2007)
4. McGinty, S., Vo, T.T.N., Meere, M.G., McKee, S., McCormick, C.: Some design considerations for polymer-free drug-eluting stents: a mathematical approach. Acta Biomater. **18**, 213–225 (2015)
5. Medtronic plc: Medtronic drug-filled stent shows promising initial results following first implants. Medtronic Press Release. Available via DIALOG. http://newsroom.medtronic.com/phoenix.zhtml?c=251324&p=irol-newsArticle&ID=2096128 (2015). Cited 3 Oct 2016
6. Wilcox, J.N.: Progress with a zotarolimus-filled stent technology and preclinical results. Medtronic Press Release. Available via DIALOG. http://www.tctmd.com/privateAssets/0/114/709599/725792/8ad0f2dd-6ad5-41f5-821c-3c0304461300.pdf (2015). Cited 3 Oct 2016

Minisymposium: Charge Transport in Semiconductor Materials: Emerging and Established Mathematical Topics

Patricio Farrell, Dirk Peschka, and Nella Rotundo

Description

Modern electronic devices would be unthinkable without semiconductor materials. Few inventions have shaped our modern society like they have. There are emerging fields such as organic semiconductors, where there is still the demand to develop models and new theory, and there are many well-established fields, e.g. particle detector simulations, where it is mostly required to develop advanced numerical methods.

The aim of this minisymposium was to encourage communication between both worlds, applied mathematics (or industrial mathematics) and the field of fundamental research on numerical methods, mathematical modeling, and optimal control problems related to semiconductor (charge transport) models.

The talks included in the minisymposium were the following:

- René Pinnau. TU Kaiserslautern, Department of Mathematics (Germany). *Combining Asymptotic Analysis and Optimization in Semiconductor Design.*
- Patricio Farrell. Weierstrass Institute Berlin (Germany). *Scharfetter-Gummel Schemes for Non-Boltzmann Statistics.*
- Klaus Gärtner. Università della Svizzera italiana, Institute of Computational Sciences (Italy). *Drift-diffusion simulations of silicon detectors for astro- and high-energy- physics.*
- Matthias Liero. Weierstrass Institute Berlin (Germany). *On $p(x)$-Laplace thermistor models describing electrothermal feedback in organic semiconductor devices.*

P. Farrell (✉) • D. Peschka • N. Rotundo
Weierstrass Institute Berlin (WIAS), Berlin, Germany
e-mail: farrell@wias-berlin.de; peschka@wias-berlin.de; rotundo@wias-berlin.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_12

- Victor Burlakov. Mathematical Institute, OCCAM, Oxford (UK). *Multi-Scale Kinetics of Photo-Excited Charges in Organo-Lead Perovskites.*
- Wil Schilders. TU Eindhoven (Netherlands). *Numerical methods for the simulation of organic and anorganic.*
- Axel Fischer. Institut für Angewandte Photophysik, TU Dresden (Germany). *Multi-physics simulation of organic devices using LTspice for first principle understanding.*
- Matthias Auf der Maur. University of Rome Tor Vergata, Rome (Italy). *Multi-scale approaches for electronic device simulation.*
- Dirk Peschka. Weierstrass Institute Berlin (Germany). *Toward the optimization of Ge microbridges.*

Comparison of Scharfetter-Gummel Flux Discretizations Under Blakemore Statistics

Patricio Farrell, Thomas Koprucki, and Jürgen Fuhrmann

Abstract We discretize the semiconductor device equations assuming a Blakemore distribution function using a finite volume scheme and compare three thermodynamically consistent Scharfetter-Gummel type flux discretizations, namely the exact solution to a two-point boundary value probem and two fluxes incorporating certain averages. In order to do this, we simulate an n-i-n semiconductor device and study the electron densities as well as the total current. While the diffusion-enhanced flux approximation using logarithmic averaging of the nonlinear diffusion enhancement behaves somewhat similarly to the exact solution of the two-point boundary value problem (the generalized Scharfetter-Gummel scheme), the scheme based on averaging the inverse activity coefficients scheme exhibits a noticeably different behavior.

1 Introduction

The semiconductor device equations (also known as the van Roosbroeck system) have become a standard model to describe the flow of electron and holes in a semiconductor device. They are very well understood if one may presume Boltzmann statistics as the governing law describing the dependency of the electron and hole densities on the corresponding chemical potentials. The use of non-Boltzmann statistics raises challenging questions concerning the space discretization. The goal of this paper is to study the influence of three Scharfetter-Gummel [9] type of flux approximations used in a finite volume discretization of the van Roosbroeck system when assuming more general statistics. In particular, we will focus on the Blakemore distribution function and compare simulation results of a device that consists of an n-doped, intrinsic and another n-doped region (n-i-n). All three schemes are thermodynamically consistent and avoid artificial steady state dissipation [1].

P. Farrell (✉) • T. Koprucki • J. Fuhrmann
Weierstrass Institute (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
e-mail: farrell@wias-berlin.de; koprucki@wias-berlin.de; fuhrmann@wias-berlin.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_13

2 Van Roosbroeck System

The (stationary) van Roosbroeck system describes the charge carrier flow and the electrostatic potential in a semiconductor device. It consists of three nonlinear partial differential equations: one for the electrostatic potential ψ and two continuity equations for the electron and hole densities (n and p):

$$-\nabla \cdot (\varepsilon_0 \varepsilon_r \nabla \psi) = q\,(p - n + C)\,, \tag{1a}$$

$$\nabla \cdot \mathbf{j}_n = qR, \tag{1b}$$

$$\nabla \cdot \mathbf{j}_p = -qR. \tag{1c}$$

We consider a homogeneous material and some domain $\Omega \subseteq \mathbb{R}^d$ for $d \in \{1, 2, 3\}$. The constants q, ε_0 and ε_r denote the elementary charge, the vacuum dielectric permittivity and the relative permittivity of the semiconductor, respectively. The recombination rate R depends on both densities. The doping profile C may vary spatially.

The electron density and the hole density are related to the electric potential ψ as well as the quasi Fermi potentials of electrons and holes φ_n and φ_p via

$$n = N_c \mathscr{F}\left(\frac{q(\psi - \varphi_n) - E_c}{k_B T}\right) \quad \text{and} \quad p = N_v \mathscr{F}\left(\frac{q(\varphi_p - \psi) + E_v}{k_B T}\right). \tag{1d}$$

For the purpose of the paper we will assume *Blakemore statistics* [2], i. e.

$$\mathscr{F}(x) = (\exp(-x) + \gamma)^{-1} \qquad \text{with} \qquad \gamma = 0.27. \tag{1e}$$

The Boltzmann approximation is given by $\mathscr{F}(x) = \exp(x)$. The temperature T, the conduction and valence band densities of states N_c and N_v as well as the conduction and valence band-edge energies E_c and E_v are assumed constant. The parameter k_B denotes the Boltzmann constant. The current densities in (1b) and (1c) can be written in drift-diffusion form

$$\mathbf{j}_n = -q\mu_n n \nabla \psi + qD_n \nabla n, \qquad \mathbf{j}_p = -q\mu_p p \nabla \psi - qD_p \nabla p. \tag{1f}$$

The diffusion coefficients D_n and D_p are related to the carrier mobilities μ_n and μ_p by a *generalized* Einstein relation

$$\frac{D_n}{\mu_n} = \frac{k_B T}{q} g\left(\frac{n}{N_c}\right), \qquad \frac{D_p}{\mu_p} = \frac{k_B T}{q} g\left(\frac{p}{N_v}\right) \tag{2}$$

with a density dependent nonlinear factor (*diffusion enhancement factor*)

$$g(\xi) = \xi \left(\mathscr{F}^{-1}\right)'(\xi),$$

leading in general to a nonlinear diffusion coefficient. For the Boltzmann approximation, we see that $g(\xi) = 1$ and for the Blakemore distribution function $g(\xi) \geq 1$. The system (1) is supplied with mixed Dirichlet-Neumann boundary conditions [5].

3 Finite Volume Discretizations

We discretize the domain Ω using the Voronoï box based finite volume method. One obtains the control volumes surrounding each given collocation point $\mathbf{x}_K$ by joining the circumcenters of the simplices containing it. Let ∂K denote the boundary of the control volume K, and $|\xi|$ the measure of a geometrical object ξ. For each control volume K, we integrate Eq. (1b) and apply Gauss's theorem to the integral of the flux divergence. Restricting our considerations to the electron transport equation, we obtain for a single node K

$$\begin{aligned} 0 &= \int_{\partial K} \mathbf{j}_n \cdot \mathbf{n} + \int_K R \, d\mathbf{x} = \sum_{L \text{ neighbor of } K} \int_{\partial K \cap \partial L} \mathbf{j}_n \cdot \mathbf{n}_{KL} ds + \int_K R \, d\mathbf{x} \\ &\approx \sum_{L \text{ neighbor of } K} \frac{|\partial K \cap \partial L|}{|\mathbf{x}_K - \mathbf{x}_L|} j_{n,KL} + |K| R(n_K, p_K), \end{aligned} \tag{3}$$

where $\mathbf{n}$ is the internal unit normal to ∂K and $\mathbf{n}_{KL}$ is the internal unit normal to the interface $\partial K \cap \partial L$ for each neighbor L of K. The values n_K, p_K are the densities n and p at the collocation points $\mathbf{x}_K$, and $j_{n,KL}$ are approximations of the (scaled by the edge length) normal currents through $\partial K \cap \partial L$. In the same manner the discretization of the Poisson equation can be obtained. More details are in [5].

4 Flux Discretizations

For the Boltzmann approximation the classical Scharfetter-Gummel scheme has been derived by solving a two-point boundary value problem resulting from a projection of the continuity equation (neglecting recombination) onto the discretization edge $\overline{\mathbf{x}_K \mathbf{x}_L}$ assuming constant current $j = \mathbf{j}_n \cdot \mathbf{n}_{KL}$ and constant electrical field with boundary values $n|_{\mathbf{x}_K} = n_K, n|_{\mathbf{x}_L} = n_L$. Its nondimensionalized version is given by

$$j_{\mathrm{SG}} = B\Big(\frac{\psi_L - \psi_K}{U_T}\Big) n_L - B\Big(-\frac{\psi_L - \psi_K}{U_T}\Big) n_K, \tag{4}$$

see [9], where $B(x) = \frac{x}{\exp(x)-1}$ is the Bernoulli function and $U_T = k_B T/q$ the thermal voltage. Furthermore, $\psi|_{\mathbf{x}_K} = \psi_K$ and $\psi|_{\mathbf{x}_L} = \psi_L$ denote the electrostatic potential at node K or L, respectively.

For Blakemore statistics $\mathscr{F}(x) = (\exp(-x) + \gamma)^{-1}$ with $\gamma = 0.27$, it is possible to also obtain an exact (albeit implicitly given) solution to the two-point boundary

value problem [7], leading to the generalized Scharfetter-Gummel flux j_{GENSG}:

$$j_{\mathrm{GENSG}} = B\left(\gamma j_{\mathrm{GENSG}} + \frac{\psi_L - \psi_K}{U_T}\right) e^{\eta_L} - B\left(-\left[\gamma j_{\mathrm{GENSG}} + \frac{\psi_L - \psi_K}{U_T}\right]\right) e^{\eta_K}, \tag{5}$$

where $\eta|_{\mathbf{x}_K} = \eta_K, \eta|_{\mathbf{x}_L} = \eta_L$ for

$$\eta = \eta\left(\psi, \varphi_n\right) = \frac{q\left(\psi - \varphi_n\right) - E_c}{k_B T}. \tag{6}$$

Due to its derivation it is equivalent to the exact solution of the point boundary value problem. Hence, we consider this scheme the best possible. Unfortunately, it is relatively expensive to obtain a flux approximation in the framework of the finite volume discretization using this scheme [4].

We will compare this exact flux discretization with two schemes that approximate the flux by averaging certain quantities and thus are cheaper to implement. Bessemoulin-Chatard derived a finite volume scheme for convection-diffusion problems by averaging the nonlinear diffusion term appropriately [1]. This idea was cast in physical framework in [8], introducing a logarithmic average of the nonlinear diffusion enhancement

$$g_{KL} = \frac{\eta_L - \eta_K}{\log \mathscr{F}\left(\eta_L\right) - \log \mathscr{F}\left(\eta_K\right)}$$

along the discretization edge. Using the generalized Einstein relation (2), one immediately observes that the diffusion enhancement g can be seen as a modification factor for the thermal voltage U_T. Replacing U_T in the Scharfetter-Gummel expression (4) by $U_T^* = U_T g_{KL}$, we deduce the *diffusion enhanced* flux

$$j_{\mathrm{DESG}} = -g_{KL}\left(\mathscr{F}\left(\eta_K\right) B\left(-\frac{\psi_L - \psi_K}{U_T g_{KL}}\right) - \mathscr{F}\left(\eta_L\right) B\left(\frac{\psi_L - \psi_K}{U_T g_{KL}}\right)\right). \tag{7}$$

Instead of replacing the thermal voltage by a suitable average along the edge, we can approximate $\mathscr{F}(\eta)$ along the edge by an exponential (the *activity* e^{η} corresponding to the Boltzmann approximation) and modify the density of states N_c accordingly by multiplication with the averaged inverse activity coefficient. In [6] it is proposed to use:

$$j_{\mathrm{IACT}} = -\beta_{KL}\left(B\left(-\frac{\psi_L - \psi_K}{U_T}\right) e^{\eta_K} - B\left(\frac{\psi_L - \psi_K}{U_T}\right) e^{\eta_L}\right). \tag{8}$$

For the *inverse activity coefficient* (a degeneracy factor) we choose the arithmetic mean

$$\beta_{KL} = \frac{1}{2}\left(\frac{\mathscr{F}\left(\eta_K\right)}{\exp(\eta_K)} + \frac{\mathscr{F}\left(\eta_L\right)}{\exp(\eta_L)}\right). \tag{9}$$

5 Comparison for an n-i-n Example

We will compare the three current approximations with the help of an n-i-n example. For the simulations we used the ddfermi software package [3]. The setup is shown in Fig. 1. The distribution of densities are plotted in [8]. Due to its structure, we can perform all calculations in 1D. The following GaAs parameters have been used for the following calculations: band gap $E_c - E_v = 1.424\,\text{eV}$, density of states $N_c = 4.7 \times 10^{17}\,\text{cm}^{-3}$, $N_v = 9 \times 10^{18}\,\text{cm}^{-3}$, mobilities $\mu_n = 8500\,\text{cm}^2/(\text{Vs})$, $\mu_p = 400\,\text{cm}^2/(\text{Vs})$, relative permittivity $\varepsilon_r = 12.9$, n-doping $N_d = 1.65 \times 10^{18}\,\text{cm}^{-3}$, temperature $T = 298.15\text{K}$ and no recombination $R = 0$.

On a sequence of finer meshes we compute the L_∞ error of the electron densities for four different biases, shown in Fig. 2. A bias is the difference in applied voltage at

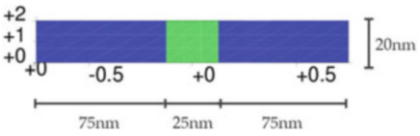


Fig. 1 The device setup: two n-doped regions separated by an intrinsic region

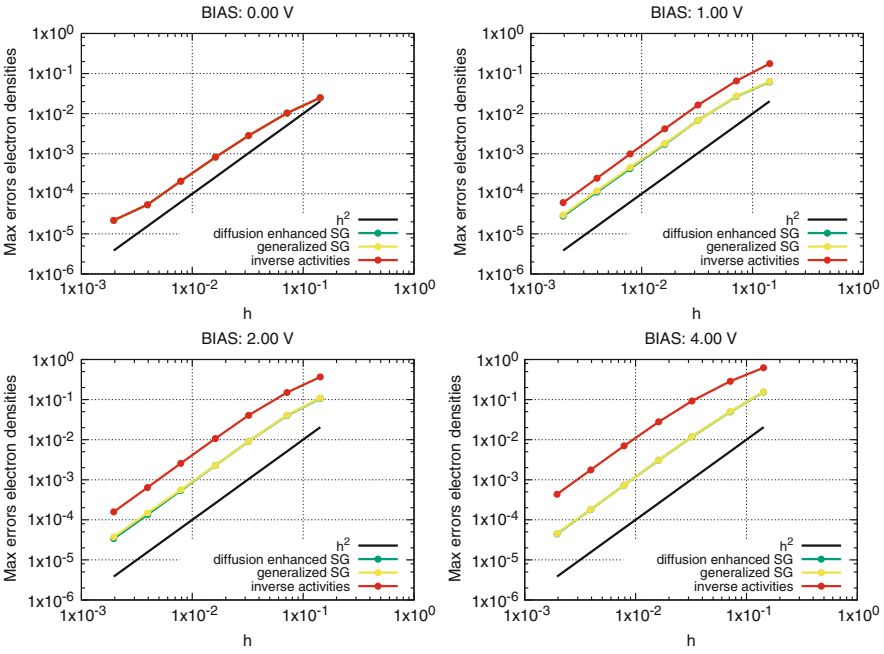


Fig. 2 Convergence analysis for the L_∞ error of the electron densities on a sequence of finer meshes for four different biases (electrostatic potential differences at the boundary). The first seven refinement levels are shown. The reference solution is computed on a very fine grid (refinement level 12) using the generalized Scharfetter-Gummel scheme. For each new refinement level the uniform mesh size is halved. Due to the thermodynamic consistency all schemes agree for zero bias

the boundary of the device. While there is almost no notable difference between the generalized Scharfetter-Gummel and the diffusion enhanced Scharfetter-Gummel scheme. The inverse activity scheme, however, converges more slowly. Since all schemes eventually converge quadratically, its convergence constant must be bigger.

Finally, we compare the total current for a set of different bias values. That is, we study the absolute value of the relative error for the IV curve. The error is shown for the previously introduced schemes and two different refinement levels in Fig. 3. It is computed with respect to the generalized Scharfetter-Gummel solution on a very fine grid (refinement level 12). As before, the generalized and diffusion enhanced Scharfetter Gummel schemes seem to perform similarly. The error for the inverse activity discretization changes sign. Hence it is nonmonotone. Unlike for the other schemes the total current I does not converge for large biases U.

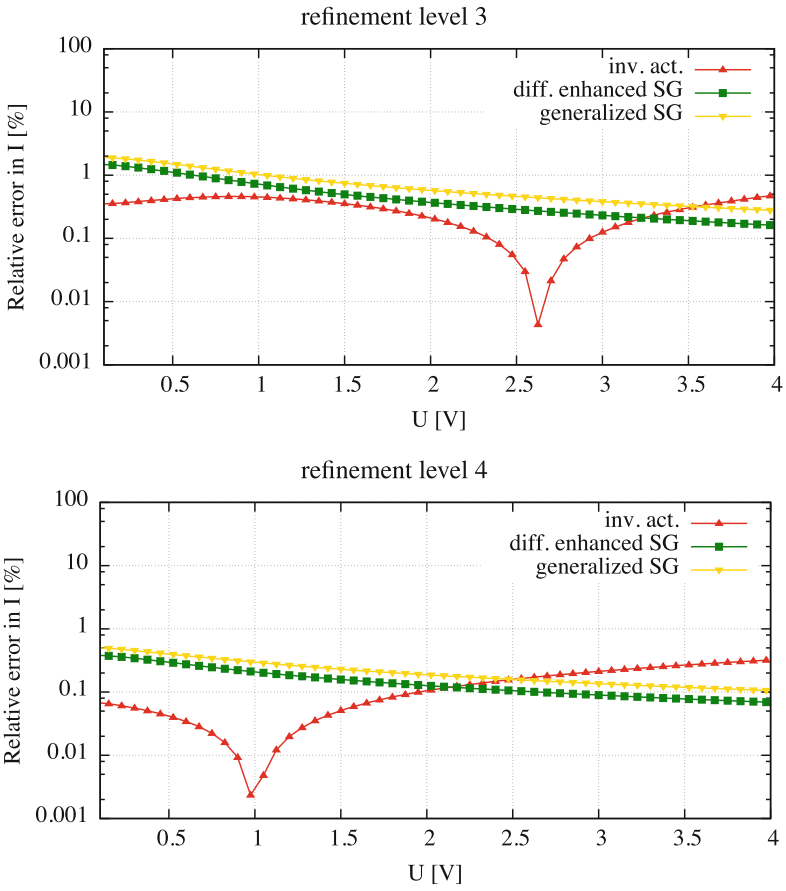


Fig. 3 Relative errors for the total current I depending on the bias U for two different refinement levels

6 Conclusion

We compared three different flux approximation schemes for a finite volume discretization of the van Roosbroeck system assuming Blakemore statistics. All schemes converge quadratically. However, whereas the generalized and diffusion enhanced Scharfetter-Gummel behave similarly, the inverse activity scheme converges more slowly. For the overall current the inverse activity scheme also exhibits a different behaviour. The error is nonmonotone for large voltages and does not converge for large biases.

References

1. Bessemoulin-Chatard, M.: A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme. Numer. Math. **121**, 637–670 (2012)
2. Blakemore, J.S.: The parameters of partially degenerate semiconductors. Proc. Phys. Soc. Lond. A **65**, 460–461 (1952)
3. Doan, D.H., Farrell, P., Fuhrmann, J., Kantner, M., Koprucki, T., Rotundo, N.: ddfermi - a drift-diffusion simulation tool. Version: 0.1.0. doi:http://doi.org/10.20347/WIAS.SOFTWARE.14
4. Eymard, R., Fuhrmann J., Gärtner K.: A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local dirichlet problems. Numer. Math. **102**, 463–495 (2006)
5. Farrell, P., Rotundo, N., Doan, D.H., Kantner, M., Fuhrmann, J., Koprucki, T.: Electronics: numerical methods for drift-diffusion models. In: Pipreck, J. (ed.) Handbook of Optoelectronic Device Modeling and Simulation. Taylor & Francis, UK (to appear 2017)
6. Fuhrmann, J.: Comparison and numerical treatment of generalised Nernst-Planck models. Comput. Phys. Commun. **196**, 166–178 (2015)
7. Koprucki, T., Gärtner, K.: Discretization scheme for drift-diffusion equations with strong diffusion enhancement. Opt. Quant. Electron. **45**, 791–796 (2013)
8. Koprucki, T., Rotundo, N., Farrell, P., Doan, D.H., Fuhrmann, J.: On thermodynamic consistency of a Scharfetter-Gummel scheme based on a modified thermal voltage for drift-diffusion equations with diffusion enhancement. Opt. Quant. Electron. **47**, 1327–1332 (2015)
9. Scharfetter, D.L., Gummel, H.K.: Large signal analysis of a silicon Read diode. IEEE Trans. Electron. Dev. **16**, 64–77 (1969)

A PDE Model for Electrothermal Feedback in Organic Semiconductor Devices

Matthias Liero, Axel Fischer, Jürgen Fuhrmann, Thomas Koprucki, and Annegret Glitzky

Abstract Large-area organic light-emitting diodes are thin-film multilayer devices that show pronounced self-heating and brightness inhomogeneities at high currents. As these high currents are typical for lighting applications, a deeper understanding of the mechanisms causing these inhomogeneities is necessary. We discuss the modeling of the interplay between current flow, self-heating, and heat transfer in such devices using a system of partial differential equations of thermistor type, that is capable of explaining the development of luminance inhomogeneities. The system is based on the heat equation for the temperature coupled to a $p(x)$-Laplace-type equation for the electrostatic potential with mixed boundary conditions. The $p(x)$-Laplacian allows to take into account non-Ohmic electrical behavior of the different organic layers. Moreover, we present analytical results on the existence, boundedness, and regularity of solutions to the system. A numerical scheme based on the finite-volume method allows for efficient simulations of device structures.

MSC 2010: 35J92, 65M08, 35J57, 35Q79, 80M12

1 Introduction

Nowadays organic (i.e. carbon-based) semiconductors are extensively used in smartphone displays and increasingly in TV screens. Lighting applications are of great

M. Liero (✉) • J. Fuhrmann • T. Koprucki • A. Glitzky
Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany
e-mail: liero@wias-berlin.de; fuhrmann@wias-berlin.de; koprucki@wias-berlin.de; glitzky@wias-berlin.de

A. Fischer
Dresden Integrated Center for Applied Physics and Photonic Materials (IAPP), Technische Universität Dresden, 01062 Dresden, Germany
e-mail: axel.fischer@iapp.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_14

interest due to the fascinating properties of organic light-emitting diodes (OLEDs), e.g. large-area surface emission, semi-transparency, flexibility. However, lighting applications require a much higher brightness than displays and hence higher currents are necessary. These cause substantial Joule self-heating accompanied by unpleasant brightness inhomogeneities of the panels. An appropriate simulation tool can help to validate cost-efficient device concepts by accounting for nonlinear self-heating effects.

The appearance of the inhomogeneities, which show a saturation of the luminance in the panel center, can neither be explained with high sheet resistances of the optically transparent electrode nor with a degradation of the organic materials due to high temperatures. In [5, 6] it was demonstrated that the complex interplay of temperature activated transport of the charge carriers and heat flow in the device lead to inhomogeneous current distributions resulting in inhomogeneous luminance.

Indeed, applying a voltage to an organic semiconductor device induces a current flow which leads to a power dissipation by Joule heating and hence also a temperature rise. As higher temperatures improve the conductivity in organic materials, higher currents occur. Thus, a positive feedback loop develops that can result in a complete destruction of the device by thermal runaway if the generated heat cannot be dispersed into the environment. The temperature dependence of the conductivity is often [5] modeled by an exponential law of Arrhenius type, which features an activation energy that is linked to the energetic disorder in the organic material.

Devices that show such an electrothermal interplay are called thermistors with negative temperature coefficient. The current-voltage characteristics of such thermistors are S-shaped with regions of negative differential resistance (S-NDR), where currents increase despite of decreasing voltages, see Fig. 1 (left). A thermistor-like behavior of organic semiconductors induced by self-heating has been verified for the organic material C_{60} in [5] and for organic materials used in OLEDs in [6].

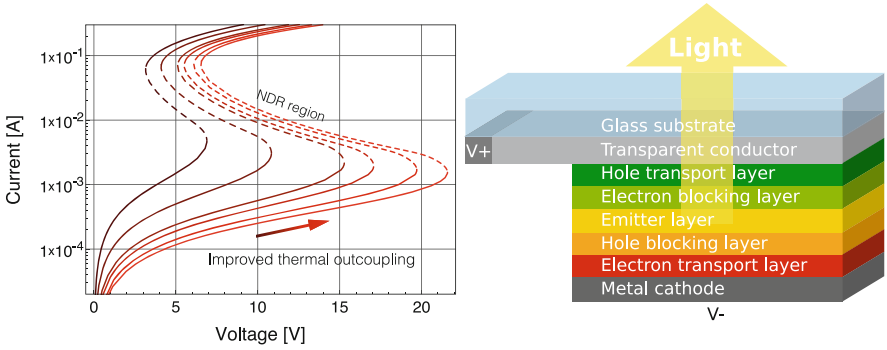


Fig. 1 *Left*: Simulated current-voltage characteristics for different thermal outcoupling regimes, regions of negative differential resistance are dashed. *Right*: Schematic picture of OLED stack

However, due to the huge aspect ratios of OLED panels, such devices cannot be regarded as a single spatially homogeneous thermistor device, but rather as an array of thermistor devices that can exchange heat. In particular, the self-heating, and hence also the local differential resistance, is now a collective property of neighboring thermistors.

In [6] we demonstrated that a new operation mode appears that is not present in the spatially homogeneous case: Eventually, the S-NDR regime has propagated through the whole panel. Now, regions in the panel do not warm up enough to remain in an S-NDR mode so that both, local voltages and currents, decrease although the externally applied voltage is still increasing. This "switched-back" regime is ultimately the reason for the observed saturation of luminance in the panel center, in fact, one even has to expect a decline of luminance as the local currents are decreasing.

In the following we present a mathematical model for the current and heat flow in organic semiconductor devices consisting of a coupled PDE system for the electrostatic potential and the temperature, see Sect. 2. This PDE modeling approach gives much more flexibility concerning variations in geometry and material composition than for example network models. Moreover, our model contains also a PDE-type description of the active organic zone whereas in other PDE simulation approaches the organic layer is reduced to the information from empirical characteristics. The crucial feature of our PDE model is that the current flow equation is of $p(x)$-Laplace type, where the exponent $p(x)$ takes the non-Ohmic behavior of the organic layers into account. In particular, the exponent is in general discontinuous as the different functional layers exhibit different power laws. In Sect. 3, we summarize the analytical results of [2, 3, 7] and present in Sect. 4 a numerical scheme that allows us to efficiently simulate the current and heat flow in OLEDs. One of the major challenges is the derivation of a stable scheme for the $p(x)$-Laplacian, which we address by using a hybrid finite-volume/finite-element approach.

2 Modeling of Current and Heat Flow

A typical OLED device structure is depicted in Fig. 1 (right). To describe the interplay of current and heat flow in organic semiconductor devices the following PDE model was proposed in [8]. It consists of the current flow equation for the electrostatic potential φ and the heat equation for the temperature T

$$\begin{aligned} -\nabla \cdot S(x, T, \nabla\varphi) &= 0, \\ -\nabla \cdot \big(\lambda(x)\nabla T\big) &= H(x, T, \nabla\varphi) \end{aligned} \qquad \text{on } \Omega \subset \mathbb{R}^d, \tag{1}$$

where S gives the electrical flux, λ is the heat conductivity, and H is the Joule heat term. The system is complemented by Dirichlet and no-flux boundary conditions for the potential φ at the contacts Γ_D and the insulating parts Γ_N of the boundary, respectively, and Newton boundary conditions for the temperature to describe the

coupling to the environment

$$\begin{aligned} \varphi = \varphi^D \text{ on } \Gamma_D, \qquad S(x,T,\nabla\varphi)\cdot\nu = 0 \text{ on } \Gamma_N, \\ -\lambda(x)\nabla T\cdot\nu = \gamma(x)(T-T_\mathrm{a}) \text{ on } \Gamma = \partial\Omega, \end{aligned} \tag{2}$$

where $T_\mathrm{a} > 0$ denotes the fixed ambient temperature and γ is the heat transfer coefficient. The special features of the model are the Arrhenius-like temperature law as well as the non-Ohmic current-voltage relations incorporated by a power law in the current expression S, namely

$$S(x,T,\nabla\varphi) = \kappa_0(x)F(x,T)|\nabla\varphi|^{p(x)-2}\nabla\varphi, \quad F(x,T) = \exp\Big[-\frac{E_\mathrm{act}(x)}{k_\mathrm{B}}\Big(\frac{1}{T}-\frac{1}{T_\mathrm{a}}\Big)\Big].$$

Here, κ_0 and E_act are material dependent effective conductivity and an activation energy, respectively, and k_B denotes Boltzmann's constant. The exponent p of the power law is also material dependent, taking different power laws in the various layers of the device into account. Hence, the first equation in (1) becomes of $p(x)$-Laplacian type with discontinuous measurable exponent p. In particular, we have $p(x) \equiv 2$ in Ohmic materials such as electrodes and $p(x) > 2$ (e.g. $p(x) = 9.7$, see [6]) in organic layers. The Joule heat term in the second equation of (1) takes the form

$$H(x,T,\nabla\varphi) = \big(1-\eta(x,T,\nabla\varphi)\big)\kappa_0(x)F(x,T)|\nabla\varphi|^{p(x)},$$

where $\eta(x,T,\nabla\varphi) \in [0,1]$ represents the light-outcoupling factor. Note that H is a priori only an L^1 function which hampers the analytical treatment of the system.

3 Analytical Results

The first and foremost question is, whether the system in (1), (2) is well-defined, i.e. existence of (weak) solutions. Due to the definition of the flux function S with spatially dependent exponent p it is clear, that we cannot work in the classical Lebesgue and Sobolev spaces. Following [4], we introduce for measurable exponents p with

$$1 < p_- := \operatorname{ess\,inf}_{x\in\Omega} p(x) \le \operatorname{ess\,sup}_{x\in\Omega} p(x) =: p_+ < \infty \tag{3}$$

the generalized Lebesgue space $L^{p(\cdot)}(\Omega)$ as the set of all measurable functions f with finite modular $\rho_{p(\cdot)}(f) := \int_\Omega |f(x)|^{p(x)}\,\mathrm{d}x < \infty$, which equipped with the corresponding Luxemburg norm is a reflexive and separable Banach space. Moreover, we define the generalized Sobolev space $W^{1,p(\cdot)}(\Omega) := \{f \in W^{1,p_-}(\Omega) : \nabla f \in L^{p(\cdot)}(\Omega)^d\}$ being also a separable and reflexive Banach space with the

norm $\|f\|_{W^{1,p(\cdot)}} := \|f\|_{W^{1,p-}} + \|\nabla f\|_{L^{p(\cdot)}}$. By $W_D^{1,p(\cdot)}(\Omega) \subset W^{1,p(\cdot)}(\Omega)$ we denote the subspace of functions with Dirichlet value 0 at Γ_D. The following theorem guarantees the existence of solutions, we refer to [2, 3] for details.

Theorem 1 *Besides an L^∞ setting for the material parameters we suppose (3), $T_\mathrm{a} > 0$, and $d \geq 2$. Then the $p(x)$-Laplace thermistor problem (1), (2) has a solution (φ, T) with $\varphi \in \varphi^D + W_D^{1,p(\cdot)}(\Omega)$ and $T \in W^{1,q}(\Omega)$ for all $q \in [1, \frac{d}{d-1})$. Especially, T is an entropy solution to the heat equation and $T \geq T_\mathrm{a}$ a.e. in Ω.*

There are two strategies to prove the existence result without improved regularity of the Joule heating term H. In [2] a more constructive way is proposed, where an truncation H_ε of the Joule heat term is applied such that it remains bounded. The regularized problems are solved by Galerkin approximation. By verifying a priori estimates not depending on the truncation parameter ε and a suitable passage to the limit the existence result for (1), (2) is obtained. The second method is based on Schauder's fixed point theorem using the set $\mathscr{M} := \{T \in L^1(\Omega) : \|T\|_{W^{1,1}} \leq \tilde{c},\ T \geq T_\mathrm{a}\}$ and the concept of entropy solutions for the heat equation with Newton boundary conditions, see [3].

An alternative approach to verify the existence of solutions to (1), (2) is to improve the regularity of the Joule heating term H such that the concept of weak solutions for the $p(x)$-Laplace thermistor problem is applicable. In [7], we proved higher regularity of the gradient of the electrostatic potential in two spatial dimensions and for $p(x) \geq 2$ piecewise constant, and such that the localized geometric situation is covered by one of the eight cases in Fig. 2. The proof is based on a localization argument, Caccioppoli estimates, and a Gehring-type lemma. With this, we obtain $\nabla\varphi \in L^{sp(\cdot)}(\Omega)$ and $H(\cdot, T, \nabla\varphi) \in L^s(\Omega)$ for some $s > 1$. Defining a fixed point map for the temperature on the set $\mathscr{M} = \{T \in W^{1,q}(\Omega) : \|T\|_{W^{1,q}} \leq c_q,\ T \geq T_\mathrm{a}\}$ for some $q > 2$ Schauder's fixed point theorem leads to the following improved existence result.

Theorem 2 *Let additionally $d = 2$, $\varphi^D \in W^{1,\infty}(\Omega)$ and $p(x) \geq 2$ be piecewise constant. We assume that the geometric structure can be locally transferred to the situation of one of the eight cases in Fig. 2, where $p_B > p_A$. Then the problem (1), (2) has a solution (φ, T) with $\varphi \in \varphi^D + W_D^{1,sp(\cdot)}(\Omega)$ and $T \in W^{1,q}(\Omega)$ for some $s > 1$ and $q > 2$. Especially, T is continuous and fulfils $T_\mathrm{a} \leq T \leq c_T$.*

Note that Schauder's theorem does not give uniqueness of fixed points and hence of solutions. However, uniqueness of solutions for the $p(x)$-Laplace thermistor system (1), (2) cannot be expected due to the hysteretic behavior induced by the positive feedback with respect to temperature described in the introduction. Even

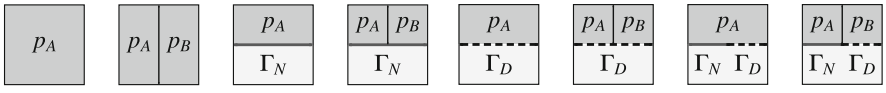


Fig. 2 Model sets with different exponents $p_A < p_B$ and different types of boundary conditions

in the spatially homogeneous setting of self-heating in organic devices (see [5] and [8, Sect. 2.1]) S-shaped current-voltage characteristics occur. Here, three different currents are possible for the same applied voltage.

4 Numerical Scheme

As we have to deal with piecewise constant functions $p(x)$, we subdivide the computational domain $\bar{\Omega} = \bigcup_{r\in\mathscr{R}} \bar{\Omega}_r$ into disjoint subdomains coinciding with the regions of continuity of the coefficients. We call the boundary between two neighboring regions *hetero interface*. To solve the PDE system (1), (2) numerically we use a finite-volume method on Voronoi cells as control volumes which are constructed based on a Delaunay grid. The latter is supposed to be boundary conforming with respect to exterior boundaries and hetero interfaces, see e.g. [9]. We assume that each control volume contains a collocation point $x_K \in \overline{\Omega}$.

We apply Gauss's theorem to the integral of the flux divergence to obtain for the current flow equation in (1) the flux balance with further subdivision into contributions from adjacent subdomains:

$$0 = \int_K \nabla \cdot S(x,T,\nabla\varphi)\,\mathrm{d}x = \sum_{r\in\mathscr{R}} \sum_{L\sim K} \int_{\partial K\cap\partial L\cap\Omega_r} \kappa_{0,r} F_r(T) |\nabla\varphi|^{p_r-2} \nabla\varphi \cdot \nu_{KL}\,\mathrm{da}, \tag{4}$$

where $L \sim K$ indicates that L is adjacent to K and ν_{KL} is the unit normal vector pointing from K into L. Note that the normal flux over a surface $\partial K \cap \partial L \cap \Omega_r$ does not only depend on the normal components of $\nabla\varphi$ but on the modulus of the full gradient. To take this into account we compute the approximation of $|\nabla\varphi|^2$ on $\partial K \cap \partial L \cap \Omega_r$ as the average squared norms of the P1 finite element gradients $\nabla_\tau\varphi$ over the set $\mathscr{T}_{K,L,r}$ of all simplices (triangles in 2D) in the underlying Delaunay triangulation adjacent to the edge $\overline{x_K x_L}$ and belonging to Ω_r :

$$\left|\nabla\varphi\right|^2|_{\partial K\cap\partial L\cap\Omega_r} \approx G^2_{K,L,r}(\varphi) := \frac{\sum_{\tau\in\mathscr{T}_{K,L,r}} |\tau||\nabla_\tau\varphi|^2}{\sum_{\tau\in\mathscr{T}_{K,L,r}} |\tau|}.$$

This approach allows us to introduce an approximation of the right-hand side of (4) consisting in replacing the surface integral by a simple quadrature, and the gradient projection by a finite difference expression leading for $T_{KL} = (T_K+T_L)/2$ to

$$0 = \sum_{r\in\mathscr{R}} \sum_{L\sim K} \frac{|\partial K\cap\partial L\cap\Omega_r|}{|x_K - x_L|} \kappa_{0,r} F_r(T_{KL}) G_{K,L,r}(\varphi)^{p_r-2}(\varphi_L-\varphi_K).$$

The same approach of calculating the conductivity in the Joule heat term is combined with the approach suggested in [1] which allows us to calculate the Joule

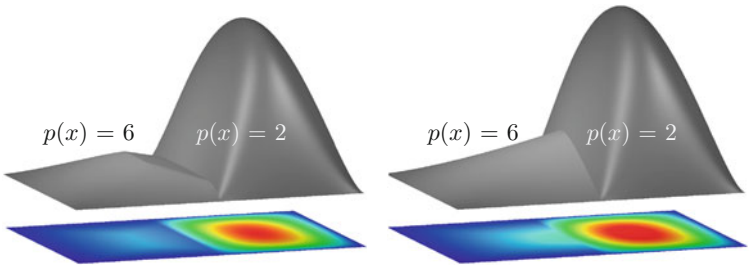


Fig. 3 Discrete solutions of $p(x)$-Laplace equation with constant right-hand side, Dirichlet boundary conditions, and piecewise constant $p(x)$. *Left*: Wrong approximation with two local maxima due to gradient averaging ignoring the hetero interface. *Right*: Correct approximation

heating approximation by edge contributions: Applying Gauss's theorem leads to

$$\begin{aligned} 0 &= \int_K \big(\nabla\cdot(\lambda(x)\nabla T) + H(x,T,\nabla\varphi)\big)\,\mathrm{d}x \\ &= \sum_{L\sim K}\int_{\partial K\cap\partial L} \lambda(x)\nabla T\cdot \nu_{KL}\,\mathrm{da} + \int_K \big(1-\eta(x)\big)\kappa_0(x)F(x,T)|\nabla\varphi|^{p(x)}\,\mathrm{d}x, \end{aligned} \tag{5}$$

and the suggested approach yields the approximation of the heat flow equation on K

$$\begin{aligned} 0 = \sum_{r\in\mathscr{R}}\sum_{L\sim K}\Big(&\frac{|\partial K\cap\partial L\cap\Omega_r|}{|x_K-x_L|}\lambda_r(T_L-T_K) \\ &+\frac{1}{2}\frac{|\partial K\cap\partial L\cap\Omega_r|}{|x_K-x_L|}(1-\eta_r)\kappa_{0,r}F_r(T_{KL})G_{K,L,r}(\varphi)^{p_r-2}(\varphi_L-\varphi_K)^2\Big). \end{aligned}$$

It is important to take care of the hetero interface when calculating the average of the gradient norm. An averaging over all simplices adjacent to a given edge regardless of the hetero region they belong to leads to an artificial diffusion along the hetero interface (see Fig. 3) which does not diminish with grid refinement. The validity of this approach and the way it has been implemented hinges on the fact that all the measures $|\partial K\cap\partial L\cap\Omega_r|$ can be calculated from contributions from each simplex which at the hetero interfaces have to stay nonnegative. It is guaranteed by the boundary conforming Delaunay property of the underlying triangulation.

Acknowledgements A.G. and M.L. gratefully acknowledge the funding received via Research Center MATHEON supported by ECMath in project D-SE2.

References

1. Bradji, A., Herbin, R.: IMA J. Numer. Anal. **28**(3), 469 (2008). doi:10.1093/imanum/drm030
2. Bulíček, M., Glitzky, A., Liero, M.: SIAM J. Math. Anal. **48**, 3496 (2016). doi:10.1137/16M1062211
3. Bulíček, M., Glitzky, A., Liero, M.: Thermistor systems of $p(x)$-laplace-type with discontinuous exponents via entropy solutions. Discrete Contin. Dyn. Syst. S **10**(4), 697–713 (2017). doi:10.3934/dcdss.2017035
4. Diening, L., Harjulehto, P., Hästö, P., Ružička, M.: Lebesgue and Sobolev Spaces with Variable Exponents. Lecture Notes in Mathematics, vol. 2017. Springer, Berlin (2011). doi:10.1007/978-3-642-18363-8
5. Fischer, A., Pahner, P., Lüssem, B., Leo, K., Scholz, R., Koprucki, T., Gärtner, K., Glitzky, A.: Phys. Rev. Lett. **110**, 126601/1 (2013). doi:10.1103/PhysRevLett.110.126601
6. Fischer, A., Koprucki, T., Gärtner, K., Brückner, J., Lüssem, B., Leo, K., Glitzky, A., Scholz, R.: Adv. Funct. Mater. **24**, 3367 (2014). doi:10.1002/adfm.201303066
7. Glitzky, A., Liero, M.: Nonlinear Anal. Real World Appl. **34**, 536 (2017). doi:10.1016/j.nonrwa.2016.09.015
8. Liero, M., Koprucki, T., Fischer, A., Scholz, R., Glitzky, A.: Z. Angew. Math. Phys. **66**(6), 2957 (2015). doi:10.1007/s00033-015-0560-8
9. Si, H., Gärtner, K., Fuhrmann, J.: Comput. Math. Math. Phys. **50**(1), 38 (2010). doi:10.1134/S0965542510010069

Minisymposium: Computational Electromagnetism

Ana Alonso Rodríguez and Ruben Specogna

Description

Electromagnetic phenomena is one of the fundations of modern technology and developing models, approximation methods and software tools for electromagnetics computations have a direct impact and great relevance in industry. In the past thirty years computational electromagnetism has became also a prolific and strengthened research field on applied mathematics. The aim of this minysimposium was to bring together a small group of engineers and mathematicians working on common interests to present advanced topics on modeling and discretization of electromagnetic fields and to discuss the effective impact on industrial applications and software design of these advanced techniques.

The talks included in the minisymposium were the following:

- Stefan Kurz. Graduate School of Excellence Computational Engineering, Darmstadt (Germany), and Tampere University of Technology, DEE-Electromagnetics, Tampere (Finland). *Structure Preserving Mesh Coupling for Maxwell's Equations*.
- Holger Heumann. INRIA, EPI CASTOR, Sophia Antipolis (France). *Overlapping mortar methods for axisymmetric plasma equilibria in tokamaks*.
- Andrei Kolyshkin. Riga Technical University, Riga (Latvia). *Eddy current testing models for the analysis of corrosion effects in metal plates*.

A. Alonso Rodríguez (✉)
Department of Mathematics, University of Trento, Trento, Italy
e-mail: ana.alonso@unitn.it

R. Specogna
Polytechnic Department of Engineering and Architecture, University of Udine, Udine, Italy
e-mail: ruben.specogna@uniud.it

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_15

- Pilar Salgado. Department of Applied Mathematics, Universidade de Santiago de Compostela (Spain). *Numerical simulation and optimization of induction heating in forge applications*.
- Maryam Khaksar Ghalati. CMUC, Department of Mathematics, University of Coimbra (Portugal). *Stability and convergence of fully explicit leap-frog DG scheme for solving 2D Maxwell's equations in anisotropic materials*.
- David González-Peñas. Departamento de Matemática Aplicada e Instituto Tecnológico de Matemática Industrial (ITMATI). Universidade de Santiago de Compostela (Spain). *Thermo-electromagneto-hydrodynamic numerical simulation of the removal of volatile impurities in molten metals*.
- Alain Bossavit. GeePs (ex. LGEP), Centrale-Supelec, Gif-sur-Yvette (France). *Topics in Magnetic Force Theory: Some avatars of the Helmholtz formula*.

FEMs on Composite Meshes for Plasma Equilibrium Simulations in Tokamaks

Holger Heumann, Francesca Rapetti, and Minh Duy Truong

Abstract We rely on a combination of different finite element methods on composite meshes, for the simulation of axisymmetric plasma equilibria in tokamaks. One mesh with Cartesian quadrilaterals covers the vacuum chamber and one mesh with triangles discretizes the region outside the chamber. The two meshes overlap in a narrow region around the chamber. This approach gives the flexibility to achieve easily and at low cost higher order regularity for the approximation of the flux function in the area that is covered by the plasma, while preserving accurate meshing of the geometric details in the exterior. The continuity of the numerical solution across the boundary of each subdomain is enforced by a new mortar-like projection.

1 Introduction

The possibility of using composite meshes in finite element (FE) simulations of industrial problems is a recurrent topic [4, 7, 12–14]. Composite meshes are involved as soon as the global discretization of a PDE combines discretizations on local (overlapping or non-overlapping) subdomains, each suitably triangulated by non-matching grids. The reason for using composite meshes are various: fitting the geometry or the local smoothness of the solution, resolving multiple scales in regions with irregular data, using fast solvers on structured grids or a divide-and-conquer/domain decomposition approach to very large problems on parallel machines.

In the present case, we are looking for a simple and practical approach to introduce in certain parts of the computational domain FE functions that are not only continuous, but have also first order, second order or higher order continuous derivatives. In general it is very difficult to introduce FE spaces over simplicial

H. Heumann • M.D. Truong
INRIA Sophia Antipolis Méditerranée, CASTOR 2004 Route des Lucioles BP 93, 06902 Sophia Antipolis, France
e-mail: Holger.Heumann@inria.fr; truongminhduy91@yahoo.com

F. Rapetti (✉)
Department of Mathematics, Univ. de Nice, Parc Valrose, 06108 Nice cedex 02, France
e-mail: Francesca.Rapetti@unice.fr

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_16

unstructured meshes with such properties. On the other hand, if we work with Cartesian meshes this becomes very simple. It is sufficient to use tensor products of spline spaces with sufficiently high regularity. So, as it is naive to expect that technical devices can be entirely triangulated with Cartesian meshes, we introduce composite meshes involving Cartesian meshes in those subdomains where we want high regular FE representations and triangular unstructured meshes in those subdomains where we want conformity with the geometry.

The industrial application we consider concerns the free boundary plasma equilibrium in tokamaks for nuclear fusion [2], mathematically described by the force balance and Maxwell's equations in the eddy-current approximation. By symmetry considerations, the free boundary plasma equilibrium problem can be reduced to a scalar semi-linear elliptic one for the poloidal flux. As the magnetic field and the current density are tangential to the level sets of the poloidal flux, the precise calculation of the level set distribution for the poloidal flux is fundamental in tokamak science. Hence, it is important to have good approximations not only of the poloidal flux but also of its derivatives.

In Fig. 1 we show a sketch of the cross section of a tokamak. It contains the geometrical details such as coils, passive structures and the iron core, that need to be accurately resolved by a triangulation (Fig. 2). A peculiarity of the free boundary plasma equilibrium problem is the unknown plasma domain, that is implicitly given as the domain that is bounded by the largest level set that it not intersected by the limiter. At the beginning, in tokamak devices the plasma was always attached to the limiter, later the focus shifted towards devices with so called diverters that allow to create plasmas that are completely detached from material. The boundary of such a plasma is characterized by saddle points, also called X-points, of the poloidal flux. The magnetic axis, located at the maximum of the poloidal flux, is another important critical point to be computed. So, here again, continuous higher derivatives of approximations of the poloidal flux, are very desirable. Hence, it is beneficial to combine triangular and Cartesian meshes for such equilibrium problems (Fig. 2).

In [9], we showed that the numerical calculation of free boundary plasma equilibria highly benefits from approximating the poloidal flux through some higher regular FE functions in the interior of the limiter. In the present paper, we rather analyse the precision of the proposed approach, by varying the discretization parameters. We thus compute the approximation error between the computed and

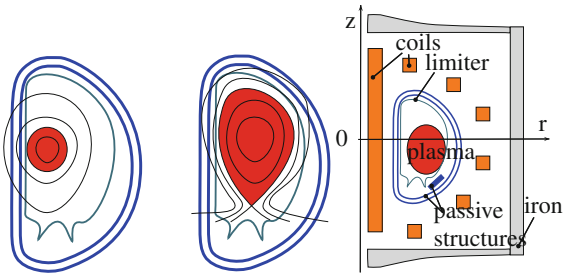


Fig. 1 *Right*: Geometric description of the tokamak in the poloidal plane. *Left and middle*: Sketch for characteristic plasma shapes. The plasma boundary touches the limiter (*middle*) or the plasma is enclosed by a flux line that goes through an X-point (*right*)

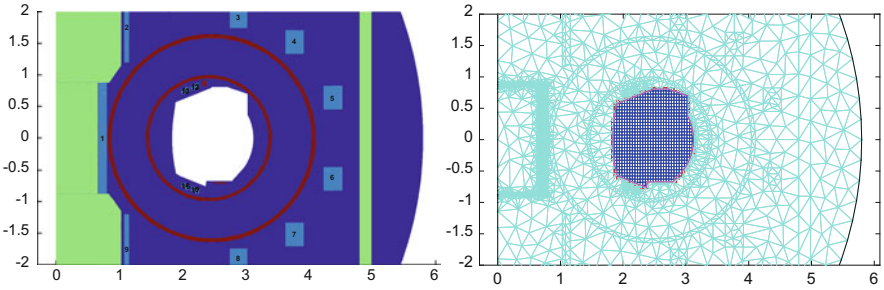


Fig. 2 *Left*: detail view for the WEST tokamak, with the iron core (*green*), the passive structures (*red*) and the various coils (*light blue*). *Right*: the composite meshes for the WEST tokamak

the synthetic solution of a model problem for the same method adopted in [9], by varying, for example, the local polynomial degree in the subdomains, the size of the overlap between the meshes, the local size of the mesh elements. Indeed, FE methods on composite meshes are widely used in practice, but their theoretical foundation is fairly limited in the literature. Therefore, we report here experimental convergence results for different discretization schemes involving composite meshes.

The outline is the following: In Sect. 2 we recall the classical mortar element method (MEM) for overlapping meshes in order to introduce, in the following Sect. 3, a modified method (MEM-M) that simplifies the implementation by avoiding integrals over cut elements. We then present experimental convergence results.

2 The Mortar Element Method (MEM) with Overlapping Meshes

We consider the following Poisson problem for the unknown ψ in the bounded domain $\Omega \subset \mathbb{R}^n$ with boundary $\Gamma = \partial\Omega$:

$$-\nabla \cdot (\nabla \psi) = f \quad \text{in } \Omega \quad \text{and} \quad \psi_{|\partial\Omega} = \psi_0 \quad \text{in } \Gamma\,, \tag{1}$$

where ∇ ($\nabla\cdot$) is the gradient (divergence) operator in $\mathbb{R}^n$. The right-hand side f and the Dirichlet data ψ_0 are given. Let $L^2(\Omega)$, be the functional space of measurable functions on Ω that are square integrable in Ω and $H^1(\Omega) = \{u \in L^2(\Omega),\ \nabla u \in L^2(\Omega)^2\}$ the Hilbert space endowed with the norm $\|u\|^2_{H^1(\Omega)} = \|u\|^2_\Omega + |u|^2_{H^1(\Omega)}$ where $|u|^2_{H^1(\Omega)} = \|\nabla u\|^2_\Omega$. Let $\Omega^{\text{in}} \subset \Omega$ be a subdomain with $\Omega^{\text{in}} \cap \Gamma = \emptyset$ and $\Omega^{\text{ex}} = \Omega \setminus \Omega^{\text{in}}$ the complement of Ω^{in} in Ω. Further, the boundary of Ω^{in}, $\gamma := \partial\Omega^{\text{in}}$, is the interface between Ω^{ex} and Ω^{in}. To formulate (1) as a variational problem in a domain decomposition framework, let us introduce the functional

space

$$\mathscr{H}_g = \{(v, w) \in H^1(\Omega^{\text{ex}}) \times H^1(\Omega^{\text{in}}),\ v_{|\Gamma} = g,\ v_{|\gamma} = w_{|\gamma}\}.$$

Then, the weak formulation of (1) is: Find $(\psi^{\text{ex}}, \psi^{\text{in}}) \in \mathscr{H}_{\psi_0}$ s.t. for all $(v, w) \in \mathscr{H}_0$

$$\int_{\Omega^{\text{ex}}} \nabla\psi^{\text{ex}} \cdot \nabla v \,\mathrm{d}\mathbf{x} + \int_{\Omega^{\text{in}}} \nabla\psi^{\text{in}} \cdot \nabla w \,\mathrm{d}\mathbf{x} = \int_{\Omega^{\text{ex}}} f\, v \,\mathrm{d}\mathbf{x} + \int_{\Omega^{\text{in}}} f\, w \,\mathrm{d}\mathbf{x}\,. \tag{2}$$

We wish to introduce different types of meshes $\mathscr{T}^{\text{ex}}$ and $\mathscr{T}^{\text{in}}$ in the two subdomains Ω^{ex} and Ω^{in}. To achieve a maximum of flexibility we do not expect the meshes $\mathscr{T}^{\text{ex}}$ and $\mathscr{T}^{\text{in}}$ to be conforming with Ω^{ex} and Ω^{in}. More precisely, we denote by Ω_h^{ex} and Ω_h^{in} the domains covered by the mesh elements of $\mathscr{T}^{\text{ex}}$ and $\mathscr{T}^{\text{in}}$, respectively, and we only require that $\Omega^{\text{ex}} \subset \Omega_h^{\text{ex}} \subset \Omega$, $\Gamma \subset \partial\Omega^{\text{ex}}$ and $\Omega^{\text{in}} \subset \Omega_h^{\text{in}} \subset \Omega$. Hence the approximation of (2) enters into the framework of overlapping domain decomposition methods. Let $\gamma^{\text{ex}} = \partial\Omega_h^{\text{ex}} \setminus \Gamma$ and $\gamma^{\text{in}} = \partial\Omega_h^{\text{in}}$ be the two boundaries of Ω_h^{ex} and Ω_h^{in} in Ω that replace the interface γ. Then we introduce the space

$$\mathscr{V}_g = \{(v, w) \in \mathscr{V}^{\text{ex}} \times \mathscr{V}^{\text{in}},\ v_{|\Gamma} = \Pi^{\text{Dir}} g,\ v_{|\gamma^{\text{ex}}} = \Pi^{\text{ex}} w,\ w_{|\gamma^{\text{in}}} = \Pi^{\text{in}} v\}\,,$$

where $\mathscr{V}^{\text{ex}}$ and $\mathscr{V}^{\text{in}}$ are $H^1(\Omega_h^{\text{ex}})$ and $H^1(\Omega_h^{\text{in}})$ conforming FE spaces defined over $\mathscr{T}^{\text{ex}}$ and $\mathscr{T}^{\text{in}}$. The operators Π^{Dir}, Π^{ex} and Π^{in} are projections onto the Dirichlet trace spaces $\mathscr{V}_\Gamma = \text{tr}_{|\Gamma}\, \mathscr{V}^{\text{ex}}$, $\mathscr{V}_\gamma^{\text{ex}} := \text{tr}_{|\gamma^{\text{ex}}}\, \mathscr{V}^{\text{ex}}$ and $\mathscr{V}_\gamma^{\text{in}} := \text{tr}_{|\gamma^{\text{in}}}\, \mathscr{V}^{\text{in}}$. The MEM with overlapping domains [1, 3, 10] applied to (2) reads: Find $(\psi^{\text{ex}}, \psi^{\text{in}}) \in \mathscr{V}_{\psi_0}$ such that

$$\mathsf{a}_s^{\text{ex}}(\psi^{\text{ex}}, v) + \mathsf{a}_t^{\text{in}}(\psi^{\text{in}}, w) = \ell_s^{\text{ex}}(f, v) + \ell_t^{\text{in}}(f, w) \quad \forall (v, w) \in \mathscr{V}_0\,, \tag{3}$$

where

$$\mathsf{a}_s^{\text{M}}(\psi, v) := \int_{\Omega_h^{\text{M}}} \nabla\psi \cdot \nabla v \,\mathrm{d}\mathbf{x} - \int_{\Omega_h^{\text{ex}} \cap \Omega_h^{\text{in}}} s\, \nabla\psi \cdot \nabla v \,\mathrm{d}\mathbf{x}\,,$$

$$\ell_s^{\text{M}}(f, v) := \int_{\Omega_h^{\text{M}}} f\, v \,\mathrm{d}\mathbf{x} - \int_{\Omega_h^{\text{ex}} \cap \Omega_h^{\text{in}}} s f\, v \,\mathrm{d}\mathbf{x}\,,$$

for $\text{M} = \text{ex}$ and $\text{M} = \text{in}$. Optimal convergence results are available when $s + t = 1$ and Π^{ex}, Π^{in}, are the L^2 projections onto $\text{tr}_{|\gamma^{\text{ex}}}\, \mathscr{V}^{\text{ex}}$, $\text{tr}_{|\gamma^{\text{in}}}\, \mathscr{V}^{\text{in}}$, respectively [1, 3]. However, two very restrictive disadvantages occur with the formulation (3):

1. The assembling of the stiffness matrices associated to $\mathsf{a}_s^{\text{ex}}(\cdot, \cdot)$ and $\mathsf{a}_t^{\text{in}}(\cdot, \cdot)$ involves products of basis functions defined on different meshes. Similarly, the assembling of the load vectors corresponding to $\ell_s^{\text{ex}}(f, \cdot)$ and $\ell_t^{\text{in}}(f, \cdot)$ involves integration over intersections of elements from different meshes.

2. The stability of MEMs requires the projections Π^{ex} and Π^{in} to be stable in $H^{\frac{1}{2}}$. The obvious choice of L^2 projections involves again surface integrals of products of basis functions defined on different meshes.

In the following section we will introduce two mortar-like mappings different from the standard L^2 projection, that allow to choose $s = t = 0$ in (3) and hence avoid the assembling of the stiffness matrix for basis functions on two different meshes.

3 A Modified Mortar Mapping (MEM-M)

We recall that the FE spaces $\mathcal{V}^{\text{ex}}$ and $\mathcal{V}^{\text{in}}$ can be represented as direct sums $\mathcal{V}^{\text{ex}} = \mathcal{V}_\circ^{\text{ex}} \oplus \mathcal{V}_\gamma^{\text{ex}}$ and $\mathcal{V}^{\text{in}} = \mathcal{V}_\circ^{\text{in}} \oplus \mathcal{V}_\gamma^{\text{in}}$ where $\mathcal{V}_\gamma^{\text{ex}} = \text{tr}_{|\gamma^{\text{ex}}}\, \mathcal{V}^{\text{ex}}$ and $\mathcal{V}_\gamma^{\text{in}} = \text{tr}_{|\gamma^{\text{in}}}\, \mathcal{V}^{\text{in}}$ are the earlier introduced trace spaces of $\mathcal{V}^{\text{ex}}$ and $\mathcal{V}^{\text{in}}$. Let us define two mappings $\Pi_f^{\text{ex}} \psi^{\text{in}}$ for $\psi^{\text{in}} = \psi_\circ^{\text{in}} + \psi_\gamma^{\text{in}}$, with $\psi_\circ^{\text{in}} \in \mathcal{V}_\circ^{\text{in}}$, $\psi_\gamma^{\text{in}} \in \mathcal{V}_\gamma^{\text{in}}$ and $\Pi_f^{\text{in}} \psi^{\text{ex}}$ for $\psi^{\text{ex}} = \psi_\circ^{\text{ex}} + \psi_\gamma^{\text{ex}}$, with $\psi_\circ^{\text{ex}} \in \mathcal{V}_\circ^{\text{ex}}$, $\psi_\gamma^{\text{ex}} \in \mathcal{V}_\gamma^{\text{ex}}$ with:

$$\Pi_f^{\text{ex}} \psi^{\text{in}} := \Pi^{\text{ex}}(\psi_\gamma^{\text{in}} + \Psi_\circ^{\text{in}}), \text{ with } \mathsf{a}_0^{\text{in}}(\Psi_\circ^{\text{in}}, w) = \ell_0^{\text{in}}(f, w) - \mathsf{a}_0^{\text{in}}(\psi_\gamma^{\text{in}}, w)\ \forall w \in \mathcal{V}_\circ^{\text{in}},$$

where Π^{ex} is either the L^2-projection or standard nodal interpolation operator onto $\mathcal{V}_\gamma^{\text{ex}}$ and Π_f^{in} analogously defined. We then introduce the space

$$\mathcal{V}_{gf} = \{(v, w) \in \mathcal{V}^{\text{ex}} \times \mathcal{V}^{\text{in}},\ v_{|\Gamma} = \Pi^{\text{Dir}} g,\ v_{|\gamma^{\text{ex}}} = \Pi_f^{\text{ex}} w,\ w_{|\gamma^{\text{in}}} = \Pi_f^{\text{in}} v\},$$

and obtain the following modified version of the MEM for overlapping meshes: Find $(\psi^{\text{ex}}, \psi^{\text{in}}) \in \mathcal{V}_{\psi_0 f}$ such that

$$\mathsf{a}_0^{\text{ex}}(\psi^{\text{ex}}, v) + \mathsf{a}_0^{\text{in}}(\psi^{\text{in}}, w) = \ell_0^{\text{ex}}(f, v) + \ell_0^{\text{in}}(f, w) \quad \forall (v, w) \in \mathcal{V}_{0,0}, \tag{4}$$

A similar approach with the lowest order FE spaces in the non-destructive testing context has been adopted in [5, 6]. It can be shown that (4) is equivalent to the following formulation. Find $(\psi^{\text{ex}}, \psi^{\text{in}}) \in \mathcal{V}^{\text{ex}} \times \mathcal{V}^{\text{in}}$, $\psi_{|\Gamma}^{\text{ex}} = \Pi^{\text{Dir}} g$ such that:

$$\begin{aligned} &\mathsf{a}_0^{\text{ex}}(\psi^{\text{ex}}, v) + \mathsf{a}_0^{\text{in}}(\psi^{\text{in}}, w) = \ell_0^{\text{ex}}(f, v) + \ell_0^{\text{in}}(f, w) \quad \forall (v, w) \in \mathcal{V}_\circ^{\text{ex}} \times \mathcal{V}_\circ^{\text{in}}, v_{|\Gamma = 0} \\ &\psi_{|\gamma^{\text{ex}}}^{\text{ex}} = \Pi^{\text{in}} \psi^{\text{in}}, \quad \psi_{|\gamma^{\text{in}}}^{\text{in}} = \Pi^{\text{ex}} \psi^{\text{ex}}, \end{aligned} \tag{5}$$

which corresponds to the numerical zoom formulation in [8]. When the Π^{ex} and Π^{in} are interpolation operators and $\mathcal{V}^{\text{ex}}$ and $\mathcal{V}^{\text{in}}$ are lowest order Lagrangian FE spaces we can recall an optimal convergence result from [11, Theorem 1] for the error in the L^∞-norm.

4 Numerical Experiments

For the numerical experiments, we consider a rectangular domain $\Omega = [-1, 1]^2$ and define Ω^{in} as the polygon with vertices $(-0.125, 0.5)$, $(0.375, 0.25)$, $(0.375, -0.375)$, $(0, -0.5)$, $(-0.375, -0.375)$, and $(-0.5, 0.25)$. The meshes $\mathscr{T}^{\text{in}}$ and $\mathscr{T}^{\text{ex}}$ for the interior and exterior domain will be a Cartesian mesh and a triangular mesh. For simplicity we prefer to take $\Omega_h^{\text{ex}} = \Omega^{\text{ex}} = \Omega \setminus \Omega^{\text{in}}$. For the numerical test, we choose the data $f(x, y)$ and ψ_0 such that $\psi(x, y) = \cos(\pi x)\sin(\pi y)$ is the solution of (1). If h_{ex} (h_{in}) is the maximal diameter of elements in $\mathscr{T}^{\text{ex}}$ ($\mathscr{T}^{\text{in}}$), and p_{ex} (p_{in}) the local polynomial degree of the FE spaces $\mathscr{V}^{\text{ex}}$ ($\mathscr{V}^{\text{in}}$), one has optimal convergence if, for a smooth solution, the approximation error in the $H^1(\Omega_h^{\text{ex}})$ and $H^1(\Omega_h^{\text{in}})$-norms behaves as $O(h^{p-1})$, with $h = \max(h_{\text{ex}}, h_{\text{in}})$ and $p = \min(p_{\text{ex}}, p_{\text{in}})$ (in $L^2(\Omega_h^{\text{ex}})$ and $L^2(\Omega_h^{\text{in}})$-norms one dares to obtain $O(h^p)$). To keep the presentation as clear as possible we show in the following figures always the maximum between the error in Ω_h^{ex} and that in Ω_h^{in}.

We consider two different pairings of FE spaces $\mathscr{V}^{\text{ex}}$-$\mathscr{V}^{\text{in}}$. The first denoted with P1-Q1 uses lowest order linear FEs over $\mathscr{T}^{\text{ex}}$ and lowest order bilinear FEs over $\mathscr{T}^{\text{in}}$. The second pair, denoted with P2-Q3 uses quadratic FEs over $\mathscr{T}^{\text{ex}}$ and bicubic FEs over $\mathscr{T}^{\text{in}}$. The elements of P2-Q3 are not only continuous on Ω_h^{in} and Ω_h^{ex} but have also continuous gradients on Ω_h^{in}.

We focus on the overlapping MEM-M (4) which uses the modified mortar mappings and is equivalent to (5). We also analyse the influence on the error curves of using either L^2 projections or interpolation to realize the gluing across γ^{ex} and γ^{in} for the MEM-M. We start with the case where Ω_h^{in} has minimal overlap with Ω_h^{ex} (see Fig. 3, left). Thus γ^{in} is adapted with the refinements in Ω_h^{in} as shown in Fig. 3. Convergence results with MEM-M are presented in Fig. 4. The convergence rate with MEM-M is optimal for the error in the H^1-norm. The results look slightly better if we apply the interpolation instead of the L^2 projection in the definition of the mortar mapping.

Next, we study the convergence rates for MEM-M when Ω_h^{in} has a large overlap with Ω_h^{ex} (see Fig. 3, right). Note that both γ^{ex} and γ^{in} remain fixed during the refinements in Ω_h^{in}. Once again, the MEM-M yields optimal convergence rate in the H^1 norm (see Fig. 4). Moreover in the case of larger overlap we observe even optimal convergence in the L^2-norm. There is no qualitative difference between MEM-M based on the L^2-projection or on the interpolation.

With the classical overlapping MEM (3) with the parameters s and t set to zero (MEM-0), the convergence rates in the H^1 and L^2 norms are not optimal in the case of minimal overlap between Ω_h^{in} and Ω_h^{ex}. The MEM-0 does not yield convergence in the case of a large overlap between Ω_h^{in} and Ω_h^{ex}.

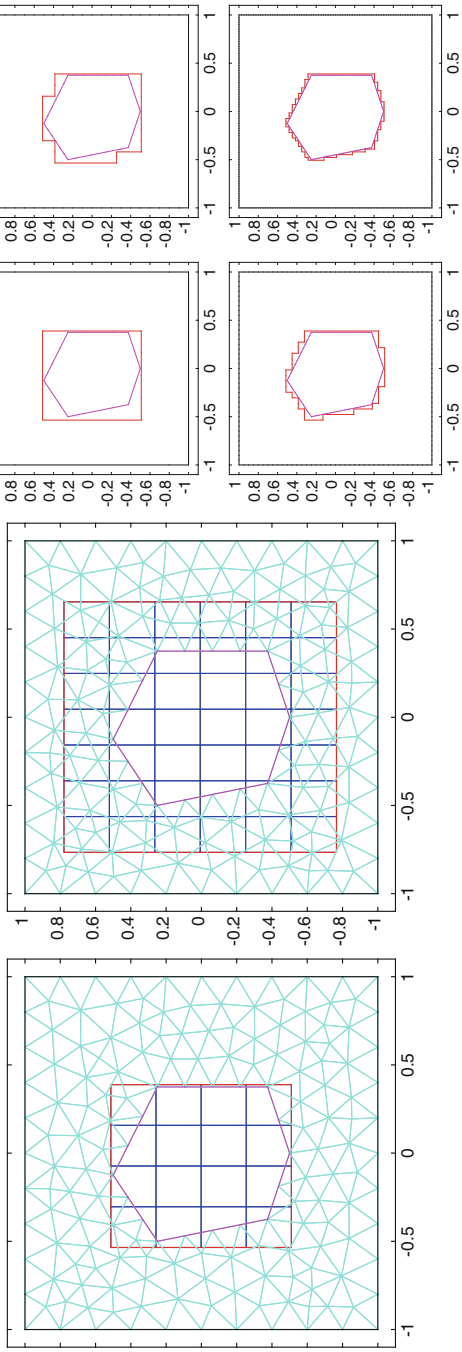


Fig. 3 *Left and Center*: Sketch of the two different settings. We may choose Ω_h^{in} to have a minimal overlap with Ω_h^{ex} (*left*), that is γ^{ex} is contained in the layer of elements of $\mathscr{T}^{\text{in}}$ which define γ^{in}. Otherwise, we say that Ω_h^{in} has a large overlap with Ω_h^{ex} (*center*). *Right*: Adaptive definition of γ^{in} in the case of minimal overlap between Ω_h^{in} and Ω_h^{ex}. Note that, γ^{ex} (*magenta*) remains fixed, while γ^{in} (*red*) changes due to the refinements in Ω_h^{in}. The interior edges of elements of $\mathscr{T}^{\text{ex}}$ and $\mathscr{T}^{\text{in}}$ are omitted for clearness

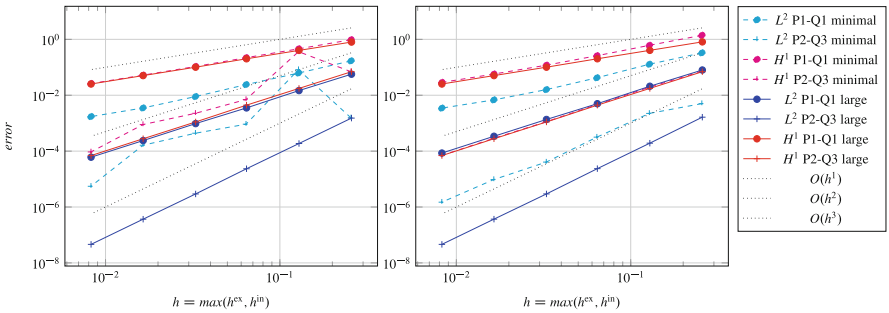


Fig. 4 Convergence in L^2 and H^1 of the scheme MEM-M using L^2-projection (*left*) or nodal interpolation (*right*)

References

1. Achdou, Y., Maday, Y.: The mortar element method with overlapping subdomains. SIAM J. Numer. Anal. **40**(2), 601–628 (2002)
2. Blum, J.: Numerical Simulation and Optimal Control in Plasma Physics. Wiley/Gauthier-Villars Series in Modern Applied Mathematics. Wiley, Chichester (1989)
3. Cai, X.-C., Dryja, M., Sarkis M.: Overlapping nonmatching grid mortar element methods for elliptic problems. SIAM J. Numer. Anal. **36**(2), 581–606 (1999)
4. Chesshire, G., Henshaw, W.D.: Composite overlapping meshes for the solution of partial differential equations. J. Comput. Phys. **90**(1), 1–64 (1990)
5. Christophe, A., Santandrea, L., Rapetti, F., Krebs, G., Le Bihan, Y.: An overlapping non-matching grid mortar element method for Maxwell's equations. IEEE Trans. Magn. **50**(2), 409–412 (2014).
6. Christophe, A., Le Bihan, Y., Rapetti, F.: A mortar element approach on overlapping non-nested grids: application to eddy current non-destructive testing. Appl. Math. Comput. **267**, 71–82 (2015)
7. Fish, J., Belytschko, T.: Elements with embedded localization zones for large deformation problems. Comput. Struct. **30**(1), 247–256 (1988)
8. Hecht, F., Lozinski, A., Pironneau, O.: Numerical zoom and the Schwarz algorithm. In: Domain decomposition methods in science and engineering XVIII, pp 63–73. Springer, Berlin (2009)
9. Heumann, H., Rapetti, F.: A finite element method with overlapping meshes for free-boundary axisymmetric plasma equilibria in realistic geometries. RR-8916, Inria SAM (2016).
10. Kuznetsov,Y.A.: Overlapping domain decomposition with non-matching grids. In: Recent developments in domain decomposition methods and flow problems. GAKUTO International Series. Mathematical Sciences and Applications, vol. 11, pp. 62–71. Gakkōtosho, Tokyo (1998)
11. Lozinski, A., Pironneau, O.: Numerical zoom for advection diffusion problems with localized multiscales. Numer. Methods Partial Differ. Equ. **27**(1), 197–207 (2011)
12. Mote, C.D., Jr.: Global-local finite element. Int. J. Numer. Methods Eng. **3**, 565–574 (1971)
13. Parter, S.V.: On the overlapping grid method for elliptic boundary value problems. SIAM J. Numer. Anal. **36**(3), 819–852 (1999)
14. Steger, J.L., Benek, J.A.: On the use of composite grid schemes in computational aerodynamics. Comput. Methods Appl. Mech. Eng. **64**, 301–320 (1987)

Eddy Current Testing Models for the Analysis of Corrosion Effects in Metal Plates

Valentina Koliskina, Andrei Kolyshkin, Rauno Gordon, and Olev Märtens

Abstract Direct method for the solution of eddy current testing problem for the case where an air core coil is located above a conducting two-layer plate with a flaw in the form of a cylindrical inclusion with reduced electrical conductivity is presented in the paper. Semi-analytical approach (the TREE method) is used to construct the solution of the system of equations for the components of the vector potential. The flaw is assumed to be symmetric with respect to the coil. Numerical calculations are performed using the proposed model and Comsol Multiphysics software. The obtained values of the change in impedance of the coil for both methods are found to be in a good agreement. The proposed model can be used for the assessment of the effect of corrosion in metal plates.

1 Introduction

Eddy current methods are widely used to analyze the effect of corrosion in metal plates [2]. Mathematical methods that are used to analyze such problems can be divided into two broad categories: (a) numerical methods and (b) analytical and semi-analytical methods. Methods from category (a) include, for example, finite element methods or finite difference methods (see [4, 9]). Analytical methods can be used for the cases where conducting medium is unbounded with respect to one or two spatial coordinates. Such models are discussed in detail in [1, 7]. Corroded samples have finite sizes so that truly analytical approach cannot be used. However, the domain of applicability of analytical methods can be extended using the following idea described in [8]: it is assumed that the electromagnetic field is exactly zero at a sufficiently large distance from the source of alternating current. In this case the domain where the problem has to be solved becomes finite and

V. Koliskina • A. Kolyshkin (✉)
Riga Technical University, 1 Kalku street, Riga, LV 1658, Latvia
e-mail: v.koliskina@gmail.com; andrejs.koliskins@rtu.lv

R. Gordon • O. Märtens
Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
e-mail: rauno.gordon@ttu.ee; olev.martens@ttu.ee

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_17

method of separation of variables can be used to solve the system of equations for the vector potential in each of the regions of interest. Examples where such an approach (known as the TREE method in the literature) is used for solution of eddy current testing problems with asymmetric flaws can be found in [5, 6]. In the present paper we present semi-analytical solution of direct eddy current testing problem for a two-layer plate. The plate contains a flaw in the form of a cylindrical inclusion with reduced electrical conductivity. The axis of the inclusion coincides with the axis of an air core coil. The formula for the change in impedance of the coil is derived. Numerical computations show good agreement of the proposed model with the computations performed using Comsol Multiphysics software.

2 Mathematical Analysis

Suppose an air core coil carrying alternating current of frequency f is located above a conducting nonmagnetic two-layer plate. The conductivites of the upper and lower layers are σ_1 and σ_2, respectively. The lower layer is infinite in the vertical direction.The upper layer has a flaw in the form of a circular cylinder of radius c. The flaw has the two layers in the vertical direction: the upper layer where $-d_1 < z < 0$ and the lower layer where $-d_1 < z < -d_2$. The upper layer of the flaw has zero conductivity while the lower layer has a reduced electrical conductivity $\sigma_3 < \sigma_1$. This is a typical situation in corroded samples [2]. It is assumed that the axis of the flaw coincides with the axis of the coil. The geometrical parameters of the coil are as follows: the inner and outer radii are r_1 and r_2, respectively, z_1 is the distance from the bottom of the coil to the upper layer of the conducting plate, $z_2 - z_1$ is the height of the coil and N is the number of turns. Due to axial symmetry only the azimuthal component of the vector potential is not equal to zero. Thus, in the system of cylindrical polar coordinates (r, φ, z) centered at the axis of the coil at the point located on the upper surface of the plate we have

$$\mathbf{A} = A(r, z) \exp(j\omega t)\mathbf{e}_{\boldsymbol{\varphi}}, \tag{1}$$

where $\omega = 2\pi f, j = \sqrt{-1}$ and $\mathbf{e}_\varphi$ is the unit vector in the azimuthal direction. We denote by R_0, R_1, R_2 and R_3 the regions where $z > 0$, $-d_1 < z < 0$, $-d_1 - d_2 < z < -d_1$ and $z < -d_1 - d_2$, respectively. The component of the vector potential in region R_i is denoted by $A_i, i = 0, 1, 2, 3$. All the components of the vector potential are equal to zero at $r = b$:

$$A_i(b, z) = 0, \quad i = 0, 1, 2, 3. \tag{2}$$

Since the problem is linear it is natural to use the superposition principle. First, we consider a single-turn coil of radius r_0 located at distance h above the conducting plate. The problem is solved for the case of the single-turn coil and then the solution

is integrated over the cross-section of the coil in order to find the vector potential for the coil of finite dimensions.

The system of equations for the components of the vector potential in regions R_i has the form

$$\frac{\partial^2 A_0}{\partial r^2} + \frac{1}{r}\frac{\partial A_0}{\partial r} - \frac{A_0}{r^2} + \frac{\partial^2 A_0}{\partial z^2} = -\mu_0 I \delta(r - r_0)\delta(z - h), \tag{3}$$

$$\frac{\partial^2 A_1}{\partial r^2} + \frac{1}{r}\frac{\partial A_1}{\partial r} - \frac{A_1}{r^2} - j\omega\sigma_1(r)\mu_0 A_1 + \frac{\partial^2 A_1}{\partial z^2} = 0, \tag{4}$$

$$\frac{\partial^2 A_2}{\partial r^2} + \frac{1}{r}\frac{\partial A_2}{\partial r} - \frac{A_2}{r^2} - j\omega\sigma_2(r)\mu_0 A_2 + \frac{\partial^2 A_2}{\partial z^2} = 0, \tag{5}$$

$$\frac{\partial^2 A_3}{\partial r^2} + \frac{1}{r}\frac{\partial A_3}{\partial r} - \frac{A_3}{r^2} - j\omega\sigma_3\mu_0 A_3 + \frac{\partial^2 A_3}{\partial z^2} = 0, \tag{6}$$

where $\delta(x)$ is the Dirac delta function, μ_0 is the magnetic constant, $\sigma_1(r) = \sigma_1$ if $c < r < b$ and $\sigma_1(r) = 0$ if $0 < r < c$. Similarly, $\sigma_2(r) = \sigma_1$ if $c < r < b$ and $\sigma_2(r) = \sigma_3$ if $0 < r < c$. The current in the coil is given by

$$\mathbf{I}^e = I \exp(j\omega t)\mathbf{e}_\varphi. \tag{7}$$

The boundary conditions are

$$A_0 \mid_{z=0} = A_1^{air} \mid_{z=0}, \quad \frac{\partial A_0}{\partial z} \mid_{z=0} = \frac{\partial A_1^{air}}{\partial z} \mid_{z=0}, \quad 0 < r < c, \tag{8}$$

$$A_0 \mid_{z=0} = A_1^{con} \mid_{z=0}, \quad \frac{\partial A_0}{\partial z} \mid_{z=0} = \frac{\partial A_1^{con}}{\partial z} \mid_{z=0}, \quad c < r < b, \tag{9}$$

$$A_1^{air} \mid_{z=-d_1} = A_2^{red} \mid_{z=-d_1}, \quad \frac{\partial A_1^{air}}{\partial z} \mid_{z=-d_1} = \frac{\partial A_2^{red}}{\partial z} \mid_{z=-d_1}, \quad 0 < r < c, \tag{10}$$

$$A_1^{con} \mid_{z=-d_1} = A_2 \mid_{z=-d_1}, \quad \frac{\partial A_1^{con}}{\partial z} \mid_{z=-d_1} = \frac{\partial A_2}{\partial z} \mid_{z=-d_1}, \quad c < r < b, \tag{11}$$

$$A_2^{red} \mid_{z=-d_1-d_2} = A_3 \mid_{z=-d_1-d_2}, \quad \frac{\partial A_2^{red}}{\partial z} \mid_{z=-d_1-d_2} = \frac{\partial A_3}{\partial z} \mid_{z=-d_1-d_2}, \quad 0 < r < c, \tag{12}$$

$$A_2 \mid_{z=-d_1-d_2} = A_3 \mid_{z=-d_1-d_2}, \quad \frac{\partial A_2}{\partial z} \mid_{z=-d_1-d_2} = \frac{\partial A_3}{\partial z} \mid_{z=-d_1-d_2}, \quad c < r < b. \tag{13}$$

Here the notations A_1^{air} and A_1^{con} are used to denote the vector potentials in region R_1 where $0 < r < c$ and $c < r < b$, respectively. Similar notation is used for the region

of reduced electrical conductivity: A_2^{red} in the region $0 < r < c$. The vector potential and its derivative with respect to r are continuous at $r = c$:

$$A_1^{con} \mid_{r=c} = A_1^{air} \mid_{r=c}, \quad \frac{\partial A_1^{con}}{\partial r} \mid_{r=c} = \frac{\partial A_1^{air}}{\partial r} \mid_{r=c}, \tag{14}$$

$$A_2^{red} \mid_{r=c} = A_2 \mid_{r=c}, \quad \frac{\partial A_2^{red}}{\partial r} \mid_{r=c} = \frac{\partial A_2}{\partial r} \mid_{r=c}. \tag{15}$$

In addition, the vector potential is assumed to be bounded at infinity in the vertical direction:

$$A_0 \mid_{z\to+\infty} = 0, \quad A_3 \mid_{z\to-\infty} = 0. \tag{16}$$

The solution in region R_0 has the form

$$A_0(r,z) = \sum_{i=1}^{\infty} D_{1i} e^{-\lambda_i z} J_1(\lambda_i r) + \frac{\mu_0 I r_0}{b^2} \sum_{i=1}^{\infty} \frac{J_1(\lambda_i r_0)}{\lambda_i J_0^2(\lambda_i b)} e^{-\lambda_i |z-h|} J_1(\lambda_i r), \tag{17}$$

where $\lambda_i = \alpha_i / b$, $\alpha_i, i = 1, 2, \ldots$ are the roots of the equation $J_1(\alpha) = 0$, J_0 and J_1 are the Bessel functions of the first kind of orders 0 and 1, respectively. The solution in region R_1 is

$$A_1^{con}(r,z) = \sum_{i=1}^{\infty} [D_{2i} e^{p_i z} + D_{3i} e^{-p_i z}] J_1(q_i r), \tag{18}$$

$$A_1^{air}(r,z) = \sum_{i=1}^{\infty} [(D_{4i} J_1(p_i r) + D_{5i} Y_1(p_i r)) e^{p_i z} + (D_{6i} J_1(p_i r) + D_{7i} Y_1(p_i r)) e^{-p_i z}], \tag{19}$$

where q_i is the separation constant, $p_i = \sqrt{q_i^2 + j\omega\sigma_1\mu_0}$ and Y_1 is the Bessel function of the second kind of order 1. Similarly, the solution in region R_2 is

$$A_2(r,z) = \sum_{i=1}^{\infty} [D_{8i} e^{s_i z} + D_{9i} e^{-s_i z}] J_1(q_i r), \tag{20}$$

$$A_2^{red}(r,z) = \sum_{i=1}^{\infty} [(D_{10i} J_1(s_i r) + D_{11i} Y_1(s_i r)) e^{s_i z} + (D_{12i} J_1(s_i r) + D_{13i} Y_1(s_i r)) e^{-s_i z}], \tag{21}$$

where $s_i = \sqrt{q_i^2 + j\omega\sigma_3\mu_0}$. Finally, the solution in region R_3 has the form

$$A_3(r,z) = \sum_{i=1}^{\infty} D_{14i} e^{u_i z} J_1(q_i r), \tag{22}$$

where $u_i = \sqrt{q_i^2 + j\omega\sigma_2\mu_0}$.

Using the boundary conditions (8)–(13) we obtain the unknown coefficients $D_{1i}-D_{14i}$. The interface conditions (14)–(15) give the equation for the unknown complex eigenvalues p_i:

$$p_i J_1'(p_i c)T'(q_i c) = q_i J_1(p_i c)T'(q_i c), \tag{23}$$

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 technical details of the derivation are not shown here for brevity and can be found, for example, in [3] where a similar problem with two cylindrical flaws is solved. The induced vector potential in air due to the presence of the conducting plate is

$$A_0^{ind}(r, z, r_0, h) = \sum_{i=1}^{n} D_{1i} e^{-\lambda_i z} J_1(\lambda_i r), \tag{24}$$

where only the first n terms of the series are taken into account.

Using the superposition principle we obtain the induced vector potential in air due to currents in the whole coil:

$$A_{0coil}^{ind}(r, z) = \int_{r_1}^{r_2} \int_{z_1}^{z_2} A_0^{ind}(r, z, r_0, h) dr_0 dh. \tag{25}$$

The change in impedance of the coil is computed as follows (see [8]):

$$Z^{ind} = \frac{i\omega}{I^2} \int_V \mathbf{A}_{0coil}^{ind} \cdot \mathbf{I}\, dV. \tag{26}$$

Using (24)–(26) we obtain

$$Z^{ind} = \frac{2j\omega\pi\mu_0 N^2}{(r_2 - r_1)^2 (z_2 - z_1)^2} \sum_{m=1}^{n} \frac{(e^{-\lambda_m z_2} - e^{-\lambda_m z_1})}{\lambda_m^3} \int_{\lambda_m r_1}^{\lambda_m r_2} \xi J_1(\xi)\, d\xi$$
$$\times \sum_{i=1}^{n} Y_{mi} \frac{(e^{-\lambda_i z_2} - e^{-\lambda_i z_1})}{\lambda_i^3} \int_{\lambda_i r_1}^{\lambda_i r_2} \eta J_1(\eta)\, d\eta. \tag{27}$$

(the formulas for the elements of the matrix Y are bulky and are not shown here).

3 Numerical Results

Formula (27) is used to compute the change in impedance of the coil for the following parameters of the problem: $r_1 = 2.5\,\text{mm}$, $r_2 = 5.5\,\text{mm}$, $z_1 = 0.3\,\text{mm}$, $z_2 = 2.6\,\text{mm}$, $d_1 = 1\,\text{mm}$, $d_2 = 1\,\text{mm}$, $b = 55\,\text{mm}$, $\sigma_1 = 0.5\,\text{Ms/m}$, $\sigma_2 = 7\,\text{Ms/m}$,

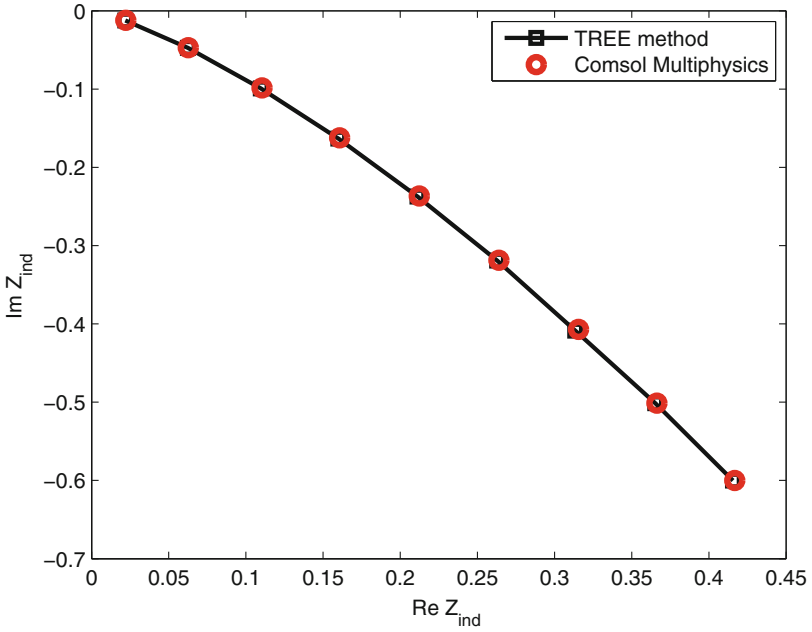


Fig. 1 The change in impedance of the coil for the nine frequencies

$c = 3$ mm. Nine frequencies are used for the calculations: from 1 to 9 kHz with the stepsize of 1 kHz. The results of computations are shown in Fig. 1. The solid curve represents theoretical calculations with the TREE method. In addition, the problem is solved using Comsol Multiphysics software. The points on the graph represent the values computed using Comsol Multiphysics. As can be seen from the graph, both calculations are in a good agreement.

Acknowledgements This work was partially supported by the grant 632/2014 by the Latvian Council of Science.

References

1. Antimirov, M.Ya., Kolyshkin, A.A., Vaillancourt, R.: Mathematical Models for Eddy Current Testing. CRM, Montréal (1997)
2. He, Y., Tian, G., Alamin, M., Simm, A., Kackson, P.: Steel corrosion characterization using pulsed eddy current systems. IEEE Sensors J. **12**, 2113–2120 (2012)
3. Koliskina, V., Kolyshkin, A.: Mathematical model for eddy current testing of metal plates with two cylindrical flaws. In: Leonowicz, Z. (ed.) Proceedings of the 15th International Conference on Environment and Electrical Engineering, pp. 374–377. Elsevier, Institute of Electrical and Electronics Engineers (2015)

4. Rodriguez, A.A., Valli, A.: Eddy Current Approximation of Maxwell Equations. Springer, New York (2010)
5. Skarlatos, A., Theodoulidis, T.: Solution to the eddy-current induction problem in a conducting half-space with a vertical cylindrical borehole. Proc. Roy. Soc. Lond. **468**, 1758–1777 (2012)
6. Skarlatos, A., Theodoulidis, T.: Calculation of the eddy-current flow around a cylindrical through-hole in a finite thickness plate. IEEE Trans. Magn. **51**(9), (2015). doi:10.1109/TMAG.2015.2426676
7. Tegopoulos, J.A., Kriezis, E.E.: Eddy Currents in Linear Conducting Media. Elsevier, Amsterdam (1985)
8. Theodoulidis, T.P., Kriezis, E.E.: Eddy Current Canonical Problems (with Applications to Nondestructive Evaluation). Tech Science Press, New York (2006)
9. Touzani, R., Rappaz, J.: Mathematical Models for Eddy Currents and Electrostatics. Springer, New York (2014)

Convergence of a Leap-Frog Discontinuous Galerkin Method for Time-Domain Maxwell's Equations in Anisotropic Materials

Adérito Araújo, Sílvia Barbeiro, and Maryam Khaksar Ghalati

Abstract We present an explicit leap-frog discontinuous Galerkin method for time-domain Maxwell's equations in anisotropic materials and establish its convergence properties. We illustrate the convergence results of the fully discrete scheme with numerical tests. This work was developed in the framework of a more general project that aims to develop a computational model to simulate the electromagnetic wave's propagation through the eye's structures in order to create a virtual optical coherence tomography scan (Santos et al., 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 8147–8150, 2015).

1 Introduction

The human retina is a complex structure in the eye that is responsible for the sense of vision. It is part of the central nervous system and it is composed by several layers, namely the outer nuclear layer that comprises the cells bodies of light sensitive photoreceptors cells, rods and cones (see Fig. 1). For many diseases that affect the eye, the diagnosis is not straightforward. The sensitivity of this structure makes medical analysis particularly complicated. Most of the diagnoses are made either by direct observation, with the possible injection of dyes, to enhance certain parts of the organ, or by numbing the eye and directly measuring its inner pressure or thickness. There are a number of eye-related pathologies that can be identified by the detailed analysis of the retinal layers.

Optical Coherence Tomography (OCT) is an increasingly popular noninvasive technique that has been successfully used as a diagnostic tool in ophthalmology in the past decades. This method allows the assessment of the human retina *in vivo* and has been shown to provide functional information. By analysing data

A. Araújo (✉) • S. Barbeiro • M.K. Ghalati
CMUC, Department of Mathematics, University of Coimbra, Apartado 3008, EC Santa Cruz, 3001-501 Coimbra, Portugal
e-mail: alma@mat.uc.pt; silvia@mat.uc.pt; maryam@mat.uc.pt

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_18

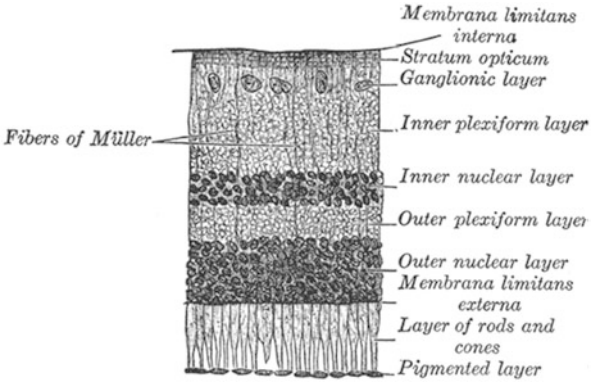


Fig. 1 Section of retina. Henry Gray. Anatomy of the human body. Philadelphia: Lea & Febiger, 1918 (in public domain at Bartleby.com)

acquired through OCT, several retinal pathologies, such as diabetic retinopathy, or macular edema, can be detected in their early stages, before noticeable morphologic alterations on the retina [9, 11]. As OCT standard techniques only provide structural information [10], it is necessary to expand OCT data analysis to account for both structural and functional information. OCT provides also the possibility of evaluating different elements in measuring the retinal nerve fibre layer (RNFL), namely the tendency of RNFL thinning in glaucoma and other diseases that involve optic nerve atrophy. Waveguides with induced anisotropy may worth to be modeled as they could play a role in biological waveguides [7].

Maxwell's equations are a fundamental set of partial differential equations which describe electromagnetic wave interactions with materials. Here we shall consider the time domain Maxwell's equations in the transverse electric (TE) mode, as in [5], where the only non-vanishing components of the electromagnetic fields are E_x, E_y and H_z. Assuming no conductivity effects, the equations in the non-dimensional form are

$$\epsilon \frac{\partial E}{\partial t} = \nabla \times H, \qquad \mu \frac{\partial H}{\partial t} = -\text{curl}\, E, \qquad \text{in } \Omega \times (0, T_f], \tag{1}$$

where $E = (E_x, E_y)$ represent the electric field components and $H = (H_z)$ the magnetic field components. This equations are set and solved on the bounded polygonal domain $\Omega \subset \mathbb{R}^2$. The electric permittivity of the medium, ϵ, and the magnetic permeability of the medium, μ, are varying in space, being ϵ an anisotropic tensor

$$\epsilon = \begin{pmatrix} \epsilon_{xx} & \epsilon_{xy} \\ \epsilon_{yx} & \epsilon_{yy} \end{pmatrix}. \tag{2}$$

We assume that electric permittivity tensor ϵ is symmetric and uniformly positive definite for almost every $(x, y) \in \Omega$, and it is uniformly bounded with a strictly positive lower bound, *i.e.*, there are constants $\underline{\epsilon} > 0$ and $\overline{\epsilon} > 0$ such that, for almost

every $(x, y) \in \Omega$, $\underline{\epsilon}|\xi|^2 \leq \xi^T \epsilon(x, y)\xi \leq \overline{\epsilon}|\xi|^2, \forall \xi \in \mathbb{R}^2$. We also assume that there are constants $\underline{\mu} > 0$ and $\overline{\mu} > 0$ such that, for almost every $(x, y) \in \Omega$, $\underline{\mu} \leq \mu(x, y) \leq \overline{\mu}$.

Equations (1) must be complemented by initial conditions $E(x, y, 0) = E_0(x, y)$ and $H(x, y, 0) = H_0(x, y)$, $(x, y) \in \Omega$, and by proper boundary conditions. Here we consider the first order Silver-Müller absorbing boundary conditions

$$n \times E = c\mu n \times (H \times n) \qquad \text{on } \partial\Omega, \tag{3}$$

where $n = (n_x, n_y)^T$ is the unit outward normal vector to the boundary and c is the speed with which a wave travels along the direction of the unit normal, defined, using the effective permittivity $\epsilon_{eff} = \det(\epsilon)/(n^T \epsilon n)$ ([5]), by $c = 1/\sqrt{\mu \epsilon_{eff}}$.

Despite the relevance of the anisotropic case, most of the numerical formulations have been restricted to isotropic and dispersive materials [4, 8]. In this work we consider a model with anisotropic permittivity tensors. We will consider a nodal discontinuous Galerkin (DG) method for the space discretization [4] and a leap-frog method for the time integration. The main contribution of this work is to propose a fully explicit second order in time and arbitrary high order in space method.

2 A Leap-frog Discontinuous Galerkin Method

Assume that the computational domain Ω is a bounded polygonal set that is partitioned into K triangular elements T_k such that $\overline{\Omega} = \cup_k T_k$. For simplicity, we consider that the resulting mesh $\mathcal{T}_h$ is conforming. The finite element space is then taken to be $V_N = \{v \in L^2(\Omega)^3 : v|_{T_k} \in P_N(T_k)^3\}$, where $P_N(T_k)$ denotes the space of polynomials of degree less than or equal to N on T_k. On each element T_k, the solution fields are approximated by vector polynomial functions $(\tilde{E}_x(x, y, t), \tilde{E}_y(x, y, t), \tilde{H}_z(x, y, t))$.

The approximate fields are allowed to be discontinuous across element boundaries. In this way, we introduce the notation for the jumps of the field values across the interfaces of the elements, $[\tilde{E}] = \tilde{E}^- - \tilde{E}^+$ and $[\tilde{H}] = \tilde{H}^- - \tilde{H}^+$, where the superscript "+" denotes the neighboring element and the superscript "−" refers to the local cell. Furthermore we introduce, respectively, the cell-impedances and cell-conductances $Z^\pm = \mu^\pm c^\pm$ and $Y^\pm = \left(Z^\pm\right)^{-1}$. At the outer cell boundaries we set $Z^+ = Z^-$.

The coupling between elements is introduced via the numerical flux, defined by

$$n \cdot \left(F(\tilde{q}) - F^*(\tilde{q})\right) = \begin{pmatrix} \dfrac{-n_y}{Z^+ + Z^-}\left(Z^+[\tilde{H}_z] - \alpha\left(n_x[\tilde{E}_y] - n_y[\tilde{E}_x]\right)\right) \\ \dfrac{n_x}{Z^+ + Z^-}\left(Z^+[\tilde{H}_z] - \alpha\left(n_x[\tilde{E}_y] - n_y[\tilde{E}_x]\right)\right) \\ \dfrac{1}{Y^+ + Y^-}\left(Y^+\left(n_x[\tilde{E}_y] - n_y[\tilde{E}_x]\right) - \alpha[\tilde{H}_z]\right) \end{pmatrix}.$$

The parameter $\alpha \in [0,1]$ can be used to control dissipation. Taking $\alpha = 0$ yields a non dissipative central flux while $\alpha = 1$ corresponds to the classic upwind flux.

In order to discretize the boundary conditions, using the same kind of approach as in [3], we consider, for upwind fluxes $Z^-\tilde{H}_z^+ = n_x\tilde{E}_y^+ - n_y\tilde{E}_x^+$ or equivalently $\tilde{H}_z^+ = Y^-(n_x\tilde{E}_y^+ - n_y\tilde{E}_x^+)$ and, for central fluxes $Z^-\tilde{H}_z^+ = (n_x\tilde{E}_y^- - n_y\tilde{E}_x^-)$ and $Y^-(n_x\tilde{E}_y^+ - n_y\tilde{E}_x^+) = \tilde{H}_z^-$. This is equivalent to consider, for both upwind and central fluxes, $\alpha = 1$ for numerical flux at the outer boundary and $[\tilde{E}_x] = \tilde{E}_x^-$, $[\tilde{E}_y] = \tilde{E}_y^-$ and $[\tilde{H}_z] = \tilde{H}_z^-$.

To define a fully discrete scheme, we divide the time interval $[0,T]$ into M subintervals by points $0 = t^0 < t^1 < \cdots < t^M = T$, where $t^m = m\Delta t$, Δt is the time step size and $T + \Delta t/2 \leq T_f$. The unknowns related to the electric field are approximated at integer time-stations t^m and are denoted by $\tilde{E}_k^m = \tilde{E}_k(.,t^m)$. The unknowns related to the magnetic field are approximated at half-integer time-stations $t^{m+1/2} = (m+\frac{1}{2})\Delta t$ and are denoted by $\tilde{H}_k^{m+1/2} = \tilde{H}_k(.,t^{m+1/2})$. With the above setting, we can now formulate the leap-frog DG method: given an initial approximation $(\tilde{E}_{x_k}^0, \tilde{E}_{y_k}^0, \tilde{H}_{z_k}^{1/2}) \in V_N$, for each $m = 0,1,\ldots,M-1$, find $(\tilde{E}_{x_k}^{m+1}, \tilde{E}_{y_k}^{m+1}, \tilde{H}_{z_k}^{m+1/2}) \in V_N$ such that, $\forall (u_k, v_k, w_k) \in V_N$,

$$\left(\epsilon_{xx}\frac{\tilde{E}_{x_k}^{m+1} - \tilde{E}_{x_k}^m}{\Delta t} + \epsilon_{xy}\frac{\tilde{E}_{y_k}^{m+1} - \tilde{E}_{y_k}^m}{\Delta t}, u_k\right)_{T_k} = \left(\partial_y \tilde{H}_{z_k}^{m+1/2}, u_k\right)_{T_k} + \left(\frac{-n_y}{Z^+ + Z^-}\left(Z^+[\tilde{H}_z^{m+1/2}] - \alpha\left(n_x[\tilde{E}_y^m] - n_y[\tilde{E}_x^m]\right)\right), u_k\right)_{\partial T_k}, \tag{4}$$

$$\left(\epsilon_{yx}\frac{\tilde{E}_{x_k}^{m+1} - \tilde{E}_{x_k}^m}{\Delta t} + \epsilon_{yy}\frac{\tilde{E}_{y_k}^{m+1} - \tilde{E}_{y_k}^m}{\Delta t}, v_k\right)_{T_k} = -\left(\partial_x \tilde{H}_{z_k}^{m+1/2}, v_k\right)_{T_k} + \left(\frac{n_x}{Z^+ + Z^-}\left(Z^+[\tilde{H}_z^{m+1/2}] - \alpha\left(n_x[\tilde{E}_y^m] - n_y[\tilde{E}_x^m]\right)\right), v_k\right)_{\partial T_k}, \tag{5}$$

$$\left(\mu\frac{\tilde{H}_{z_k}^{m+3/2} - \tilde{H}_{z_k}^{m+1/2}}{\Delta t}, w_k\right)_{T_k} = \left(\partial_y \tilde{E}_{x_k}^{m+1} - \partial_x \tilde{E}_{y_k}^{m+1}, w_k\right)_{T_k} + \left(\frac{1}{Y^+ + Y^-}\left(Y^+(n_x[\tilde{E}_y^{m+1}] - n_y[\tilde{E}_x^{m+1}]) - \alpha[\tilde{H}_z^{m+1/2}]\right), w_k\right)_{\partial T_k}, \tag{6}$$

where $(\cdot,\cdot)_{T_k}$ and $(\cdot,\cdot)_{\partial T_k}$ denote the classical $L^2(T_k)$ and $L^2(\partial T_k)$ inner-products. The boundary conditions are considered as described in the previous section.

We want to emphasize that the scheme (4)–(6) is fully explicit in time, in opposition to [6], where the scheme is defined with the upwind fluxes involving the unknowns E_k^{m+1} and $H_k^{m+3/2}$ and to [2], where the scheme that is defined with

the central fluxes leads to a locally implicit time method in the case of Silver-Müller absorbing boundary conditions.

In Theorem 1 we state the convergence properties of the method (4)–(6), assuming that the mesh is regular in the sense that there is a constant $\tau > 0$ such that for all $T_k \in \mathscr{T}_h$, $h_k/\tau_k \leq \tau$, where τ_k denotes the maximum diameter of a ball inscribed in T_k, h_k is the diameter of the triangle $T_k \in \mathscr{T}_h$, and h is the maximum element diameter.

Theorem 1 *Let us consider the leap-frog DG method (4)–(6) complemented with the discrete Silver-Müller absorbing boundary condition and suppose that the solution of the Maxwell's equations (1) complemented by (3) has the following regularity:* $E_x, E_y, H_z \in L^\infty(0, T_f; H^{s+1}(\Omega))$, $\frac{\partial E_x}{\partial t}, \frac{\partial E_y}{\partial t}, \frac{\partial H_z}{\partial t} \in L^2(0, T_f; H^{s+1}(\Omega) \cap L^\infty(\partial\Omega))$ *and* $\frac{\partial^2 E_x}{\partial t^2}, \frac{\partial^2 E_y}{\partial t^2}, \frac{\partial^2 H_z}{\partial t^2} \in L^2(0, T_f; H^1(\Omega))$, $s \geq 0$. *If the time step* Δt *satisfies the stability condition stated in Theorem 4.1 in [1], i.e.,*

$$\Delta t < \tilde{C} \frac{\min\{\underline{\epsilon}, \underline{\mu}\}}{\max\{C_E, C_H\}} \min\{h_k\},$$

where

$$C_E = N^2 + (N+1)(N+2)\left(5 + \frac{4\alpha + 1}{2\min\{Z_k\}}\right),$$

$$C_H = N^2 + (N+1)(N+2)\left(5 + \frac{2\alpha + 1}{\min\{Y_k\}}\right),$$

where $\tilde{C}$ *is a positive constant independent of h and N, then*

$$\max_{1 \leq m \leq M} \left(\|E^m - \tilde{E}^m\|_{L^2(\Omega)} + \|H_z^{m+1/2} - \tilde{H}_z^{m+1/2}\|_{L^2(\Omega)} \right) \leq C(\Delta t + h^{\min\{s,N\}}) + C\left(\|E^0 - \tilde{E}^0\|_{L^2(\Omega)} + \|H_z^{1/2} - \tilde{H}_z^{1/2}\|_{L^2(\Omega)} \right),$$

where C is a generic positive constant independent of Δt *and the mesh size h.*

We want to remark that we only get first order convergence in time. A possible way to recover second order convergence is to consider a locally implicit time scheme (see *e.g.* [2]). In order to keep efficiency, we propose an alternative which is explicit and second order convergent in time. The scheme is defined as following: for each time step solve (4)–(6) and save the solution in the variables $(\tilde{\tilde{E}}_{x_k}^{m+1}, \tilde{\tilde{E}}_{y_k}^{m+1}, \tilde{\tilde{H}}_{z_k}^{m+3/2})$. Then the numerical solution $(\tilde{E}_{x_k}^{m+1}, \tilde{E}_{y_k}^{m+1}, \tilde{H}_{z_k}^{m+3/2})$ is com-

puted replacing in (4)–(6) the numerical flux by the following expression

$$\begin{pmatrix} \dfrac{-n_y}{Z^+ + Z^-}\left(Z^+[\tilde{H}_z^{m+1/2}] - \alpha\left(n_x\dfrac{[\tilde{E}_y^m]+[\tilde{\tilde{E}}_y^{m+1}]}{2} - n_y\dfrac{[\tilde{E}_x^m]+[\tilde{\tilde{E}}_x^{m+1}]}{2}\right)\right) \\ \dfrac{n_x}{Z^+ + Z^-}\left(Z^+[\tilde{H}_z^{m+1/2}] - \alpha\left(n_x\dfrac{[\tilde{E}_y^m]+[\tilde{\tilde{E}}_y^{m+1}]}{2} - n_y\dfrac{[\tilde{E}_x^m]+[\tilde{\tilde{E}}_x^{m+1}]}{2}\right)\right) \\ \dfrac{1}{Y^+ + Y^-}\left(Y^+\left(n_x[\tilde{E}_y^{m+1}] - n_y[\tilde{E}_x^{m+1}]\right) - \alpha\dfrac{[\tilde{H}_z^{m+1/2}]+[\tilde{\tilde{H}}_z^{m+3/2}]}{2}\right) \end{pmatrix}.$$

3 Numerical Results

To illustrate the theoretical results, we consider the model problem (1) defined in the square $\Omega = (-1,1)^2$, complemented by initial conditions and Silver-Müller absorbing boundary conditions (3). In order to make it easier to find examples with known exact solution and consequently with the possibility to compute the error of the numerical solution, source terms were introduced. The simulation time is fixed at $T = 1$ and in all tests we assume that $\mu = 1$ and ϵ is given by (2), with $\epsilon_{xx} = 4x^2 + y^2 + 1$, $\epsilon_{yy} = x^2 + 1$ and $\epsilon_{xy} = \epsilon_{yx} = \sqrt{x^2 + y^2}$.

To illustrate the order of convergence in space, we fix $\Delta t = 10^{-5}$, except for $N = 4$ where we consider $\Delta t = 10^{-6}$. We plot of the error depending on the maximum element diameter for each mesh, where both the vertical and horizontal axis are scaled logarithmically. The numerical convergence rate is approximated by the slope of the linear regression line. In Fig. 2 we plot the discrete L^2-error of the $\tilde{E}_x$ component of electric field, considering different degrees for the polynomial approximation while the spatial mesh is refined for both central and upwind fluxes.

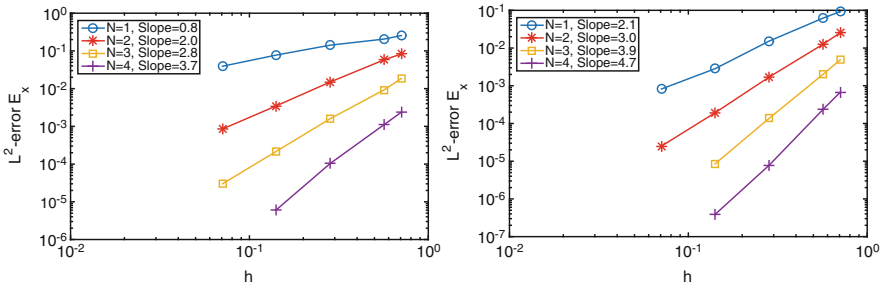


Fig. 2 $\|E_x^M - \tilde{E}_x^M\|_{L^2(\Omega)}$ *versus* h. *Left*: central flux; *Right*: upwind flux

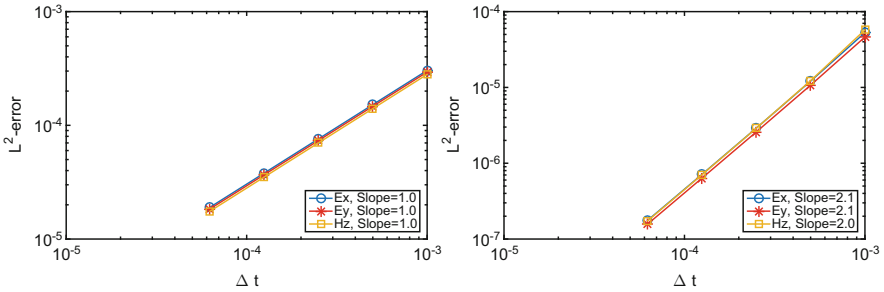


Fig. 3 $\|E_x^M - \tilde{E}_x^M\|_{L^2(\Omega)}$, $\|E_y^M - \tilde{E}_y^M\|_{L^2(\Omega)}$ and $\|H_z^M - \tilde{H}_z^M\|_{L^2(\Omega)}$ *versus* Δt. *Left*: Leap-frog; *Right*: Predictor-corrector leap-frog

For central flux, the numerical convergence rate is close to the value estimated in Theorem 1, $\mathcal{O}(h^N)$, and for upwind flux we observe higher order of convergence, up to $\mathcal{O}(h^{N+1})$ in some cases. Similar results were obtained for $\tilde{E}_y$ and $\tilde{H}_z$.

To visualize the convergence in time, the polynomials degree and the number of elements have been set to $N = 8$ and $K = 800$, respectively. The results plotted in Fig. 3 illustrate the first order convergency in time for the leap-frog method and that the second order is recovered when the predictor-corrector method is considered. These results correspond to upwind fluxes.

4 Conclusions

We have presented a leap-frog DG method for time dependent Maxwell's equations in anisotropic media considering central and upwind fluxes in the numerical scheme. The numerical results support the convergence estimates given in Theorem 1. A rigorous proof of this theorem will be presented in a forthcoming work.

Acknowledgements This work was partially supported by the Centre for Mathematics of the University of Coimbra—UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MCTES and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020; by the Portuguese Government through the BD grant SFRH/BD/51860/2012; and by the Fundação para a Ciência e a Tecnologia, I.P..

References

1. Araújo, A., Barbeiro, S., Ghalati, M.K.: Stability of a leap-frog discontinuous Galerkin method for time-domain Maxwell's equations in anisotropic materials. ArXiv:1607.06425 [math.NA] (2016)
2. Fezoui, L., Lanteri, S., Lohrengel, S., Piperno, S.: Convergence and stability of a discontinuous Galerkin time-domain method for the 3D heterogeneous Maxwell's equations on unstructured meshes. ESAIM: Math. Model. Numer. Anal. **39**(6), 1149–1176 (2005)

3. González, J.A.: A discontinuous Galerkin finite element method for the time-domain solution of Maxwell equations. PhD thesis, Universidad de Granada (2014)
4. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications. Springer, New York (2008)
5. König, M., Busch, K., Niegemann, J.: The discontinuous Galerkin time-domain method for Maxwell's equations with anisotropic materials. Photonics Nanostruct. Fundam. Appl. **8**(4), 303–309 (2010)
6. Li, J., Waters, J.W., Machorro, E.A.: An implicit leap-frog discontinuous Galerkin method for the time-domain Maxwell's equations in metamaterials. Comput. Methods Appl. Mech. Eng. **223–224**, 43–54 (2012)
7. Limeres, J., Calvo, M.L., Enoch, J.M., Lakshminarayanan, V.: Light scattering by an array of birefringent optical waveguides: theoretical foundations. J. Opt. Soc. Am. B **20**(7), 1542–1549 (2003)
8. Lu, T., Zhang, P., Cai, W.: Discontinuous Galerkin methods for dispersive and lossy Maxwell's equations and PML boundary conditions. J. Comput. Phys. **200**(2), 549–580 (2004)
9. Santos, M., Araújo, A., Barbeiro, S., Caramelo, F., Correia, A., Marques, M.I., Pinto, L., Serranho, P., Bernardes, R., Morgado, M.: Simulation of cellular changes on optical coherence tomography of human retina. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 8147–8150 (2015)
10. Schliesser, J.A., Gallimore, G., Kunjukunju, N., Sabates, N.R., Koulen, P., Sabates, F.N.: Clinical application of optical coherence tomography in combination with functional diagnostics: advantages and limitations for diagnosis and assessment of therapy outcome in central serous chorioretinopathy. Clin. Ophthalmol. (Auckland, N.Z.) **8**, 2337–2345 (2014)
11. Serranho, P., Morgado, M., Bernardes, R.: Optical coherence tomography: a concept review. In: Optical Coherence Tomography: A Clinical and Technical Update, pp. 139–156. Springer, Berlin (2012)

Topics in Magnetic Force Theory: Some Avatars of the Helmholtz Formula

Alain Bossavit

Abstract There is a great variety of formulas purporting to describe the force field inside magnetized matter. They don't always agree, which is puzzling. Starting from the Korteweg–Helmholtz formula, obtained thanks to the highly reliable Virtual Power Principle (VPP), we show how variant expressions can result from algebraic manipulations that assume, without making this explicit, extra physical hypotheses. Those we discuss here, assuming a $B = \mu H$ magnetic law, are (1) Dependence of μ on density, (2) Incompressibility of the magnetic medium in which the magnetic forces develop.

1 When Permeability Depends on Density

In eddy-current theory (where displacement currents and Coulomb forces are neglected), the force density is $J \times B + f$, where J stands for the current density and f is what is called here the 'magnetic force'. Let us assume that magnetic induction B and magnetic field strength H are related by $B = \mu H$, where permeability μ may depend on the spatial position and on the density ρ of the magnetic material. The classic Helmholtz formula [6] is then (with many notational abuses, corrected below):

$$f = -{}^1\!/_2\, |H|^2\, \nabla\mu + \nabla[\, \rho\, (\partial_\rho \mu)\, |H|^2/2], \tag{1}$$

to be understood in the sense of distributions (because of discontinuities of μ at material interfaces). This can be proved by using the VPP (see e.g. [2] for such a proof), provided one knows the magnetic energy density $\psi_{\text{mag}}(x, b, \beta)$ as a function of position x, of the ambient induction b at this point, and of the matter's 'distortion' β (the linear map that describes how material points and material vectors are placed in physical space as the effect of a virtual motion).

A. Bossavit (✉)
GeePs (ex. LGEP), Centrale-Supelec, Gif-sur-Yvette, France
e-mail: bossavit@lgep.supelec.fr

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_19

1.1 Is the Very Definition of Force Ambiguous?

A first difficulty, at this stage, is that ψ_{mag} is only a part of the total energy density $\psi(x, b, \beta)$ of the contraption, the other part being (let's assume elastic behavior, for definiteness), $\psi_{\text{ela}}(x, \beta) = \psi(x, 0, \beta) = \psi(x, b, \beta) - \psi_{\text{mag}}(x, b, \beta)$. This splitting of the total energy density may look natural (the idea is that elastic behavior is insensitive to the magnetic situation), but one should be aware that a different splitting (an example is given in [3]) would lead to a different definition of f. So there are limits to the *objectivity* of the magnetic force concept: Force is well defined only *after* a magnetic compartment and an elastic one have been delimited in the magneto-elastic system under study.

Barring variant splittings is not enough, however, to account for the bewildering variety of force formulas. For instance, one often finds formulas like this (with a constant C):

$$f = C\nabla(|B|^2/2), \tag{2}$$

which features a remarkable (and attractive, in many respects) relation between force and field. (In the case of a homogeneous material, the force would be proportional to the gradient of magnetic energy.) We'll see that such formulas *can* be derived from (1), but only under some restrictive assumptions that should not be overlooked. Otherwise, (2) is misleading.

1.2 Deriving the Helmholtz Formula

Let us concentrate on the case of a homogeneous isotropic material for which the reluctivity $\nu(x)$ at point x is supposed to only depend on the density $\rho(x)$ of the material in question at this point. (Admittedly, this is a bit academic, since linear isotropic B–H laws are only crude approximations for most materials, but results can easily be transposed to electrostatics, where $D = \epsilon E$ is much more common. Since the polarization $P = D - \epsilon_0 E$ plausibly depends on the density of available dipoles, a dependence of ϵ on density is to be expected.) The reluctivity of the vacuum is denoted by ν_0.

To be definite, suppose that a body made of this material is plunged into the magnetic field generated by some distant DC source. Other mechanical forces (for instance, gravity and whatever reaction force is needed to counteract it) may act on this matter, and the result of their action combined with magnetic body forces is some reference configuration for the material, characterized by a reluctivity field ν_{ref} (points where $\nu_{\text{ref}}(x) \neq \nu_0$ are inside the body, points where $\nu_{\text{ref}}(x) = \nu_0$ are outside) and by the density field ρ_{ref} (with $\rho_{\text{ref}}(x) > 0$ inside, $\rho_{\text{ref}}(x) = 0$ outside).

Being interested in magnetic body forces in *that* configuration, nothing more, we may ignore the stress–strain characteristic, as well as its dependence on the

ambient field, provided the part of the coupled law that obviously matters, namely the relation between reluctivity and density, is known. Let us therefore introduce a function $\tilde{\nu}$ from the real-positive half-line to itself, with the following interpretation: $\tilde{\nu}(\rho)$ is the reluctivity of a chunk of our material when its state of strain confers the density ρ to it. (In the reference state, therefore, $\nu_{\mathrm{ref}}(x) = \tilde{\nu}(\rho_{\mathrm{ref}}(x))$.) We shall have use for the first and second derivative of $\tilde{\nu}$, denoted by $\partial_\rho \tilde{\nu}$ and $\partial_{\rho\rho}\tilde{\nu}$.

Let u be a smooth velocity field, compactly supported. Consider the virtual motion $x \to t\,u(x)$ and denote by $\nu_u(t, y)$ the reluctivity at point y and time t during this motion. (By $\nu_u(t)$ we shall mean the scalar *field* $y \to \nu_u(t, y)$, whose time-derivative we denote by $\partial_t \nu_u(t)$. Its spatial derivative will be $\nabla \nu_u(t)$.) If the reluctivity didn't depend on density, one would have $\nu_u(t, x + t\,u(x)) = \nu_u(0, x) = \nu_{\mathrm{ref}}(x)$. But from time 0 to time t, a small piece of matter surrounding x will increase its volume by a factor $1 + t\,(\mathrm{div}\,u)(x)$, or equivalently, the density $\rho_{\mathrm{ref}}(x)$ at time 0 will become $\rho_{\mathrm{ref}}(x)/[1+t\,(\mathrm{div}\,u)(x)]$ at time t (all this up to terms in t^2). Accordingly, the reluctivity of this particle will be $\tilde{\nu}(\rho_{\mathrm{ref}}(x)/[1 + t\,(\mathrm{div}\,u)(x)])$, so we shall have

$$\nu_u(t, x + t\,u(x)) \simeq \tilde{\nu}\left(\frac{\rho_{\mathrm{ref}}(x)}{1 + t\,(\mathrm{div}\,u)(x)}\right). \tag{3}$$

Now, the virtual power is minus the t-derivative, at $t = 0$, of the magnetic energy $W(t) = \int \nu_u(t)|B|^2/2$, where B is the induction field in the reference state, i.e., at time 0. (Remember that we are talking of the *partial* derivative of $\Psi(t, B)$ with respect to t, so that B doesn't change.) Therefore, $\int u \cdot f = -\int \partial_t \nu_u(0)|B|^2/2$. So we must access $\partial_t \nu_u(0)$, the derivative with respect to time, taken at time 0, of the reluctivity as it evolves under the virtual motion.

This is done by differentiating (3) with respect to t. The derivative of the left-hand side is, thanks to the chain rule,

$$\partial_t \nu_u(t, x) + u(x) \cdot \nabla \nu_u(t, x).$$

The derivative of the right-hand side is

$$-\partial_\rho \tilde{\nu}(\rho_{\mathrm{ref}}(x))\,\rho_{\mathrm{ref}}(x)\,(\mathrm{div}\,u)(x)/[1 + t\,(\mathrm{div}\,u)(x)]^2.$$

For $t = 0$ we find, getting rid of t and x,

$$-\partial_t \nu_u = u \cdot \nabla \nu_u + \partial_\rho \tilde{\nu}(\rho_{\mathrm{ref}})\,\rho_{\mathrm{ref}}\,\mathrm{div}\,u,$$

hence, after an integration by parts, and using $\nu_u(0) = \nu_{\mathrm{ref}} = \tilde{\nu}(\rho_{\mathrm{ref}})$,

$$-\int \partial_t \nu_u |B|^2/2 = \int u \cdot [{}^1\!/_2\,|B|^2\,\nabla(\tilde{\nu}(\rho_{\mathrm{ref}})) - \nabla(\rho_{\mathrm{ref}}\,(\partial_\rho \tilde{\nu})(\rho_{\mathrm{ref}})\,|B|^2/2)]$$

at $t = 0$. Since the left-hand side is $\int u \cdot f$, we finally obtain :

$$f = {}^1\!/_2\,|B|^2\,\nabla[\tilde{\nu}(\rho_{\mathrm{ref}})] - \nabla[\rho_{\mathrm{ref}}\,(\partial_\rho \tilde{\nu})(\rho_{\mathrm{ref}})\,|B|^2/2]. \tag{4}$$

In terms of H and μ, and with $\tilde{\mu}$ defined by $\tilde{\mu}(\rho) = 1/\tilde{\nu}(\rho)$, this would read

$$f = -{}^1\!/_2\, |H|^2\, \nabla[\tilde{\mu}(\rho_{\text{ref}})] + \nabla[\rho_{\text{ref}}\, (\partial_\rho \tilde{\mu})(\rho_{\text{ref}})\, |H|^2/2], \tag{5}$$

the same result as in (1), up to small notational differences. This is the classic Helmholtz expression. (Let's recall that the gradients in (4) and (5) are to be understood as distributions, as discussed in [4].)

We have kept in (4) and (5) a few square brackets that may appear superfluous. They help to parse correctly. Compare for instance $\nabla(\tilde{\nu}(\rho_{\text{ref}}))$ and $\nabla\tilde{\nu}(\rho_{\text{ref}})$. In the process of parsing $\nabla\tilde{\nu}(\rho_{\text{ref}})$, one might construe $\nabla\tilde{\nu}$ as the spatial gradient of some field $\tilde{\nu}$, which would make no sense: $\tilde{\nu}$ is a mere real-to-real function, not a scalar field. On the other hand, $\nabla(\tilde{\nu}(\rho_{\text{ref}}))$ is unambiguously the spatial gradient of $\tilde{\nu}(\rho_{\text{ref}})$, which *is* a scalar field, being the composition of the real-to-real function $\tilde{\nu}$ with the scalar field ρ_{ref}. These remarks should be enough to keep us alert (and make notational abuses feasible) in what comes now.

1.3 Transforming the Helmholtz Formula

Noticing that

$$\nabla[\tilde{\nu}(\rho_{\text{ref}}) - \rho_{\text{ref}}\, (\partial_\rho \tilde{\nu})(\rho_{\text{ref}})] = -\rho_{\text{ref}}\, (\partial_{\rho\rho} \tilde{\nu})(\rho_{\text{ref}})\, \nabla\rho_{\text{ref}},$$

we transform (4) this way (cf. [8, 9]):

$$\begin{aligned} f &= {}^1\!/_2\, |B|^2\, \nabla[\tilde{\nu}(\rho_{\text{ref}}) - \rho_{\text{ref}}\, (\partial_\rho \tilde{\nu})(\rho_{\text{ref}})] - \rho_{\text{ref}}\, (\partial_\rho \tilde{\nu})(\rho_{\text{ref}})\, \nabla(|B|^2/2) \\ &= -{}^1\!/_2\, [|B|^2 \rho_{\text{ref}}\, (\partial_{\rho\rho} \tilde{\nu})(\rho_{\text{ref}})\, \nabla\rho_{\text{ref}} + \rho_{\text{ref}}\, (\partial_\rho \tilde{\nu})(\rho_{\text{ref}})\, \nabla(|B|^2)], \end{aligned}$$

which calls attention on the first part, proportional to $(\partial_{\rho\rho} \tilde{\nu})(\rho_{\text{ref}})$, of the magnetic force. The second derivative of $\tilde{\nu}$ has no reason to vanish, so we find no advantage in this formula with respect to (4), except when $(\partial_{\rho\rho} \tilde{\nu})(\rho_{\text{ref}})$ is small enough to justify the following approximation:

$$f \simeq -{}^1\!/_2\, \rho_{\text{ref}}\, (\partial_\rho \tilde{\nu})(\rho_{\text{ref}})\, \nabla(|B|^2), \tag{6}$$

in which one may see some heuristic value [7]. Observe however that if the reference density ρ_{ref} is uniform, (6) says just what (4) says.

The interest of this algebraic manipulation, which thus appears limited, may lie elsewhere. As said already, determining the function $\tilde{\nu}$ is not a matter of theorization, but of measurement. Such measurements require some a priori model for the behavior of $\tilde{\nu}$, with adjustable parameters that will be identified by an appropriate experiment. It's legitimate to make this model as simple as possible: a linear function $\tilde{\nu}(\rho) = \alpha - \beta\rho$, where α and β are the parameters to determine. (We expect β to

be positive in the case of a paramagnetic material, hence the choice of sign.) Then $\partial_{\rho\rho}\tilde{\nu} = 0$ and

$$f = \beta\, \rho_{\text{ref}} \nabla(|B|^2/2), \tag{7}$$

according to (6). One gets α and β by measuring the magnetic susceptibility for a series of values of the density, a difficult enough task for not to insist on determining the second derivative of $\tilde{\nu}$. Then (7) is as accurate as measurements allow, and simple.

Let us now lift the homogeneity restriction,[1] when magnetized matter in the system is made of several distinct materials, or why not, of a continuous infinity of them. This is appropriate when the function $\tilde{\nu}$ depends on temperature, say, but also in more contrived situations. Imagine for instance a composite material, made by mixing two homogeneous ones characterized by their own characteristics $\tilde{\nu}_1$ and $\tilde{\nu}_2$, and besides, mixed in different proportions at different locations. In such a case, we may as well imagine a continuum of possible materials, indexed by a label m that takes its values in some set $\mathscr{M}$, in such a way that an $\mathscr{M}$-valued function $x \to m(x)$ describes how space is filled with these materials (one of them being the air or the vacuum). We introduce now a function $\tilde{\nu}(m, \rho)$ to describe the magnetic behavior of this variety of materials. As before, we are interested in the force field in the reference configuration $\{m_{\text{ref}}, \rho_{\text{ref}}\}$, in which the material present at point x is $m_{\text{ref}}(x)$ with the density $\rho_{\text{ref}}(x)$.

The sequel is a mere repetition: Instead of (3), we have

$$\nu_u(t, x + t\,u(x)) \simeq \tilde{\nu}(m_{\text{ref}}(x), \rho_{\text{ref}}(x)/[1 + t\,(\operatorname{div} u)(x)]), \tag{8}$$

we still want to access $\partial_t \nu_u(0)$, and the derivative of the right-hand side of (8) at $t = 0$ is still $-\partial_\rho \tilde{\nu}(m_{\text{ref}}, \rho_{\text{ref}})\, \rho_{\text{ref}} \operatorname{div} u$, so there is little change *in fine* with respect to (4). We find

$$f = {}^1\!/_2\, |B|^2\, \nabla[\tilde{\nu}(m_{\text{ref}}, \rho_{\text{ref}})] - \nabla[\, \rho_{\text{ref}}\, (\partial_\rho \tilde{\nu})(m_{\text{ref}}, \rho_{\text{ref}})\, |B|^2/2]. \tag{9}$$

But the buck stops there: Owing to the dependence of $\tilde{\nu}$ on m, new terms of the form $\partial_m \tilde{\nu}\, \nabla\rho$ or $\rho\, \partial_{\rho m} \tilde{\nu}\, \nabla\rho$ appear, that spoil the simplicity of (6). We are finally much better off with the following simplified version of (9):

$$f = {}^1\!/_2\, |B|^2\, \nabla\nu - \nabla[\, \rho\, \partial_\rho \nu\, |B|^2/2], \tag{10}$$

where both the dependence of ν on m and ρ and the notational abuse (ν for the quite different $\tilde{\nu}$) are understood.

[1] We shall not attempt to deal with anisotropy here. For this, one should make $\tilde{\nu}$, no longer a scalar but a symmetric 3×3 matrix, a function of the symmetrized gradient $\nabla_S u$ instead of $\operatorname{div} u$, subject to a condition of invariance with respect to rotations that comply with material frame indifference.

Thus, the proportionality of force with the gradient of magnetic energy, as observed in (7) for a homogeneous material, doesn't hold in general.

2 When the Medium is Incompressible: Kelvin Force

Another example is about the so-called Kelvin force, $f = \mu_0 M \cdot \nabla H$, when the B–H law is presented as $B = \mu_0(H + M)$. (Recall that $M \cdot \nabla H$ and $\nabla H \cdot M$ stand for, in coordinates with respect to an orthonormal basis and using Einstein's convention, $M^j \partial_j H^i$ and $M^j \partial_i H^j$ respectively. One has $M \cdot \nabla H - \nabla H \cdot M = (\mathrm{rot}\, H) \times M$.)

Kelvin's force is commonplace in the study of ferrofluids [10]. Here, its derivation from the Helmholtz force relies on an additional hypothesis, namely the incompressibility of the fluid. First, suppose $\mathrm{rot}\, H = 0$ everywhere. Then $M \cdot \nabla H = \nabla H \cdot M$ (the so-called 'Maxwell force'). Comparing this with the Helmholtz expression, and assuming $\partial_\rho \mu = 0$ there, for the sake of simplicity, one finds that 'Kelvin' − 'Helmholtz' = $\nabla((H \cdot M)/2)$, a gradient. If the ferrofluid is incompressible, one may apply the VPP to only those virtual velocities u such that $\mathrm{div}\, u = 0$, and by integration by parts, the virtual power associated with the Kelvin and the Helmholtz expressions are then the same. So using Kelvin doesn't make a difference as regards the final solution of the coupled problem. Yet the 'Kelvin force' *is not the real force* in that case.

2.1 Deriving the Kelvin Formula

Let's develop what precedes. A typical reference is [5], by Engel and Friedrichs.

These authors take $B = \mu_0(H + M)$ as B–H law, without being more specific about M, and they postulate a "Maxwell tensor", $T^{ij} = H^i B^j - {}^1\!/_2\, \mu_0 |H|^2 \delta^{ij}$ in coordinates ($H \otimes B - {}^1\!/_2\, \mu_0 |H|^2 \delta$ in compact form). Suppose $\mathrm{rot}\, H = 0$. Then, with "$\widehat{=}$" for "defined as",

$$f^i \mathrel{\widehat{=}} \partial_j T^{ij} = \mu_0 (M^j \partial_j) H^i \equiv \mu_0 M \cdot \nabla H,$$

the Kelvin force.

Proof If $\mathrm{rot}\, H = 0$, then $H \cdot \nabla H = \nabla H \cdot H = {}^1\!/_2\, \nabla(|H|^2)$. Since $\mathrm{div}\, B = 0$, one has $\mathrm{div}(H \otimes B) = B \cdot \nabla H$, hence

$$\mathrm{div}(H \otimes B) = \mu_0 M \cdot \nabla H + \mu_0 H \cdot \nabla H = \mu_0 M \cdot \nabla H + {}^1\!/_2\, \mu_0 \nabla(|H|^2),$$

showing that (remark that $\mathrm{div}(g\delta) = \nabla g$, for a scalar g)

$$\mathrm{div}[H \otimes B - {}^1\!/_2\, \mu_0 |H|^2 \delta] = \mu_0\, M \cdot \nabla H. \qquad \square$$

But no reason is given, in the cited paper, to consider the field f thus defined as "the force". It seems that, for the authors, "the (classical) Maxwell tensor comes first", then comes the force as its divergence, which is, arguably, putting things heels over head. The fact that M could be just anything, here, is enough to raise suspicion. Indeed, in case $M = \chi H$, so that $B = \mu H$ with $\mu = (1 + \chi)\mu_0$, we know that the force[2] is not given by $\mu_0 M \cdot \nabla H \equiv \mu_0 \chi H \cdot \nabla H$ but by $-{}^1\!/_2\, |H|^2 \nabla \mu \equiv -{}^1\!/_2\, \mu_0 |H|^2 \nabla \chi$. We didn't use $\mathrm{rot}\, H = 0$, however, which seems to be a standard assumption in ferrofluid studies. If so, $\mu_0 \chi H \cdot \nabla H = \mu_0 \chi \nabla H \cdot H = {}^1\!/_2\, \mu_0 \chi \nabla(|H|^2)$, to be compared with $-{}^1\!/_2\, \mu_0 |H|^2 \nabla \chi$. These expressions are quite different, but *their integrals* over all space, and more generally, over any domain containing the support of χ, *are the same,* as a simple integration by parts will show. The Kelvin force, therefore, correctly predicts the total force over a magnetic body, even though its local value can differ considerably from that of the Helmholtz force (which we know is correct, if one ignores magnetostrictive effects such as the dependence of μ on local strain). The difference is (up to the factor $\mu_0/2$) the gradient of $\chi |H|^2$, which of course cannot be neglected.

This expression we just derived for the Kelvin force when $\mathrm{rot}\, H = 0$, namely ${}^1\!/_2\, \mu_0 \chi \nabla(|H|^2)$, seems to be more popular than the Ur one, $\mu_0\, M \cdot \nabla H$. This is probably due to the physicist's intuition that each of these magnetic particles ferrofluids are made of behaves like a magnetic dipole[3]: The factor $\chi(x)$ can be seen as a measure of the density of particles at point x, and they all tend to move towards regions of greater magnetic field [1], which makes the $\nabla(|H|^2)$ term plausible. (Up to a point, for the force on such a dipole of moment m, immersed in a field H, is $m \cdot \nabla H$: no square of $|H|$ there yet. To make it appear, consider that the dipoles tend to align with the ambient field, hence a proportionality between M and H, but let's be wary of such reasonings: Passing from a description of the behavior of individual magnetic particles to a law about their collective behavior is a homogenization problem of some magnitude, perhaps not really solved so far.)

References

1. Bailey, R.L.: Lesser known applications of ferrofluids. J. Magn. Magn. Mater. **39**, 178–182 (1983)
2. Bossavit, A.: On forces in magnetic matter. IEEE Trans. Magn. **50**, 229–232 (2014)
3. Bossavit, A.: Discrete magneto-elasticity: a geometrical approach. IEEE Trans. Magn. **46**, 3485–3491 (2010)
4. Bossavit, A.: Bulk forces and interface forces in assemblies of magnetized pieces of matter. IEEE Trans. Magn. **52**, 7003504 (2016)
5. Engel, A., Friedrichs, R.: On the electromagnetic force on a polarizable body. Am. J. Phys. **70**, 428–432 (2002)

[2] Which derives from the tensor $H \otimes B - {}^1\!/_2\, \mu\, |H|^2 \delta$, not from $H \otimes B - {}^1\!/_2\, \mu_0\, |H|^2 \delta$.

[3] One made of two close magnetic charges of opposite signs, not of a small current loop.

6. von Helmholtz, H.: Über die auf das innere Magnetisch oder dielectrisch Polarisirter Körper wirkender Kräfte. Ann. d. Phys. und Chem. **13**, 385–406 (1881)
7. Heyrendt, L.: Études des actions des forces magnétiques volumiques créées par un champ magnétique intense sur des fluides à seuil (Thesis, University of Lorraine). http://www.theses.fr/2012LORR0224. (2012). Accessed 25 Nov 2016
8. Liu, M.: Range of validity for the Kelvin force. Phys. Rev. Lett. **84**, 2762 (2000)
9. Luo, W., Du, T., Huang, J.: Novel convective instabilities in a magnetic field. Phys. Rev. Lett. **82**, 4134–4137 (1999)
10. Odenbach, S., Liu, M.: Invalidation of the Kelvin force in ferrofluids. Phys. Rev. Lett. **86**, 1615–1631 (2001)

Minisymposium: Computational Methods for Finance and Energy Markets

E. Jan W. ter Maten and Matthias Ehrhardt

Description

This minisymposium is an activity of the ECMI Special Interest Group on Computational Finance [2]. The SIG was launched at ECMI-2014 in Taormina (June 9–13, 2014) and (together with the ITN STRIKE network [3]) organized several sessions of a minisymposium in Computational Finance. The corresponding reporting can be found in [1]. At ECMI-2016 we brought again together twelve speakers.

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 of this minisymposium was to present most recent developments of effective and robust numerical schemes for solving linear and nonlinear problems arising from the mathematical theory of pricing financial derivatives and related financial products. These approaches vary in departing directly from the system of stochastic differential equations (involving SABR dynamics) to approaches for the derived partial differential equations. The SDE group focusses on fast Monte-Carlo methods involving COS methods and techniques for GPU processors. The PDE group discussed efficient finite difference methods (varying in involving techniques like sparse time integrators by alternating direction methods, or boundary element methods), also combined with pseudospectral methods. Further speakers concisely discussed aspects like Credit Value Adjustment, Counterparty Credit Risk and effects due to uncertainty in parameters.

In the recent years we observe an increasing interest in mathematical methods for energy markets as well. The rapid changes in energy trading within the last two

E.J.W. ter Maten (✉) • M. Ehrhardt
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: terMaten@math.uni-wuppertal.de; Ehrhardt@math.uni-wuppertal.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_20

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. Existing modelling tools and numerical methods for this application field face several new kinds of challenges.

Coupled with highly inelastic demand and a variety of supply side constraints, the lack of energy storage can result in sudden price spikes and high, time-varying volatility. Mean reversion rates and typical seasonal patterns exhibit a complex multi-scale nature with respect to the time variable. Calibration is an essential ingredient to obtain realistic models. The models need to have a detailed specification with the various quantiles being related to multiple factors through coefficients which have dynamic properties themselves related to some of the exogenous factors.

The talks included in the minisymposium were the following:

- Álvaro Leitao, TU Delft and CWI-Amsterdam (the Netherlands). *On an efficient one and multiple time-step Monte Carlo simulation of the SABR model*.
- Qiang Feng, CWI-Amsterdam (the Netherlands). *Credit Value Adjustment, Wrong Way Risk and Bermudan Options*.
- Mark Cummins, DCU Business School, Dublin City University (Ireland). *Model Risk in Gas Storage Valuation: Joint Calibration-Estimation Risk Measurement*.
- María Suárez-Taboada, Universidade da Coruña, A Coruña (Spain). *Uncertainty Quantification and Heston model*.
- Andreas Binder, MathConsult GmbH, Linz (Austria). *Developing a tool-chain for computational finance in a regulated universe*.
- Zaza van der Have, TU Delft and CWI-Amsterdam (the Netherlands). *The COS method for option valuation under the SABR dynamics*.
- Michael Schürle, University of St. Gallen (Switzerland). *Optimization of hydro storage systems and indifference pricing of power contracts*.
- Florentina Paraschiv, University of St. Gallen (Switzerland). *Estimation and application of fully parametric multifactor quantile regression with dynamic coefficients*.
- Jose G. López-Salas, Universidade da Coruña, A Coruña (Spain). *Parallel stratified regression Monte-Carlo scheme for BSDEs with applications in finance*.
- Christian Hendricks, Bergische Universität Wuppertal (Germany). *Hybrid finite difference / pseudospectral methods for stochastic volatility models*.
- E. Jan W. ter Maten, Bergische Universität Wuppertal (Germany). *Stochastic volatility models calibration with Model Order Reduction*.
- María del Carmen Calvo-Garrido, Universidade da Coruña, A Coruña (Spain). *PDE modeling and numerical methods for swing option pricing in electricity markets*.

References

1. Ehrhardt, M., Günther, M., Maten, E.J.W. ter: ECMI SIG on computational finance. ECMI Newsl. **56**, 20–22. http://www.mafy.lut.fi/EcmiNL/issues.php?action=viewart&ID=333 (2014). Accessed 10 Feb 2017
2. Ehrhardt, M., Maten, E.J.W. ter: SIG computational finance. ECMI Annual Report 2015, European Consortium for Mathematics in Industry, p. 50. http://www-amna.math.uni-wuppertal.de/ecmi-sig-cf/ (2016). Accessed 10 Feb 2017
3. Multi-ITN STRIKE: Novel Methods in Computational Finance. http://www.itn-strike.eu/. Accessed 10 Feb 2017

Efficient Multiple Time-Step Simulation of the SABR Model

Álvaro Leitao, Lech A. Grzelak, and Cornelis W. Oosterlee

Abstract In this work, we present a multiple time-step Monte Carlo simulation technique for pricing options under the *Stochastic Alpha Beta Rho (SABR)* model. The proposed method is an extension of the one time-step Monte Carlo method that we proposed in Leitao et al. (Appl. Math. Comput. 293: 461–479, 2017). We call it the *mSABR method*. A highly efficient method results, with many interesting and nontrivial components, like Fourier inversion for the sum of log-normals, stochastic collocation, Gumbel copula, correlation approximation, that are not yet seen in combination within a Monte Carlo simulation. The present multiple time-step Monte Carlo method is especially useful for long-term or for exotic options. This paper is a short version of an already published paper (Leitao et al. On an efficient multiple time-step Monte Carlo simulation of the SABR model. Quantitative Finance.

Á. Leitao (✉)
TU Delft, Delft Institute of Applied Mathematics, Delft, The Netherlands

CWI-Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
e-mail: A.LeitaoRodriguez@tudelft.nl

L.A. Grzelak
ING, Quantitative Analytics, Amsterdam, The Netherlands

TU Delft, Delft Institute of Applied Mathematics, Delft, The Netherlands
e-mail: L.A.Grzelak@tudelft.nl

C.W. Oosterlee
CWI-Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

TU Delft, Delft Institute of Applied Mathematics, Delft, The Netherlands
e-mail: c.w.oosterlee@cwi.nl

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_21

1 Almost Exact SABR Simulation

The SABR model [5] is defined by the following SDE system, with independent Brownian motions $\mathrm{d}\hat{W}_S(t)$ and $\mathrm{d}\hat{W}_\sigma(t)$,

$$\begin{aligned} \mathrm{d}S(t) &= \sigma(t)S^\beta(t)\left(\rho\mathrm{d}\hat{W}_\sigma(t) + \sqrt{1-\rho^2}\mathrm{d}\hat{W}_S(t)\right), \quad S(0) = S_0\exp(rT), \\ \mathrm{d}\sigma(t) &= \alpha\sigma(t)\mathrm{d}\hat{W}_\sigma(t), \qquad\qquad\qquad\qquad\quad \sigma(0) = \sigma_0. \end{aligned} \tag{1}$$

Here $S(t) = \bar{S}(t)\exp(r(T-t))$ denotes the forward price of the asset $\bar{S}(t)$, with r the interest rate, S_0 the spot price, T the maturity and $\sigma(t)$ a stochastic volatility process, with $\sigma(0) = \sigma_0$. The model parameters are $\alpha > 0$ (the volatility of volatility), $0 \leq \beta \leq 1$ (the variance elasticity) and ρ (the correlation coefficient).

Based on the work by Islah [6], an analytic approximation for the *cumulative distribution function* (CDF) of the SABR conditional process has been obtained. For some generic time interval $[s,t]$, $0 \leq s < t \leq T$, assuming $S(s) > 0$, the conditional cumulative distribution for forward $S(t)$ with an absorbing boundary at $S(t) = 0$, given $\sigma(s)$, $\sigma(t)$ and $\int_s^t \sigma^2(z)\mathrm{d}z$, is given by

$$Pr\left(S(t) \leq K | S(s) > 0, \sigma(s), \sigma(t), \int_s^t \sigma^2(z)\mathrm{d}z\right) = 1 - \chi^2(a; b, c), \tag{2}$$

where

$$a = \frac{1}{\nu(t)}\left(\frac{S(s)^{1-\beta}}{(1-\beta)} + \frac{\rho}{\alpha}(\sigma(t) - \sigma(s))\right)^2, \qquad c = \frac{K^{2(1-\beta)}}{(1-\beta)^2\nu(t)},$$

$$b = 2 - \frac{1 - 2\beta - \rho^2(1-\beta)}{(1-\beta)(1-\rho^2)}, \qquad \nu(t) = (1-\rho^2)\int_s^t \sigma^2(z)\mathrm{d}z,$$

and $\chi^2(x; \delta, \lambda)$ is the non-central chi-square cumulative distribution function. This formula is exact in the case of $\rho = 0$ and constitutes an *approximation* otherwise.

1.1 SABR Monte Carlo Simulation

To perform an "almost exact" multiple time-step Monte Carlo method for the SABR model, several steps need to be performed:

- *Simulation of the SABR volatility process, $\sigma(t)$ given $\sigma(s)$.* By Eq. (1), the stochastic volatility process of the SABR model exhibits a log-normal distribution. As well known, the solution is a *geometric Brownian motion*.
- *Simulation of the SABR integrated variance process, $\int_s^t \sigma^2(z)\mathrm{d}z | \sigma(s), \sigma(t)$.* This conditional distribution is not available in closed-form.

- *Simulation of the SABR forward price process.* The forward price $S(t)$ can be simulated by inverting the CDF in Eq. (2).

In Sect. 2.2, we propose a procedure to sample $\int_s^t \sigma^2(z)\mathrm{d}z|\sigma(s), \sigma(t)$ based on the Gumbel copula. For this, the CDF of the integrated variance given volatility, $\sigma(s)$, (as a marginal distribution) must be derived. In Sect. 2.1, a recursive technique to obtain an approximation of this CDF is presented.

2 Components of the mSABR Method

In this section, we discuss the different components of the mSABR method. Hereafter, we denote the SABR integrated variance process by $Y(s,t) := \int_s^t \sigma^2(z)\mathrm{d}z$.

2.1 CDF of $\int_s^t \sigma^2(z)\mathrm{d}z|\sigma(s)$ using the COS Method

We present a technique to approximate the CDF of $Y(s,t)|\sigma(s)$, i.e. $F_{Y|\sigma(s)}$. We will work in the log-space, so an approximated CDF of $\log Y(s,t)|\log\sigma(s)$, $F_{\log Y|\log\sigma(s)}$, will be derived. We have employed this technique in the one time-step SABR method [8], but the multiple time-step version is more involved. We approximate $Y(s,t)$ by its discrete equivalent, i.e.

$$Y(s,t) := \int_s^t \sigma^2(z)\mathrm{d}z \approx \sum_{j=1}^{M} \Delta t \sigma^2(t_j) =: \hat{Y}(s,t)$$

where M is the number of discrete time-points, $t_j = s + j\Delta t, j = 1,\ldots,M$ and $\Delta t = \frac{t-s}{M}$. $\hat{Y}(T)$ is subsequently transformed to the logarithmic domain, being $f_{\log\hat{Y}|\log\sigma(s)}$ the *probability density function* (PDF) of $\log\hat{Y}(s,t)|\log\sigma(s)$. This PDF is found by approximating the associated characteristic function, $\phi_{\log\hat{Y}|\log\sigma(s)}$, and applying a Fourier inversion procedure. Based on the work in [9], we can define a recursive procedure to recover the characteristic function of $f_{\log\hat{Y}|\log\sigma(s)}$. We start by defining the sequence,

$$R_j = \log\left(\frac{\sigma^2(t_j)}{\sigma^2(t_{j-1})}\right), \quad j = 1,\ldots,M, \tag{3}$$

where R_j is the logarithmic increment of $\sigma^2(t)$ between t_j and t_{j-1}. As the volatility process follows log-normal dynamics, the R_j are independent and identically distributed, i.e. $R_j \overset{\mathrm{d}}{=} R$. By the definition of R_j in Eq. (3), we write $\sigma^2(t_j)$ as

$$\sigma^2(t_j) = \sigma^2(t_0)\exp(R_1 + R_2 + \cdots + R_j). \tag{4}$$

At this point, a backward recursion procedure in terms of R_j will be defined by which we can recover $\phi_{\log \hat{Y}|\log\sigma(s)}(u)$. We define

$$Y_1 = R_M, \quad Y_j = R_{M+1-j} + Z_{j-1}, \quad j = 2, \ldots, M, \tag{5}$$

with $Z_j = \log(1 + \exp(Y_j))$.

By Eqs. (4) and (5), the discrete integrated variance can be expressed as

$$\hat{Y}(s,t) = \sum_{i=1}^{M} \sigma^2(t_i)\Delta t = \Delta t \sigma^2(s) \exp(Y_M). \tag{6}$$

From Eq. (6) and by applying the definition of characteristic function, we determine $\phi_{\log \hat{Y}|\log\sigma(s)}$, as follows

$$\phi_{\log \hat{Y}|\log\sigma(s)}(u) = \mathbb{E}[\exp(iu \log \hat{Y}(s,t))|\log\sigma(s)] = \exp\left(iu \log(\Delta t \sigma^2(s))\right) \phi_{Y_M}(u).$$

We have reduced the computation of $\phi_{\log \hat{Y}|\log\sigma(s)}$ to the computation of ϕ_{Y_M}. An accurate and efficient way of approximating ϕ_{Y_M} was derived in [8], which is employed also here. Once the approximation of ϕ_{Y_M}, $\hat{\phi}_{Y_M}$, has been derived, we can recover $f_{\log \hat{Y}|\log\sigma(s)}$ from $\phi_{\log \hat{Y}|\log\sigma(s)}$ by employing the COS method [3], as follows

$$f_{\log \hat{Y}|\log\sigma(s)}(x) \approx \frac{2}{\hat{b}-\hat{a}} \sum_{k=0}^{N-1}{}' C_k \cos\left((x-\hat{a})\frac{k\pi}{\hat{b}-\hat{a}}\right),$$

with

$$C_k = \Re\left(\phi_{\log \hat{Y}|\log\sigma(s)}\left(\frac{k\pi}{\hat{b}-\hat{a}}\right)\exp\left(-i\frac{\hat{a}k\pi}{\hat{b}-\hat{a}}\right)\right),$$

where N is the number of COS terms, $[\hat{a}, \hat{b}]$ is the support of $\log \hat{Y}(s,t)|\log\sigma(s)$ and the prime $'$ and $\Re$ symbols mean division of the first term in the summation by two and taking the real part of the complex-valued expressions in the brackets, respectively. $F_{\log \hat{Y}|\log\sigma(s)}$ can be obtained by integration.

The characteristic function $\phi_{\log \hat{Y}|\log\sigma(s)}$ and PDF $f_{\log \hat{Y}|\log\sigma(s)}$ need to be derived for each sample of $\log\sigma(s)$ and this makes the computation of $F_{\log \hat{Y}|\log\sigma(s)}$ (and also its subsequent inversion) very expensive. In order to overcome this drawback, an efficient sampling alternative based on Lagrange interpolation is employed here, the *Stochastic Collocation Monte Carlo sampler* (SCMC) [4]. The implementation details of SCMC method within this framework can be found in [7].

2.2 Copula-Based Simulation of $\int_s^t \sigma^2(z)\mathrm{d}z|\sigma(t),\sigma(s)$

In this section we present an algorithm for the simulation of the integrated variance given $\sigma(t)$ and $\sigma(s)$ by means of a copula. In order to obtain the joint distribution, we require the marginal distributions. In our case, the required CDFs are $F_{\log Y|\log\sigma(s)}$ and $F_{\log\sigma(t)|\log\sigma(s)}$. In Sect. 2.1, an approximated CDF of $\log Y|\log\sigma(s)$, $F_{\log\hat{Y}|\log\sigma(s)}$, has been derived. $F_{\log\sigma(t)|\log\sigma(s)}$ is known analytically.

For any copula some measure of the correlation between the marginal distributions is needed. We will employ the Pearson correlation coefficient for $\log Y(s,t)$ and $\log\sigma(t)$. Let first consider the following approximation

$$\log\int_s^t \sigma^2(z)\mathrm{d}z \approx \int_s^t \log\sigma^2(z)\mathrm{d}z = 2\int_s^t \log\sigma(z)\mathrm{d}z,$$

where the logarithm and the integral are interchanged. After some algebraic manipulations, an approximation of the Pearson correlation coefficient is obtained as

$$\mathscr{P}_{\log Y,\log\sigma(t)} \approx \frac{t^2-s^2}{2\sqrt{\left(\frac{1}{3}t^4+\frac{2}{3}ts^3-t^2s^2\right)}}. \tag{7}$$

Some copulas do not employ, in their definition, Pearson's coefficient as the correlation measure. Archimedean copulas employ the Kendall's τ rank correlation coefficient. A relation between the Pearson's and Kendall's τ coefficients is available.

In this work, we will use Archimedean Gumbel copula, based on the analysis performed in [8]. We have empirically shown that the Gumbel copula performs most accurately in this context. for a variety of SABR parameter values. The formal definition of the Archimedean Gumbel copula, considering $F_1\ldots,F_d \in [0,1]^d$ as the marginal distributions, reads

$$C_\theta(F_1\ldots,F_d) = \exp\left(-\left(\sum_{i=1}^{d}(-\log(F_i))^\theta\right)^{1/\theta}\right), \tag{8}$$

where $(-\log(\cdot))^\theta$, $\theta>0$ is the so-called *generator function* of the Gumbel copula. In order to calibrate parameter θ, a relation between θ and the rank correlation coefficient Kendall's τ can be employed, i.e. $\theta = 1/(1-\tau)$.

So, we have derived approximations for all the components required to apply our copula-based technique for the integrated variance simulation. The algorithm to sample $\int_s^t \sigma^2(z)\mathrm{d}z$ given $\sigma(t)$ and $\sigma(s)$ then consists of the following steps:

1. Determine $F_{\log\sigma(t)|\log\sigma(s)}$ (analytically).
2. Determine $F_{\log\hat{Y}|\log\sigma(s)}$ (Sect. 2.1).
3. Determine the correlation between $\log Y(s,t)$ and $\log\sigma(t)$ by Eq. (7).

4. Generate correlated uniforms, $U_{\log\sigma(t)|\log\sigma(s)}$ and $U_{\log\hat{Y}|\log\sigma(s)}$ by Eq. (8).
5. From $U_{\log\sigma(t)|\log\sigma(s)}$, invert the $F_{\log\sigma(t)|\log\sigma(s)}$ to get the samples of $\log\sigma(t)|\log\sigma(s)$.
6. From $U_{\log\hat{Y}|\log\sigma(s)}$, invert the $F_{\log\hat{Y}|\log\sigma(s)}$ to get the samples of $\log\hat{Y}|\log\sigma(s)$.
7. The samples of $\sigma(t)|\sigma(s)$ and $\int_s^t \sigma^2(z)\mathrm{d}z|\sigma(t),\sigma(s)$ are obtained by taking exponentials.

2.3 Simulation of $S(t)$ given $S(s)$, $\sigma(s)$, $\sigma(t)$ and $\int_s^t \sigma^2(z)\mathrm{d}z$

We complete the mSABR method by the conditional sampling of $S(t)$. The most commonly used techniques can be classified in two categories: direct inversion of the SABR distribution function given in Eq. (2) and moment-matching approaches. The direct inversion procedure has a higher computational cost because of the evaluation of the non-central χ^2 distribution. However some recent developments make this computation affordable. In [2], the authors proposed a forward asset simulation based on a combination of moment-matching (Quadratic Gaussian) and enhanced direct inversion procedures. We employ this technique also here.

As already pointed out by Andersen [1], the almost exact simulation of the asset price in some stochastic volatility models can result in loss of the martingale property, due to the approximation of a continuous process by its discrete equivalent. This is especially seen when the size of time-step is large and for certain configurations of the SABR parameters, like small β values, close-to-zero initial asset values S_0, high *vol-vol* parameter α or large initial volatility σ_0. We employ for the mSABR method a simple but effective numerical martingale correction, as follows

$$S(t) = S(t) - \frac{1}{n}\sum_{i=1}^{n} S_i(t) + \mathbb{E}[S(t)] = S(t) - \frac{1}{n}\sum_{i=1}^{n} S_i(t) + S_0,$$

where $S_i(t)$ represents the $i-$th Monte Carlo sample.

3 Numerical Experiments

In this section, we benchmark the mSABR method by pricing options under the SABR dynamics. We will consider the following model configuration: $S_0 = 0.5$, $\sigma_0 = 0.5$, $\alpha = 0.4$, $\beta = 0.5$, $\rho = 0$ and $T = 4$. Regarding the mSABR parameters, we set $M = 3000$ and $N = 150$. The number of employed Monte Carlo paths is 1.000.000. The strikes $K_i \in \{0.37, 0.41, 0.45, 0.5, 0.55, 0.61, 0.67\}$

We first perform a convergence study in terms of the number of time steps. In Table 1, the *implied volatilities* obtained by the mSABR method for several choices of the number of time steps, m, are presented. We observe by the errors in basis

Table 1 Implied volatility, increasing m

Strikes	K_1	K_2	K_3	K_4	K_5	K_6	K_7
Reference	73.34%	71.73%	70.17%	N/A	67.23%	65.87%	64.59%
$m = T/4$	73.13%	71.75%	70.41%	69.11%	67.85%	66.64%	65.48%
Error(bp)	−21.51	2.54	24.38	N/A	61.71	76.66	89.26
$m = T$	73.25%	71.67%	70.14%	68.66%	67.24%	65.89%	64.62%
Error(bp)	−9.56	−5.93	−2.79	N/A	0.92	2.21	3.17
$m = 4T$	73.34%	71.73%	70.18%	68.67%	67.24%	65.87%	64.58%
Error(bp)	0.15	0.58	0.78	N/A	0.43	0.04	−0.48

Table 2 Pricing barrier options with mSABR. Prices ×100

Strikes	K_1	K_2	K_3	K_4	K_5	K_6	K_7
	$B = 1.2$						
MC	6.2868	5.4577	4.6330	3.8263	3.0551	2.3358	1.6869
mSABR	6.3264	5.4890	4.6566	3.8436	3.0677	2.3452	1.6953
MSE	5.3196×10^{-8}						

points (bp) that our technique converges very rapidly in m. This implies that the mSABR method is much faster than the Taylor-based Monte Carlo methods.

A pricing experiment of barrier options with the mSABR method is now carried out. The *up-and-out* call option is considered here, with barrier level, $B > S_0$. As a reference, a Monte Carlo method with very fine Euler time discretization is employed. The number of time-steps for the mSABR scheme is set to $m = 12T$. In Table 2, the obtained prices and the *mean squared error* (MSE) are presented.

4 Conclusions

We have proposed an accurate and robust multiple time-step Monte Carlo method to simulate the SABR model. A copula methodology to generate samples from the conditional SABR integrated variance process has been introduced. The marginal distribution of the integrated variance has been derived by employing Fourier techniques. We have dealt with an increasing computational complexity, due to the multiple time-steps and the almost exact simulation, by applying an interpolation based on stochastic collocation. This has resulted in an efficient SABR simulation scheme. We have numerically shown the convergence, where the mSABR method requires only very few time steps to achieve high accuracy. As a multiple time-step method, it can be employed to price path-dependent options.

Acknowledgements Supported by the EU 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. Andersen, L.B.G.: Efficient simulation of the Heston stochastic volatility model. J. Comput. Finance **11**, 1–22 (2008)
2. Chen, B., Oosterlee, C.W., van der Weide, H.: A low-bias simulation scheme for the SABR stochastic volatility model. Int. J. Theor. Appl. Finance **15**, 1250016-1–1250016-37 (2012)
3. Fang, F., Oosterlee, C.W.: A novel pricing method for European options based on Fourier-cosine series expansions. SIAM J. Sci. Comput. **31**, 826–848 (2008)
4. Grzelak, L.A., Witteveen, J.A.S., Suárez-Taboada, M., Oosterlee, C.W.: The stochastic collocation Monte Carlo sampler: highly efficient sampling from "expensive" distributions. Available at SSRN. https://ssrn.com/abstract=2529691 (2015)
5. Hagan, P.S., Kumar, D., Lesniewski, A.S., Woodward, D.E.: Managing smile risk. Wilmott Mag. January, 84–108 (2002)
6. Islah, O.: Solving SABR in exact form and unifying it with LIBOR market model. Available at SSRN. http://ssrn.com/abstract=1489428 (2009)
7. Leitao, A., Grzelak, L.A., Oosterlee, C.W.: On an efficient multiple time-step Monte Carlo simulation of the SABR model. Quantitative Finance. http://dx.doi.org/10.1080/14697688.2017.1301676
8. Leitao, A., Grzelak, L.A., Oosterlee, C.W.: On a one time-step Monte Carlo simulation approach of the SABR model: application to European options. Appl. Math. Comput. **293**, 461–479 (2017)
9. Zhang, B., Oosterlee, C.W.: Efficient pricing of European-style Asian options under exponential Lévy processes based on Fourier cosine expansions. SIAM J. Financ. Math. **4**, 399–426 (2013)

Uncertainty Quantification and Heston Model

María Suárez-Taboada, Jeroen A.S. Witteveen, Lech A. Grzelak, and Cornelis W. Oosterlee

Abstract In this paper, we study the impact of the parameters involved in Heston model by means of Uncertainty Quantification. The Stochastic Collocation Method already used for example in computational fluid dynamics, has been applied throughout this work in order to compute the propagation of the uncertainty from the parameters of the model to the output. The well-known Heston model is considered introduced and parameters involved in the Feller condition are taken as uncertain due to their important influence on the output. Numerical results where the Feller condition is satisfied or not are shown as well as a numerical example with real market data.

1 Introduction

In general, stochastic differential equations governing the prices of certain financial products do not always have an analytical solution. The required numerical approximations used do not contain enough information about the reliability of the output of a certain mathematical model. Furthermore, randomness can be present in different inputs such as different model parameters, which consequently introduces uncertainty into the solution of the model. These issues make mathematical models in computational finance a good framework to apply Uncertainty Quantification. Stochastic volatility models are the focus of this work. Our objective is to study the propagation of the uncertainty from the input towards its solution.

In many financial markets as equity and interest rate, the so-called volatility smile appears as an important feature of pricing models, being the reflection of

J.A.S. Witteveen

M. Suárez-Taboada (✉)
Department of Mathematics, University of A Coruña, A Coruña, Spain
e-mail: maria.suarez3@udc.es

L.A. Grzelak • C.W. Oosterlee
Delft University of Technology, DIAM, Delft, Netherlands
e-mail: L.A.Grzelak@tudelft.nl; c.w.oosterlee@cwi.nl

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_22

the no constant behaviour in market volatilities for options at different strike prices with a single expiration date. It is well-known that the classical Black-Scholes framework is not able to reproduce the volatility smile observed in the market, due to the constant volatility assumed. Stochastic volatility models are an important breakthrough in financial research to avoid this drawback. Among these more sophisticated models, the Heston model [5] has become one of the most used for practitioners being highly efficient due to its easy computational implementation by means of numerical methods involving Fourier transforms [2, 3] and it will be the focus of this work.

As long as the Heston model contains open parameters, a calibration procedure is important to fit these parameters to market data. That is, prices obtained with the corresponding mathematical model (in this case, Heston) should match the ones observed in the market. More precisely, if we have a set of dates and market prices, a suitable calibration method will provide a set of parameters at each date considered. Once we have these data, their distributions can be established. This will make pricing more reliable because we consider a distribution for the parameters. Thus, a general vision on their uncertainty as well as their impact in the pricing could be obtained.

This paper is organized as follows: Firstly, the mathematical framework for Uncertainty Quantification is introduced as well as the method considered for computing the propagation of uncertainty, the Stochastic Collocation Method. Next, some numerical results for test where the Feller condition is satisfied or not. Finally, one with real market data will be showed.

2 Mathematical Framework

Throughout this section, some aspects and concepts about Uncertainty Quantification (UQ) will be described following the notation used by [6, 8]. In order to study the propagation of the uncertainty from the model parameters to the output, the nonintrusive Stochastic Collocation Method [10] is used as it has the attractive advantage of treating an existing solver as a black box.

2.1 Stochastic Collocation Model

First of all, we assume a complete probability space $(\Omega, \mathscr{F}, \mathbb{Q})$ and finite time horizon [0,T] where Ω is the set of all possible realizations of the stochastic process between 0 and T. The information structure is represented by an augmented filtration $\bar{\mathscr{F}}_t, t \in [0, T]$ with $\mathscr{F}_T$ the sigma algebra of distinguishable events at time T, and $\mathbb{Q}$ is the risk-neutral probability measure on elements of $\mathscr{F}$.

Next, a set $\omega = \{\omega_1, \cdots, \omega_{n_\xi}\} \in \Omega \subset \mathbb{R}^{n_\xi}$ of events in the complete probability space and a set $\boldsymbol{\xi} = \{\xi^{(1)}, \cdots, \xi^{(n_\xi)}\}$ of parameters involved in the mathematical

model are considered with parameter space $\Xi \subset \mathbb{R}^n$. For every sample $\boldsymbol{\xi}(\omega)$ the output should be computed by a suitable numerical solver. In general, a nonintrusive uncertainty quantification method consists of a sampling method and a standard interpolation method such as Lagrange interpolation. The sampling method selects the sampling points and returns the sampled values, being computed by solving the problem considered for the realization of the random parameter vector. It is interesting that the computational cost of the sample interpolation in UQ is almost depreciable compared to the cost of expensive existing solvers.

Let us suppose that our problem can be formulated as follows

$$\mathscr{L}u(\boldsymbol{x}, t, \boldsymbol{\xi}) = f(\boldsymbol{x}, t), \quad (\boldsymbol{x}, t) \in X \times (0, T), \quad T > 0 \tag{1}$$

with appropriate initial and boundary conditions. Next, $u(\boldsymbol{x}, t, \boldsymbol{\xi})$ is considered as the output of interest such as the price of a financial product with maturity T, which does not only depend on the spatial coordinates $\boldsymbol{x} \in X \subset R^{n_X}$ with $n_X \in \{1, 2, 3\}$ and time $t \in \mathbb{R}^+$, but also on a vector of n_ξ random parameters $\boldsymbol{\xi} = \{\xi^{(1)}, \cdots, \xi^{(n_\xi)}\} \in \Xi$. The random parameters $\boldsymbol{\xi}$ have a known joint probability density function $f_\xi(\boldsymbol{\xi})$ with finite variance. Each uncertain input parameter corresponds to one additional random parameter $\xi^{(i)}$ and to a dimension in the associated probability space Ω.

The stochastic collocation method is based on one-dimensional Lagrange polynomial interpolation of the output $u(\boldsymbol{x}, t, \xi_j)$ at N Gauss quadrature points ξ_j in Ξ

$$u(\boldsymbol{x}, t, \xi)^i = \sum_{j=1}^{N} u_j(\boldsymbol{x}, t)^i L_j(\xi), \quad L_j(\xi) = \prod_{k=1, k\neq j}^{N} \frac{\xi - \xi_k}{\xi_j - \xi_k}, \tag{2}$$

where $u_j(\boldsymbol{x}, t) = u(\boldsymbol{x}, t, \xi_j)$ are the solutions of the governing equations for ξ_j, and $L_j(\xi)$ are the Lagrange polynomials of degree $N - 1$ for which holds $L_j(\xi_k)) = \delta_{jk}$. Then the main objectives of UQ are the probability distribution and the resulting approximation of the statistical moments of the output distribution.

The stochastic collocation method can be extended to multiple inputs using a tensor product of quadrature points. However, if the number of uncertain parameters increases, approximations based on tensor product grids suffer from the curse of dimensionality since the number of collocation points in a tensor grid grows exponentially. In these cases, we can consider sparse tensor product spaces as first proposed by Smolyak [9].

3 Numerical Results

In the literature several numerical methods have been used to solve the mathematical models stated for pricing financial products such as finite differences [1], finite elements [7] or Monte Carlo [4]. We consider a class of highly efficient numerical

pricing techniques, the class of Fourier-based numerical integration methods, being the COS-method one of them [2, 3].

If we focus on pricing a financial instrument, such a put option, for every sample of the parameters involved in the mathematical model governing the price of the product, we will obtain a value given by the numerical solver chosen. Then, the propagation of the uncertainty from the model parameters to the output is dominated by the computational cost of the numerical solver. Due to this, it is important to choose an efficient numerical method with a low computational cost. In this work, put options have been considered.

3.1 Test: Mixed Feller Condition

An important feature of the Heston model is the stochastic volatility which allows to reproduce the implied volatility smile present in many financial markets. Every parameter has a specific effect on the implied volatility curve generated by the dynamics so its interesting to study UQ for the implied volatility as well as for the prices of certain financial products.

If $2\kappa\bar{\upsilon} > \gamma^2$, the so-called Feller condition is satisfied being difficult to achieve in practice so in this work both cases will be considered. This will allow us to take into account different financial scenarios. The set of random parameters where UQ is applied to is $\xi = \{\kappa, \bar{\upsilon}, \gamma, \upsilon_0, \rho_{x,\upsilon}\}$.

As a first approach and in order to simplify the results to visualize some conclusions, only parameters κ and γ have been considered. Thus, we obtain a two dimensional problem for UQ. Tests where Feller condition is satisfied or not have been carried out, being mixed Feller condition test the one chosen to present in order to give us an insight in the impact of the variability of κ and γ on the output of the problem stated. The deterministic values for the parameters can be seen in Table 1.

The results have been computed using nine interpolation points in each direction of the domain, that is eighty-one collocations points. However, having in mind a higher dimensional problem, a Smolyak approximation has been considered in order to reduce the computational time so the number of points is reduced to twenty-nine. Although in this case the improvement is not very noticeable, an important time reduction is expected for higher dimensions, because an increase of the number of uncertain parameters turns into an increase of the number of collocation points.

In the numerical results to follow, κ will vary in [0.4,11.0] and γ in [0.05, 0.5]. Thus, the Feller condition can be either satisfied or not at sample points considered in these domains. In this case where the Feller condition is satisfied at some points of the domain considered for κ and γ, we see a highly satisfactory performance

Table 1 Deterministic parameter setting for the put option

S_0	K	τ	r	$\bar{\upsilon}$	υ_0	$\rho_{x,\upsilon}$
100	140	2.0	0.05	0.05	0.05	−0.9

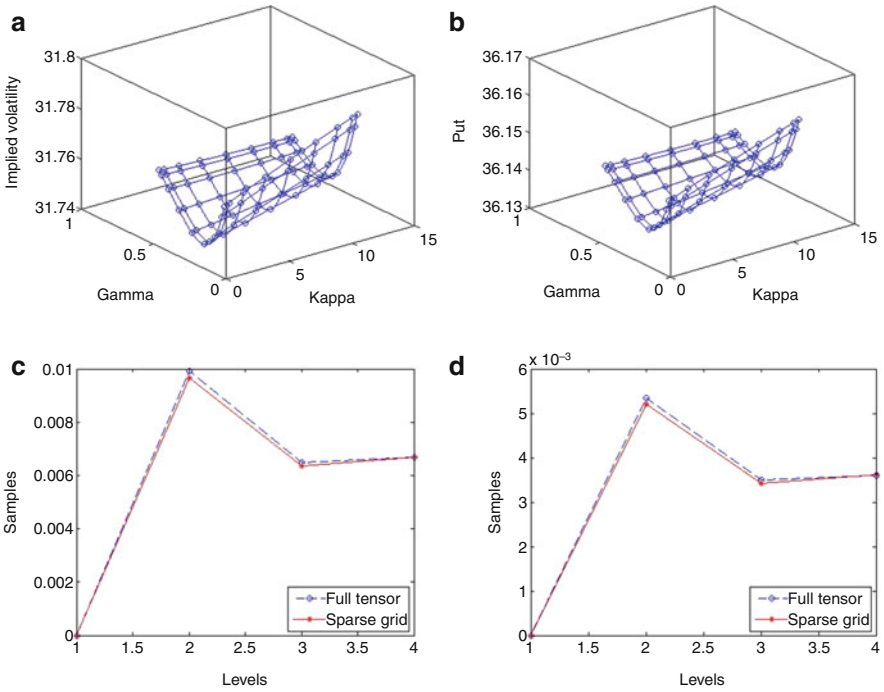


Fig. 1 Mixed Feller condition. (**a**) Implied volatility. (**b**) Put prices. (**c**) Std. Dev. implied vol. (**d**) Std. Dev. put prices

of the interpolation method in Fig. 1a,b. In Fig. 1c,d the standard deviations for the implied volatility and put prices are shown.

3.2 Numerical Test with Real Market Data

A numerical test where different values for the whole set of parameters is shown. These values are taken from real market data and some pre-processing is needed in order to apply UQ to this particular case. The calibration based on optimization over the market data results in the set of values of the parameters and the corresponding distributions. We will deal with a five- dimensional problem as five parameters are considered so the dimensionality of the problem is increased with respect to the numerical test showed in the previous section. The computational cost is about 8 minutes as the Smolyak approximation is considered for the post-processing where stochastic collocation method is involved. Throughout this test the put option deterministic parameters are the same as the ones of the previous analytical tests collected in Table 1.

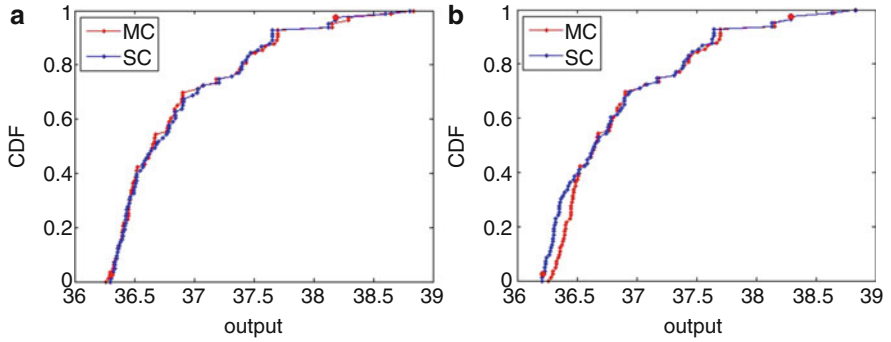


Fig. 2 Cumulative density function (full tensor and sparse grid). (**a**) $\omega = 2$ (Full tensor). (**b**) $\omega = 5$ (Sparse grid)

Table 2 Convergence sparse grid $\omega = 1, 2, 3, 4, 5$

	$\omega = 1$	$\omega = 2$	$\omega = 3$	$\omega = 4$	$\omega = 5$
Points	11	61	241	801	2433
Mean	3.7765e + 01	3.7338e + 01	3.6779e + 01	3.6926e + 01	3.6842e + 01
Std. Dev.	8.5236e−01	6.0479e − 01	6.8795e−01	5.94229e − 01	6.1646e − 01
Mean error	8.8542e − 01	4.5846e − 01	9.9622e−02	4.6040e−02	3.7463e − 02
Std. Dev. error	2.5416e − 01	6.5885e − 03	8.97506e−02	3.9754e−03	1.8257e − 02
L_1 error	9.0194e − 01	4.6129e − 01	1.4612e−01	5.5369e−02	4.9122e − 02

Monte Carlo solutions mean 3.6879e+01 and standard deviation 5.9820e−01

The cumulative probability distribution function (CDF) of the put option prices is computed by a Monte Carlo simulation of the data points and by the Stochastic Collocation method. In Fig. 2 the approximations for the full tensor for level $\omega = 2$ and the sparse grid for level $\omega = 5$ are shown to compare the results with a comparable number of points (2433 points and 3125 points, respectively).

The convergence behavior is compared in Table 2 with the Monte Carlo solution as function of the level ω and the number of points. The mean and standard deviation are integral quantities and therefore the mean and standard deviation errors can converge non-monotonically due to cancellation of errors. The L_1 error is added as a stronger convergence measure, which shows a monotonic convergence in this case.

The conclusion of this section is that the accuracy of the sparse grids interpolation technique is comparable to that of the full tensor grids as function of the number of points. The number of points for most of the levels is larger than the 84 data points in the Monte Carlo simulation. However, the number of computations in the Stochastic Collocation method is independent of the number of data points. Stochastic Collocation could therefore be faster than Monte Carlo simulation at an acceptable accuracy, in cases in which a larger set of market data points is available.

The good performance of the method is seen as both cumulative density functions are close one to each other.

4 Conclusions

The impact of the uncertainty provided by the randomness of the parameters in the Heston model is studied by means of Uncertainty Quantification. In particular, the Stochastic Collocation Method is used to study the propagation of the uncertainty from the input of the model to the solution.

It is challenging to apply techniques already used in fluid dynamics to computational finance and on the other hand, the good performance of this kind of methods is stated through the tests considered in this paper. Firstly, the analytical tests show an insight of the impact of the variability of κ and γ on the output of the problem. For mixed Feller condition, due to the large parameter range, the response is even more nonlinear and the standard deviation is largest than for the tests where Feller condition is satisfied.

Then, a real market data test is considered where the number of model parameters increases turning it into a five-dimensional problem for Uncertainty Quantification. In consequence, despite suffering from the curse of dimensionality as the number of collocation points also increases, Smolyak sparse grid Stochastic Collocation Method is used to reduce the computational cost of the method.

Acknowledgements This paper has been ERCIM "Alain Bensoussan Fellowship Programme" and partially funded by MCINN (Project MTM2010–21135–C02-01)

References

1. Achdou, I., Pironneau O.: Computational Methods for Option Pricing. SIAM, Philadelphia (2005)
2. Fang F., Oosterlee C.W.: A novel pricing method for european options based on fourier-cosine series expansions. SIAM J. Sci. Comput. **31**, 826–848 (2008)
3. Fang, F., Oosterlee C.W.: Pricing early-exercise and discrete-barrier options by Fourier-cosine series expansions. Numer. Math. **114**, 27–62 (2009)
4. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, Berlin (2003)
5. Heston, S.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)
6. Loeven, G.J.A., Witteveen Jeroen, A.S., Bijl, H.: Probabilistic collocation: an efficient nonintrusive approach for arbitrarily distributed parametric uncertainties. In: 45th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada (2007)
7. Marcozzi, M.D.: On the valuation of Asian options by variational methods. SIAM J. Sci. Comput. **24**(4), 1124–1140 (2003)
8. Nobile, F., Tempone, R., Webster, C.G.: A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Sci. Numer. Anal. **46**, 2309–2345 (2008)
9. Smolyak, S.A.: Quadrature and interpolation formulas for tensor products of certain classes of functions, Dokl. Akad. Nauk SSSR **4**, 240–243 (1963)
10. Xiu, D., Hesthaven, J.S.: High order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**, 1118–1139 (2005)

Reduced Models in Option Pricing

José P. Silva, E. Jan W. ter Maten, Michael Günther, and Matthias Ehrhardt

Abstract We consider the computational efficiency of the backward vs. forward approaches and compare these with the respective ones resulting from a parametric reduced order model, whose speed-up can be put to good use in the calibration of the underlying dynamics. We apply a global Proper Orthogonal Decomposition in the time domain to obtain the reduced basis and the Modified Craig-Sneyd ADI and Chang-Cooper schemes to numerically solve the partial differential equations. The numerical results are presented for the Black-Scholes and Heston models.

1 Introduction

In computational finance option prices can be studied by stochastic differential equations (SDEs) involving Brownian motion [1, 5, 11]. This usually leads to Monte Carlo simulation approaches. Assuming proper hedging and suitable boundary conditions the option prices can also be shown to satisfy a system of time-dependent partial differential equations (PDEs), which are our starting point. In the case of the 1D Black-Scholes model, we have the backwards PDE [11]

$$\frac{\partial u}{\partial t}(t,S) = -\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 u}{\partial S^2}(t,S) - (r-q)\,S\frac{\partial u}{\partial S}(t,S) + ru(t,S)\,, \tag{1}$$

for a call option $u = u(t,S)$, where $0 \le t \le T$, $0 \le S \le S_{\max}$ and whose terminal and boundary conditions depend on the payoff. E.g. for a call option with "strike" K we have $u(T,S) = \max(S-K,0)$, $0 \le K \le S_{\max}$. Furthermore, $\sigma^2 = \sigma_u^2$ is the variance (volatility) of u, r is the risk-free interest and q is the dividend rate.

J.P. Silva • E.J.W. ter Maten (✉) • M. Günther • M. Ehrhardt
Bergische Universität Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany
e-mail: Jose.Pedro.Moreira.da.Silva@gmail.com; termaten@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de; ehrhardt@math.uni-wuppertal.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_23

Introducing a stochastic square root variance model with a mean reverting process one we obtain again a backwards 2D-PDE, the 2D Heston model [11]

$$\frac{\partial u}{\partial t} = -\frac{1}{2}\nu S^2 \frac{\partial^2 u}{\partial S^2} - \rho\sigma\nu S \frac{\partial^2 u}{\partial \nu \partial S} - \frac{1}{2}\sigma^2 \nu \frac{\partial^2 u}{\partial \nu^2} - (r-q)S\frac{\partial u}{\partial S} - \kappa(\theta - \nu)\frac{\partial u}{\partial \nu} + ru, \tag{2}$$

for $u = u(t, S, v)$, $0 \leq S \leq S_{\max}$, $0 \leq \nu \leq \nu_{\max}$, where $v = \sigma_u^2$ follows a Cox-Ingersoll-Ross stochastic process with an own variance $\sigma^2 = \sigma_v^2$; ρ is the correlation between the Wiener processes for u and for ν. θ is the long-term mean of variance for ν. Similar as before, $u(T, S, \nu) = \max(S - K, 0)$, $0 \leq K \leq S_{\max}$.

For (1) and (2) one needs efficient time integration methods like the Modified Craig-Sneyd (MCS) ADI method [6, 7].

The time-varying density distribution of the option satisfies a dual equation (Fokker-Planck). Then the option can be obtained by a post-processing step by taking a proper integral with the density solution of the dual equation. Clearly for solving the dual PDE one is interested only in non-negative outcomes. This imposes extra conditions to time-integrators. The Chang-Cooper scheme can be guaranteed to generate non-negative solutions [8, 9].

Sensitivity is closely related to dual approaches. In finance the Dupire equation has become famous [1, 4, 5]. It expresses the sensitivity of the option prices with respect to the expiration time.

Model Order Reduction (MOR) is a technique used to derive a low-order, highly accurate approximation of the solution of a problem by projecting the original, full order problem onto a subspace of (much) smaller dimension while still capturing the dominant dynamics of the Full Order Model (FOM). We use a Global Proper Orthogonal Decomposition approach with uniform sampling to obtain our Reduced Order Model (ROM).

2 Direct Backward and Adjoint Forward Models

Equations (1) and (2) are both of the form

$$\begin{cases} \partial_t u\,(t,x) & = Lu\,(t,x) \\ u\,(T,x) & = u_T\,(x) \end{cases} \qquad (t,x) \in \mathscr{Q} \equiv [t_0, T] \times \Omega$$

and equipped with appropriate boundary conditions. Multiplying the above equation by a sufficiently regular test function p and integrating by parts over $\mathscr{Q}$, we obtain

$$\int_{\mathscr{Q}} p\,(\partial_t - L)\,u = \int_{\mathscr{Q}} u\left(-\partial_t + L^*\right)p + \int_{\Omega} pu\,\Big|_{t=t_0}^{t=T} \tag{3}$$

where L^* is the adjoint operator of L. Let us choose the test function p as the solution of the adjoint equation $\partial_t p = -L^* p$ with initial condition $p(x, t_0) = p_0(x) = \delta(x - x_0)$. Due to Lagrange's identity and as p satisfies the adjoint equation, we obtain

$$\int_\Omega p(x, T)\, u(x, T)\, dx = \int_\Omega p(x, t_0)\, u(x, t_0)\, dx = u(x_0, t_0)\,. \tag{4}$$

When we scale p such that $\int_\Omega p(T, x)\, dx = 1$, we can interpret p as a probability density function. For (1) and (2) the adjoint equations $\partial_t p = -L^* p$ have the form

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left(A_i(t, \boldsymbol{x})\, p\right) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}\left(\sum_k B_{ik}(t, \boldsymbol{x})\, B_{jk}(t, \boldsymbol{x})\, p\right). \tag{5}$$

For (1) we have

$$\boldsymbol{A} = [(r - q)\, S]\,, \quad B = \left[\sqrt{\sigma} S\right],$$

resulting in the forward PDE for $p(t, S)$

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial S^2}\left(\sigma^2 S^2 p\right) - \frac{\partial}{\partial S}\left((r - q)\, Sp\right) \tag{6}$$

with initial condition $p(0, S) = \delta(S - S_0)$ and zero-flux boundary conditions.

For (2), with $x = (S, \nu)$, we obtain

$$\boldsymbol{A} = \begin{bmatrix} (r - q)\, S \\ \kappa(\theta - \nu) \end{bmatrix}, \quad B = \begin{bmatrix} \sqrt{\nu} S & 0 \\ \sigma\sqrt{\nu}\rho & \sigma\sqrt{\nu}\sqrt{1 - \rho^2} \end{bmatrix}$$

yielding

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial S^2}\left(\nu S^2 p\right) + \frac{\partial^2}{\partial S \partial \nu}\left(\rho\sigma\nu S p\right) + \frac{1}{2}\frac{\partial^2}{\partial \nu^2}\left(\sigma^2 \nu p\right) \tag{7}$$

$$- \frac{\partial}{\partial S}\left((r - q)\, Sp\right) - \frac{\partial}{\partial \nu}\left(\kappa(\theta - \nu)\, p\right), \tag{8}$$

with initial condition

$$p(0, x) = p(0, S, \nu) = \delta(S - S_0)\, \delta(\nu - \nu_0)$$

and zero-flux (reflecting) boundary conditions. The adjoint PDEs (6) and (8) are also called Fokker-Planck equations.

Summarizing, we can calculate $u(t, x)$ directly by solving first the PDE (1) or (2) backward in time, but also by solving first the adjoint, forward PDE (6) or (8),

respectively, for $p(t,x)$, and, next, applying a postprocessing step (4). We note that in the last step quadrature is needed. The integrand is a product of a smooth function $p(T,x)$ with a piecewise linear terminal function $u(T,x)$—the last one can be interpolated exactly.

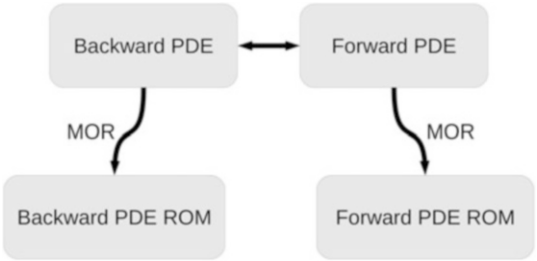
The adjoint solution $p(t,x)$ can be seen as a conditional probability $p(t,x|t_0,x_0)$. Hence, in solving the dual PDE one is interested only in non-negative outcomes. This imposes extra conditions to time-integrators [8]. The Chang-Cooper scheme can be guaranteed to generate non-negative solutions [8, 9]. Additionally, one is interested in conserving the probability in time:

$$\int_{\Omega} p\,(t,x)\,dx = 1. \tag{9}$$

If one likes to determine $u(t,x)$ for several K or T, or to determine the sensitivities of U with respect to K and T, the adjoint integration becomes favourable over the direct backwards time integration for $u(t,x)$.

Both, to the backward models as well as to the adjoint forward models MOR can be applied. We use a Global Proper Orthogonal Decomposition approach [10] together with a uniform sampling to obtain our Reduced Order Model (ROM) [12], similar to in [2].

We have then the following setup



3 Numerical Results

Due to the difference in the behaviour of the solutions of the backward and forward equations, different methods were selected for each of the two cases. In the backward equation case, plenty of methods have been proposed (see [13] and [7]) and from these, we selected the class of ADI schemes for its speed, especially in the high-dimensional cases. In particular, the ADI MCS scheme, which offers a second-order consistency in time and whose stability has been well established [6]. In the forward (adjoint) case, the conservation of probability imposes an extra condition and, accordingly, we chose the Chang-Cooper scheme with BDF2 [3, 9] , which is also second-order consistent in time.

To demonstrate the speed-up improvement, we choose to price a European Put Option with the 1D Black-Scholes model and with the Heston Model, choosing the parameters in a range of $\pm 20\%$ around the reference values in Table 1.

It is important to note that this range of parameters always *guarantees that the Feller condition is fulfilled*, i.e. $\frac{2\kappa\theta}{\sigma^2} > 1$. In the Black-Scholes case we choose $n_x = 1280$ and in the Heston Model one we choose $n_x = 64$ points per dimension for the spatial discretization, with $n_t = 40$ for the temporal discretization in both cases. This choice of temporal discretization points gives us enough snapshots to build a sufficiently large basis for the ROM. The spatial and temporal steps, h_i and k, are defined as $h_i = \frac{x_{i,max} - x_{i,min}}{n_x}$ and $k = \frac{T}{n_t}$. We evaluate the price of the option at-the-money, i.e. at $S = K$ and at the long-term mean value, $v = \theta$, with $K = 120$. For the Black-Scholes case we choose the domain $S \in [0, 4K]$ and for the Heston Model $(S, v) \in [0, 4K] \times [0, 1]$.

We collect the snapshots at every time step and at a discretized grid for the parametric domain. We discretize the parametric domain $\mathscr{P} = \prod_{i=1}^{n} [\theta_{i,min}, \theta_{i,max}]$ with a uniform grid p_k containing 5 equally distanced points per parameter, including the extreme values. We will have then 2^5 and 4^5 parameter vectors to generate our snapshots for the Black-Scholes and Heston Model, respectively. For the evaluation of the ROM accuracy, we compare the reduced and the full models at the midpoints of the parametric grid, p_l, which results for the Black-Scholes and for the Heston Model in 2^4 and 4^4 evaluations, respectively. The computational time for both models taking the FOM as reference for the speed-up, is presented in Table 2.

In Fig. 1 we observe the maximum absolute price error between the ROM and FOM forward models over all the parameter vectors corresponding to the comparison points.

$$error = \max_{\mu \in p_l} |price_{FOM}(\mu) - price_{ROM}(\mu)|$$

Table 1 Center of parameters' range

(a) Black-Scholes	
r	σ
0.03	0.3

(b) Heston Model			
r	θ	κ	σ
0.03	0.16	4	0.3

Table 2 ROM numerical speed-up

	Black-Scholes	Heston
Backward FOM	1	1
Backward ROM	3.6x	3x
Forward FOM	1	1
Forward ROM	10x	1.5x

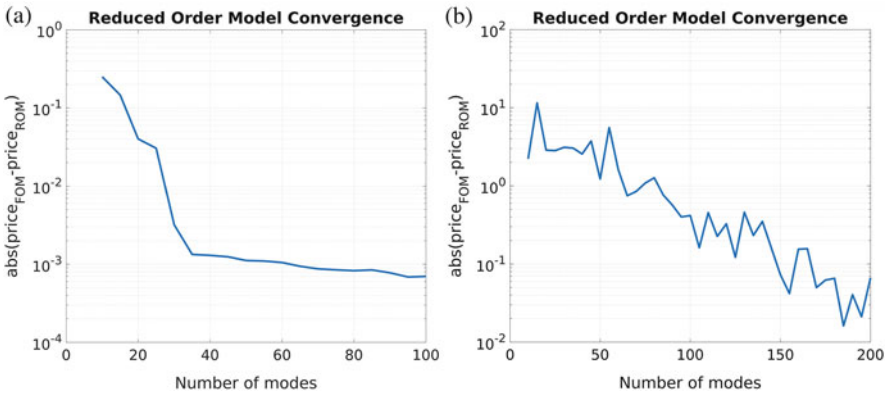


Fig. 1 Error between FOM and ROM. (**a**) Black-Scholes. (**b**) Heston

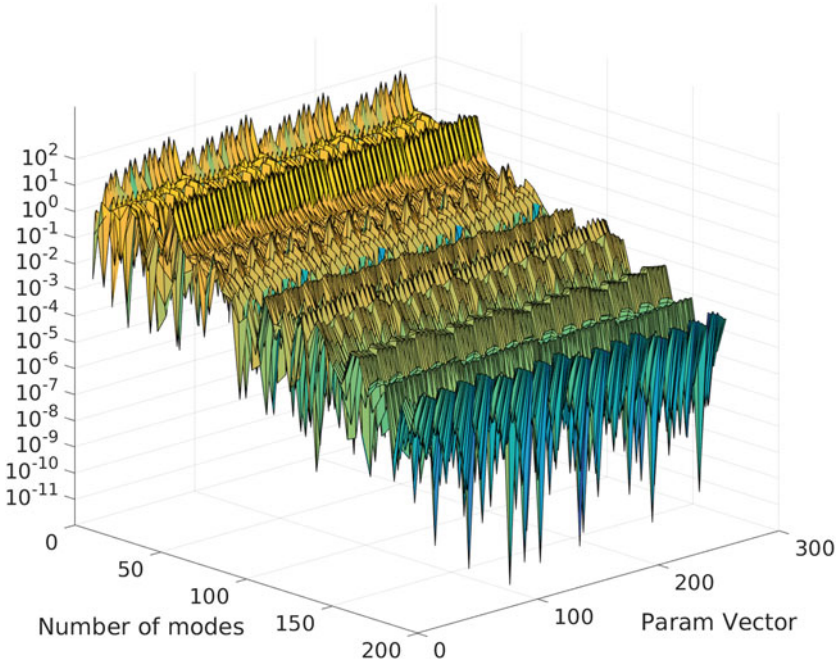


Fig. 2 Validity of the ROM across comparison points

For the Heston case, we need around 150 modes to span a subspace which approximates the FOM in the selected range of the parameters. In Fig. 2 we observe that the model shows a uniform error across the 256 vectors of parameters, proving its applicability in the whole parametric domain and not only for the parameters which were used for the snapshots.

4 Conclusion

The proposed MOR approach shows an improvement in speeding-up the time integration to solve both the backward and the forward (adjoint) PDE, which, in practice, would be cumulatively reflected in the effective calibration time. For a least-squares minimization run in Matlab® for the backward and comparable forward equations, the authors found a speed-up of $3.3\times$, $5.1\times$, and $12\times$ for the backward ROM, forward FOM and forward ROM, with 20 option prices to calibrate. Note that those are speed-ups in comparison with the backward FOM, which is already a very efficient one to solve numerically with the operator splitting schemes. Despite the already reasonable improvement, we expect an even greater improvement by tuning the model in different areas such as the location of the temporal snapshots, the inclusion of the mixed derivative term, selection of training points based on a greedy approach and the use of non-uniform grids.

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, http://www.itn-strike.eu). 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–Society for Industrial and Applied Mathematics, Philadelphia (2005)
2. Balajewiecz, M., Toivanen, J.: Reduced order models for pricing American options under stochastic volatility and jump-diffusion models. Procedia Comput. Sci. **80**, 734–743 (2016)
3. Chang, J., Cooper, G.: A practical difference scheme for Fokker-Planck equations. J. Comput. Phys. **6**(1), 1–16 (1970)
4. Dupire, B.: Pricing with a smile. Risk **7**(1), 18–20 (1994)
5. Günther, M., Jüngel, A.: Finanzderivate mit MATLAB, 2nd edn. Vieweg+Teubner Verlag and Springer Fachmedien Wiesbaden GmbH, Wiesbaden (2010)
6. Haentjens, T.: ADI schemes for the efficient and stable numerical pricing of financial options via multidimensional partial differential equations. Ph.D. thesis, Universiteit Antwerpen (2013)
7. in 't Hout, K., Toivanen, J.: Application of operator splitting methods in finance. In: Glowinski, R., Osher, S.J., Yin, W. (Eds): Splitting Methods in Communication, Imaging, Science, and Engineering. Series Scientific Computation, Springer International Publishing Switzerland, 541–575 (2017)
8. Le Floc'h, F.: Positive second order finite difference methods on Fokker-Planck equations with Dirac initial data–application in finance. SSRN-id 2605160 (2015)
9. Mohammadi, M., Borzì, A.: Analysis of the Chang–Cooper discretization scheme for a class of Fokker–Planck equations. J. Numer. Math. **23**(3), 271–288 (2015)
10. Paul-Dubois-Taine, A., Amsallem, D.: An adaptive and efficient greedy procedure for the optimal training of parametric reduced-order models. Int. J. Numer. Methods Eng. **102**(5), 1262–1292 (2014)
11. Seydel, R.U.: Tools for Computational Finance. Universtext 5, Springer, London (2012)

12. Silva, J., ter Maten, E.J.W., Günther, M., Ehrhardt, M.: Proper orthogonal decomposition in option pricing: basket options and Heston model. In: G. Russo, V. Capasso, G. Nicosia, V. Romano (eds.) Progress in Industrial Mathematics at ECMI 2014. Mathematics in Industry Series, vol. 22. Springer, Berlin (2016)
13. Tavella, D., Randall, C.: Pricing Financial Instruments: The Finite Difference Method. John Wiley & Sons Inc., New York (2000)

Minisymposium: Differential Equation Models of Propagation Processes

András Bátkai and Peter L. Simon

Description

The aim of this mini symposium was to present results about modeling different propagation processes by using ODEs and PDEs. Epidemic propagation on networks was considered with focusing on the relation between the structure of the network and the qualitative behavior of the solutions of the corresponding differential equations. Old and new product formulas for evolution equations were surveyed. Applications to numerical analysis and operator theoretic properties of the evolution equation were given.

The talks included in the minisymposium were the following:

- Diana Knipl, Department of Mathematics, University College London (UK). *Rich dynamics in simple disease spread models on travel networks.*
- Gergely Röst, Institute of Mathematics, University of Szeged (Hungary). *Impact of non-Markovian recovery on network epidemics.*
- Joan Saldaña, Universitat de Girona (Spain). *Density-dependent diffusion rates and the initial phase of epidemics on heterogeneous metapopulations.*
- András Bátkai, Bergische Universität Wuppertal (Germany). *Product formulas for non-autonomous evolution equations.*
- Peter L. Simon, Institute of Mathematics, Eötvös Loránd University, Budapest Hungary. *Approximating epidemic propagation on networks by low-dimensional ODEs.*

A. Bátkai (✉)
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: batka@cs.elte.hu

P.L. Simon
Institute of Mathematics, Eötvös Loránd University, Budapest, Hungary
e-mail: simonp@cs.elte.hu

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_24

Variance of Infectious Periods and Reproduction Numbers for Network Epidemics with Non-Markovian Recovery

Gergely Röst, István Z. Kiss, and Zsolt Vizi

Abstract For a recently derived pairwise model of network epidemics with non-Markovian recovery, we prove that under some mild technical conditions on the distribution of the infectious periods, smaller variance in the recovery time leads to higher reproduction number when the mean infectious period is fixed.

1 Introduction

Networks provide a useful paradigm to incorporate contact patterns and various heterogeneities within a population [8, 9]. The basic ingredients of such models are nodes and links, usually representing individuals and the contacts between them, but they may represent also groups of individuals (such as the population at some geographic location), and the connectedness of these groups (such as transportation routes [6, 7]). In simple disease outbreak models, the status of an individual can be susceptible (S), infected (I) or recovered (R). A key parameter associated with most epidemic models is the basic reproduction number (denoted by $\mathscr{R}_0$), which denotes the expected number of secondary infections generated by a typical infected individual introduced into a fully susceptible population [2]. The reproduction number is also a threshold quantity: if $\mathscr{R}_0 < 1$ the epidemic will die out, while if $\mathscr{R}_0 > 1$ the disease will spread. Another important measure of epidemic severity is the final epidemic size, which is the total number of individuals who become

G. Röst (✉)
Bolyai Institute, University of Szeged, Szeged, Hungary
e-mail: rost@math.u-szeged.hu

I.Z. Kiss
Department of Mathematics, University of Sussex, Brighton, UK
e-mail: i.z.kiss@sussex.ac.uk

Z. Vizi
Bolyai Institute, University of Szeged, Szeged, Hungary
Robert Bosch Ltd., Engineering Centre of Budapest, Budapest, Hungary
e-mail: vizi@math.u-szeged.hu

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_25

infected during the time course of the epidemic. These two quantities are often connected via the so-called final size relation.

Pairwise models have been successfully used to approximate stochastic epidemics on networks and represent an improvement on compartmental models. The former are formulated in terms of the expected values for the number of susceptible ($[S]$), infected ($[I]$) and recovered ($[R]$) nodes, which depend on the expected values of (SS) pairs ($[SS]$) and (SI) pairs ($[SI]$). Introducing the usual notations.

- $[X](t)$ for the expected number of nodes in state X at time t,
- $[XY](t)$ for the expected number of links connecting a node in state X to another in state Y, and
- $[XYZ](t)$ for the expected number of triplets in state $X - Y - Z$,

where, $X, Y, Z \in \{S, I, R\}$, and by summing up all possible transitions, the pairwise model reads as

$$\begin{aligned}
[\dot{S}](t) &= -\tau[SI](t),\\
[\dot{I}](t) &= \tau[SI](t) - \gamma[I](t),\\
[\dot{SS}](t) &= -2\tau[SSI](t),\\
[\dot{SI}](t) &= \tau[SSI](t) - \tau[ISI](t) - \tau[SI](t) - \gamma[SI](t),
\end{aligned} \tag{1}$$

where τ is the per contact infection rate and γ is the recovery rate. Here $[S] + [I] + [R] = N$ is the total number of nodes in the network, and only those equations are listed which are necessary to derive a complete self-consistent system. The equations for links contain triplets, thus we have to break the dependence on higher order terms to obtain a closed system. The closure approximation formula $[XSY] = \frac{n-1}{n}\frac{[XS][SY]}{[S]}$, where n is the average number of links per node, leads to the self-consistent system [3]

$$\begin{aligned}
[\dot{S}](t) &= -\tau[SI](t),\\
[\dot{I}](t) &= \tau[SI](t) - \gamma[I](t),\\
[\dot{SS}](t) &= -2\tau\frac{n-1}{n}\frac{[SS](t)[SI](t)}{[S](t)},\\
[\dot{SI}](t) &= \tau\frac{n-1}{n}\left(\frac{[SS](t)[SI](t)}{[S](t)} - \frac{[SI](t)[SI](t)}{[S](t)}\right) - (\tau+\gamma)[SI](t).
\end{aligned} \tag{2}$$

Closing at the level of pairs with the approximation $[XY] = n[X]\frac{[Y]}{N}$, one obtains the so called mean-field model (or compartmental model)

$$\dot{S}(t) = -\tau\frac{n}{N}S(t)I(t), \quad \dot{I}(t) = \tau\frac{n}{N}S(t)I(t) - \gamma I(t), \tag{3}$$

with basic reproduction number

$$\mathscr{R}_0 = \frac{n}{N}\tau\mathbb{E}(\mathscr{I})S_0, \tag{4}$$

where, $\mathbb{E}(\mathscr{I}) = 1/\gamma$ is the expected infectious period. There are many results for the Markovian case [1, 3, 5], for example, the final epidemic size is given by

$$\frac{s_\infty^{\frac{1}{n}} - 1}{\frac{1}{n-1}} = \frac{n-1}{N}\frac{\tau}{\tau+\gamma}[S]_0\left(s_\infty^{\frac{n-1}{n}} - 1\right), \tag{5}$$

where $[S]_0$ is the number of susceptible individuals at time $t = 0$ and $s_\infty = [S]_\infty/[S]_0$, where $[S](\infty) = [S]_\infty$.

2 Non-Markovian Recovery

The Markovianity of the recovery process is a strong simplifying assumption. For many epidemics, the infectious period has great importance and it is measured empirically. Recently, pairwise approximations of the SIR dynamics with non-Markovian recovery have been derived, see [4, 10–12]. In the special of fixed recovery time σ, the mean-field model is given by

$$S'(t) = -\tau\frac{n}{N}S(t)I(t), \qquad I'(t) = \tau\frac{n}{N}S(t)I(t) - \tau\frac{n}{N}S(t-\sigma)I(t-\sigma), \tag{6}$$

while the pairwise model turned out to be [4]

$$\begin{aligned}
[\dot{S}](t) &= -\tau[SI](t),\\
[\dot{SS}](t) &= -2\tau\frac{n-1}{n}\frac{[SS](t)[SI](t)}{[S](t)},\\
[\dot{I}](t) &= \tau[SI](t) - \tau[SI](t-\sigma),\\
[\dot{SI}](t) &= \tau\frac{n-1}{n}\frac{[SS](t)[SI](t)}{[S](t)} - \tau\frac{n-1}{n}\frac{[SI](t)[SI](t)}{[S](t)} - \tau[SI](t)\\
&\quad -\tau\frac{n-1}{n}\frac{[SS](t-\sigma)[SI](t-\sigma)}{[S](t-\sigma)}e^{-\int_{t-\sigma}^{t}\tau\frac{n-1}{n}\frac{[SI](u)}{[S](u)}+\tau du}.
\end{aligned} \tag{7}$$

Both systems are now delay differential equations rather than ordinary differential equations, as is the case for Markovian epidemics. Considering a general distribution for the recovery period, the pairwise model can be formulated as a system of integro-differential equations [11, 12], which we omit here. In [4], the following

final epidemic size relation has been derived:

$$\frac{s_\infty^{\frac{1}{n}} - 1}{\frac{1}{n-1}} = \frac{n-1}{N}\left(1 - e^{-\tau\sigma}\right)[S]_0 \left(s_\infty^{\frac{n-1}{n}} - 1\right). \tag{8}$$

3 The Pairwise Reproduction Number and Recovery Times

In [4], a newly introduced basic reproduction-like number is defined as $\mathscr{R}_0^p = \frac{n-1}{N}\left(1 - e^{-\tau\sigma}\right)[S]_0$, which appears also in (8). It has also been shown, that for arbitrary infectious periods, the basic reproduction number of the pairwise model is

$$\mathscr{R}_0^p = \frac{n-1}{N}\left(1 - \mathscr{L}[f_{\mathscr{I}}](\tau)\right)[S]_0, \tag{9}$$

where $\mathscr{L}[\cdot]$ is the Laplace transform and $f_{\mathscr{I}}$ is the probability density function of the recovery process given by the random variable $\mathscr{I}$. Numerical tests and analytical results have both confirmed that, in general, the following implicit relation for the final epidemic size holds

$$\frac{s_\infty^{\frac{1}{n}} - 1}{\frac{1}{n-1}} = \mathscr{R}_0^p \left(s_\infty^{\frac{n-1}{n}} - 1\right) = \frac{n-1}{N}\left(1 - \mathscr{L}[f_{\mathscr{I}}](\tau)\right)[S]_0 \left(s_\infty^{\frac{n-1}{n}} - 1\right). \tag{10}$$

Notice that while $\mathscr{R}_0$ depends on the expected value only, see (4), the pairwise reproduction number (9) uses the complete density function, thus the average length of the infectious period does not determine exactly the reproduction number. As a consequence, for an epidemic we have to know as precisely as possible the shape of the distribution. We shall analyze how the basic reproduction number (9), which is not only an epidemic threshold but also determines the final size via (10), depends on the variance of the recovery time distribution. In [10], using gamma, lognormal and uniform distributions we showed that once the mean infectious period is fixed, smaller variance in the infectious period gives a higher reproduction number and consequently a more severe epidemic. Next we generalize this result without restricting ourselves to special distributions.

4 Relationship Between the Variance and the Reproduction Number

In this section we give some simple conditions which may guarantee that smaller variance induces higher pairwise reproduction number. We consider a random variable $\mathscr{I}$ corresponding to recovery times with probability density functions $f_{\mathscr{I}}(t)$, cumulative distribution function $F_{\mathscr{I}}(t) = \int_0^t f_{\mathscr{I}}(s)ds$ and we shall use

the integral function of the CDF $\mathscr{F}_{\mathscr{I}}(t) := \int_0^t F_{\mathscr{I}}(s)ds$. Clearly, $\frac{d^2}{dt^2}\mathscr{F}_{\mathscr{I}}(t) = \frac{d}{dt}F_{\mathscr{I}}(t) = f_{\mathscr{I}}(t)$. Moreover, $F_{\mathscr{I}}(0) = \mathscr{F}_{\mathscr{I}}(0) = 0$.

Theorem 1 *Consider two random variables $\mathscr{I}_1$ and $\mathscr{I}_2$ such that*

$$\mathbb{E}(\mathscr{I}_1) = \mathbb{E}(\mathscr{I}_2) < \infty, \tag{11}$$

and

$$\mathrm{Var}(\mathscr{I}_1) < \mathrm{Var}(\mathscr{I}_2) < \infty. \tag{12}$$

Assume that

$$\lim_{t\to\infty} t^3 f_{\mathscr{I}_j}(t) = 0, \quad j \in \{1, 2\}, \tag{13}$$

and for all $t > 0$,

$$\mathscr{F}_{\mathscr{I}_1}(t) \neq \mathscr{F}_{\mathscr{I}_2}(t) \tag{14}$$

holds. If $\mathscr{I}_1$ and $\mathscr{I}_2$ represent the recovery time distribution, then for the corresponding reproduction numbers the relation $\mathscr{R}^p_{0,\mathscr{I}_1} > \mathscr{R}^p_{0,\mathscr{I}_2}$ holds.

Proof Using assumption (11), we deduce

$$\begin{aligned}\int_0^\infty t\,(f_{\mathscr{I}_1}(t) - f_{\mathscr{I}_2}(t)) &= [t(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t))]_0^\infty - \int_0^\infty (F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t))dt \\ &= \lim_{t\to\infty} t(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t)) - [\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)]_0^\infty \\ &\overset{[*]}{=} -\lim_{t\to\infty} (\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)) = 0\end{aligned}$$

thus

$$\lim_{t\to\infty} (\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)) = 0. \tag{15}$$

To see [∗], i.e. $\lim_{t\to\infty} t(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t)) = 0$, we need some algebraic manipulations:

$$\begin{aligned}\lim_{t\to\infty} t(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t)) &= \lim_{t\to\infty} \frac{F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t)}{\frac{1}{t}} \overset{\text{L'H}}{=} \lim_{t\to\infty} \frac{f_{\mathscr{I}_1}(t) - f_{\mathscr{I}_2}(t)}{-\frac{1}{t^2}} \\ &= -\lim_{t\to\infty} t^2(f_{\mathscr{I}_1}(t) - f_{\mathscr{I}_2}(t) \overset{(13)}{=} 0,\end{aligned}$$

where L'H refers to the L'Hospital rule. From assumption (12), we have

$$\begin{aligned}\mathrm{Var}(\mathscr{I}_1) = \mathbb{E}(\mathscr{I}_1^2) - (\mathbb{E}(\mathscr{I}_1))^2 &< \mathbb{E}(\mathscr{I}_2^2) - (\mathbb{E}(\mathscr{I}_2))^2 = \mathrm{Var}(\mathscr{I}_2)\\ &\overset{(11)}{\Rightarrow} \mathbb{E}(\mathscr{I}_1^2) < \mathbb{E}(\mathscr{I}_2^2).\end{aligned}$$

or equivalently $\int_0^\infty t^2(f_{\mathscr{I}_1} - f_{\mathscr{I}_2})dt < 0$. We can carry out some calculation on the left-hand side of this inequality:

$$\begin{aligned}\int_0^\infty t^2(f_{\mathscr{I}_1} - f_{\mathscr{I}_2})dt &= [t^2(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t))]_0^\infty - 2\int_0^\infty t(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t))dt\\ &= \lim_{t\to\infty} t^2(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t)) - 2[t(\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t))]_0^\infty\\ &+ 2\int_0^\infty \mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)dt\\ &\overset{[**]}{=} -2\lim_{t\to\infty} t(\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)) + 2\int_0^\infty \mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)dt\\ &\overset{[**]}{=} 2\int_0^\infty \mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)dt,\end{aligned}$$

consequently

$$\int_0^\infty \mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)dt < 0 \tag{16}$$

To prove [∗∗], i.e. $\lim_{t\to\infty} t^2(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t)) = \lim_{t\to\infty} t(\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)) = 0$, we have

$$\begin{aligned}\lim_{t\to\infty} t(\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)) &= \lim_{t\to\infty} \frac{\mathscr{F}_{\mathscr{I}_1}(t) - \mathscr{F}_{\mathscr{I}_2}(t)}{\frac{1}{t}} \overset{\mathrm{L'H}}{=} \lim_{t\to\infty} \frac{F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t)}{-\frac{1}{t^2}}\\ &= -\lim_{t\to\infty} t^2(F_{\mathscr{I}_1}(t) - F_{\mathscr{I}_2}(t))\\ &\overset{\mathrm{L'H}}{=} \lim_{t\to\infty} \frac{f_{\mathscr{I}_1}(t) - f_{\mathscr{I}_2}(t)}{\frac{2}{t^3}} = \frac{1}{2}\lim_{t\to\infty} t^3(f_{\mathscr{I}_1}(t) - f_{\mathscr{I}_2}(t))\\ &\overset{(13)}{=} 0.\end{aligned}$$

Since $F_{\mathscr{I}}(t) \geq 0, t \geq 0$ and monotone increasing, the integral function of CDF $\mathscr{F}_{\mathscr{I}}(t)$ is monotone increasing and convex. Using (14) and (16), we obtain

$$\mathscr{F}_{\mathscr{I}_1}(t) < \mathscr{F}_{\mathscr{I}_2}(t), \tag{17}$$

for all $t > 0$. Clearly, for $\mathscr{R}^p_{0,\mathscr{I}_1} > \mathscr{R}^p_{0,\mathscr{I}_2}$, it is enough to prove, that $\mathscr{L}[f_{\mathscr{I}_1}](\tau) < \mathscr{L}[f_{\mathscr{I}_2}](\tau)$, i.e. $\int_0^\infty e^{-\tau t}(f_{\mathscr{I}_1}(t) - f_{\mathscr{I}_2}(t))dt < 0$. First, we perform some algebraic

manipulation on the left-hand side:

$$\begin{aligned}\int_0^\infty e^{-\tau t}(f_{\mathscr{I}_1}(t)-f_{\mathscr{I}_2}(t))dt &= [e^{-\tau t}(F_{\mathscr{I}_1}(t)-F_{\mathscr{I}_2}(t))]_0^\infty \\ &\quad +\tau\int_0^\infty e^{-\tau t}(F_{\mathscr{I}_1}(t)-F_{\mathscr{I}_2}(t))dt \\ &= \tau[e^{-\tau t}(\mathscr{F}_{\mathscr{I}_1}(t)-\mathscr{F}_{\mathscr{I}_2}(t))]_0^\infty \\ &\quad +\tau^2\int_0^\infty e^{-\tau t}(\mathscr{F}_{\mathscr{I}_1}(t)-\mathscr{F}_{\mathscr{I}_2}(t))dt \\ &\overset{(15)}{=} \tau^2\int_0^\infty e^{-\tau t}(\mathscr{F}_{\mathscr{I}_1}(t)-\mathscr{F}_{\mathscr{I}_2}(t))dt.\end{aligned}$$

In conclusion, we have

$$\tau^2\int_0^\infty e^{-\tau t}(\mathscr{F}_{\mathscr{I}_1}(t)-\mathscr{F}_{\mathscr{I}_2}(t))dt \overset{(17)}{<} 0,$$

therefore $\mathscr{L}[f_{\mathscr{I}_1}](\tau) < \mathscr{L}[f_{\mathscr{I}_2}](\tau)$, which gives $\mathscr{R}^p_{0,\mathscr{I}_1} > \mathscr{R}^p_{0,\mathscr{I}_2}$.

5 Conclusion

Our previous works already indicated that for pairwise models not only the mean, but higher order properties of the distribution of the recovery times have an impact on the outcome of the epidemic. We derived useful threshold quantities for non-Markovian recovery in [4]. In [10], we showed that for particular distribution families (typically two parameter families such as gamma, lognormal, and uniform distribution), smaller variance leads to higher reproduction number within the same family when the mean is fixed. Our new result in this study allows as to make comparisons between distributions of different kinds. To show the usefulness of Theorem 1, as an example, we consider $\mathscr{I}_1 \sim \text{Exp}(\gamma)$ and $\mathscr{I}_2 \sim \text{Fixed}\left(\frac{1}{\gamma}\right)$, i.e. $f_{\mathscr{I}_1}(t) = \gamma e^{-\gamma t}, t \geq 0$ and $f_{\mathscr{I}_2}(t) = \delta\left(t-\frac{1}{\gamma}\right)$, where $\delta(t)$ denotes the Dirac delta function. Clearly, we obtain $\mathscr{F}_{\mathscr{I}_1}(t) = t + \frac{1}{\gamma}e^{-\gamma t} - \frac{1}{\gamma}$ and $\mathscr{F}_{\mathscr{I}_2}(t) = t - \frac{1}{\gamma}$, thus there is no $t_0 > 0$, such that $\mathscr{F}_{\mathscr{I}_1}(t_0) = \mathscr{F}_{\mathscr{I}_2}(t_0)$. Since $\mathbb{E}(\mathscr{I}_1) = \mathbb{E}(\mathscr{I}_2) = \frac{1}{\gamma}$, $\frac{1}{\gamma^2} = \text{Var}(\mathscr{I}_1) > \text{Var}(\mathscr{I}_2) = 0$ and the other conditions of Theorem 1 are satisfied, we find $\mathscr{R}^p_{0,\mathscr{I}_1} < \mathscr{R}^p_{0,\mathscr{I}_2}$.

Acknowledgements Research was supported by Hungarian Scientific Research Fund OTKA K109782

References

1. Carlos, L., Juher, D., Saldaña, J.: On the early epidemic dynamics for pairwise models. J. Theor. Biol. **352**, 71–81 (2014)
2. Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J.: On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J. Math. Biol. **28**(4), 365–382 (1990)
3. Keeling, M.J.: The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. Lond. B **266**, 859–867 (1999)
4. Kiss, I.Z., Röst, G., Vizi, Z.: Generalization of pairwise models to non-Markovian epidemics on networks. Phys. Rev. Lett. **115**(7), 078701 (2015)
5. Kiss, I.Z., Miller, J.C., Simon, L.P.: Mathematics of Epidemics on Networks – From Exact to Approximate Models. Springer, Berlin (2017)
6. Knipl, D., Röst, G.: Large number of endemic equilibria for disease transmission models in patchy environment. Math. Biosci. **258**, 201–222 (2014)
7. Nakata, Y., Röst, G.: Global analysis for spread of infectious diseases via transportation networks. J. Math. Biol. **70**(6), 1411–1456 (2015)
8. Newman, M.E.J.: Spread of epidemic disease on networks. Phys. Rev. E **66**(1), 016128 (2002)
9. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., Vespignani, A.: Epidemic processes in complex networks. Rev. Mod. Phys. **87**(3), 925 (2015)
10. Röst, G., Vizi, Z., Kiss, I.Z.: Impact of non-Markovian recovery on network epidemics. In: Mondaini, R.P. (ed.) BIOMAT 2015, pp. 40–53. World Scientific (2016)
11. Röst, G., Vizi, Z., Kiss, I.Z.: Pairwise approximation for SIR type network epidemics with non-Markovian recovery. arXiv:1605.02933 (2016)
12. Wilkinson, R.R., Ball, F.G., Sharkey, K.J.: The relationships between message passing, pairwise, Kermack-McKendrick and stochastic SIR epidemic models. arXiv:1605.03555 (2016)

Minisymposium: Effective Solutions for Industry Using Mathematical Technology

José Durany, Wenceslao González, Peregrina Quintela, Jacobo de Uña, and Carlos Vázquez

Description

The main aim of this symposium was to present five of the most successful projects of ITMATI (Technological Institute of Industrial Mathematics) oriented to industrial customers, as a result of an intense relationship with our environment since our

J. Durany (✉)
Department of Applied Mathematics II, University of Vigo, 36310 Vigo, Spain

Technological Institute for Industrial Mathematics ITMATI, 15782 Santiago de Compostela, Spain
e-mail: durany@dma.uvigo.es

W. González
Department of Statistics and Operations Research, University of Santiago, 15782 Santiago de Compostela, Spain

Technological Institute for Industrial Mathematics ITMATI, Santiago de Compostela, Spain
e-mail: wenceslao.gonzalez@usc.es

P. Quintela
Department of Applied Mathematics, University of Santiago, 15782 Santiago de Compostela, Spain

Technological Institute for Industrial Mathematics ITMATI, Santiago de Compostela, Spain
e-mail: peregrina.quintela@usc.es

J. de Uña
Department of Statistics and Operations Research, University of Vigo, 36310 Vigo, Spain

Technological Institute for Industrial Mathematics ITMATI, Santiago de Compostela, Spain
e-mail: jacobo@uvigo.es

C. Vázquez
Department of Mathematics, University of A Coruña, 15071 A Coruña, Spain

Technological Institute for Industrial Mathematics ITMATI, Santiago de Compostela, Spain
e-mail: carlosv@udc.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_26

formal origin in February 2013, focusing in win-win engagements with all our stakeholders. We focused in projects developed from specific needs of companies, reason of our conviction to promote the Transference of Applied Knowledge to the Industry. Specifically,

- to facilitate and promote the application of mathematical techniques and methods in the productive sector,
- to develop new technological knowledge to improve the competitiveness of the companies in the field of technology and innovation,
- to provide services of support to business innovation,
- to spread the technological knowledge of the ITMATI,
- to boost academic and scientific collaborations with universities and centres of research and transfer and enhance interdisciplinary,
- to help strengthen the relationship between source of knowledge and business.

This minisymposium emphasized the importance of this mathematical transference applied to the resolution of industrial problems. In particular, five success stories between research groups in industrial mathematics and companies were presented. Namely:

Part I: Applied Mathematics

- A. Souto-Serantes. FerroAtlántica, Spain. *The mathematical modeling in the R & D Department of FerroAtlántica*
  Several projects have been undertaken in recent years that have led to exchanges with leading research centers and universities around the world. We have chosen as an example for this session the Electromagnetic casting furnace model (EMC) where the electromagnetic model and the thermodynamic model were done in detail. We exposed their results and the consequences of them on the industrial project.

- Svenn Anton Halvorsen. Teknova AS, Norway. *Practical industrial mathematics between industry and academia.*
  Metallurgical Scale-up project included general thermal and electric scale-up, with application for electrodes and slag furnaces. The examples given in the talk showed that mathematical modelling is a very valuable tool to study metallurgical processes. Consequently, good cooperation between industry and academia, combining industrial experience, metallurgy and mathematics is mandatory.

- J. García San Luis, J.F. Rodríguez Calo. Repsol Technology Center, Spain. *Fostering successful partnerships between industry and academia. ITMATI-Repsol research center.*
  In the joint research unit Repsol-ITMATI, the company Repsol pursues a competitive advantage in dealing with different problems of mathematical programming, and tries to obtain that advantage by developing and implementing mathematical and computational methods in software tools, with immediate application in their business units. The main keys to the success of this collaboration were presented.

Part II: Statistics and Operations Research

- K. Wuerzburg, A. Casal. Inova, Spain. *Applications of smart data management and analytic systems in the maritime industry.*
  InovaLabs has successfully applied data management techniques in maritime industries, helping companies to implement energy efficient and intelligent systems. Two of these systems are a demand forecast tool for an innovative clean power supply for seagoing vessels at port, and an intelligent decision support tool to predict and analyse the impact of energy-efficient and emission reduction measures at port infrastructure.
- A. Lorenzo, J.L. Sáiz. INAER, Spain. *Mathematics in the service of emergencies.*
  INAER is an enterprise of the Babcock group that provides air services, including air services in emergencies such as wildfire, medical emergencies, search and rescue missions. Acquiring data will not suffice: the cornerstone of the progress forward will be developing the capacity for processing all these data and obtain reliable information, sometimes in real time. We can distinguish three main lines of work: process and fusion of spatial data from the surface of the Earth related to properties of the terrain or the surface of the sea, process of geolocation data in real time and, after action, statistical analysis of mission data. The main contributions coming from statistical techniques and optimization have been highlighted.

Practical Industrial Mathematics: Between Industry and Academia

Svenn Anton Halvorsen

Abstract Teknova is a small research institute located in the southern part of Norway. We collaborate closely with several metallurgical companies, among others, within industrial mathematics. Two cases from challenging, interdisciplinary projects are presented: One case where the proper non-dimensionalized equation is applied to derive the electrical regimes relevant for smelting furnaces, and one showing how a mathematical point of view largely can enhance simulations based on metallurgical background.

We have experienced that many factors are required to succeed in our projects within industrial mathematics: Trust, technical/industrial competence, broad mathematical competence, translator skills, and proper network. Our role is between industry and academia. We will not succeed without proper and extensive cooperation with 'both worlds'.

1 Introduction

Teknova is a small research institute located in the southern part of Norway. We collaborate closely with several metallurgical companies, among others, within industrial mathematics. Industrial problems are normally interdisciplinary and can be very challenging.

Cases from two major projects at Teknova will be discussed, *Metallurgical Scale-up* and *Electrical Conditions in Metal Processes (ElMet)*. The case from *Metallurgical Scale-up* demonstrates how the non-dimensionalized equation for the time-harmonic electric field can be applied to define the relevant electrical regimes, while the case from the latter project shows how a mathematical point of view largely can enhance simulations based on metallurgical background.

Both projects include mathematical expertise from University of Oxford and University of Santiago the Compostela (USC), in addition to excellent metallurgical

S.A. Halvorsen (✉)
Teknova AS, Kristiansand, Norway
e-mail: sah@teknova.no

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_27

competence from The Norwegian University of Science and Technology (NTNU) and Norwegian metallurgical companies.

2 Metallurgical Scale-Up

The project, *Metallurgical Scale-up*, ended in 2015. Here, we have mainly demonstrated how comparatively simple mathematical analysis can supply valuable scale-up information. The studies include general thermal and electric scale-up, with application to electrodes and slag furnaces [3]. The study of slag furnaces included basic mathematical analysis and numerical simulations [1, 3, 5].

For general time-harmonic electromagnetic problems, the possible regimes were derived as follows: Assume harmonic time variation, constant parameters, and apply Ohm's law. The equation for the electric field can then be written as [3, 5]:

$$\nabla^2 \mathbf{E} = i\omega\sigma_e\mu\mathbf{E} - \frac{\omega^2}{c^2}\mathbf{E} \tag{1}$$

where $\mathbf{E}$ is the electric field vector, i is the imaginary unit, ω is the angular frequency, σ_e is the electrical conductivity, μ is the magnetic permeability, and c is the speed of light. A similar equation can be derived for the magnetic field.

Following the derivation in [4], Eq. (1) is non-dimensionalized by introducing the tilde-variables $x = L\tilde{x}$, $y = L\tilde{y}$, $z = L\tilde{z}$, $\mathbf{E} = E_0\tilde{\mathbf{E}}$; where L is a characteristic length scale and E_0 is a characteristic field strength. The non-dimensional equation can be written as:

$$\tilde{\nabla}^2\tilde{\mathbf{E}} - 2i\left(\frac{L}{\delta}\right)^2\tilde{\mathbf{E}} + \left(2\pi\frac{L}{\lambda}\right)^2\tilde{\mathbf{E}} = 0 \tag{2}$$

where $\tilde{\nabla}^2$ is the non-dimensional Laplacian, δ is the skin depth defined by:

$$\delta = \sqrt{\frac{2}{\omega\sigma_e\mu}} \tag{3}$$

and λ is the electromagnetic wavelength given by:

$$\lambda = \frac{2\pi c}{\omega} \tag{4}$$

If $(2\pi L/\lambda)^2$ is sufficiently small, the last term in Eq. (2) can be neglected, and if $(L/\delta)^2$ is small, then the second term can also be dropped. Hence, there will be

different regimes for the electric field depending on the scale parameter, L:

- Electromagnetic waves, $\lambda \ll L$.
- Alternating current (AC), $\lambda \gg L$.
  - High frequency, $\delta \ll L$.
  - Low frequency, $\delta \approx L$.
- Direct current (DC), $\delta \gg L$.

For electrical smelting furnaces, $\lambda \gg L$, electromagnetic waves can safely be neglected and the third term in Eq. (2) can be dropped.

Analysing non-dimensional equations is a powerful tool to get an overview of independent model parameters and possible regimes for the problem involved. Additional studies are required to reveal the details.

Consider, for instance, the criterion for the DC approximation: δ should be much greater than the scaling parameter, L. A study of the electrical current in an insulated cylinder has revealed that *much* greater, is not required. For practical purposes, the DC approximation is reasonable to predict the current distribution even if the skin depth is as small as the cylinder radius [5]. This indicates that the DC approximation can be relevant for metallurgical applications as long as the skin depth is not significantly smaller than the appropriate scaling parameter, L. Relevant studies are recommended to reveal the practical limitation for typical cases.

3 Electrical Conditions in Metal Processes (ElMet)

The *ElMet* project started in 2015 and will continue through 2019. As the name implies, this project has focus on the electrical conditions in metallurgical processes. Research is needed to understand the effects of 3-phase alternating current, including how the associated power distribution governs the chemical reactions and temperature distribution.

Previous simulations within metallurgy have often applied direct current (DC) to study large, 3-phase, smelting furnaces. One example is Dhainaut who performed electrical 3D simulations for a typical submerged, 3-electrode, arc furnace for the production of high carbon ferromanganese [2]. Instead of performing AC simulations, Dhainaut solved the DC equations and chose

> only one instant of time where one electrode has a voltage of 100 V while the two others have a voltage of −50 V.

This corresponds to the time instance when the current is at its *maximum* in one electrode, and is evenly distributed to the two other electrodes. As the simulations are valid for “only one instant of time”, it is not straightforward to interpret the results for real furnaces. Clearly, there is too much electrical power released around one electrode, and too little around the other two.

For the simulations, Dhainaut assigned different electrical conductivities to six regions in the furnace: Mixture at 400, 800, and 1200 °C, coke beds at 1500 °C, electrodes and metal bath.

For the electrodes and the metal, the corresponding skin depths are significantly less than the electrode radius and the metal height, respectively. Hence, the DC approximation is not valid in these regions. These regions have conductivities some orders of magnitude higher than the remaining part of the furnace. They influence the current and power distribution in the furnace significantly by being such high conductivity regions, but their actual conductivity values are of minor influence in the coke beds and the charge mixture regions. The high conductivities also imply that a low fraction of the electric power is dissipated in the electrodes or the metal.

For the coke beds, the skin depth is larger than the distance between the electrodes, and for the mixture regions the skin depths are significantly larger than the furnace diameter. Hence, a DC approximation is probably a proper approximation for these regions.

In the low frequency, or DC, approximation, the electric field will be the gradient of an electric potential. Ohm's law and current conservation are applied to derive the well-known equation for the electric potential:

$$\nabla(\sigma_e \nabla V) = 0, \tag{5}$$

where V is the electric potential.

As boundary conditions for Eq. (5) we apply the electric potential (voltage) at the top (clamp position) of each electrode, and zero normal current at all other boundaries.

The electric conductivities will depend on the position (material composition) and the temperature, but not on the electric potential. Hence, Eq. (5) is *linear* with respect to V.

From a mathematical point of view, it is then natural to define three DC basis problems for Eq. (5) with the following voltages at the top of the three electrodes:

$$U_1 = 1, U_2 = 0, U_3 = 0, \tag{6}$$

$$U_1 = 0, U_2 = 1, U_3 = 0, \tag{7}$$

$$U_1 = 0, U_2 = 0, U_3 = 1. \tag{8}$$

Zero normal currents are still assumed at all other boundaries.

Let U_{1a}, U_{2a}, and U_{3a} be the amplitudes of the electric potentials for each electrode and let $\mathbf{E}_1(x, y, z)$, $\mathbf{E}_2(x, y, z)$, and $\mathbf{E}_3(x, y, z)$ be the real-valued electric fields for the respective basis problems. Then we can combine the basis solutions to get the E-field for the AC problem [4]:

$$\mathbf{E} = U_{1a}\mathbf{E}_1 + U_{2a}\mathbf{E}_2 e^{-i\frac{2}{3}\pi} + U_{3a}\mathbf{E}_3 e^{+i\frac{2}{3}\pi} \tag{9}$$

where the electrical phase shift between the electrodes have been taken into account.

Equation (9) is written in the phasor form, i.e. it should be multiplied by exp ($i\omega t$) to show the complex time variation.

Based on Eq. (9) we can now compute the current and power distribution at *any* time instance, and in addition the time averaged power distribution.

We have defined three basis problems, but it is only required to solve for two of them. Then the solution of the third one is easily computed by combining the other two. For the symmetric case, with same voltage amplitudes and same conditions around each of the electrodes, it is sufficient to solve one DC problem for one half of the furnace. The solution for the other half is found by mirroring and the solutions for basis problems 2 and 3 follows by rotation. [4]

The main computational cost consists of solving the equations involved, while post-processing is cheap. Hence, we have the following summary:

- The metallurgist applied one DC calculation to get insight into the conditions for "one instance of time".
- Mathematical insight has shown that:
  - Two DC calculations are sufficient to find the solution at *any* time instance, and in addition the time averaged power distribution.
  - For the symmetric case, only one DC calculation for half of the furnace is required.

The AC method presented in this paper ignores electromagnetic induction and is therefore an approximation for the conditions in an electric furnace. Based on our mathematical analysis, it should give a proper approximation for the current and power distributions in the region between the electrodes and the metal, but it is not suitable for details within large electrodes or in the metal [4].

Within the *ElMet* project the method will be compared with full electromagnetic computations to determine its limitations.

4 Concluding Remarks

The examples show that mathematical modelling is a very valuable tool to study metallurgical processes. It should be emphasized that the results in these projects are based on extensive cooperation. The projects are truly interdisciplinary and involves the following elements:

- Mathematical modelling (Teknova, University of Oxford, University of Santiago de Compostela)
- Process understanding/metallurgy (Norwegian University of Science and Technology - NTNU, industrial partners)
- Industrial experience (Industrial partners)

At Teknova we have experienced that many factors are required to succeed in our projects within industrial mathematics:

- *Trust*—Industrial information is essential, and the company must trust that it will not be misused.
- *Technical competence*—Within the relevant fields, to fully understand the industrial jargon.
- *Broad mathematical competence*—To do proper mathematical modelling.
- *Translator skills*—The ability to 'translate' the industrial problem to a mathematical model, and interpret the model results in a language understood within the industrial company.
- *Network*—Know who can be consulted whenever own competence needs to be supplemented.

Our role is between industry and academia. We will not succeed without proper and extensive cooperation with 'both worlds'.

Acknowledgements This paper is published as part of the project Electrical Conditions and their Process Interactions in High Temperature Metallurgical Reactors (ElMet), in short *Electrical Conditions in Metal Processes (ElMet)*, with financial support from The Research Council of Norway and the companies Alcoa, Elkem and Eramet.

The project *Metallurgical Scale-up* was funded by The Regional Research Fund Agder (RFF Agder).

References

1. Bermúdez, A., Cregan, V., Gómez, D., Rial, A., Vázquez, R., Halvorsen, S.: Electric scale-up for a slag heating furnace. In: Bermúdez, A., Manteiga, W., Quintela, P., Vilar, J. (eds.) Proceedings of the 97th European Study Group with Industry (ESGI), Santiago de Santiago de Compostela, Spain (2015). http://math-in.net/97ESGI/imagenes/97ESGI.pdf
2. Dhainaut, M.: Simulation of the electric field in a submerged arc furnace. In: Proceedings of the Tenth International Ferroalloy Congress (INFACON X), Cape Town, South Africa, pp. 605–613 (2004)
3. Halvorsen, S., Schlanbusch, R., Shinkevich, S.: Mathematical models for metallurgical scale-up. In: Proceedings of the fourteenth International Ferroalloys Congress (INFACON XIV), Kiev, Ukraine, pp. 588–597 (2015)
4. Halvorsen, S., Olsen, H., Fromreide, M.: An efficient simulation method for current and power distribution in 3-phase electrical smelting furnaces. IFAC-PapersOnLine **49**(20), 167–172 (2016). doi:http://dx.doi.org/10.1016/j.ifacol.2016.10.115. http://www.sciencedirect.com/science/article/pii/S2405896316316780; 17th Symposium on Control, Optimization and Automation in Mining, Mineral and Metal Processing (IFAC MMM), 2016 Vienna, Austria, 31 August–2 September 2016
5. Schlanbusch, R., Halvorsen, S., Shinkevich, S., Gómez, D.: Electrical scale-up of metallurgical processes. In: Proceedings of the 2014 COMSOL Conference, Cambridge, UK, pp. 7–31 (2014). https://www.comsol.com/paper/electrical-scale-up-of-metallurgical-processes-18321

Minisymposium: EU-MATHS-IN: Success Stories of Mathematical Technologies in Societal Challenges and Industry

Peregrina Quintela and Antonino Sgalambro

Description

The development of new products, production processes or improvements in the society 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.

On the other hand, the European Commission, in order to bring together resources and knowledge across different fields, technologies and disciplines, including social sciences and the humanities, has defined a challenges-based approach.

The network-of-networks named EU-MATHS-IN, sponsored by the European Mathematical Society (EMS) and the European Consortium for Mathematics in Industry (ECMI), aims to become a dedicated one-stop-shop and service unit to coordinate and facilitate the required exchanges in the field of application-driven

P. Quintela (✉)
Math-in, Spanish Network for Mathematics and Industry and University of Santiago de Compostela, c/ Lope Gómez de Marzoa, s/n. Campus Vida, 15782 Santiago de Compostela, Spain
e-mail: peregrina.quintela@math-in.net

A. Sgalambro
SM[i]2—Sportello Matematico per l'Industria Italiana, Roma, Italy

c/o Istituto per le Applicazioni del Calcolo "Mauro Picone", National Research Council of Italy, Roma, Italy
e-mail: a.sgalambro@iac.cnr.it

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_28

mathematical research and its exploitation for innovations in industry, science and society. The organization has been established in Amsterdam on Nov. 27, 2013 with a joint meeting of the stakeholders: currently, 14 national networks are members of the EU-MATHS-IN foundation.

The minisymposium was divided into three sessions, classified according to the Societal Challenges established in the EU Framework Programme for Research and Innovation H2020 and the NACE based Economic Activities classification:

- Logistics and Transport; Mechanics and Mechatronics; Agriculture and Fishing.
- Energy and Environment.
- Textiles, Clothing and Footwear; Materials; Biomedicine and Health Care; Information and Communication Technology.

The first session hosted a brief introduction on the goals and strategies of the EU-MATHS-IN, a European network of networks in industrial mathematics.

The talks included in the minisymposium were the following:

- G. di Pillo. ACTOR SRL-Spinoff of Sapienza University of Rome (Italy). *A two-objective optimization of ship itineraries for a cruise company.*
- R. Lenz. Zuse Institute Berlin (Germany). *Optimization approach for a new problem in the gas transport industry.*
- Z. Horváth. Széchenyi István University; T. Jakubík. Széchenyi István University and Audi Hungária Motors Ltd (Hungary). *Modelling, simulation and optimization of a vehicle exhaust system with simultaneous heat radiation and free convection.*
- N. Budko. DIAM, Delft University of Technology (The Netherlands). *Modelling oxygen consumption in germinating seeds.*
- R. Fontanges. Labex AMIES (France). *Wind velocity field approximation from sparse data for new wind farms installation.*
- J.F. Rodríguez Calo. Repsol Technology Center (Spain). *Automatic identification of kinetic models in industrial reaction systems.*
- W. T. Lee. University of Portsmouth (United Kingdom). *Mathematical modelling of a wave energy converter.*
- A. Lozano. Basque Center for Applied Mathematics and CIC EnerguiGUNE; B. Escribano. Basque Center for Applied Mathematics (Spain). *First principles based modeling of prospective materials for Sodium-Ion batteries.*
- E. Carrizosa. IMUS, University of Sevilla (Spain). *Optimal design of solar power tower systems.*
- N. Marheineke. FAU Erlangen-Nürnberg (Germany). *Aerodynamic web forming: process simulation and material properties.*
- R. Fontanges. Labex AMIES (France). *Control and measurements of nanomaterials by aggregation and analysis of different sources of images.*
- P. Maass. ZeTeM, Universität Bremen (Germany). *Mathematical concepts for hyperspectral imaging. Applications in digital pathology.*
- R. Benini. Strecof Dynamics srls (Rome). *The TVD project: a dynamical assessment platform for the learning evaluation.*

The following institutions were involved in the organization of this minisymposium:

- EU-Maths-IN—Stichting European Service Network of Mathematics for Industry and Innovation.
- AMIES—Agence pour les mathématiques en interaction avec l'entreprise et la société. France.
- Czech Network for Mathematics in Industry. Czech Republic.
- EU-MATHS-IN.se—Swedish Network for Mathematics in Industry. Sweden.
- KoMSO—Komitee für mathematische Modellierung, Simulation und Optimierung. Germany.
- HU-maths-in—Hungarian Service Network for Mathematics in Industry and Innovations. Hungary.
- IMNA—Industrial Mathematics Network for Austria. Austria.
- MACSI—Mathematics Applications Consortium for Science and Industry. Ireland.
- Math-in—Spanish Network for Mathematics & Industry. Spain.
- Norwegian Network of Mathematics for Industry and Innovation. Norway.
- PL-MATHS-IN—Polish Service Network for Mathematics in Industry and Innovations. Poland.
- PT-MATHS- IN—Portuguese Network for Mathematics & Industry. Portugal.
- PWN—Platform Wiskunde Nederland. Netherlands.
- SM[i]2—Sportello Matematico per l'Industria Italiana (Mathematical desk for Italian Industry). Italy.
- Smith Institute—Smith Institute for industrial mathematics and systems engineering. United Kingdom.

Modeling Oxygen Consumption in Germinating Seeds

Neil Budko, Bert van Duijn, Sander Hille, and Fred Vermolen

Abstract The consumption of oxygen by a germinating seed is assumed to be a good indicator of seed vitality and can potentially be used to predict the germination time. With the current availability of relatively simple single-seed respiration measurement methods and more oxygen consumption data opportunities emerge for detailed analysis of the underlying mechanisms relating respiration to germination processes. Due to the complex (structural and physiological) nature of seeds experimental analysis alone is very difficult. Mathematical modeling may provide an insight into the relationship between the germination of seeds and respiration. We have approached this problem by considering the population dynamics of mitochondria in seeds subject to limited oxygen supply and present a simple but rigorous and easily testable mathematical model that can handle large amounts of data and is interpretable in terms of the effective biological parameters of the seeds.

1 Introduction

When a healthy dry seed is brought in contact with water it rapidly restores its normal biological functionality and eventually germinates. Although the event of germination is poorly defined, it is commonly associated with the moment when the expanding embryo makes the radicle penetrate the seed coat [3]. Besides the germination percentage, important parameters related to germination of a seed batch are the mean germination time and the speed of germination reflecting the homogeneity of the batch. In agronomical and horticultural practice a fast and homogenous germination is desirable for many crops. Since the time of germination of seeds in seed batches often varies significantly, much effort is devoted to the selection and treatments of seeds to achieve simultaneous germination (within hours

N. Budko (✉) • F. Vermolen
Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
e-mail: n.v.budko@tudelft.nl; f.j.vermolen@tudelft.nl

B. van Duijn • S. Hille
Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands
e-mail: bert.vanduijn@fytagoras.com; shille@math.leidenuniv.nl

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_29

rather than days of each other). The rate of oxygen consumption by germinating seeds is considered to be one of the promising parameters for the monitoring of the seed's status and a candidate for predicting germination and seed vigor [4, 15].

The consumption of oxygen can nowadays be measured on the level of individual seeds and with great precision. In a typical setup a significant number of randomly selected seeds is placed in small sealed air-tight test tubes containing enough water, a specific substrate and air. The oxygen content in each test tube is monitored remotely with the help of a fluorescent-dye technique [8, 15, 16]. Such an experiment produces a large amount of oxygen partial pressure data for large numbers of individual seeds.

The germination and the associated respiration of seeds are very complex processes [2, 6, 10]. Currently oxygen consumption profiles are usually simply fitted with low-order polynomials or piecewise-linear curves and the germination behavior is predicted on the basis of certain features of these curves [4, 15]. There is an understandable dissatisfaction with this approach as the parameters of these polynomials have no clear biological meaning. In line with this, for instance, the possibility of predicting the germination time from test-tube oxygen measurements is still open to debate [3].

This paper presents a simple but rigorous and easily testable mathematical model that, on the one hand, can handle large amounts of data, while, on the other hand, is interpretable in terms of the effective biological parameters of the seed. Our model is based on ordinary differential equations (ODE) popular in biological growth and proliferation studies [11, 14] and describes the two major stages of the germination process: the initial stage of repair characterized by water uptake, the initiation of metabolism and the subsequent stage of growth with its increased metabolism. We also derive a direct mathematical relation between the seed parameters retrieved in test-tube experiments and the expected behavior of these seeds in more realistic conditions. Finally, a phase-space technique is proposed to predict the distribution of plant sizes as a function of time.

We would like to mention that this project has been initiated at the 90th European Study Group Mathematics With Industry (SWI 2013, Leiden) [5].

2 Single Seed Model

Assuming that the main consumers of oxygen in seeds are mitochondria and that all mitochondria consume oxygen at the same rate, which is proportional to the concentration of oxygen [9], we arrive at the following equation:

$$\frac{dc}{dt} = -\alpha n(c - c_{\min}), \quad c(0) = c_0, \tag{1}$$

where c denotes the oxygen partial pressure, and n is the number of mitochondria. This equation takes into account that the oxygen consumption slows down and stops when the oxygen pressure drops to a certain minimal level $c_{\min}$. In physiology this is generally referred to as the COP (critical oxygen pressure) level, i.e. the oxygen pressure at which oxygen becomes the limiting factor for the respiration rate. Further, assuming that the population of mitochondria is growing at a rate that is proportional to both the number of mitochondria and the concentration of oxygen we arrive at the second equation:

$$\frac{dn}{dt} = \beta n(c - c_{\min}), \quad n(0) = n_0. \tag{2}$$

The rate coefficients α and β are both non-negative. Expressing β as $\beta = \rho\alpha$, one can think of ρ as an oxygen to mitochondrial biomass conversion factor.

We neglect the temporal changes in α, assuming that it is a constant. However, the proliferation coefficient β does strongly depend on time, i.e. on the physiological state of the seed and its mitochondria. Indeed, although there is some argument about the rate of the biogenesis of mitochondria in germinating seeds, the common understanding is that repair needs to be done to the cells and membranes of the embryo before these cells and their mitochondria can be fully functional and replicate [1, 7, 13]. The most simple way to account for this fact is to set $\beta = 0$ for $0 \le t \le t_r$, i.e. for the initial period of metabolism, and then let $\beta > 0$ be constant for $t > t_r$, where t_r denotes the repair time.

Equations (1) and (2) admit explicit solutions for both the repair and the growth stages, with the oxygen level at time t given by:

$$c(t) = \begin{cases} c_{\min} + (c_0 - c_{\min}) \exp(-\alpha n_0 t), & 0 \le t \le t_r; \\ c_{\min} + \dfrac{(c_r - c_{\min})\,[\beta(c_r - c_{\min}) + \alpha n_0]}{\beta(c_r - c_{\min}) + \alpha n_0 \exp\left[(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)\right]}, & t > t_r, \end{cases} \tag{3}$$

where c_r is the concentration of oxygen at the end of the repair stage. This function can be used to determine the unknown parameters by fitting experimental oxygen data. Having done so one can evaluate the following function:

$$\frac{n(t)}{n_0} = \begin{cases} 1, & 0 \le t \le t_r; \\ 1 + \beta(c_r - c_{\min})/(\alpha n_0) \\ \quad - \dfrac{\beta(c_r - c_{\min})\,[\beta(c_r - c_{\min}) + \alpha n_0]\,/(\alpha n_0)}{\beta(c_r - c_{\min}) + \alpha n_0 \exp\left[(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)\right]}, & t > t_r, \end{cases} \tag{4}$$

thus obtaining the curve of the relative growth of the mitochondrial population in a closed test tube.

A currently proposed technique for estimating the so-called Relative Germination Time (RGT) amounts to finding the point of fastest decline rate in the oxygen consumption curve and extrapolating the tangential line at that point until it crosses the time axis [15].

The explicit solution (3) provides the link between the RGT and the model parameters. Obviously, the point of maximum decline in $c(t)$ will be somewhere in the growth-stage part of the graph, where $t > t_r$. Straightforward calculations show that the tangential line to the graph of $c(t)$ drawn at the point of maximum in the second derivative of $c(t)$ crosses the time axis at

$$\mathrm{RGT} = t_r + \frac{2 + \ln\left[\frac{\beta(c_r - c_{\min})}{\alpha n_0}\right]}{\beta(c_r - c_{\min}) + \alpha n_0} + \frac{4\beta c_{\min}}{[\beta(c_r - c_{\min}) + \alpha n_0]^2}. \tag{5}$$

Substitution of $t = \mathrm{RGT}$ in (4) reveals that RGT does not correspond to any particular growth factor in the mitochondrial population. Moreover, the RGT determined from the measurements in a sealed test tube, where the oxygen is gradually depleted, may significantly deviate from the open-air germination time.

Instead of using the standard RGT one could, for example, associate the test-tube germination time t_g with a fixed growth factor of the mitochondrial population, i.e., $n(t_g)/n_0 = k = \text{const}$. Such a choice would be based on the assumption that the total number of mitochondria inside the plant (embryo and radicle) is proportional to its volume and that the radicle will protrude the seed coating as soon as the mitochondrial population has grown by a certain factor.

The supply of oxygen in vivo, for example, in a standard germination test on paper, is virtually unlimited. Since $c(t) = c_0$ is no longer a limiting factor, the overall rate of both the repair and the growth processes will be different from what is observed in test tubes, which will affect the time of germination as well. However, since the fundamental parameters of the model depend on individual seeds rather than their environment, one can use the values of these parameters obtained in test-tube experiments for predicting the growth and germination of seeds in vivo.

The unlimited oxygen supply can be modeled by fixing the concentration of oxygen in (2) at its initial value $c_0 - c_{\min}$ for all $t \geq 0$. Hence, the predicted relative growth of the mitochondrial population in the open air will be given by

$$\frac{\tilde{n}(t)}{n_0} = \begin{cases} 1, & 0 \leq t \leq \tilde{t}_r; \\ \exp\left[\beta(c_0 - c_{\min})(t - \tilde{t}_r)\right], & t > \tilde{t}_r, \end{cases} \tag{6}$$

where the parameters β, c_0, and $c_{\min}$ are obtained, e.g., by data-fitting in a test-tube experiment. This exponential formula, obviously, describes only the early stages in the seed development, where other limiting factors (water supply, minerals, etc.) do not yet play any major role.

We assume that the open-air repair time $\tilde{t}_r$ featured in (6) is directly related to the test-tube t_r and in this paragraph we derive the relation. Being essentially a chemical reaction, the process of repair requires a certain amount (M mass units) of oxygen to be consumed by the embryo and the initial population of mitochondria. In a test tube with air volume V this amount of oxygen is consumed in t_r hours and equals

$$M = V\,[c_0 - c_r] = V\,[c_0 - c_{\min}]\left[1 - e^{-\alpha n_0 t_r}\right]. \tag{7}$$

In the open air during the repair process the oxygen is simply absorbed at the constant rate $\alpha n_0\,[c_0 - c_{\min}]$, and the amount of oxygen absorbed in $\tilde{t}_r$ hours inside an equivalent volume of space surrounding the seed is $\tilde{M} = V\,[c_0 - c_{\min}]\,\alpha n_0 \tilde{t}_r$. Assuming that $\tilde{M} = M$ we obtain the following relation between the test-tube and the open-air repair times:

$$\tilde{t}_r = \frac{1}{\alpha n_0}\left[1 - e^{-\alpha n_0 t_r}\right]. \tag{8}$$

Finally, defining the open-air germination time $\tilde{t}_g(k)$ as the time when the population of mitochondria has increased by the factor k, i.e., $\tilde{n}(\tilde{t}_g)/n_0 = k$, from (6) and (8) we obtain $\tilde{t}_g(k)$ explicitly as

$$\tilde{t}_g(k) = \tilde{t}_r + \frac{\ln k}{\beta[c_0 - c_{\min}]} = \frac{1}{\alpha n_0}\left[1 - e^{-\alpha n_0 t_r}\right] + \frac{\ln k}{\beta[c_0 - c_{\min}]}. \tag{9}$$

Measurements of oxygen consumption at the single-seed level were performed on barley seeds (*Hordeum vulgare* Var. Esterel 2003) with the Q2 measuring technology [16] as described previously [4, 15]. Typical oxygen consumption profiles for three barley seeds are shown in Fig. 1 (upper-left plot). The data were fitted with the present model using a nonlinear least-squares minimization algorithm. Predictions for test-tube and open-air relative growth of the mitochondrial populations are presented in the upper-right plot. Lower plots of Fig. 1 show the histograms of the reconstructed effective biological parameters αn_0, β, and t_r for a batch of barley seeds.

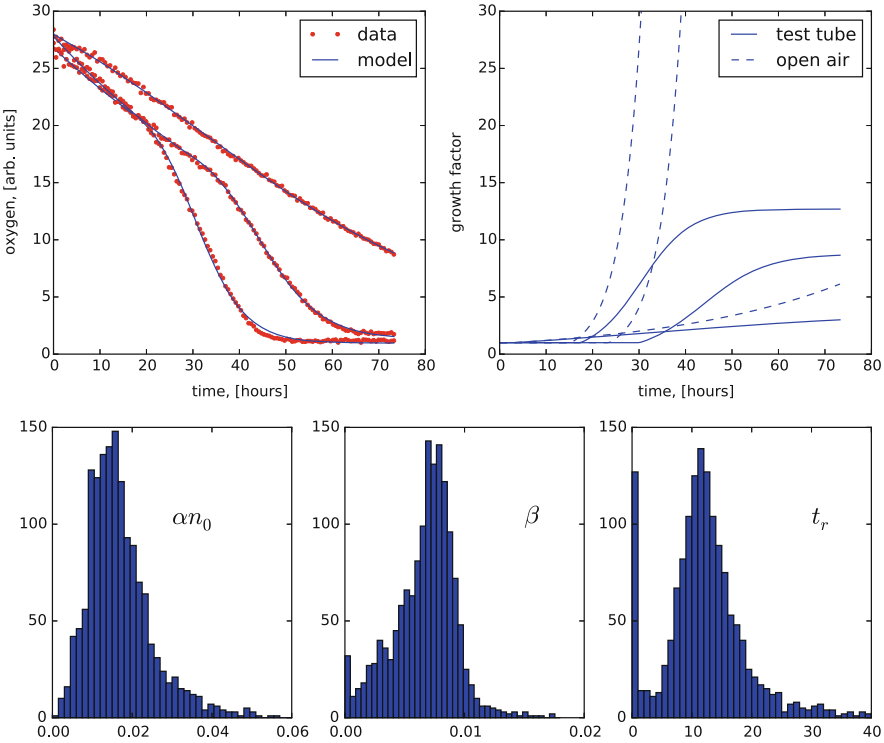


Fig. 1 Oxygen data for three barley seeds fitted with the present model (*top*, *left*); Predicted relative growth of mitochondrial populations (*top*, *right*); Distributions of biological parameters for a batch of barley seeds (*bottom*)

3 Seed Batch Model

From the practical point of view one is often interested in the distribution of germination times or plant sizes (within seeds) as a function of time. The related inverse problem is also very important: how to achieve a desired plant size and germination time distribution by seed selection and treatment strategy? To solve these problems one could, in principle, resort to extensive Monte-Carlo simulations. Here we propose an alternative analytical phase-space approach inspired by the Lifshitz-Slyozov-Wagner method [12, 17] and reminiscent of the Liouville equation of Hamiltonian dynamics.

Our main assumption is that the size of a plant during germination is proportional to the size of its mitochondrial population. Consider a simplified model for the open-air growth of the mitochondrial population in the i-th seed from a batch of P seeds all having the same repair times t_r, but possibly different initial populations

of mitochondria n_{0i} and proliferation coefficients β_i:

$$\frac{dn_i}{dt} = \beta_i(c_0 - c_{\min})n_i, \quad n_i(\tilde{t}_r) = n_{0i}, \quad i = 1, \ldots, P. \tag{10}$$

At $t = t_r$ each plant can be represented by a point with coordinates (n_{0i}, β_i) in the (n, β) phase plane. Let the local density of these points be described by a smooth distribution function $u_0(n, \beta)$ with the properties

$$\int_0^\infty \int_0^\infty u_0(n, \beta)\, dn\, d\beta = P, \quad \text{and} \quad \int_0^\infty u_0(n, \beta)\, d\beta = \Psi_0(n); \tag{11}$$

where $\Psi_0(n)$ is the distribution of sizes of initial mitochondrial populations among the seeds. For $t > t_r$ the points representing the seeds will move along individual trajectories $(n_i(t), \beta_i)$ over the phase plane. Hence, the local density of points will change, while their total number P will remain constant. The time-evolution of the distribution function $u(n, \beta, t)$ corresponding to the growth model (10) and subject to the conservation law $P = \text{const}$ is described by the following Cauchy problem:

$$\frac{\partial u}{\partial t} + \beta(c_0 - c_{\min})n\frac{\partial u}{\partial n} + \beta(c_0 - c_{\min})u = 0, \quad u(n, \beta, 0) = u_0(n, \beta). \tag{12}$$

This problem can be solved with the method of characteristics explicitly:

$$u(n, \beta, t) = e^{-\beta(c_0 - c_{\min})t} u_0\left(ne^{-\beta(c_0 - c_{\min})t}, \beta\right). \tag{13}$$

The desired distribution $\Psi(n, t)$ of plant sizes at time t can be computed by (numerically) integrating the distribution $u(n, \beta, t)$ over β. This technique can be easily extended to higher-dimensional distribution functions involving other parameters, such as t_r and α, or even time-varying environmental conditions and treatment of plants.

References

1. Arciuch, V.G.A., Elguero, M.E., Poderoso, J.J., Carreras, M.C.: Mitochondrial regulation of cell cycle and proliferation. Antioxid. Redox Signal. **16**(10), 1150–1180 (2012)
2. Bewley, J.D.: Seed germination and dormancy. Plant Cell **9**, 1055–1066 (1997)
3. Bewley, J.D., Bradford, K.J., Hilhorst, H.W.M., Nonogaki, H.: Seeds: Physiology of Development, Germination and Dormancy, 3rd edn. Springer, New York/Heidelberg/Dordrecht/London (2013)
4. Bradford, K.J., Bello, P., Fu, J.-C., Barros, M.: Single-seed respiration: a new method to assess seed quality. Seed Sci. Technol. **41**, 420–438 (2013)
5. Budko, N., Corbetta, A., van Duijn, B., Hille, S., Krehel, O., Rottschäfer, V., Wiegman, L., Zhelyazov, D.: Oxygen transport and consumption in germinating seeds. In: Proceedings of the 90th European Study Group Mathematics with Industry, Leiden, 28 January–1 February 2013, pp. 5–30

6. Cannell, M.G.R., Thornley, J.H.M.: Modelling the components of plant respiration: some guiding principles. Ann. Bot. **85**(1), 45–54 (2000)
7. Carrie, C., Murcha, M.W., Giraud, E., Ng, S., Zhang, M.F., Narsai, R., Whelan, J.: How do plants make mitochondria? Planta (2012). doi:10.1007/s00425-012-1762-3
8. Gerritsen, H.C., Sanders, R., Draaijer, A., Ince, C., Levine, Y.K.: Fluorescence lifetime imaging of oxygen in living cells. J. Fluoresc. **7**(1), 11–15 (1997)
9. Gnaiger, E., Steinlechner-Maran, R., Méndez, G., Eberl, T., Margreiter, R.: Control of mitochondrial and cellular respiration by oxygen. J. Bioenerg. Biomembr. **27**(6), 583–596 (1995)
10. James, W.O., James, A.L.: The respiration of barley germinating in the dark. New Phytol. **39**(2), 145–177 (1940)
11. Kozusko, F., Bourdeau, M.: A unified model of sigmoid tumour growth based on cell proliferation and quiescence. Cell Prolif. **40**, 824–834 (2007)
12. Lifshitz, I.M., Slyozov, V.V.: The kinetics of precipitation from supersaturated solid solutions. J. Phys. Chem. Solids **19**, 35–50 (1961)
13. Logan, D.C., Harvey Millar, A., Sweetlove, L.J., Hill, S.A., Leaver, C.J.: Mitochondrial biogenesis during germination in maize embryos. Plant Physiol. **125**, 662–672 (2001)
14. Merchant, D.J., Kuchler, R.B., Munyo, W.: Population dynamics in suspension cultures of an animal cell strain. J. Biochem. Microbiol. Technol. Eng. **II**, 253–266 (1960)
15. Van Asbrouck, J., Taridno, P.: Using the single seed oxygen consumption measurement as a method of determination of different seed quality parameters for commercial tomato seed samples. Asian J. Food Agro-Ind. **2**, S88–S95 (2009)
16. Van Duijn, A., Koenig, W.: Measuring metabolic rate changes. Patent 2001 EP1134583 (A1) WO0169243 (A1) US2004033575 (A1) CA2403253 (A1) DE60108480T
17. Wagner, C.: Theorie der Alterung von Niederschlägen durch Umlösen (Ostwald-Reifung). Z. Elektrochem. **65**, 581–591 (1961)

Mathematical Modelling of a Wave-Energy Converter

William Lee, Michael Castle, Patrick Walsh, Patrick Kelly, and Cian Murtagh

Abstract We report progress towards developing a mathematical model that can be used to optimise the design of a novel power take off unit for a wave energy generator. We show that the power take off unit can be considered as a non-smooth, dissipative dynamical system. We derive equations of motion using the Lagrangian framework, incorporating a Rayleigh dissipation function and discuss a procedure for generating approximate analytical solutions.

1 Introduction

Limerick Wave Ltd, are developing a new Power Take Off (PTO) system [1] which they aim to optimise. The design takes alternative movement induced by water waves and converts them into continuous movement via the use of a freewheeling system. In order to give an estimation for the power output of the Wave Energy Converter (WEC), one must derive the equation of motion for the angular velocity of the flywheel for a given set of conditions. See Fig. 1.

Three modes of motion are to be taken into account: positive sticking, negative sticking, and freewheeling. Positive (Negative) sticking occurs when the fly wheel is being driven by the upward (downward) motion of arm induced by waves. These occur under different conditions and each have their own constraints applied.

W. Lee (✉)
MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland

Department of Mathematics, University of Portsmouth, Portsmouth, UK
e-mail: william.lee@port.ac.uk

M. Castle
Department of Mathematics, University of Portsmouth, Portsmouth, UK
e-mail: castlemichael95@gmail.com

P. Walsh • P. Kelly
Limerick Wave, Limerick, Ireland
e-mail: patrick.walsh@ul.ie; patrick.kelly@ul.ie

C. Murtagh
Seapower, Unit 3, Pier Road, Enniscrone, Co. Sligo, Ireland
e-mail: cian.murtagh@seapower.ie

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_30

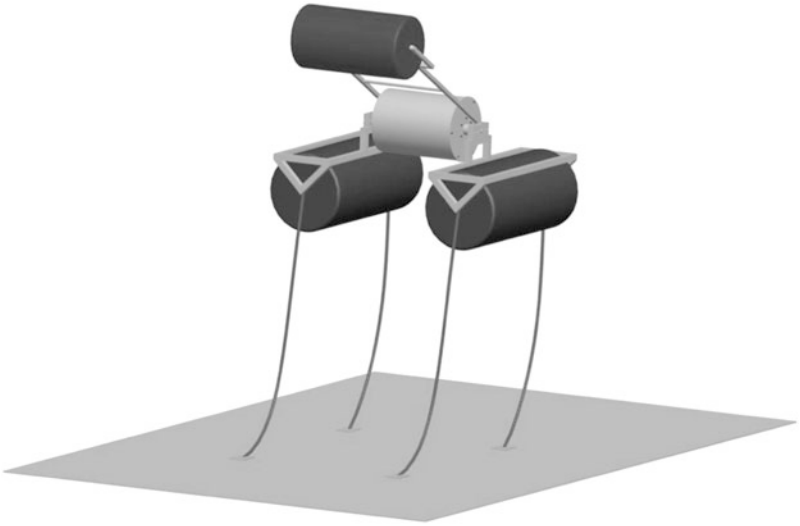


Fig. 1 Wave Energy Converter (WEC) under development by Limerick Wave

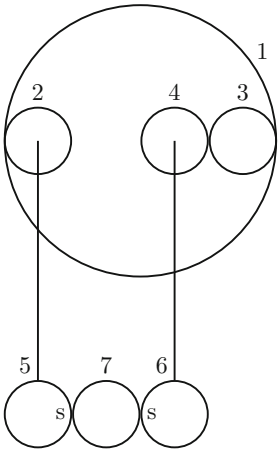


Fig. 2 Epicyclic gearing system used in the Power Take Off (PTO). Sprags are indicated by a 's'

Dissipative and non-conservative forces act on the system, with non-smooth dynamics due to sprags [4] as well as referred inertia [2].

Analytical solutions are required in order to optimise sizing. The Lagrangian method was used, along with a dissipation function [3] in order to incorporate the non-conserved forces into the model. Although starting with seven coordinates, one for each of the gears, applying constraints instead bring us to two, with a further reduction available for both positive and negative sticking. See Fig. 2.

After finding general solutions to the equations of motion for the different states, we must then discover the transition times between the different modes in order to perform the correct power integral. The motion is periodic, and hence we need only to describe the motion of the flywheel between three transition times.

2 Equations of Motion

A derivation of the equations of motion for all three states, including the conditions for transitioning between them are given in this section.

To start, we evaluate the kinetic energies and potential for the conserved forces giving us the Lagrangian

$$\mathcal{L} = \frac{1}{2}\left(mL^2 + I_1 + \frac{I_2T_1^2}{T_2^2} + \frac{I_3T_1^2}{T_3^2} + \frac{I_4T_1^2}{T_4^2} + \frac{I_5T_1^2}{T_2^2} + \frac{I_6T_1^2}{T_4^2}\right)\dot{\theta}_1^2 + \frac{1}{2}I_7\dot{\theta}_7^2 - \frac{1}{2}kL^2\theta_1^2 \tag{1}$$

where m is the mass of the float, I_i is the inertia of gear i, T_i is the number of teeth of gear i, k is the spring constant of the float and L is the length of the arm. The Lagrangian incorporates the constraints

$$\theta_1 = \frac{\theta_2T_2}{T_1} = \frac{\theta_3T_3}{T_1} = -\frac{\theta_4T_4}{T_1} = \frac{\theta_5T_2}{T_1} = -\frac{\theta_6T_4}{T_1} = \frac{y}{L} \tag{2}$$

where θ_i is the angular displacement of gear i and y is the vertical displacement of the float, which hold in all three phases.

In order to apply the Lagrangian operator successfully we need to make use of Rayleigh's Dissipation Function [3], R, below which allows us to include the non-conservative forces acting on the system. R is given by

$$R = \frac{1}{2}k\dot{y}^2 + \tau\dot{\theta}_7 + LF\cos(\omega t)\dot{y}. \tag{3}$$

The equations of motion of the system are then given by

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = Q_i. \tag{4}$$

In the case of a freewheeling system this gives us two equations of motion.

$$I_7\ddot{\theta}_7 = \tau \tag{5}$$

and

$$A\ddot{\theta}_7 + CL^2\dot{\theta}_7 + KL^2\theta_7 = LF\cos(\omega t) \tag{6}$$

where $A = mL^2 + I_1 + \frac{I_2T_1^2}{T_2^2} + \frac{I_3T_1^2}{T_3^2} + \frac{I_4T_1^2}{T_4^2} + \frac{I_5T_1^2}{T_2^2} + \frac{I_6T_1^2}{T_4^2}$. Note we need only θ_7 in order to determine the motion during this phase.

Upon applying the additional constraint $\theta_1 = \frac{T_4T_7}{T_1T_6}\theta_7$ we achieve the positive sticking equation of motion

$$A\left(\frac{T_4T_7}{T_1T_6}\right)\ddot{\theta}_7 + CL^2\left(\frac{T_4T_7}{T_1T_6}\right)\dot{\theta}_7 + KL^2\left(\frac{T_4T_7}{T_1T_6}\right)\theta_7 = LF\cos(\omega t) + \frac{T_1T_6}{T_4T_7}\tau \tag{7}$$

Similarly using $\theta_1 = -\frac{T_2T_7}{T_1T_5}\theta_7$ as the additional constraint, we get

$$A\left(\frac{T_2T_7}{T_1T_5}\right)\ddot{\theta}_7 + CL^2\left(\frac{T_2T_7}{T_1T_5}\right)\dot{\theta}_7 + KL^2\left(\frac{T_2T_7}{T_1T_5}\right)\theta_7 = LF\cos(\omega t) + \frac{T_1T_6}{T_4T_7}\tau \tag{8}$$

for our negative sticking equation of motion.

Transition times, as well as constants of integration must then be found. The condition for which the flywheel leaves Positive or Negative sticking and instead begins to freewheel is that when the kinetic energy it would have when freewheeling is greater than it would be during a sticking phase. That is, when

$$\ddot{\theta}_7^{FW} > \ddot{\theta}_7^{PS} \tag{9}$$

$$\ddot{\theta}_7^{FW} > \ddot{\theta}_7^{NS} \tag{10}$$

where $\dot{\theta}^{FW}$ is the angular velocity when freewheeling whilst $\dot{\theta}^{PS}$ and $\dot{\theta}^{NS}$ are the angular velocities of positive and negative sticking respectively.

When calculating the transition points these should be interpreted as the values of $\ddot{(\theta)}_7^{mode}$ if the system was constrained to be in mode(=FW, PS, NS)

The transition point is that at which equality holds. We must also have continuity of velocity at all transition times. There is a total of four equations of the type

$$\dot{\theta}_7^{FW} = \dot{\theta}_7^{PS} \tag{11}$$

$$\dot{\theta}_7^{FW} = \dot{\theta}_7^{NS} \tag{12}$$

For the final equation, we must have continuity of displacement. Due to the periodic nature, we must have that the path of the velocity ends where it begins, giving us the final equation.

$$\theta_{7,1}^{FW} = \theta_{7,2}^{FW} \tag{13}$$

where the subscripts signify the difference between the expressions.

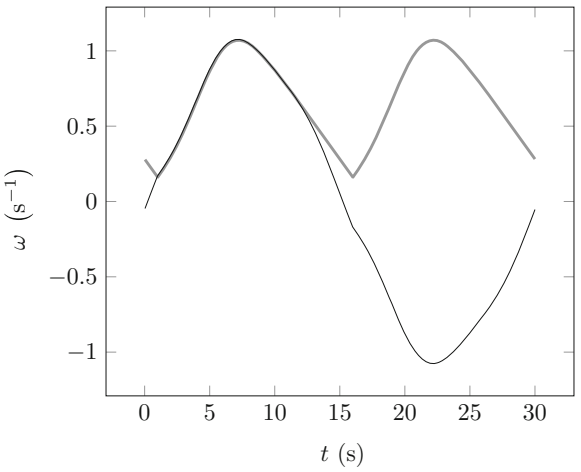


Fig. 3 Angular velocities. *Thick grey line*: $\dot{\theta}_7$ against time. *Black line* $\dot{\theta}_1$ against time (corrected by a gearing ratio to be comparable to $\dot{\theta}_7$)

Finally, by integrating the rate of work done against flywheel resistance over a single period, T and averaging, we can calculate the average power generation, $\bar{P}$,

$$\bar{P} = \frac{1}{T}\int_0^T \tau\dot{\theta}_7 dt \tag{14}$$

A procedure for developing analytic approximations to solutions of these equations has been developed. With the reduced model each of the equations of motion, as well as the transition times can be solved analytically after making use of small angle approximations and application of Taylor's expansion formula. The transient restarts at each transition requiring us to only review a small section of the motion.

3 Results

Numerical solutions to the equations of motion are shown in Fig. 3. The figure shows $\dot{\theta}_7$ and $\dot{\theta}_1$ corrected by a factor of the gearing ratio $\frac{T_1T_6}{T_4T_7} = \frac{T_1T_5}{T_2T_7}$ to allow direct comparison between the angular velocities.

4 Conclusion

The Limerick Wave PTO is a complex dynamical system with dissipative, non-smooth dynamics. We have shown that the equations of motion of this system can be modelled using a Lagrangian and a dissipation function. Future work will focus on optimising the power output as a function of design parameters.

Acknowledgements WTL acknowledges the support of MACSI, the Mathematics Applications Consortium for Science and Industry (www.macsi.ul.ie), funded by the Science Foundation Ireland Investigator Award 12/IA/1683. MC acknowledges the funding support from the London Mathematical Society, via its Undergraduate Research Bursaries in Mathematics Scheme, that has contributed to this publication. PW, PK and CM acknowledges the support of Enterprise Ireland via the Innovation Voucher scheme.

References

1. Burns, S., Chapwanya, M., Cummins, C., Dellar, P., Giddings, J., Giounanlis, P., Hicks, P., et al.: Limerick Wave: using flywheel technology to convert the power of the waves to electricity. Technical report 686. http://www.maths-in-industry.org/miis/ (2013)
2. Gerstmayr, J., Ambrósio J.A.C.: Component mode synthesis with constant mass and stiffness matrices applied to flexible multibody systems. Int. J. Numer. Methods Eng. **73**, 1518–1546 (2008)
3. Goldstein, H., Poole, C., Safko, J.: Classical Mechanics, pp. 22–24. Pearson Education International, Addison Wesley (2002)
4. Richter, H., Maynard R.E.: Hybrid-dynamical modelling, characterization, and experimental verification of a free-wheeling clutch. Proc. Inst. Mech. Eng. I J. Syst. Control Eng. **224**, 361–372 (2010)

Aerodynamic Web Forming: Pareto-Optimized Mass Distribution

Nicole Marheineke, Sergey Antonov, Simone Gramsch, and Raimund Wegener

Abstract In the technical textile industry an objective of the airlay process is the production of high quality nonwoven fabrics with the minimal use of fiber raw material. Since a process simulation of the multi-scale two-phase problem is very computationally expensive, we deduce an efficiently evaluable surrogate model to handle the multi-criteria optimization task.

1 Introduction

Aerodynamic web forming addresses a broad spectrum of applications for the produced nonwoven materials. Airlay fabrics range from insulation and filter materials over automotive and mattress felts to medical and hygiene products. In the airlay process elastic staple fibers leave from a rotating drum into a turbulent air flow. Suctioning onto a perforated moving conveyor belt leads to the forming of a random three-dimensional web structure, cf. Fig. 1, [1, 4]. As the final nonwoven fabrics have different shapes, an objective in industry is the reduction of material waste by controlling the mass distribution on the conveyor belt through a systematic non-uniform fiber deposition. A typical quality criterion for the fabrics (or the fabric-associated web regions) is thereby a homogeneous mass distribution. This optimization task of multiple criteria is topic in the paper. We particularly investigate whether spatial variations of the conveyor belt's porosity can be used as control parameter.

N. Marheineke (✉)
FAU Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: marheineke@math.fau.de

S. Antonov • S. Gramsch • R. Wegener
Fraunhofer ITWM, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
e-mail: sergey.antonov@itwm.fraunhofer.de; simone.gramsch@itwm.fraunhofer.de; raimund.wegener@itwm.fraunhofer.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_31

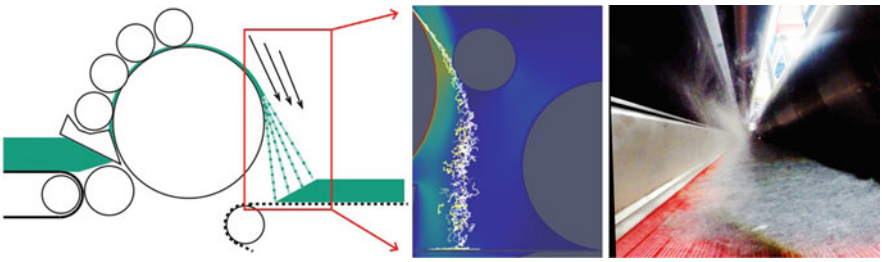


Fig. 1 Airlay process: aerodynamic web forming zone with elastic fiber dynamics in turbulent flow and deposition onto a moving conveyor belt. Sketch with simulation result (*left*), photo (*right*)

2 Process Simulation and Optimization with Surrogate Model

The airlay process is characterized by the machine direction (MD) and its cross-direction (CD) where the conveyor belt lies in the MD-CD plane. In the process the turbulent fiber suspension is dilute and the fibers mainly follow the mean flow streamlines without influencing the flow or interacting with each other. The conveyor belt affects the air flow and hence the deposition of the suspended fibers. To study the feasibility of controlling the fiber web's mass distribution by means of the conveyor belt porosity (permeability, respectively) we consider the following exemplary set-up. We desire a periodic lay-down in MD: half the web region should be filled in a homogeneous way (fabric), half should be as empty as possible since it will be not further used (waste). Therefore, we consider conveyor belts with periodic permeability patterns in MD. Due to this design the process properties are homogeneous in CD apart from negligible boundary effects.

Process Simulation with Fiber Dynamics and Deposition A process simulation of the multi-scale two-phase problem can be performed as described in [4]. Thereby, we particularly model a fiber asymptotically as an elastic Kirchhoff beam that is capable of large, geometrically nonlinear deformations and driven by stochastic aerodynamic drag forces due to the surrounding turbulent flow. The flow is described by a statistic k-ϵ turbulence model. The conveyor belt is realized as porous medium using Darcy's law with a periodic porosity pattern. Let I denote the periodic cell and $p : I \to \mathbb{R}_0^+$ be the relative belt permeability, $\int_I p \, \mathrm{d}x = 1$. To the moving conveyor belt we perform a transient (2d) simulation of the unloaded flow. In the longtime behavior the flow becomes periodic and can be expressed by a finite sequence of time snapshots. Since the belt speed is comparably small to the magnitude of the mean flow velocity, we consider the fiber behavior with respect to a single snapshot. The resulting fiber mass distribution is obtained from computing the dynamics and deposition of a large sample of n fibers per time snapshot, regarding their lay-down with respect to one fiber end point and projecting these MD lay-down positions in

the periodic cell I. The discrete cumulative numerical results $\hat{f}$ are collected in a histogram.

As reference scenario we use a conveyor belt with a permeable stripe pattern p_0 —open and closed in the intended regions for fabric and waste, respectively. Moreover, we consider two different types of fiber raw material (i.e., light bicomponent (PES/PET) and heavier solid (PES) fibers); see Figs. 2 and 3 for the corresponding mean air flow in an airlay plant geometry and $\hat{f}_0$ associated to fiber samples with $n_{bico} = n_{solid} = 1000$. For more information to the airlay process and the fiber raw material we refer to [4]. For the fiber-flow computations at industrial scale we apply the commercial CFD software ANSYS Fluent for the flow and the licensable research software FIDYST of the Fraunhofer ITWM for the fibers (cf. details to model and numerics in [5]).

Surrogate Model for Mass Distribution As process simulations of the fibers' behavior in the turbulent flow and their deposition are very computationally demanding we deduce a surrogate model for optimizing the mass distribution of the resulting fiber web. We model the mass distribution $f \in \mathscr{C}(I, \mathbb{R}_0^+)$ as periodic convolution of the relative belt permeability p and a kernel function k, i.e.,

$$f[p,k](x) = (p \star k)(x), \qquad x \in I. \tag{1}$$

The kernel $k : I \to \mathbb{R}_0^+$ might be interpreted as a fiber deposition distribution that we can determine from the described fiber-flow (reference) simulation of the process. The identification of k is an inverse and ill-posed problem that we solve by help of a Tikhonov regularization [3] with regularization parameter $0 < \epsilon_k \ll 1$, presupposing sufficient regularity of the involved functions,

$$\operatorname{argmin}_{k \geq 0} K(k), \qquad K(k) = \|f[p_0,k] - \hat{f}_0\|^2_{\mathscr{L}^2(I)} + \epsilon_k \|f'[p_0,k]\|^2_{\mathscr{L}^2(I)}. \tag{2}$$

Considering the referential permeable stripe pattern p_0, the identified kernel functions for the bicomponent and solid fibers are visualized in Fig. 4 (left). The corresponding mass distributions f_0 of (1) are seen in Fig. 3. The optimization problem (2) is numerically solved with the MATLAB routine `fmincon`.

Pareto Optimization Regarding the periodic cell $I = I_f \cup I_w$ with $I_f \cap I_w = \emptyset$ and $|I_f| = |I_w| = |I|/2$, we aim a mass distribution f that satisfies the homogeneity demands on the fabric in I_f and should be almost zero in the waste-associated region I_w. Hence, we introduce the deviation from homogeneity S and the waste A as optimization criteria

$$S(f) = \langle f \rangle^{-1} \sqrt{\int_{I_f} (f - \langle f \rangle)^2 \, \mathrm{d}x}, \qquad \langle f \rangle = \int_{I_f} f \, \mathrm{d}x \tag{3a}$$

$$A(f) = \int_{I_w} f \, \mathrm{d}x \tag{3b}$$

and formulate the following pareto optimization task.

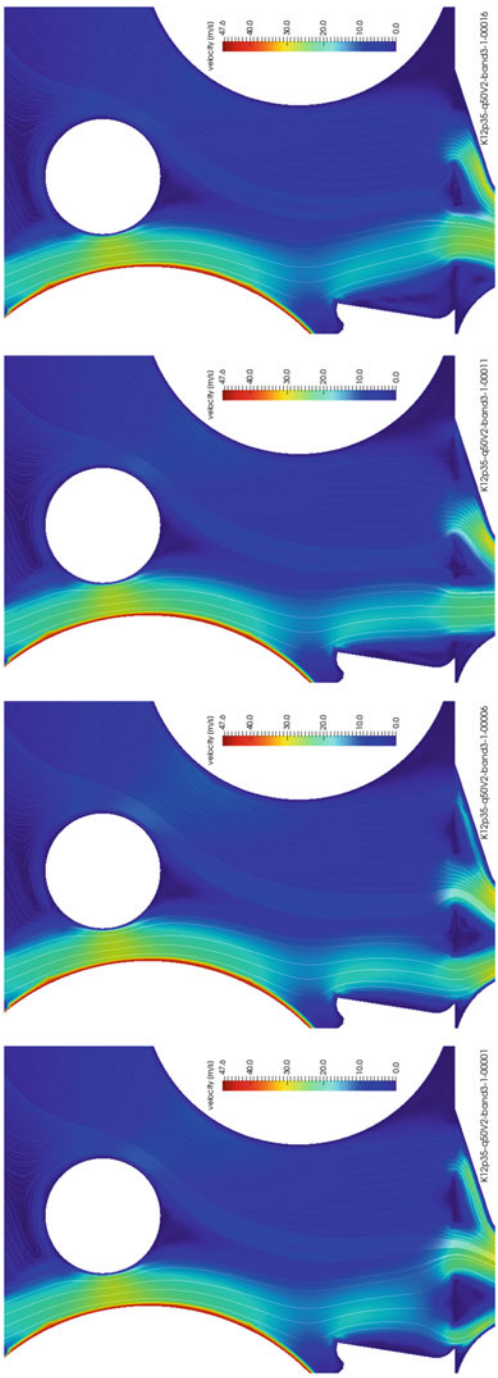


Fig. 2 Flow streamlines due to moving conveyor belt with permeable stripe pattern p_0 for different time snapshots over one time period. Machine parts are displayed in *white*, flow is colored by mean velocity magnitude

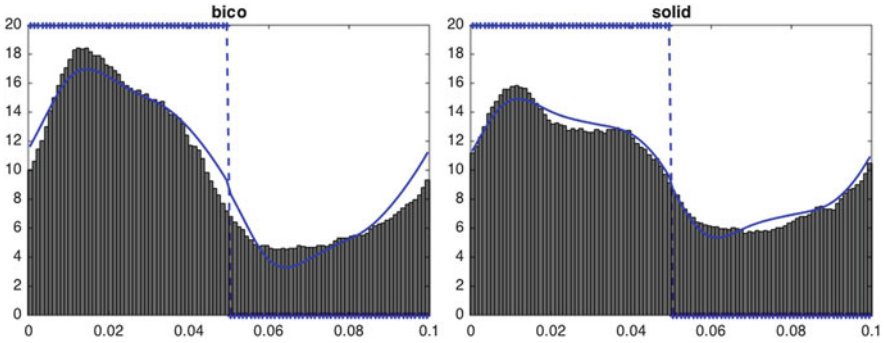


Fig. 3 Mass distribution in I for referential belt permeability p_0 (*dashed line*, marker *plus symbol*): $\hat{f}_0$ from process simulation (histogram) and f_0 from surrogate model (1) after kernel identification (*solid line*). Bicomponent (*left*) and solid fibers (*right*)

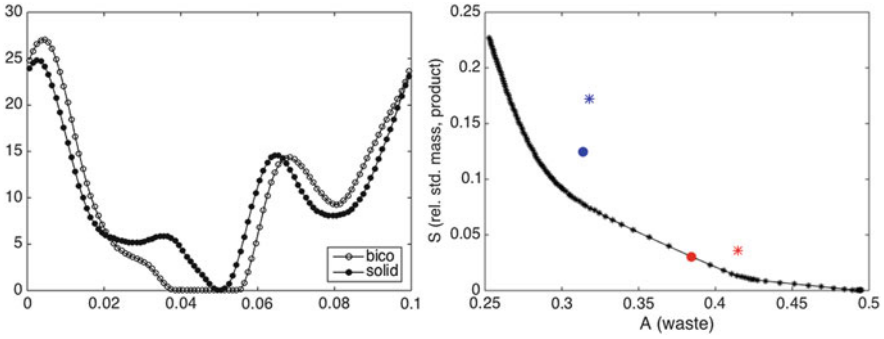


Fig. 4 *Left*: Kernel function k (2) for bicomponent (*open circle*) and solid (*filled circle*) fibers. *Right*: Pareto-front (*black line*, *star*) wrt. the two optimization criteria: A (waste) and S (deviation from homogeneity). Values of the reference (Fig. 1, *blue*) and a pareto-optimized scenario (Fig. 5, *red*); the results belong to the surrogate model (*filled circle*) and the process simulation (*star*)

Let $\mathscr{P} = \mathscr{C}^1(I, \mathbb{R}_0^+)$ be the space of considered permeability functions, k the kernel function of (2) and f the mass distribution wrt. to the surrogate model (1). Find the boundary of the set $\mathscr{M}(\mathscr{P}) \subset \mathbb{R}^2$ that is given by

$$\mathscr{M}(\mathscr{P}) = \{(a, s) \in \mathbb{R}^2 \mid (a, s) = (A(f[p, k]), S(f[p, k]) \text{ for } p \in \mathscr{P}\}.$$

We are interested in the pareto front. As subset of $\partial\mathscr{M}(\mathscr{P})$ it consists of all permeability functions for which an improvement (minimization) with respect to both optimization criteria A and S is not possible. For $\mathscr{P} = \mathscr{C}^1(I, \mathbb{R}_0^+)$ the multi-criteria optimization is an ill-posed problem. Since we deal here with only two criteria, we can consider a convex combination of the cost functionals A and S (3)

and determine the pareto front as solution of a sequence of regularized minimization problems

$$\begin{aligned} &\min_{p\geq 0} J_\alpha(p), \\ &J_\alpha(p) = \alpha\, A(f[p,k]) + (1-\alpha)\, S(f[p,k]) + \epsilon \|p'\|_{\mathscr{L}^2(I)} \quad \text{for } \alpha \in [0,1] \\ &\text{subjected to (1)} \end{aligned} \tag{4}$$

with calibrated kernel k (2) and regularization parameter $0 < \epsilon \ll 1$. For more sophisticated strategies in cases of more criteria see [2] and references within.

In the considered set-up we deal with two types of fiber raw material and thus use $A = (A_{bico} + A_{solid})/2$ and $S = (S_{bico} + S_{solid})/2$. The pareto front is visualized in Fig. 4 (right). It is determined by means of the MATLAB routine `fmincon` using an equidistant grid for the convex parameter α in (4). Figure 4 shows (A, S) for the referential belt permeability p_0 (cf. Fig. 1). We observe a difference between the results from the full process simulation (original model) and the surrogate model. This is not surprising and originated in the identification/calibration of the kernel. Note that also the evaluation of the discrete simulation data in the histogram has an impact on the result. One could think about smoothing the underlying data. However, the convolution with the kernel function is already a smoothing. As optimized scenario on the pareto front we choose a small deviation S from homogeneity with moderate waste A. The respective belt permeability p_{opt} is given in Fig. 5. Figure 5 also shows the mass distribution with respect to the surrogate model f_{opt} and the one $\hat{f}_{opt}$ obtained from the process simulation. The validation result is very promising.

To assess the impact of the optimization, we consider the material saving. Let M be the total mass of raw material that is required for manufacturing a homogeneous web in I_f with a constant belt permeability. The effective material in the fabric-associated region is $M_f = M/2$. Using a pareto-optimized belt permeability implies

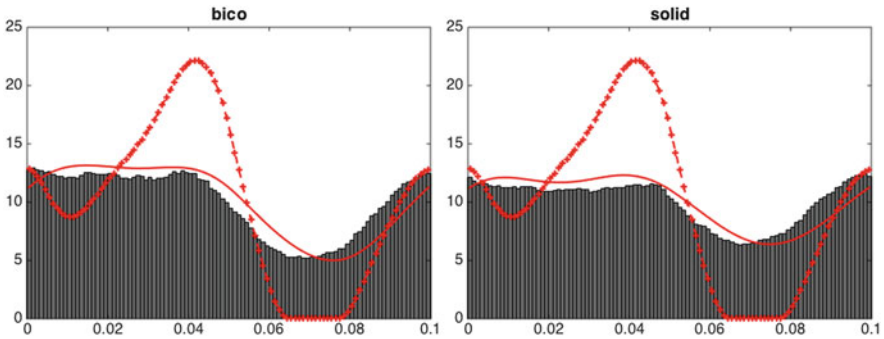


Fig. 5 Mass distribution in I for a pareto-optimized belt permeability p_{opt} (*dashed line*, marker +): f_{opt} from calibrated surrogate model (*solid line*) and $\hat{f}_{opt}$ from validating process simulation (histogram). Bicomponent (*left*) and solid fibers (*right*)

a reduction of waste. Keeping M_f fixed, the relative material saving is given by

$$\frac{M - M_{opt}}{M} = 1 - \frac{1}{2(1-A)}$$

with $A \in [0, 0.5]$ of (3). In the chosen scenario of Fig. 5 we have $A \approx 0.4$ and thus a material saving of approximately 16.7%.

3 Discussion and Outlook

The feasibility study provides a proof of concept that the belt permeability as control allows for the optimization of the fiber web's mass distribution. The effect is obtained on spatial scales that correspond to the typical lay-down range of the fibers. Hence, suctioning in combination with different porosity patterns of the conveyor belt can be applied for fine-tuning the nonwoven's quality lately in the deposition zone of the process. In view of certain fabrics' shapes, a systematic non-uniform fiber deposition on larger scales is aimed for. Here, one might think about a process design where the raw material supply or the air flow in the whole aerodynamic web forming zone act as control quantities.

Acknowledgements The financial support of the German Bundesministerium für Bildung und Forschung, Project OPAL 05M13, is acknowledged.

References

1. Albrecht, W., Fuchs, H., Kittelmann, W. (eds.): Nonwoven Fabrics: Raw Materials, Manufacture, Applications, Characteristics, Testing Processes. Wiley, New York (2006)
2. Ehrgott, M.: Multicriteria Optimization. Springer, Heidelberg (2000)
3. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Springer, Berlin (1996)
4. Gramsch, S., Klar, A., Leugering, G., Marheineke, N., Nessler, C., Strohmeyer, C., Wegener, R.: Aerodynamic web forming: process simulation and material properties. J. Math. Ind. **6**(13), 1–23 (2016)
5. Wegener, R., Marheineke, N., Hietel, D.: Virtual production of filaments and fleeces. In: Neunzert, H., Prätzel-Wolters, D. (eds.) Currents in Industrial Mathematics: From Concepts to Research to Education, pp. 103–162. Springer, Heidelberg (2015)

Minisymposium: Finite Volume Schemes for Degenerate Problems

Mazen Saad

Description

This minisymposium was a response to the growing use of finite volume schemes combined with finite element schemes in degenerate problems and was intended to bring together developers and researchers from academia and industry. Degenerate problem arises from multifluid flow in porous media or in medicine model as the breast cancer or bone healing model.

Saturations of oil, gas and water, or densities of cells are positive quantities, hence it is crucial to propose numerical schemes ensuring the physical properties as the positiveness and conservation. The goal of this minisymposium was to shed light on recent advances, for instance, in nonlinear control volume finite element scheme or nonlinear monotone scheme for degenerate problems with anisotropic and heterogeneous diffusion tensor over general meshes.

The talks included in the minisymposium were the following:

- Cindy Guichard. Université Pierre et Marie Curie, Paris (France). *A degenerate coupled parabolic problem arising in the study of the root-nutrient system.*
- Moustafa Ibrahim. Ecole Centrale de Nantes (France). *Nonlinear CVFE scheme for simulating a breast cancer model.*
- Mladen Jurak. University of Zagreb (Croatia). *On Double Porosity Model for Two-Phase Flow in Porous Media.*
- Mazen Saad. Ecole Centrale Nantes (France). *Numerical analysis for partially miscible two-phase two component flows in porous media.*

M. Saad (✉)
Ecole Centrale de Nantes, Laboratoire de Mathématiques Jean Leray, Nantes, France
e-mail: mazen.saad@ec-nantes.fr

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_32

Convergence of a Nonlinear Control Volume Finite Element Scheme for Simulating Degenerate Breast Cancer Equations

Françoise Foucher, Moustafa Ibrahim, and Mazen Saad

Abstract In this paper, a positive nonlinear CVFE scheme is considered for the simulation of an anisotropic degenerate breast cancer model. This scheme includes the use of the finite element method combined with the Godunov scheme to approximate the diffusion terms over a primal mesh, and it uses a nonclassical upwind finite volume scheme to approximate the other terms over a barycentric dual mesh. This scheme ensures the discrete maximum principle and it converges to a weak solution without any restriction on the transmissibility coefficients. Numerical experiment is supplied in order to show the efficiency of the scheme to simulate an anisotropic breast cancer model over a general mesh.

1 The Anisotropic Degenerate Breast Cancer Model

Let Ω be an open bounded polygonal and connected subset of $\mathbb{R}^d$, $d = 2,\ 3$ and let $t_{\rm f} > 0$ be a fixed finite time. We denote by $Q_{t_{\rm f}} = \Omega \times (0, t_{\rm f})$ and $\Sigma_{t_{\rm f}} = \partial\Omega \times (0, t_{\rm f})$. We are interested in a modified degenerate nonlinear system [4] modeling the breast cancer development given by the set of equations

$$\begin{cases} \partial_t f = -\eta_f v f - \rho_1 f & \text{in } Q_{t_{\rm f}}, \\ \partial_t q = \lambda_q q(1-A) - \eta_q u q + \rho_1 f - \rho_2 q & \text{in } Q_{t_{\rm f}}, \\ \partial_t r = \lambda_r r(1-A) - \eta_r u r + \rho_2 q - \rho_3 r & \text{in } Q_{t_{\rm f}}, \\ \partial_t s = \lambda_s s(1-A) - \eta_s u s + \rho_3 r - \rho_4 s & \text{in } Q_{t_{\rm f}}, \\ \partial_t u - \operatorname{div}\left(\Lambda(\mathbf{x}) a(u) \nabla u - \Lambda(\mathbf{x}) \chi(u) \nabla f\right) = \lambda_u u(1-A) + \rho_4 s & \text{in } Q_{t_{\rm f}}, \\ \partial_t v - \operatorname{div}\left(D(\mathbf{x}) \nabla v\right) = g(u, v) & \text{in } Q_{t_{\rm f}}, \end{cases} \tag{1}$$

F. Foucher • M. Saad
Ecole Centrale de Nantes, UMR 6629 CNRS, Laboratoire de Mathématiques Jean Leray, F-44321 Nantes, France
e-mail: francoise.foucher@ec-nantes.fr; mazen.saad@ec-nantes.fr

M. Ibrahim (✉)
Math and Science Division, American College of the Middle East, 220, Dasman 15453, Kuwait
e-mail: moustafa.ibrahim@acm.edu.kw

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_33

We add to this system the homogeneous zeros-flux boundary conditions on Σ_{t_f} given by

$$(\Lambda(\mathbf{x})a(u)\nabla u - \Lambda(\mathbf{x})\chi(u)\nabla f)\cdot\mathbf{n} = 0, \qquad D(\mathbf{x})\nabla v\cdot\mathbf{n} = 0, \tag{2}$$

where $\mathbf{n}$ is the unit normal vector to $\partial_t\Omega$ outward to Ω. Finally, we consider the initial conditions on Ω given by:

$$w(\mathbf{x},t) = w_0(\mathbf{x}), \qquad \text{in } \Omega, \qquad w = f,\ q,\ r,\ s,\ u,\ v. \tag{3}$$

In the above model, we represent by f a fraction of healthy breast stem cells in the breast tissue, by q the cells with LOH in TSG_1, by r the cells with inactivated TSG_1, and by s the cells with LOH in TSG_2. The density of the tumor cells and the enzymes concentration which they produce are represented by $u = u(\mathbf{x},t)$ and $v = v(\mathbf{x},t)$ respectively. Next, $A = u+f+q+r+s$ represents the total tissue and cell population including tumor cells, $a(u)$ is a density-dependent diffusion coefficient, and $\Lambda(\mathbf{x})$ is the diffusion tensor in a heterogeneous medium. Furthermore, the function $\chi(u)$ represents the enzymes sensitivity to the healthy breast stem cells, and $D(\mathbf{x})$ is the diffusion tensor for v. The function $g(u,v)$ describes the production and degradation of the enzymes by the tumor cells; here, we assume it is a nonlinear function given by

$$g(u,v) = \alpha u(1-v) - \beta v, \qquad \alpha,\beta \geq 0, \tag{4}$$

where α is the production rate and β is the degradation rate. Finally, we denote λ_w the production rate of w for $w = q,\ r,\ s,\ u$ and by η_w the death rate of w for $w = f,\ q,\ r,\ s$.

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, $a(u) \geq 0$ for $0 \leq u \leq 1$, and $u(0) = u(1) = 0$.

(A2) The sensitivity $\chi : [0,1] \longrightarrow \mathbb{R}$ is a continuous function such that, $\chi(0) = \chi(1) = 0$, and $\chi(u) > 0$ for $0 < u < 1$. Furthermore, we assume that there exists a function $\mu \in \mathcal{C}\left([0,1];\mathbb{R}^+\right)$, such that $\mu(u) = \dfrac{\chi(u)}{a(u)}$ for all $u \in (0,1)$ 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) We assume that the production rate λ_q of q is greater than the mutation probability ρ_1. Furthermore, we assume that: $\lambda_r \geq \rho_2,\ \lambda_s \geq \rho_3,\ \lambda_u \geq \rho_4$.

(A5) The initial functions u_0, v_0, and f_0 satisfy: $u_0,\ v_0 \in L^2(\Omega)$ and $f_0 \in H^1(\Omega)$. In addition, we assume that: $0 \leq f_0,\ q_0,\ r_0,\ s_0,\ u_0 \leq 1, \quad v_0 \geq 0$ in Ω,

In the sequel, we use the Lipschitz continuous function $\xi : [0,1] \to \mathbb{R}$ defined by $\xi(u) := \int_0^u \sqrt{a(s)}\,\mathrm{d}s$, and the following set of functions: $\eta(v)$, $p(v)$, $\Gamma(v)$

and $\Phi(v)$ defined in $\mathbb{R}$ by $\eta(v) = \max(0, \min(v, 1))$, $p(v) = \int_1^v \frac{1}{\eta(s)}\,\mathrm{d}s$, $\Gamma(v) = \int_1^v p(s)\,\mathrm{d}s$, $\Phi(v) = \int_0^v \frac{1}{\sqrt{\eta(s)}}\,\mathrm{d}s$. Hereafter, we adopt the convention: $\eta(v)p(v) = 0$ when $v \leq 0$.

We are interested to numerically investigate the anisotropic breast cancer model. For instance, the finite volume scheme has been studied by Foucher et al. [6] for the degenerate isotropic case. The scheme adopted here is inspired from the nonlinear scheme used in [2] where it has been employed to discretize an anisotropic Keller-Segel model.

2 The Nonlinear CVFE Scheme for System (1)–(3)

The discretization of system (1)–(3) 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 $\mathcal{T}$ be a conforming triangulation of the domain Ω. We define $h_{\mathcal{T}}$ and $\theta_{\mathcal{T}}$ to be the size and the regularity of the mesh Ω defined by: $h_{\mathcal{T}} := \max_{T\in\mathcal{T}} h_T$ and $\theta_{\mathcal{T}} = \max_{T\in\mathcal{T}} \frac{h_T}{\rho_T}$, 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 $\mathcal{V}$ the set of vertices of the triangulation $\mathcal{T}$ and by $\mathcal{E}$ the set of edges of $\mathcal{T}$. For every vertex $K \in \mathcal{V}$ (located at position $\mathbf{x}_K$), we denote by $\mathcal{E}_K$ the subset of $\mathcal{E}$ 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 $K \in \mathcal{V}$ by connecting the barycenters $\mathbf{x}_T$ of the triangles $T \in \mathcal{T}_K$ with the barycenters $\mathbf{x}_\sigma$ of the edges $\sigma \in \mathcal{E}_K$. We denote by $\mathcal{H}_{\mathcal{T}}$ the usual $\mathbb{P}_1$-finite element space defined by

$$\mathcal{H}_{\mathcal{T}} = \{\phi \in \mathcal{C}^0\left(\overline{\Omega}\right); \phi_{|_T} \in \mathbb{P}_1(\mathbb{R}), \forall T \in \mathcal{T}\}$$

and by $(\varphi_K)_{K\in\mathcal{V}}$ its canonical basis. Furthermore, we consider the discrete control volumes space $\mathcal{X}_{\mathcal{M}}$ defined by $\mathcal{X}_{\mathcal{M}} = \{\phi : \Omega \longrightarrow \overline{\mathbb{R}}, \phi_{|_{\omega_K}} \text{ is constant }, \forall K \in \mathcal{V}\}$.

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.

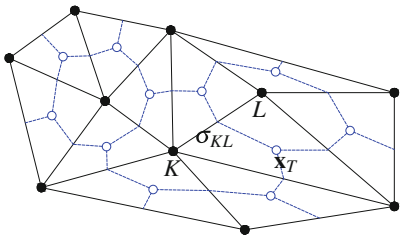


Fig. 1 Triangular mesh $\mathcal{T}$ and Donald dual mesh $\mathcal{M}$: dual volumes, vertices, interfaces

We set $t^n = n\Delta t$ for $0 \leq n \leq N+1$, and introduce the space and time discrete spaces

$$\mathcal{H}_{\mathcal{T},\Delta t} = \{\phi \in L^\infty(Q_{t_f}); \phi(\mathbf{x},t) = \phi\left(\mathbf{x}, t^{n+1}\right) \in \mathcal{H}_{\mathcal{T}}, \forall t \in (t^n, t^{n+1}], 0 \leq n \leq N\},$$

$$\mathcal{X}_{\mathcal{M},\Delta t} = \{\phi \in L^\infty(Q_{t_f}); \phi(\mathbf{x},t) = \phi\left(\mathbf{x}, t^{n+1}\right) \in \mathcal{X}_{\mathcal{M}}, \forall t \in (t^n, t^{n+1}], 0 \leq n \leq N\}.$$

Let m_K be the 2-dimensional Lebesgue measure of ω_K for every $K \in \mathcal{V}$. The nonlinear CVFE scheme for the discretization of system (1)–(3), is given by the following set of equations: for all $K \in \mathcal{V}$

$$w^0_{\mathcal{M}}(\mathbf{x}) = w^0_K = \frac{1}{m_K}\int_{\omega_K} w_0(\mathbf{y})\,\mathrm{d}\mathbf{y}, \qquad \text{where } w = f,\ q,\ r,\ s,\ u,\ v. \tag{5}$$

and for all $n \in \{0,\ldots,N\}$

$$\begin{aligned}
\frac{f_K^{n+1}-f_K^n}{\Delta t} &= -\eta_f v_K^{n+1} f_K^{n+1} - \rho_1 f_K^{n+1},\\
\frac{q_K^{n+1}-q_K^n}{\Delta t} &= \lambda_q q_K^{n+1}(1-A_K^{q,n+\frac{1}{2}}) - \eta_q u_K^n q_K^{n+1} + \rho_1 f_K^{n+1} - \rho_2 q_K^{n+1},\\
\frac{r_K^{n+1}-r_K^n}{\Delta t} &= \lambda_r r_K^{n+1}(1-A_K^{r,n+\frac{1}{2}}) - \eta_r u_K^n r_K^{n+1} + \rho_2 q_K^{n+1} - \rho_3 r_K^{n+1},\\
\frac{s_K^{n+1}-s_K^n}{\Delta t} &= \lambda_s s_K^{n+1}(1-A_K^{s,n+\frac{1}{2}}) - \eta_s u_K^n s_K^{n+1} + \rho_3 r_K^{n+1} - \rho_4 s_K^{n+1},\\
m_K\frac{u_K^{n+1}-u_K^n}{\Delta t} &+ \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right) - \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL}\mu_{KL}^{n+1} a_{KL}^{n+1}(f_K^{n+1}-f_L^{n+1})\\
&= m_K\left(\lambda_u u_K^{n+1}(1-A_K^{n+1}) + \rho_4 s_K^{n+1}\right),\\
m_K\frac{v_K^{n+1}-v_K^n}{\Delta t} &+ \sum_{\sigma_{KL}\in\mathcal{E}_K} D_{KL}\eta_{KL}^{n+1}\left(p\left(v_K^{n+1}\right) - p\left(v_L^{n+1}\right)\right)\\
&= m_K\left(\alpha u_K^n(1-v_K^{n+1}) - \beta v_K^{n+1}\right),
\end{aligned} \tag{6}$$

where

$$A_K^{q,n+\frac{1}{2}} = f_K^{n+1} + q_K^{n+1} + r_K^n + s_K^n + u_K^n, \quad A_K^{r,n+\frac{1}{2}} = f_K^{n+1} + q_K^{n+1} + r_K^{n+1} + s_K^n + u_K^n,$$

$$A_K^{s,n+\frac{1}{2}} = f_K^{n+1} + q_K^{n+1} + r_K^{n+1} + s_K^{n+1} + u_K^n,\ A_K^{n+1} = f_K^{n+1} + q_K^{n+1} + r_K^{n+1} + s_K^{n+1} + u_K^{n+1}.$$

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 (see [2]) by

$$a_{KL}^{n+1} = \begin{cases} \max_{s\in I_{KL}^{n+1}} a(s) & \text{if } \Lambda_{KL} \geq 0, \\ \min_{s\in I_{KL}^{n+1}} a(s) & \text{if } \Lambda_{KL} < 0, \end{cases} \quad \eta_{KL}^{n+1} = \begin{cases} \max_{s\in J_{KL}^{n+1}} \eta(s) & \text{if } D_{KL} \geq 0, \\ \min_{s\in J_{KL}^{n+1}} \eta(s) & \text{if } D_{KL} < 0. \end{cases} \tag{7}$$

Finally, μ_{KL}^{n+1} is set to be equals to

$$\mu_{KL}^{n+1} = \begin{cases} \mu_\downarrow\left(u_K^{n+1}\right) + \mu_\uparrow\left(u_L^{n+1}\right), & \text{if } \Lambda_{KL}\left(f_K^{n+1} - f_L^{n+1}\right) \geq 0, \\ \mu_\uparrow\left(u_K^{n+1}\right) + \mu_\downarrow\left(u_L^{n+1}\right), & \text{if } \Lambda_{KL}\left(f_K^{n+1} - f_L^{n+1}\right) < 0. \end{cases}$$

The functions $\mu_\uparrow$ and $\mu_\downarrow$ are deduced from the function μ introduced in assumption (A2) and given by $\mu_\uparrow(z) = \int_0^z (\mu'(s))^+\,\mathrm{d}s$, and $\mu_\downarrow(z) = -\int_0^z (\mu'(s))^-\,\mathrm{d}s$. Note that the nonlinear CVFE scheme (6) is locally conservative on the dual median mesh $\mathcal{M}$.

3 Discrete Estimates, Existence and Convergence of the Scheme

In the sequel, we denote by ζ a continuous real-valued function defined on $\mathbb{R}_+$ such that $\zeta(0) \geq 0$, and we consider its continuous extension by $\zeta(0)$ on $\mathbb{R}_-^*$.

Lemma 1 *Let $\left(w_K^n\right)_{K\in\mathcal{T}, n\in\{0,\ldots,N+1\}}$ be a solution to the following equation*

$$\frac{w_K^{n+1} - w_K^n}{\Delta t} = \zeta(w_K^{n+1}) - \gamma_n w_K^{n+1} + \sigma_n, \qquad \text{in } Q_{t_\mathrm{f}}, \tag{8}$$

where γ_n, $\sigma_n \geq 0$ for all $n \in \{0, \ldots, N+1\}$. Assume that $w_K^0 \geq 0$ for all $K \in \mathcal{V}$. Then, for all $K \in \mathcal{V}$ and all $n \in \{0, \ldots, n+1\}$, we have $w_K^n \geq 0$.

Sketch of the Proof We take a dual element ω_K such that $w_K^{n+1} = \min_{L\in\mathcal{V}}\{w_L^{n+1}\}$, then we multiply Eq. (8) by $-w_K^{n+1}$. Using the extension by $\zeta(0) \geq 0$ of the continuous function ζ for $w \leq 0$ and the nonnegativity of w_K^n (induction hypothesis), one can deduce that $w_K^{n+1} \geq 0$

We denote by $\mathcal{U}$ the vector-valued function defined by $\mathcal{U} = (f, q, r, s, u, v)$, and by $\mathcal{U}_{\mathcal{M},\Delta t}$ the corresponding piecewise constant reconstruction in $\mathcal{X}_{\mathcal{M},\Delta t}$.

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 *Let* $\left(\mathcal{U}_K^{n+1}\right)_{K\in\mathcal{V},n\in\{0,\ldots,N\}}$ *be a solution to the CVFE scheme* (5)–(6). *Then, for all n, and all* $K \in \mathcal{V}$, *we have* $v_K^n \geq 0$, *and* $0 \leq w_K^n \leq 1$, *for* $w = f, q, r, s, u$.

Now we give some discrete properties on the CVFE scheme (5)–(6).

Proposition 2 *For all* $n \geq 0$, *there exists a constant C independent of h such that*

$$\iint_{Q_{t_f}} \Lambda \nabla v_{\mathcal{T},\Delta t} \cdot \nabla v_{\mathcal{T},\Delta t} \mathrm{d}\mathbf{x}\,\mathrm{d}t = \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL}(v_K^{n+1} - v_L^{n+1})^2 \leq C. \tag{9}$$

$$\sum_{\sigma_{KL}\in\mathcal{E}} |\Lambda_{KL}| \left(f_K^{n^\star+1} - f_L^{n^\star+1}\right)^2 + \rho_1 \sum_{n=0}^{n^\star} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |\Lambda_{KL}| \left(f_K^{n+1} - f_L^{n+1}\right)^2 \leq C. \tag{10}$$

$$\sum_{K\in\mathcal{V}} m_K \left(u_K^{n^\star+1}\right)^2 + \sum_{n=0}^{n^\star} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2 \leq C. \tag{11}$$

We rely on the topological degree argument based on the above energy estimates in order to get the following existence result (see e.g. [1]).

Proposition 3 *Assume that* $\Delta t < \min\{\frac{1}{\lambda_q}, \frac{1}{\lambda_r}, \frac{1}{\lambda_s}\}$ *then there exists at least one solution* $\left(\mathcal{U}_K^{n+1}\right)_{K\in\mathcal{V}}$ *to the scheme* (5)–(6).

We are in a position to state the main result on the convergence of the nonlinear CVFE scheme (5)–(6) towards a weak solution of the continuous system (1)–(3) as the time and space discretization steps go to zero.

Theorem 1 *For all* $q \in [1, \infty)$, *the discrete solution* $\left(\mathcal{U}_{\mathcal{M}_m,\Delta t_m}\right)_m$ *converges (up to an unlabeled subsequence) in* $L^q\left(Q_{t_f}\right)$ *towards a weak solution of the continuous system* (1)–(3) *as* $m \to \infty$.

Sketch of the Proof We follow the same steps used in [5, 7] based on the use of Kolmogorov compactness criterion.

4 Numerical Result

In this section, we show a numerical simulation of the development of anisotropic breast cancer over a 2-D domain. For the numerical test, we have chosen $\eta_f = \eta_q = \eta_r = \eta_s = 0.5$. As estimated in [3, 4], we take $\lambda_u = 0.75$, $\lambda_q = 7.5 \times 10^{-3}$, $\lambda_r = 7.575 \times 10^{-3}$, and $\lambda_s = 0.0375$. We also define as in [3, 4] the mutation rates ρ_i in line with the mutation probabilities stated in the introduction $\rho_1 = 2 \times 10^{-7}$, $\rho_2 = 5 \times 10^{-6}$, $\rho_3 = 10^{-3}$, and $\rho_4 = 3 \times 10^{-2}$. We fix $\Delta t = 0.01$, $\alpha = 0.1$, $\beta = 0$, $a(u) = d_u u(1-u)$, $d_u = 0.0005$, $\mu(u) = \frac{\zeta}{d_u} u(1-u)$, $\chi(u) = d_v \times (u(1-u))^2$, and $d_v = 0.005$. Furthermore, we assume that initially all the cells and chemicals

Fig. 2 Initial condition for the healthy stem cells f: $0.7 \leq f_0 \leq 1$ (*left*) and for the other cell densities (*right*)

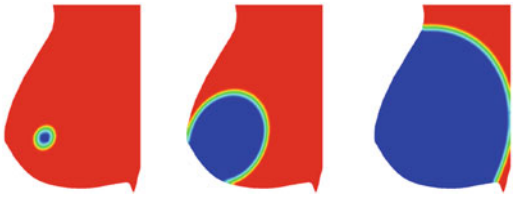


Fig. 3 Evolution of healthy stem cells f at $t = 13$: $0.03 \leq f \leq 1$ (*left*), at $t = 17.4$: $0 \leq f \leq 1$ (*center*), and at $t = 25$: $0 \leq f \leq 1$ (*right*)

Fig. 4 Evolution of tumor cells u at $t = 13$: $0 \leq u \leq 0.697$ (*left*), at $t = 17.4$: $0 \leq u \leq 1$ (*center*), and at $t = 25$: $0 \leq u \leq 1$ (*right*)

are equal to zero over the domain unless the healthy stem cells which are defined by regions, and we assume zero-flux boundary conditions. For instance, the healthy stem cell density f is initially defined by $f_0(\mathbf{x}, \mathbf{y}) = 0.7$ in the square region given by $(\mathbf{x}, \mathbf{y}) \in [2,\ 2.45] \times [2.85,\ 3.25]$ and 1 otherwise (see Fig. 2). The diffusion tensors are defined, for all $\mathbf{x}$, by $\Lambda(\mathbf{x}) = \begin{pmatrix} 7 & 2 \\ 2 & 10 \end{pmatrix}$, $D(\mathbf{x}) = d_v \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Figures 3 and 4 show the evolution of each of the healthy stem cells f and the tumor cells u at different moments. We see that after being formed, the tumor invades the healthy tissue by destroying the healthy stem cells f and it continues to spread quickly to the whole breast.

References

1. Cances, C., Guichard, C.: Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. Math. Comput. **85**(298), 549–580 (2016)
2. Cances, C., Ibrahim, M., Saad, M.: Positive nonlinear CVFE scheme for degenerate anisotropic Keller-Segel system. https://hal.archives-ouvertes.fr/hal-01119210 (2015)
3. Enderling, H., Anderson, A.R., Chaplain, M.A., Munro, A.J., Vaidya, J.S.: Mathematical modelling of radiotherapy strategies for early breast cancer. J. Theor. Biol. **241**(1), 158–171 (2006)
4. Enderling, H., Chaplain, M.A., Anderson, A.R., Vaidya, J.S.: A mathematical model of breast cancer development, local treatment and recurrence. J. Theor. Biol. **246**(2), 245–259 (2007)

5. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. 7, pp. 713–1020 (2000)
6. Foucher, F., Ibrahim, M., Saad, M.: Numerical analysis of a finite volume scheme for the simulation of a nonlinear degenerate breast cancer model (submitted for publication, 2016)
7. Gallouët, T., Latché, J.C.: Compactness of discrete approximate solutions to parabolic PDEs – application to a turbulence model. Commun. Pure Appl. Anal. **11**(6), 2371–2391 (2012)

Minisymposium: Fluid Instabilities and Transport Phenomena in Industrial Processes

Ricardo Barros

Description

This minisymposium focused on the mathematical modeling and analysis of processes occurring in industry (for example, food and drink, pharmaceutical, and oil sector). These were some of the challenging industrial problems currently under investigation at the Mathematics Application Consortium for Science and Industry (MACSI) in Ireland. The talks covered the development of innovative mathematical models that were intended to help industry optimize and improve processes. Models were devised for mass transfer in particulate systems, e.g. solid-liquid extraction, particle-laden viscous flows, and stability/instability problems in liquid bridges and two-phase flows. Asymptotic, analytical, and numerical techniques were used to investigate the solutions of the proposed mathematical models.

The talks included in the minisymposium were the following:

- K. Moroney. MACSI, University of Limerick (Ireland). *Asymptotic analysis of a coffee extraction model for a fixed coffee bed.*
- M. Haynes. MACSI, University of Limerick (Ireland). *Equivalence of stability methods for a horizontal liquid bridge.*
- R. Barros. MACSI, University of Limerick (Ireland). *Particle-laden viscous suspensions in an inclined channel.*
- S. Kaar. MACSI, University of Limerick (Ireland). *Separation of bubbles in Guinness leading to instability waves.*

R. Barros (✉)
MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
e-mail: ricardo.barros@ul.ie

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_34

- W. Arne. Fraunhofer ITWM (Germany). *Viscoelastic law for Cosserat rod models with application in rotational spinning processes.*
- J. Duley. MACSI, University of Limerick (Ireland). *A model for heterogeneous nucleation and short term growth of molecular clusters.*

Mathematical Modelling of Waves in Guinness

Simon Kaar, William Lee, and Stephen O'Brien

Abstract We provide a simple two-dimensional model of bubbly two-phase flow which can be used to investigate why waves form and propagate downward while a pint of Guinness is settling. We start out with the basic equations of the two-phase flow and use the large timescale difference of beer convection and rising bubbles in order to treat the convection flow as quasistatic. Using this argument we further simplify the two-phase mixture equations to that of a single liquid whose density varies with bubble concentration. A stability analysis shows that waves can occur through an instability analogous to the Kelvin-Helmholtz instability which forms in parallel shear flow. We provide a description of the form of these waves, and compare them to observations. Our theory provides a platform for the description of waves in more general bubbly two-phase shear flows.

1 Introduction

Separation processes are important in the chemical engineering of food, drink and pharmaceuticals. The settling of a freshly poured pint of Guinness or other stout beers provides an easily observable example of the complex flow patterns that can occur in even simple, gravity-driven separation. In a separating pint of Guinness the bubbles appear to descend in waves at the glass—this is readily observable since the white bubbles contrast strongly with the dark beer. Benilov et al. [1] explained the origin of the sinking bubbles by a simple physical mechanism of enhanced settling known as the Boycott effect [2] and confirmed their results by CFD simulation. Assuming the existence of the downwards flow responsible for the sinking bubbles, Robinson et al. [6] used a one dimensional models to propose a mechanism of wave formation due to a roll wave type instability. A framework which explains both

S. Kaar • S. O'Brien
MACSI, University of Limerick, Limerick, Ireland
e-mail: simon.kaar@ul.ie; stephen.obrien@ul.ie

W. Lee (✉)
MACSI, University of Limerick, Limerick, Ireland
Department of Mathematics, University of Portsmouth, Portsmouth, UK
e-mail: william.lee@port.ac.uk

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_35

sinking bubbles and the resulting waves is proposed here. We set up a similar model to that studied by Benilov et al., but investigate the stability of the resulting flow patterns. This differs from the approach taken by Robinson et al. in two respects: firstly the model is two dimensional rather than one dimensional, secondly sinking bubbles emerge naturally in this framework and do not need to be imposed on the model.

As is well known, in a typical glass of stout beer there is a circulation: the bubbly beer rises, except for a narrow band at the perimeter where it descends. The width of this faster downward flow is related to the rising bubbles and their simultaneous motion away from the side of the glass. We show below that the resulting shear flow is unstable and will generate observable waves.

A well known phenomena seen in parallel shear flows with continuously varying distributions of density and velocity is appearance of waves which form at positions where the shear is greatest. Waves may grow in time and lead to overturning billows. These breaking internal waves generate mixing over a height a little shorter than their wavelength. This phenomenon is known as the Kelvin-Helmholtz instability [4].

Across each horizontal cross-section, the bubbles are forced upward through the beer due to buoyancy, leaving pure beer beneath. However, for an upwardly inclined glass, the rising bubbles produce a band of pure beer at the perimeter also, which sinks due to its increased average density. It will be shown that this evolving horizontal density arrangement of bubbles, the void fraction $\alpha(y)$, where y is the horizontal dimension, gives rise to a convection circulation flow. Interestingly, this natural progress of the bubbles gives rise to the distributions of density and velocity within the glass which leads to the instability mentioned above. We created a mathematical model of this progression to determine at which velocity shear arrangement we first see waves.

It was important to verify that bubbles rise as a swarm in the manner assumed above. It is well known that a uniform distribution in the density or void fraction α of bubbles in a fluid is susceptible to voidage disturbances in the form of density waves. Bubbles move in accordance with a mass continuity law which is hyperbolic and shock forming. In our case the equation is non-linear as the bubble velocity is itself a function of the void fraction. This is well known and the exact form of non-linearity is determined empirically [7]. The bubbles rise at this velocity and using Rankine-Hugoniot jump conditions [5] it was shown that the trailing edge of rising bubbles moves as a shock at the same velocity.

In this paper we consider the early stages of settling where there exists a uniform distribution of bubbles throughout the glass. We build a mathematical model of the settling process, the evolving circulation flow and for the onset of waves brought about by this shear flow. We first developed the full two-phase mixture momentum equations from which we determined the dominant behaviours. This highlighted a macro timescale for circulation flow and a micro-timescale for bubble motion horizontally.

The macro scale equation, the value of the circulation velocity along the axis of the glass, as being and order of magnitude greater than the micro, the rise velocity

of bubbles relative to the beer, lead to the simplification of a quasistatic problem. At each instant the evolving density and velocity field is being treated as in equilibrium. Variation in $\alpha(y)$ horizontally drives the circulation due to buoyancy in a forward coupling [3]. The reverse coupling was not considered. This is the typical approach for problems where α is small.

To model the circulatory flow and investigate its stability we created a simplified model for the governing equations by treating the bubbly beer as a single liquid with a density field which varies according to the bubble void fraction. This model predicted a buoyancy driven base flow on which a linear stability analysis was performed. The procedure used was analogous to the derivation of the Orr-Sommerfeld equation [4] which is a fourth order differential equation whose dependent variable $\phi(y)$, is the amplitude of oscillation of the horizontal velocity component of the bubbly beer flow. In our case we calculated an extended form of this stability equation as we additionally accounted for variable density by including for variable density $\alpha(y)$. This resulted in an eigenproblem consisting of a set of two coupled differential equations. This generalised eigenproblem was solved numerically.

2 The Mathematical Model

We consider the case in which motions of bubbles perpendicular to the walls of the glass and the resulting circulatory flow parallel to the walls of the glass can be considered independently. This simplification can be effected due to the different timescales of the two motions. Bubble motion perpendicular to the walls of the glass is a slow process due to the small size of the bubbles, therefore we can treat the circulatory flow as being in equilibrium on the timescale over which density variations occur.

2.1 Bubble Volume Fraction

Here we take the bubble volume fraction to be

$$\alpha = \alpha_0 \left(1 - \frac{\cosh(y/w)}{\cosh(L/w)}\right), \tag{1}$$

where $\alpha_0 = 0.05$ is the typical bubbly beer fraction, L the distance from the centreline to the perimeter of the glass and w a shape parameter describing the width of the bubble depleted region close to the wall of the glass. By incrementing w we simulate the horizontal propagation of bubbles.

2.2 Single Liquid Equations

In modelling the circulatory flow parallel to the walls of the glass we assume the motion of both bubbles and beer can be described by a single velocity field. This assumption is again due to the small size of the bubbles: the motion of the bubbles relative to the fluid will be much smaller in magnitude than the circulatory flow.

The governing equations for the single liquid layer are the Navier Stokes and bubble conservation equations. These can be non-dimensionalised using length, velocity and time scales taken to be L the distance from the centreline to glass, U_0 a velocity scale derived by balancing viscous and buoyancy forces, and $t_0 = L/U_0$. In nondimensional form and component notation these are

$$\frac{\partial u_i}{\partial x_i} = 0, \tag{2}$$

$$\frac{\partial \alpha}{\partial t} + \frac{\partial}{\partial x_i}(\alpha u_i) = 0, \tag{3}$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{x_j} = -\frac{\partial p}{\partial x_i} - \frac{\alpha \hat{g}_i}{\mathrm{Fr}} + \frac{1}{\mathrm{Re}} \frac{\partial^2 u_i}{\partial x_j^2}, \tag{4}$$

where u_i are the velocity components, p is a reduced pressure, α the bubble void fraction, $\hat{g}_i$ is a unit vector in the direction of gravity and Fr and Re are Froude and Reynolds numbers, respectively.

2.3 Base Velocity

We assume a steady state parallel shear flow where $U(y)$ is the steady state base velocity. The single liquid equations (2)–(4), reduce to

$$0 = -\frac{\partial P}{\partial x} - \frac{A}{\mathrm{Fr}} + \frac{1}{\mathrm{Re}} \frac{\partial^2 U}{\partial y^2}, \tag{5}$$

where A is the dimensionless bubble density derived from Eq. (1), shown in Fig. 1, which together with no slip boundary conditions at the wall and a condition (due to incompressibility) of zero net flow through a horizontal surface, result in the velocity profile shown in Fig. 2.

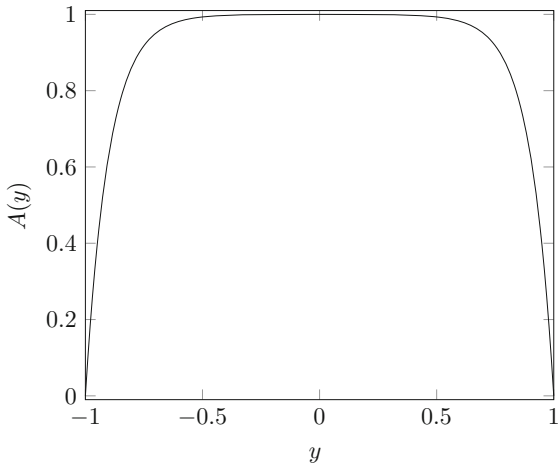


Fig. 1 Base bubble volume fraction, A

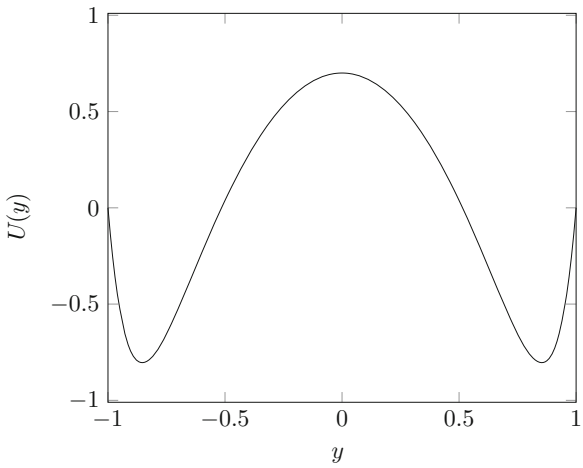


Fig. 2 Base flow field, U

3 Linear Stability Analysis

The governing equations (2)–(4) were linearised about the base velocity $U(y)$, base bubble density $A(y)$ and base pressure $P(x, y)$. Substitutions were

$$u = U + fe^{ik(x-\lambda t)}, \tag{6}$$

$$v = \phi e^{ik(x-\lambda t)}, \tag{7}$$

$$p = P + \pi e^{ik(x-\lambda t)}, \tag{8}$$

$$\alpha = A + \nu e^{ik(x-\lambda t)}, \tag{9}$$

where ϕ, f, π and ν are disturbance amplitudes.

Following a similar procedure to that used to derive the Orr-Sommerfeld equation [4], f and π can be eliminated to produce an eigenproblem consisting of two coupled ordinary differential equations

$$\left(\frac{\partial^2\phi}{\partial y^2} - k^2\phi\right)\lambda = U\frac{\partial^2\phi}{\partial y^2} - \phi\left(\frac{\partial^2 U}{\partial y^2} + \mathrm{Re}k^2 U\right) + \frac{1}{Fr}\frac{\partial \nu}{\partial y}, \\ - \frac{1}{ik\mathrm{Re}}\left(\frac{\partial^4\phi}{\partial y^4} - 2k^2\frac{\partial^2\phi}{\partial y^2} + k^4\phi\right) \tag{10}$$

$$\nu\lambda = \nu U + \frac{1}{ik}\frac{\partial A}{\partial y}, \tag{11}$$

where $\phi(y)$ is the amplitude of cross-stream velocity disturbances and $\nu(y)$ the amplitude of density disturbances, k is the wavenumber.

This eigenvalue problem can be solved numerically. A second order finite difference method was used to set up a block matrix generalised eigenvalue problem

$$\begin{pmatrix} \mathbf{M}_{\psi\psi} & \mathbf{M}_{\psi\nu} \\ \mathbf{M}_{\nu\psi} & \mathbf{M}_{\nu\nu} \end{pmatrix}\begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{\nu} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{N}_{\psi\psi} & \mathbf{N}_{\psi\nu} \\ \mathbf{N}_{\nu\psi} & \mathbf{N}_{\nu\nu} \end{pmatrix}\begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{\nu} \end{pmatrix}. \tag{12}$$

The eigenvalue problem was solved for one base density $A(y)$ and one k value and this resulted in an unstable mode whose amplitude is shown in Fig. 3. This shows qualitative features in common with the waves observed in a settling pint of Guinness, namely regions of enhanced and reduced bubble density close to the walls of the glass.

4 Conclusion

We have developed a simple framework for investigating the phenomena of sinking waves of bubbles observed in a settling glass of Guinness. Time independent solutions of our model show a downwards flow and thus sinking bubbles close the wall of the glass. In addition we have shown that this flow is unstable and that one unstable mode consists of bubble density waves similar to those seen in a pint of settling Guinness. In future work we plan a systematic survey of parameter space to determine the exact point at which the flow becomes unstable.

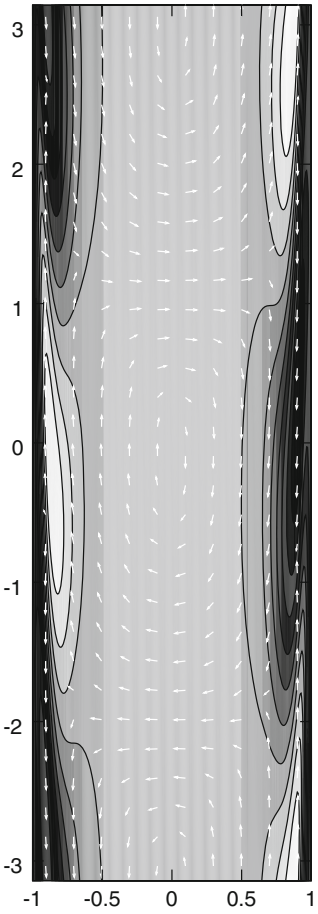


Fig. 3 Eigenfunction plot for flow velocity vector and density. Bubble density is shown by the colouring with *dark colour* indicating low bubble density. The velocity field is shown by *white arrows* (only the perturbing velocity field is shown, not the base velocity)

Acknowledgements The authors acknowledge the support of MACSI, the Mathematics Applications Consortium for Science and Industry (www.macsi.ul.ie), funded by the Science Foundation Ireland Investigator Award 12/IA/1683.

References

1. Benilov, E.S., Cummins, C.P., Lee, W.T.: Why do bubbles in Guinness sink? Am. J. Phys. **81**, 88 (2013)
2. Boycott, A.E.: Sedimentation of blood corpuscles. Nature **104**, 532 (1920)
3. Brennen, C.E.: Fundamentals of Multiphase Flow. Cambridge University Press, Cambridge (2005)
4. Drazin, P.G., Reid, W.H.: Hydrodynamic Stability. Cambridge University Press, Cambridge (2004)

5. Rankine, W.J.M.: On the thermodynamic theory of waves of finite longitudinal disturbances. Philos. Trans. R. Soc. **160**, 277–288 (1870)
6. Robinson, M., Fowler, A.C., Alexander, A.J., O'Brien, S.B.G.: Waves in Guinness. Phys. Fluids **20**, 067101 (2008)
7. Wallis, G.B.: One-Dimensional Two-Phase Flow. McGraw-Hill, New York (1969)

Viscoelastic Cosserat Rod Model for Spinning Processes

Walter Arne, Nicole Marheineke, and Raimund Wegener

Abstract Embedded in the special Cosserat theory (slender-body theory) we propose an incompressible viscoelastic rod model that covers viscous and elastic behavior in the asymptotic limits. Its applicability is demonstrated in an industrial fiber spinning process.

1 Introduction

In spinning processes a hot liquid jet is extruded from a nozzle and mechanically or aerodynamically stretched. By cooling down it solidifies and forms an elastic inextensible thin fiber. Process simulation and design require an appropriate material modeling that covers the full range from the viscous jet behavior at the nozzle to the elastic behavior of the final fiber. Making use of a Maxwell-like relaxation and a generalized Kirchhoff constraint, we propose a viscoelastic special Cosserat rod model that contains the well-known incompressible viscous and elastic rods [1, 7, 9] as asymptotic limits. We show its applicability for an industrial process.

2 Viscoelastic Cosserat Rod and Its Limit Models

A special Cosserat rod [1] in the three-dimensional Euclidean space $\mathbb{E}^3$ is characterized by a curve $\mathbf{r} : Q \to \mathbb{E}^3$ and an orthonormal director triad $\{\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}\} : Q \to SO(3)$ specifying the position and the orientation of the cross-sections. In $Q = \{(s,t) \in \mathbb{R}^2 \,|\, s \in [s_a(t), s_b(t)],\ t > 0\}$, s denotes the material parameter and t the time. The behavior of the rod is generally described by the kinematic and

W. Arne • R. Wegener
Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
e-mail: walter.arne@itwm.fraunhofer.de; raimund.wegener@itwm.fraunhofer.de

N. Marheineke (✉)
FAU Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: marheineke@math.fau.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_36

dynamic equations

$$\begin{aligned} \partial_t \mathbf{r} &= \mathbf{v}, & \partial_s \mathbf{r} &= \boldsymbol{\tau} \\ \partial_t \mathbf{d_k} &= \boldsymbol{\omega} \times \mathbf{d_k}, & \partial_s \mathbf{d_k} &= \boldsymbol{\kappa} \times \mathbf{d_k} \\ (\rho A)\, \partial_t \mathbf{v} &= \partial_s \mathbf{n} + \mathbf{f} \\ \partial_t \mathbf{h} &= \partial_s \mathbf{m} + \boldsymbol{\tau} \times \mathbf{n} \end{aligned} \tag{1}$$

with tangent $\boldsymbol{\tau}$, curvature $\boldsymbol{\kappa}$, linear and angular velocity $\mathbf{v}$, $\boldsymbol{\omega}$. Here, (ρA) denotes the line density. Constitutive laws for the inner forces $\mathbf{n}$ and couples $\mathbf{m}$, a description of the outer forces $\mathbf{f}$, a geometric model for the angular momentum $\mathbf{h}$ as well as appropriate initial and boundary conditions specify and complete (1).

Remark 1 Alternatively to the Lagrangian (material) description used in (1) the rod model can be also formulated in other parametrizations, e.g., in the Eulerian (spatial) parameterization with the arc-length as parameter. Since the transformation is in general time-dependent such a re-formulation implies convective terms in the equations [9]. A re-parametrization is especially advantageous if stationarity can be achieved (cf. Sect. 3).

Remark 2 To discuss the material laws it is convenient to formulate (1) in the director basis. Therefore, we introduce a fixed outer orthonormal basis $\{\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}\}$, then $\mathbf{z} = \sum z_k \mathbf{d_k} = \sum \bar{z}_k \mathbf{e_k} \in \mathbb{E}^3$ with the respective component triples $\mathsf{z} = (z_1, z_2, z_3)$ and $\bar{\mathsf{z}} = (\bar{z}_1, \bar{z}_2, \bar{z}_3) \in \mathbb{R}^3$. The director basis is associated with an orthogonal matrix $\mathsf{D} = (D_{ij}) = (\mathbf{d_i} \cdot \mathbf{e_j}) \in SO(3)$. For the coordinate tupels, the relations $\mathsf{D} \cdot \bar{\mathsf{z}} = \mathsf{z}$ as well as $\mathsf{D} \cdot \partial_t \bar{\mathsf{z}} = \partial_t \mathsf{z} + \omega \times \mathsf{z}$ and $\mathsf{D} \cdot \partial_s \bar{\mathsf{z}} = \partial_s \mathsf{z} + \kappa \times \mathsf{z}$ hold.

Characteristic Numbers In dimensionless form the rod model is specified by characteristic numbers. To make (1) dimensionless we choose typical values for line density $(\rho A)_0$, velocity magnitude v_0, rod diameter d_0 and length r_0. Due to the two length scales the model framework contains the so-called slenderness ratio

$$\epsilon = \frac{d_0}{r_0}.$$

We obtain the dimensionless model quantities from the dimensional ones (indicated by *) by scaling with respective reference values (indicated by $_r$), i.e., $\mathsf{x} = \mathsf{x}^\star / x_r$. We particularly use

$$s_r = r_0, \quad t_r = \frac{r_0}{v_0}, \quad (\rho A)_r = (\rho A)_0, \quad (\tau_3 A)_r = d_0^2, \quad A_r = d_0^2, \quad f_r = (\rho A)_0 \frac{v_0^2}{r_0}$$

$$r_r = r_0, \quad \kappa_r = \frac{1}{r_0}, \quad v_r = v_0, \quad \omega_r = \frac{v_0}{r_0}, \quad n_r = (\rho A)_0 v_0^2, \quad m_r = (\rho A)_0 v_0^2 r_0.$$

Apart from the mass line density (ρA), the volume line density is here introduced as $(\tau_3 A)$. This notation is motivated from the fact that we exclusively consider

Kirchhoff-type materials ($\tau_1 = \tau_2 = 0$) in this paper. Hence, $(\tau_3 A)/\tau_3 = A$ is the actual cross-sectional area at (s, t).

In view of the material modeling we deal with three further characteristic numbers. For viscous behavior we introduce the typical dynamical viscosity μ_0 which yields two viscosity scales and implies the Reynolds number Re. For elastic behavior we introduce the typical Young's modulus E_0 which yields two velocity scales and implies the Mach number Ma. The relation of Young's E and shear modulus G is thereby expressed by the Poisson number ν_P, i.e., $E/G = 2(1 + \nu_P)$. Finally, for viscoelastic behavior we introduce the typical relaxation time θ_0 that we choose in dependence of μ_0 and E_0. This results in two time scales and implies the Deborah number De as function of Re and Ma, i.e.,

$$\mathrm{Re} = \frac{(\rho A)_0 v_0 r_0}{d_0^2 \mu_0}, \qquad \mathrm{Ma} = \frac{v_0}{d_0}\sqrt{\frac{(\rho A)_0}{E_0}}, \qquad \mathrm{De} = \frac{\mu_0/E_0}{r_0/v_0} = \frac{\mathrm{Ma}^2}{\mathrm{Re}}.$$

The reference values are taken as

$$\mu_r = \mu_0, \quad E_r = E_0, \quad G_r = E_0, \quad \theta_r = \theta_0.$$

In three-dimensional incompressible material modeling the relation $\mu/\theta = G = E/3$ holds as $\nu_P = 1/2$.

Viscoelastic Rod For the simulation and design of industrial fiber spinning processes we aim a rod model for viscoelastic incompressible materials that satisfies the viscous and elastic behaviors in the asymptotic limit cases. We propose the following viscoelastic rod model in Lagrangian parametrization that is formulated in the director basis (cf. Remark 2), i.e.,

$$\begin{aligned}
\mathsf{D}\cdot\partial_t\bar{\mathsf{r}} &= \mathsf{v}, & \mathsf{D}\cdot\partial_s\bar{\mathsf{r}} &= \tau\\
\partial_t\mathsf{D} &= -\omega\times\mathsf{D}, & \partial_s\mathsf{D} &= -\kappa\times\mathsf{D}\\
\partial_t((\rho A)\mathsf{v}) + (\rho A)\omega\times\mathsf{v} &= \partial_s\mathsf{n} + \kappa\times\mathsf{n} + \mathsf{D}\cdot\bar{\mathsf{f}}\\
\partial_t(\epsilon^2(\rho A)\frac{(\tau_3 A)}{\tau_3}\mathsf{M}_\mathsf{i}\cdot\omega) + \epsilon^2(\rho A)\frac{(\tau_3 A)}{\tau_3}\omega\times(\mathsf{M}_\mathsf{i}\cdot\omega) &= \partial_s\mathsf{m} + \kappa\times\mathsf{m} + \tau\times\mathsf{n}
\end{aligned}$$

with material law

$$\begin{aligned}
\tau_1 = \tau_2 = 0, \qquad \partial_t\tau_3 &= \frac{\mathrm{Re}}{3}\frac{1}{\mu}\frac{\tau_3^2}{(\tau_3 A)}n_3\\
\mathrm{De}\,\theta\,\partial_t\mathsf{m} + \mathsf{m} &= \frac{3\epsilon^2}{\mathrm{Re}}\mu\frac{(\tau_3 A)^2}{\tau_3^3}\mathsf{M}_\mathsf{v}\cdot\partial_t\kappa,
\end{aligned} \tag{2}$$

and inertia- and viscosity-associated matrices M_i, M_v. In case of homogeneous circular cross-sections these matrices have the special form $\mathsf{M}_\mathsf{i} = \mathsf{L}_2$ and $\mathsf{M}_\mathsf{v} = \mathsf{L}_{2/3}$

with $\mathsf{L_a} = \mathrm{diag}(1,1,a)/(4\pi)$, $a \in \mathbb{R}$. Note that (ρA), $(\tau_3 A)$, μ and θ are either prescribed material functions or modeled by further evolution equations.

In the limit $\mathrm{De} \to 0$ we obtain the viscous incompressible rod model of [2, 7] since in (2) the material equation for the inner couple m is a Maxwell-like relaxation of the viscous constitutive law and the inner forces n are determined by the generalized Kirchhoff constraint with the viscous relation between the tangential force n_3 and the elongation τ_3. The underlying viscous laws were originally proposed in [7, 8] and investigated as well as successfully applied to fiber spinning in, among others, [3, 4], i.e.,

$$\tau_1 = \tau_2 = 0, \qquad \partial_t \tau_3 = \frac{\mathrm{Re}}{3}\frac{1}{\mu}\frac{\tau_3^2}{(\tau_3 A)} n_3, \qquad \mathsf{m} = \frac{3\epsilon^2}{\mathrm{Re}}\mu\frac{(\tau_3 A)^2}{\tau_3^3}\mathsf{M_v}\cdot\partial_t\kappa. \tag{3}$$

The material functions of the viscous rod are (ρA), $(\tau_3 A)$ and μ.

In the limit $\mathrm{De} \to \infty$, $\mathrm{Re} \to 0$ with $\mathrm{De}\,\mathrm{Re} = \mathrm{Ma}^2$ and $\mu/\theta = E/3$ we obtain an inextensible, unshearable elastic Euler-Bernoulli beam model, considering the incompressible case $\nu_P = 1/2$. Thereby, not only the material laws change but also the balance equation for the angular momentum, i.e.,

$$\begin{aligned} \partial_t(\epsilon^2(\rho A)A\,\mathsf{M_i}\cdot\omega) + \epsilon^2(\rho A)A\,\omega\times(\mathsf{M_i}\cdot\omega) &= \partial_s\mathsf{m} + \kappa\times\mathsf{m} + \tau\times\mathsf{n} \\ \tau = \tau_3^\circ\mathsf{e}_3, \qquad \mathsf{m} &= \frac{\epsilon^2}{\mathrm{Ma}^2}\frac{EA^2}{\tau_3^\circ}\mathsf{M_e}\cdot(\kappa-\kappa^\circ). \end{aligned} \tag{4}$$

In this elastic limit we have $\partial_t((\tau_3 A)/\tau_3) = 0$ implying a time-independent cross-sectional area A. Moreover $\mathsf{M_v}$ becomes the elasticity-associated matrix $\mathsf{M_e}$ which admits the form $\mathsf{M_e} = \mathsf{L}_{1/(1+\nu_P)}$ in case of homogeneous circular cross-sections, $\nu_P = 1/2$. The material functions of the elastic rod are (ρA), A, E as well as τ_3° and κ°, where τ_3° and κ° are the referential quantities (integration constants). It is convenient to insert the Kirchhoff constraint $\tau = \tau_3^\circ\mathsf{e}_3$ in the remaining model equations. The classical Kirchhoff beam [1] (also known as Kirchhoff-Love equations) where the inertia term in the angular momentum balance is additionally neglected is contained in the resulting elastic rod model. It is given by the slenderness-low Mach number limit as $\epsilon \to 0$, $\mathrm{Ma} \to 0$ and $\epsilon/\mathrm{Ma} = const$ (cf. [5]).

Remark 3 In applications temperature effects often play a crucial role. The Cosserat rod system is then extended by a balance for the energy (temperature T), i.e.,

$$c_p\,\partial_t((\rho A)T) = q$$

with heat capacity c_p and heat source q. Consequently, the material functions θ, μ and E are modeled as temperature-dependent.

3 Industrial Application and Numerical Results

In the exemplarily considered spinning process hot polymer (polypropylene) jets are continuously extruded from the nozzles in vertical (gravitational) direction. In the machine they are stretched and cooled down by a surrounding air flow. The machine has a channel-like geometry with symmetric air inlets at the sides, the spinning set-up is time-independent and homogeneous in channel direction. The respective unloaded stationary two-dimensional flow field is visualized in Fig. 1.

Fiber-Flow Simulation A stationary Eulerian description is the basis for the fiber-flow simulations that we perform similarly to the procedures (models and methods) stated in [3]. For simplicity we neglect here the impact of the fibers on the flow and apply in this sense an one-way coupling. The air flow is described by the incompressible isothermal Navier-Stokes equations and computed with the commercial software tool ANSYS Fluent. We have here the inlet speed $v_{air} = 0.5$ [m/s] and the temperature $T_{air} = 290$ [K] (Fig. 1). The fibers' behavior is driven by gravity and aerodynamics. Focusing on their thinning and solidifying, we consider fiber jets of fixed length l with prescribed inflow properties (straight, non-stretched) at the nozzle and stress-free end, in particular we have $(\rho A)_{noz} = 1.13 \cdot 10^{-4}$ [kg/m], $v_{noz} = 2.1 \cdot 10^{-4}$ [m/s], $T_{noz} = 496$ [K], $d_{noz} = 4.0 \cdot 10^{-4}$ [m] and $l = 3.85$ [m]. In this Eulerian setting the Cosserat rod model becomes a boundary value problem of spatial ordinary differential equations which is extended by an additional balance for the temperature T (cf. Remark 3). For the used models of the aerodynamic heat source q_{air} and the drag forces f_{air} we refer to [3, 9]. We prescribe the material functions θ, μ and E of the polypropylene jets

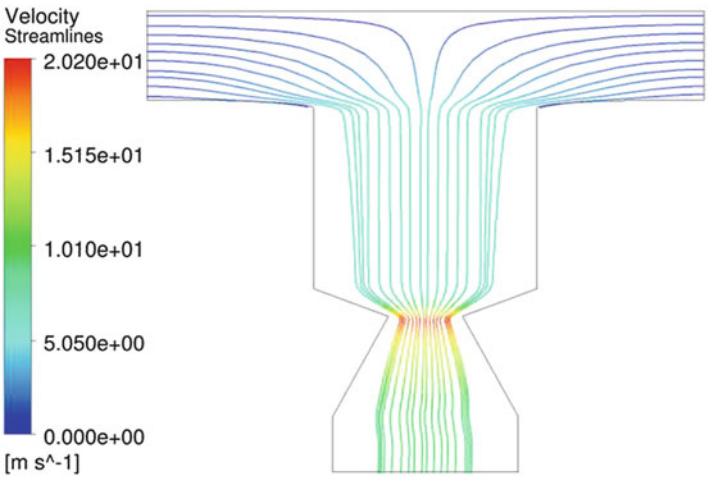


Fig. 1 Aerodynamic flow field in machine geometry of a spinning process, homogeneity in third direction. Streamlines are *colored* with respect to the velocity magnitude

as temperature-dependent. Thereby, we particularly apply an Arrhenius law for the viscosity $\mu(T) = c_1 \exp(c_2/T)$ [kg/(m s)] and a constant Young's modulus $E = 10^9$ [kg/(m s^2)], which implies a model for the relaxation time due to the incompressibility, $\theta(T) = 3\mu(T)/E$ [s]. The numerical computation of the fiber jets is performed in MATLAB using the routine `bvp4c.m`. This is a collocation method with a Runge-Kutta scheme of fourth order that we combine with a continuation strategy (homotopy approach) [9].

Viscoelastic Fiber Behavior In the spinning process the machine geometry has a constriction where the velocity of the air flow increases and through which the fiber jets are pulled. At this point the fibers are solidified and show an inextensible elastic behavior. Due to their bending ability they pass the constriction without hitting the machine walls.

Modeling the fiber jets as viscous rods (3) as done in [3, 4] is certainly adequate close to the nozzle. However, the numerical results in Fig. 2 (top) clearly stress the fact that the jets are not capable of changing and adapting their curvatures with respect to the streamlines of the flow field. After the jets' entering the machine geometry, the curvature keeps the same. Thus, it is no surprising that

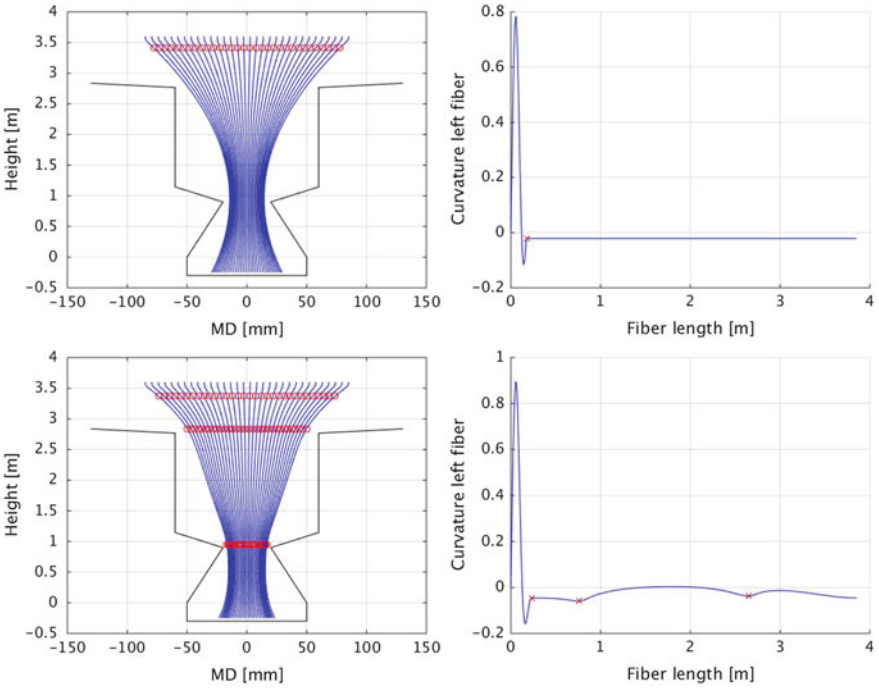


Fig. 2 *Left*: Fibers' behavior in machine geometry of a spinning process (32 fiber jets). *Right*: Curvature of the left fiber in the bundle. The *red markers* indicate the positions of the specific curvature behavior in the machine. Viscous fibers (*top*); viscoelastic fibers (*bottom*)

process simulations might fail because of predicting wrong fiber-wall collisions. Our proposed viscoelastic rod model (2) overcomes this limitation. By the continuous temperature-dependent transition from the viscous to the elastic inextensible behavior, it allows for the practically relevant bending effects of the fibers. As shown in Fig. 2 (bottom), the fiber curvature changes appropriately along the machine geometry.

4 Conclusion

Summing up, our proposed viscoelastic Cosserat rod model enables the monolithic numerical simulation and design of spinning processes from nozzle to deposition, since it covers incompressible viscous as well as inextensible elastic behavior in the asymptotic limits. Certainly one could also think of developing a compressible rod model (cf. [6]). Replacing the generalized Kirchhoff constraint by a Maxwell-like relaxation of the inner forces might result in a Timoshenko-like beam model in the elastic limit. But such a beam model contains shear and expansion waves and is not adequate for the thin fibers in the considered application at hand.

Acknowledgements This work has been supported by German DFG, project 251706852, MA 4526/2-1, WE 2003/4-1.

References

1. Antman, S.S.: Nonlinear Problems of Elasticity. Springer, New York (2006)
2. 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)
3. Arne, W., Marheineke, N., Schnebele, J., Wegener, R.: Fluid-fiber-interactions in rotational spinning process of glass wool manufacturing. J. Math. Ind. **1**, 2 (2011)
4. Arne, W., Marheineke, N., Meister, A., Schiessl, S., Wegener, R.: Finite volume approach for the instationary Cosserat rod model describing the spinning of viscous jets. J. Comput. Phys. **294**, 20–37 (2015)
5. Baus, F., Klar, A., Marheineke, N., Wegener, R.: Low Mach-number–slenderness limit for elastic Cosserat rods (2015). arXiv:1507.03432
6. Meligan, K.: Asymptotical and Numerical Analysis of Viscoelastic Material Laws for Fiber-Modelling. Master's thesis, FAU Erlangen-Nürnberg (2016)
7. Ribe, N.M.: Coiling of viscous jets. Proc. R. Soc. Lond. A **2051**, 3223–3239 (2004)
8. Ribe, N.M., Habibi M., Bonn, D.: Stability of liquid rope coiling. Phys. Fluids **18**, 084102 (2006)
9. Wegener, R., Marheineke, N., Hietel, D.: Virtual production of filaments and fleeces. In: Neunzert, H., Prätzel-Wolters, D. (eds.) Currents in Industrial Mathematics: From Concepts to Research to Education, pp. 103–162. Springer, Heidelberg (2015)

Minisymposium: Masters in Industrial Mathematics. Overview and Analysis of Graduates and Business Collaborators

Elena Vázquez-Cendón, Carlos Vázquez, José Durany, Manuel Carretero, and Fernando Varas

Description

To educate in Industrial Mathematics is one of the main objectives of ECMI. In the past editions of the ECMI conferences the different training opportunities offered by institutions have been analyzed.

The aim of this minisymposium was to propose a space for debate in which those responsible for the different titles in Industrial Mathematics could share with researchers in the field the new contributions that are being offered on the international scene.

We suggested that presentations dealt with the visions graduates had once their education was over and they entered the working world. Moreover, in many of these qualifications companies collaborate as teachers, internship tutors and suggesting master dissertations. We think it is extremely important to share and analyze the

E. Vázquez-Cendón (✉)
University of Santiago de Compostela, Santiago de Compostela, Spain
e-mail: elena.vazquez.cendon@usc.es

C. Vázquez
University of A Coruña, A Coruña, Spain
e-mail: carlosv@udc.es

J. Durany
University of Vigo, Vigo, Spain
e-mail: durany@dma.uvigo.es

M. Carretero
Carlos III University of Madrid, Madrid, Spain
e-mail: manuel.carretero@uc3m.es

F. Varas
Technical University of Madrid, Madrid, Spain
e-mail: fernando.varas@upm.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_37

opinions of graduates and companies under the ECMI umbrella as we believe collaboration in industrial mathematics is so important.

The talks included in the minisymposium were the following:

- T. Kauranne. Lappeenranta University of Technology (Finland). *International and double degree master studies and industrial collaboration in Technomathematics: a Finnish-Russian-African-Indian perspective.*
- E. Lindner. Institute of Computational Mathematics, Johannes Kepler University Linz (Austria). *Technical mathematics and industrial mathematics at Johannes Kepler University Linz (JKU).*
- O. López-Pouso, Department of Applied Mathematics, Universidade de Santiago de Compostela (Spain). *ECCUM, the creation of a Central Asian MSc with European cooperation.*
- W. Okrasiński. Faculty of Pure and Applied Mathematics, Wrocław University of Technology (Poland). *Interdisciplinary approach to the mathematical modelling in a high school – an example from Poland.*
- E. Vázquez-Cendón. Department of Applied Mathematics, Universidade de Santiago de Compostela (Spain). *Master in Industrial Mathematics (M2i). Learning 4.0 cross-institutional education to solve real problems in industry.*
- S. Abreu. Department of Mathematics, Polytechnic of Porto and LEMA, Laboratory for Mathematical Engineering, Porto (Portugal). *The Master of Applied Mathematics to Engineering and Finance of the School of Engineering, Polytechnic of Porto, Portugal.*
- T. Mendonça. Departamento de Matemática, Faculdade de Ciências da Universidade do Porto (Portugal). *The Master degree of Mathematical Engineering at the University of Porto.*
- J.A. Ferreira. CMUC, Department of Mathematics of University of Coimbra, (Portugal). *Training mathematician in the University of Coimbra: an ECMI master program.*

The Master Degree on Applied Mathematics to Engineering and Finance, School of Engineering, Polytechnic of Porto

Stella Abreu, José Matos, Manuel Cruz, Sandra Ramos, and Jorge Santos

Abstract The Master on Applied Mathematics to Engineering and Finance of the School of Engineering of the Polytechnic of Porto, Portugal, started its first edition in the scholar year of 2012/2013. This course is divided into four semesters (two scholar years). It emerged in an engineering school that has a strong tradition of dialog and collaboration with industry and business partners. In the second year, students must attend the subject of Dissertation/Project/Internship. This subject may be developed in a business or industrial environment. We will present some successful stories on applying this methodology in the above-mentioned master course, showing the benefits and the difficulties that arise from the collaboration with industry and business partners.

1 Introduction

The area of industrial mathematics has seen good developments in the last years, in Portugal. The role of academic institutions in this development is important. This master course was proposed for approval in the end of 2011 by the School of Engineering of the Polytechnic of Porto (ISEP). It emerged in accordance with the concerns of international organizations such as the Organisation for Economic Co-operation and Development (OECD). One of the recommendations of its Report on Mathematics in Industry from 2008 is (cf. [1], p. 18):

> Developing curriculum options that prepare the students for a career at the interface of mathematics and industry.

We also share the opinion, stated in that Report, that the curricula (cf. [1], p. 18):

> .. should be designed to demonstrate that the interaction of mathematics and industry leads to exciting research opportunities and benefits.

S. Abreu (✉) • J. Matos • M. Cruz • S. Ramos • J. Santos
ISEP/LEMA, School of Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal
e-mail: sau@isep.ipp.pt; jma@isep.ipp.pt; mbc@isep.ipp.pt; sfr@isep.ipp.pt; jms@isep.ipp.pt

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_38

The curriculum of this course takes into consideration that the candidates may arrive with different academic profiles, looking for broadening their knowledge and professional reconversion. It was designed to be attended by graduates, not only in Mathematics, but also in Engineering, Economic Sciences, Management and Accounting.

The main purpose of this course is to specialize professionals in Risk Analysis, Modelling and Computational Mathematics, Data Analysis and Classification, Actuarial Mathematics and Advanced Statistic Techniques, in order to solve technological and financial management problems of organizations.

Along the course, students develop skills in mathematical modelling, study and comparison of models, checking and interpreting data, presentation of results and implementing solutions. They also develop the ability to work with several computational tools, in order to solve problems.

2 The Curriculum

The curriculum of the course is described in Tables 1, 2, and 3. It integrates a subject on Cases Study and three Seminar subjects on applied mathematics to engineering and finance. These four subjects were conceived to make the master students aware of the wide scope of mathematics in the real world. Due to their background, some of the students may have had little contact with real world direct applications of mathematics, besides the traditional examples from banks, insurance or IT companies.

The subjects for the first semester of the first year are presented in Table 1.

In addition to the compulsory subjects, students must choose two options in the list: Database, Innovation and Entrepreneurship, Financial Investments, Data Mining, Neural Networks. All options of this and of the following semesters have 6 ECTS.

The Seminar I subject brings to the classroom professionals from several industrial areas that employ mathematicians. We also invite players from successful partnerships between industry and mathematics research centers. The main purpose for doing this is to convey to students the different points of view when looking at

Table 1 Subjects in the first semester of the first year

Subject	ECTS
Statistical experimentation and data analysis	6
Graphs and algorithms	4
Numerical methods in engineering	5
Seminar I	3
Option 1	6
Option 2	6

Table 2 Subjects in the second semester of the first year

Subject	ECTS
Analysis and risk management	6
Machine learning	5
Operations research, planning and optimization	5
Game theory and financial mathematics	5
Seminar II	3
Option	6

Table 3 Subjects in the first semester of the second year

Subject	ECTS
Cases Study	5
Seminar III	3
Option	6

a challenge: the industry point of view and the mathematical researchers' point of view.

In the second semester of the first year, the subjects are listed in Table 2.

One option must be chosen from the list: Multivariate Statistics for Engineering and Finance, Project Management, Process Modelling and Simulation, Forecasting Methods.

The Seminar II subject is almost entirely devoted to research related tools. For example, an introduction to LaTex, and making scientific presentations is usual on this subject. Main rules for writing a master thesis, citation rules, ethics and behavior on internships, soft skills, as well as other practical questions are also addressed, in order to prepare the students for the last part of the master.

The subjects for the first semester of the second year are listed in Table 3.

Students must also choose one option in the list: Data Warehouse and Analytical Processing, Actuarial and Financial Calculus, Imaging Metrology, Optimization and Decision Methods in Power Systems.

Seminar III is devoted to scientific research. As so, each student chooses a scientific paper (often related to some topic that was appealing to the student in Seminar I) to be read and to present at the end of the semester.

At the same time, students are also engaged on the subject of Cases Study, where they are exposed to some real world problems that may be addressed on a mathematical point of view. Support from researchers that worked on those industrial problems is welcome. In this sense, the preparation of this subject should be made in close relation with research groups that work in industrial mathematics. For students that, at this stage, already have an internship or a research area assigned, this subject may give an additional insight on their future work.

3 Graduates and Business Collaborators

At the beginning of the second year, besides the subjects presented in Table 3, students must attend the subject of Dissertation/ Project/ Internship, during two semesters, with a total of 46 ECTS. This subject may be developed in a business or industrial environment, in accordance with the Recommendations to Academia of the European Science Foundation, cf. [2], p. 33:

> Master programs should provide the option of specific problem solving on industrial internships with support from interdisciplinary scientists if necessary.

Our students are strongly motivated to attend the Internship because it prepares them for a career at the interface of mathematics and industry.

We will present some successful stories on applying this methodology.

3.1 Collaboration with a Portuguese Retail Bank

One of the students chose to do the internship in a Bank. He was submitted to tests and an interview before being accepted by that institution.

It was a 9-month full-time internship. The supervision was shared between two supervisors, one at the Bank and another one at ISEP. The bureaucracy due to some confidentiality issues was one of the constraints in this internship as it is usual with internships dealing with sensitive data regarding institutional or personal confidential information.

This work took place in the models area and automated decision processes of the Bank. Its goal was to develop a model to prevent and mitigate the risk of credit default. Foreseeing the potential credit default is critical in the way that it allows the Bank to implement preventive actions near customers.

The internship had three main stages:

- To understand the business, the Bank's informational system and the existing processes to prevent credit default.
- To make the analysis of a new database and to structure the information in order to integrate it in the Bank's informational system and to understand the information constraints.
- To develop a new model from the combination of the Bank's information and the new database, with the aim of making the credit default prevention even more effective.

The new model was based on the **scoring model** because it is a predictive and understandable modeling technique widely used and accepted in banking. It allows the introduction of expert rules to ensure business requirements (legal or operative).

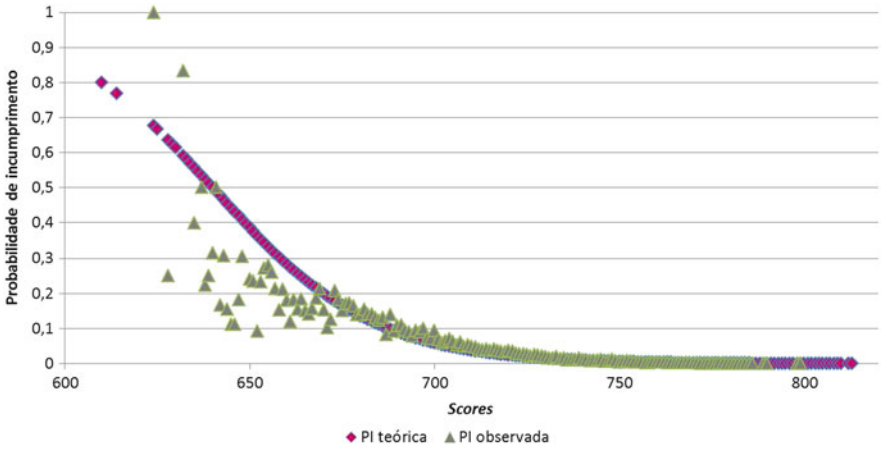


Fig. 1 Comparison between the theoretical (*diamonds*) and the observed (*triangles*) credit default. Score values below 670 correspond only to 0.6% of the population

The resulting new model was tested and it proved to be predictive, statistically stable and responding to the proposed objective. In Fig. 1, a comparison between the theoretical and the observed credit default is shown.

The new model was integrated in the Bank's informational system.

3.2 Collaboration with a Leading Company in the Portuguese Energy Sector

This was a 9-month, part-time internship. The supervision was shared between two supervisors, one at the Company and another one at ISEP.

The purpose of this work was to study, based on open access data, the behavior of the electric energy prices in Iberian Electricity Market (MIBEL), identifying the causes and constraints of this behavior.

The study included historical data resulting from the observation of different variables during a time period, such as historic values of prices and several types of energy production and weather predictions for wind speed, temperature, among others.

The internship took place in the Statistical and Econometric areas. A procedure that identifies variables with explanatory causality in the price model was developed.

The main difficulties were:

- The outcome variable (electric energy prices) is extremely sensitive to unexpected events, like economic and meteorological events that made it very challenging to find variables that explain the prices.

- Obtaining data. All data used in the study are open access data, available for specific time periods and with frequency variable.

The results and achievements were:

- A procedure to identify variables with explanatory causality in price was developed.
- The variables with causality relation with the price were incorporated in a mathematical model developed to forecast the electric energy prices. The model presents a good performance in terms of prediction errors.

3.3 Collaboration with a Multinational Industrial Corporation

One of our most fruitful collaborations was a full-time internship in a Multinational Industrial Corporation, with the duration of 9 months. After this internship, the collaboration (with ISEP and with the student) continued and still remains.

This internship was supervised by one person at the Company and two at ISEP. The goal was the development of a model for stock management for the automotive spare parts market.

The Operations Department is the main group of the Company which is responsible for the stock management in business. It processes daily a big amount of data in order to give an efficient response to the Company needs in terms of stock. The internship took place in the forecast and modelling areas. An automatic and configurable model for stock management was developed. It allows desirable quality conditions, namely, the service ratio.

The main challenges of this work were:

- The heterogeneity of data, in terms of demand, made it very challenging to find a model that fits all the references used by the Company.
- Second, the scalability and complexity of the model were a concern as there were more than 200,000 references to be analysed and the outputs had to be calculated in some acceptable time.
- The model should adapt easily if there will be a need to change the Company policy in terms of quality of service parameters.

The results and achievements were:

- A new mathematical model for stock management of spare parts was developed. The model incorporated different forecast methods in order to accommodate the large variability in demand of the references.
- A software tool was developed to implement automatically all the features that the Company asked in the beginning of the project, as well as some upgrades made during the development phase. An easy to use graphical interface was produced to operate the software.

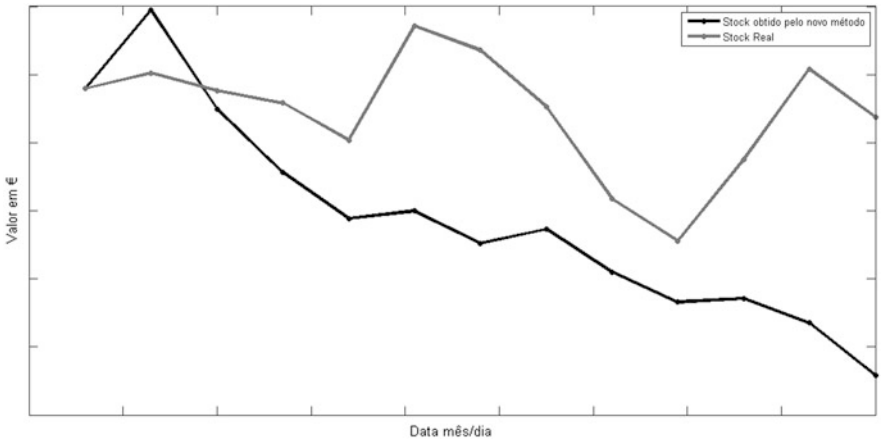


Fig. 2 Comparison between real stock (*gray line*) and stock obtained by new method (*black line*)

- The new model was integrated in the Company system. This integration has resulted in the reduction of stock, preserving, or increasing often, the service ratio (see Fig. 2).

Acknowledgements We thank LEMA (Engineering Mathematical Laboratory) for supporting this work.

References

1. Organisation for Economic Co-operation and Development. In: Report on Mathematics in Industry. Global Science Forum (2008). https://www.oecd.org/science/sci-tech/41019441.pdf. Cited 19 Oct 2016
2. European Science Foundation. In: Forward Look Mathematics and Industry. European Science Foundation (2010). www.esf.org/index.php?id=6264. Cited 19 Oct 2016

Minisymposium: Mathematical Modeling and Simulation for Nanoelectronic Coupled Problems (nanoCOPS)

Sebastian Schöps and Lihong Feng

Description

In this minisymposium recent advances in mathematical modeling, uncertainty quantification (UQ) and model order reduction (MOR) for electronic devices were presented. They are the research results from the cooperation between the academic and industrial partners in the framework of nanoCOPS ('Nanoelectronic Coupled Problems Solutions'). It is a collaborative research project within the research program 'Information and communication technologies' of the Seventh Framework Programme for Research and Technological Development (FP7) funded by the European Union.

The project addresses the development of efficient simulation/design tools for Power-MOS devices, with applications in energy harvesting, that involve couplings between electromagnetics, heat, stress, and radio-frequency circuitries in wireless communication, which involve electro-thermal coupling, electromagnetic-circuit-heat coupling and multirate behaviour, together with analogue-digital signals. Efficient mathematical modeling, UQ and MOR approaches act as important roles in the design processes, and were the focus of this minisymposium.

S. Schöps (✉)
Technische Universität Darmstadt, Graduate School of Computational Engineering, Dolivostraße 15, 64293 Darmstadt, Germany
e-mail: schoeps@gsc.tu-darmstadt.de

L. Feng
Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany
e-mail: feng@mpi-magdeburg.mpg.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_39

The talks included in the minisymposium were the following:

- Roland Pulch, Piotr Putek, Herbert de Gersem. Universität Greifswald (Germany), Bergische Universität Wuppertal (Germany) and Technische Universität Darmstadt (Germany). *Identification of probabilistic input data for modeling uncertainties.*
- Nicodemus Banagaaya, Peter Benner and Lihong Feng. Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg (Germany). *PMOR for Linear Systems with Many Inputs.*
- David Duque, Sebastian Schöps and Herbert De Gersem. Technische Universität Darmstadt (Germany). *An electric circuit description of the finite integration technique for electromagnetic problems.*
- Rick Janssen, E. Jan W. ter Maten, Piotr Putek, Roland Pulch, Caren Tischendorf, Wim Schoenmaker. NXP (the Netherlands), Bergische Universität Wuppertal (Germany), Universität Greifswald (Germany), Humboldt University Berlin (Germany) and Magwel NV (Belgium). *Nanoelectronic coupled problem solutions: methods and applications.*
- Peter Meuris, Wim Schoenmaker, Christian Strohm and Caren Tischendorf. Magwel NV (Belgium) and Humboldt University of Berlin (Germany). *Holistic Coupled Field and Circuit Simulation.*

Identification of Probabilistic Input Data for a Glue-Die-Package Problem

Roland Pulch, Piotr Putek, Herbert De Gersem, and Renaud Gillon

Abstract In mathematical models, physical or geometrical parameters often involve uncertainties due to measurement errors, estimations or imperfections of an industrial production. An uncertainty quantification can be performed by a stochastic description, where parameters are substituted by random variables or random processes. The probability distributions of the parameters have to be predetermined as an input to the stochastic model. However, the variability of input parameters often cannot be measured directly, whereas the output quantities are available. We consider a test problem from nanoelectronics, where a piece of glue connects a die and a package. The geometrical parameters as well as the material parameters are uncertain for the piece of glue. We fit the input probability distributions of the random parameters to measurements of the output, which represents a kind of inverse problem. For this purpose, a minimization problem is defined including a piecewise linear approximation of the cumulative distribution functions. We present numerical results for this test problem.

R. Pulch (✉)
Institute for Mathematics and Computer Science, Ernst-Moritz-Arndt-Universität Greifswald, Walther-Rathenau-Str. 47, 17489 Greifswald, Germany
e-mail: roland.pulch@uni-greifswald.de

P. Putek
Chair of Applied Mathematics and Numerical Analysis, Bergische Universität Wuppertal, Gauß-Str. 29, 42119 Wuppertal, Germany
e-mail: putek@math.uni-wuppertal.de

H. De Gersem
Computational Electromagnetics Laboratory (TEMF), Technische Universität Darmstadt, Schloßgartenstr. 8, 64283 Darmstadt, Germany
e-mail: degersem@temf.tu-darmstadt.de

R. Gillon
ON Semiconductor, Westerring 15, 9700 Oudenaarde, Belgium
e-mail: renaud.gillon@onsemi.com

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_40

1 Introduction

In mathematical models of physical or technical applications, the involved parameters often exhibit uncertainties. Concerning nanoelectronics, miniaturization causes variations of parameters in an industrial production, see [2, 9]. On the one hand, variations can be examined by a worst-case-corner-analysis, see [11]. On the other hand, uncertainty quantification is often based on stochastic models, where random variables or random processes replace varying parameters, see [10].

For a numerical simulation of a stochastic model, the probability distributions of the parameters have to be predetermined. Measurements of the input parameters are often not feasible. Alternatively, measurements of output quantities are accessible. Thus we want to fit probabilistic input data to measurements of outputs. This task can be seen as an inverse problem or a backward problem instead of the forward problem, where inputs are propagated to outputs.

We discuss a test problem relevant for the production of electronic circuits. A simplified model of a piece of glue between a die and a package is considered, which is also discussed briefly in [8]. The material parameters and the geometrical parameters of the glue are uncertain. The aim is to identify the probability distributions of five random parameters.

Unknown parameters of a probability distribution can be determined by the method of moments, the maximum likelihood principle or Bayesian estimations, see [5]. Polynomials of Gaussian-distributed variables are fitted by Fleishman's approach, cf. [3], for example. However, these methods require that the type of the probability distribution (uniform, Gaussian, etc.) is known a priori. Alternatively, we construct a numerical technique for the fitting of arbitrary cumulative distribution functions, where just the continuity of the function is assumed. The method employs a piecewise linear approximation of the cumulative distribution function. The approach yields a minimization problem, more precisely a nonlinear least squares problem. We apply pseudo random numbers, see [7], for both the evaluation of the objective function by sampling and the determination of grid points in the domain of the unknowns.

Finally, we apply this numerical method to a glue-die-package problem. Therein, measurements are still generated artificially. The results demonstrate the feasibility of the technique.

2 Problem Definition

In the production of electronic devices, glue is applied to hold a die and a package together. On the one hand, metallic parts and semiconducting parts of devices can often be produced with a high precision. On the other hand, a piece of glue represents a cheap substance, where both the material parameters and the formation

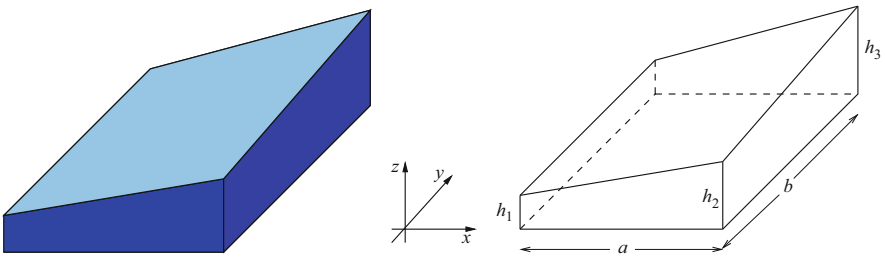


Fig. 1 Simplified geometry of a piece of glue connecting a die and a package (figure from [8])

Table 1 Parameters in glue-die-package problem

Parameter	Symbol	Unit	Minimum value	Maximum value
Thermal conductivity	λ	(W/m K)	1.6	2.4
Volumetric heat capacity	ρc	(J/m^3 K)	$2.5 \cdot 10^6$	$3.75 \cdot 10^6$
Average height	h_0	(m)	$5.0 \cdot 10^{-4}$	$7.5 \cdot 10^{-4}$
First angle	α	(rad)	0.	0.175
Second angle	β	(rad)	0.	0.175

undergo significant variations. The material properties as well as the geometry are important for the heat transfer between the layers.

We consider a simplified geometry for a piece of glue, as shown in Fig. 1. The bottom layer is a fixed rectangle with the lengths a, b. The top layer is not parallel due to the variations. Five parameters specify the setting: the thermal conductivity λ and the volumetric heat capacity ρc as material parameters, the average height h_0 and the two angles α, β between the bottom and the top layer in x- and y-direction, respectively, as geometrical parameters. Table 1 illustrates the parameters.

The output data consists of measurements in L different patches between the bottom layer and the top layer. The thermal conductance G and the heat capacitance C for each patch read as

$$G_\ell = \frac{\lambda\, a_\ell\, b_\ell}{h_\ell} \qquad \text{and} \qquad C_\ell = \frac{\rho c\, h_\ell\, a_\ell\, b_\ell}{3} \tag{1}$$

for $\ell = 1, \ldots, L$. Therein, a_ℓ, b_ℓ denote the constant lengths of the rectangle for a single patch and h_ℓ is the average height in the patch. It holds that

$$h_\ell = h_0 + \tan\alpha\, x_\ell + \tan\beta\, y_\ell \doteq h_0 + \alpha\, x_\ell + \beta\, y_\ell \tag{2}$$

with known coordinates x_ℓ, y_ℓ for $\ell = 1, \ldots, L$. The approximation (2) is inserted into the model functions (1).

In our test problem, we predetermine independent uniform distributions with fixed intervals for each input parameter. The minimum and maximum values of the intervals are given in Table 1. The model functions (1) allow for the simulation of

measurements, where measurement errors can be added to test the techniques. Now our aim is to recover the original probability distributions from the measurements only.

3 Fitting of Cumulative Distribution Functions

Although uniform distributions are assumed for the input parameters in Sect. 2, we design a method for the approximation of an arbitrary continuous cumulative distribution function. Consequently, the strategy can also be applied for problems, where the type of the input probability distributions is unknown a priori.

We make the following assumptions. On the one hand, k output quantities are given by a model function $f : \Pi \to \mathbb{R}^k$, i.e.,

$$y = f(p), \qquad y_i = f_i(p_1, \ldots, p_n) \qquad \text{for } i = 1, \ldots, k \tag{3}$$

with the input parameters $p \in \Pi \subset \mathbb{R}^n$. The parameter domain is a cuboid

$$\Pi = \prod_{i=1}^{n} [p_{i,\min}, p_{i,\max}] \subset \mathbb{R}^n. \tag{4}$$

If the range of a parameter is infinite, then a truncation to a compact interval is required. Given a realization of the input parameters, the outputs can be evaluated using (3). On the other hand, the output quantities can be measured simultaneously, which yields the measurements $\{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(M)}\} \subset \mathbb{R}^k$ for different unknown realizations of the input parameters.

The random variables are assumed to be independent. We consider a single parameter $p \in [p_{\min}, p_{\max}]$. Let $F : [p_{\min}, p_{\max}] \to [0, 1]$ be its cumulative distribution function. We construct a piecewise linear approximation of the cumulative distribution function by

$$\tilde{F}(p) := \begin{cases} 0 & \text{for } p < q_1 \\ \frac{j-1}{m-1} \cdot \frac{q_{j+1}-p}{q_{j+1}-q_j} + \frac{j}{m-1} \cdot \frac{p-q_j}{q_{j+1}-q_j} & \text{for } p \in [q_j, q_{j+1}),\ j \in \{1, \ldots, m-1\} \\ 1 & \text{for } p \geq q_m \end{cases} \tag{5}$$

using discrete points

$$p_{\min} \leq q_1 < q_2 < \cdots < q_m \leq p_{\max} \tag{6}$$

for some integer m. This construction yields a strictly monotone function $\tilde{F}$. The grid points $q_1, \ldots, q_m$ represent the unknowns now. For given starting values, an arbitrary number N of samples for p can be generated by the approximation (5).

Therefore the inverse function $\tilde{F}^{-1}$ is evaluated at uniformly distributed (pseudo) random numbers in $[0, 1]$.

The above construction is used for each random variable separately. For simplicity, we arrange the same number m for all input parameters, which is not necessary in general. Let $\mathscr{K} := \Pi^m \subset \mathbb{R}^{m \times n}$ with the parameter domain (4). The unknown degrees of freedom are written as a matrix $Q \in \mathscr{K}$, whose ith column is associated to the ith parameter. Let $\mathscr{K}' \subset \mathscr{K}$ be the subset of all matrices satisfying the ascending order (6) in each column. The ordering can be ensured by a transformation of the problem to variables $\Delta q_j := q_{j+1} - q_j$, which are required to be positive.

On the one hand, the parameter-dependent output quantities (3) are measured simultaneously. The range $y_i \in [y_{i,\min}, y_{i,\max}]$ is partitioned into r_i equidistant intervals for $i = 1, \ldots, k$. Let M be the total number of measurements and s_{ij} for $j = 1, \ldots, r_i$ be the number of measurements in the jth interval for the ith output quantity.

On the other hand, we consider N realizations of the tuple of random parameters, which are computed by inverse functions $\tilde{F}_1^{-1}, \ldots, \tilde{F}_n^{-1}$ assuming some starting values in $\mathscr{K}'$. The realizations $\{p^{(1)}, \ldots, p^{(N)}\}$ yield associated output quantities $\{y^{(1)}, \ldots, y^{(N)}\}$ using the model function (3). Let t_{ij} for $j = 1, \ldots, r_i$ be the number of values in the jth interval for the ith output quantity.

Comparing the ratio from sampling and the ratio from measurements for each interval and each output quantity, a nonlinear least squares problem appears for the unknown grid points in the approximations (5). The problem reads as

$$\min_{Q \in \mathscr{K}'} \sum_{i=1}^{k} \sum_{j=1}^{r_i} \left(\frac{s_{ij}}{M} - \frac{t_{ij}(Q)}{N} \right)^2 \tag{7}$$

The dimensionality of the unknowns is nm. The number M of measurements is often restricted in practice. In contrast, the number N of realizations can be chosen arbitrarily large limited by the computational effort for the evaluations (3) only.

Unfortunately, typical numerical methods, where the values $\Delta q_j > 0$ are considered as unknowns, yield poor approximations of this minimization problem. This behavior occurs both for derivative-based algorithms like the Levenberg-Marquardt method, see [1], and derivative-free algorithms like the Nelder-Mead technique, cf. [6]. Hence we construct a simple alternative, which can be applied only if the evaluations of the model function (3) are relatively cheap. The idea is to evaluate the objective function in (7) on a finite grid in $\mathscr{K}'$ and to determine the minimum value on the grid. An advantage of this approach is that starting values for the unknown solution are not required.

Although nm unknowns appear, the dimensions are not independent of each other. The problem consists of n groups of unknowns. For the ith group, we generate a trial set $\{q_1, \ldots, q_m\}$ by uniformly distributed (pseudo) random numbers in $[p_{i,\min}, p_{i,\max}]$. Afterwards the values are sorted in ascending order. An arbitrary number N_{grid} of grid points in $\mathscr{K}'$ can be produced by this approach.

4 Numerical Results

We apply the technique designed in Sect. 3 to the test problem from Sect. 2 now. The required computations were done in the software package MATLAB [4]. In particular, we utilized the built-in routine `rand` for the generation of all pseudo random numbers.

In the model problem, we arrange $L = 8$ patches. Hence the number of output quantities (1) becomes $k = 2L = 16$. We produce artificial measurements of the outputs by samples from the uniform distributions of the input parameters using pseudo random numbers. Moreover, measurement errors up to 0.1% are included by adding pseudo random numbers of this magnitude. We employ $M = 100$ measurements in total. The ranges $G \in [0, 0.015]$ and $C \in [0, 0.001]$ are applied for each patch. Furthermore, we select $r_i = 10$ for all $i = 1, \ldots, n$.

The minimum and maximum values of the uniform distributions, see Table 1, are considered as unknowns. Thus slightly larger intervals $[p_{i,\min}, p_{i,\max}]$ are arranged in the numerical method. The number $m = 20$ is used within the discretization of each cumulative distribution function (5) for all parameters. We produce $N = 1000$ samples of the output quantities for each evaluation of the objective function.

The simple variant of our numerical method for the nonlinear least squares problem (7) is applied using $N_{\text{grid}} = 10{,}000$ grid points in $\mathscr{K}'$. Figure 2 illustrates the true cumulative distribution functions of the input parameters and the resulting approximations. We observe a moderate approximation of the exact distributions. However, our approach does not require a priori assumptions on the type of the distribution, except for the minimum and maximum values. Other numerical techniques may result in worse approximations.

5 Outlook

At a later stage, the package-die-glue problem shall be simulated in a more realistic and relevant form. Firstly, the artificially computed measurements will be replaced by measured data. Secondly, the simple model functions (1) will be substituted by parabolic partial differential equations describing the heat transfer.

Acknowledgements This work is part of the project nanoCOPS (nanoelectronic COupled Problems Solutions) supported by the European Union in the FP7-ICT-2013-11 program (grant agreement no. 619166). The authors are indebted to Dr. Jan ter Maten (Bergische Universität Wuppertal, Germany), Prof. Volkmar Liebscher (Ernst-Moritz-Arndt-Universität Greifswald, Germany) and Dr. Tanja Clees (Fraunhofer-Institut SCAI, Sankt-Augustin, Germany) for helpful discussions.

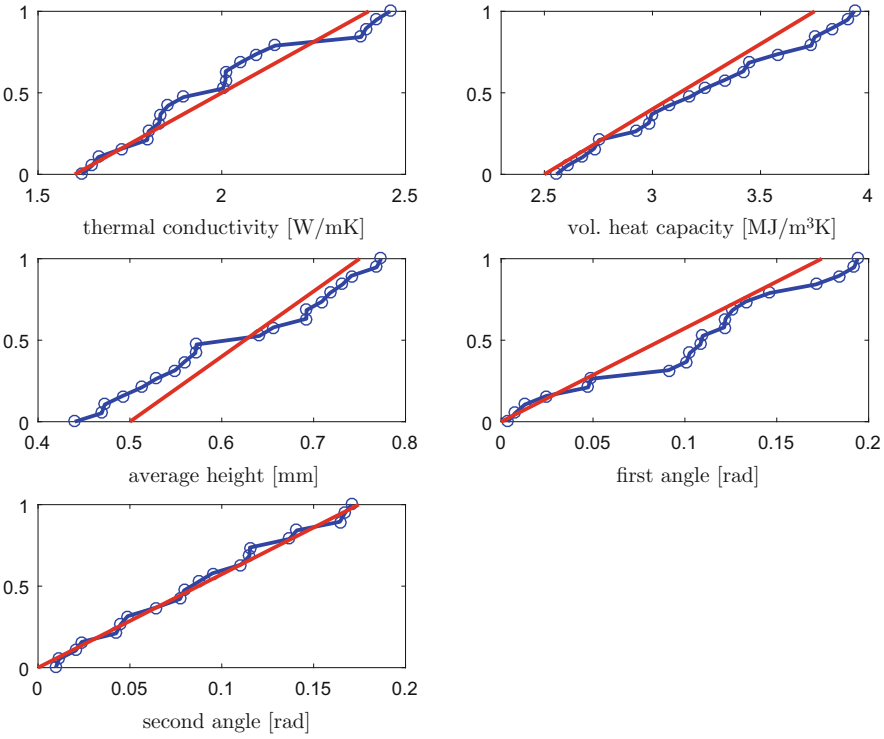


Fig. 2 Exact cumulative distribution functions (*red lines*) and piecewise linear approximations (*blue lines*) for the five input parameters

References

1. Gavin, H.P.: The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems (2016). http://people.duke.edu/~hpgavin/ce281/lm.pdf. Cited 20 Dec 2016
2. Li, P., Liu, F., Li, X., Pileggi, L.T., Nassif, S.R.: Modeling interconnect variability using efficient parametric model order reduction. In: Design, Automation and Test Conference in Europe, pp. 958–963 (2005)
3. Luo, H.: Generation of Non-normal Data – A Study of Fleishman's Power Method. Research Report, Department of Statistics, Uppsala University, Sweden (2011)
4. MATLAB, version 8.6.0 (R2015b), The Mathworks Inc., Natick, Massachusetts (2015)
5. Montgomery, D.C., Runger, G.C.: Applied Statistics and Probability for Engineers, 3rd edn. Wiley, Hoboken (2003)
6. Nelder, J., Mead, R.: A simplex method for function minimization. Comput. J. **7**, 308–313 (1965)
7. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
8. Pulch, R., Putek, P., ter Maten, E.J.W., Duque, D., Schöps, S., Römer, U., Casper, T., Yue, Y., Feng, L., Benner, P., Schoenmaker, W.: Public Report on Uncertainty Quantification. nanoCOPS project (2016). http://www.fp7-nanocops.eu. Cited 20 Dec 2016

9. Wang, Z., Lai, X., Roychowdhury, J.: PV-PPV: parameter variability aware, automatically extracted, nonlinear time-shifted oscillators macromodels. In: Proceedings of the IEEE Design Automation Conference, pp. 142–147 (2007)
10. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)
11. Zhang, H., Chen, T.H., Ting, M.Y., Li, X.: Efficient design-specific worst-case corner extraction for integrated circuits. In: DAC '09 Proceedings of the 46th Annual Design Automation Conference, pp. 386–389 (2009)

Parametric Model Order Reduction for Electro-Thermal Coupled Problems with Many Inputs

Nicodemus Banagaaya, Peter Benner, and Lihong Feng

Abstract The modified BDSM-ET method is a model order reduction (MOR) technique which was developed to reduce non-parametric electro-thermal (ET) coupled problems with many inputs. The method leads to block-wise sparse reduced-order models (ROMs) which are accurate and computationally cheaper compared to the existing MOR methods. In this work, we extend the modified BDSM-ET method to parametrized ET coupled problems with many inputs.

1 Introduction

Spatial discretization of electro-thermal (ET) coupled problems considering parameter variations leads to a parametrized nonlinear quadratic dynamical system of the following form:

$$\mathbf{E}(\mu)\mathbf{x}'(t) = \mathbf{A}(\mu)\mathbf{x}(t) + \mathbf{x}(t)^T\mathcal{F}(\mu)\mathbf{x}(t) + \mathbf{B}(\mu)\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \tag{1a}$$

$$\mathbf{y}(t) = \mathbf{C}(\mu)\mathbf{x}(t) + \mathbf{D}(\mu)\mathbf{u}(t), \tag{1b}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector and $\mathbf{E}(\mu) \in \mathbb{R}^{n\times n}$ is uniformly singular, indicating that (1) is a system of differential-algebraic equations (DAEs) for every parameter μ. Here, $\mathbf{A}(\mu) \in \mathbb{R}^{n\times n}, \mathbf{B}(\mu) \in \mathbb{R}^{n\times m}, \mathbf{C}(\mu) \in \mathbb{R}^{\ell\times n}, \mathbf{D}(\mu) \in \mathbb{R}^{\ell\times m}$, and $\mathcal{F}(\mu) = \left[\mathbf{F}_1(\mu)^T, \ldots, \mathbf{F}_n(\mu)^T\right]^T \in \mathbb{R}^{n\times n\times n}$ is a 3-way tensor of n parametrized matrices $\mathbf{F}_i(\mu) \in \mathbb{R}^{n\times n}$. Each element in $\mathbf{x}(t)^T\mathcal{F}(\mu)\mathbf{x}(t) \in \mathbb{R}^n$ is a scalar $\mathbf{x}(t)^T\mathbf{F}_i(\mu)\mathbf{x}(t) \in \mathbb{R}, i = 1, \ldots, n$. The vector $\mu \in \mathbb{R}^d$ represents the parameter variations which may arise from material properties, system configurations, etc. In practice, these systems have very large dimensions compared to the number of inputs and outputs. Despite the ever increasing computational power, simulation of such systems in acceptable time is very difficult, in particular if multi-query tasks are

N. Banagaaya (✉) • P. Benner • L. Feng
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany
e-mail: banagaaya@mpi-magdeburg.mpg.de; benner@mpi-magdeburg.mpg.de; feng@mpi-magdeburg.mpg.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_41

required. This calls for application of parametric model order reduction (pMOR). pMOR replaces (1) by a parametric reduced-order model (pROM):

$$\mathbf{E}_r(\mu)\mathbf{x}_r'(t) = \mathbf{A}_r(\mu)\mathbf{x}_r(t) + \mathbf{x}_r(t)^T\mathcal{F}_r(\mu)\mathbf{x}_r(t) + \mathbf{B}_r(\mu)\mathbf{u}(t), \quad \mathbf{x}_r(0) = \mathbf{x}_{r_0}, \tag{2a}$$

$$\mathbf{y}_r(t) = \mathbf{C}_r(\mu)\mathbf{x}_r(t) + \mathbf{D}(\mu)\mathbf{u}(t), \tag{2b}$$

where $\mathbf{E}_r(\mu) = \mathbf{V}^T\mathbf{E}(\mu)\mathbf{V}$, $\mathbf{A}_r(\mu) = \mathbf{V}^T\mathbf{A}(\mu)\mathbf{V}$, $\mathbf{B}_r(\mu) = \mathbf{V}^T\mathbf{B}(\mu)$, $\mathbf{C}_r(\mu) = \mathbf{C}(\mu)\mathbf{V}$. Using Proposition 1, the reduced nonlinear term is obtained by reformulating

$$\mathbf{V}^T\big(\mathbf{x}(t)^T\big(\mathbf{V}^T\mathcal{F}(\mu)\,\mathbf{V}\big)\mathbf{x}(t)\big) \quad \text{as} \quad \mathbf{x}_r(t)^T\mathcal{F}_r(\mu)\mathbf{x}_r(t),$$

where $\mathcal{F}_r(\mu) = \big[\mathbf{F}_{r_1}(\mu)^T, \ldots, \mathbf{F}_{r_r}(\mu)^T\big]^T \in \mathbb{R}^{r\times r\times r}$ is a 3-way tensor with parametrized matrices $\mathbf{F}_{r_i}(\mu) \in \mathbb{R}^{r\times r}$, $\mathbf{V}^T\mathcal{F}(\mu)\mathbf{V} := \big[\mathbf{V}^T\mathbf{F}_1(\mu)^T\mathbf{V}, \ldots, \mathbf{V}^T\mathbf{F}_n(\mu)^T\mathbf{V}\big]^T$.

Proposition 1 *Let $\boldsymbol{W} = \big(\boldsymbol{w}_{ij}\big) \in \mathbb{R}^{n\times r}$ be a parameter-independent matrix, $\mathbf{x}_r \in \mathbb{R}^r$, and $\tilde{\mathcal{F}}(\mu) = \big[\tilde{\mathbf{F}}_1(\mu)^T, \ldots, \tilde{\mathbf{F}}_n(\mu)^T\big]^T \in \mathbb{R}^{r\times r\times n}$ be a 3-way parameterized tensor, then there exist a 3-way parameterized tensor $\mathcal{F}_r(\mu) \in \mathbb{R}^{r\times r\times r}$, such that:*

$$\boldsymbol{W}^T\left(\mathbf{x}_r^T\tilde{\mathcal{F}}(\mu)\mathbf{x}_r\right) = \mathbf{x}_r^T\mathcal{F}_r(\mu)\mathbf{x}_r, \quad \textit{where} \quad \mathcal{F}_r(\mu) = \big[\hat{\mathbf{F}}_1(\mu)^T, \ldots, \hat{\mathbf{F}}_r(\mu)^T\big]^T$$

with $\hat{\mathbf{F}}_j(\mu) = \sum_{i=1}^{n} \mathbf{w}_{ij}\tilde{\mathbf{F}}_i(\mu) \in \mathbb{R}^{r\times r}$, $j = 1, \ldots, r$.

The proof is analogous to the proof of Proposition 1 in [3] for the non-parametric case. The projection matrix $\mathbf{V} \in \mathbb{R}^{n\times r}$ which is valid for all parameters μ in the desired range, and for arbitrary inputs $\mathbf{u}(t)$, can be constructed using, e.g., the implicit moment-matching pMOR method from [4]. However, direct application of this approach to system (1) may produce pROMs which are inaccurate or difficult to simulate. Moreover, if m is large, standard pMOR methods lead to very large and dense ROMs. In Sect. 2, we review the pMOR method for ET coupled problems with many inputs. Section 3 extends the modified BDSM-ET method from [2] to parametrized ET coupled problems with many inputs. Finally, we present numerical experiments and conclusions. For simplicity, we remove (t) for time dependent variables in the next sections.

2 pMOR for ET Coupled Problems

In this section, we review the pMOR method for ET coupled problems proposed in [3]. We consider a parametric structure which is quite common in the nanoelectronics coupled problems, i.e., the ET coupled problems take the form of (1) with

system matrices and tensor structures as follows:

$$\mathbf{E}(\mu) = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{E}_T(\mu) \end{pmatrix}, \quad \mathbf{A}(\mu) = \begin{pmatrix} \mathbf{A}_v & 0 \\ 0 & \mathbf{A}_T(\mu) \end{pmatrix}, \quad \mathbf{B}(\mu) = \begin{pmatrix} \mathbf{B}_v(\mu) & 0 \\ 0 & \mathbf{B}_T(\mu) \end{pmatrix},$$
$$\mathbf{C}(\mu) = \begin{pmatrix} \mathbf{C}_v(\mu) & \mathbf{C}_T(\mu) \end{pmatrix}, \quad \mathbf{D}(\mu) = \begin{pmatrix} \mathbf{D}_v(\mu) & \mathbf{D}_T(\mu) \end{pmatrix}, \quad \mathbf{u} = \left[\mathbf{u}_v^T, \mathbf{u}_T^T\right]^T,$$
$$\mathcal{F}(\mu) = \left(0, \ldots, 0, \mathbf{F}_{n_v+1}(\mu)^T, \ldots, \mathbf{F}_n(\mu)^T\right)^T, \tag{3}$$

where $\mathbf{F}_i(\mu) = \begin{pmatrix} \mathbf{F}_{v_i}(\mu) & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}$, $\mathbf{F}_{v_i}(\mu) \in \mathbb{R}^{n_v\times n_v}$, $i = n_v+1, \ldots, n$, $\mathbf{A}_v(\mu) \in \mathbb{R}^{n_v\times n_v}$, $\mathbf{B}_v(\mu) \in \mathbb{R}^{n_v\times \tilde{m}}$, $\mathbf{E}_T(\mu) \in \mathbb{R}^{n_T\times n_T}$, $\mathbf{A}_T(\mu) \in \mathbb{R}^{n_T\times n_T}$, $\mathbf{B}_T(\mu) \in \mathbb{R}^{n_T\times \tilde{m}}$, $\mathbf{C}_v(\mu) \in \mathbb{R}^{\ell\times n_v}$, $\mathbf{C}_T(\mu) \in \mathbb{R}^{\ell\times n_T}$, $\mathbf{D}_v(\mu) \in \mathbb{R}^{\ell\times \tilde{m}}$, $\mathbf{D}_T(\mu) \in \mathbb{R}^{\ell\times \tilde{m}}$, and $\mathbf{u}_v, \mathbf{u}_T \in \mathbb{R}^{\tilde{m}}$, $\tilde{m} = m/2$. Thus, substituting the above matrices and the tensor $\mathcal{F}(\mu)$ into (1) leads to an equivalent decoupled system given by

$$\mathbf{A}_v(\mu)\mathbf{x}_v = -\mathbf{B}_v(\mu)\mathbf{u}_v, \tag{4a}$$
$$\mathbf{E}_T(\mu)\mathbf{x}_T' = \mathbf{A}_T(\mu)\mathbf{x}_T + \mathbf{x}_v^T\mathcal{F}_T(\mu)\mathbf{x}_v + \mathbf{B}_T(\mu)\mathbf{u}_T, \quad \mathbf{x}_T(0) = \mathbf{x}_{T_0}, \tag{4b}$$
$$\mathbf{y} = \mathbf{C}_v(\mu)\mathbf{x}_v + \mathbf{C}_T(\mu)\mathbf{x}_T + \mathbf{D}_v(\mu)\mathbf{u}_v + \mathbf{D}_T(\mu)\mathbf{u}_T, \tag{4c}$$

where $\mathcal{F}_T(\mu) = \left(\mathbf{F}_{T_1}(\mu)^T, \ldots, \mathbf{F}_{T_{n_T}}(\mu)^T\right)^T \in \mathbb{R}^{n_v\times n_v\times n_T}$, $\mathbf{F}_{T_j}(\mu) = \mathbf{F}_{v_{n_v+j}}(\mu)$, $j = 1, \ldots, n_T$ and $\mathbf{F}_{v_i}(\mu)$ is as defined earlier. Equations (4a) and (4b) are the electrical and thermal subsystems, respectively. We can observe that the nonlinear term $\mathbf{x}_v^T\mathcal{F}_T(\mu)\mathbf{x}_v$ can be treated as part of the thermal input, since it is known after first simulating the electrical subsystem. The output can be obtained through (4c). pMOR replaces the decoupled system (4) with a reduced-order decoupled system

$$\mathbf{A}_{v_r}(\mu)\mathbf{x}_{v_r} = -\mathbf{B}_{v_r}(\mu)\mathbf{u}_v, \tag{5a}$$
$$\mathbf{E}_{T_r}(\mu)\mathbf{x}_{T_r}' = \mathbf{A}_{T_r}(\mu)\mathbf{x}_{T_r} + \mathbf{x}_{v_r}^T\mathcal{F}_{T_r}(\mu)\mathbf{x}_{v_r} + \mathbf{B}_{T_r}(\mu)\mathbf{u}_T, \quad \mathbf{x}_{T_r}(0) = \mathbf{x}_{T_{r_0}}, \tag{5b}$$
$$\mathbf{y}_r = \mathbf{C}_{v_r}(\mu)\mathbf{x}_{v_r} + \mathbf{C}_{T_r}(\mu)\mathbf{x}_{T_r} + \mathbf{D}_v(\mu)\mathbf{u}_v + \mathbf{D}_T(\mu)\mathbf{u}_T, \tag{5c}$$

where $\mathbf{A}_{v_r}(\mu) \in \mathbb{R}^{r_v\times r_v}$, $\mathbf{B}_{v_r}(\mu) \in \mathbb{R}^{r_v\times \tilde{m}}$, $\mathbf{E}_{T_r}(\mu) \in \mathbb{R}^{r_T\times r_T}$, $\mathbf{A}_{T_r}(\mu) \in \mathbb{R}^{r_T\times r_T}$, $\mathbf{B}_{T_r}(\mu) \in \mathbb{R}^{r_T\times \tilde{m}}$, $\mathbf{C}_{v_r}(\mu) \in \mathbb{R}^{\ell\times r_v}$, $\mathbf{C}_{T_r}(\mu) \in \mathbb{R}^{\ell\times r_T}$, $\mathcal{F}_{T_r}(\mu) \in \mathbb{R}^{r_v\times r_v\times r_T}$, such that the reduced order is $r = r_v + r_T \ll n$. This pMOR method leads to simple and accurate ROMs for ET coupled problems. However, if $\tilde{m}$ is large, it may lead to ROMs which are very large and dense. The same problem was already observed in non-parametric systems, and could be dealt with using the superposition principle, see [2].

3 pMOR for ET Coupled Problems with Many Inputs

In this section, we extend the modified BDSM-ET method in [2] to parametrized ET coupled problems. We propose to apply the superposition principle to the parametrized subsystems (4a) and (4b) separately. Then, the standard pMOR method from [4] can be applied to each subsystem. Assume that $\mathbf{B}_v(\mu), \mathbf{B}_T(\mu)$ have no zero columns, so that they can be split into $\mathbf{B}_k(\mu) = \sum_{i=1}^{\tilde{m}} \mathbf{B}_{k_i}(\mu)$, $k = v, T$, where $\mathbf{B}_{k_i}(\mu)$ are column rank-1 parametric matrices defined as

$$\mathbf{B}_{k_i}(:, j) = \begin{cases} \mathbf{b}_{k_i}(\mu) \in \mathbb{R}^{n_k}, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases} \quad i = 1, \ldots, \tilde{m}.$$

By the superposition principle and using the above splitting of $\mathbf{B}_v(\mu)$, the electrical subsystem (4a) can be decomposed into $\tilde{m}$ subsystems

$$\mathbf{A}_v(\mu)\mathbf{x}_{v_i} = -\mathbf{B}_{v_i}(\mu)\mathbf{u}_v, \quad \mathbf{y}_v = \mathbf{C}_v(\mu)\mathbf{x}_{v_i}, \; i = 1, \ldots, \tilde{m}. \tag{6}$$

Here and below, blkdiag($\mathbf{M}$) denotes the block-diagonal matrix with the matrix $\mathbf{M}$ on its diagonal. Then the $\tilde{m}$ parametrized subsystems in (6) can be equivalently transformed into a block-diagonal system of dimension $\tilde{n}_v$ given by

$$\mathscr{A}_v(\mu)\xi_v = -\mathscr{B}_v(\mu)\mathbf{u}_v, \quad \mathbf{y}_v = \mathscr{C}_v(\mu)\xi_v + \mathbf{D}(\mu)\mathbf{u}, \tag{7}$$

where $\xi_v = (\mathbf{x}_{v_1}^T, \ldots, \mathbf{x}_{v_{\tilde{m}}}^T)^T \in \mathbb{R}^{\tilde{n}_v}$, $\tilde{n}_v = \tilde{m}n_v$, $\mathscr{B}_v(\mu) = (\mathbf{B}_{v_1}(\mu)^T, \ldots, \mathbf{B}_{v_{\tilde{m}}}(\mu)^T)^T \in \mathbb{R}^{\tilde{n}_v \times \tilde{m}}$, $\mathscr{C}_v(\mu) = (\mathbf{C}_v(\mu), \ldots, \mathbf{C}_v(\mu)) \in \mathbb{R}^{\ell \times \tilde{n}_v}$, $\mathscr{A}_v(\mu) = \text{blkdiag}(\mathbf{A}_v(\mu), \ldots, \mathbf{A}_v(\mu)) \in \mathbb{R}^{\tilde{n}_v \times \tilde{n}_v}$. The next step is to reduce the dimension of (7). This is done by replacing (7) with a much smaller ROM

$$\mathscr{A}_{v_r}(\mu)\xi_{v_r} = -\mathscr{B}_{v_r}(\mu)\mathbf{u}_v, \quad \mathbf{y}_{v_r} = \mathscr{C}_{v_r}(\mu)\xi_{v_r} + \mathbf{D}_v(\mu)\mathbf{u}, \tag{8}$$

where $\mathscr{A}_{v_r}(\mu) = \mathbf{V}_{\mathbf{v}}^T \mathscr{A}_v(\mu)\mathbf{V}_v$, $\mathscr{B}_{v_r}(\mu) = \mathbf{V}_v^T \mathscr{B}_v(\mu)$, $\mathscr{C}_{v_r}(\mu) = \mathscr{C}_v(\mu)\,\mathbf{V}_v$, $\xi_v = \mathbf{V}_v \xi_{v_r} \in \mathbb{R}^{r_v}$ and the projection matrix $\mathbf{V}_v = \text{blkdiag}(\mathbf{V}_v^{(1)}, \mathbf{V}_v^{(2)}, \ldots, \mathbf{V}_v^{(m)}) \in \mathbb{R}^{\tilde{m}n \times r_v}$, $r_v \ll n_v$. The projection matrices $\mathbf{V}_v^{(i)} \in \mathbb{R}^{n_V \times r_{v_i}}$, $r_v = \sum r_{v_i}$ can be constructed by applying the existing pMOR methods to each system in (6).

Finally, we reduce the thermal subsystem (4b). Applying the superposition principle to the electrical subsystem (4a) introduces $\left(\sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}^T\right) \mathcal{F}_T(\mu) \left(\sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}\right)$ into the thermal subsystem (4b), i.e. $\mathbf{x}_v$ is replaced by $\sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}$ in the nonlinear part. In order to obtain a sparse tensor, the approximation $\left(\sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}^T\right) \mathcal{F}_T(\mu) \left(\sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}\right) \approx \sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}^T \mathcal{F}_T(\mu)\mathbf{x}_{v_i}$ is introduced into the thermal subsystem. Simulation results in Sect. 4 show that the error caused by this approximation is acceptable. Since $\sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}^T \mathcal{F}_T(\mu)\mathbf{x}_{v_i}$ can be considered as an extra input for the thermal subsystem,

the superposition principle still applies to the thermal subsystem. Thus (4b) can be approximated as

$$\mathbf{E}_T(\mu)\mathbf{x}_T' = \mathbf{A}_T(\mu)\mathbf{x}_T + \xi_v^T \mathcal{W}_T(\mu)\xi_v, + \mathbf{B}_T(\mu)\mathbf{u}_T, \quad \mathbf{x}_T(0) = \mathbf{x}_{T_0}, \tag{9a}$$

$$\mathbf{y}_T = \mathbf{C}_T(\mu)\mathbf{x}_T + \mathbf{D}_T(\mu)\mathbf{u}_T. \tag{9b}$$

Here, we have used the equality $\sum_{i=1}^{\tilde{m}} \mathbf{x}_{v_i}^T \mathcal{F}_T(\mu)\mathbf{x}_{v_i} = \xi_v^T \mathcal{W}_T(\mu)\xi_v$, where $\mathcal{W}_T(\mu) = \left[\mathbf{W}_{T_1}(\mu)^T, \ldots, \mathbf{W}_{T_{n_T}}(\mu)^T\right]^T \in \mathbb{R}^{\tilde{n}_v \times \tilde{n}_v \times n_T}$, $\mathbf{W}_{T_i}(\mu)$=blkdiag$(\mathbf{F}_{T_i}(\mu), \ldots, \mathbf{F}_{T_i}(\mu)) \in \mathbb{R}^{\tilde{n}_v \times \tilde{n}_v}$, $\mathbf{F}_{T_i}(\mu) \in \mathbb{R}^{n_v \times n_v}$, and ξ_v is defined as in (7). We can see that the reduced state in (8) introduces an approximation into (9) leading to

$$\mathbf{E}_T(\mu)\mathbf{x}_T' = \mathbf{A}_T(\mu)\mathbf{x}_T + \xi_{v_r}^T \hat{\mathcal{W}}_T(\mu)\xi_{v_r} + \mathbf{B}_T(\mu)\mathbf{u}_T, \quad \mathbf{x}_T(0) = \mathbf{x}_{T_0}, \tag{10a}$$

$$\mathbf{y}_T = \mathbf{C}_T(\mu)\mathbf{x}_T + \mathbf{D}_T(\mu)\mathbf{u}_T, \tag{10b}$$

where $\hat{\mathcal{W}}_T(\mu) = \mathbf{V}_\mathbf{v}^T \mathcal{W}_T(\mu)\mathbf{V}_v := \sum_{i=1}^{\tilde{m}} \mathbf{V}_v^{(i)^T} \mathcal{F}_T(\mu)\mathbf{V}_v^{(i)} \in \mathbb{R}^{r_v \times r_v \times n_T}$. Here, $\mathbf{V}_v^{(i)^T} \mathcal{F}_T(\mu)\mathbf{V}_v^{(i)} := \left[\mathbf{V}_v^{(i)^T} \mathbf{F}_{T_1}(\mu)^T \mathbf{V}_v^{(i)}, \ldots, \mathbf{V}_v^{(i)^T} \mathbf{F}_{T_{n_T}}(\mu)^T \mathbf{V}_v^{(i)}\right]^T$.

By superposition principle and using the splitting of $\mathbf{B}_T(\mu)$, the thermal subsystem system (10) can also be split into $\tilde{m}$ subsystems

$$\mathbf{E}_T(\mu)\mathbf{x}_{T_1}' = \mathbf{A}_T(\mu)\mathbf{x}_{T_1} + \xi_{v_r}^T \hat{\mathcal{W}}_T(\mu)\xi_{v_r} + \mathbf{B}_{T_1}(\mu)\mathbf{u}_T, \quad \mathbf{x}_{T_1}(0) = \mathbf{x}_{T_0}, \tag{11a}$$

$$\mathbf{E}_T(\mu)\mathbf{x}_{T_i}' = \mathbf{A}_T(\mu)\mathbf{x}_{T_i} + \mathbf{B}_{T_i}(\mu)\mathbf{u}_T, \quad \mathbf{x}_{T_i}(0) = 0,\ i = 2, \ldots, \tilde{m}, \tag{11b}$$

$$\mathbf{y}_T = \big(\mathbf{C}(\mu), \ldots, \mathbf{C}(\mu)\big)\xi_T + \mathbf{D}_T(\mu)\mathbf{u}_T, \tag{11c}$$

where $\xi_T = (\mathbf{x}_{T_1}^T, \ldots, \mathbf{x}_{T_{\tilde{m}}}^T)^T \in \mathbb{R}^{\tilde{n}_T}$, $\tilde{n}_T = \tilde{m}n_T$. System (11) can be reformulated as a block-diagonal structured system of dimension $\tilde{n}_T$ given by

$$\begin{aligned} \mathscr{E}_T(\mu)\xi_T' &= \mathscr{A}_T(\mu)\xi_T + \xi_{v_r}^T \tilde{\mathcal{W}}_T(\mu)\xi_{v_r} + \mathscr{B}_T(\mu)\mathbf{u}_T, \\ \mathbf{y}_T &= \mathscr{C}_T(\mu)\xi_T + \mathbf{D}_T(\mu)\mathbf{u}_T, \quad \xi_T(0) = \big(\mathbf{x}_{T_0}^T, 0, \ldots, 0\big)^T, \end{aligned} \tag{12}$$

where $\tilde{\mathcal{W}}_T(\mu) = \begin{pmatrix} \hat{\mathcal{W}}_T(\mu) \\ 0 \end{pmatrix} \in \mathbb{R}^{r_v \times r_v \times \tilde{n}_T}$, $\mathscr{E}_T(\mu) = \text{blkdiag}(\mathbf{E}_T(\mu), \ldots, \mathbf{E}_T(\mu))$, $\mathscr{A}_T(\mu) = \text{blkdiag}(\mathbf{A}_T(\mu), \ldots, \mathbf{A}_T(\mu)) \in \mathbb{R}^{\tilde{n}_T \times \tilde{n}_T}$, $\mathscr{C}_T(\mu) = (\mathbf{C}_T(\mu), \ldots, \mathbf{C}_T(\mu)) \in \mathbb{R}^{\ell \times \tilde{n}_T}$, $\mathscr{B}_T(\mu) = (\mathbf{B}_{T_1}{}^T(\mu), \ldots, \mathbf{B}_{T_{\tilde{m}}}{}^T(\mu))^T \in \mathbb{R}^{\tilde{n}_T \times \tilde{m}}$. The corresponding ROM of (12) is given by

$$\begin{aligned} \mathscr{E}_{T_r}(\mu)\xi_{T_r}' &= \mathscr{A}_{T_r}(\mu)\xi_T + \xi_{v_r}^T \mathcal{W}_{T_r}(\mu)\xi_{v_r} + \mathscr{B}_{T_r}(\mu)\mathbf{u}_T, \\ \mathbf{y}_{T_r} &= \mathscr{C}_{T_r}(\mu)\xi_{T_r} + \mathbf{D}_T(\mu)\mathbf{u}_T, \quad \xi_{T_r}(0) = \big(\mathbf{x}_{T_{r_0}}^T, 0, \ldots, 0\big)^T, \end{aligned} \tag{13}$$

where $\mathscr{E}_{T_r}(\mu) = \mathbf{V}_T^T \mathscr{E}_T(\mu)\mathbf{V}_T \in \mathbb{R}^{r_T \times r_T}$, $\mathscr{A}_{T_r}(\mu) = \mathbf{V}_T^T \mathscr{A}_T(\mu)\mathbf{V} \in \mathbb{R}^{r_T \times r_T}$, $\mathscr{B}_{T_r}(\mu) = \mathbf{V}_T^T \mathscr{B}_T(\mu) \in \mathbb{R}^{r_T \times \tilde{m}}$, $\mathscr{C}_{T_r}(\mu) = \mathscr{C}_T(\mu)\mathbf{V}_T \in \mathbb{R}^{\ell \times r_T}$, $\mathbf{V}_T = \text{blkdiag}(\mathbf{V}_T^{(1)}, \ldots, \mathbf{V}_T^{(\tilde{m})}) \in \mathbb{R}^{\tilde{n}_T \times r_T}$. Using Proposition 1, the reduced nonlinear term $\xi_{v_r}^T \mathcal{W}_{T_r}(\mu)\xi_{v_r}$, where $\mathcal{W}_{T_r}(\mu) = \left[\hat{\mathbf{W}}_{T_1}(\mu)^T, \ldots, \hat{\mathbf{W}}_{T_{r_T}}(\mu)^T\right]^T \in \mathbb{R}^{r_v \times r_v \times r_T}$ with block-diagonal matrices $\hat{\mathbf{W}}_{T_i}(\mu) \in \mathbb{R}^{r_v \times r_v}$, $i = 1, \ldots, r_T$, can be reformulated from $\mathbf{V}_T^T \left(\xi_{v_r}^T \tilde{\mathcal{W}}_T \xi_{v_r}\right)$. Note that by construction, $\mathcal{W}_{T_r}(\mu)$ is a sparse tensor since it has block-wise sparse $\hat{\mathbf{W}}_{T_i}(\mu)$ slices. The projection matrices $\mathbf{V}_T^{(1)}$ and $\mathbf{V}_T^{(i)}$, $i = 2, \ldots, \tilde{m}$ can be constructed by applying existing pMOR methods on the subsystem (11a) and each in (11b), respectively. In this work, we use the implicit moment-matching pMOR method proposed from [4], this implies that $\mathbf{V}_T^{(1)}$ can be constructed by ignoring the nonlinear part of (11a). This approach allows us to automatically construct $\mathbf{V}_v$ and $\mathbf{V}_T$, respectively, using the global a posteriori error bound [3], and it is used to produce the simulation results in the next section.

The main difference from the modified BDSM-ET method in [2] for non-parametric systems resides in treating the electrical subsystem after superposition. For non-parametric systems, the electrical subsystems are non-parametric algebraic systems, which can only be reduced using special techniques, such as the Gaussian elimination method [2]. For parametric systems, the electrical subsystems are parametric, which can be directly treated using the existing pMOR methods, so that automatic pMOR is possible.

4 Numerical Experiments

All simulations were done using MATLAB®Version 2012b on a Laptop with 6 GB RAM, CPU@ 2.00 GHz. The sparse tensor computations were performed with the MATLAB Tensor Toolbox [1]. All used pMOR methods are defined as follows: IpMOR-ET denotes the standard pMOR method for ET coupled problems proposed in [3]. Modified IpBDSM-ET denotes the method proposed in Sect. 3. We consider electro-thermal simulation of a power-MOS device model. It is a parameterized quadratic system described in (1) with matrices and tensor in the form of (3) and $n =$ 13,216 state variables. There are $m = 6$ inputs and $\ell = 12$ outputs. The matrices $\mathbf{E}(\mu), \mathbf{A}(\mu), \mathbf{B}(\mu), \mathbf{C}(\mu), \mathbf{D}(\mu)$ and the tensor $\mathbf{F}(\mu)$ exhibit a parameter dependence of an affine form given by $\mathbf{M}(\mu) = \mathbf{M}_0 + \mu\mathbf{M}_1$, where $\mathbf{M}(\mu)$ indicates any parametrized matrix or tensor in (1) and $\mathbf{M}_i$ $(i = 0, 1)$ are parameter-independent coefficients. For this example, $\mu = \sigma$ describes the electrical conductivity of the power-MOS device. This system is decoupled into the form (4) with $n_v = 1660$ electrical equations, and $n_T = 11{,}556$ thermal equations. Both pMOR methods were able to reduce the electrical and thermal equations to 4 and 52 equations, respectively. Hence, the power-MOS device model is reduced from 13,216 to 56 leading to a reduction rate of 99.6 as shown in Table 1.

Table 1 Simulation efficiency comparison

pMOR methods	Original model			Reduced order	Speed-up	Reduction rate (%)	Error
	n	m	ℓ	r			
IpMOR-ET	13,216	6	12	56	1365.1	99.6	1.4×10^{-6}
Modified IpBDSM-ET	13,216	6	12	56	1376.2	99.6	3.3×10^{-5}

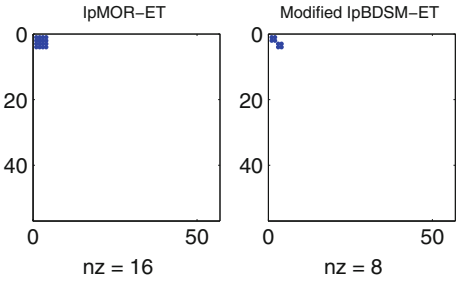


Fig. 1 Comparison of the sparsity of the first nonzero slice of the reduced tensors

Next, we simulate the ROMs in Table 1 within the time interval $t \in [0, 2 \times 10^{-6}]$ and the electrical conductivity $\sigma \in [10^7, 10^8]$. We discretized the time and the electrical conductivity into 100 and 50 grid points, respectively. In Table 1, we compare the simulation efficiency of the pMOR methods. However, the advantage of the IpBDSM-ET method should be significant if there are many inputs and outputs ($m, l \approx 100$), which has been shown in [2] for a non-parametric system with 408 inputs and 816 outputs. Furthermore, the modified IpBDSM-ET leads to sparser ROMs than the IpMOR method as illustrated in Fig. 1.

5 Conclusions

We have proposed an IpBDSM-ET method for ET coupled problems with many inputs. By construction, the pMOR method produces sparse yet accurate ROMs as compared to the standard IpMOR-ET method. Future work will include the extension of the method to more general settings regarding the coupling structure.

References

1. Bader, B.W., Kolda, T.G.: Matlab tensor toolbox version 2.6 (2015). Available via DIALOG. http://www.sandia.gov/~tgkolda/TensorToolbox. Cited 03 Mar 2016
2. Banagaaya, N., Feng, L., Schoenmaker, W., Meuris, P., Gillon, R., Benner, P.: Sparse model order reduction for electro-thermal problems with many inputs. In: Proceedings of Scientific Computing in Electrical Engineering, SCEE 2016, St. Wolfgang, 3–7 Oct (2016, to appear)

3. Banagaaya, N., Feng, L., Schoenmaker, W., Meuris, P., Benner. P.: An Index-aware Parametric Model Order Reduction for Parametrized Quadratic DAEs. Appl. Math. Comput, in press. doi:10.1016/j.amc.2017.04.024
4. Feng, L., Benner, P.: A robust algorithm for parametric model order reduction based on implicit moment matching. In: Alfio, Q., Gianluigi, R. (eds.) Reduced Order Methods for Modeling and Computational Reduction, pp. 159–186. Springer, Heidelberg (2014)

Nanoelectronic Coupled Problem Solutions: Uncertainty Quantification of RFIC Interference

Piotr Putek, Rick Janssen, Jan Niehof, E. Jan W. ter Maten, Roland Pulch, Bratislav Tasić, and Michael Günther

Abstract Due to the key trends on the market of RF products, modern electronics systems involved in communication and identification sensing technology impose requiring constraints on both reliability and robustness of components. The increasing integration of various systems on a single die yields various on-chip coupling effects, which need to be investigated in the early design phases of Radio Frequency Integrated Circuit (RFIC) products. Influence of manufacturing tolerances within the continuous down-scaling process affects the output characteristics of electronic devices. Consequently, this results in a random formulation of a direct problem,

P. Putek
Bergische Universität Wuppertal, Chair of Applied Mathematics and Numerical Analysis, Gaußstraße 20, 42119 Wuppertal, Germany

Ernst-Moritz-Arndt-Universität Greifswald, Institute for Mathematics and Computer Science, Walther-Rathenau-Straße 47, 17489 Greifswald, Germany
e-mail: Piotr.Putek@math.uni-wuppertal.de

R. Janssen • J. Niehof • B. Tasić
NXP Semiconductors B.V., High Tech Campus 46, 5656 AE Eindhoven, The Netherlands
e-mail: Rick.Janssen@nxp.com; Jan.Niehof@nxp.com; Bratislav.Tasic@nxp.com

E.J.W. ter Maten (✉)
Bergische Universität Wuppertal, Chair of Applied Mathematics and Numerical Analysis, Gaußstraße 20, 42119 Wuppertal, Germany

Eindhoven University of Technology, Chair of Mathematics and Computer Science (CASA), 5600 MB Eindhoven, The Netherlands
e-mail: Jan.ter.Maten@math.uni-wuppertal.de; E.J.W.ter.Maten@tue.nl

R. Pulch
Ernst-Moritz-Arndt-Universität Greifswald, Institute for Mathematics and Computer Science, Walther-Rathenau-Straße 47, 17489 Greifswald, Germany
e-mail: pulchr@uni-greifswald.de

M. Günther
Bergische Universität Wuppertal, Chair of Applied Mathematics and Numerical Analysis, Gaußstraße 20, 42119 Wuppertal, Germany
e-mail: Michael.Guenther@math.uni-wuppertal.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_42

whose solution leads to robust and reliable simulations of electronics products. Therein, the statistical information can be included by a response surface model, obtained by the Stochastic Collocation Method (SCM) with Polynomial Chaos (PC). In particular, special emphasis is given to both the means of the gradient of the output characteristics with respect to parameter variations and to the variance-based sensitivity, which allows for quantifying impact of particular parameters to the variance. We present results for the Uncertainty Quantification of an integrated RFCMOS transceiver design.

1 Introduction

Modern mixed-signal and radio frequency (RF) integrated circuits (ICs) increasingly show the integration of various systems on a singe die [4, 5]. The integration involves both noisy parts, the so-called aggressors, and sensitive parts, the so-called victims and thus challenge the intellectual property blocks (IPs) to provide their proper and interference-free functioning. The integration goes hand in hand with progressive down scaling with impact on various parameters. The statistical variations, resulting from manufacturing tolerances of industrial processes, could lead to the acceleration of migration phenomena in semiconductor devices and finally can cause a thermal destruction of devices due to thermal runaway [7, 9, 10, 12].

Moreover, unintended RF coupling, which can occur both as a result of industrial imperfections and as a consequence of the integration process, might additionally downgrade the quality of products and their performance or even be dangerous for safety of both environment and the end users [2]. Meeting the specification requirements for electromagnetic compatibility standards and issues related to interference between IPs at early design stages allows for avoiding expensive re-spins and for the consecutive decrease of the time-to-market cycle. In this phase proper floorplanning and grounding strategies are studied [5]. It allows for the identification, quantification and prediction of cross-domain coupling. Figures 1 and 2 show a floorplan setup and a testbench model, which includes the key elements. Among the coupling paths investigated in [5] were (i) the exposed diepad and downbonds, (ii) the splitter cells, (iii) the substrate, and (iv) the air. We analyze the exposed diepad vias and downbonds paths with respect to a number of model parameter variations including the number of downbonds, the number of ground pins, and the number of exposed diepad vias. Cross-domain transfer functions $\mathbf{y}$ from the digital to the analogue RF domain are studied with respect to input variations. We have a sinusoidal component of $|X|$, an angular frequency $\omega := 2\pi f$ and a phase $\phi := \arg(X)$ as input to a linear time-invariant system and, with corresponding output as $|Y|$ and $\phi_Y := \arg(Y)$, the frequency response of the transfer function and the phase shift are defined by $G(\omega) = |Y|/|X| =: |H(\mathrm{i}\omega)|$ and $\phi(\omega) := \phi_Y - \phi_X = \arg(H(\mathrm{i}\omega))$, respectively.

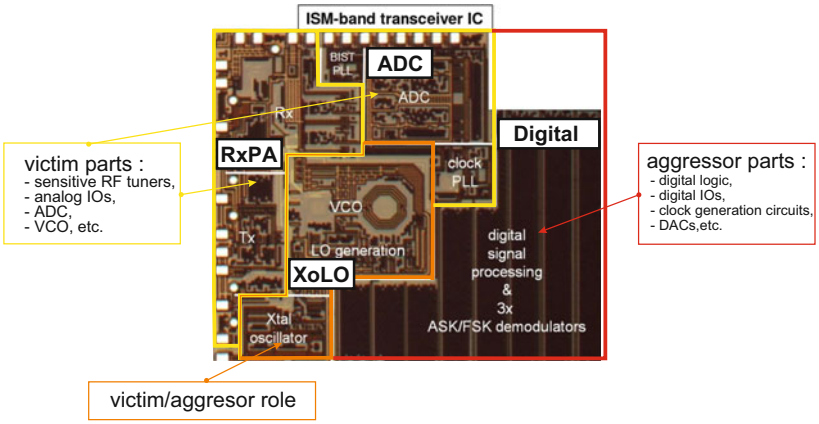


Fig. 1 Chip architecture with domains indicated [5]: Floorplan model for isolation and grounding strategies [2]

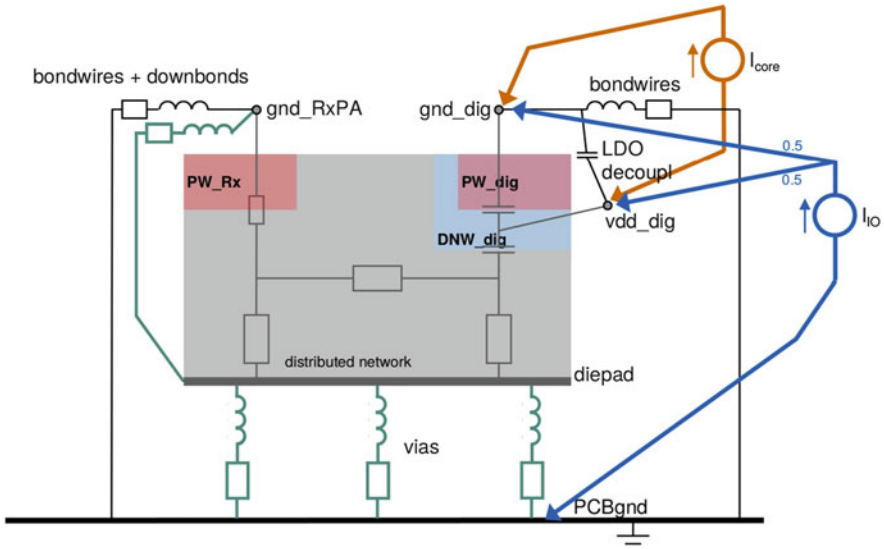


Fig. 2 Chip architecture with domains indicated [5]: Testbench model for an RFIC isolation problem

2 Stochastic Modeling

We apply stochastic modeling for a floorplan model with grounding strategies. The physical design, shown in Fig. 2, involves on-chip coupling effects, chip-package interaction, substrate coupling, leading to co-habitation issues. Consequently, a

direct problem is governed by a system of time-harmonic random-dependent Partial Differential Equations, derived from Maxwell's equations

$$\begin{cases} \nabla \cdot [\epsilon(\boldsymbol{\chi}) \nabla \Phi(\boldsymbol{\chi}) + \mathrm{i}\epsilon(\boldsymbol{\chi})\, \omega \mathbf{A}(\boldsymbol{\chi})] = \rho(\boldsymbol{\chi}) \\ \nabla \times (\nu(\boldsymbol{\chi}) \nabla \times \mathbf{A}(\boldsymbol{\chi})) = \mathbf{J}(\boldsymbol{\chi}) + \omega^2 \epsilon(\boldsymbol{\chi}) \left(\mathbf{A}(\boldsymbol{\chi}) - \frac{\mathrm{i}}{\omega} \nabla \Phi(\boldsymbol{\chi})\right) \\ \nabla \cdot \mathbf{A}(\boldsymbol{\chi}) + \mathrm{i}\omega k \Phi(\boldsymbol{\chi}) = 0 \\ \nabla \cdot \mathbf{J}(\boldsymbol{\chi}) + \mathrm{i}\omega \rho(\boldsymbol{\chi}) = 0, \end{cases} \tag{1}$$

equipped with suitable initial and boundary conditions. Here, $\boldsymbol{\chi} := (\mathbf{x}, f, \boldsymbol{\xi})$ is defined on $D \times D_{\mathrm{F}} \times \Xi$ with $D = D_1 \cup D_2 \cup D_3$ being a bounded domain in $\mathbb{R}^3$, composed of regions such as metal, insulator and semiconductor, respectively. D_{F} represents the frequency spectrum and Ξ is a multidimensional domain of physical parameters. The charge density ρ is represented by $\rho = q(n - p - N_{\mathrm{D}})$ on D_3 and 0 otherwise (on $D_{1,2}$); the current density $\mathbf{J}$ is defined as $\mathbf{J}_{D_1} = -\sigma(\nabla \Phi + \mathrm{i}\epsilon\, \omega \mathbf{A})$, $\mathbf{J}_{D_2} = 0$ and $\mathbf{J}_{D_3} = \mathbf{J}_n + \mathbf{J}_p$. Here, σ and ϵ are the electric conductivity and the permittivity. Φ is the scalar electric potential, while $\mathbf{A}$ is the magnetic vector potential. $\mathbf{J}_n$ and $\mathbf{J}_p$ denote electron and hole current densities, whereas n and p represent electron and hole concentrations. N_{D} refers to the doping concentration, k is a constant that depends on the scaling scenario. In order to obtain the solution of an integral equation formulation of (1), ADS/Momentum© from Keysight Technologies, http://www.keysight.com, has been used. Therein, Green's functions are applied to model the proper behavior of the substrate [3]. In our simulations, the Quasi-Static Mode is used, which provides accurate electromagnetic simulation performance in RF for the geometrically complex and electrically small designs.

3 Uncertainty Quantification

For Uncertainty Quantification (UQ), a type of SCM compound with the PC expansion has been used. In this respect, some parameters $z(\boldsymbol{\xi}) \in \Xi$ in the model (1) have been modified by random variables

$$z(\boldsymbol{\xi}) = [z_{\mathrm{dwnbond}}(\xi_1),\ z_{\mathrm{exp}}(\xi_2),\ z_{\mathrm{Xolo}}(\xi_3),\ z_{\mathrm{RxPa}}(\xi_4)], \tag{2}$$

where $\boldsymbol{\xi}$ is defined on the probability triple $(\Omega, \mathscr{F}, \mathbb{P})$ [13]. We assume a joint (uniform) probability density function $g : \Xi \to \mathbb{R}$ associated with $\mathbb{P}$ and that y is a quadratically integrable function. Then, a response surface model of y, in the form of a truncated series of the PC expansion [13], reads as

$$y(f, z) \doteq \sum_{i=0}^{N} v_i(f) \Psi_i(z), \tag{3}$$

with a priori unknown coefficient functions v_i and predetermined basis polynomials Ψ_i with the orthogonality property $\mathbb{E}\left[\Psi_i \Psi_j\right] = \delta_{ij}$. Here, $\mathbb{E}$ is the expected value, associated with $\mathbb{P}$. Specifically, for the calculation of the unknown coefficients v_i, we applied a pseudo-spectral approach with the Stroud-3 formula [10, 12, 14]. Within SCM, first the solution at each (deterministic) quadrature node $\boldsymbol{z}^{(k)}$, $k = 1, \ldots, K$ of the system (1) is determined, resulting in approximations for the v_i in the form of

$$v_i(f) \doteq \sum_{k=1}^{K} y\left(f, \boldsymbol{z}^{(k)}\right) \Psi_i\left(\boldsymbol{z}^{(k)}\right) w_k, \quad w_k \in \mathbb{R}. \tag{4}$$

Finally, the moments are approximated by, cf. [13],

$$\mathbb{E}\left[y\left(f, \boldsymbol{z}\right)\right] \doteq v_0(f), \quad \mathrm{Var}\left[y\left(f, \boldsymbol{z}\right)\right] \doteq \sum_{i=1}^{N} |v_i(f)|^2 \tag{5}$$

assuming $\Psi_0 = 1$. In order to investigate the impact of each uncertain parameter on the output variation, we performed a variance-based sensitivity analysis. The Sobol decomposition yields normalized variance-based sensitivity coefficients [6, 11]

$$S_j := \frac{\mathrm{V}_j^d}{\mathrm{Var}(y)} \quad \text{with} \quad \mathrm{V}_j^d := \sum_{i \in I_j^d} |v_i|^2, \ \ j = 1, \ldots, q, \tag{6}$$

with sets $I_j^d := \{j \in \mathbb{N} : \Psi_j(z_1, \ldots, z_q)$ is not constant in z_j and degree$(\Psi_i) \leq d\}$, where d is the maximum degree of the polynomials. We will have $d = 3$ and $q = 4$. Note that $0 \leq S_j \leq 1$. A value close to 1 means a large contribution to the variance. Differentiating (3) with respect to z_k gives $\partial y / \partial z_k$ at any value of $\mathbf{z}$. The z_k-th mean sensitivity is obtained by integrating over the whole parameter space [13].

4 Numerical Example and Conclusions

The model, shown schematically in Fig. 2, has been simulated within the frequency range from 1 MHz to 10 GHz. We performed UQ analysis using [1] for the frequency response functions y_2 and y_3,[1] which have been defined as (see also Fig. 2)

$$y_2 = |\mathrm{CplXolo}| := \frac{|\mathrm{gnd\_xolo - PCBgnd}|}{|\mathrm{Vdd\_dig - gnd\_dig}|}, \quad y_3 = |\mathrm{CplRx}| := \frac{|\mathrm{gnd\_rx - PCBgnd}|}{|\mathrm{Vdd\_dig - gnd\_dig}|}. \tag{7}$$

[1] $y_1 = |\mathrm{CplADC}|$ has been neglected due to its insensitivity w.r.t. the input variations.

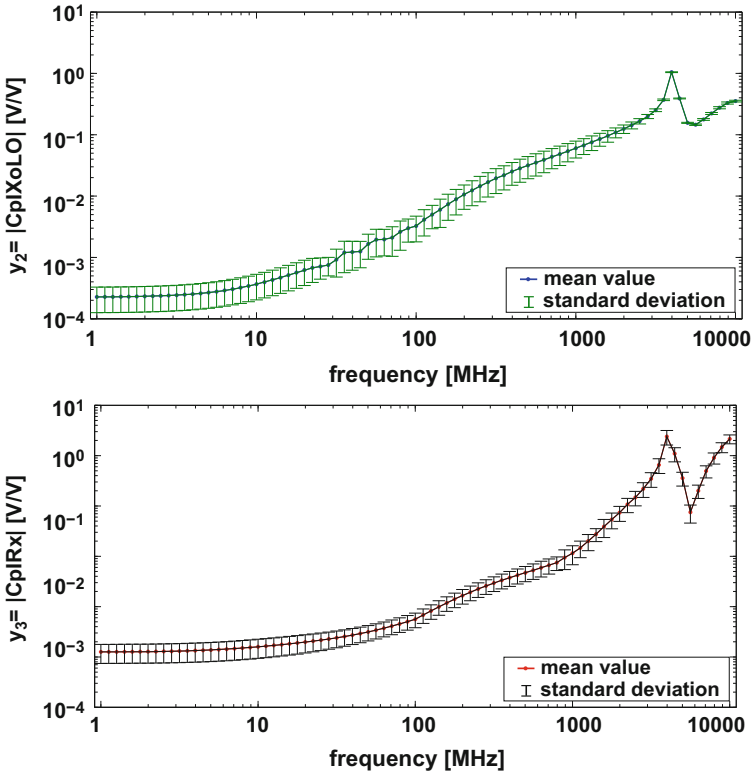


Fig. 3 Mean and standard deviation values of the modulus of the cross-domain frequency response transfer functions y_2 and y_3, calculated for the testbench model under input uncertainties

The results in terms of statistical moments have been depicted in Fig. 3. Here, we assumed that the input variations are described by a joint uniform discrete distribution, which describes numbers of parallel connected impedances. Therefore, in this case, the particular numbers of connected branches are generated using the range of discrete random variables as: $N_{\text{downbonds}} \in \langle 1, 10\rangle$, $N_{\text{exp}} \in \langle 1, 20\rangle$, $N_{\text{XoLO}} \in \langle 1, 8\rangle$, $N_{\text{RxPA}} \in \langle 1, 12\rangle$, thus $N := (N_{\text{downbonds}}, N_{\text{exp}}, N_{\text{XoLO}}, N_{\text{RxPA}})$, $R := (R_{\text{downbonds}}, R_{\text{exp}}, R_{\text{XoLO}}, R_{\text{RxPA}})$, $L := (L_{\text{downbonds}}, L_{\text{exp}}, L_{\text{XoLO}}, L_{\text{RxPA}})$. Consequently, the particular impedances z are defined as follows: $z(\omega) = [(R_1 + \mathrm{i}\omega L_1)/N_1, (R_2 + \mathrm{i}\omega L_2)/N_2, (R_3 + \mathrm{i}\omega L_3)/N_3, (R_4 + \mathrm{i}\omega L_4)/N_4]$, where $R_1 = 50.0[\mathrm{m}\Omega]$ and $L_1 = 0.1[\mathrm{nH}]$; $R_2 = 1.0[\mathrm{m}\Omega]$ and $L_2 = 0.1[\mathrm{nH}]$; $R_3 = 100.0[\mathrm{m}\Omega]$ and $L_3 = 2.0[\mathrm{nH}]$; $R_4 = 100.0[\mathrm{m}\Omega]$ and $L_4 = 2.0[\mathrm{nH}]$. The variance-based

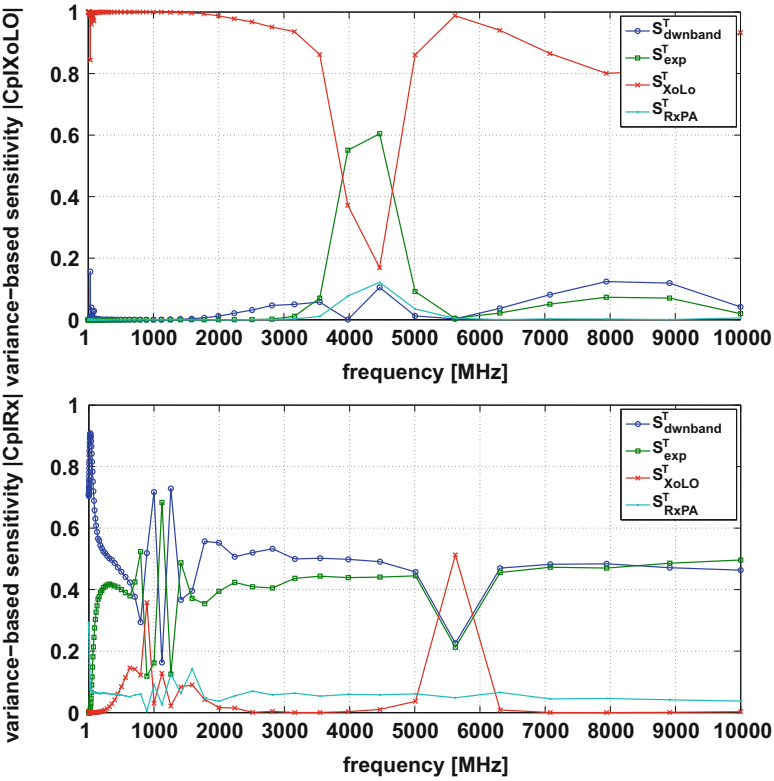


Fig. 4 Variance-based sensitivity performed for the testbench model. Due to the normalization, a value close to 1 means a large ('dominant') contribution to the variance

sensitivity coefficients, shown in Fig. 4, allow to find the most influential parameters contributing to the variance, whereas the mean gradients of y are presented in Fig. 5. Based on this analysis we further developed a regularized Gauss-Newton algorithm, which allows for finding robust optimized values of the considered parameters with minimum variation around the mean of an appropriate objective function [8].

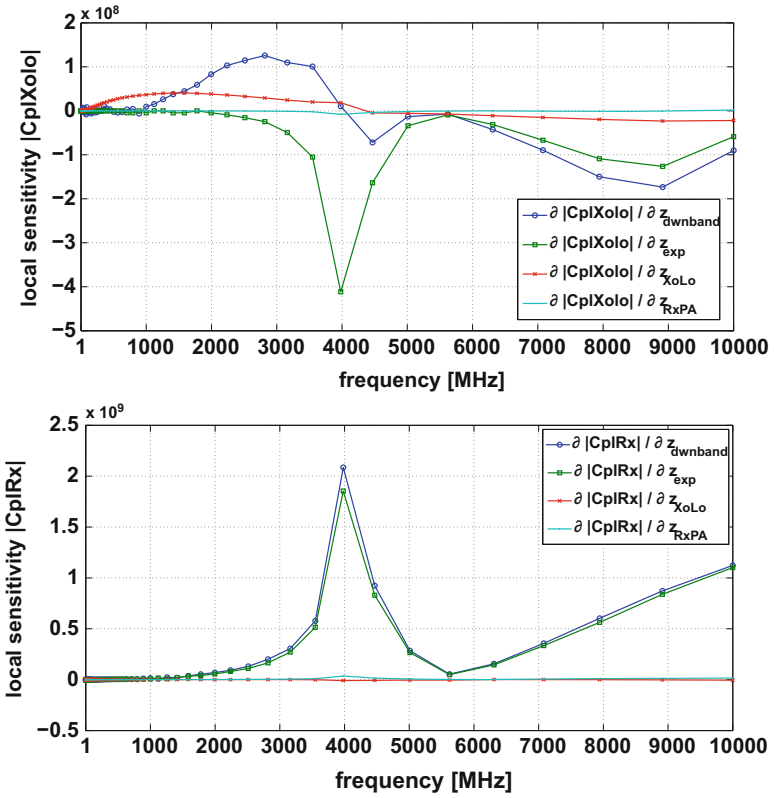


Fig. 5 Mean gradient sensitivity analysis performed for the testbench model. Shown are the means of the coordinates of the gradient of y with respect to $\mathbf{z}$

Acknowledgements The nanoCOPS (Nanoelectronic COupled Problems Solutions) project [12] is supported by the European Union in the FP7-ICT-2013-11 Program under the grant agreement number 619166, http://fp7-nanocops.eu/.

References

1. Dakota 6.2. https://dakota.sandia.gov/, Sandia National Laboratories (2015)
2. Di Bucchianico, A., ter Maten, E.J.W., Pulch, R., Janssen, R., Niehof, J., Hanssen, J., Kapora, S.: Robust and efficient uncertainty quantification and validation of RFIC isolation. Radioengineering **23**, 308–318 (2014)
3. Gharpurey, R., Meyer, R.G.: Modeling and analysis of substrate coupling in integrated circuits. IEEE J. Solid-State Circuits **31**(3), 344–353 (1996)
4. Kapora, S., Hanssen, M., Niehof, J., Sandifort, Q.: Methodology for interference analysis during early design stages of high-performance mixed-signal ICs. In: Proceedings of 2015 10th International Workshop on the Electromagnetic Compatibility of Integrated Circuits (EMC Compo), Edinburgh, 10–13 November, pp. 67–71 (2015)

5. Niehof, J., van Sinderen, J.: Preventing RFIC interference issues: A modeling methodology for floorplan development and verification of isolation- and grounding strategies. In: Proceedings SPI-2011, 15th IEEE Workshop on Signal Propagation on Interconnects, Naples, pp. 11–14 (2011)
6. Pulch, R., ter Maten, E.J.W., Augustin, F.: Sensitivity analysis and model order reduction for random linear dynamical systems. Math. Comput. Simulat. **111**, 80–95 (2015)
7. Putek, P., Meuris, P., Günther, M., ter Maten, E.J.W., Pulch, R., Wieers, A., Schoenmaker, W.: Uncertainty quantification in electro-thermal coupled problems based on a power transistor device. IFAC-PapersOnLine **48**, 938–939 (2015)
8. Putek, P., Janssen, R., Niehof, J., ter Maten, E.J.W., Pulch, R., Günther, M.: Robust optimization of an RFIC isolation problem under uncertainties. In: Langer, U., Amrhein, W., Zulehner, W. (eds.) Scientific Computing in Electrical Engineering (SCEE 2016). Series Mathematics in Industry. Springer (2017)
9. Putek, P., Meuris, P., Pulch, R., ter Maten, E.J.W., Günther, M., Schoenmaker, W., Deleu, F., Wieers, A.: Shape optimization of a power MOS device under uncertainties. In: Proceedings DATE-2016, Design, Automation and Test in Europe, Dresden, pp. 319–324 (2016)
10. Putek, P., Meuris, P., Pulch, R., ter Maten, E.J.W., Schoenmaker, W., Günther, M.: Uncertainty quantification for robust topology optimization of power transistor devices. IEEE Trans. Magn. **52**(3), 1700104 (2016)
11. Sudret, B.: Global sensitivity analysis using polynomial chaos expansions. Rel. Eng. Syst. Saf. **93**(7), 964–979 (2008)
12. ter Maten, E.J.W., et al.: Nanoelectronic coupled problems solutions – nanoCOPS: modelling, multirate, model order reduction, uncertainty quantification, fast fault simulation. J. Math. Ind. **7**(2), 19 pp. (2016)
13. Xiu, D.: Numerical Methods for Stochastic Computations – A Spectral Method Approach. Princeton University Press, Princeton (2010)
14. Xiu, D., Hesthaven, J.: High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**, 1118–1139 (2005)

Minisymposium: Mathematical Modeling of Charge Transport in Graphene and Low Dimensional Structure

Antonino La Magna, Giovanni Mascali, and Vittorio Romano

Description

The minisymposium was concerned with the mathematical modeling and simulation of charge transport in graphene and other 2D materials and in structures, like double gate MOSFETs, nano-ribbons and nano-wires, where the presence of confinement effects allows for the formal description of the carrier flow as that of a two dimensional or one dimensional electron gas. Graphene, thanks to its peculiar electrical properties and mechanical properties, is considered as one of the most promising materials for future electron devices. Lately, it has also been realized that, by increasing the miniaturization of devices, hot-spots are observed, that is zones with very high crystal temperature due to the release of energy by high energetic electrons. The effect is particularly relevant in materials with reduced dimensionality and confined structures. For these reasons, the minisymposium foresaw the discussion of the following arguments: ab initio calculations to furnish the correct band structures for the materials, e.g. doped graphene or graphene nanoribbons; thermal effects in the crystal lattice; analytical properties of the Schroedinger equation; Monte Carlo simulations; Wigner transport equation; numerical schemes

A. La Magna (✉)
IMM-CNR, Stradale Primosole, Catania, Italy
e-mail: antonino.lamagna@imm.cnr.it

G. Mascali
Dipartimento di Matematica, Università della Calabria and INFN-Gruppo c. Cosenza, Cosenza, Italy
e-mail: g.mascali@unical.it

V. Romano
Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: romano@dmi.unict.it

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_43

for the charge carrier transport equation; fluid models deduced from the kinetic transport equations.

The talks included in the minisymposium were the following:

- G. Mascali. Università della Calabria and INFN-Gruppo c. Cosenza (Italy). *Semiconductor thermal conductivity and how it is affected by embedding nanoparticles.*
- F. Vecil. Université Blaise Pascal (France). *Implementation on a high-performance computing platform of a deterministic solver for Double-Gate MOSFETs.*
- V. Romano. Universiy of Catania (Italy). *Deterministic and stochastic solutions of the Boltzmann equation for charge transport in graphene on substrates.*
- O. Muscato. Universiy of Catania (Italy). *Low-field electron mobility in silicon nanowires obtained via hydrodynamic simulations.*
- A. La Magna. IMM-CNR, Stradale Primosole, Catania (Italy). *Atomistic dynamics and electronic transport in hydrogenated graphene.*
- I. Deretzis. IMM-CNR, Stradale Primosole, Catania (Italy. *Selective band structure unfolding for the determination of the p-type character of oxygen-treated few-layer MoS2.*
- N. Rotundo. WIAS Berlin (Germany). *On some extension of energy-drift-diffusion models.*
- G. Alí. Università della Calabria (Italy). *Some analytical results for a two-temperature energy-transport model for semiconductors.*

Low-Field Electron Mobility in Silicon Nanowires

Orazio Muscato, Tina Castiglione, and Armando Coco

Abstract Silicon nanowires (SiNWs) are quasi-one-dimensional structures in which electrons are spatially confined in two directions and they are free to move in the orthogonal direction. The subband decomposition and the electrostatic force field are obtained by solving the Schrödinger—Poisson coupled system. The electron transport along the free direction can be tackled using a hydrodynamic model, formulated by taking the moments of the multisubband Boltzmann equation. We shall introduce an extended hydrodynamic model where closure relations for the fluxes and production terms have been obtained by means of the Maximum Entropy Principle of Extended Thermodynamics, and in which the main scattering mechanisms such as those with phonons and surface roughness have been considered. By using this model, the low field mobility for a Gate-All-Around (GAA) SiNW transistor has been evaluated.

1 Introduction

Silicon nanowire transistors (SNWTs) are considered an interesting alternative architecture to the conventional planar technology for devices, because of their improved electrostatic control of the channel via the gate voltage and the consequent suppression of short channel effects. Actual SNWTs are intrinsically three-dimensional, and spatial symmetries in nonidealized devices cannot be invoked to reduce the spatial degrees of freedom. In addition, it is well known that the behavior of field-effect transistors is dominated by electrostatics, which has therefore to be accurately simulated in order to reproduce the device electrical behavior. For this reason, simulations based on the self-consistent solution of Schrödinger-Poisson and transport equations in three dimensions represent a required tool to understand device behavior and extract design guidelines. The solution of the Schrödinger

O. Muscato (✉) • T. Castiglione
Dipartimento di Matematica e Informatica, University of Catania, Viale A. Doria 6, Catania, Italy
e-mail: orazio.muscato@unict.it; castiglione@dmi.unict.it

A. Coco
Oxford Brookes University, Oxford, UK
e-mail: acoco@brookes.ac.uk

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_44

equation, in particular, provides the correct density of states that is significantly different from the bulk density of states, due to transversal quantum confinement, and has an obvious direct impact on the charge density. Along the free direction of the wire the electron transport can be described using a semiclassical formulation based on the 1-D multisubband Boltzmann Transport Equation (MBTE), supposing to have channel lengths greater than 10 nm [24]. To solve the MBTE is not an easy task also from the numerical point of view, because it forms an integro-differential system in two dimensions in phase-space and one in time, with a complicate collisional operator. The Monte Carlo method provides a stochastic solution of the MBTE, although affected by statistical noise [4, 19–22]. In alternative, by taking moments of the MBTE one can obtain a hierarchy of balance equations, providing a good engineering-oriented approach. The modeling of the high-order moments (fluxes) and the moments on the collisional operator (production terms) which play an important role, can be tackled by means of the Maximum Entropy Principle (hereafter MEP) of Extended Thermodynamics [3], which has been widely used for semiconductor modeling [2, 5–8, 11–14, 16, 18]. This principle furnishes the form of the distribution functions that makes the best use of the knowledge of a finite number of moments, in such a way this distribution is useful to determine, analytically without any fitting procedure, the higher-order moments and the production terms. In this way, recently, an extended hydrodynamic model has been obtained [1, 9, 10, 15, 17] and we shall use it in order to evaluate the low-field mobility for a Gate-All-Around SiNW transistor.

2 Quantum Confinement and Electronic Transport

In bulk Silicon (Si) electrons which contribute mainly to the charge transport are those which belong to six equivalent valleys near to the X-point of the Brilloiun zone. In SiNW the band structure is altered with respect to the bulk case depending on the cross-section wire dimension, the atomic configuration, and the crystal orientation. Tight-binding (TB) calculations shows that the six equivalent Δ conduction valleys of the bulk Si are split into two groups because of the quantum confinement [25]. The subbands related to the four **unprimed** valleys Δ_4 ($[0 \pm 10]$ and $[00 \pm 1]$ orthogonal to the wire axis) are projected into a unique valley in the Γ point of the one-dimensional Brillouin zone. Therefore a SiNW is a direct band-gap semiconductor. The subbands related to the **primed** valleys Δ_2 ($[\pm 100]$ along the wire axis) are found at higher energies and exhibit a minimum, located at $k_x = \pm 0.37\pi/a_0$, and the energy gap between the Δ_4 and Δ_2 bottom valley is 117 meV. Moreover, from the energy dispersion relation $E(k)$ obtained from the TB, one can evaluate the effective mass m^* in the parabolic spherical band approximation, obtaining $m^*_{\Delta_2} = 0.94$, $m^*_{\Delta_4} = 0.27$ (in rest electron mass units).

In an ideal quantum wire, the system is assumed sufficiently long along the axis of the wire that translational invariance holds. For a quantum wire with linear expansion in x-direction, and confined in the plane $y - z$, the normed wave function

$\phi(x, y, z)$ can be written in the form

$$\phi(x, y, z) = \chi_l^{\mu}(y, z)\frac{e^{ik_x x}}{\sqrt{L_x}} \tag{1}$$

where μ is the valley index (one Δ_4 valley and two Δ_2 valleys), $l = 1, N_{sub}$ the subband index, $\chi_l^{\mu}(y, z)$ is the wave function of the l-th subband and μ-th valley, and the term $e^{ik_x x}/\sqrt{L_x}$ describes an independent plane wave in x-direction confined to the normalization length, $0 \leq x \leq L_x$, with wave number k_x.

Let us consider a set of conduction electrons moving in the wire. Any one electron will react to the confining potential $U(y, z)$ and to the presence of all other free electrons in the system. According to the *Hartree approximation*, we shall assume that a given electron feels the total potential $V_{tot}(x, y, z) = U(y, z) - eV(x, y, z)$, where V is the electrostatic potential and e is the absolute value of the electric charge. The spatial confinement in the (y, z) plane is governed by the Schrödinger-Poisson system (SP)

$$\begin{cases} H[V]\,\chi_{lx}^{\mu}[V] = \varepsilon_{lx}^{\mu}[V]\,\chi_{lx}^{\mu}[V] \\ \\ H[V] = -\dfrac{\hbar^2}{2m_{\mu}^*}\left(\dfrac{\partial^2}{\partial y^2} + \dfrac{\partial^2}{\partial z^2}\right) + V_{tot}(x, y, z) \\ \\ \nabla \cdot [\epsilon_0 \epsilon_r \nabla V(x, y, z)] = -e(N_D - N_A - n[V]) \\ \\ n[V](x, y, z, t) = \sum_{\mu}\sum_{l} \rho_l^{\mu}(x, t)|\chi_{lx}^{\mu}[V](y, z, t)|^2 \end{cases} \tag{2}$$

where $(2)_1$ is the Schrödinger equation in the Effective Mass Approximation (EMA), $(2)_3$ is the Poisson equation, and N_D, N_A are the assigned doping profiles (due to donors and acceptors). The electron density $n[V]$ is given by $(2)_4$, where $\rho_l^{\mu}(x, t)$ is the linear density in the μ-valley and l-subband which must be evaluated by the transport model (hydrodynamic/kinetic) in the free movement direction. In the following we shall assume a quadratic cross-section of the wire (having total dimension $L_y = L_z$) where the wire is surrounded by an oxide layer which gives rise to a deep potential barrier having $U = 4.05\,\text{eV}$. The SP system forms a coupled nonlinear PDEs, which it is usually solved by using an adaptive iteration scheme [23]. The solution gives the electrostatic potential V, as well as the eigenvalues (or subband energies) ε_{lx}^{μ} and the eigenvectors (or electron envelope wavefunctions) χ_{lx}^{μ} as a function of the unconfined x direction. Then the total electron energy in the μ-valley and l-subband is $E_l^{\mu} = \varepsilon_{lx}^{\mu} + \dfrac{\hbar^2 k_x^2}{2m_{\mu}^*}$.

The electronic transport, along the free motion direction, is governed by the following extended hydrodynamic model which has been obtained by taking

moments of the MBTE [10]

$$\frac{\partial \rho_l^\mu}{\partial t} + \frac{\partial (\rho_l^\mu V_l^\mu)}{\partial x} =$$

$$\rho_l^\mu \sum_{l'} \mathscr{C}_\rho(\mu, l, \mu, l') + \rho_l^\mu \sum_{l', \mu' \neq \mu} \mathscr{C}_\rho(\mu, l, \mu', l') \tag{3}$$

$$\frac{\partial (\rho_l^\mu V_l^\mu)}{\partial t} + \frac{2}{m_\mu^*} \frac{\partial (\rho_l^\mu W_l^\mu)}{\partial x} + \frac{\rho_l^\mu}{m_\mu^*} \frac{\partial \varepsilon_l^\mu}{\partial x} =$$

$$\rho_l^\mu \sum_{l'} \mathscr{C}_V(\mu, l, \mu, l') + \rho_l^\mu \sum_{l', \mu' \neq \mu} \mathscr{C}_V(\mu, l, \mu', l') \tag{4}$$

$$\frac{\partial (\rho_l^\mu W_l^\mu)}{\partial t} + \frac{\partial (\rho_l^\mu S_l^\mu)}{\partial x} + \rho_l^\mu \frac{\partial \varepsilon_l^\mu}{\partial x} V_l^\mu =$$

$$\rho_l^\mu \sum_{l'} \mathscr{C}_W(\mu, l, \mu, l') + \rho_l^\mu \sum_{l', \mu' \neq \mu} \mathscr{C}_W(\mu, l, \mu', l') \tag{5}$$

$$\frac{\partial (\rho_l^\mu S_l^\mu)}{\partial t} + \frac{\partial (\rho_l^\mu F_l^\mu)}{\partial x} + \frac{3\rho_l^\mu}{m_\mu^*} \frac{\partial \varepsilon_l^\mu}{\partial x} W_l^\mu =$$

$$\rho_l^\mu \sum_{l'} \mathscr{C}_S(\mu, l, \mu, l') + \rho_l^\mu \sum_{l', \mu' \neq \mu} \mathscr{C}_S(\mu, l, \mu', l') \tag{6}$$

where (3)–(6) represent balance equations for the density (ρ_l^μ), momentum (V_l^μ), energy (W_l^μ) and energy-flux (S_l^μ) respectively. This system is not closed due the presence of the high-order moment F_l^μ as well as of the production terms (i.e. the RHS). By using the MEP one determines the functional form of the distribution function and consequently one obtains $F_l^\mu = 6(W_l^\mu)^2/m_\mu^*$ and the terms $\mathscr{C}_\rho, \mathscr{C}_V, \mathscr{C}_W, \mathscr{C}_S$, which have been calculated taking into account scattering with acoustic and optical phonons, as well as surface roughness scattering (SRS). The above closed system is of hyperbolic type.

3 Low-Field Mobility

In order to test our extended hydrodynamic model, we have considered a Gate-All-Around (GAA) SiNW transistor. This is a Silicon nanowire with an added gate wrapped around it, in such a way we have a three contact device with source, drain, and gate. The channel has been homogeneously doped to $N_D = 3\ 10^{15}\,\mathrm{cm}^{-3}$ and

it is very long ($L_x = 120\,\text{nm}$) with respect to the transversal dimensions ($L_y = L_z = 10\,\text{nm}$), and the oxide thickness is 1 nm. The low-field mobility, as measured in long-channel devices at low fields, has a surprisingly strong influence on the performance of short-channel devices. For this reason, it is still used as a figure of merit for benchmarking different technology options and device architectures. In order to evaluate this parameter, we fix an uniform electric field along the channel of 1 kV/cm, and then we compute the mobility as the ratio between the average electron velocity and the driving field, as function of the gate voltage V_g.

The average electron velocity has been obtained with the following steps: (a) we fix V_g and solve the Schrödinger-Poisson system (2); (b) once this solution has been obtained, the energies ε_l^μ and wave functions χ_l^μ for each subband are exported into the hydrodynamic model, and the linear density ρ_l^μ is used as initial condition; (c) the hydrodynamic model (3)–(6) is solved and, in the stationary regime, the average velocity has been calculated. In Fig. 1 we show the low-field mobility as function of the gate voltage V_g, obtained by including/excluding the SRS mechanism.

From this figure its is clear how the SRS is the key scattering mechanism as it yields a very strong dependence of the low-field electron mobility on the effective field. The obtained results are in agreement with those obtained by means of Monte Carlo simulations [4].

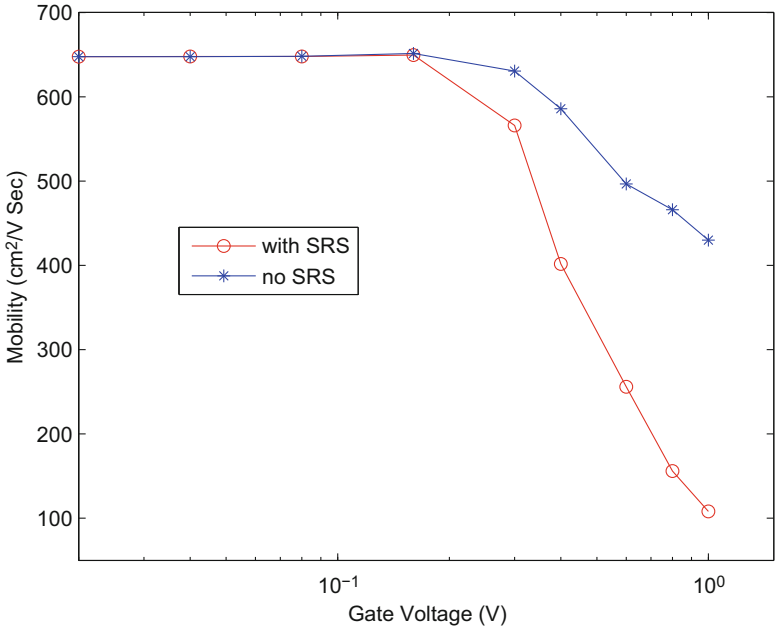


Fig. 1 Low-field mobility versus gate voltage, obtained with/without Surface Roughness Scattering mechanism

Acknowledgements We acknowledge the support of the Università degli Studi di Catania, FIR 2014 "Charge Transport in Graphene and Low dimensional Structures: modeling and simulation" and the National Group of Mathematical Physics (GNFM-INdAM), "Progetto giovani 2015".

References

1. Castiglione, T., Muscato, O.: Non-parabolic band hydrodynamic model for silicon quantum wires. J. Comput. Theor. Transport **46**(3), 186–201 (2017)
2. Di Stefano, V., Muscato, O.: Seebeck effect in silicon semiconductors. Acta Appl. Math. **122**(1), 225–238 (2012)
3. Lebon, G., Jou, D., Casas-Vázquez, J.: Understanding Non-equilibrium Thermodynamics. Springer, Berlin (2008)
4. Lenzi, M., Palestri, P., Gnani, E., Reggiani, S., Gnudi, A., Esseni, D., Selmi, L., Baccarani, G.: Investigation of the transport properties of silicon nanowires using deterministic and Monte Carlo approaches to the solution of the Boltzmann transport equation. IEEE Trans. Electron Devices **55**(8), 2086–2096 (2008)
5. Majorana, A., Mascali, G., Romano V.: Charge transport and mobility in monolayer graphene. J. Math. Ind. **7**, 4 (2017)
6. Mascali, G.: A hydrodynamic model for silicon semiconductors including crystal heating. Eur. J. Appl. Math. **26**(4), 477–496 (2015)
7. Mascali, G.: A new formula for Silicon thermal conductivity based on a hierarchy of hydrodynamical models. J. Stat. Phys. **163**(5), 1268–1284 (2016)
8. Mascali, G.: Thermal conductivity reduction by embedding nanoparticles. J. Comput. Electron. **16**(1), 180–189 (2017). doi:10.1007/s10825-016-0934-y
9. Muscato, O., Castiglione, T.: Electron transport in silicon nanowires having different cross-sections. Commun. Appl. Ind. Math. **7**(2), 8–25 (2016)
10. Muscato, O., Castiglione, T.: A hydrodynamic model for silicon nanowires based on the maximum entropy principle. Entropy **18**(10), 368 (2016)
11. Muscato, O., Di Stefano,V.: Local equilibrium and off-equilibrium thermoelectric effects in silicon semiconductors. J. Appl. Phys. **110**(9), 093706 (2011)
12. Muscato, O., Di Stefano, V.: An Energy Transport Model describing heat generation and conduction in silicon semiconductors. J. Stat. Phys. **144**, 171–197 (2011)
13. Muscato, O., Di Stefano, V.: Hydrodynamic modeling of the electro-thermal transport in silicon semiconductors. J. Phys. A: Math. Theor. **44**, 105501 (2011)
14. Muscato, O., Di Stefano, V.: Heat generation and transport in nanoscale semiconductor devices via Monte Carlo and hydrodynamic simulations. COMPEL **30**(2), 519–537 (2011)
15. Muscato, O., Di Stefano, V.: Hydrodynamic modeling of silicon quantum wires. J. Comput. Electron. **11**(1), 45–55 (2012)
16. Muscato, O., Di Stefano,V.: Electro-thermal behaviour of a sub-micron silicon diode. Semicond. Sci. Technol. **28**, 025021 (2013)
17. Muscato, O., Di Stefano, V.: Hydrodynamic simulation of a $n^+ - n - n^+$ silicon nanowire. Contin. Mech. Thermodyn. **26**, 197–205 (2014)
18. Muscato, O., Di Stefano,V.: Electrothermal transport in silicon carbide semiconductors via a hydrodynamic model. SIAM J. Appl. Math. **75**(4), 1941–1964 (2015)
19. Muscato, O., Wagner, W., Di Stefano, V.: Properties of the steady state distribution of electrons in semiconductors. Kinet. Relat. Models **4**(3), 809–829 (2011)
20. Muscato, O., Di Stefano, V., Wagner, W.: A variance-reduced electrothermal Monte Carlo method for semiconductor device simulation. Comput. Math. Appl. **65**(3), 520–527 (2013)
21. Ramayya, E.B., Knezevic, I.: Self-consistent Poisson-Schrödinger-Monte Carlo solver: electron mobility in silicon nanowires. J. Comput. Electron. **9**, 206–210 (2010)

22. Ryu, H.: A multi-subband Monte Carlo study on dominance of scattering mechanisms over carrier transport in sub-10-nm Si nanowire FETs. Nanoscale Res. Lett. **11**(1), 36 (2016)
23. Trellakis, A., Galik, A.T., Pacelli, A., Ravaioli, U.: Iteration scheme for the solution of the two-dimensional Schrödinger-Poisson equations in quantum structures. J. Appl. Phys. **81**(12), 7880–7884 (1997)
24. Wang, J., Lundstrom, M.: Does source-to-drain tunneling limit the ultimate scaling of MOSFETs? In: IEEE IEDM Technical Digest, pp. 707–710 (2002)
25. Zheng, Y., Rivas, C., Lake, R., Alam, K., Boykin, T.B., Klimeck, G.: Electronic properties of silicon nanowires. IEEE Trans. Electron Devices **52**(6), 1097–1103 (2005)

On Some Extension of Energy-Drift-Diffusion Models: Gradient Structure for Optoelectronic Models of Semiconductors

Alexander Mielke, Dirk Peschka, Nella Rotundo, and Marita Thomas

Abstract We derive an optoelectronic model based on a gradient formulation for the relaxation of electron-, hole- and photon-densities to their equilibrium state. This leads to a coupled system of partial and ordinary differential equations, for which we discuss the isothermal and the non-isothermal scenario separately.

1 Introduction

Our goal is to construct hybrid energy-drift diffusion models for semiconductors by including optic effects of microscopic or quantum-mechanical nature. For this, we want to apply the modeling approach of GENERIC (General Equations for Non-Equilibrium Reversible Irreversible Coupling), cf. [10]. As a first step in this direction, we study in this work how to formulate a model for optoelectronic processes in semiconductors as a generalized gradient system. We will discuss the isothermal case and the non-isothermal case. For both cases we will consider the admissible state variables $\boldsymbol{q}$ in a state space $\boldsymbol{Q}$, and determine a suitable thermodynamical functional $\Phi : \boldsymbol{Q} \to \mathbb{R}$, that drives all the optoelectronic processes, as well as a potential $\Psi^* : \boldsymbol{Q} \times \boldsymbol{Q}^* \to [0, \infty)$ encoding the different additive dissipative processes in the material. With the aid of these functionals Φ and Ψ^*, the evolution of the state $\boldsymbol{q}$ is governed by the rate equation

$$\dot{\boldsymbol{q}} = \mathrm{D}_\eta \Psi^*(\boldsymbol{q}; -\mathrm{D}_q \Phi(\boldsymbol{q})) \quad \text{in } \boldsymbol{Q} \tag{1}$$

where $\dot{\boldsymbol{q}}$ denotes the time-derivative of the state vector $\boldsymbol{q}$ and $\mathrm{D}_\eta \Psi^*(\boldsymbol{q}; \boldsymbol{\eta})$ denotes the derivative of $\Psi^*(\boldsymbol{q}; \boldsymbol{\eta})$ with respect to the variable $\boldsymbol{\eta}$, and similarly, $\mathrm{D}_q \Phi(\boldsymbol{q})$ the derivative of $\Phi(\boldsymbol{q})$ with respect to $\boldsymbol{q}$.

It was first shown in [5, 11, 12] that certain diffusion processes can be written as gradient flows of the form (1) with Φ given by the free energy or the entropy

A. Mielke • D. Peschka (✉) • N. Rotundo • M. Thomas
Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany
e-mail: mielke@wias-berlin.de; peschka@wias-berlin.de; rotundo@wias-berlin.de; thomas@wias-berlin.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_45

functional. In [7, 8] it was proven that also *reaction*-diffusion systems can be cast in the gradient structure of (1) and in [4, 7] this approach was adapted to the drift-diffusion and recombination processes arising in semiconductors.

In this contribution we will show that also the optoelectronic models treated in, e.g., [2, 13, 14] fall into the framework of (1), see [4, 7, 8]. These models consist of transport equations for the densities of charge carriers, electrons n and holes p, and describe their motion in the device driven by diffusion and drift in a self-consistent electrical field $\boldsymbol{E} = -\nabla\varphi_c$. In addition, electrons, holes and photons are generated or annihilated according to different radiative and non-radiative processes, coupled to the equations in a thermodynamically consistent way. The number of photons generated by such kinds of radiative recombination mechanisms is determined by a photon rate equation, which arises from the corresponding reaction kinetics; this feature so far was not considered in [4, 7, 8].

To incorporate the photon rate equation to our gradient flow formulation we make use of the ideas of [15]. Therein, the key idea is to extend the classical thermodynamic treatment of electromagnetic radiation beyond the purely thermal black-body radiation towards *luminescent radiation* which is observed, e.g., in diodes and lasers. It is assumed that optoelectronic excitations in a semiconductor lead to an equilibrium of electron/hole chemical potentials μ_c, μ_v with the photon chemical potential μ_γ so that $\mu_\gamma = \mu_c + \mu_v$. This leads to a modified radiation formula following Bose-Einstein statistics for the number of photons per volume and energy interval

$$\rho(E)f(E,\mu) = \frac{E^2}{\pi^2}\left(\frac{n_r}{\hbar c}\right)^3 \left(\exp\left(\tfrac{E-\mu_\gamma}{k_B\theta}\right) - 1\right)^{-1} \approx \frac{E^2}{\pi^2}\left(\frac{n_r}{\hbar c}\right)^3 \exp\left(-\tfrac{E-\mu_\gamma}{k_B\theta}\right),$$

with $E = \hbar\omega$ beyond Planck distribution in equilibrium, where $\mu_\gamma \equiv 0$ for photons, *cf.* [15]. From this, by standard thermodynamics, we can compute entropy and energies as functions of photon density γ and temperature θ, *cf.*, [6]. This approach allows us to adapt the framework of [4] to optoelectronics. In order to present simple closed-form expressions in this contribution, we assume that electron-, hole-, and photon-distributions can be approximated by Boltzmann distributions.

2 Gradient Structure: From Functionals to Differential Equations

Thermodynamic Functionals The optoelectronic system consists of negatively charged electrons n in the conduction band and positively charged holes p in the valence band. The density of states is parabolic with band edges E_c, E_v and effective masses m_c^*, m_v^* for conduction and valence band. Furthermore we have a photon density γ in the system with constant refractive index n_r and quadratic density of states. While the electron and hole densities n, p are space-dependent, it is assumed that the photon density γ is constant in space. This leads to the following effective

densities of state for electrons, holes, and photons

$$\bar{n} = N_c = 2\left(\frac{m_c^* k_B \theta}{2\pi\hbar^2}\right)^{3/2}, \quad \bar{p} = N_v = 2\left(\frac{m_v^* k_B \theta}{2\pi\hbar^2}\right)^{3/2}, \quad \bar{\gamma} = \frac{2}{\pi^2}\left(\frac{n_r k_B \theta}{\hbar c}\right)^3, \tag{2}$$

and to the intrinsic carrier density $n_i^2 = \bar{n}\bar{p}\exp\big((E_v - E_c)/(k_B\theta)\big)$. In (2) by $\hbar$ we denote the Dirac constant, k_B is the Boltzmann constant, and θ is the absolute temperature. We use the notation $\boldsymbol{c} = (c_1, c_2, c_3) = (n, p, \gamma)$ and $\bar{\boldsymbol{c}} = (\bar{c}_1, \bar{c}_2, \bar{c}_3) = (\bar{n}, \bar{p}, \bar{\gamma})$, where all $c_i, \bar{c}_i$ are densities with units of $(\text{length})^{-3}$. Assuming the Boltzmann distributions for charge carriers and photons, we define the internal energy $\mathscr{U}$, the entropy $\mathscr{S}$, and the free energy $\mathscr{F}$ of the system as

$$\mathscr{U}(\boldsymbol{c}, \theta) = \int_\Omega U \,\mathrm{d}x, \qquad \mathscr{S}(\boldsymbol{c}, \theta) = \int_\Omega S \,\mathrm{d}x, \qquad \mathscr{F}(\boldsymbol{c}, \theta) = \int_\Omega F \,\mathrm{d}x, \tag{3a}$$

where $\Omega \subset \mathbb{R}^3$ is an open bounded domain occupied by the semiconductor. Moreover, $F(\boldsymbol{c}, \theta) = U(\boldsymbol{c}, \theta) - \theta S(\boldsymbol{c}, \theta)$, where the densities U and S are given by

$$U(\boldsymbol{c}, \theta) = \frac{\varepsilon}{2}|\nabla\varphi_c|^2 + k_B\theta\left(\tfrac{3}{2}(n+p) + 3\gamma\right) + E_c n - E_v p + c_V\theta, \tag{3b}$$

$$S(\boldsymbol{c}, \theta) = -k_B\left[n\left(\log\tfrac{n}{\bar{n}} - \tfrac{5}{2}\right) + p\left(\log\tfrac{p}{\bar{p}} - \tfrac{5}{2}\right) + \gamma\left(\log\tfrac{\gamma}{\bar{\gamma}} - 4\right)\right] + c_V \log\theta, \tag{3c}$$

in which $c_V\theta$ and $c_V \log\theta$ constitute lattice contributions to the internal energy in (3b) and the entropy in (3c), where c_V denotes the heat capacity of the lattice. The contribution to the entropy by electrons and holes are given by the Sackur-Tetrode equation. The internal energy density U in (3b) contains the electrostatic potential φ_c, which depends implicitly on $\boldsymbol{c}$ as it is determined by the Poisson equation

$$-\nabla\cdot\varepsilon\nabla\varphi_c = e(C + p - n), \tag{3d}$$

where C is the given concentration of dopants, e is the elementary charge and ε denotes the dielectric permittivity. Now we define dual dissipation functionals $\Psi^*(\boldsymbol{q}; \boldsymbol{\eta}) = \Psi^*_{\text{rec}}(\boldsymbol{q}; \boldsymbol{\eta}) + \Psi^*_{\text{diff}}(\boldsymbol{q}; \boldsymbol{\eta})$, for which we separately discuss additive contributions due to recombinations and diffusion.

Dissipation by Recombinations with Detailed Balance Between electrons, holes, and photons we consider different types of recombinations r with stoichiometric coefficients $\boldsymbol{\alpha}^r = (\alpha_1^r, \alpha_2^r, \alpha_3^r)$ and $\boldsymbol{\beta}^r = (\beta_1^r, \beta_2^r, \beta_3^r)$ in the sense of [7]. As can be found in, e.g., [2, 3, 13] recombinations, characteristic for optoelectronic semiconductor materials, can be written in the form

$$\alpha_1^r n + \alpha_2^r p + \alpha_3^r \gamma \underset{k_-^r}{\overset{k_+^r}{\rightleftharpoons}} \beta_1^r n + \beta_2^r p + \beta_3^r \gamma.$$

In particular, this form includes non-radiative recombination-generation processes such as Auger- and Shockley-Read-Hall, as well as radiative emission-absorption, both spontaneous and stimulated. More precisely, their specific form is

$$\begin{aligned}
n+p &\rightleftharpoons \emptyset && \text{thus} && \boldsymbol{\alpha}^1-\boldsymbol{\beta}^1=(1,1,0)^\top && \text{recombination-generation},\\
n+p+\gamma &\rightleftharpoons 2\gamma && \text{thus} && \boldsymbol{\alpha}^2-\boldsymbol{\beta}^2=(1,1,-1)^\top && \text{stimulated absorption-emission},\\
n+p &\rightleftharpoons \gamma && \text{thus} && \boldsymbol{\alpha}^3-\boldsymbol{\beta}^3=(1,1,-1)^\top && \text{spontaneous absorption-emission},
\end{aligned}$$

each of them with forward reaction rate k_+^r and backward reaction rate k_-^r. Assuming detailed balance there is a steady state $\hat{\boldsymbol{c}}$ so that $k_r(\boldsymbol{q})=k_+^r(\boldsymbol{q})\,\hat{\boldsymbol{c}}^{\boldsymbol{\alpha}^r}=k_-^r(\boldsymbol{q})\,\hat{\boldsymbol{c}}^{\boldsymbol{\beta}^r}$. This notation means $\hat{\boldsymbol{c}}^{\boldsymbol{\alpha}^r}=\hat{n}^{\alpha_1^r}\hat{p}^{\alpha_2^r}\hat{\gamma}^{\alpha_3^r}$. Following [7, 8] and using $\widetilde{\boldsymbol{\eta}}=(\eta_n,\eta_p,\eta_\gamma)$ a suitable dual dissipation functional for the recombinations is given by

$$\Psi^*_{\mathrm{rec}}(\boldsymbol{q};\boldsymbol{\eta})=\int_\Omega \frac{1}{2}\widetilde{\boldsymbol{\eta}}\cdot\mathbb{H}(\boldsymbol{q})\widetilde{\boldsymbol{\eta}}\,\mathrm{d}x,\quad \mathbb{H}(\boldsymbol{q})=\sum_{r=1}^{3}\Lambda_r(\boldsymbol{q})(\boldsymbol{\alpha}^r-\boldsymbol{\beta}^r)\otimes(\boldsymbol{\alpha}^r-\boldsymbol{\beta}^r), \tag{4a}$$

$$\Lambda_r(\boldsymbol{q})=k_r(\boldsymbol{q})\,\ell\left(\left(\frac{\boldsymbol{c}}{\hat{\boldsymbol{c}}}\right)^{\boldsymbol{\alpha}^r},\left(\frac{\boldsymbol{c}}{\hat{\boldsymbol{c}}}\right)^{\boldsymbol{\beta}^r}\right),\quad \text{where}\quad \ell(x,y)=\begin{cases}\frac{x-y}{\log x-\log y} & x\neq y,\\ y & x=y.\end{cases} \tag{4b}$$

Note that $\mathbb{H}$, defined in (4a), is symmetric by construction and positive semidefinite on the stoichiometric subspace (for details see [7]).

Isothermal Model In the isothermal case we have $\boldsymbol{q}=\boldsymbol{c}$ and the gradient dynamics with fixed $\theta=\theta_\star$, is driven by the free energy $\mathscr{F}(\boldsymbol{c})\equiv\mathscr{F}(\boldsymbol{c},\theta_\star)$, so that the evolution of the state variables $\boldsymbol{q}=\boldsymbol{c}$ is given by $\dot{\boldsymbol{c}}=\mathrm{D}_{\boldsymbol{\eta}}\Psi^*\big(\boldsymbol{c};-\mathrm{D}\mathscr{F}(\boldsymbol{c})\big)$. Using a projector $\mathrm{P}\in\mathbb{R}^{2\times3}$ such that $\mathrm{P}\boldsymbol{\eta}=(\eta_n,\eta_p)^\top$, we introduce the dual dissipation potential for diffusion as follows

$$\Psi^*_{\mathrm{diff}}(\boldsymbol{q};\boldsymbol{\eta})=\int_\Omega\frac{1}{2}\nabla\mathrm{P}\boldsymbol{\eta}\cdot\mathbb{M}(\mathbf{q})\nabla\mathrm{P}\boldsymbol{\eta}\,\mathrm{d}x,\qquad \mathbb{M}(\boldsymbol{q})=\frac{1}{e}\begin{pmatrix}n\mu_n & 0\\ 0 & p\mu_p\end{pmatrix}, \tag{5}$$

where $\nabla\mathrm{P}\boldsymbol{\eta}\cdot\mathbb{M}\nabla\mathrm{P}\boldsymbol{\eta}$ means $\sum_i^d\partial_i(\mathrm{P}\boldsymbol{\eta})^\top\mathbb{M}\partial_i\mathrm{P}\boldsymbol{\eta}$. Here $\mu_n,\mu_p>0$ represent electron and hole mobilities. The dual dissipation potential comprising recombination and diffusion is given by $\Psi^*(\boldsymbol{q};\boldsymbol{\eta})=\Psi^*_{\mathrm{diff}}(\boldsymbol{q};\boldsymbol{\eta})+\Psi^*_{\mathrm{rec}}(\boldsymbol{q};\boldsymbol{\eta})$. Integrating by parts we find that the derivative of Ψ^* is

$$\langle\mathrm{D}_{\boldsymbol{\eta}}\Psi^*(\boldsymbol{q};\boldsymbol{\eta}),\boldsymbol{\nu}\rangle=\int_\Omega\boldsymbol{\nu}^\top\left[-\mathrm{P}^\top\nabla\cdot\mathbb{M}(\boldsymbol{q})\nabla\mathrm{P}\boldsymbol{\eta}+\mathbb{H}(\boldsymbol{q})\boldsymbol{\eta}\right]\mathrm{d}x \tag{6}$$

where $\boldsymbol{\nu}=(\nu_n,\nu_p,\nu_\gamma)$ with ν_n,ν_p are general functions and ν_γ is constant. When performing integration by parts we used appropriate boundary conditions for the

boundary terms to vanish. By the properties of $\mathbb{M}$ we clearly see that $\mathrm{D}_{\boldsymbol{q}}\Psi^*(\boldsymbol{q};\cdot) : \boldsymbol{Q}\times\boldsymbol{Q}^* \to \mathbb{R}$ is a symmetric and positive semidefinite operator. Due to (3) we have

$$\partial_{\boldsymbol{c}} F(\boldsymbol{c},\theta_\star) = k_{\mathrm{B}}\theta_\star \begin{pmatrix} \log(n/\bar{n}) \\ \log(p/\bar{p}) \\ \log(\gamma/\bar{\gamma}) \end{pmatrix} + \begin{pmatrix} E_{\mathrm{c}} - e\varphi_c \\ e\varphi_c - E_{\mathrm{v}} \\ 0 \end{pmatrix}. \tag{7}$$

Then (6) and (7) lead to the abstract form

$$\dot{n} = -\nabla\cdot\boldsymbol{j}_n - \left[\Lambda_1(\partial_n F + \partial_p F) + (\Lambda_2+\Lambda_3)(\partial_n F + \partial_p F - \partial_\gamma F)\right], \tag{8a}$$

$$\dot{p} = -\nabla\cdot\boldsymbol{j}_p - \left[\Lambda_1(\partial_n F + \partial_p F) + (\Lambda_2+\Lambda_3)(\partial_n F + \partial_p F - \partial_\gamma F)\right], \tag{8b}$$

$$|\Omega|\dot{\gamma} = \int_\Omega (\Lambda_2+\Lambda_3)\left[\partial_n F + \partial_p F - \partial_\gamma F\right]\mathrm{d}x, \tag{8c}$$

where the currents $\boldsymbol{j}_n$ and $\boldsymbol{j}_p$ are defined as

$$\begin{pmatrix} \boldsymbol{j}_n \\ \boldsymbol{j}_p \end{pmatrix} = -\mathbb{M}\nabla \begin{pmatrix} \partial_n F \\ \partial_p F \end{pmatrix} = -\mathbb{M}\nabla \begin{pmatrix} k_{\mathrm{B}}\theta_\star \log(n/\bar{n}) + E_{\mathrm{c}} - e\varphi_c \\ k_{\mathrm{B}}\theta_\star \log(p/\bar{p}) - E_{\mathrm{v}} + e\varphi_c \end{pmatrix}.$$

Using (4b) we get

$$\dot{n} = -\nabla\cdot\boldsymbol{j}_n - R_{\mathrm{nr}} - g(\boldsymbol{q})\left(\tfrac{np}{n_i^2} - \tfrac{\gamma}{\bar{\gamma}}\right), \tag{9a}$$

$$\dot{p} = -\nabla\cdot\boldsymbol{j}_p - R_{\mathrm{nr}} - g(\boldsymbol{q})\left(\tfrac{np}{n_i^2} - \tfrac{\gamma}{\bar{\gamma}}\right), \tag{9b}$$

$$|\Omega|\dot{\gamma} = -\left[\int_\Omega \tfrac{g(\boldsymbol{q})}{\bar{\gamma}}\,\mathrm{d}x\right]\gamma + \int_\Omega g(\boldsymbol{q})\frac{np}{n_i^2}\,\mathrm{d}x, \tag{9c}$$

where $\hat{n}\hat{p} = n_i^2(\theta_\star)$ and $\hat{\gamma} = \bar{\gamma}$. The term $R_{\mathrm{nr}} = k_1(\boldsymbol{q})(np/n_i^2 - 1)$, represents non-radiative recombination-generation processes. The rate of optical transitions $g(\boldsymbol{q}) = (\gamma/\bar{\gamma})k_2(\boldsymbol{q}) + k_3(\boldsymbol{q})$ includes stimulated and spontaneous absorption-emission and resembles optical gain in lasers. An equilibrium state of (9) is characterized by $n_{\mathrm{eq}}p_{\mathrm{eq}} = n_i^2$ and $\gamma_{\mathrm{eq}} = \bar{\gamma}$. To determine the equilibrium carrier densities $n_{\mathrm{eq}}, p_{\mathrm{eq}}$ requires to solve the Poisson equation (3d), which is nonlinear due to the insertion of $p_{\mathrm{eq}} = n_i\exp(-e\varphi_c/(k_{\mathrm{B}}\theta_\star))$ and $n_{\mathrm{eq}} = n_i\exp(e\varphi_c/(k_{\mathrm{B}}\theta_\star))$. With no-flux boundary conditions the gradient system (9) then implies a monotonous decay of the free energy towards this equilibrium

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}(\boldsymbol{c}(t)) = \langle \mathrm{D}_c\mathscr{F}, \dot{\boldsymbol{c}}\rangle = -\langle \mathrm{D}_c\mathscr{F}, \mathrm{D}_{\boldsymbol{\eta}}\Psi^*(\boldsymbol{c};\mathrm{D}_c\mathscr{F})\rangle \le 0 \tag{10}$$

as a consequence of the positive semi-definiteness of $\mathrm{D}_{\boldsymbol{\eta}}\Psi^*(\boldsymbol{q};\cdot)$.

Non-isothermal Model In the non-isothermal case it is advantageous to start with the internal energy $u = U(\boldsymbol{c}, \theta)$ as an independent variable, *cf.* [1]. Then, we solve (3b) for $\theta = \Theta(\boldsymbol{c}, u)$ and set $\bar{\boldsymbol{q}} = (\boldsymbol{c}, u) \in \boldsymbol{Q}$. We redefine energy and entropy as $\bar{\mathscr{U}}(\bar{\boldsymbol{q}}) = \int_\Omega u \, \mathrm{d}x$, and $\bar{\mathscr{S}}(\bar{\boldsymbol{q}}) = \mathscr{S}\big(\boldsymbol{c}, \Theta(\boldsymbol{c}, u)\big)$. For any functional $\Phi : \boldsymbol{Q} \to \mathbb{R}$ the derivative $\mathrm{D}_{\bar{\boldsymbol{q}}}\Phi(\bar{\boldsymbol{q}})$ now has four components. Therefore, the dual dissipation potential $\bar{\Psi}^*$ in the non-isothermal case can be obtained by extending the dual dissipation potential Ψ^* of the isothermal case as follows

$$\bar{\Psi}^*(\bar{\boldsymbol{q}}; \boldsymbol{\eta}) = \int_\Omega \nabla(\mathrm{P}\boldsymbol{\eta}) \cdot \bar{\mathbb{M}}(\bar{\boldsymbol{q}}) \nabla \mathrm{P}\boldsymbol{\eta} + \boldsymbol{\eta} \cdot \bar{\mathbb{H}}(\bar{\boldsymbol{q}}) \boldsymbol{\eta} \, \mathrm{d}x, \tag{11a}$$

$$\text{where} \qquad \bar{\mathbb{M}} = \Theta \begin{pmatrix} \mathbb{M} & \mathbb{M}_{\mathrm{cross}} \\ \mathbb{M}_{\mathrm{cross}}^\top & \mathbb{M}_u \end{pmatrix} \quad \text{and} \quad \bar{\mathbb{H}} = \Theta \begin{pmatrix} \mathbb{H} & 0 \\ 0 & 0 \end{pmatrix}, \tag{11b}$$

where $\mathrm{P} \in \mathbb{R}^{3\times 4}$ is a projector $\mathrm{P}\boldsymbol{\eta} = (\eta_n, \eta_p, \eta_u)$. Observe that $\mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{U}}(\boldsymbol{c}, u) = (\boldsymbol{0}, 1)^\top$, so that the integrand in (11a) vanishes and thus also $\mathrm{D}_{\boldsymbol{\eta}}\bar{\Psi}^*(\bar{\boldsymbol{q}}; \mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{U}}(\bar{\boldsymbol{q}})) = 0$. In turn, this implies the conservation of the internal energy $\bar{\mathscr{U}}(\bar{\boldsymbol{q}}(t))$ as

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\bar{\mathscr{U}} = \langle \mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{U}}, \dot{\bar{\boldsymbol{q}}} \rangle = \langle \mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{U}}, \mathrm{D}_{\boldsymbol{\eta}}\bar{\Psi}^*(\bar{\boldsymbol{q}}; \mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{S}}) \rangle = \langle \mathrm{D}_{\boldsymbol{\eta}}\bar{\Psi}^*(\bar{\boldsymbol{q}}; \mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{U}}), \mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{S}} \rangle = 0,$$

where the last step follows from the symmetry of $\mathrm{D}_{\boldsymbol{\eta}}\bar{\Psi}^*(\bar{\boldsymbol{q}}; \cdot)$. The evolution of the thermo-optoelectronic system is given by $\dot{\bar{\boldsymbol{q}}} = \mathrm{D}_{\boldsymbol{\eta}}\bar{\Psi}^*(\bar{\boldsymbol{q}}; \mathrm{D}_{\bar{\boldsymbol{q}}}\bar{\mathscr{S}}(\bar{\boldsymbol{q}}))$ with the negative entropy as the driving functional. Since it is more common to write the equations using the temperature θ as a variable, we now reverse the previous change of variables and replace u with $U(\boldsymbol{c}, \theta)$ and $\Theta(\boldsymbol{c}, u)$ with θ and set $\Xi(\boldsymbol{c}, \theta) = (\boldsymbol{c}, U(\boldsymbol{c}, \theta))$. This generates the transformed dual dissipation potential w.r.t. the variables $\boldsymbol{q} = (\boldsymbol{c}, \theta)$ by $\Psi^*(\boldsymbol{c}, \theta; \boldsymbol{\eta}) = \bar{\Psi}^*(\Xi(\boldsymbol{c}, \theta); A\boldsymbol{\eta})$ inducing $\dot{\boldsymbol{q}} = \mathrm{D}_{\boldsymbol{\eta}}\Psi^*(\boldsymbol{q}; D_q\mathscr{S})$, where

$$A = \big(\mathrm{D}_{(c,\theta)}\Xi\big)^{-\top} = \begin{pmatrix} \mathbb{I}_{3\times 3} & -\frac{\partial_{\boldsymbol{c}} U}{\partial_\theta U} \\ 0 & \frac{1}{\partial_\theta U} \end{pmatrix}, \qquad A(\boldsymbol{q})\partial_{\boldsymbol{q}} S(\boldsymbol{q}) = \tfrac{1}{\theta} \begin{pmatrix} -\partial_{\boldsymbol{c}} F(\boldsymbol{c}, \theta) \\ 1 \end{pmatrix}.$$

The derivative $\mathrm{D}_{\boldsymbol{\eta}}\Psi^*$ is as in (6), but with a new projector P, we replaced $\boldsymbol{q} = \boldsymbol{c}$ by $\boldsymbol{q} = (\boldsymbol{c}, \theta)$, $\boldsymbol{\eta}$ by $\mathrm{A}\boldsymbol{\eta}$, and $\mathbb{H}, \mathbb{M}$ by $\bar{\mathbb{H}}, \bar{\mathbb{M}}$. In addition to (8) but with temperature-dependent coefficients we obtain the following equation for the temperature

$$(\partial_\theta U)\dot{\theta} = -\nabla \cdot \boldsymbol{j}_\theta + (\partial_n U)(\nabla \cdot \boldsymbol{j}_n) + (\partial_p U)(\nabla \cdot \boldsymbol{j}_p) + (\partial_{\boldsymbol{c}} U) \cdot \mathbb{H}\partial_{\boldsymbol{c}} F, \tag{12}$$

with $(\partial_{\boldsymbol{c}} U)^\top = (-e\varphi_c + E_\mathrm{c}, e\varphi_c - E_\mathrm{v}, 0) + \frac{1}{2}k_\mathrm{B}\theta(3, 3, 6)$ and the currents

$$\begin{pmatrix} \boldsymbol{j}_n \\ \boldsymbol{j}_p \\ \boldsymbol{j}_\theta \end{pmatrix} = \bar{\mathbb{M}}\nabla \frac{1}{\theta} \begin{pmatrix} \partial_n F \\ \partial_p F \\ -1 \end{pmatrix} \equiv \bar{\mathbb{M}}\nabla \frac{1}{\theta} \begin{pmatrix} k_\mathrm{B}\theta \log(n/\bar{n}) + E_\mathrm{c} - e\varphi_c \\ k_\mathrm{B}\theta \log(p/\bar{p}) - E_\mathrm{v} + e\varphi_c \\ -1 \end{pmatrix}. \tag{13}$$

Using $\Lambda_r(\boldsymbol{c},\theta)$ from (4b) as in (9) we get the explicit form

$$\dot{n} = -\nabla\cdot\boldsymbol{j}_n - R_{\rm nr} - g(\boldsymbol{q})\left(\tfrac{np}{n_i^2} - \tfrac{\gamma}{\bar\gamma}\right), \tag{14a}$$

$$\dot{p} = -\nabla\cdot\boldsymbol{j}_p - R_{\rm nr} - g(\boldsymbol{q})\left(\tfrac{np}{n_i^2} - \tfrac{\gamma}{\bar\gamma}\right), \tag{14b}$$

$$|\Omega|\dot\gamma = -\left[\int_\Omega \tfrac{g(\boldsymbol{q})}{\bar\gamma}\,\mathrm{d}x\right]\gamma + \int_\Omega g(\boldsymbol{q})\frac{np}{n_i^2}\,\mathrm{d}x, \tag{14c}$$

$$\hat{c}_V\dot\theta = -\nabla\cdot\boldsymbol{j}_\theta + \sum_{i=1}^{2}\partial_{c_i}U(\nabla\cdot\boldsymbol{j}_{c_i}) + \alpha_{\rm nr}R_{\rm nr} + \alpha_{\rm r}g\left(\tfrac{np}{n_i^2} - \tfrac{\gamma}{\bar\gamma}\right), \tag{14d}$$

where $\alpha_{\rm nr} = (\partial_n U + \partial_p U) = 3k_{\rm B}\theta + E_{\rm g}$ and $\alpha_{\rm r} = (\partial_n U + \partial_p U + \partial_\gamma U) = 6k_{\rm B}\theta + E_{\rm g}$, with the band gap is $E_{\rm g} = E_{\rm c} - E_{\rm v}$. The heat capacity $\hat{c}_V \equiv (\partial_\theta U) = c_V + k_{\rm B}\big(\frac{3}{2}(n+p) + 3\gamma\big) + E_{\rm c}'(\theta)n - E_{\rm v}'(\theta)p$ is usually dominated by c_V. Note in general $E_{\rm c}, E_{\rm v}, \bar{n}, \bar{p}$ depends on space and temperature, which will then consistently generate extra drift terms in the current in (13). Also observe that the same calculations as in (10) with $\mathrm{D}\mathscr{S}$ instead of $-\mathrm{D}\mathscr{F}$ provide the production of entropy $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{S}(\boldsymbol{q}(t)) \geq 0$.

3 Discussion

The non-isothermal model derived here is in the spirit of [2], but our approach is focussed on a concise derivation in the framework of gradient structures. Since a term-by-term comparison is beyond the scope of this paper let us mention a few key differences. The model in [2] is for a semiconductor laser, for which additionally light is spatially localized in a mode density χ solving a Helmholtz equation. Even though one can easily modify the functionals to create similar looking terms, e.g., g is replaced by $g|\chi|^2$ in the emission-absorption, the coherent, and in this sense more luminescent, character of light makes a thermodynamic approach using entropies more elusive. We believe that using the GENERIC formalism to couple the charge transport to Hamiltonian system which are either more microscopic, e.g., quantum mechanical [9], or to the classical Maxwell equation might help to clarify the situation. For the shortness of the presentation we assumed homogeneous natural boundary conditions of vanishing normal fluxes and obtained the decay of the free energy. Obviously the model is built so that it also supports electrical pumping of a photon field, i.e., an energy flux through non-homogeneous boundary conditions.

Acknowledgements This research was partially supported by DFG via project B4 in SFB 787 and by the Einstein Foundation Berlin via the MATHEON project OT1 in ECMath. The authors are grateful to M. Liero, A. Glitzky and T. Koprucki for fruitful discussions.

References

1. Albinus, G., Gajewski, H., Hünlich, R.: Thermodynamic design of energy models of semiconductor devices. Nonlinearity **15**(2), 367 (2002)
2. Bandelow, U., Gajewski, H., Hünlich, R.: Fabry-Perot Lasers: Thermodynamics-Based Modeling, pp. 63–85. Springer, New York (2005)
3. Chuang, S.: Physics of Optoelectronic Devices. Wiley Series in Pure and Applied Optics, vol. 22. Wiley, New York (1995)
4. Glitzky, A., Mielke, A.: A gradient structure for systems coupling reaction–diffusion effects in bulk and interfaces. Z. Angew. Math. Phys. **64**(1), 29–52 (2013)
5. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
6. Landau, L., Lifshitz, E.: Statistical Physics, vol. 5: Course of Theoretical Physics. Pergamon Press, Oxford (1980)
7. Mielke, A.: A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems. Nonlinearity **24**(4), 1329–1346 (2011)
8. Mielke, A.: On thermodynamical couplings of quantum mechanics and macroscopic systems. In: Mathematical Results in Quantum Mechanics: Proceedings of the QMath12 Conference, pp. 331–348. World Scientific (2015)
9. Mittnenzweig, M., Mielke, A.: An entropic gradient structure for Lindblad equations and GENERIC for quantum systems coupled to macroscopic models. arXiv preprint (2016). ArXiv:1609.05765
10. Öttinger, H.C.: Beyond Equilibrium Thermodynamics. Wiley, Hoboken, NJ (2005)
11. Otto, F.: Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory. Arch. Ration. Mech. Anal. **141**(1), 63–103 (1998)
12. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**(1–2), 101–174 (2001)
13. Peschka, D., Thomas, M., Glitzky, A., Nürnberg, R., Gärtner, K., Virgilio, M., Guha, S., Schroeder, T., Capellini, G., Koprucki, T.: Modeling of edge-emitting lasers based on tensile strained germanium microstrips. IEEE Photonics J. **7**(3), 1–15 (2015)
14. Wachutka, G.K.: Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **9**(11), 1141–1149 (1990)
15. Würfel, P.: The chemical potential of radiation. J. Phys. C Solid State Phys. **15**(18), 3967 (1982)

Minisymposium: Mathematics in Nanotechnology

Timothy G. Myers

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 focussed 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 minisymposium we intended to bring together a selection of speakers who discussed a broad range of topics relevant to nanoscience and who were able to demonstrate the relevance of mathematics to this research field.

The talks included in the minisymposium were the following:

- Tim Myers and Francesc Font (Centre de Recerca Matemàtica, Spain and University of Limerick, Ireland). *Can you trust mathematics at the nanoscale?*
- Helena Ribera (Centre de Recerca Matemàtica, Spain). *Model for nanoparticle melting with a Newton cooling condition and size-dependent latent heat.*
- Clemens Heitzinger (TU Wien, Austria). *Stochastic partial differential equations for the modeling of nanowire and nanopore sensors.*
- David Gómez-Castro (Complutense University of Madrid, Spain). *Nanocatalysis: homogenization of critical scale two-phase composites with nonlinear reaction.*
- Gary O'Keeffe (University of Limerick, Ireland). *Modelling the efficiency of a nanofluid based direct absorption solar collector.*

T.G. Myers (✉)
Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C, 08193 Bellaterra, Barcelona, Spain
e-mail: tmyers@crm.cat

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_46

A Model for Nanoparticle Melting with a Newton Cooling Condition and Size-Dependent Latent Heat

Helena Ribera and Timothy G. Myers

Abstract In this paper we study the melting of a spherical nanoparticle. To match with experimental data, the model includes several new features such as size-dependent latent heat and a cooling boundary condition at the boundary. Melt temperature variation and density change are also included. A novel form of Stefan condition is used to determine the position of the melt front. Results show that melting times can be significantly faster than those predicted by previous theoretical models, primarily due to the latent heat variation.

1 Introduction

Nanoparticles have a high ratio of surface to volume atoms, which makes them behave differently to their bulk counterparts: examples include enhanced mechanical strength; enhanced solar radiation absorption and superparamagnetism. A well-known nanoscale property is the decrease in the phase change temperature with particle size [4]. Variation in surface tension with radius has been reported too [13], though less noticeable. The focus of this paper is in the latent heat decrease with radius. Lai et al. [8] presented the first calorimetry measurements of the melting process of nanometer-sized tin particles, ranging from $5-50\,\text{nm}$ in diameter. They found a reduction of up to 70% from the bulk latent heat for the smaller sized particles. The molecular dynamic simulations/experiments of [2, 5, 7] observed the same qualitative behavior.

The mathematical modelling of phase change is termed the Stefan problem. A number of assumptions for mathematical convenience are discussed in [1, Table 1.1]. The theoretical study [6] considered melting point depression. In all of these studies the outer boundary temperature was assumed constant (greater than the melt temperature). Font and Myers [6] included density variation and melting point depression in their model. Back [3, §7.1–7.4] replaced the standard latent heat

H. Ribera (✉) • T.G. Myers
Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C, 08193 Bellaterra, Barcelona, Spain
e-mail: hribera@crm.cat; tmyers@crm.cat

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_47

expression in the energy balance with a size-dependent function from [8], and showed decreases in melt times.

In this paper we extend the previous works to produce a more realistic melting model. Specifically, we incorporate the variation of latent heat, melt temperature and density, and impose a physically realistic boundary condition. One final novelty in this work concerns the form of Stefan condition presented in [9].

2 Latent Heat Variation

Recent studies [8, 11, 14] have used functions involving $1/R$ (where R is the particle radius) to model size-dependent latent heat. All three models involve a single fitting parameter and exhibit some agreement with data but do not show the right speed of decay to the bulk value. Motivated by this, we propose

$$L_m = L_m^* \left(1 - e^{-C\frac{R}{R_c}}\right), \tag{1}$$

where C is our fitting parameter. To ensure C takes reasonable values we also introduce the capillary length $R_c = \sigma_{sl}^*/(\rho_s L_m^*)$, where L_m^* and σ_{sl}^* are the bulk latent heat and solid-liquid surface tension, respectively. Figure 1 shows a comparison of this exponential model with [8, 11, 14] and experimental data.

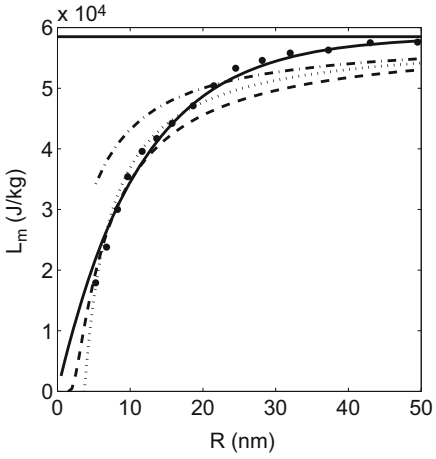


Fig. 1 Latent heat for a tin nanoparticle as a function of the radius for data (*dots*) and models of Lai et al. [8] (*dashed line*), Xiong et al. [14] (*dotted line*), Shin et al. [11] (*dash-dotted line*) and our relation (1) (*solid line*)

3 Mathematical Model

The physical situation considered here follows that described in a previous paper [6]. A nanoparticle with initial radius R_0 is subjected to an external heat source. Melting begins at the outer boundary and progresses inwards. The solid-liquid interface is denoted $R(t)$. Since the liquid and solid densities are different, the outer boundary $R_b(t)$ moves, where $R_b(0) = R_0$. The temperature in each phase is described by the standard heat equations

$$\rho_l c_l \left(\frac{\partial T}{\partial t} + v \frac{\partial T}{\partial r} \right) = k_l \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial T}{\partial r} \right), \qquad R(t) < r < R_b(t), \tag{2}$$

$$\rho_s c_s \frac{\partial \theta}{\partial t} = k_s \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \theta}{\partial r} \right), \qquad 0 < r < R(t), \tag{3}$$

where T and θ denote the temperature in the liquid and solid respectively, ρ_i, c_i and k_i are the densities, the specific heats and the conductivities respectively. The index notation $i = s, l$ refers to the solid, liquid phases. The velocity v at which heat is advected in Eq. (2) is given in [6, Eq. (8)].

The heat equation in the solid is solved over the region $0 \leq r \leq R(t)$. The position of the melt front, $R(t)$, is determined by the Stefan condition [9]

$$\left(\rho_s \left[L_m + \frac{v(R,t)^2}{2} \right] + \frac{2\sigma_{sl}^*}{R} \right) \frac{\mathrm{d}R}{\mathrm{d}t} = k_s \left. \frac{\partial \theta}{\partial r} \right|_{r=R} - k_l \left. \frac{\partial T}{\partial r} \right|_{r=R}. \tag{4}$$

The terms in the brackets represent the time-dependent latent heat, kinetic energy and the energy required to create new surface.

The boundary conditions are

$$-k_l \left. \frac{\partial T}{\partial r} \right|_{r=R_b} = h(T(R_b,t) - T_H), \quad T(R,t) = \theta(R,t) = T_m(t), \quad \theta_r(0,t) = 0, \tag{5}$$

where T_H is the temperature in the surrounding environment, T_m^* is the bulk melting temperature, $T_m(t)$ the melt temperature, and h the heat transfer coefficient. Following [6], we set $\theta(r,0) = T_m(0)$.

The position of the outer boundary is determined via $v(R_b) = dR_b/dt$ in equation [6, Eq. (8)] and integrating to give $R_b = \left(R_0^3 \rho - R^3(\rho - 1)\right)^{1/3}$ where $R(0) = R_b(0) = R_0$.

The fixed temperature boundary condition of previous studies is the limit of the first boundary condition in (5)a) as $h \to \infty$, but there exists a limit to the heat transfer beyond which the material vaporises. For comparison with previous models we choose the highest possible value for h which still permits thermodynamic stability [10]. For tin, $h_{\max} \approx 4.7 \times 10^9$ W/m^2 K.

The melt temperature $T_m(t)$ is approximated by the Gibbs-Thomson equation [12],

$$T_m(t) = T_m^* \left(1 - \frac{2\sigma_{sl}^*}{\rho_s L_m^* R}\right). \tag{6}$$

Substituting the parameters from [10, Table 1] into (6) we observe that $T_m(t)$ becomes negative for $R < 0.31$ nm. However, this is not an issue because our model is valid for R where continuum theory holds (i.e., $R > 2$–5 nm).

We scale the model using the dimensionless variables

$$\begin{gathered} \hat{T} = \frac{T - T_m^*}{\Delta T}, \quad \hat{\theta} = \frac{\theta - T_m^*}{\Delta T}, \quad \hat{T}_m = \frac{T_m - T_m^*}{\Delta T}, \quad \hat{L}_m = \frac{L_m}{L_m^*}, \\ \hat{r} = \frac{r}{R_0}, \quad \hat{R} = \frac{R}{R_0}, \quad \hat{R}_b = \frac{R_b}{R_0}, \quad \hat{t} = \frac{k_l}{\rho_l c_l R_0^2} t. \end{gathered} \tag{7}$$

We drop the hat notation and the resulting Stefan condition is

$$\rho\beta \left[L_m(t) + \frac{\alpha}{R}\right] \frac{\mathrm{d}R}{\mathrm{d}t} + \gamma \left(\frac{\mathrm{d}R}{\mathrm{d}t}\right)^3 = k \frac{\partial \theta}{\partial r}\bigg|_{r=R} - \frac{\partial T}{\partial r}\bigg|_{r=R}, \tag{8}$$

where the Stefan number $\beta = L_m^*/(c_l \Delta T)$, $\gamma = (1-\rho)^2 \rho_s \kappa_l^3/(2\Delta T k_l R_0^2)$, $\alpha = 2 * \sigma_{sl}^* s/(L_m^* \rho_s R_0)$, and $\kappa_l = k_l/(\rho_l c_l)$. Note, β is typically large due to melting point depression where only a very small increase above the melt temperature is sufficient to induce complete particle melting.

4 Perturbation Solution

Dividing (8) by β yields $\mathrm{d}R/\,\mathrm{d}t \approx 0$ (for sufficiently large β). Physically this implies that the large Stefan number solution corresponds to slow melting (slow compared to heat transfer in the material). As melting is our focus here, we thus rescale time via $\tau = \epsilon t$ where $\epsilon = 1/\beta \ll 1$. The Stefan condition becomes

$$\rho \left[L_m(t) + \frac{\alpha}{R}\right] \frac{\mathrm{d}R}{\mathrm{d}\tau} + \gamma \epsilon^3 \left(\frac{\mathrm{d}R}{\mathrm{d}\tau}\right)^3 = k \frac{\partial \theta}{\partial r}\bigg|_{r=R} - \frac{\partial T}{\partial r}\bigg|_{r=R}. \tag{9}$$

With the new time-scale the time derivatives in the two heat equations are now multiplied by ϵ. Hence, on the melting time-scale the heat equations are approximately pseudo-steady state (note, the boundary conditions still depend on time).

To solve for the liquid temperature we define the expansion $T = T_0 + \epsilon T_1 + \mathcal{O}(\epsilon^2)$. Upon substituting this expression into the corresponding heat equation with the new time scale and solving, we obtain

$$T_0(r,\tau) = F_1(\tau) + \frac{F_2(\tau)}{r}, \tag{10}$$

$$T_1(r,\tau) = \frac{r^2}{6}\frac{dF_1}{d\tau} + \frac{r}{2}\frac{dF_2}{d\tau} - \frac{F_3(\tau)}{r} + \frac{R^2 R_\tau F_2(\tau)(\rho-1)}{2r^2} + F_4(\tau), \tag{11}$$

where $F_1(\tau)$, $F_2(\tau)$, $F_3(\tau)$ and $F_4(\tau)$ are all functions of $R(\tau)$. We note that $F_{i_\tau} = R_\tau \overline{F}_i$, for all i. Similarly for the solid temperature we write $\Theta = \Theta_0 + \epsilon\Theta_1 + \mathcal{O}(\epsilon^2)$ and obtain solutions

$$\theta_0(r,\tau) = -\frac{\Gamma}{R}, \qquad \theta_1(r,\tau) = \frac{\Gamma}{6}\frac{\rho c}{k}\left(\frac{r^2-R^2}{R^2}\right)R_\tau. \tag{12}$$

Introducing these solutions into the Stefan condition (9) and via $F_{i_\tau} = R_\tau \overline{F}_i$, we obtain a cubic equation for speed of the melt front, R_τ,

$$\epsilon^3\gamma\left(\frac{dR}{d\tau}\right)^3 + \left(\rho\left[L_m(t) + \frac{\alpha}{R}\right] + \epsilon\left[\frac{R\overline{F}_1}{3} + \frac{\overline{F}_2}{2} + \frac{\overline{F}_3}{R^2} - \frac{F_2(\rho-1)}{R} - \frac{\Gamma\rho c}{3R}\right]\right)\frac{dR}{d\tau} - \frac{F_2}{R^2} = 0. \tag{13}$$

To follow the evolution of the nanoparticle we solve (13) via MATLAB, and then integrate the resultant first order differential equation for $R(\tau)$, subject to $R(0) = 1$.

To verify the accuracy of the perturbation solution we follow the numerical scheme described in Font and Myers [6] to solve the full problem.

5 Results

In this section we present the results of the model. In all cases we use data for tin, provided in [10, Table 1]. For comparison with the fixed boundary temperature model we also impose the maximum heat flux discussed earlier. Note, in all figures we plot R down to the non-dimensional equivalent of 2 nm.

In Fig. 2 we plot the variation of the radius $R(t)$ for Stefan number $\beta = 10$ with initial radius $R_0 = 10$, and for $\beta = 100$ with $R_0 = 100$ nm. The perturbation and numerical solutions are represented by the solid and dashed lines, respectively. As expected, large β solutions show greater agreement with the numerical solution. The graphs show that the evolution of the radius $R(t)$ is accurately predicted by the perturbation solution.

We note that melting times are almost doubled when using Stefan condition (8) as opposed to the standard condition from the literature. This can be explained by the fact that the current model predicts a melt rate of the order $R_t \propto 1/L_m(t)$, whereas

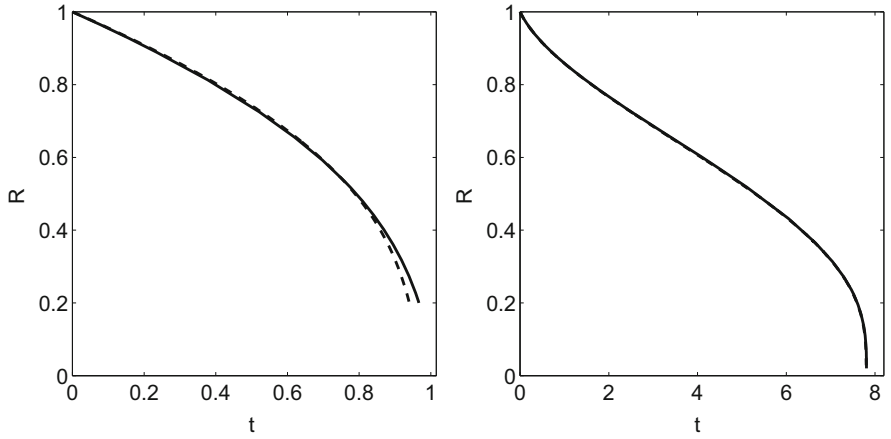


Fig. 2 Melting front evolution of a tin nanoparticle for perturbation (*solid line*) and numerical (*dashed line*) solutions. *Left*: $\beta = 10$, $R_0 = 10\,\text{nm}$. *Right*: $\beta = 100$, $R_0 = 100\,\text{nm}$

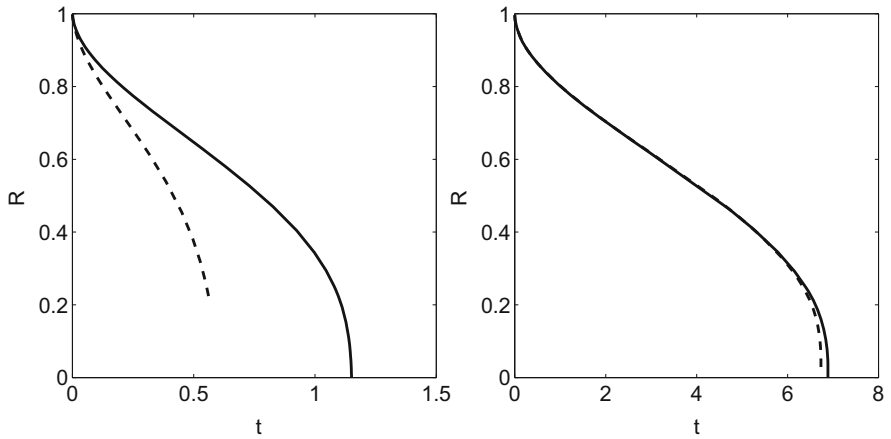


Fig. 3 Melt front position with a Newton cooling boundary condition (*solid line*) and fixed temperature boundary condition (*dashed line*), $\beta = 100$. *Left*: $R_0 = 10\,\text{nm}$. *Right*: $R_0 = 100\,\text{nm}$

previous treatments have order $R_t \propto 1/L_m^*$. Obviously, since $L_m(t) \to L_m^*$ as the radius increases the difference in results will decrease with an increase in initial particle size.

In previous models the boundary condition imposed was $T(R_b, t) = T_H$ instead of the Newton cooling condition used here. In Fig. 3 we show the difference in melting for the perturbation solution subject to the cooling condition (5) (solid line) and fixed temperature boundary condition (dashed line), both with $T_H = 507.6\,\text{K}$. For the 10 nm particle the change in boundary condition results in melting almost three times slower than with a fixed temperature. When $R_0 = 100\,\text{nm}$ the melting time increases by only 13.5%.

6 Conclusions

In this paper we presented a model to describe the melting of spherically symmetric nanoparticles which uses a new Stefan condition presented in [9] and a Newton cooling condition at the boundary. We included size-dependent latent heat and demonstrated better agreement with experimental data compared to previous models. We also included size-dependent melt temperature, and accounted for the difference in solid and liquid densities. The approximate analytical and full numerical solutions showed excellent agreement.

The inclusion of new Stefan condition with size-dependent latent heat resulted in a substantial decrease in melting times. We verified that the dominant factor in this is the size dependence of the latent heat rather than the Stefan condition. Nonetheless, the latter factor still has some impact on melting times.

Replacing the standard fixed temperature boundary condition with the more realistic cooling condition yields longer melting times. The fact that there is no longer an infinite melt rate initially also allowed us to neglect the kinetic energy term in the Stefan condition, provided the ratio between densities is close to unity.

Acknowledgements The authors acknowledge that the research leading to these results has received funding from "la Caixa" Foundation. TM acknowledges financial support from the Ministerio de Ciencia e Innovación grant MTM2014-56218.

References

1. Alexiades, V., Solomon, A.D.: Mathematical Modeling of Melting and Freezing Processes. Hemisphere, Washington, DC (1992)
2. Bachels, T., Güntherodt, H.-J., Schäfer, R.: Melting of isolated tin nanoparticles. Phys. Rev. Lett. **85**, 1250–1253 (2000)
3. Back, J.M.: Stefan problems for melting nanoscaled particles. Ph.D. thesis, U. Queensland (2014). http://eprints.qut.edu.au/79905/1/Julian_Back_Thesis.pdf
4. Buffat, P., Borel, J.-P.: Size effect on the melting temperature of gold particles. Phys. Rev. A **13**, 2287–2298 (1976)
5. Ercolessi, F., Andreoni, W., Tosatti, E.: Melting of small gold particles: mechanism and size effects. Phys. Rev. Lett. **66**, 911–914 (1991)
6. Font, F., Myers, T.G., Mitchell, S.L.: A mathematical model for nanoparticle melting with density change. Microfluid. Nanofluid. **18**, 233–243 (2014)
7. Jiang, H., Moon, K.-S., Dong, H., Hua, F., Wong, C.: Size-dependent melting properties of tin nanoparticles. Chem. Phys. Lett. **429**, 492–496 (2006)
8. Lai, S., Guo, J., Petrova, V., Ramanath, G., Allen, L.: Size-dependent melting properties of small tin particles: nanocalorimetric measurements. Phys. Rev. Lett. **77**, 99–102 (1996)
9. Myers, T.G.: Mathematical modelling of phase change at the nanoscale. Int. Commun. Heat Mass Transfer **76**, 59–62 (2016)
10. Ribera, H., Myers, T.G.: A mathematical model for nanoparticle melting with size-dependent latent heat and melt temperature. Microfluid. Nanofluid. **20**(11), 147 (2016)
11. Shin, J.-H., Deinert, M.R.: A model for the latent heat of melting in free standing metal nanoparticles. J. Chem. Phys. **140**, 164707 (2014)

12. Sun, J., Simon, S.: The melting behavior of aluminum nanoparticles. Thermochim. Acta **463**, 32–40 (2007)
13. Tolman, R.C.: The effect of droplet size on surface tension. J. Chem. Phys. **17**, 333 (1949)
14. Xiong, S., Qi, W., Cheng, Y., Huang, B., Wang, M., Li, Y.: Universal relation for size dependent thermodynamic properties of metallic nanoparticles. Phys. Chem. Chem. Phys. **13**, 10652–10660 (2011)

The Stochastic Drift-Diffusion-Poisson System for Modeling Nanowire and Nanopore Sensors

Leila Taghizadeh, Amirreza Khodadadian, and Clemens Heitzinger

Abstract We use the stochastic drift-diffusion-Poisson system to model charge transport in nanoscale devices. This stochastic transport equation makes it possible to describe device variability, noise, and fluctuations. We present—as theoretical results—an existence and local uniqueness theorem for the weak solution of the stochastic drift-diffusion-Poisson system based on a fixed-point argument in appropriate function spaces. We also show how to quantify random-dopant effects in this formulation. Additionally, we have developed an optimal multi-level Monte-Carlo method for the approximation of the solution. The method is optimal in the sense that the computational work is minimal for a given error tolerance.

1 Introduction

Stochastic partial differential equations make it possible to describe the propagation of random effects through PDE models. The motivation for considering the fully stochastic version of the drift-diffusion-Poisson system stems from the fact that device variability, noise, and fluctuations are becoming increasingly important as the characteristic lengths of semiconductor devices have been shrunk deep into the nanometer length scale. The model equations include classical devices such as FinFETs, but also more general simulation problems such as field-effect and nanopore sensors.

The investigation of such a system of stochastic PDE (see Sect. 2) poses new theoretical and numerical challenges. The numerical challenges are due to the large number of stochastic dimensions. We summarize a theoretic result in Sect. 3 and present numerical results in Sect. 4. The conclusions are given in Sect. 5.

L. Taghizadeh (✉) • A. Khodadadian • C. Heitzinger
Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstrasse 8–10, 1040 Vienna, Austria
e-mail: Leila.Taghizadeh@TUWien.ac.at; Amirreza.Khodadadian@TUWien.ac.at; Clemens.Heitzinger@TUWien.ac.at

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_48

2 The Stochastic Model Equations

Suppose that $(\Omega, \mathbb{A}, \mathbb{P})$ is a probability space and let $D \subset \mathbb{R}^d$, $d \leq 3$, be a bounded and convex domain. The σ-algebra $\mathbb{A}$ is assumed to be rich enough to support any random variable we encounter. The whole physical domain D is partitioned into three subdomains with different physical properties and hence different model equations in order to include a large range of applications. The first subdomain D_{Si} consists of the (silicon) nanowire and acts as the transducer of the sensor; in this subdomain, the drift-diffusion-Poisson equations describe charge transport. The semiconductor is surrounded by a dielectric layer (usually an oxide) which comprises the second subdomain D_{ox}, where just the Poisson equation holds. The third subdomain D_{liq} is the aqueous solution containing cations and anions and the Poisson-Boltzmann equation holds. The boundary layer at the sensor surface is responsible for the recognition of the target molecules. In the case of field-effect sensors, solving a homogenization problem at the surface, the manifold Γ, gives rise to two interface conditions (1e)–(1f) for the Poisson equation between the semiconductor and the liquid [6]. In summary, the domain is partitioned as

$$D = D_{\mathrm{Si}} \cup D_{\mathrm{ox}} \cup D_{\mathrm{liq}}.$$

The model equations $\forall \omega \in \Omega$ are the boundary-value problem

$$-\nabla \cdot (A(x,\omega)\nabla V(x,\omega)) \tag{1a}$$

$$= qC_{\mathrm{dop}}(x,\omega) - qn_i(\mathrm{e}^{V(x,\omega)/U_T}u(x,\omega) - \mathrm{e}^{-V(x,\omega)/U_T}v(x,\omega)) \quad \text{in } D_{\mathrm{Si}}, \tag{1b}$$

$$-\nabla \cdot (A(x,\omega)\nabla V(x,\omega)) = 0 \quad \text{in } D_{\mathrm{ox}}, \tag{1c}$$

$$-\nabla \cdot (A(x,\omega)\nabla V(x,\omega)) = -2\eta \sinh(\beta(V(x,\omega) - \Phi(x,\omega))) \text{in } D_{\mathrm{liq}}, \tag{1d}$$

$$V(0+,y,\omega) - V(0-,y,\omega) = \alpha(y,\omega) \quad \text{on } \Gamma, \tag{1e}$$

$$A(0+)\partial_x V(0+,y,\omega) - A(0-)\partial_x V(0-,y,\omega) = \gamma(y,\omega) \quad \text{on } \Gamma, \tag{1f}$$

$$U_T n_i \nabla \cdot (\mu_n \mathrm{e}^{V(x,\omega)/U_T}\nabla u(x,\omega)) \tag{1g}$$

$$= n_i \frac{u(x,\omega)v(x,\omega) - 1}{\tau_p(\mathrm{e}^{V(x,\omega)/U_T}u(x,\omega) + 1) + \tau_n(\mathrm{e}^{-V(x,\omega)/U_T}v(x,\omega) + 1)} \quad \text{in } D_{\mathrm{Si}}, \tag{1h}$$

$$U_T n_i \nabla \cdot (\mu_p \mathrm{e}^{-V(x,\omega)/U_T}\nabla v(x,\omega)) \tag{1i}$$

$$= n_i \frac{u(x,\omega)v(x,\omega) - 1}{\tau_p(\mathrm{e}^{V(x,\omega)/U_T}u(x,\omega) + 1) + \tau_n(\mathrm{e}^{-V(x,\omega)/U_T}v(x,\omega) + 1)} \quad \text{in } D_{\mathrm{Si}}, \tag{1j}$$

$$\alpha(y,\omega) = M_\alpha(V(y,\omega)) \quad \text{in } \Gamma, \tag{1k}$$

$$\gamma(y,\omega) = M_\gamma(V(y,\omega)) \quad \text{in } \Gamma, \tag{1l}$$

$$V(x,\omega) = V_D(x) \qquad \text{on } \partial D_D, \tag{1m}$$

$$\mathbf{n}\cdot\nabla V(x,\omega) = 0 \qquad \text{on } \partial D_N, \tag{1n}$$

$$u(x,\omega) = u_D(x), \quad v(x,\omega) = v_D(x) \qquad \text{on } \partial D_{D,\mathrm{Si}}, \tag{1o}$$

$$\mathbf{n}\cdot\nabla u(x,\omega) = 0, \quad \mathbf{n}\cdot\nabla v(x,\omega) = 0 \qquad \text{on } \partial D_{N,\mathrm{Si}}, \tag{1p}$$

where $A(x,\omega)$, the permittivity, is a random field with $x \in \mathbb{R}^d$, and $\omega \in \Omega$. Ω in the probability space $(\Omega, \mathbb{A}, \mathbb{P})$ denotes the set of elementary events, i.e., the sample space, $\mathbb{A}$ the σ-algebra of all possible events, and $\mathbb{P}\colon \mathbb{A} \to [0,1]$ is a probability measure. $V(x,\omega)$ is the electrostatic potential and $q > 0$ is the elementary charge, $C_{\mathrm{dop}}(x,\omega)$ is the doping concentration, D_n and D_p are the diffusion coefficients, μ_n and μ_p are the mobilities. The constant n_i is the intrinsic charge density, and τ_n and τ_p are the lifetimes of the free carriers. η is the ionic concentration, the constant β equals $\beta := q/(k_B T)$ in terms of the Boltzmann constant k_B and the temperature T, and Φ is the Fermi level. Here $\mathbf{n}$ denotes the unit outward normal vector on the boundary.

The functions α and γ are essentially given by the dipole-moment and the surface-charge densities of the boundary layer; in general, we write them as the functional $M_\alpha(V)$ and $M_\gamma(V)$ of the potential V for a fully self-consistent formulation. In the case of biosensors, they can be computed using the Metropolis Monte-Carlo method in the constant-voltage ensemble [2] for inclusion in a fully self-consistent loop [1, 7] or by solving the nonlinear Poisson-Boltzmann equation [5]. In the case of gas sensors, α and γ are given by systems of ordinary differential equations for surface reactions [3, 7].

We assume that the Einstein relations $D_n = U_T\mu_n$ and $D_p = U_T\mu_p$ hold, where the constant U_T is the thermal voltage. Therefore, it is beneficial to change variables from the electron and hole concentrations n and p in the usual form of the drift-diffusion-Poisson system to the Slotboom variables u and v defined by

$$n(x,\omega) =: n_i \mathrm{e}^{V(x,\omega)/U_T} u(x,\omega),$$
$$p(x,\omega) =: n_i \mathrm{e}^{-V(x,\omega)/U_T} v(x,\omega).$$

3 Analysis of the Stochastic Model Equations

In this section, we state a theorem on the existence and local uniqueness of solutions to the stochastic drift-diffusion-Poisson system whose proof will be published elsewhere [4]. The proof is based on a fixed-point argument for the existence, and the implicit-function theorem for the local uniqueness.

Theorem 1 (Existence and Local Uniqueness) *Suppose that assumptions on the regularity of the data and of the domain hold. Furthermore, assume that there exists a constant* $\mathbb{R} \ni K \geq 1$ *such that* $\frac{1}{K} \leq u_D(x), v_D(x) \leq K$ *holds for every* $x \in \partial D_{\mathrm{Si},D}$ *and that* $\underline{C} := \inf_{x\in D} C_{\mathrm{dop}}(x,\omega) \leq C_{\mathrm{dop}}(x,\omega) \leq \sup_{x\in D} C_{\mathrm{dop}}(x,\omega) =: \overline{C}$ *holds for every* $\omega \in \Omega$.

Then, for every right-hand side $f(x,\omega) \in L^2(\Omega; L^2(D))$ *and for all Dirichlet boundary conditions* $V_D, u_D, v_D \in H^{1/2}(\partial D)$, *there exists a locally unique weak solution*

$$(V(x,\omega), u(x,\omega), v(x,\omega), \alpha(x,\omega), \gamma(x,\omega)) \in \left(L^2(\Omega; H^1(D) \cap L^\infty(D\times\Omega)\right)$$
$$\times \left(L^2(\Omega; H^1(D_{\mathrm{Si}})) \cap L^\infty(D_{\mathrm{Si}}\times\Omega)\right)^2 \times \left(L^2(\Omega; H^1(\Gamma)) \cap L^\infty(\Gamma\times\Omega)\right)^2$$

of the stochastic boundary-value problem (1). *For every* $\omega \in \Omega$, *it satisfies the* L^∞*-estimate*

$$\begin{aligned}
\underline{V} \leq V(x,\omega) \leq \overline{V} \qquad & \text{in } D,\\
\frac{1}{K} \leq u(x,\omega) \leq K \qquad & \text{in } D_{\mathrm{Si}},\\
\frac{1}{K} \leq v(x,\omega) \leq K \qquad & \text{in } D_{\mathrm{Si}},
\end{aligned}$$

where

$$\underline{V} := \min\Big(\inf_{\partial D_D} V_D, \Phi, U_T \ln\big(\frac{1}{2Kn_i}(\underline{C} + \sqrt{\underline{C}^2 + 4n_i^2})\big)\Big),$$
$$\overline{V} := \max\Big(\sup_{\partial D_D} V_D, \Phi, U_T \ln\big(\frac{K}{2n_i}(\overline{C} + \sqrt{\overline{C}^2 + 4n_i^2})\big)\Big).$$

4 Numerical Results

Because of the slow convergence of the vanilla Monte-Carlo approach and the computational effort of calculating charge transport by solving a system of partial differential equations, the development of efficient numerical methods is of great importance.

The multi-level Monte-Carlo (MLMC) estimator is defined as

$$\hat{u}_{h_L} = \frac{1}{M_0}\sum_{i=1}^{M_0} u_{h_0}^{(i)} + \sum_{\ell=1}^{L}\frac{1}{M_\ell}\sum_{i=1}^{M_\ell}(u_{h_\ell}^{(i)} - u_{h_{\ell-1}}^{(i)}),$$

where u_{h_ℓ} and M_ℓ are finite-element solution and number of evaluations at level ℓ, respectively. h_ℓ denotes the mesh size of finite-element approximation at level ℓ for the mesh refinement, which is achieved by $h_\ell = r^{-\ell} h_0$, where h_0 denotes the mesh size of the coarsest triangulation and $r > 1$ is independent of ℓ. To develop an optimal MLMC finite-element method, we minimize the computational effort to satisfy a given error tolerance ε using the root mean-square error (RMSE).

The convergence order α is assumed for the discretization error

$$\|\mathbb{E}[u - u_{h_\ell}]\|_{H^1(D)} \leq C_1 h_\ell^\alpha \quad \exists C_1 \in \mathbb{R}^+, \tag{2}$$

the convergence order β is assumed for

$$\sigma[u_{h_\ell} - u_{h_{\ell-1}}] \leq C_0 h_{\ell-1}^\beta \quad \exists C_0 \in \mathbb{R}^+, \tag{3}$$

and it is assumed that the estimate $\sigma[u_{h_0}] \leq C_{00}$ holds, where $\sigma[u] := \|\mathbb{E}[u] - u\|_{L^2(\Omega;H^1(D))}$ is the standard deviation and $\mathbb{E}[\cdot]$ denotes the expected value.

Then it can be shown that the RMSE satisfies the inequality

$$\text{RMSE} \leq \frac{C_{00}}{\sqrt{M_0}} + C_0 \sum_{\ell=1}^{L} \frac{1}{\sqrt{M_\ell}} h_{\ell-1}^\beta + C_1 h_L^\alpha. \tag{4}$$

To obtain an optimal method, we solve the optimization problem

$$\begin{aligned} \underset{M_\ell, h_\ell}{\text{minimize}} \quad & \sum_{\ell=0}^{L} \sum_{k=1}^{4} \mu_k M_\ell h_\ell^{-\gamma_k} \\ \text{subject to} \quad & \frac{C_{00}}{\sqrt{M_0}} + C_0 \sum_{\ell=1}^{L} \frac{1}{\sqrt{M_\ell}} h_{\ell-1}^\beta + C_1 h_L^\alpha \leq \varepsilon \end{aligned} \tag{5}$$

and determine the optimal hierarchies $(L, \{M_\ell\}_{\ell=0}^L, \{h_\ell\}_{\ell=0}^L)$. The optimal number $L \in \mathbb{N}$ of levels is also unknown a priori. The model of the work includes—on each level—four different terms, which are the work estimates for the Poisson and drift-diffusion equations considering assembly of and solving the system matrix for each of the equations. Hence the terms $\mu_k M_\ell h_\ell^{-\gamma_k}$ denote the computational cost at level ℓ for the k-th part of solving the system.

We use the interior-point method to solve this nonlinear problem and optimize the hierarchies. Therefore, we find optimal parameters yielding an efficient numerical method for solving the stochastic model equations (1).

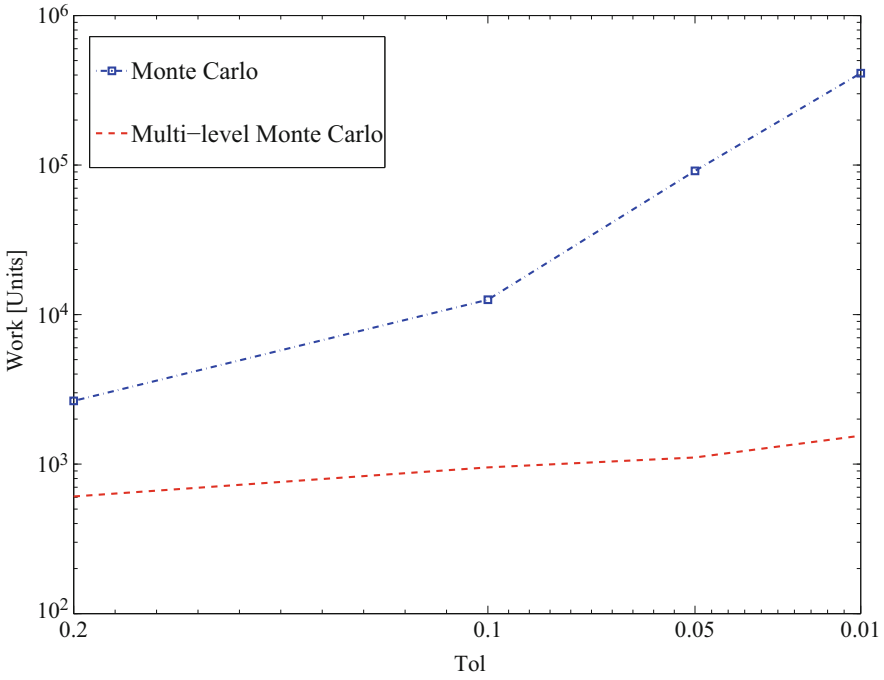


Fig. 1 Comparison of total computational work for MC and MLMC for various given error tolerances

Figure 1 shows the computational work for the optimal MLMC and the optimal MC methods (as the straightforward method). The optimal MLMC method is orders of magnitude faster.

Due to the various sources of device variability, noise, and fluctuations, all the coefficients in the model equation are in general random. Especially the random location of impurity atoms is a crucial source of stochasticity in all nanoscale transistors such as FinFETs. Random-dopant effects are the major limiting factor of transistor performance nowadays, and random dopants are included in the model equations. In previous work [1], we discussed how modifying device parameters such as ionic concentration, surface charge, and doping improves nanowire sensor response. In this work, we calculate the electrical current and discuss the effect of random-dopant fluctuations on the performance of nanoscale devices. Figure 2 shows the electrical current as a function of dopant concentration. This figure indicates that a higher density results in higher conductance and therefore in a higher current. Furthermore, Figs. 3 and 4 illustrate the effect of using different gate voltages for transducers, which are doped with phosphorus and arsenic atoms, on the electrical current.

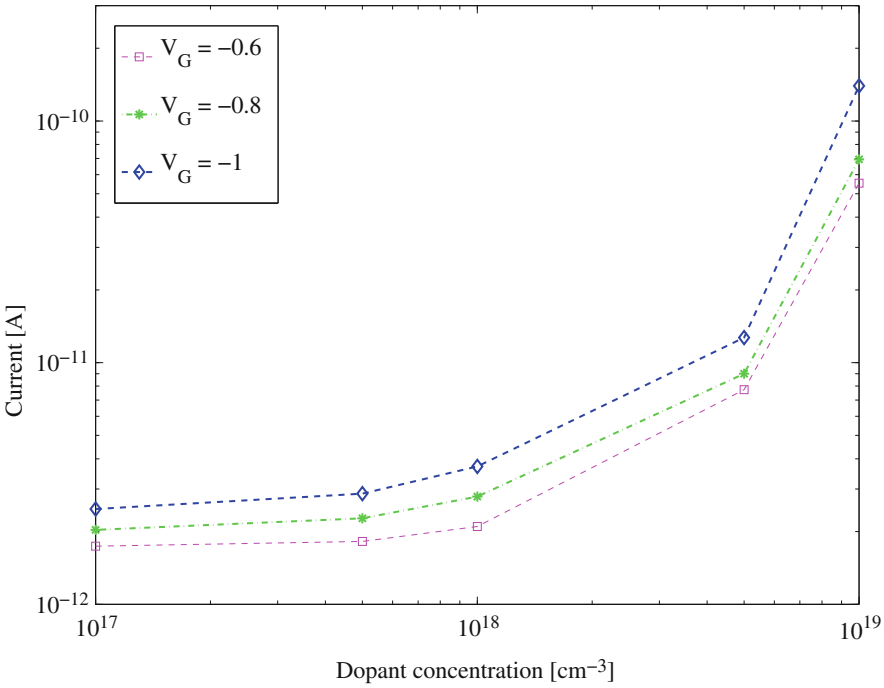


Fig. 2 Expected value of electrical current as a function of doping concentration for different gate voltages through a nanowire doped with boron

5 Conclusions

In this paper, we introduced the stochastic drift-diffusion-Poisson system to model various stochastic effects in nanoscale devices. The system of stochastic partial differential equations used here is general enough to describe many stochastic effects, as all coefficients are stochastic. The development of optimal numerical methods is ongoing work and addresses the challenge posed by the large number of stochastic dimensions in these types of problem. The number of dimensions can be in the hundreds or thousands, while the computational expense posed by one evaluation is the solution of a system of deterministic partial differential equations.

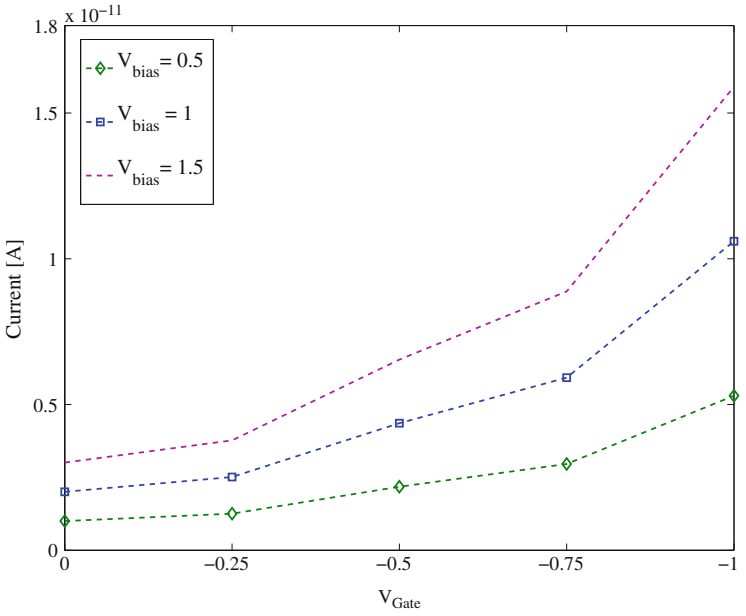


Fig. 3 The electrical current for different bias voltage versus gate voltage for phosphorus impurities for a nanowire with 100 nm diameter for $C_{\text{dop}} = 10^{18}\,\text{cm}^{-3}$

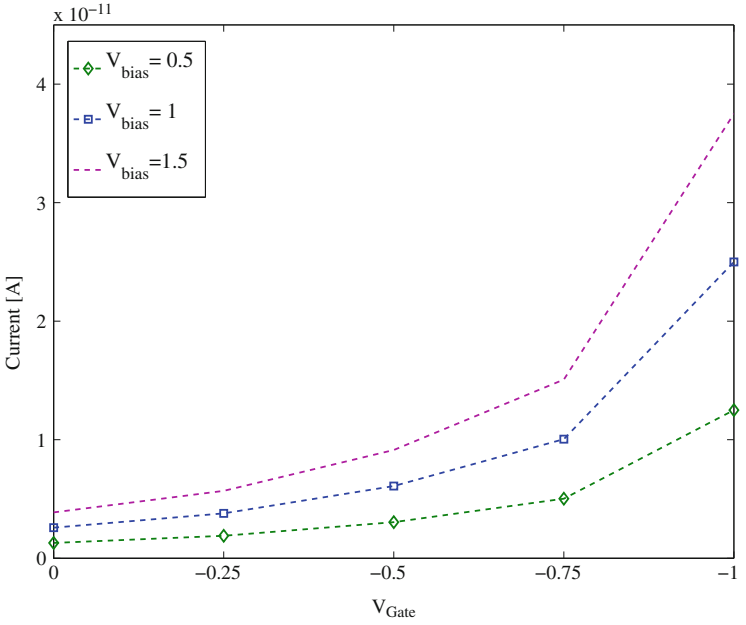


Fig. 4 The electrical current for different bias voltage versus gate voltage for arsenic impurities for a nanowire with 120 nm diameter for $C_{\text{dop}} = 10^{18}\,\text{cm}^{-3}$

Acknowledgements The authors acknowledge support by the FWF (Austrian Science Fund) START project no. Y660 *PDE Models for Nanotechnology*.

References

1. Baumgartner, S., Heitzinger, C., Vacic, A., Reed, M.A.: Predictive simulations and optimization of nanowire field-effect PSA sensors including screening. Nanotechnology **24**, 225503 (2013)
2. Bulyha, A., Heitzinger, C.: An algorithm for three-dimensional Monte-Carlo simulation of charge distribution at biofunctionalized surfaces. Nanoscale **3**, 1608–1617 (2011)
3. Fort, A., Rocchi, S., Serrano-Santos, M.B., Spinicci, R., Vignoli, V.: Surface state model for conductance responses during thermal-modulation of SnO-based thick film sensors: part I–model derivation. IEEE Trans. Instrum. Meas. **55**, 2102–2106 (2006)
4. Heitzinger, C., Taghizadeh, L.: Existence and local uniqueness for the stochastic drift-diffusion-Poisson system. Submitted for publication.
5. Heitzinger, C., Liu, Y., Mauser, N.J., Ringhofer, C., Dutton, R.W.: Calculation of fluctuations in boundary layers of nanowire field-effect biosensors. J. Comput. Theor. Nanosci. **7**, 2574–2580 (2010)
6. Heitzinger, C., Mauser, N.J., Ringhofer, C.: Multiscale modeling of planar and nanowire field-effect biosensors. SIAM J. Appl. Math. **70**, 1634–1654 (2010)
7. Tulzer, G., Baumgartner, S., Brunet, E., Mutinati, G.C., Steinhauer, S., Köck, A., Barbano, P.E., Heitzinger, C.: Kinetic parameter estimation and fluctuation analysis of CO at SnO_2 single nanowires. Nanotechnology **24**, 315501 (2013)

A Mathematical Proof in Nanocatalysis: Better Homogenized Results in the Diffusion of a Chemical Reactant Through Critically Small Reactive Particles

Jesús Ildefonso Díaz and David Gómez-Castro

Abstract We consider a reaction-diffusion in which the reaction takes place on the boundary of the reactive particles. In this sense the particles can be thought of as a catalysts that produce a change in the ambient concentration w_ε of a reactive element. It is known that depending on the size of the particles with respect to their periodic repetition there are different homogeneous behaviors. In particular, there is a case in which the kind of nonlinear reaction kinetics changes and becomes more smooth. This case can be linked with the strange behaviors that arise with the use of nanoparticles. In this paper we show that concentrations of a catalyst are always higher when nanoparticles are applied.

1 Introduction

We consider a reaction-diffusion problem in which the reaction takes place on the boundary of the inclusions. In this sense the inclusions can be though as a catalysts that produce as change in the ambient concentration w_ε of a reactive element. This is standardly modeled as

$$\begin{cases} -\Delta w_\varepsilon = 0 & \Omega_\varepsilon, \\ \partial_\nu w_\varepsilon + \varepsilon^{-\gamma} g(w_\varepsilon) = 0 & S_\varepsilon, \\ w_\varepsilon = 1 & \partial\Omega, \end{cases} \tag{1}$$

J.I. Díaz • D. Gómez-Castro (✉)
Dpto. de Matemática Aplicada and Instituto de Matemática Interdisciplinar, Universidad Complutense de Madrid, Plaza de las Ciencias 3, 28040 Madrid, Spain
e-mail: jidiaz@ucm.es; dgcastro@ucm.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_49

where g is a nondecreasing function such that $g(0) = 0$, Ω_ε is a perforated domain, $\partial\Omega_\varepsilon = S_\varepsilon \cup \partial\Omega$. R. Aris defined (see, e.g., [1]) the effectiveness of a reactor Ω as

$$\eta_\varepsilon = \frac{1}{|S_\varepsilon|}\int_{S_\varepsilon} g(w_\varepsilon)dx. \tag{2}$$

In the case where the particles are large (in a sense that would be precised later), the problem can be though homogenous, as $\Omega_\varepsilon \to \Omega$ then $w_\varepsilon \to w$ (in a sense that would be precised later), where the homogenized problem results

$$\begin{cases} -\Delta w + Ag(w) = 0 & \Omega, \\ w = 1 & \partial\Omega, \end{cases} \tag{3}$$

for a certain constant A. In this setting Aris defined the effectiveness for the homogenized problem as

$$\eta = \frac{1}{|\Omega|}\int_{\Omega} g(w)dx. \tag{4}$$

This kind of problems, when g is not Lipschitz, has been shown to develop, in some cases, a region of positive measure $\{x \in \Omega : u(x) = 0\}$. This region, which is sometimes known as a dead core, has been studied in [2, 5].

Nonetheless, when the holes are of a sufficiently small size with respect to their repetition, the behaviour of the limit changes and becomes

$$\begin{cases} -\Delta w + Bh(w) = 0 & \Omega, \\ w = 1 & \partial\Omega, \end{cases} \tag{5}$$

and h is a new nonlinearity, which we will introduce later, and $B > 0$ is a constant.

This change in behaviour, which is related to the pioneering paper [3], will be linked to new surprising properties that arise with the use of nanoparticles (see [11]). In this setting, the correct definition for the effectiveness of the limit problem is unclear.

The aim of this paper is to show that homogenized problem is more effective in the case associated with nanoparticles than the other cases. It represents a mathematical proof of some experimental facts in the literature.

2 Statement of Results

Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 2$, with a smooth boundary $\partial\Omega$ and let $Y = (-1/2, 1/2)^n$. Denote by $G_0 = B_1(0)$ the unit ball centered at the origin. For $\delta > 0$ and $\varepsilon > 0$ we define $\delta B = \{x \,|\, \delta^{-1}x \in B\}$ and $\widetilde{\Omega}_\varepsilon = \{x \in \Omega \,|\, \rho(x, \partial\Omega) > 2\varepsilon\}$. Let

$$a_\varepsilon = C_0\varepsilon^\alpha, \tag{6}$$

where $\alpha > 1$ and C_0 is a given positive number. Define

$$G_\varepsilon = \bigcup_{j\in\Upsilon_\varepsilon}(a_\varepsilon G_0 + \varepsilon j) = \bigcup_{j\in\Upsilon_\varepsilon} G_\varepsilon^j,$$

where $\Upsilon_\varepsilon = \{j \in \mathbb{Z}^n : (a_\varepsilon G_0 + \varepsilon j) \cap \overline{\widetilde{\Omega}}_\varepsilon \neq \emptyset\}$, $|\Upsilon_\varepsilon| \cong d\varepsilon^{-n}$, $d = const > 0$, $\mathbb{Z}^n$ is the set of vectors z with integer coordinates. The reference cell is represented by Fig. 1.

Define $Y_\varepsilon^j = \varepsilon Y + \varepsilon j$, where $j \in \Upsilon_\varepsilon$ and note that $\overline{G}_\varepsilon^j \subset \overline{Y}_\varepsilon^j$ and center of the ball G_ε^j coincides with the center of the cube Y_ε^j. Our "microscopic domain" is defined as

$$\Omega_\varepsilon = \Omega \setminus \overline{G_\varepsilon}, \qquad S_\varepsilon = \partial G_\varepsilon, \qquad \partial\Omega_\varepsilon = \partial\Omega \cup S_\varepsilon,$$

which can be represented as in Fig. 2.

We define the space $W^{1,p}(\Omega_\varepsilon, \partial\Omega)$ be the completion, with respect to the norm of $W^{1,p}(\Omega_\varepsilon)$, of the set of infinitely differentiable functions in $\overline{\Omega}_\varepsilon$ equal to zero in a neighborhood of $\partial\Omega$.

We are interest in understating the comparison of the limits of (1) when $\alpha \in (1, \frac{n}{n-2})$ and $\alpha = \frac{n}{n-2}$, which are known as the subcritical and critical cases in homogenization. The case $\alpha = 1$ was studied in [4]. In order to do this, we consider

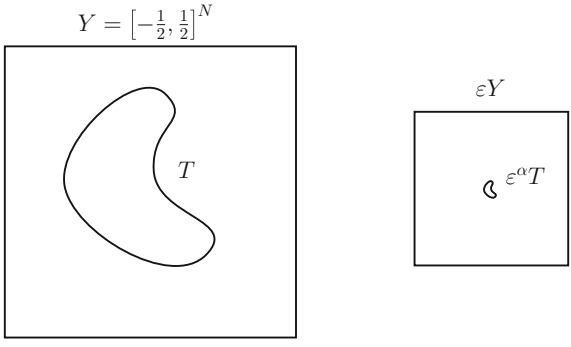


Fig. 1 The reference cell Y and the scalings by ε and ε^α, for $\alpha > 1$. Notice that, for $\alpha > 1$, $\varepsilon^\alpha T$ (for a general particle shaped as T) becomes smaller relative to εY, which scales as the repetition. In our case T will be a ball $B_1(0)$

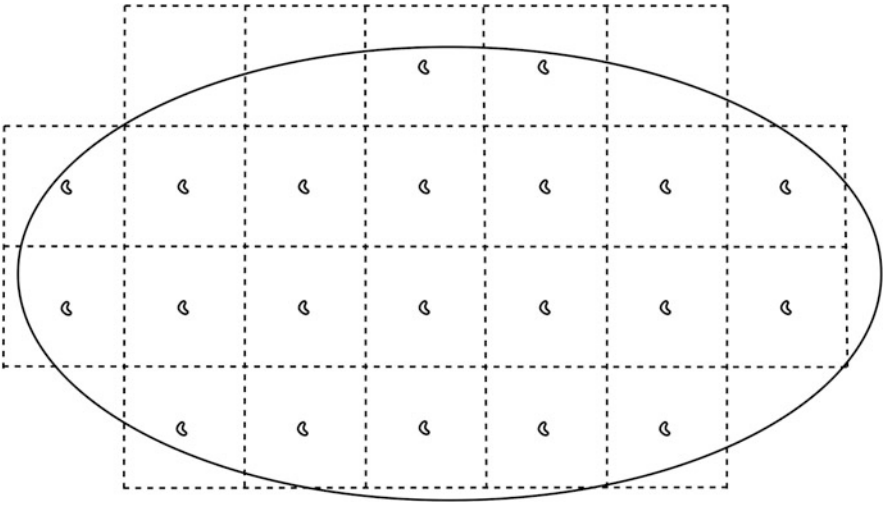


Fig. 2 The fixed bed reactor, i.e., the domain Ω_ε

the change in variable $u = 1 - w$, $\sigma(u) = g(1) - g(1-u)$ we have

$$\begin{cases} -\Delta u_\varepsilon = 0 & \Omega_\varepsilon, \\ \partial_\nu u_\varepsilon + \varepsilon^{-\gamma}\sigma(u_\varepsilon) = \varepsilon^{-\gamma} g(1) & S_\varepsilon, \\ u_\varepsilon = 0 & \partial\Omega. \end{cases} \tag{7}$$

The direct study of the family of solutions $(u_\varepsilon)_{\varepsilon>0}$ is difficult, since they are not defined in the same domain. We consider a family of linear extension operators (see [10])

$$P_\varepsilon : \{u \in H^1(\Omega_\varepsilon) : u = 0, \partial\Omega\} \to H_0^1(\Omega) \tag{8}$$

such that

$$\|\nabla P_\varepsilon u\|_{L^2(\Omega)} \leq \|\nabla u\|_{L^2(\Omega_\varepsilon)}. \tag{9}$$

We define the different possible limits u:

- If $\alpha \in \left(1, \frac{n}{n-2}\right)$ then $u = u_{\text{non-crit}}$, which for $A = C_0^{n-1}\omega_n$ satisfies

$$\begin{cases} -\Delta u_{\text{non-crit}} + A\sigma(u_{\text{non-crit}}) = Ag(1) & \Omega, \\ u_{\text{non-crit}} = 0 & \partial\Omega. \end{cases} \tag{10}$$

- If $\alpha = \frac{n}{n-2}$ then $u = u_{\text{crit}}$, which for $B = (n-2)C_0^{n-2}\omega_n = \frac{n-2}{C_0}A$ satisfies

$$\begin{cases} -\Delta u_{\text{crit}} + BH(u_{\text{crit}}) = 0 & \Omega, \\ u_{\text{crit}} = 0 & \partial\Omega, \end{cases} \tag{11}$$

  where H is the solution of the functional equation

$$\frac{n-2}{C_0}H(s) = \sigma(s - H(s)) - g(1). \tag{12}$$

We will start by indicating that, in the sense of maximal monotone graphs, in the particular case of $\sigma(u) = g(1) - g(1-u)$ one has

Lemma 1 *Let σ be a maximal monotone graph, then the solution H of (12) is given by*

$$H(u) = -\left(g^{-1}\left(\frac{n-2}{C_0}\cdot\right) + Id\right)^{-1}(1-u). \tag{13}$$

Hence $H(u) \leq 0$ for every $u \in [0,1]$.

Remark 1 Notice that, in particular, in Eq. (5) we have

$$h(w) = \left(g^{-1}\left(\frac{n-2}{C_0}\cdot\right) + Id\right)^{-1}(w) \tag{14}$$

which is a nondecreasing function such that $h(0) = 0$.

Lemma 2 *Let σ be a bounded maximal monotone graph of $[0,1] \times \mathbb{R}$, then H is non-expansive in $[0,1]$ (and hence Lipschitz continuous).*

Proof If $\sigma \in \mathscr{C}^1([0,1])$, differentiating (12) with respect to s we derive

$$H'(s) = \frac{\sigma'(s - H(s))}{\frac{n-2}{C_0} + \sigma'(s - H(s))} \in (0,1). \tag{15}$$

Hence,

$$|H(t) - H(s)| \leq |t - s| \tag{16}$$

for all $t, s \in [0,1]$. If σ is a maximal monotone graph, let $\sigma_\delta \in \mathscr{C}^1([0,1])$ be an approximation in the sense of maximal monotone graphs $\sigma_\delta \to \sigma$. In particular, $H_\delta \to H$ pointwise, and hence

$$|H(t) - H(s)| \leq |t - s| \tag{17}$$

which concludes the proof. □

We have the following homogenization result.

Theorem 1 ([12]) *Let $\alpha > 1$, $\gamma = \alpha(n-1) - n$, $\sigma \in \mathscr{C}^1(\mathbb{R})$ be such that $\sigma(0) = 0$,*

$$0 < k_1 \leq \sigma'(s) \leq k_2, \tag{18}$$

and let u_ε be the weak solution of (7). Then, the extension $P_\varepsilon u_\varepsilon$ converge as $\varepsilon \to 0$

$$P_\varepsilon u_\varepsilon \to \begin{cases} u_{non\text{-}crit} & \text{if } \alpha \in \left(1, \frac{n}{n-2}\right), \\ u_{crit} & \text{if } \alpha = \frac{n}{n-2}, \end{cases} \tag{19}$$

strongly in $W_0^{1,p}(\Omega)$ for $1 \leq p < 2$ and weakly in $H_0^1(\Omega)$.

Since, in our case $0 \leq u_\varepsilon \leq 1$ then we can have a simple corollary:

Corollary 1 *Let $\sigma \in \mathscr{C}([0,1])$, nondecreasing and such that $\sigma(0) = 0$, then (19) holds weakly in $H_0^1(\Omega)$.*

Proof Applying the estimates in [9] we check that $(P_\varepsilon u_\varepsilon)$ is bounded in $H_0^1(\Omega)$, hence there exists a limit $\hat{u}$ such that, up to a subsequence, $P_\varepsilon u_\varepsilon \to \hat{u}$ strongly in L^2.

Let σ_δ be such that it satisfies Theorem 1 and $\sigma_\delta \to \sigma$ in $\mathscr{C}([0,1])$ as $\delta \to 0$. Let $u_{\varepsilon,\delta}$ the solution of (7) with σ_δ. We can check, again with estimates in [9] that

$$\|u_\varepsilon - u_{\varepsilon,\delta}\|_{L^2(\Omega^\varepsilon)} \leq C\|\sigma - \sigma_\delta\|_{\mathscr{C}([0,1])}. \tag{20}$$

Passing to the limit as $\varepsilon \to 0$

$$\|\hat{u} - u_\delta\|_{L^2(\Omega)} \leq C\|\sigma - \sigma_\delta\|_{\mathscr{C}([0,1])}. \tag{21}$$

On the other hand, applying the theory of maximal monotones graphs, it is easy to check that H_δ the solution of (12) with σ_δ satisfies $H_\delta \to H$ in the sense of maximal monotone graphs. In both cases, $u_\delta \to u$ in L^2 where u is the solution of the problem with H or σ. Therefore $u_\varepsilon \to u$ weakly as $\varepsilon \to 0$. □

It is known already (see [7]) that, for the non critical cases the effectiveness behaves as expected

$$\eta_\varepsilon \to \eta, \qquad \text{as } \varepsilon \to 0. \tag{22}$$

However, in the noncritical case the reaction kinetics changes. Therefore it is not clear whether it is natural to define the effectiveness in the usual way. Nonetheless, we can give a pointwise comparison inequality. Let u_{crit} and $u_{\text{non-crit}}$ be the solutions of (11) and (10):

Theorem 2 *Let $\sigma \in \mathscr{C}([0,1])$ be such that $\sigma(0) = 0$. Then*

$$u_{crit} \leq u_{non\text{-}crit}. \tag{23}$$

Proof Since $H(s) \leq 0$ we have that

$$BH(s) = (n-2)C_0^{n-2}\omega_n H(s) = C_0^{n-1}\omega_n \frac{n-2}{C_0} H(u) \tag{24}$$

$$= C_0^{n-1}\omega_n \left(\sigma(s - H(s)) - g(1)\right) = A\left(\sigma(s - H(s)) - g(1)\right) \tag{25}$$

$$\geq A\left(\sigma(s) - g(1)\right). \tag{26}$$

Therefore, applying the comparison principle, $u_{\text{crit}} \leq u_{\text{non-crit}}$. □

Remark 2 Therefore, if $g \in \mathscr{C}([0,1])$, the concentration w in the critical case is always larger than in the non critical cases

$$w_{\text{crit}} \geq w_{\text{non-crit}}. \tag{27}$$

Hence, we get a pointwise better reaction.

Remark 3 The basic convergence result given in Theorem 1 has been proved in many different cases. In particular, for non smooth σ in the form of a root or a Heaviside function and nonlinear diffusion in the form of a p-Laplacian see [6]. The case of Signorini type boundary conditions can be found in [8].

Acknowledgements The research of D. Gómez-Castro is supported by a FPU fellowship from the Spanish government. The research of J.I. Díaz and D. Gómez-Castro was partially supported by the project ref. MTM 2014-57113-P of the DGISPI (Spain).

References

1. Aris, R., Strieder, W.: Variational Methods Applied to Problems of Diffusion and Reaction. Springer Tracts in Natural Philosophy, vol. 24. Springer, New York (1973)
2. Bandle, C., Sperb, R., Stakgold, I.: Diffusion and reaction with monotone kinetics. Nonlinear Anal. Theory Methods Appl. **8**(4), 321–333 (1984)
3. Cioranescu, D., Murat, F.: A strange term coming from nowhere. In: Cherkaev, A., Kohn, R. (eds.) Topics in Mathematical Modelling of Composite Materials, pp. 45–94. Springer Science+Business Media, LLC, New York (1997)
4. Conca, C., Díaz, J.I., Liñán, A., Timofte, C.: Homogenization in chemical reactive flows. Electron. J. Differ. Equ. **40**, 1–22 (2004)
5. Díaz, J.I.: Nonlinear Partial Differential Equations and Free Boundaries. Pitman, London (1985)
6. Díaz, J.I., Gómez-Castro, D., Podol'skii, A.V., Shaposhnikova, T.A.: Homogenization of the p-Laplace operator with nonlinear boundary condition on critical size particles: identifying the strange terms for some non smooth and multivalued operators. Dokl. Math. **94**(1), 387–392 (2016)
7. Díaz, J.I., Gómez-Castro, D., Timofte, C.: The effectiveness factor of reaction-diffusion equations: homogenization and existence of optimal pellet shapes. J. Elliptic Parabolic Equ. **2**, 117–127 (2016)

8. Díaz, J.I., Gómez-Castro, D., Podol'skii, A.V., Shaposhnikova, T.A.: Homogenization of variational inequalities of Signorini type for the p-Laplacian in perforated domains when $p \in (1, 2)$. Dokl. Math. **95**, 151–156 (2017)
9. Podol'skii, A.V.: Homogenization limit for the boundary value problem with the p-Laplace operator and a nonlinear third boundary condition on the boundary of the holes in a perforated domain. Dokl. Math. **82**(3), 942–945 (2010)
10. Podol'skii, A.V.: Solution continuation and homogenization of a boundary value problem for the p-Laplacian in a perforated domain with a nonlinear third boundary condition on the boundary of holes. Dokl. Math. **91**(1), 30–34 (2015)
11. Schimpf, S., Lucas, M., Mohr, C., Rodemerck, U., Brückner, A., Radnik, J., Hofmeister, H., Claus, P.: Supported gold nanoparticles: in-depth catalyst characterization and application in hydrogenation and oxidation reactions. Catal. Today **72**(1), 63–78 (2002)
12. Zubova, M.N., Shaposhnikova, T.A.: Homogenization of boundary value problems in perforated domains with the third boundary condition and the resulting change in the character of the nonlinearity in the problem. Differ. Equ. **47**(1), 78–90 (2011)

The Effect of Depth-Dependent Velocity on the Performance of a Nanofluid-Based Direct Absorption Solar Collector

Gary J. O'Keeffe, Sarah L. Mitchell, Tim G. Myers, and Vincent Cregan

Abstract In this paper we present a two-dimensional model for the efficiency of an inclined nanofluid-based direct absorption solar collector. The model consists of a system of two differential equations; a radiative transport equation describing the propagation of solar radiation through the nanofluid and an energy equation. The Navier-Stokes equations are solved by applying no-slip boundary conditions to give a depth-dependent velocity profile. The heat source term is approximated via a power law approximation. The energy equation is solved numerically and the resulting solution is then used to calculate the efficiency. We investigate the effect of depth-dependent velocity and show how it affects the temperature, and thus the efficiency.

1 Introduction

The solar energy industry has experienced phenomenal growth in recent years due to both technological improvements resulting in cost reductions and government policies supportive of renewable energy development and utilisation [6]. Nanofluid-based direct absorption solar collectors (NDASCs) use a colloidal solution of nanoparticles and a base fluid to absorb incident sunlight. This method improves on collector efficiency compared to traditional collectors [2, 3, 5].

We present a model to predict the effect of a NDASC. The model consists of a system of partial differential equations describing the conservation of mass, momentum and energy. A heat source term is obtained via the radial flux integral, which is approximated via the method described by Cregan and Myers [2]. The resulting solution is then used to investigate the efficiency of the collector subject to variation in model parameters. Cregan and Myers approxi-

G.J. O'Keeffe (✉) • S.L. Mitchell
Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
e-mail: gary.okeeffe@ul.ie; sarah.mitchell@ul.ie

T.G. Myers • V. Cregan
Centre de Recerca Matemática, Campus de Bellaterra, Barcelona, Spain
e-mail: tmyers@crm.cat; vcregan@crm.cat

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_50

mate velocity in the collector using the average fluid velocity [2]. We instead propose a depth-dependent expression for the velocity, which leads to a model for the nanofluid temperature, and thus a more realistic description of collector efficiency.

First we first describe the various components of the model such as the optical and physical properties of nanofluid, the heat source term, conservation of mass and energy, and the heat loss due to Newton cooling at the upper panel of the collector. We then simplify the model via dimensional analysis and present a numerical solution to the reduced model. Finally, the model results are discussed.

2 Model

We model the solar collector as a two-dimensional channel on an inclined plane. As shown in Fig. 1, x^* is the coordinate in the downstream direction and z^* is the perpendicular distance downwards from the upper surface (* denotes a dimensional variable). Note that $H \ll L$, and the nanofluid flow is laminar and in the x^* direction. The collector makes an angle of θ to the horizontal. We assume that the underside of the collector ($z^* = H$) is completely insulated and there is heat loss due to Newton cooling at the upper panel, whilst the temperature of the surrounding air, T^*_{Amb}, remains constant. The nanofluid enters the collector

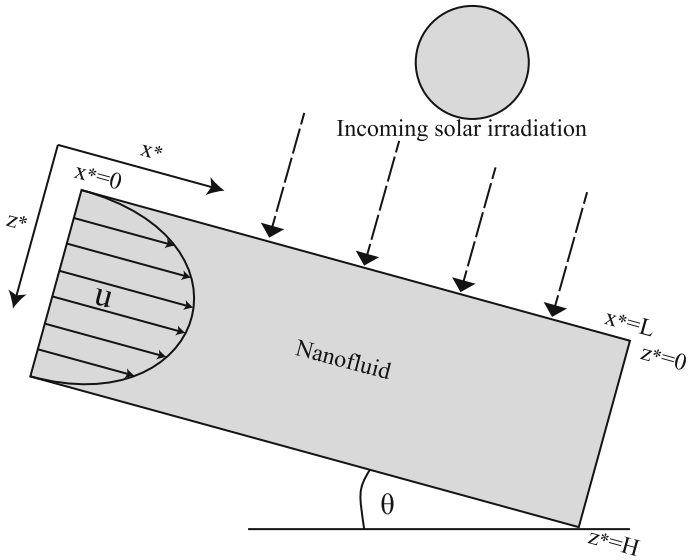


Fig. 1 Schematic of nanofluid based solar collector as modelled in this paper

with initial temperature T_I^*, and absorbs solar radiation volumetrically. The change in normally incident solar spectral flux due to attenuation of the nanofluid is dissipated as heat release and described by the one-dimensional radiative transport equation

$$\frac{\partial J_\lambda^*}{\partial z^*} = -K_e J_\lambda^*, \tag{1}$$

where $J_\lambda^*(z^*, \lambda^*)$ is the spectral intensity and λ^* is the wavelength. The nanofluid extinction coefficient accounts for light attenuation in the nanofluid, and is defined to be

$$K_e = K_a^{bf} + K_s^{bf} + K_a^{np} + K_s^{np}, \tag{2}$$

where K_a and K_s are the absorption and scattering coefficients respectively, and are calculated using the same approach used by Cregan and Myers [2]. The subscripts *bf*, *nf* and *np* denote base fluid, nanofluid and nanoparticle respectively. Integrating (1) and applying the boundary condition $J(z = 0, \lambda) = J_0$ yields

$$J_\lambda^* = (1 - R) J_0 e^{-K_e z^*}, \tag{3}$$

where R is the approximate fraction of reflected solar radiation, and J_0 is the concentrated normal incident solar radiation at the upper surface of the collector, approximated by Planck's black body relation (see (2.11) in [2]). Upon integrating (3) over the entire spectrum we obtain the radiative flux

$$P^*(z^*) = \int_0^\infty (1 - R) J_0 e^{-K_e z^*} \, \mathrm{d}\lambda. \tag{4}$$

This expression is highly non-linear in wavelength, and so must be integrated numerically. We use the approximation for (4) first proposed by Cregan and Myers [2], i.e.,

$$\int_0^\infty J_0 e^{-K_e z^*} \, \mathrm{d}\lambda \approx \frac{(1 - R) G_s}{\left(1 + \frac{\beta_0}{2R} z^*\right)^{\beta_1}}, \tag{5}$$

β_0 and β_1 are fitting parameters and G_s is the solar intensity at the top of the collector.

Conservation of momentum is expressed using the steady state Navier-Stokes equations for laminar, incompressible fluid flow down an inclined plane under the action of gravity

$$(\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{1}{\rho_{nf}} \nabla p^* - \frac{\mu_{nf}}{\rho_{nf}} \nabla^2 \mathbf{u} = \mathbf{g}, \tag{6}$$

where μ_{nf} is the viscosity of the nanofluid, $\mathbf{u} = (u^*, 0)$, u^* is acting in the x^* direction, p^* the pressure, and $\mathbf{g}$ is gravity. The nanofluid density is calculated using classical mixing theory: $\rho_{nf} = \phi\rho_{np} + (1-\phi)\rho_{bf}$, where ϕ is the particle volume fraction of nanoparticles in the nanofluid. Via the lubrication approximation for a thin film of liquid [1], (6) reduces to

$$-\frac{\partial p^*}{\partial x^*} + \mu_{nf}\frac{\partial^2 u^*}{\partial z^{*2}} + \rho_{nf} g \sin\theta = 0, \quad -\frac{\partial p^*}{\partial z^*} + \rho_{nf} g \cos\theta = 0. \tag{7}$$

Integrating the latter of these equations and applying $p^*(0) = p_a^*$ (where p_a is the ambient pressure), leads to $p^* = p_a^* + \rho_{nf} g z^* \cos\theta$. Integrating the first equation in (7) and upon applying the no-slip boundary conditions at the upper and lower surfaces of the collector yields

$$u^* = \frac{\rho_{nf} g H^2 \sin\theta}{2\mu_{nf}}\left(\frac{z^*}{H} - \frac{z^{*2}}{H^2}\right). \tag{8}$$

The steady state heat equation is

$$\rho_{nf} c_{p,nf}\mathbf{u}\cdot\nabla T^* = k_{nf}\nabla^2 T^* + q, \tag{9}$$

where $T^*(x^*, z^*)$ is the nanofluid temperature, $c_{p,nf}$ is the specific heat capacity and is calculated using classical mixing theory:

$$c_{p,nf} = \frac{\phi\rho_{np}c_{p,np} + (1-\phi)\rho_{bf}c_{p,bf}}{\rho_{nf}}, \tag{10}$$

and the nanofluid thermal conductivity is given by

$$k_{nf} = \frac{k_{bf}}{\left(1-\phi^{1/3}\right)^2}\left[(1-\phi) + \phi\frac{\rho_{np}c_{p,np}}{\rho_{bf}c_{p,bf}}\right]\frac{n-1}{2(n+1)}\left[\frac{1+\phi^{1/3}}{2} - \frac{1}{n+1}\right]^{-1}, \tag{11}$$

where $n = 2.233$ [4]. The heat source term, q^*, is the derivative of the radiative flux:

$$q^* = -\frac{\mathrm{d}P^*}{\mathrm{d}z^*} = \frac{(1-R)G_S\beta_0\beta_1}{\left(1+\beta_0\frac{z^*}{H}\right)^{\beta_1+1}}. \tag{12}$$

Introducing the velocity from (8) into (9) leads to

$$\frac{c_{p,nf}\rho_{nf}^2 g\sin\theta}{2\mu_{nf}}(Hz^* - z^{*2})\frac{\partial T^*}{\partial x^*} = k_{nf}\frac{\partial^2 T^*}{\partial z^{*2}} + \frac{(1-R)G_S\beta_0\beta_1}{\left(1+\beta_0\frac{z^*}{H}\right)^{\beta_1+1}}, \tag{13}$$

with boundary conditions,

$$\left.\frac{\partial T^*}{\partial z^*}\right|_{z^*=H} = 0, \qquad k_{nf}\left.\frac{\partial T^*}{\partial z^*}\right|_{z^*=0} = h_s\left(\left.T^*\right|_{z^*=0} - T^*_{Amb}\right), \tag{14}$$

where h_s is the surface heat transfer coefficient, and the inlet condition is $T^*(0, z^*) = T^*_I$. The model is non-dimensionalised via $x^* = Lx$, $z^* = Hz$, $T^* = T\Delta T + T^*_I$. The temperature variation in the nanofluid is driven by the source term, and hence the temperature scale is chosen to be

$$\Delta T = \frac{(1-R)G_S L\beta_0\beta_1}{HUc_{p,nf}\rho_{nf}}, \tag{15}$$

where $U = \rho_{nf} g_{nf} H^2 \sin\theta/2\mu_{nf}$. We now rewrite (13) in dimensionless form as

$$(z - z^2)\frac{\partial T}{\partial x} = \gamma\frac{\partial^2 T}{\partial z^2} + \frac{1}{(1+\beta_0 z)^{\beta_1 - 1}}, \tag{16}$$

where $\gamma = k_{nf}L/c_{p,nf}\rho_{nf}UH^2$. The dimensionless boundary and initial conditions are

$$\left.\frac{\partial T}{\partial z}\right|_{z=1} = 0, \qquad \left.\frac{\partial T}{\partial z}\right|_{z=0} = \mathrm{Nu}\left(T|_{z=0} - \hat{T}\right), \qquad T(0, z) = 0, \tag{17}$$

where $\mathrm{Nu} = (Hh_s)/k_{nf}$ is the ratio of convective to conductive heat transfer at the upper panel.

Collector performance, η, is the ratio of usable thermal energy to incident solar energy and we modify the definition for η from [5] and [7]. Allowing for depth-dependent fluid velocity in a two dimensional channel, we modify the efficiency expression in [2] to obtain

$$\eta = \frac{\rho_{nf}c_{p,nf}}{G_S L}\int_0^H (T^*_O(z^*) - T^*_I)u(z^*)\,\mathrm{d}z^*, \tag{18}$$

where T^*_O is the temperature at the outlet. In dimensionless form we can rewrite (18) as

$$\eta = \beta_0\beta_1\int_0^1 T_O(z)(z - z^2)\,\mathrm{d}z. \tag{19}$$

3 Results

In this section we compare our numerical simulation to the corresponding results of Cregan and Myers [2]. In particular, we solved (16) via the MATLAB routine 'Pdepe' for the nanofluid temperature, and hence the efficiency using

Table 1 Solar collector and nanofluid parameters and physical constants from Cregan and Myers [2]

Quantity	Symbol	Value	Units
Collector length/width/height	$L/W/H$	$1, 1, 0.0012$	m
Collector angle of inclination	θ	$\pi/9$	–
Viscosity (H_2O)	μ_{bf}	10^{-3}	$\mathrm{kg\,m^{-1}\,s^{-1}}$
Surface heat transfer coefficient	h_s	6.43	$\mathrm{W\,m^{-2}\,K^{-1}}$
Air/inlet nanofluid temperature	T^*_{Amb}/T^*_I	298.15/308.15	K
Incident solar radiation	G_S	1000	$\mathrm{W\,m^{-2}}$
Gravity	g	9.8	$\mathrm{m\,s^{-2}}$
Density water/aluminium	ρ_{bf}/ρ_{np}	1000/2700	$\mathrm{kg\,m^{-3}}$
Conductivity water/aluminium	k_{bf}/k_{np}	0.609/247	$\mathrm{W\,m^{-1}\,K^{-1}}$
Specific heat capacity water/aluminium	$c_{p,bf}/c_{p,np}$	4187/900	$\mathrm{J\,kg^{-1}\,K^{-1}}$
Fitting parameters (for $\phi = 0.006$)	β_0/β_1	10.60/1.14	–

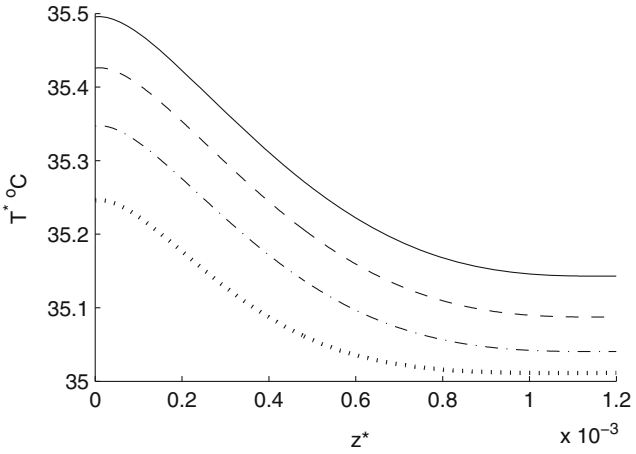


Fig. 2 Temperature profiles for a water-aluminium nanofluid at $x = 0.25L$ (*dotted line*), $x = 0.5L$ (*dash-dotted line*), $x = 0.75L$ (*dashed line*) and $x = L$ (*solid line*) for $\phi = .006$, $R = 0.35$

(19). Table 1, taken from [2], lists the parameter values for a water-aluminium nanofluid.

Figure 2 shows the temperature profile as a function of depth at various points along the channel. The influence of solar radiation is greatest near $z^* = 0$, leading to the highest temperatures near the upper panel. Further, the temperature increases as the nanofluid flows through the channel.

Figure 3a shows that the inclusion of depth-dependent velocity can have a noticeable effect on the efficiency of a NDASC under certain conditions. As collector height increases, our model performs different to the model proposed by Cregan and Myers [2]. The depth-velocity expressed in (8) will be smaller close to $z^* = 0$ than the constant velocity from [2] and so heat is dispersed at a slower

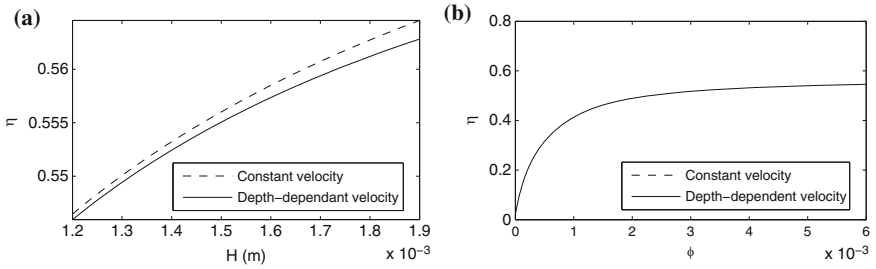


Fig. 3 Collector efficiency versus collector height and nanofluid particle volume fraction, $R = 0.35$. (**a**) Efficiency versus collector height, $\phi = 0.006$. (**b**) Efficiency versus particle volume fraction

rate, resulting in a lower efficiency. Both models indicate a similar initial increase in efficiency with H, but as H gets larger our model sees a slower increase in efficiency. There are also circumstances where the difference between our model and Cregan and Myers' model [2] is minimal. In Fig. 3b both models exhibit the same behaviour where a large initial increase in efficiency is followed by a more gradual increase with the addition of more nanoparticles.

4 Conclusion

In this paper we developed a two-dimensional model for the efficiency of an inclined nanofluid-based direct absorption solar collector. Solving the conservation of momentum equations yielded a depth-dependent velocity solution. This solution for the velocity was included in the energy equation which lead to a more realistic model than that proposed by Cregan and Myers [2]. The heat source term was approximated via a power law approximation. The energy equation was solved numerically and the resulting solution was then used to calculate the efficiency.

The inclusion of depth-dependent velocity had a noticeable effect on NDASC efficiency under certain conditions. Heat is dispersed at a slower rate in the solar collector when flow is depth-dependent. As collector height increases, our model predicts a less efficient collector than the model proposed by Cregan and Myers [2].

Acknowledgements Authors would like to acknowledge the funding provided by the Irish Research Council (GOIPG/2014/887), and the Mathematics for Industry Network COST Action TD1409 (COST-STSM-TD1409-30261), which made this research possible.

References

1. Acheson, D.J.: Elementary Fluid Dynamics. Oxford University Press, Oxford (1990)
2. Cregan, V., Myers, T.G.: Modelling the efficiency of a nanofluid direct absorption solar collector. Int. J. Heat Mass Transf. **90**, 505–514 (2015)
3. Lenert, A., Wang, E.N.: Optimization of nanofluid volumetric receivers for solar thermal energy conversion. Sol. Energy **86**, 253–265 (2012)
4. Myers, T.G., MacDevette, M.M., Ribera, H.: A time-dependent model to determine the thermal conductivity of a nanofluid. J. Nanopart. Res. **15**, 1–11 (2013)
5. Otanicar, T.P., Phelan, P.E., Prasher, R.S., Rosengarten, G., Taylor, R.A.: Nanofluid-based direct absorption solar collector. J. Renew. Sust. Energy **2**, 033102 (2010)
6. Timilsina, G.R., Kurdgelashvili, L., Narbel, P.A.: Solar energy: markets, economics and policies. Renew. Sust. Energy Rev. **16**, 449–465 (2012)
7. Tyagi, H., Phelan, P., Prasher, R.: Predicted efficiency of a low-temperature nanofluid-based direct absorption solar collector. J. Sol. Energy Eng. **131**, 041004 (2009)

Minisymposium: Maths for the Digital Factory

Dietmar Hömberg

Description

Around one in ten of all enterprises in the EU-27s non-financial business economy were classified to manufacturing in 2009, a total of 2.0 million enterprises. The manufacturing sector employed 31 million persons in 2009, generated 5.812 billion Euro of turnover and 1.400 billion Euro of value added. (source: Eurostat). In the last 5 or 10 years all industrialised countries have launched initiatives related to digital manufacturing, sometimes also referred to as Industry 4.0 (in Europe) or Smart Manufacturing (US).

The minisymposium provided case studies showing how mathematics are able to contribute towards digital manufacturing, topics include the coupling of machine, process and workpiece for milling processes, the optimal control of material flow problems on conveyor belts, and the discrete geometric modeling of slender flexible structures for interactive assembly simulation in automotive industry.

The talks included in the minisymposium were the following:

- M. Pfirsching. Department of Mathematics, University of Mannheim (Germany). *Material flow problems on conveyor belts using a multi-scale model hierarchy.*
- A. Schmidt. Center for Industrial Mathematics and MAPEX Center for Materials and Processes, University of Bremen (Germany). *Modelling, simulation, and optimization of thermal distortions from milling processes.*

D. Hömberg (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
e-mail: dietmar.hoemberg@wias-berlin.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_51

- M. Roller. Institute for Industrial Mathematics, Kaiserslautern (Germany). *Discrete geometric modeling of slender flexible structures for interactive assembly simulation in automotive industry.*
- M. A. González. GKN Driveline Vigo, Vigo (Spain). *Numerical simulation of induction heating on ferromagnetic materials parts.*

Modelling, Simulation, and Optimization of Thermal Deformations from Milling Processes

Alfred Schmidt, Carsten Niebuhr, Daniel Niederwestberg, and Jost Vehmeyer

Abstract During a machining process, the produced heat results in thermomechanical deformation of the workpiece and thus an incorrect material removal by the cutting tool, which may exceed given tolerances.

We present a numerical model including a finite element simulation for thermomechanics, a dexel model for material removal, and a process model for forces and heating produced by the machining tool.

For minimization of the final shape deviation, this forward model defines the constraints for an optimal control problem. Main control variables are the process parameters and the path of the machining tool. These are varied according to a compensation and optimization approach.

1 Introduction

The heat produced by a machining process results in thermomechanical deformation of the workpiece and thus an incorrect material removal by the cutting tool. When the resulting deviation from a desired shape is too large, an additional finishing process step is needed. Especially when machining thin walled parts for lightweight structures, the deviations may be quite large. The goal of a process optimization is to minimize the deviation from the desired shape during respectively after the milling process. Thus, the additional process step for corrections can be avoided, the quality of workpieces is improved, and overall speed of production can be increased.

We present a numerical model which is built by coupling the finite element simulation for thermomechanics, a dexel model for material removal, and a process

A. Schmidt (✉) • C. Niebuhr • J. Vehmeyer
Center for Industrial Mathematics and MAPEX Center for Materials and Processes, University of Bremen, Bremen, Germany
e-mail: schmidt@math.uni-bremen.de; niebuhr@math.uni-bremen.de; vehmeyer@math.uni-bremen.de

D. Niederwestberg
Institute of Production Engineering and Machine Tools, Leibniz Universität Hannover, Hannover, Germany
e-mail: niederwestberg@ifw.uni-hannover.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_52

model for forces and heating produced by the machining tool. This combined model defines the constraints for an optimal control problem. Main control variables are the process parameters and the path of the machining tool, which are varied according to the compensation and optimization approach described in the final section.

The material removal leads to a time-dependent domain, where the most relevant boundary conditions for the thermomechanics are given especially on the moving part of the boundary, where the cutting tool engages. At the same place, deformations lead to incorrect material removal. Thus, good approximations of the time-dependent domain and of the PDE solutions near the moving boundary are important challenges for the simulation and optimization.

2 Model and Numerical Method

We propose a macroscopic model for a milling or drilling process which consists of a coupled system of geometry approximation, contact zone analysis, and thermomechanics [3], see Fig. 1. The time-dependent workpiece geometry which results from material removal by the cutting tool is approximated and computed by dexel fields in all three coordinate directions. Analysis of the current tool contact zone via the dexel model gives local chip thickness and other process parameters, which enter a process model that computes local heat sources and forces at the dexel end points of the contact zone [4]. These represent boundary data for the thermomechanics simulation by the finite element method, which computes

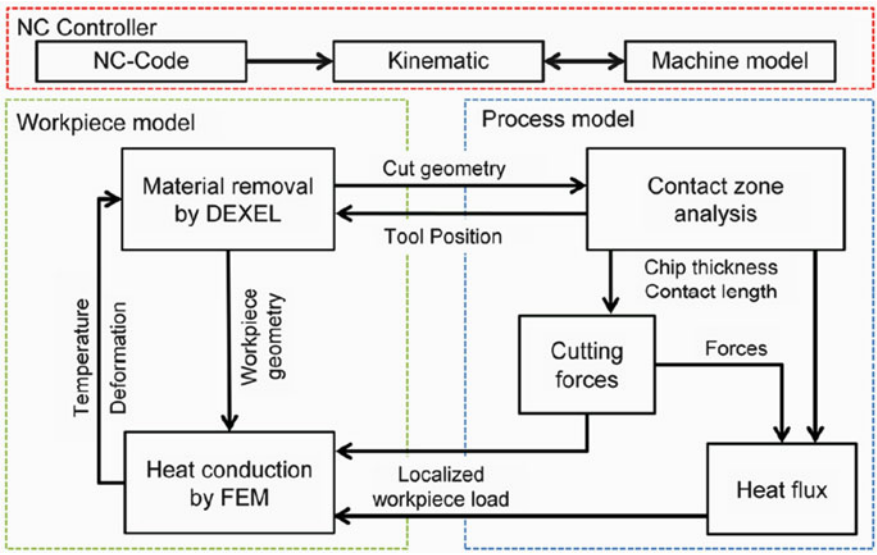


Fig. 1 Components of the numerical model, with information exchange

temperature and deformations by solving the time-dependent heat equation and quasi-stationary elasticity. Deformation near the contact zone is included in the calculation of material removal, while local temperature is one of the parameters of the process model.

2.1 Handling of the Time-Dependent Workpiece Geometry

The time-dependent domain due to the material removal is reflected in the finite element method via the dexel fields. Figure 2 depicts in a simple 2D situation, how the dexel cut points are used to define the domain as a piecewise (bi-) linear subgraph, and the finite element mesh for the current domain consists of all elements of the (locally refined) mesh which intersect with the subgraph. As the boundary of the discrete domain Γ_h and the graph Γ do not coincide, the boundary data (heat flux and forces) have to be transferred and adjusted accordingly [6].

2.2 Simulation of Milling Processes

We demonstrate and verify the numerical method by application to the milling of a thin walled part from a block clamped via two bored holes, used as a reference process. Figure 3 shows geometry, temperature, and (amplified) deformation at one

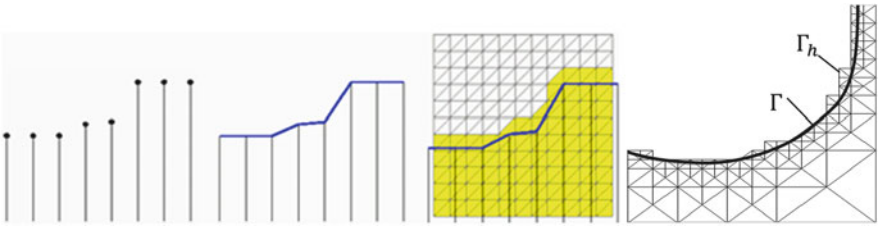


Fig. 2 Generation of finite element subdomain from dexel model, 2D sketch. *Left*: dexel endpoints, subgraph, and selection of mesh elements. *Right*: Boundaries of exact and discrete subdomains

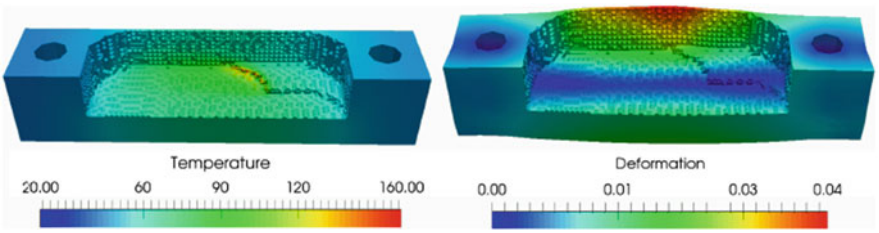


Fig. 3 Temperature [$^\circ C$] and deformation [*mm*], shown amplified by a factor of 100, during the milling of the reference workpiece. Deformation is large particularly at the thin wall

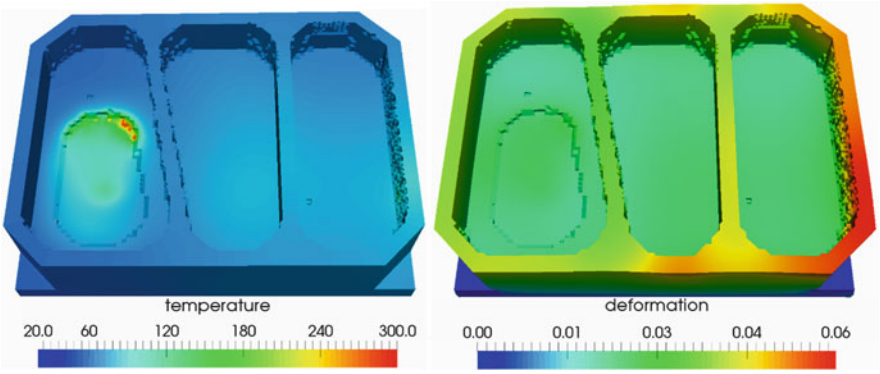


Fig. 4 Milling of a thin walled workpiece: Geometry, temperature ([$^\circ C$], *left*), and deformation ([*mm*], *right*, amplified by factor 100) during the process

time during the process. Comparison shows a good agreement with experimental data [5], for both temperature and deformation.

The reference process is a simple model for more involved milling processes with thin walled workpieces for lightweight structures, see Fig. 4 for an example.

3 Compensation and Optimization

When the shape deviation of the workpiece due to thermomechanical deformations during the process is above a given tolerance, a costly postprocessing step is needed, or the process parameters have to be adjusted. The optimization procedure for parameters and machining tool path is based on the numerical model described above.

Let $y \in Y$ denote the state (geometry, deformation, temperature, ...) and u the control (process parameters, tool path, ...) from a set of admissible controls U_{ad}. The numerical model defines the forward (control-to-state) operator $y = A_s(u)$. Typically, the control enters the model via the material removal and the process model, defining the domain and the boundary conditions for the thermomechanics subproblem [8]. Given an error functional $J(y, u)$, the optimal control problem is

$$\min_{y \in Y,\, u \in U_{ad}} J(y, u) \quad \text{subject to} \quad y = A_s(u).$$

As the state is high-dimensional and the control-to-state operator involving the time-dependent milling process simulation is computationally very expensive, the standard optimization approach with solution of dual problems for computation of descent directions et cetera is not applied here, but instead a greedy approach is used, where the high dimensional problem is divided into lower dimensional subproblems.

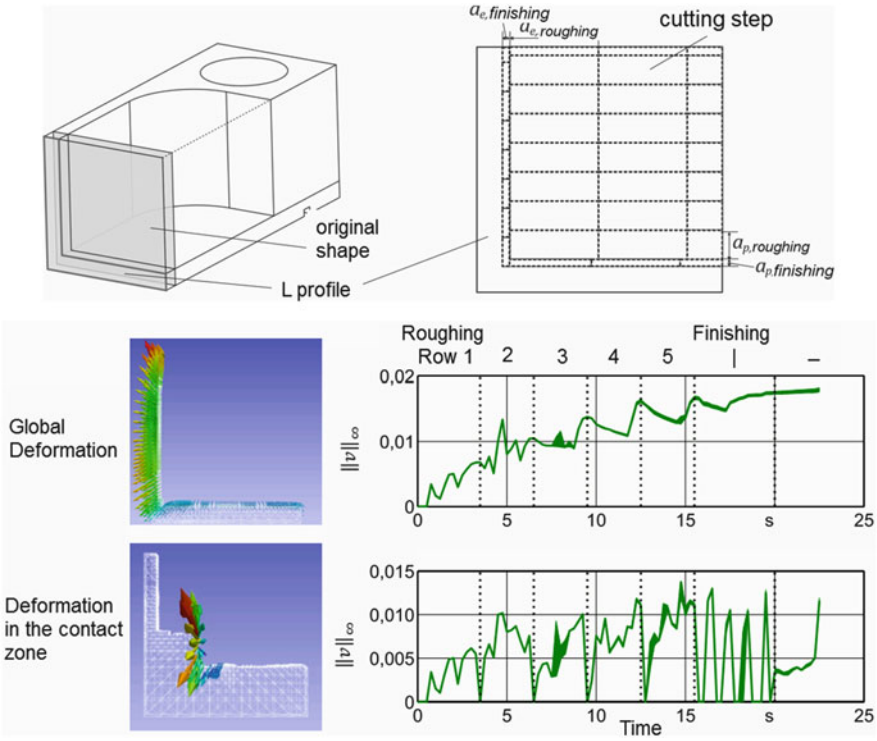


Fig. 5 Model process for optimization of L-shaped domain (*top*), comparison of global (*middle*) and local (*bottom*) error criteria

This can involve a subdivision of the complete process time into smaller time intervals as well as a separate consideration of different process parameter sets [1].

The selection of a reasonable optimization criterion is demonstrated for a narrow, finally L-shaped slice of the reference workpiece with roughing and finishing process steps, see top of Fig. 5. Examples of two different norms for optimization criteria based on the deformation are also depicted in Fig. 5. The middle part shows the global deformation of the workpiece at one time on the left and the maximum norm of global deformation over time, the lower part shows on the left the deformation in the contact zone, where the milling tool removes material, at one time and the maximum norm of deformation in the contact zone over time. While the final goal is to optimize the desired shape globally, the only influence of deformations is via the cutting of the tool, which is reflected by the more local criterion. Some results of a process parameter optimization are presented in [1].

A direct compensation approach by adaption of the tool path and axis is presented in [2, 7]. The tool can be shifted (and rotated) according to the local thermomechanical deformation at the current contact zone. As the deformation is typically not constant over the contact zone, a suitable averaging procedure has to

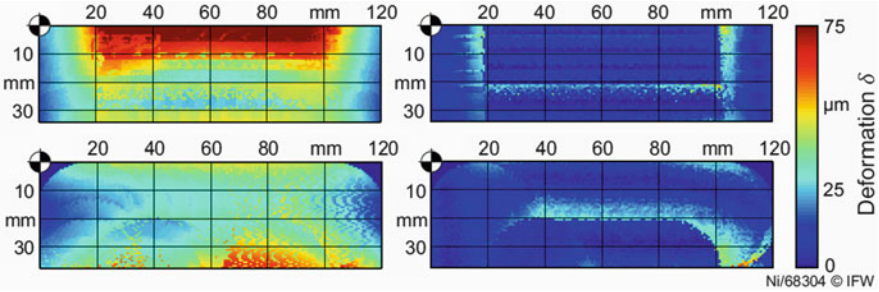


Fig. 6 Compensation for the reference workpiece: Deformation in side view (*top*) and top view (*bottom*), on the left after milling with original tool path, on the right with compensated tool path

be applied in order to compute a corrected tool axis. Based on the current forward simulation, this compensation approach can be interpreted as a descent step in the optimization procedure. Numerical and experimental verifications show a very good reduction of final shape errors for the reference process [1, 2, 7], see Fig. 6.

Acknowledgements We thank Dietmar Hömberg for organizing MS37 "Maths for the digital factory" where we could report about the subject.

The presented results have been obtained within the research project "Thermomechanical Deformation of Complex Workpieces in Drilling and Milling Processes" (DE 447/90-3, MA 1657/21-3) within the DFG Priority Program 1480 "Modeling, Simulation and Compensation of Thermal Effects for Complex Machining Processes". The authors would like to thank the DFG for its financial and organizational support of the project.

References

1. Denkena, B., Maaß, P., Schmidt, A., Niederwestberg, D., Vehmeyer, J., Niebuhr, C., Gralla, P.: Thermomechanical deformation of complex workpieces in milling and drilling processes. In: Biermann, D., Hollmann, F. (eds.) Thermal Effects in Complex Machining Processes, Final Report of the DFG Priority Programme 1480. Springer Lecture Notes in Production Engineering Series (2017, to appear)
2. Denkena, B., Niederwestberg, D.: Tool path adaption for compensation of thermal workpiece deflections based on virtual machining. In: Altintas, Y. (ed.) Proceedings of the 4th International Conference on Virtual Machining Process Technology, VMPT (2015)
3. Denkena, B., Schmidt, A., Henjes, J., Niederwestberg, D., Niebuhr, C.: Modeling a thermomechanical NC-simulation. Procedia CIRP **8**, 69–74 (2013)
4. Denkena, B., Vehmeyer, J., Niederwestberg, D., Maaß, P.: Identification of the specific cutting force for geometrically defined cutting edges and varying cutting conditions. Int. J. Mach. Tools Manuf. **82–83**, 42–49 (2014)
5. Denkena, B., Schmidt, A., Maaß, P., Niederwestberg, D., Niebuhr, C., Vehmeyer, J.: Prediction of temperature induced shape deviations in dry milling. Procedia CIRP **31**, 340–345 (2015)
6. Niebuhr, C., Niederwestberg, D., Schmidt, A.: Finite element simulation of macroscopic machining processes – Implementation of time-dependent domain and boundary conditions. ZeTeM Report 14–01, p. 14, University of Bremen (2014)

7. Niederwestberg, D.: Prognose und Kompensation der temperaturbedingten Werkstückverlagerungen beim Trockenfräsen. PhD Thesis, Leibniz Universität Hannover (2016)
8. Schmidt, A., Bänsch, E., Jahn, M., Luttmann, A., Niebuhr, C., Vehmeyer, J.: Optimization of engineering processes including heating in time-dependent domains. In: Bociu, L., Desideri, J.A., Habbal A. (eds.) System Modeling and Optimization, 27th IFIP TC 7 Conference, CSMO 2015, Sophia Antipolis, June 29–July 3, 2015. Revised Selected Papers. Springer IFIP AICT Series, vol. 494, pp. 452–461 (2017)

Minisymposium: MODCLIM: Erasmus+ Project

Matylda Jabłońska-Sabuka

Description

MODCLIM is an integrated research training course and problem solving workshop for challenging mathematical and computational problems from industry and applied sciences. During each of the years of the project, a group of 20–30 PhD and advanced undergraduate (MSc) level students are trained, and 4–6 industrial problems from leading edge technological development are worked through. The project unites a number of ECMI partners (all of whom have experience in running a Masters programme in industrial mathematics and have contacts with industry in their region) but is also the first one of this kind in the Sub-Mediterranean area and an important step towards cross-mediterranean collaboration in industrial mathematics.

The purpose is to train MS students and PhD students in research skills of applied mathematics. The project concentrates on practical problems coming from real world and results in development of research capacity in applied mathematics, spur science and technology education. Moreover, MODCLIM prepares the students (many of whom being future academicians) in a mild way for Study Groups. The difference is that with MODCLIM the students get a proper introductory course and training tasks related to the real life modeling problem, whereas during a Study Group participants would have to search and understand the same concepts in just under a week.

The partners involved in MODCLIM are all ECMI nodes. The project was created in the spirit of ECMI's Modeling Weeks and Study Groups. The idea was

M. Jabłońska-Sabuka (✉)
Department of Computational Engineering, Lappeenranta University of Technology, Lappeenranta, Finland
e-mail: matylda.jablonska-sabuka@lut.fi

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_53

to expose the students to real industrial problems, including the training school necessary for them to grasp the problem the students would work on.

The talks included in the minisymposium were the following:

- Juan Rocha. University of Las Palmas Gran Canaria (Spain). *Modeling Clinic for Industrial Mathematics: A collaborative project under Erasmus+ program.*
- Matylda Jabłońska-Sabuka. Lappeenranta University of Technology (Finland). *The importance of synergies among mathematical modeling activities: MODCLIM - ESGI example.*
- Florian Gensheimer. Mathematical Institute, University of Koblenz-Landau (Germany). *MODCLIM 2015: A student's point of view.*
- Thomas Götz. University of Koblenz, Landau (Germany). *MODCLIM 2016 - Review of the Modeling Clinic at Koblenz University".*
- Giovanni Russo. Department of Mathematics and Computer Science, University of Catania (Italy). *MODCLIM Training school, Catania 14–27 February 2016.*

Modeling Clinic for Industrial Mathematics: A Collaborative Project Under Erasmus+ Program

Agnieszka Jurlewicz, Claudia Nunes, Giovanni Russo, Juan Rocha, Matti Heilio, Matylda Jablonska-Sabuka, Nada Khoury, Poul Hjorth, Susana Serna, and Thomas Goetz

Abstract *Modeling Clinic for Industrial Mathematics* (MODCLIM) is a Strategic Partnership for the Development of Training Workshops and Modeling Clinic for Industrial Mathematics, funded through the European Commission under the

A. Jurlewicz (✉)
Faculty of Pure and Applied Mathematics, WUST, Wrocław, Poland
e-mail: Agnieszka.Jurlewicz@pwr.edu.pl

C. Nunes
Mathematics Department, IST, Lisboa, Portugal
e-mail: cnunes@math.ist.utl.pt

G. Russo
Department of Mathematics and Informatics, UNICT, Catania, Italy
e-mail: russo@dmi.unict.it

J. Rocha
Department of Mathematics, ULPGC, Las Palmas, Spain
e-mail: jrocha@dma.ulpgc.es

M. Heilio • M. Jablonska-Sabuka
Department of Mathematics and Physics, LUT, Lappeenranta, Finland
e-mail: matti.heilio@lut.fi; Matylda.Jablonska-Sabuka@lut.fi

N. Khoury
CUCID, ULPGC, Las Palmas, Spain
e-mail: nada.khoury@fpct.ulpgc.es

P. Hjorth
Department of Applied Mathematics and Computer Science, DTU, Kongens Lyngby, Denmark
e-mail: pghj@dtu.dk

S. Serna
Department of Mathematics, UAB, Birmingham, AL 35294, USA
e-mail: serna@mat.uab.es

T. Goetz
Mathematical Institute, UNIKOLD, Mainz, Germany
e-mail: goetz@uni-koblenz.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_54

Erasmus Plus Program, Key Action 2: Cooperation for innovation and the exchange of good practices.

MODCLIM develops a project focused on mathematical technologies needed for the progress of industry and novel engineering solutions. The project pretends to contribute in addressing the challenges that industrial mathematics, as a part of the next generation of methodologies in research and development (R&D) and knowledge management, represents for university education, curriculum development, training practices and research collaboration.

MODCLIM project introduces into European higher education a novel concept in academic research training and education of industrial mathematics by combining elements from earlier successful ideas tested worldwide. It is organized in cycles, each consisting of three interconnected steps: (1) Training school, (2) Intermediate online-project, and (3) Modeling Clinic. The project's aim is to provide an integrated learning experience for the students, thus ensuring a proper connection between these three steps.

In this paper we expose an overall idea about MODCLIM project, and a brief description of two MODCLIM cycles organized in 2015 and 2016.

1 Introduction

Progress of industry and novel engineering solutions are increasingly depending on computational approach, modeling and simulation. Mathematical technologies are needed to produce effective design processes, accelerate test cycles, spur innovation, support systems integration schemes and control the production models. Industrial mathematics is among the next generation methodologies in R&D and knowledge management. This presents a challenge for university education, curriculum development, training practices and research collaboration. With the main aim to contribute addressing these challenges, MODCLIM (Modeling Clinic for Industrial Mathematics, see [4]) project develops and integrates a research training course and a modeling clinic workshop for the challenging mathematical and computational problems from industry and applied sciences.

The MODCLIM project concept is a novel approach that assimilates features from prior European experience (see [1] and [3]) and successful innovative models overseas, like RIPS's model from USA (see [5]). The project is organized in cycles, each consisting of three interconnected steps: (1) Training school, (2) Intermediate online-project, and (3) Modeling Clinic. The purpose of the MODCLIM cycle is to train a group of 20–25 PhD and advanced undergraduate (MSc, BSc) level students, addressing different industrial problems from leading edge technological development. The topics addressed should be different in each cycle, although all of them have to be focused on mathematical modeling of industrial problems and the training of future researchers through intensive courses and problem solving. It will result in development of research capacity in applied mathematics, spur science and technology education and inspire curriculum updates. Moreover, these results

will contribute to promote technological development in the participating countries and modernize the education and applied research agenda in the participating departments. For students, the project will help increase their career and employment perspectives.

There are great many universities and hundreds of industries in Europe and neighboring partner countries that could implement a similar configuration of training courses and modeling clinic workshops in order to establish a bridge between academic mathematics, computational technology and the real R&D world (see [2]). The MODCLIM project intends to provide a model case and template for such arrangements.

2 MODCLIM Framework and Consortium

The *Strategic Partnership for the Development of Training Workshops and Modeling Clinic for Industrial Mathematics* (acronym MODCLIM) is a strategic partnership for higher education project, approved in 2014 by Erasmus+ program, Key Action 2: Cooperation for innovation and the exchange of good practices. Under this framework, the MODCLIM Consortium is conformed by the following Institutions: **Technical University of Denmark** (hereinafter DTU, Denmark), **Instituto Superior Técnico Lisboa** (hereinafter IST, Portugal), **Lappeenranta University of Technology** (hereinafter LUT, Finland), **Universitat Autónoma de Barcelona** (hereinafter UAB, Spain), **University of Bristol** (hereinafter UOB, England), **Universitá Degli Studi di Catania** (hereinafter UNICT, Italy), **University of Koblenz-Landau** (hereinafter UNIKOLD, Germany), **University of Las Palmas de Gran Canaria** (hereinafter ULPGC, Spain) and **Wroclaw University of Science and Technology** (hereinafter WUST, Poland). The Consortium is supported by the following Associated Entities: **University of Djelfa** (Algeria), **Mohammed V University of Rabat** (Morocco) and **Instituto Tecnológico de Canarias** (ITC, Spain).

The project is led by the University of Las Palmas de Gran Canaria, and managed through the University Center for International Development Cooperation (CUCID), as well as the Department of Mathematics and SIANI Institute. However, the project Academic Coordinator is the representative from Lappeenranta University of Technology, Dr. Matti Heillio. In general, the most relevant aspect regarding the Consortium is that the partners are very active members, and share the activities and responsibilities between them.

3 Structure of the MODCLIM Activities

MODCLIM Project activities have a form of three-step cycles finished with Final Project Reports. MODCLIM applies the ECTS credit system in accordance to the Bologna process, so that a student who completes the whole activities within the cycle could recognize a maximum of 10 ECTS. Each cycle consists of:

1. **Training school**, where during 2 weeks students have intensive lecture courses and project work on 4 topics (2 topics per week). Projects are R&D problems requiring mathematical modelling, computer simulation and innovative analysis. Projects emerge from industry, society and applied sciences. Topics range from engineering, novel materials and manufacturing, energy and environment to bioprocesses and medical technology, quantitative methods of data analysis, etc.
2. **Intermediate online-project work**, where during 2 months students work in teams (on-line from their home countries) on a project assignment. In this second stage the students choose between four projects assignments offered by the lecturers in order to create four intermediate project teams. Each intermediate project topic is connected to the Training School lecture course—so it could be considered as an extended exercise or application project based on the learning obtained at the lectures.
3. **Modeling Clinic Workshop** that is a 1 week session carried out in study group format and focusing on the cracking of the actual industrial problem. In addition to the Training school students (PhD, MS), also other mathematical scientists, faculty members, and postdocs participate workshop serving as project team leaders and industry liaisons representing the company submitting the problem.
4. **Final project report** preparation, where each students' team summarizes the results obtained during the Modeling Clinic supervised by the instructor.

The partner universities that send a lecturer to the Training School, should also bring a project to the Modelling Clinic which is connected to the Training School in such a way that the intensive dose of theory partially reveals some ways to solve the problem and provides tools to tackle the project. Moreover, the projects for the Modeling Clinic should exhibit "Industrial mathematics" or "real world impact" in the broad sense.

Two such cycles were organized in the MODCLIM framework along the years 2015 and 2016. In the next sections we provide more details about their content and results.

4 MODCLIM Cycle 2015

MODCLIM project was launched during the kick off meeting held at the ULPGC from the 1st until the 2nd of December 2014. During this meeting the Consortium agreed on the activities to develop along 2015, as well as the process to follow in order to propose and select topics, lecturers/instructors, students, etc. Moreover, the hosting universities for the Training Schools and Modelling Clinic Workshops, for both cycles 2015 and 2016, were approved.

As for the MODCLIM Cycle 2015, from December 2014 until early February 2015 the Consortium was able to define the topics, select the lecturers and select the students for the **Training School 2015**. This event took place in the ULPGC

from the 15th until the 28th of March 2015. The courses, topics and lecturers for this event were the following:

1. *Introduction to continuum mechanics with application to the medical treatment of eye (corneal topography)*. Dr. Łukasz Płociniczak - WUST.
2. *Modelling and optimization of wind farms/ solar farms*. Dr. Gustavo Montero, Dr. Felipe Diaz, Dr. Eduardo Barrera - ULPGC.
3. *Mathematical models for semiconductor industry*. Dr. Vittorio Romano - UNICT.
4. *Ecomathematics, mathematical models in ecology/ environmental engineering*. Dr. Virpi Junttila - LUT.

Once the Training School 2015 had finished, each student was assigned to one of the four **intermediate project teams/intermediate online-projects 2015** according to her/his preferences. From April to June of 2015, the four intermediate project teams/intermediate online-projects 2015 were developed with the same topics as the corresponding courses of the Training School 2015.

Modelling Clinic Workshop 2015 was co-organized and merged with the *European Study Group with Industry (ESGI 112)*, within the event entitled **Workshop on Computational and Modelling Problems for Industry** which was held in LUT, from the 7th until the 11th of September 2015. This event was organized in collaboration with LUT and St Petersburg State Polytechnic University. The problems addressed in the workshop were the following:

1. *How to measure the intra-ocular pressure*. Dr. Łukasz Płociniczak - WUST.
2. *Thrusters with additional stators*. Dr. Joonas Sorvari - LUT.
3. *Solar farm modeling*. Dr. Felipe Diaz - ULPGC.
4. *Detection and compensation of Vortex Induced Vibrations*. Dr. Sergey Lupuleac - St Petersburg State Polytechnic University.
5. *Non-aerobic water purification reactor*. Dr. Tuomo Kauranne - LUT.
6. *Performance of a pn junction in a solar cell*. Dr. Massimo Camarda - Paul Scherrer Institute (PSI, Switzerland)
7. *Mathematical modeling of displacements and thermal stresses in anisotropic materials in cooling*. Dr. Jari Jarvinen - CSC, Finish IT Center for Science

To complete MODCLIM Cycle 2015, the students have elaborated **Final Project Reports 2015** in teams regarding to the following topics addressed during the Modelling Clinic Workshop 2015:

1. *Solar Project of MODCLIM Final Report* - Supervisor: Dr. Felipe Diaz - ULPGC;
2. *Performance of pn Junction in a Solar Cell* - Supervisor: Dr. Massimo Camarda - PSI;
3. *Eye Structure Modelling* - Supervisor: Dr. Łukasz Płociniczak - WUST;
4. *Removal of Chemical Oxygen Demand (COD) from Gelatine Manufacturing Waste Water using Anaerobic sequential Batch Reactor* - Supervisor: Dr. Tuomo Kauranne - LUT.

After the finalization of the Training School 2015 and the Modeling Clinic Workshop 2015, all the students were requested to complete a questionnaire. The results of this survey were very gratifying.

5 MODCLIM Cycle 2016

On the 9th of September 2015 the Intermediate Project Meeting was held in LUT. In this meeting, the Consortium agreed on the activities to develop in 2016 and the process to follow in order to propose and select topics, lecturers/instructors, students, etc. From November until late December of 2015 the Consortium defined the topics, selected the lecturers and selected the students for the **Training School 2016** that then took place at UNICT, from the 14th until the 27th of February 2016. The courses, topics and lecturers for this event were the following:

1. *Extracting information from data: Principal component analysis, robustness, and outlier detection*. Dr. M. Rosário Oliveira - IST, Universidade de Lisboa.
2. *Image Analysis. Focus on image restoration based on variational models*. Dr. Susana Serna - UAB.
3. *Smoothed particle hydrodynamics for non Newtonian fluids and applications*. Dr. Giuseppe Bilotta - UNICT.
4. *Dynamic Network optimization and applications in civil security*. Dr. Jan Ohst - UNIKOLD.

Once the Training School 2016 had finished, each student was assigned to one of the four **intermediate project teams/intermediate online-projects 2016** according to her/his preferences. From March to May of 2016, the four intermediate online-projects 2016 were developed with the same titles as the corresponding courses of the Training School 2016.

Modelling Clinic Workshop 2016 was held at UNIKOLD, from the 16th until the 22th of May 2016. The problems addressed during this workshop were the following:

1. *Extracting information from data: Principal component analysis, robustness, and outlier detection*. Dr. M. Rosário Oliveira - IST, Universidade de Lisboa.
2. *Super-resolution recovery of video imaging through turbulence*. Dr. Susana Serna - UAB.
3. *Smoothed particle hydrodynamics for non Newtonian fluids and applications*. Dr. Giuseppe Bilotta - UNICT.
4. *Dynamic Network optimization and applications in civil security*. Dr. Jan Ohst - UNIKOLD.

Regarding Final Project Reports 2016, the students were working on their corresponding final reports until late July 2016. The titles of the Final Project Reports 2016 are the following:

1. *Extracting information from data: Principal component analysis, robustness, and outlier detection*. Supervisor: Dr. M. Rosário Oliveira - IST.
2. *Super-resolution recovery of video imaging through turbulence*. Supervisor: Dr. Susana Serna -UAB.
3. *Paper coating problem with Newtonian fluids and with Bingham fluids*. Supervisor: Dr. Giuseppe Bilotta - UNICT.
4. *Quickest path in dynamic networks*. Supervisor: Dr. Jan Ohst - UNIKOLD.

At the time of drafting this paper, September 2016, the Consortium is working on the final project report, dissemination activities (among them MODCLIM International Dissemination Seminar, organized at WUST on 13th of September 2016), survey for the students involved in MODCLIM cycle 2016, intellectual outputs, etc. for closing MODCLIM project at the end of December 2016.

6 MODCLIM Future

The Consortium agreed on continuing the collaboration and apply for a new edition of MODCLIM project. In fact, the Consortium and the Associated Entities have been extended: Institut National des Sciences Appliquées Rouen (France) is a new partner, and St. Petersburg State Politechnical University (Russia) is a new Associated Entity.

The Consortium will continue searching for additional sources for supporting such successful learning structure; and intensify ties between the industrial sectors and the MODCLIM project, for developing new and innovative approaches and supporting the dissemination of best practices in industrial mathematics.

Acknowledgements We would like to acknowledge the support of the Erasmus+ Program for funding the MODCLIM project; as well as the support of all the universities and entities involved in MODCLIM project, and the collaboration of the teachers and students who have participated throughout the project.

References

1. European Consortium for Mathematics in Industry (ECMI) - Blog. http://www.ecmiindmath.org
2. Mathematics for Industry Network (MI-NET). http://mi-network.org
3. Mathematics in Industry. Study Groups with Industry (Univesity of Oxford). http://miis.maths.ox.ac.uk/
4. Modeling Clinic for Industrial Mathematics (MODCLIM). http://modclim.ulpgc.es
5. Research in Industrial Projects for Students (RIPS) summer program. http://www.ipam.ucla.edu/rips/

Minisymposium: Moving Boundary Problems in Industrial Applications

Brendan J. Florio

Description

Moving boundary problems appear in the modelling of a variety of physical processes including phase-change (melting and solidification) and the dissolution of a solid in a solvent and fluid dynamics. Due to this broad scope, we see moving boundary problems in many industrial applications including (but not limited to) metal-casting, nanotechnology and pharmacology. Here we presented a collection of studies from these diverse fields and we highlight various analytical and numerical techniques.

The problems presented here were perfect examples of applying mathematical models to industrial applications. Such endeavours promoted the cooperation between mathematicians and industry and contributed to the core principals of ECMI.

The talks included in the minisymposium were the following:

- F. Font. MACSI. *Influence of substrate melting on the laser-induced dewetting of nanothin films.*
- S. Mitchell. MACSI. *The effect of superheat on the macrosegregation of a continuously cast binary alloy.*
- K. Devine. MACSI. *Mathematical modelling of the formation of oscillation marks in the continuous casting of steel.*

B.J. Florio (✉)
Mathematics Applications Consortium for Science and Industry (MACSI), University of Limerick, Limerick, Ireland
e-mail: brendan.florio@ul.ie

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_55

- B. J. Florio. MACSI. *Mathematical modelling of phase change in nanowires.*
- V. Cregan. Centre de Recerca Matemàtica, Barcelona. *Nanoparticle growth via the precipitation method.*
- N. McInerney. MACSI. *A moving boundary problem in controlled release pharmaceuticals.*

Nanoparticle Growth via the Precipitation Method

V. Cregan, T.G. Myers, S.L. Mitchell, H. Ribera, and M.C. Schwarzwälder

Abstract We consider a model for the evolution of a system of nanoparticles in solution via the processes of size focusing and Ostwald ripening. The model consists of a diffusion equation for the concentration of the solution, a Stefan-type condition to track the particle-liquid interfaces and a time-dependent expression for the bulk concentration obtained via mass conservation. Based on a small dimensionless parameter we propose a pseudo-steady state model, which is solved numerically to obtain the average particle radius and standard deviation. The results are shown to be in good agreement with experimental data for cadmium selenide nanoparticles.

1 Introduction

Nanoparticles are units of matter with dimensions in the range 1–100 nm that lie at the interface of bulk materials and atomic structures. Whilst the physical and chemical properties of bulk materials tend to be independent of size, at the nanoscale, size strongly dictates material properties. Consequently, nanoparticles exhibit many advantageous size-dependent magnetic, electrical, chemical and optical properties. These properties are extremely sensitive to the particle size, and thus the ability to produce monodisperse particles that lie within a controlled size distribution is critical.

Precipitation of nanoparticles from solution is one of the most widely employed synthesis methods [3, 8]. Typically, nucleation and growth are separated, where the former is used to generate seeds for the latter. The resulting system is not in its lowest possible energy state due to the presence of small particles. Thermodynamic equilibrium is achieved by Ostwald ripening, whereby large particles grow at the

V. Cregan (✉) • T.G. Myers • H. Ribera • M.C. Schwarzwälder
Centre de Recerca Matemàtica, Campus UAB Edifici C, Bellaterra, 08193 Barcelona, Spain
e-mail: vcregan@crm.cat; tim.myerscrm@gmail.com; hribera.crm@gmail.com; mcalvo.crm@gmail.com

S.L. Mitchell
Mathematics Applications Consortium for Science and Industry, University of Limerick, Sreelane, Castletroy, Co. Limerick, Ireland
e-mail: sarah.mitchell@ul.ie

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_56

expense of the more soluble small particles. This leads to the undesirable defocusing of the particle size distribution (PSD), caused by growth and dissolution of bigger and smaller particles, respectively. The PSD can be refocused by changing the reaction kinetics by the addition of precipitating material [4].The main disadvantage of this process is that the precise relationship between particle growth and system conditions is still not fully understood.

In the following we develop a differential equation model to describe the change in a system of a nanoparticles. We begin with a single particle model in order to understand the basic growth process and then extend it to a system of N-particles. The model is solved numerically and compared to data for cadmium selenide [4].

2 Growth of a Single Particle

As shown in Fig. 1, we consider a single, spherical nanoparticle, with radius r_p^*, where * denotes dimensional quantities. As per the La Mer model [2], following nucleation we assume the system evolves solely via particle growth. The monomer concentration adjacent to the particle surface and in the far-field are c_i^* and c_b^*,

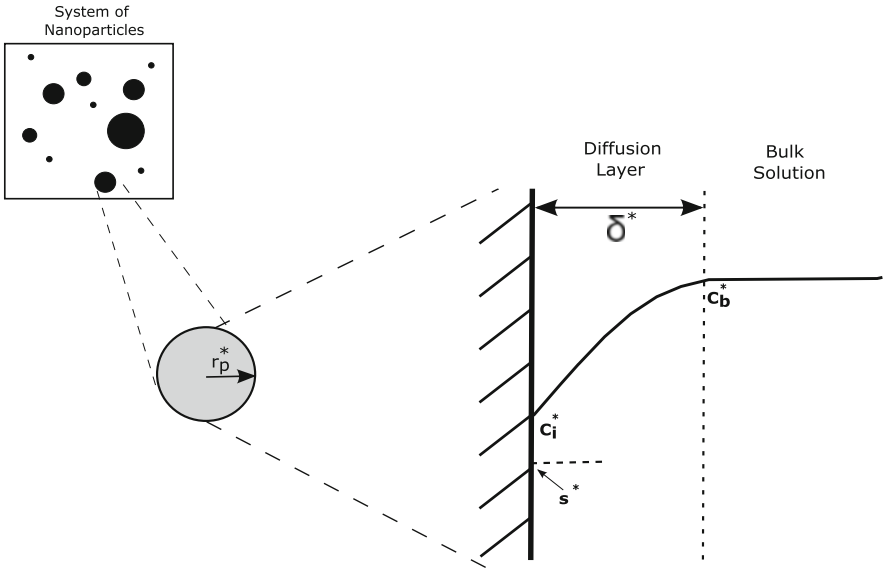


Fig. 1 Schematic of a system of nanoparticles in a solution with a zoom-in of a single nanoparticle and the surrounding monomer concentration profile where s^*, c_i^* and c_b^* are the particle solubility, the concentration at the surface of the particle and the far-field concentration, respectively [6]

respectively. The Ostwald-Freundlich condition (OFC)

$$s^* = s^*_\infty \exp\left(\frac{2\sigma V_M}{r^*_p R_G T}\right) \equiv s^*_\infty \exp\left(\frac{\alpha}{r^*_p}\right), \tag{1}$$

relates the solubility, s^*, of a particle to its radius, where s^*_∞ is the bulk material solubility, σ the interfacial energy, V_M the molar volume, R_G the universal gas constant, T the absolute temperature and $\alpha = (2\sigma V_M)/(R_G T)$. If $s^* < c^*_b$ monomer molecules diffuse from the bulk towards the particle to react with the surface and the particle grows, whereas if $s^* > c^*_b$ the particle shrinks. In addition, the OFC indicates that solubility increases as particle size decreases. The fact that smaller particles are more soluble than larger ones is the basis of Ostwald ripening.

Assuming the solution is initially well mixed with concentration $c^*_{b,0}$, the governing model is

$$\frac{\partial c^*}{\partial t^*} = \frac{D}{r^{*2}}\frac{\partial^*}{\partial r^*}\left(r^{*2}\frac{\partial c^*}{\partial r^*}\right), \tag{2}$$

$$c^*(r^*_p, t^*) = c^*_i\,, \quad c^*(\delta^*, t^*) = c^*_b(t^*)\,, \quad c^*(r^*, 0) = c^*_{b,0} \quad \text{for } r^* > r^*_p\,, \tag{3}$$

where c^* is monomer concentration, r^* is distance from the centre of the particle, t^* is time and D is the diffusion coefficient. As in [3, 6, 8], we include a diffusion layer of length δ^* around the particle, where the concentration adjusts from the c^*_i to c^*_b. We note that the 0 subscript refers to the value of the associated parameter at $t^* = 0$.

In practice, the particle surface concentration, c^*_i, is difficult to measure [6]. We eliminate it from the model using two equivalent expressions for the mass flux at the particle surface, J. Via Fick's first law (for the flux passing through a spherical surface of radius r^*) and the first-order reaction we have

$$J = 4\pi r^{*2} D\frac{\partial c^*}{\partial r^*}\,, \qquad J = 4\pi r^{*2} k(c^*_i - s^*)\,, \tag{4}$$

where k is a rate constant. Equating these expressions at the particle surface yields the condition

$$c^*_i = s^* + \frac{D}{k}\frac{\partial c^*}{\partial r^*}\bigg|_{r^*=r^*_p}. \tag{5}$$

To obtain an equation for the radius we note the change in volume of a particle

$$\frac{d}{dt^*}\left(\frac{4}{3}\pi r^{*3}_p\right) = 4\pi r^{*2}_p \frac{\mathrm{d}r^*_p}{\mathrm{d}t^*} = V_M J\,. \tag{6}$$

Substituting the first expression in (4) into (6) leads to the Stefan-type condition

$$\frac{\mathrm{d}r_p^*}{\mathrm{d}t^*} = V_M D \frac{\partial c^*}{\partial r^*}\bigg|_{r^*=r_p^*}, \qquad r_p^*(0) = r_{p,0}^*, \tag{7}$$

where $r_{p,0}^*$ is the initial particle radius.

For a single particle, provided the solution volume is sufficiently greater than that of the particle, we may assume c_b^* is approximately constant. For a large number of particles, or a small amount of solution, the bulk concentration will reduce in time. For a particle in a given volume V with radius R, mass conservation of the monomer atoms in the particle and surrounding solution is given by

$$\frac{4\pi}{3}\left[\rho_p r_{p,0}^{*3} + M_p c_{b,0}^*(R^3 - r_{p,0}^{*3})\right] = \frac{4\pi}{3}\left[\rho_p r_p^{*3} + M_p c_b^*(t)(R^3 - r_p^{*3})\right], \tag{8}$$

where ρ_p is density and M_p is molar mass.

2.1 Nondimensionalisation and Pseudo-steady State Solution

The model is nondimensionalised via

$$r = \frac{r^*}{r_{p,0}^*}, \quad r_p = \frac{r_p^*}{r_{p,0}^*}, \quad c = \frac{c^* - s_0^*}{\Delta c}, \quad c_b = \frac{c_b^* - s_0^*}{\Delta c}, \quad s = \frac{s^* - s_0^*}{\Delta c}, \tag{9}$$

where $\Delta c = c_{b,0}^* - s_0^*$ represents the driving force for particle growth and $s_0^* = s_\infty^* \exp(\alpha/r_{p,0}^*)$ is the initial solubility. The concentration and growth equations yield two time scales $\tau_D^* \sim r_{p,0}^{*2}/D$ and $\tau_R^* \sim r_{p,0}^{*2}/(V_M D \Delta c)$, respectively. As we are interested in particle growth, we use $t = t^*/\tau_R^*$. Rescaling (2), (3), (5), (7) and (8) gives

$$\epsilon \frac{\partial c}{\partial t} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial c}{\partial r}\right), \qquad \frac{\mathrm{d}r_p}{\mathrm{d}t} = \frac{\partial c}{\partial r}\bigg|_{r=r_p}, \tag{10}$$

$$c(r_p, t) = s + \mathrm{Da}\,\frac{\partial c}{\partial r}\bigg|_{r=r_p}, \quad c(\delta, t) = c_b(t), \quad c(r, 0) = 1, \quad r_p(0) = 1, \tag{11}$$

$$s = s_\infty \exp(\omega/r_p) - s_0, \qquad c_b(t) = -s_0 + c_{b,0}\left(\frac{\beta - 1}{\beta - r_p^3}\right) + c_1\left(\frac{1 - r_p^3}{\beta - r_p^3}\right), \tag{12}$$

where

$$\epsilon = V_M \Delta c\,, \quad \delta = \frac{\delta^*}{r^*_{p,0}}\,, \quad \mathrm{Da} = \frac{D}{k r^*_{p,0}}\,, \quad \omega = \frac{\alpha}{r^*_{p,0}}\,, \quad s_\infty = \frac{s^*_\infty}{\Delta c}\,, \tag{13}$$

$$s_0 = \frac{s^*_0}{\Delta c}\,, \quad c_{b,0} = \frac{c^*_{b,0}}{\Delta c}\,, \quad c_1 = \frac{\rho_p}{M_p \Delta c}\,, \quad \beta = \frac{R^{*3}}{r^{*3}_{p,0}}\,. \tag{14}$$

The dimensionless parameter ϵ is generally very small for nanoparticle growth. Talaplin et al. [7] quote $\epsilon = \mathcal{O}(10^{-6})$ for cadmium selenide (CdSe) nanoparticles. Hence, assuming $\epsilon \ll 1$ we neglect the time derivative in the diffusion equation. Upon integrating the reduced equation and applying the boundary conditions we obtain the pseudo-steady state solution

$$c = -\frac{A}{r} + B\,, \quad \text{where} \quad A = \frac{\delta r_p^2 (s - c_b)}{r_p(r_p - \delta) - \delta\,\mathrm{Da}}\,, \qquad B = s + A\left(\frac{1}{r_p} + \frac{\mathrm{Da}}{r_p}\right). \tag{15}$$

Several authors observed that $r_p \ll \delta$ [5–7], which upon applying to (15) yields

$$A = \frac{r_p^2 (c_b - s)}{\mathrm{Da} + r_p}\,, \qquad B = c_b\,. \tag{16}$$

Substituting (15) into the Stefan condition in (10) leads to the nonlinear equation

$$\frac{\mathrm{d}r_p}{\mathrm{d}t} = \frac{c_b - s}{\mathrm{Da} + r_p} \equiv \frac{c_b + s_0 - s_\infty \exp(\omega / r_p)}{\mathrm{Da} + r_p}\,. \tag{17}$$

3 Evolution of a System of N Particles

The single particle model is extended to a system of N particles subject to a normal distribution with mean initial radius $\overline{r}^*_{p,0}$ and standard deviation σ_o. The particle radii, initial radii and solubilities are denoted r^*_i, $r^*_{i,0}$ and s^*_i, respectively, where i represents the ith particle and $i = 1 \ldots N$. We rescale via (9) with $\overline{r}^*_{p,0}$ replacing $r^*_{p,0}$. Hence, in what follows, all overlined dimensionless parameters are the same as those in (13) and (14), except $\overline{r}^*_{p,0}$ replaces $r^*_{p,0}$.

Under the pseudo-steady approximation the growth of each particle is described by (17) with r_p, s_0 replaced by $r_i, s_{i,0}$. In dimensionless form we have

$$\frac{\mathrm{d}r_i}{\mathrm{d}t} = \frac{c_b - \overline{s}}{\overline{\mathrm{Da}} + r_i} \equiv \frac{c_b + \overline{s}_0 - \overline{s}_\infty \exp(\overline{\omega}/r_i)}{\overline{\mathrm{Da}} + r_i}\,. \tag{18}$$

and

$$c_b(t) = -\overline{s_0} + \overline{c}_{b,0}\left(\frac{R^3 - \sum_{i=1}^{N} r_{i,0}^{*3}}{R^3 - \overline{r}_p^3 \sum_{i=1}^{N} r_i^3}\right) + \overline{c}_1\left(\frac{\sum_{i=1}^{N} r_{i,0}^{*3} - \overline{r}_p^3 \sum_{i=1}^{N} r_i^3}{R^3 - \overline{r}_p^3 \sum_{i=1}^{N} r_i^3}\right). \tag{19}$$

4 Results

The numerical solution of (18) and (19) is compared with data from the experimental study of Peng et al. [4] for CdSe nanoparticles. Several unknown parameters (in italics in Table 1) are obtained via a parameter sweep routine similar to those used in [1, 3].

Figure 2 compares a simulation for 1000 CdSe particles (dashed lines) with the corresponding data (dash dotted lines) of [4]. Specifically, Fig. 2a and b show the evolutions of the average particle radius, $\overline{r_p}$, and the standard deviation percentage, $\sigma_{o,\%}$ ($\equiv \sigma_0/\overline{r_p}$), respectively. Initially, the solution concentration is higher than the solubility of all the particles. Thus, the particles consume material in the solution and grow, leading to an increase in $\overline{r_p}$. As the growth rate of the small particles is larger than that of the bigger particles, size focusing occurs as indicated by the rapid decrease in $\sigma_{o,\%}$ in the first 20 min. Particle growth results in the depletion of material from the solution, and hence a decrease in the concentration. Eventually, the solubility of the smaller particles drops below the bulk concentration, and thus they begin to dissolve, whilst bigger particles continue to grow. Ostwald ripening

Table 1 Physical parameters for the cadmium selenide nanoparticle synthesis method of Peng et al. [4]

Quantity	Symbol	Value	Units
Universal gas constant	R_G	8.31446	$\mathrm{J\,mol^{-1}\,K^{-1}}$
Boltzmann constant	k_b	1.38×10^{-23}	$\mathrm{m^2\,kg\,s^{-2}\,K^{-1}}$
Density	ρ_p	5816	$\mathrm{kg\,m^{-3}}$
Molar volume	V_M	3.29×10^{-5}	$\mathrm{m^3\,mol^{-1}}$
Molar mass	M_p	0.1914	$\mathrm{kg\,mol^{-1}}$
Solution temperature	T	573.15	K
Surface energy	σ	0.44	$\mathrm{J\,m^{-2}}$
Diffusion coefficient	D	10^{-9}	$\mathrm{m^2\,s^{-1}}$
Reaction rate	k	0.8797	$\mathrm{m\,s^{-1}}$
Solubility of bulk material	s_∞^*	1.64×10^{-9}	$\mathrm{mol\,m^{-3}}$
Initial bulk concentration	$c_{b,0}^*$	9.46×10^{-7}	$\mathrm{mol\,m^{-3}}$
Volume of liquid	V	5.57×10^{-12}	$\mathrm{m^3}$

The parameters in italics are obtained via a model fitting approach

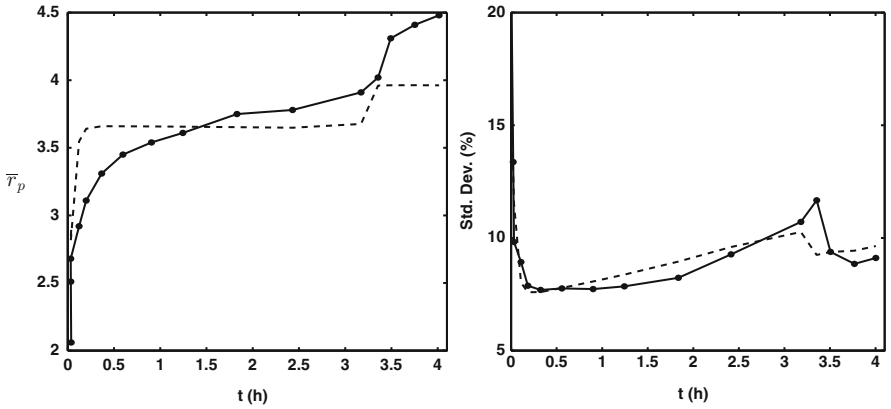


Fig. 2 Numerical simulation (*dashed line*) for (**a**) average particle radius and (**b**) standard deviation (%) compared to experimental data (*dash dotted line*) [4] for a system of 1000 CdSe nanoparticles

is observed after approximately 1 h where $\sigma_{o,\%}$ begins to gradually increase. After 190 min, additional material was added to the solution. As indicated by the decrease in $\sigma_{o,\%}$, the injection of additional material results in an almost instantaneous drop in particle size polydispersity.

5 Conclusion

We presented a model for the growth of a single nanoparticle, which was then extended to a system of N-particles. The model showed relatively good agreement with data [4] for CdSe nanoparticles. The model demonstrated both size focusing and Ostwald ripening.

Acknowledgements V. Cregan, H. Ribera and M.C. Schwarzwälder were supported by “la Caixa" foundation grant. T. Myers acknowledges the support of a Ministerio de Ciencia e Innovación Grant MTM2011-23789. S.L. Mitchell acknowledges the support of the Mathematics Applications Consortium for Science and Industry (www.macsi.ul.ie) funded by the Science Foundation Ireland grant 12/IA/1683.

References

1. Clark, M.D., Kumar, S.K., Owen, J.S., Chan, E.M.: Nano Lett. **11**(5), 1976–1980 (2011)
2. La Mer, V.K.: Ind. Eng. Chem. **44**(6), 1270–1277 (1952)
3. Mantzaris, N.V.: Chem. Eng. Sci. **60**(17), 4749–4770 (2005)

4. Peng, X., Wickham, J., Alivisatos, A.P.: J. Am. Chem. Soc. **120**(21), 5343–5344 (1998)
5. Peng, Z.A., Peng, Z.: J. Am. Chem. Soc. **123**(7), 1389–1395 (2001)
6. Sugimoto, T.: Adv. Colloid Interface Sci. **28**, 65–108 (1987)
7. Talapin, D.V., Rogach, A.L., Haase, M., Weller, H.: J. Phys. Chem. B **105**(49), 12278–12285 (2001)
8. Viswanatha, R., Sarma, D.D.: Growth of nanocrystals in solution. In: Nanomaterials Chemistry: Recent Developments and New Directions, pp. 139–170. Wiley Online Library, Hoboken (2007)

Minisymposium: New Developments in Models of Traffic and Crowds

Poul G. Hjorth and Mads Peter Sørensen

Description

Modern society is increasingly faced with problems arising from overcrowded and congested motorways, and control of large crowds (sometimes failing, resulting in casualties, even loss of life). Over-engineering or oversizing available space is not always an option, and there is a need for mathematically sophisticated solutions arising from dynamical systems modelling.

Recent studies have produced a number of detailed models for the many-body interactions that take place both in traffic flows (where vehicles adjust their motion and position of other vehicles) and in flows of crowds of pedestrians (where pedestrians adjust their motion to the motion and position of the surrounding crowd). Classic and new mathematical techniques are brought to bear to study and describe the evolution dynamics: bifurcation theory, PDE and ODE reduction models, and equation-free analysis. The minisymposium explored current models, and models in development.

The talks included in the minisymposium were the following:

- N. Forcadel, INSA de Rouen, France, *Justification of macroscopic traffic flow model by specified homogenization of microscopic models.*
- N. García-Chan, University of Guadalajara, Mexico, *Numerical Simulation of the Urban Pollution Related to Traffic Flow.*
- J.-G. Caputo, INSA de Rouen, France, *A General Microscopic Traffic Model Yielding Dissipative Shocks.*

P.G. Hjorth (✉) • M.P. Sørensen
Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Pedersens Plads B-324, DK-2800 Kgs Lyngby, Denmark
e-mail: pghj@dtu.dk; mps@dtu.dk

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_57

- P. G. Hjorth, Technical University of Denmark. *Pedestrial Dynamics from Social Force Models.*
- J. Starke, Queen Mary University of London, UK, *Analysis of Unstable Pedestrian Flows*
- M. P. Sørensen, Technical University of Denmark, *Emerging Traffic Jam Patterns in the Optimal Velocity Model for Traffic Flow.*

Numerical Simulation for Evaluating the Effect of Traffic Restrictions on Urban Air Pollution

Néstor García-Chan, Lino J. Alvarez-Vázquez, Aurea Martínez, and Miguel E. Vázquez-Méndez

Abstract Traffic flow is the main pollution source in many urban areas, so when the pollutant concentration is above the permitted level, a common public policy consists of restricting the vehicular traffic. However, this restriction presents a negative impact on the economic activities and the inhabitants mobility. For these reasons different sectors question its efficiency arguing that the pollution reduction is due to wind changes and not to a lower vehicular traffic. In this work we estimate the pollution emission rate and pollutant concentration with a novel methodology that consists of combining the 1D Lighthill-Whitham-Richards (LWR) traffic model for road networks with a 2D advection-diffusion-reaction model for air pollution. This allows us to verify the efficiency of restrictions on vehicular traffic in the framework of numerical simulations in a real urban domain: the Guadalajara Metropolitan Area in Mexico.

1 The Mathematical Model

We consider an urban domain $\Omega \subset \mathbb{R}^2$ including a road network composed of N_R unidirectional avenues (segments) that meet at a number N_J of junctions, such that the endpoints of each segment are either on the boundary of Ω or touch one of the

N. García-Chan (✉)
Departamento de Física, C.U. Ciencias Exactas e Ingenierías, Universidad de Guadalajara, 44430 Guadalajara, Mexico
e-mail: nestor.gchan@academicos.udg.mx

L.J. Alvarez-Vázquez • A. Martínez
Departamento de Matemática Aplicada II, E.I. Telecomunicación, Universidade de Vigo, 36310 Vigo, Spain
e-mail: lino@dma.uvigo.es; aurea@dma.uvigo.es

M.E. Vázquez-Méndez
Departamento de Matemática Aplicada, E.P.S., Universidade de Santiago de Compostela, 27002 Lugo, Spain
e-mail: miguelernesto.vazquez@usc.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_58

junctions. Each segment $A_i \subset \Omega,\ i = 1, \ldots, N_R$, is modelled by an interval $[a_i, b_i]$, and we denote by:

$$\begin{aligned} \sigma_i : [a_i, b_i] &\longrightarrow A_i \subset \Omega \\ s &\longmapsto \sigma_i(s) = (x_i(s), y_i(s)) \end{aligned} \tag{1}$$

a parametrization of the segment A_i preserving the sense of motion on the avenue. For simulating the pollution due to traffic flow, we combine a 1D LWR traffic model with a 2D air pollution model. This combined model is explained in the following subsections.

1.1 A Traffic Flow Model of LWR Type

In order to model the traffic flow in the road network we consider the conservation law formulation proposed by Lighthill and Whitham [5], and by Richards [8], the so-called LRW model. Then, we denote by $\rho_i(s,t) \in [0, \rho_i^{max}]$ the density of cars in the avenue A_i, by $v_i(s,t)$ the velocity, and by $Q_i(s,t) = \rho_i(s,t)v_i(s,t)$ the corresponding flux.

This model is based on assuming that the velocity v_i depends exclusively on the density ρ_i, and, consequently, there exist a function $f_i : [0, \rho_i^{max}] \to \mathbb{R}$, the so-called static relation, such that $Q_i = f_i(\rho_i)$. Thus, the conservation law of the number of vehicles is described by the following system of partial differential equations: for $i = 1, \ldots, N_R$,

$$\frac{\partial \rho_i}{\partial t} + \frac{\partial f_i(\rho_i)}{\partial s} = 0 \qquad \text{in } (a_i, b_i) \times (0, T), \tag{2a}$$

$$\rho_i(., 0) = \rho_i^0 \qquad \text{in } [a_i, b_i], \tag{2b}$$

$$\rho_k(a_k, .) = \rho_k^{in} \qquad \text{in } (0, T), \tag{2c}$$

$$\rho_l(b_l, .) = \rho_l^{out} \qquad \text{in } (0, T), \tag{2d}$$

where ρ_i^0 is the initial vehicles density, and ρ_k^{in} (respectively, ρ_l^{out}) is the inlet (respectively, the outlet) density condition for the incoming avenue k (respectively, the outgoing avenue l). The system should be completed with the conservation of the number of cars at each junction $j = 1, \ldots, N_J$, the so-called Rankine-Hugoniot relation (see [2, 4]),

$$\sum_{k=1}^{n_j} f_{v_j(k)}(\rho_{v_j(k)}(b_{v_j(k)}, .)) = \sum_{l=n_j+1}^{n_j+m_j} f_{v_j(l)}(\rho_{v_j(l)}(a_{v_j(l)}, .)) \qquad \text{in } (0, T), \tag{2e}$$

where n_j and m_j are, respectively, the number of incoming and outgoing avenues, and we assume that the incoming roads are $A_{v_j(1)}, \ldots, A_{v_j(n_j)}$, and that the outgoing ones are $A_{v_j(n_j+1)}, \ldots, A_{v_j(n_j+m_j)}$.

The initial/boundary value problem (2a)–(2e) is not yet unequivocally determined. To determine it in an unique way, we also assume the following rules over the junctions (see [2]):

R1 We know some preferences of drivers, in the sense that the traffic from incoming avenues is distributed on outgoing avenues according to fixed and normalized coefficients.

R2 Respecting rule R1, drivers make their choices so that the traffic flux is maximized.

1.2 A 2D Air Pollution Model

The concentration $\phi(x,t)$ of an air pollution indicator (for instance, carbon monoxide CO) in a domain $\Omega \subset \mathbb{R}^2$ and over a time interval $(0,T)$, is given by the following initial/boundary value problem (see, for instance, [3, 9]):

$$\begin{cases} \dfrac{\partial \phi}{\partial t} + \mathbf{u}\cdot\nabla\phi - \nabla\cdot(\mu\nabla\phi) + \kappa\phi = F & \text{in } \Omega\times(0,T), \\ \phi(.,0) = \phi^0 & \text{in } \Omega, \\ \mu\dfrac{\partial \phi}{\partial n} - \phi\,\mathbf{u}\cdot\mathbf{n} = 0 & \text{on } S^-, \\ \mu\dfrac{\partial \phi}{\partial n} = 0 & \text{on } S^+, \end{cases} \tag{3}$$

where $\mathbf{u}(x,t)$ is the wind velocity field, $\mu(x,t)$ is the CO molecular diffusion coefficient, $\kappa(x,t)$ is the rate of the (first order) reaction term, F represents the source of contaminant which, as discussed below, will be given by (4), ϕ^0 is a known function giving the initial concentration of CO, $\mathbf{n}$ denotes the unit outward normal vector to the boundary $\partial\Omega$, and we write $\partial\Omega = S^- \bigcup S^+$ where $S^- = \{(x,t) \in \partial\Omega \times (0,T) \text{ such that } \mathbf{u}\cdot\mathbf{n} < 0\}$ represents the inflow boundary and $S^+ = \{(x,t) \in \partial\Omega \times (0,T) \text{ such that } \mathbf{u}\cdot\mathbf{n} \geq 0\}$ represents the outflow boundary.

In this work, we will assume that vehicle emissions are proportional to the traffic flow given by the solution of model (2a)–(2e). Implicitly, this means considering avenues as curves in Ω (sets of measure zero) and, consequently, emissions on the avenues should be included in the model (3) through measures on these sets [6]. Thus, we propose to model the source of pollution due to vehicular traffic by

$$F = \sum_{i=1}^{N_R} \gamma_i\, \xi_{A_i}, \tag{4}$$

where γ_i is a contamination rate, and, for each $t \in [0, T]$, $\xi_{A_i}(t)$ is the Radon measure (an element of the dual of the space of continuous functions) given by

$$\begin{aligned} \xi_{A_i}(t) : \mathscr{C}(\overline{\Omega}) &\longrightarrow \mathbb{R} \\ v &\longmapsto \langle \xi_{A_i}(t), v \rangle = \int_{A_i} Q_{A_i}(., t)\, v(.)\, d\sigma, \end{aligned}$$

where $Q_{A_i}(., t)$ represents the vehicular flux on the avenue A_i. For the parametrization given by (1), we have $Q_{A_i}(\sigma_i(s), t) = f_i(\rho_i(s, t))$ with ρ_i the solution of model (2a)–(2e). In this case, and assuming the functions to be smooth enough, we have that

$$\langle \xi_{A_i}(t), v \rangle = \int_{a_i}^{b_i} f_i(\rho_i(s, t))\, v(\sigma_i(s))\, \|\sigma_i'(s)\|\, ds, \quad \forall v \in \mathscr{C}(\overline{\Omega}). \tag{5}$$

2 Numerical Experiments

As an example of urban domain Ω we chose the Guadalajara Metropolitan Area (GMA), the second largest metropolitan zone in Mexico, with a population of 4.5 million inhabitants, two millions of vehicles, and pollution levels above permitted limits (see [1, 7] and references therein). In Ω we have taken, as a first approach to the problem, a road network formed by $N_R = 12$ unidirectional avenues, with $N_J = 6$ junctions, three outlet-network roads (A_6, A_9, A_{12}) and two inlet-network roads (A_1, A_2) (see Fig. 1 left).

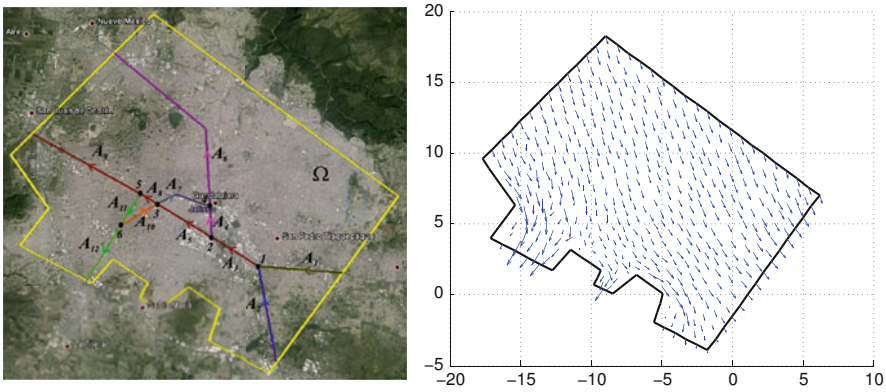


Fig. 1 Satellite photo of the Guadalajara Metropolitan Area, where we show the domain Ω for the case study and the road network under consideration (*left*). Velocity field corresponding to a wind with characteristic SW direction (*right*)

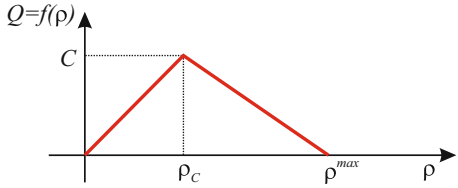


Fig. 2 Triangular Fundamental Diagram (TFD): function f providing the theoretical flow on each avenue, where the maximum is achieved at a critical density ρ_c

2.1 Input Data for the Mathematical Model

To solve the system (2a)–(2e) on the road network under consideration, first we need to set the static relations (functions $f_i(\rho_i)$), providing the theoretical flow on each avenue. In this paper we use the well-known Triangular Fundamental Diagram (TFD), which is a piecewise and triangular shape function (see Fig. 2) defined from some empirical parameters (see [11] for more details).

In the experiments shown here, we consider that the twelve avenues present a similar behaviour, so we admit that $f_1 = \ldots = f_{12}$ are given as in Fig. 2, for a critical density $\rho_C = 40.26\, u/km$, and a capacity of $C = 2013\, u/h$. These values correspond to the following experimental data: desired velocity $V^0 = 50\, km/h$, time gap between vehicles $T^{gap} = 3.3\, 10^{-4}\, h$, number of lines $I^* = 1$, and $\rho^{max} = 120\, u/km$.

On the other hand, for the air pollution model, we take $\mu = 3.5\, 10^{-8}\, km^2/h$, corresponding to an aerosol as CO, and $\kappa = 0.6\, 10^{-2}/h$, which is equivalent to 1 week of atmospheric residence of the pollutant (see [10]). We also consider $\gamma_i = 10^{-6}\, kg/(u\, km)$, no changes on meteorological conditions with a constant characteristic SW wind (see Fig. 1 right) and, for better interpretation of results, a null CO initial concentration ($\phi^0 = 0$).

2.2 Effect of Traffic Restrictions on Air Pollution

We assume a null density as initial condition in the network ($\rho_i^0 = 0,\ i = 1, \ldots, 12$), admit free traffic downstream in the outlet-network roads, and simulate the public policy of restricting the vehicular traffic by imposing different inlet boundary conditions to the LWR model.

So, we assume that the inlet boundary density is the same in both inlet-network avenues ($\rho_1^{in}(t) = \rho_2^{in}(t)$) and we suppose that it is a 24 h periodic function of wave type, with maximum values on peaks hours and minimum values in valley hours. Moreover, we consider two different wave types, a full-traffic day and a restricted-traffic day, the latter one with half the traffic of the former one (see Fig. 3 up).

Under above conditions we obtain the solution of the LWR model for two cases: during 45 days of full-traffic, and during 30 days of full-traffic plus 15 days of restricted-traffic. The output of both cases is then used for implementing the 2D

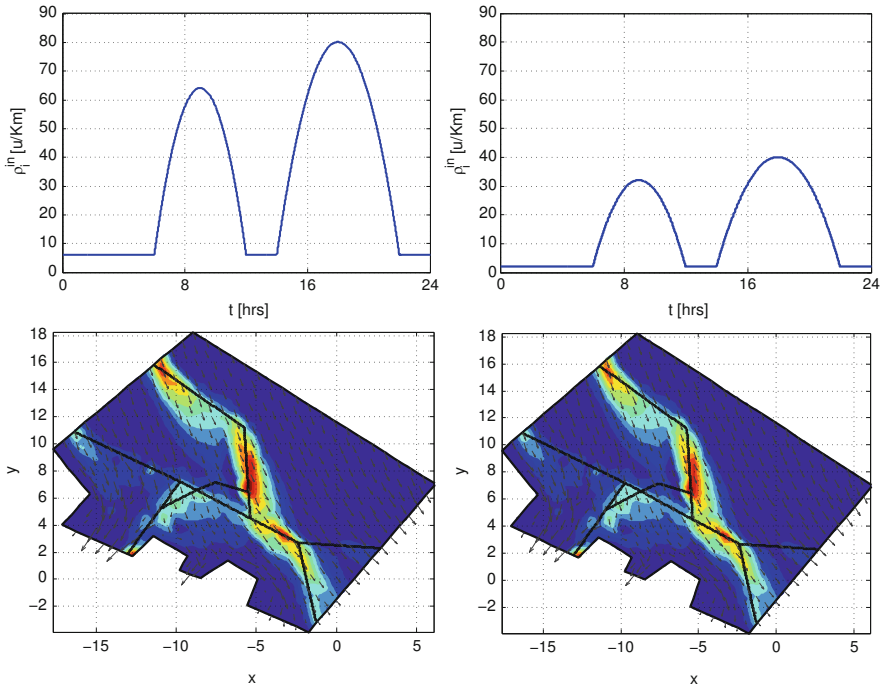


Fig. 3 Boundary conditions for a full-traffic day (*left up*) and for a restricted-traffic day (*right up*). Filled isolines of pollution over the domain Ω, corresponding to 45 days of full-traffic (*left down*) and to 30 days of full-traffic plus 15 days of restricted-traffic (*right down*)

air pollution model, obtaining the pollutant distribution (see Fig. 3 down) and the corresponding net pollutant concentration (via numerical integration). The results of the experiments show that, in this very simplified case, the policy of restricting the vehicular traffic only reaches a decrease of the net pollutant concentration of about 3% (more particular details can be seen by a comparison of the graphics shown in Fig. 3).

Acknowledgements This work was supported by Project MTM2015-65570-P of M.E.C. (Spain) and FEDER. The first author also thanks the support from Sistema Nacional de Investigadores SNI-52768 and Programa para el Desarrollo Profesional Docente PRODEP/103.5/13/6219 (Mexico).

References

1. Benítez-García, S.E., Kanda, I., Wakamatsu, S., Okazaki, Y., Kawano, M.: Analysis of criteria air pollutant trends in three Mexican metropolitan areas. Atmosphere **5**, 806–829 (2014)
2. Coclite, G.M., Garavello, M., Piccoli, M.: Traffic flow on a road network. SIAM J. Math. Anal. **36**, 1862–1886 (2005)

3. García-Chan, N., Alvarez-Vázquez, L.J., Martínez, A., Vázquez-Méndez, M.E.: On optimal location and management of a new industrial plant: numerical simulation and control. J. Frankl. Inst. Eng. Appl. Math. **351**, 1356–1371 (2014)
4. Holden, H., Risebro, H.: A mathematical model of traffic flow on a network of unidirectional roads. SIAM J. Math. Anal. **26**, 999–1017 (1995)
5. Lighthill, M.J., Whitham, G.B.: On kinetic waves II. Theory of traffic flows on long crowed roads. Proc. Roy. Soc. Lond. Ser. A. **229**, 317–345 (1955)
6. Nguyen, P.A., Raymond, J.P.: Control localized on thin structures for the linearized Boussinesq system. J. Optim. Theory Appl. **141**, 147–165 (2009)
7. Ramírez-Sánchez, H.U., Andrade-García, M.D., Bejaran, R., García-Guadalupe, M.E., Wallo-Vázquez, A., Pompa-Toledano, A.C., De la Torre-Villasenor, O.: The spatial-temporal distribution of the atmospheric polluting agents during the period 2000–2005 in the Urban Area of Guadalajara, Jalisco, Mexico. J. Hazard. Mater. **165**, 1128–1141 (2009)
8. Richards, P.I.: Shock waves on the highway. Oper. Res. **4**, 42–51 (1956)
9. Skiba, Y.N., Parra-Guevara, D.: Control of emission rates. Atmosfera **26**, 379–400 (2013)
10. Sportisse, B.: Fundamentals in Air Pollution: From Processes to Modelling. Springer, New York (2010)
11. Treiber, M., Kesting, A.: Traffic Flow Dynamics: Data, Models and Simulation. Springer, New York (2013)

A General Microscopic Traffic Model Yielding Dissipative Shocks

Yuri Borissovich Gaididei, Jean-Guy Caputo, Peter Leth Christiansen, Jens Juul Rasmussen, and Mads Peter Sørensen

Abstract We consider a general microscopic traffic model with a delay. An algebraic traffic function reduces the equation to the Aw-Rascle microscopic model while a sigmoid function gives the standard "follow the leader". For zero delay we prove that the homogeneous solution is globally stable. For a positive delay, it becomes unstable and develops dispersive and dissipative shocks. These are followed by a finite time singularity for the algebraic traffic function and by kinks for the sigmoid function.

1 Introduction

There has been a recent regain of interest in system of conservation laws to describe road traffic and crowd phenomena [1]. Such crowding phenomena also happen in bio-polymerization [2] and biology, see for example the review [3]. The traffic models have in general asymmetric coupling so that there is a preferred direction; a driver is only sensitive to cars ahead of him. Also, these models are all microscopic and few relations are established with a macroscopic model. One example is the Aw-Rascle traffic model [4], which was introduced as a microscopic version of their continuum model based on two conservation laws.

Y.B. Gaididei (✉)
Bogolyubov Institute for Theoretical Physics, Metrologichna str. 14B, 01413 Kiev, Ukraine
e-mail: ybg@bitp.kiev.ua

J.-G. Caputo
Laboratoire de Mathématiques, INSA de Rouen, 76801 Saint-Etienne du Rouvray, France
e-mail: caputo@insa-rouen.fr

P.L. Christiansen • J.J. Rasmussen
Department of Physics, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
e-mail: p.l.christiansen@mail.tele.dk; jjra@fysik.dtu.dk

M.P. Sørensen
Department of Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
e-mail: mpso@dtu.dk

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_59

Here, inspired by the microscopic model of [4], we generalize it by including delay effects and examine its dynamics. This equation reduces exactly to the "follow the leader" model we studied in [5] for a sigmoid traffic function. The Aw-Rascle model [4] corresponds to an algebraic traffic function. For both models, we show that when the delay is zero, the only solution is homogeneous so that the cars are equally spaced. For the Aw-Rascle system and for a non zero delay, we obtain a traveling wave solution formally but we could not observe it numerically. To elucidate this we examine the linear stability of the model for both traffic functions and show the influence of the different parameters. The continuum limit of our model is a Korteweg-De-Vries-Burgers (KdV-Burgers) second order partial differential equation with dissipation and dispersion. It is not equivalent to two conservation laws. The numerics confirm the dispersive KdV-Burgers dynamics in the stable and unstable case. For the algebraic traffic function, we get Burgers shocks for small delay τ, while for large τ the solution u touches 0 causing a singularity in finite time. For a sigmoid traffic function, we get again Burgers shocks for small τ. Large τ give kinks that become static as time $t \to \infty$.

2 A Generalized Discrete Traffic Model

We consider $n = 1, \ldots, N$ cars in a circle, following a discrete generalized Aw-Rascle model

$$\dot{x}_n = v_n, \qquad \dot{v}_n(t+\tau) = (v_{n+1} - v_n)\, F'(x_{n+1} - x_n) + \epsilon(c - v_n). \tag{1}$$

We included a time delay, τ, to model the reaction time of the drivers. The term in ϵ mimics the optimal velocity models because it tends to have cars move with a common velocity c. It is also a realistic term because on many highways the velocity is set.

The system above can be written as

$$\ddot{x}_n(t+\tau) = \frac{d}{dt} F(x_{n+1} - x_n) + \epsilon\,(c - \dot{x}_n). \tag{2}$$

By subtracting Eqs. (2) for $n+1$ and n, we get the evolution of the headway $u_n \equiv x_{n+1} - x_n$

$$\ddot{u}_n(t+\tau) = \frac{d}{dt}\Big(F(u_{n+1}) - F(u_n)\Big) - \epsilon\, \dot{u}_n\ . \tag{3}$$

Integrating (3),we get

$$\dot{u}_n(t+\tau) = F(u_{n+1}) - F(u_n) - \epsilon\,(u_n - \ell)\ , \tag{4}$$

where the constant $\ell = L/N$ is a mean distance between cars. It is determined from the condition

$$\sum_{n=1}^{N} u_n = L \,, \tag{5}$$

where L is the length of the road. We shall assume that the delay time τ is short. Then we expand the function $u_n(t+\tau)$ in a Taylor series in (3) to get the second order differential equation

$$\tau \ddot{u}_n(t) + \dot{u}_n(t) = F(u_{n+1}) - F(u_n) - \epsilon\,(u_n - \ell) \,. \tag{6}$$

This is the main equation that we shall analyze in this article for the Aw-Rascle and sigmoid traffic functions

$$F_1(u) = C(1 - (\frac{u}{h})^{-\gamma}), \qquad F_2(u) = C\tanh(u-h), \tag{7}$$

where $C > 0$, $\gamma > 0$ and h is the safety distance.

3 Stationary States and Traveling Waves

For $\tau = 0$ we show, for F_1 and F_2, using a Lyapunov function based on a Hamiltonian representation of the system that the only stationary states that exist for $\tau = 0$ are the homogeneous states $u_n = \ell$. For that, we use the variables $w_n = F(u_n)$, $u_n = G(w_n)$ and write (6) as

$$G'(w_n)\dot{w}_n - \Delta_1 w_n = \Delta_2 w_n - \epsilon\left(G(w_n) - \ell\right) - \tau\,\frac{d^2}{dt^2}G(w_n), \tag{8}$$

where $\Delta_1 w_n = (w_{n+1} - w_{n-1})/2$, $\Delta_2 w_n = (w_{n+1} + w_{n-1} - 2w_n)/2$. The left hand side can be derived from a Hamiltonian and the right hand side is dissipative. The Lyapunov function is built from this Hamiltonian. This result means that only the spatially homogeneous stationary solution $u_n = \ell$ can exist and that traffic jams, i.e. non homogeneous solutions can only exist as transient waves caused by a specific initial condition or boundary condition.

For $\tau \neq 0$, in our previous work [5], we found a traveling wave solution for the sigmoid traffic function F_2. The procedure, successful for F_2, can be implemented for F_1 as well and yields formally a cnoidal wave solution. We could not find this formal solution numerically for the algebraic traffic function F_1; all unstable initial conditions lead to a singularity.

4 Linear Stability Analysis

To investigate the stability of the uniform flow solution, $u_n(t) = \ell$, for $n = 1, 2, \ldots, N$, another natural step is to do a linear analysis. This also allows to see the action of the different parameters. We assume that $u_n(t) = \ell + U_n(t)$ and linearize Eqs. (6) with respect to $U_n(t)$. We obtain

$$\tau \ddot{U}_n + \dot{U}_n + \epsilon\, U_n = F'(\ell)\ (U_{n+1} - U_n)\,. \tag{9}$$

The stability is analyzed by considering solutions of (9) of the form $U_n(t) = \hat{U}_k\ \exp\{ik\,n + \sigma\, t\}, \qquad k = \frac{2\pi\, j}{N}, \qquad j = -\frac{N}{2} + 1, \ldots, \frac{N}{2}$. where σ is a complex frequency. If $Re(\sigma) > 0$ the solution is unstable. Inserting the expression above into Eq. (9) leads to the second degree equation $\tau\, \sigma^2 + \sigma + \epsilon - F'(\ell)(e^{ik} - 1) = 0$.

If $\tau = 0$, then $Re(\sigma) = -\epsilon + F'(\ell)(\cos k - 1) < 0$ for $F'(\ell) > 0$ so that there is no instability; we recover the result of the previous section.

If $2\,\tau\, F'(\ell)) > 1$, then there exists a band of wave numbers $k \in (k_-, k_+)$ such that $R = Re(\sigma_k) > 0$. The threshold values $k_\pm$ are given by

$$\sin^2\left(\frac{k_\pm}{2}\right) = \frac{2\,\tau\, F'(\ell) - 1 \pm \sqrt{(2\,\tau\, F'(\ell) - 1)^2 - 4\,\tau\,\epsilon}}{4\,\tau\, F'(\ell)}. \tag{10}$$

4.1 *Influence of the Parameters τ, ϵ and h*

The expression (10) shows how the different parameters $\tau,\ \epsilon,\ h$ affect the stability of the homogeneous state. Consider the effect of τ. The growth rate $Re(\sigma_k)$ is plotted in the left panel of Fig. 1 as a function of k for different τs. Clearly larger τs increase the instability range; however the rate saturates for large τ and remains about 0.2.

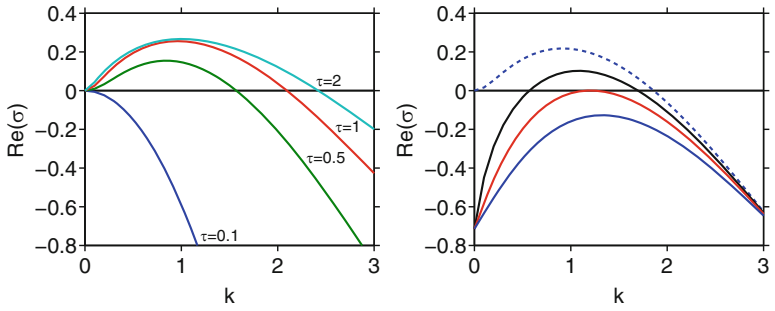


Fig. 1 Plot of the instability rate $Re(\sigma)$ as a function of the wave number $k = \frac{2\pi j}{N}$ for the Aw-Rascle model using $F_1(u)$ in Eq. (7). The *left panel* shows the effect of the delay τ on $Re(\sigma)$ for $\tau = 0.1,\ 0.5,\ 1$ and 2 using $\epsilon = 0$. On the *right panel* we consider different values of ϵ, $\epsilon = 0$ (*dashed line*), $\epsilon = 0.5$ (*black continuous*) $\epsilon = \epsilon_c = 1.15714$ (*red online*) and $\epsilon = 2.5$ (*blue online*) using $\tau = 0.7$. The other parameters are: $\gamma = 1$, $h = 1$, $C = 2$, $\ell = 1$

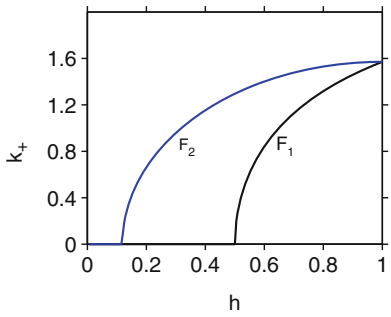


Fig. 2 Effect of the safety distance h: maximum k_+ of the instability region as a function of the safety distance h. The parameters are $C = 1,\ \ell = 1, \tau = 1, \gamma = 1$ and $\epsilon = 0$

The parameter ϵ shifts the instability region towards smaller wave-lengths and reduces the instability rate as shown in the right panel of Fig. 1. Finally we consider how the safety distance h affects the stability. Here the specific form of F_1 and F_2 changes the result (Fig. 2). We see that for $h > 0.1\ell$ solutions are unstable for the sigmoid function F_2. For the algebraic one F_1 we need $h > 0.5\ell$ to observe an instability.

5 Continuum Limit Near $k = 0$: KdV-Burgers Equation

The linear stability results show that the range of unstable numbers is close to $k = 0$. It is then useful to look at the continuum limit of the model (8). For that we write $w_n(t) = w(\mathrm{s}, t)$ where s is a continuum variable. We take $\epsilon = 0$. Equation (8) has a spatially homogeneous stationary solution $w(s) = W \equiv F(\ell)$. We assume that $w(s,t) = W + z(s,t),\ |z(s,t)| \ll W$, and substitute this expression into Eq. (8) to get

$$G'(W+z)z_t = z(s+1) - z(s) - \tau\,(G(W+z))_{tt}, \tag{11}$$

where the subscripts denote partial derivatives. The difference term can be expanded as $z(s+1) - z(s) \approx z_s + \frac{1}{2}z_{ss} + \frac{1}{6}z_{3s}$. Subscript s denotes derivative of $z(s)$ with respect to s. We expand $G"(W+z),\ G'(W+z)$ and $G(W+z)$ in Taylor series and obtain the final equation

$$(G' + G''\,z)z_t = z_s + \frac{1}{2}\,z_{ss} + \frac{1}{6}\,z_{sss} - \tau\,(G' + zG")\,z_{tt} - \tau\,(G" + zG^{(3)}))\,(z_t)^2, \tag{12}$$

where we have omitted the dependencies for clarity, and where G, G', G'' and $G^{(3)}$ are evaluated at W. In (12), one sees that at leading order $G'(W)z_t = z_s$ so that the term zz_t can be replaced by zz_s/G'. The Eq. (12) reduces then to a perturbed KdV Burgers equation [6]

$$G'z_t - z_s = -\frac{G''}{G'}\,zz_s + \frac{1}{2}\,z_{ss} + \frac{1}{6}\,z_{sss} - \tau\,(G' + zG")\,z_{tt} - \tau\,(G" + zG^{(3)})\,(z_t)^2, \tag{13}$$

Equation (13) has solutions which are dispersive and dissipative shocks. A sine initial condition for example, will sharpen into a front as for the Burgers equation, then a dispersive wave appears following KdV and gradually damps out due to the nonlinearity.

6 Numerical Results

To verify and illustrate the analysis above, we solved (6) for the two different traffic functions F_1 and F_2. We use a long wave initial condition as in [5]

$$u_n = \ell + 0.05 \sin(\frac{2\pi}{N} n). \tag{14}$$

Our main finding is that above threshold, the solutions with F_1 develop a singularity in finite time because u_n reaches 0 at some n. The solutions with F_2 tend to kinks as was evidenced in our previous article [5].

To illustrate the singular behavior observed for F_1, the Aw-Rascle model, we plot in Fig. 3 the initial conditions in (14) and $u_n(t)$ for the three different times

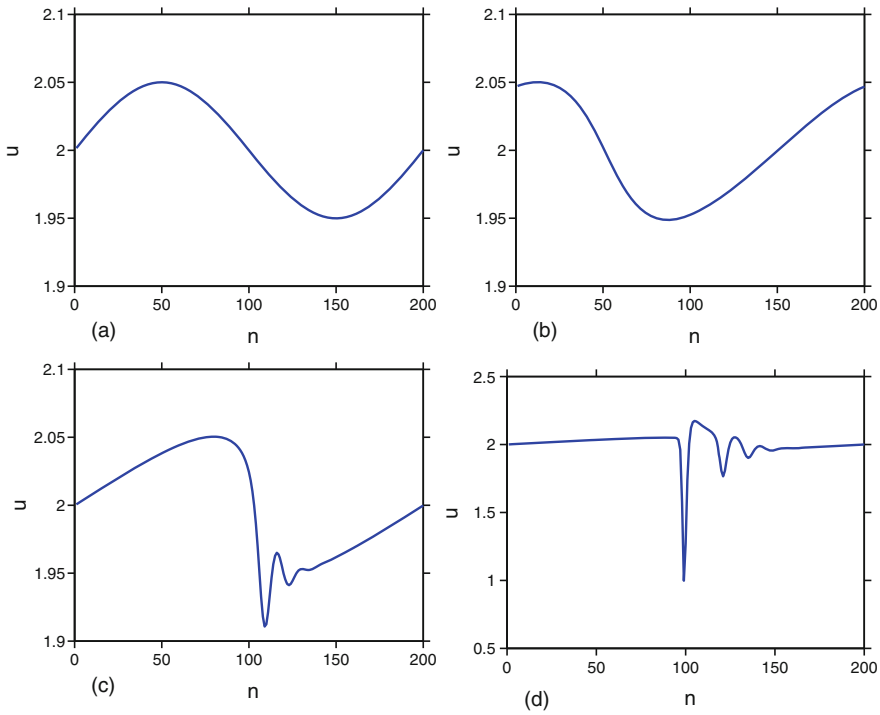


Fig. 3 Simulation results for Eq. (6) using the Aw-Rascle model $F_1(u) = C(1 - 1/u)$. Snapshots of $u_n(t)$ for (**a**) $t = 0$ (initial conditions), (**b**) $t = 500$, (**c**) $t = 1200$ and (**d**) $t = 1600$. The parameters are $N = 200$, $\tau = 1.1$, $\epsilon = 0$, $C = 2$, $\ell = 2$, $h = 1$ and $\gamma = 1$

$t = 500,\ 1200$ and 1600. The parameters are $N = 200,\ \tau = 1.1,\ \epsilon = 0.,\ C = 2,\ h = 1,\ \ell = 2$ and $\gamma = 1$. The instability develops fairly slowly, in accordance with the growth rate because $k = \frac{2\pi}{N} \approx 0.031$ is close to 0. Of course, as soon as the singularity is reached, the code breaks down. The wave propagation is from right to left and due to the form of the nonlinearity the wave pulse steepens on the back side, this is particularly evident on Fig. 3c. The dispersive and dissipative terms in Eq. (13) cannot prevent the steepening to continue and a singularity develops.

In Fig. 4 we show the computed time evolution of u_n for the second traffic function F_2, that is the sigmoid function. For the same initial condition (14), the solution develops into kinks where large u regions alternate with small u regions. These kinks become very sharp for large times due to the relatively small dissipation.

Again the wave propagates from left to right and the back steepens due to the nonlinearity. However, eventually the instability leads to a kink and an anti kink solution corresponding to the formation of two shock waves. The shock waves balance the steepening due to the nonlinearity and the dissipation/dispersion. The kinks have negative regions of the headway u_n for the parameters we use which are standard. This is non physical and care should be taken to tune the parameters

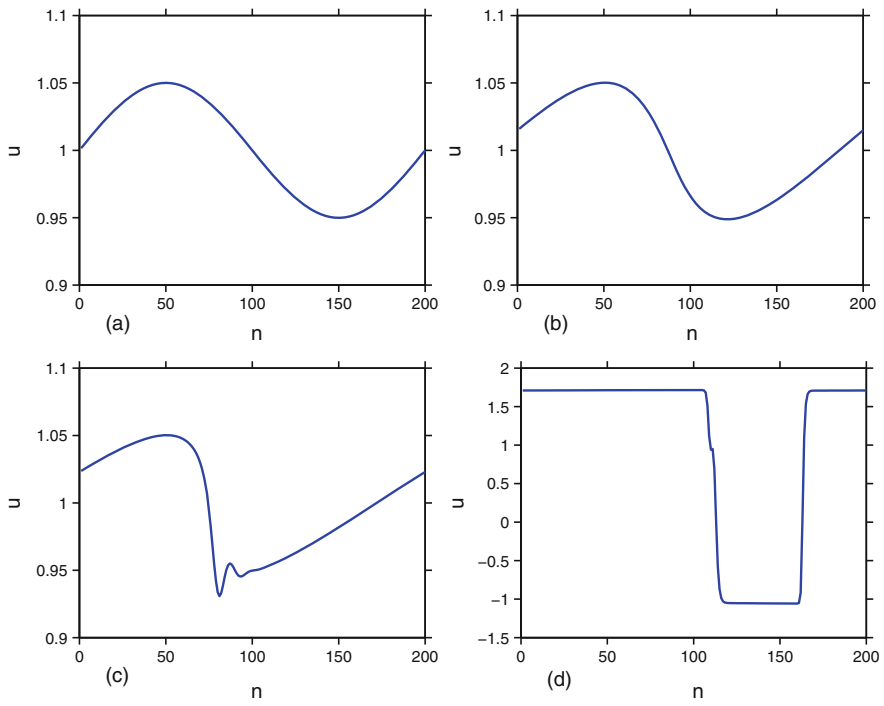


Fig. 4 Simulation results for Eq. (6) using the sigmoid model $F_2(u) = C\tanh(u-h)$. Snapshots of $u_n(t)$ for (**a**) $t = 0$ (initial conditions), (**b**) $t = 383$, (**c**) $t = 766$ and (**d**) $t = 1609$. The parameters are $N = 200$, $\tau = 1$, $\epsilon = 0$, $C = 1$, $\ell = 1$, $h = 0.2$ and $\gamma = 1$

to prevent this. For this, parameters should be chosen carefully and if possible compared to real road situations like [7].

Acknowledgements Yu.G. acknowledges a Guest Professorship funded by Civilingeniør Frederik Christiansens Almennyttige Fond and partial financial support from the National Academy of Sciences of Ukraine. He thanks the Department of Physics, Technical University of Denmark for its hospitality. J.G. C. received support from the Region of Normandy through the program Xterm.

References

1. Bellomo, N., Dogbé, C.: On the modeling of traffic and crowds: a survey of models, speculations, and perspectives. SIAM Rev. **53**(3), 409–463 (2011)
2. Macdonald, C.T., Gibbs, J.H.: Kinetics of biopolymerization on nucleic acid templates. Biopolymers **6**, 1–25 (1968)
3. Chowdhury, D., Schadschneider, A., Nishinari, K.: Physics of transport and traffic phenomena in biology: from molecular motors and cells to organisms. Phys. Life Rev. **2**, 318–352 (2005)
4. Aw, A., Klar, A., Materne, T., Rascle, N.: Derivation of continuum traffic flow models from miscroscopic follow-the-leader models. SIAM J. Appl. Math. **63**, 259–278 (2002)
5. Gaididei, Y.B., Berkemer, R., Caputo, J.G., Christiansen, P.L., Kawamoto, A., Shiga, T., Sørensen, M.P., Starke, J.: Analytical solutions of jam pattern formation on a ring for a class of optimal velocity traffic models. New J. Phys. **11**, 073012–070331 (2009)
6. Whitham, G.B.: Linear and Nonlinear Waves. Wiley, New York (1974)
7. Nakayama, A., Kikuchi, M., Shibata, A., Sugiyama, Y., Tadaki, S., Yukawa, S.: Quantitative explanation of circuit experiments and real traffic using the optimal velocity model. New. J. Phys. **18**, 043040 (2016)

Minisymposium: Nonlinear Diffusion Processes: Cross Diffusion, Complex Diffusion and Related Topics

Adérito Araújo, Sílvia Barbeiro, Ángel Durán, and Eduardo Cuesta

Description

Nonlinear diffusion equations have attracted a lot of attention over the last few years in many practical applications. After the pioneering work of Keler and Segel in the 1970s, cross-diffusion models became very popular in biology, chemistry and physics to emulate systems with multiple species. The range of application is even wider and, in particular, complex-diffusion models are of special relevance for example in the field of image processing. Meanwhile the underlying mathematical theory has been developed in a synergetic way with applications, in recent years, this topic became the focus of an intensive research within the mathematics community. In spite of the relevance of cross-diffusion models in numerous fields of application and all the mathematical activity around them, important questions remain unanswered and meaningful challenging problems still need to be addressed.

The aim of this minisymposium was to bring together researchers from the various fields, combining expertise in the theory of PDEs and competences in numerical analysis with a problem oriented perspective. The main goals were to give new insights in understanding the intimate nature of nonlinear processes, study their numerical discretizations, and develop sophisticated mathematical and computational tools that could be used efficiently in real problems, with special regards to cross diffusion and complex diffusion models.

A. Araújo (✉) • S. Barbeiro
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
e-mail: alma@mat.uc.pt; silvia@mat.uc.pt

Á. Durán • E. Cuesta
Department of Applied Mathematics, University of Valladolid, Valladolid, Spain
e-mail: angel@mac.uva.es; eduardo@mat.uva.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_60

In this symposium we gathered scientists in pure and applied mathematics, computational science and physics with proven experience in the fields of computational biology and biomedical engineering, enhancing the visibility of recent developments in the topic of nonlinear cross-diffusion models.

The talks included in the minisymposium were the following:

- A. Madzvamuse. The University of Sussex, School of Mathematical and Physical Sciences, Department of Mathematics, Brighton (United Kingdom). *Characterising the effects of cross-diffusion for reaction-diffusion models on stationary and evolving domains and surfaces.*
- V. Selgas. Department of Mathematics, Universidad de Oviedo (Spain). *Analysis of a splitting-differentiation population model leading to cross-diffusion.*
- A. Araújo. CMUC, Department of Mathematics, University of Coimbra (Portugal). *A Cross-diffusion models for image processing I. Linear case.*
- A. Durán. Department of Applied Mathematics, University of Valladolid (Spain). *Cross-diffusion models for image processing II. Nonlinear case.*
- C. Totzeck. TU Kaiserslautern (Germany). *Consensus-based global optimization.*
- L. Plociniczak. Wrocław, University of Technology, Wrocław (Poland). *Anomalous nonlinear diffusion in porous media: analytical approximations.*

Cross-Diffusion in Reaction-Diffusion Models: Analysis, Numerics, and Applications

Anotida Madzvamuse, Raquel Barreira, and Alf Gerisch

Abstract Cross-diffusion terms are nowadays widely used in reaction-diffusion equations encountered in models from mathematical biology and in various engineering applications. In this contribution we review the basic model equations of such systems, give an overview of their mathematical analysis, with an emphasis on pattern formation and positivity preservation, and finally we present numerical simulations that highlight special features of reaction-cross-diffusion models.

1 Introduction

Recently there has been a surge in the analysis and simulation of mathematical models of reaction-diffusion type in the presence of so-called *cross-diffusion*. Cross-diffusion is a process in which the gradient in the concentration or density of one chemical or biological species induces a flux (either linear or nonlinear) of another species. This notion of cross-diffusion also includes the well-known cases of chemo- and haptotaxis modelling. Accordingly, the applications of reaction-cross-diffusion systems are abundant in the literature and include pattern forming in developmental biology [11], electrochemistry [3], cancer motility [5, 8, 12] and biofilms [15]. The introduction of cross-diffusion in standard reaction-diffusion models has been shown to prevent blow-up phenomena that are associated with such systems in the absence of cross-diffusion [9]. Explicit analytic solutions to these complex and

A. Madzvamuse (✉)
School of Mathematical and Physical Sciences, Department of Mathematics, University of Sussex, Pevensey 3 5c15, Brighton, BN1 9QH, UK
e-mail: a.madzvamuse@sussex.ac.uk

R. Barreira
Polytechnic Institute of Setubal, Barreiro School of Technology, Rua Américo da Silva Marinho-Lavradio, 2839-001 Barreiro, Portugal
e-mail: raquel.barreira@estbarreiro.ips.pt

A. Gerisch
TU Darmstadt, Fachbereich Mathematik, AG Numerik und Wissenschaftliches Rechnen, Dolivostr. 15, D-64293 Darmstadt, Germany
e-mail: gerisch@mathematik.tu-darmstadt.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_61

often nonlinearly coupled systems of partial differential equations do rarely exist and thus several numerical methods have been applied to provide approximate solutions. Such methods are not only available for fixed spatial domains but also for simulations on continuously evolving spatial domains and surfaces [11]. In this contribution we present briefly the model system under consideration and discuss some state-of-the art analytical results and numerical tools as well as provide numerical solutions for some particular models.

2 A Multiple Species Reaction-Cross-Diffusion Model

Let $\Omega \subset \mathbb{R}^m$, $m \in \{1, 2, 3\}$, be a simply connected bounded domain with $\partial\Omega \in C^{0,1}$. Moreover, let $\mathbf{u} = (u_1(\mathbf{x}, t), \cdots, u_n(\mathbf{x}, t))^T$ be a vector-valued function describing n species (chemical, biological, or otherwise) at position $\mathbf{x} \in \Omega$ and time $t \in I = [0, t_F]$, $t_F > 0$. The evolution equations for a reaction-cross-diffusion model can be obtained from the application of the law of mass conservation and are given by the following generalised non-dimensional system with zero-flux boundary conditions [11]

$$\partial_t u_i = \nabla \cdot \left(\sum_{j=1}^{n} D_{ij}(\mathbf{u}) \nabla u_j \right) + \gamma f_i(\mathbf{u}), \quad i = 1, \cdots, n, \quad \mathbf{x} \in \Omega\,,\ t > 0\,, \tag{1a}$$

$$\left(\sum_{j=1}^{n} D_{ij}(\mathbf{u}) \nabla u_j \right) \cdot \nu = 0, \qquad i = 1, \cdots, n, \quad \mathbf{x} \in \partial\Omega\,,\ t > 0\,, \tag{1b}$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega\,,\ t = 0. \tag{1c}$$

In this framework, $D_{ij}(\mathbf{u})$ is the constant, linear, or nonlinear diffusion coefficient relating the jth species gradient with the flux of the ith species. γ is a non-dimensional scaling parameter describing the relative strength of the reaction kinetics $f_i(\mathbf{u})$ [13].

2.1 *Analysis of Reaction-Cross-Diffusion Models*

The general framework (1) encompasses the well-studied (Patlak-) Keller-Segel [9], Armstrong-Painter-Sherratt [2], and the Shigesada-Kawasaky-Teramoto [17] models. In many of these models, $D_{ij}(\mathbf{u})$ is highly nonlinear which makes rigorous mathematical analysis of such models, in particular if involving multiple species, challenging [4, 9, 14]—even in the absence of nonlinear reaction kinetics $f_i(\mathbf{u})$.

Analysis of Linear and Nonlinear Cross-Diffusion Models Some recent studies of reaction-diffusion systems with linear cross-diffusion show that such models enhance pattern formation for self-organised processes [7, 11]. In these works, the

following two-species reaction-cross-diffusion system was studied

$$\partial_t u = \nabla\cdot(\nabla u + d_v \nabla v) + \gamma f(u,v), \qquad \partial_t v = \nabla\cdot(d\nabla v + d_u \nabla u) + \gamma g(u,v), \tag{2}$$

where for illustrative purposes an activator-depleted model [13] with $f(u,v) = a - u + u^2 v$ and $g(u,v) = b - u^2 v$ was considered. On application of the linear stability theory, the following necessary conditions for cross-diffusion-driven instability were obtained and these generalise classical Turing diffusion-driven instability conditions in the absence of cross-diffusion

$$f_u + g_v < 0,\ f_u g_v - f_v g_u > 0,\ d - d_u d_v > 0,\ df_u + g_v - d_u f_v - d_v g_u > 0, \tag{3}$$

$$(df_u + g_v - d_u f_v - d_v g_u)^2 - 4(d - d_u d_v)\,(f_u g_v - f_v g_u)) > 0. \tag{4}$$

The above generalisation implies that in the presence of cross-diffusion, a wide range of non-standard reaction-diffusion models can give rise to patterning that is induced by linear cross-diffusion. For example, activator-inhibitor, activator-activator, inhibitor-inhibitor kinetics can give rise to patterning either in the form of long-range inhibition, short-range activation (i.e., $d > 1$), or through short-range inhibition, long-range activation ($d < 1$), or through equal-range activation and inhibition ($d = 1$).

In Fig. 1 we exhibit Turing instability parameter regions for various diffusion and cross-diffusion coefficients: (a) in the absence of cross-diffusion, (b)-(e) positive

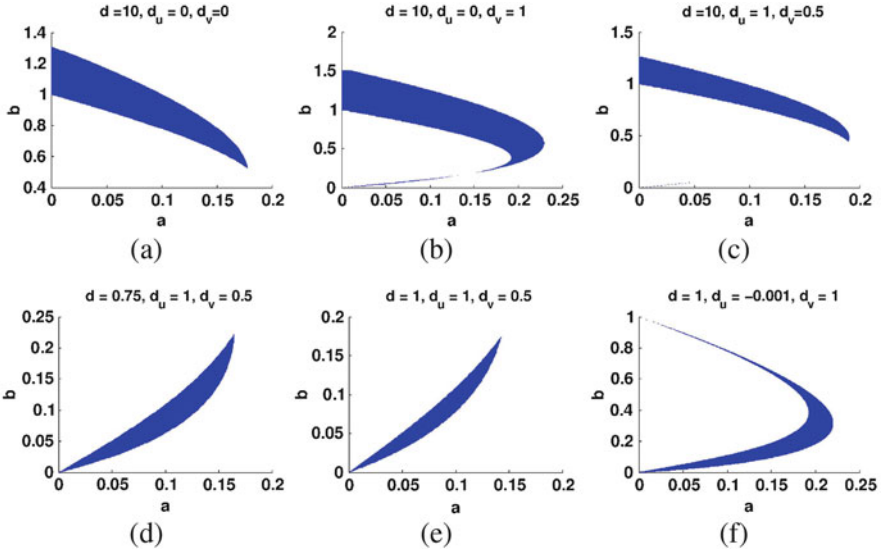


Fig. 1 Turing instability parameter regions for the reaction kinetic parameters (a, b) for the fixed diffusion and cross-diffusion parameters as given in each plot title. The blue region in each subplot is the region of Turing instability for (a, b)

cross-diffusion, and (f) negative cross-diffusion. We observe substantial changes in the range for instability in the (a, b)-parameter space as we introduce cross-diffusion, compare Fig. 1a–c. Cross-diffusion induces different parameter spaces. The non-empty parameter ranges for instability in Fig. 1d–e only exist due to the presence of cross-diffusion and they are empty in its absence.

A drawback associated with reaction-diffusion models with linear cross-diffusion is that in some cases these fail to reproduce experimental observations, hence it becomes imperative to study nonlinear cross-diffusion models [4, 7]. Methods for nonlinear analysis of such models include perturbation methods [14] and weakly nonlinear analysis [7]. Unlike linear stability analysis which studies short-time behaviour of the system close to bifurcation points, nonlinear analysis allows the study of long-time behaviour of solutions far away from bifurcation points.

Positivity of Solutions for Reaction-Cross-Diffusion Models In many applications, the solution $\mathbf{u}$ of a reaction-cross-diffusion model (1) represents species concentration or densities or other non-negative quantities. It is therefore necessary that the evolving solution $\mathbf{u}$, starting from any non-negative initial data $\mathbf{u}_0$, stays non-negative for all times. If that holds then the system is called *positivity-preserving*. For single specie models satisfying the condition $f_1(0) \geq 0$, the non-negativity of solutions is a consequence of the maximum principle [6, 16]. The situation is different and more involved in the case of systems. To this end, consider the simple case of a reaction-cross-diffusion system (1) with constant diffusion/cross-diffusion coefficients in Eq. (1a). A necessary condition for this system to be positivity-preserving is that the matrix D is diagonal, i.e. that all cross-diffusion terms vanish, see [18, Chap. 14] or [16]. That means for this system with non-vanishing cross-diffusion terms there exist non-negative initial data $\mathbf{u}_0$ such that its solution $\mathbf{u}$ does not stay non-negative. Intuitively, in the equation for u_1, if $D_{1,2}\Delta u_2 < 0$ in some part of Ω then this leads to a decrease in u_1 independent of its actual value and thus to negative u_1 values if u_1 is sufficiently small (see Fig. 2). As a consequence of this negative result, nonlinear cross-diffusion terms are required for positivity-preserving systems. We refer the reader to, for instance, [16] for corresponding conditions on D_{ij} and f_i.

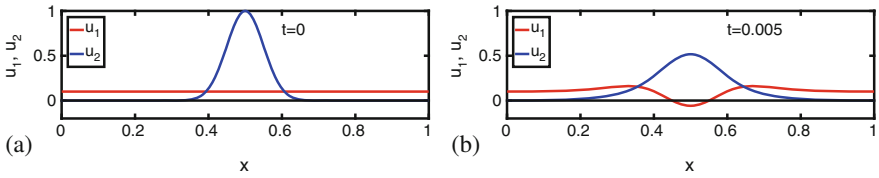


Fig. 2 (**a**) Initial data $u_1(x, 0) = 0.1$ and $u_2(x, 0) = (\frac{1}{2}(1 - \cos(2\pi x)))^{20}$ for system (1) with $D = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$, $\gamma = 1, f = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ in $\Omega = (0, 1)$. (**b**) Numerical solution of this system at $t = 0.005$

3 Numerical Methods for Solving Reaction-Cross-Diffusion Models

In many physical modelling cases, the choice of the numerical method depends crucially on the physical properties of the model system. Reaction-cross-diffusion systems typically fall under parabolic partial differential equations for which numerous numerical methods exist. The simplest and most frequently used schemes are based on finite differences, e.g., [4, 16]. For positivity-preserving models involving, for instance, taxis terms, finite volume spatial discretisations, incorporating some form of flux-limiting for ensuring positivity of the discretised system, have been studied and used extensively [8]. Mass lumping has been shown to preserve positivity for finite element-based methods which deal more naturally with complex stationary and evolving domains/manifolds/surfaces [6, 11]. Other numerical methods such as spectral and meshless methods can be employed with considerable difficulties in treating complex geometries and nonlinear boundary conditions.

Numerical Example 1: Pattern Formation In Fig. 3 we exhibit finite element solutions for the reaction-diffusion system with linear cross-diffusion for system (2) on planar domains and surfaces [10]. Spot and stripe patterns are observed.

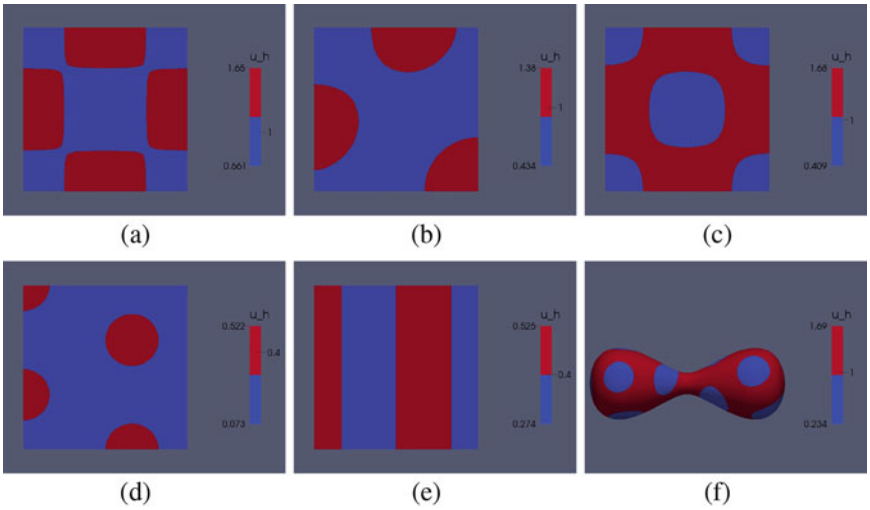


Fig. 3 Finite element simulations revealing the patterning with or without cross-diffusion for the model system (2) with model parameter values (**a**) $d = 10$, $d_u = 0$, $d_v = 0$, $a = 0.05$, $b = 1.1$; (**b**) $d = 10$, $d_u = 0$, $d_v = 1$, $a = 0.21$, $b = 0.26$; (**c**) $d = 10$, $d_u = 1$, $d_v = 0.5$, $a = 0.025$, $b = 1.1$; (**d**) $d = 0.75$, $d_u = 1$, $d_v = 0.5$, $a = 0.1$, $b = 0.08$; (**e**) $d = 1$, $d_u = 1$, $d_v = 0.5$, $a = 0.1$, $b = 0.08$; and (**f**) $d = 1$, $d_u = 1$, $d_v = 0.5$, $a = 0.1$, $b = 0.08$

Numerical Example 2: Cancer Invasion Our next example involves a cancer invasion model taken from [1, 5]. Here, the spatio-temporal interactions of two cancer cell populations, the extracellular matrix (ECM), and a matrix-degrading enzyme (MDE) are studied. Thus we define

$$\mathbf{u} = (\text{cell density } c_1, \text{cell density } c_2, \text{ECM density } v, \text{MDE concentration } m)\,.$$

The densities and concentrations are scaled such that the volume fraction of occupied space reads $\rho(\mathbf{u}) = c_1 + c_2 + v \in [0, 1]$; note that MDE does not take up space. The model *without* cross-diffusion terms reads

$$\partial_t c_1 = \nabla \cdot [D_{c,1}\nabla c_1 - c_1 \mathscr{A}_1(\mathbf{u}(t,\cdot))] + \mu_{c,1} c_1 (1 - \rho(\mathbf{u})) - M(t,\mathbf{u}) c_1\,, \tag{5a}$$

$$\partial_t c_2 = \nabla \cdot [D_{c,2}\nabla c_2 - c_2 \mathscr{A}_2(\mathbf{u}(t,\cdot))] + \mu_{c,2} c_2 (1 - \rho(\mathbf{u})) + M(t,\mathbf{u}) c_1\,, \tag{5b}$$

$$\partial_t v = -\gamma m v + \mu_2 (1 - \rho(\mathbf{u}))^+\,, \tag{5c}$$

$$\partial_t m = \nabla \cdot \left[D_m \nabla m\right] + \alpha_{c,1} c_1 + \alpha_{c,2} c_2 - \lambda m\,. \tag{5d}$$

In the above, $\mathscr{A}$ is a non-local operator modelling cell-cell and cell-matrix adhesion and $M(t,\mathbf{u})$ accounts for mutations of cells from the less to the more invasive type (see [5] for full details). Note that the diffusion constants $D_{c,1}$ and $D_{c,2}$ are small so that diffusion (cell random motility) has, compared to adhesion and cell proliferation, only a small influence on the overall dynamics of the solution of (5). The modelling of cell random motility in (5) is rather simple and interactions between the two cell types and the ECM can be expected to affect random motility when the space gets locally filled, i.e., when $\rho(\mathbf{u})$ locally tends to one. To this end, we introduce nonlinear cross-diffusion in (5) by replacing $D_{c,1}\nabla c_1$ in (5a) with $D_{c,1} c_1 \nabla \rho(\mathbf{u}) = D_{c,1} c_1 \nabla c_1 + D_{c,1} c_1 \nabla c_2 + D_{c,1} c_1 \nabla v$, and similarly for $D_{c,2}\nabla c_2$ in (5b). Numerical simulations of system (5) without and with cross-diffusion are shown in Fig. 4. The introduction of cross-diffusion leaves the overall dynamics of the solution largely unaffected, Fig. 4b and c. However, the model with cross-diffusion exhibits a slightly slower speed of invasion of the cancer cells and the interface between the two cell types is significantly sharpened. This provides for a clear qualitative change of the model behaviour which may be of practical significance. Similar results are observed in a cell-sorting model with cross-diffusion [12].

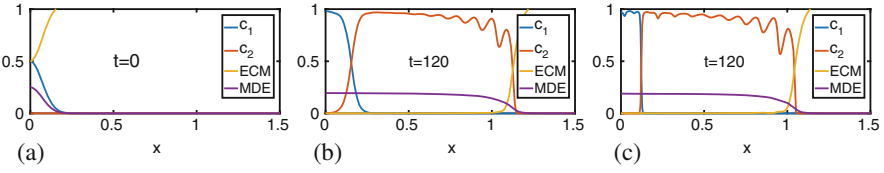


Fig. 4 (**a**) Initial data for model (5) at time $t = 0$. (**b**) Numerical solution of model (5) at time $t = 120$. (**c**) Same as (**b**) but for model (5) with cross-diffusion

4 Conclusion and Open Research Questions

In this review we have highlighted some important aspects of cross-diffusion terms in reaction-diffusion models with respect to their analytic and numerical treatment together with some specific examples. Follow-up research non-exhaustively includes nonlinear stability analysis for the investigation of patterning mechanisms far away from bifurcation points, the development of higher-order numerical schemes for reaction-cross-diffusion systems that honour their positivity preservation properties, and the development of analytical and numerical tools for reaction-cross-diffusion systems on evolving domains. This research will also be driven by the needs of developmental and cellular biologists and other scientists who frequently use cross-diffusion in their applications.

Acknowledgements AM is a Royal Society Wolfson Research Merit Award Holder generously supported by the Wolfson Foundation. All the authors (AM, RB, AG) thank the Isaac Newton Institute for Mathematical Sciences for its hospitality during the programme *Coupling Geometric PDEs with Physics for Cell Morphology, Motility and Pattern Formation*; EPSRC EP/K032208/1.

References

1. Andasari, V., Gerisch, A., Lolas, G., South, A.P., Chaplain, M.A.J.: Mathematical modeling of cancer cell invasion of tissue: biological insight from mathematical analysis and computational simulation. J. Math. Biol. **63**(1), 141–171 (2011)
2. Armstrong, N.J., Painter, K.J., Sherratt, J.A.: A continuum approach to modelling cell-cell adhesion. J. Theor. Biol. **243**(1), 98–113 (2006)
3. Bozzini, B., Lacitignola, D., Mele, C., Sgura, I.: Coupling of morphology and chemistry leads to morphogenesis in electrochemical metal growth: a review of the reaction-diffusion approach. Acta Appl. Math. **122**(1), 53–68 (2012)
4. Burger, M., Francesco, M.D., Pietschmann, J.F., Schlake, B.: Nonlinear cross-diffusion with size exclusion. SIAM J. Math. Anal. **42**(6), 2842–2871 (2010)
5. Domschke, P., Trucu, D., Gerisch, A., Chaplain, M.A.J.: Mathematical modelling of cancer invasion: implications of cell adhesion variability for tumour infiltrative growth patterns. J. Theor. Biol. **361**, 41–60 (2014)
6. Frittelli, M., Madzvamuse, A., Sgura, I., Venkataraman, C.: Lumped finite element method for reaction-diffusion systems on compact surfaces. arXiv:1609.02741 (2016)
7. Gambino, G., Lupo, S., Sammartino, M.: Effects of cross-diffusion on turing patterns in a reaction-diffusion schnakenberg model. arXiv:1501.04890 (2015)
8. Gerisch, A., Chaplain, M.: Robust numerical methods for taxis–diffusion–reaction systems: applications to biomedical problems. Math. Comput. Mod. **43**(1–2), 49–75 (2006)
9. Hittmeir, S., Jüngel, A.: Cross diffusion preventing blow-up in the two-dimensional Keller–Segel model. SIAM J. Math. Anal. **43**(2), 997–1022 (2011)
10. Madzvamuse, A., Barreira, R.: Exhibiting cross-diffusion-induced patterns for reaction-diffusion systems on evolving domains and surfaces. Phys. Rev. E **90**, 043307 (2014)
11. Madzvamuse, A., Ndakwo, H.S., Barreira, R.: Cross-diffusion-driven instability for reaction-diffusion systems: analysis and simulations. J. Math. Biol. **70**(4), 709–743 (2014)
12. Murakawa, H., Togashi, H.: Continuous models for cell–cell adhesion. J. Theor. Biol. **374**, 1–12 (2015)

13. Murray, J.D.: Mathematical Biology. I, 3rd edn. Springer, New York (2003)
14. Ni, W.M.: Diffusion, cross-diffusion, and their spike-layer steady states. Not. Am. Math. Soc. **45**(1), 9–18 (1998)
15. Rahman, K.A., Sudarsan, R., Eberl, H.J.: A mixed-culture biofilm model with cross-diffusion. Bull. Math. Biol. **77**(11), 2086–2124 (2015)
16. Rahman, K., Sonner, S., Eberl, H.: Derivation of a multi-species cross-diffusion model from a lattice differential equation and positivity of its solutions. Act. Phys. Pol. B **9**(1), 121 (2016)
17. Shigesada, N., Kawaasaki, K., Teramoto, E.: Spatial segregation of interacting species. J. Theor. Biol. **79**(1), 83–99 (1979)
18. Smoller, J.: Shock Waves and Reaction-Diffusion Equations, 2nd edn. Springer, New York (1994)

On a Splitting-Differentiation Process Leading to Cross-Diffusion

Gonzalo Galiano and Virginia Selgas

Abstract We generalize the dynamical system model proposed by Sánchez-Palencia for the splitting-differentiation process of populations to include spatial dependence. This gives rise to a family of cross-diffusion partial differential equations problems, among which we consider the segregation model proposed by Busenberg and Travis. For the one-dimensional case, we make a direct parabolic regularization of the problem to show the existence of solutions in the space of BV functions. Moreover, we introduce a Finite Element discretization of both our parabolic regularization and an alternative regularization previously proposed in the literature. Our numerical results suggest that our approach is more stable in the tricky regions where the solutions exhibit discontinuities.

1 The Splitting-Differentiation Model

We start from [17], where Sánchez-Palencia analyzes the splitting of one single species (with population density U) into two different species (with population densities U_1 and U_2), which preserve the original ecological behavior. More precisely, we assume that U satisfies a logistics law until time t^* and then, due to a number of factors, a splitting-differentiation happens at time t^*, after which (U_1, U_2) is assumed to satisfy a Lotka-Volterra system as follows:

$$U'(t) = U(t)\big(\alpha - \beta U(t)\big) \text{ for } t \in (0, t^*), \;\; U(0) = U_0 \qquad (1)$$

$$U_i'(t) = U_i(t)\big(\alpha - \beta\big(U_1(t) + U_2(t)\big)\big) \text{ for } t \in (t^*, T), \;\; U_i(t^*) = U_{i0}, \;\; (i = 1, 2) \qquad (2)$$

where we assume that $U_0, U_{i0} > 0$ and $U_{10} + U_{20} = U(t^*)$. It is important to notice that this splitting is *without differentiation* since $U_1 + U_2$ still satisfies (1) for $t \geq t^*$. Furthermore, problem (1) has a unique non-trivial equilibrium ($U_\infty = \alpha/\beta$) but, on

G. Galiano • V. Selgas (✉)
Department of Mathematics, Universidad de Oviedo, c/Calvo Sotelo, 33007 Oviedo, Spain
e-mail: galiano@uniovi.es; selgasvirginia@uniovi.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_62

the contrary, problem (2) has a continuum set of non-trivial equilibria (any $U_{1\infty} \geq 0$ and $U_{2\infty} \geq 0$ such that $U_{1\infty} + U_{2\infty} = \alpha/\beta$).

Alternatively, the splitting of populations U_1 and U_2 may involve *differentiation*, which is modelled as a perturbation in the Lotka-Volterra coefficients:

$$U_i'(t) = U_i(t)\big(\alpha_i - \big(\beta_{i1}U_1(t) + \beta_{i2}U_2(t)\big)\big) \text{ for } t \in (t^*, T), \quad U_i(t^*) = U_{i0}, \tag{3}$$

for $i = 1, 2$. Observe that, under this splitting, $U_1 + U_2$ does not satisfy, in general, problem (1) for $t \geq t^*$. Furthermore, concerning the equilibrium of the system, in [17], the author shows that the differentiation mechanism selects, in general, a unique solution for the equilibrium system, having therefore a stabilizing effect.

In this work, we want to extend the previous dynamical system models to the case of *space dependent* population densities. Thus, we first consider the dynamics of a species population density:

$$\partial_t u - \operatorname{div} J(u) = f(u) \text{ in } Q_{(0,t^*)}, \ J(u)\cdot\nu = 0 \text{ on } \Gamma_{(0,t^*)}, \ u(0,\cdot) = u_0 \text{ on } \Omega, \tag{4}$$

where $u_0 > 0$ and $\Omega \subset \mathbb{R}^N$ is a bounded Lipschitz domain, with ν the unit outwards normal vector field on the boundary $\partial\Omega$. We also denote $Q_{(0,t^*)} = (0, t^*) \times \Omega$, and its parabolic boundary by $\Gamma_{(0,t^*)} = (0, t^*) \times \partial\Omega$. We assume that the growth-competition term has the logistic form $f(u) = u(\alpha - \beta u)$, and the flow is given by $J(u) = u\nabla u + u\mathbf{q}$. Then, it is well-known that the term $u\nabla u$ captures the individuals aversion to overcrowding, while $\mathbf{q}$ is usually determined by an environmental potential, $\mathbf{q} = -\nabla\varphi$, whose minima represent attracting points for the populations. For biological background and origins of the model see [16] and references therein.

After splitting, the new two populations densities u_1 and u_2 satisfy the system:

$$\begin{aligned} &\partial_t u_i - \operatorname{div} J_i(u_1, u_2) = f_i(u_1, u_2) \text{ in } Q_{(t^*,T)}, \\ &J_i(u_1, u_2)\cdot\nu = 0 \text{ on } \Gamma_{(t^*,T)}, \ u_i(t^*,\cdot) = u_{i0} \text{ on } \Omega, \end{aligned} \tag{5}$$

for $i = 1, 2$, with u_{i0} such that $u_{10} + u_{20} = u(t^*, \cdot)$. The flows J_i and the growth-competition terms f_i are to be defined.

Definition of the Flows. Like for the dynamical system model, we assume that the possible differentiation process only takes place through the growth and the inter- and intra-competitive behavior of the new species. Thus, the split flows must satisfy

$$J_1(u_1, u_2) + J_2(u_1, u_2) = J(u_1 + u_2) = (u_1 + u_2)\nabla(u_1 + u_2) + (u_1 + u_2)\mathbf{q}.$$

This suggests a natural definition of the linear transport term of J_i, but the nonlinear diffusive term admits several reasonable decompositions. For instance, in [8] the author considered the splitting $J_i(u_1, u_2) = u_i\nabla u_i + b_i\nabla(u_1u_2) + u_i\mathbf{q}$, with $b_i \geq 0$ such that $b_1 + b_2 = 1$. Under this splitting, problem (5) takes the form of the cross-diffusion model introduced by Shigesada et al. [18], for which a thorough mathematical analysis does exist (cf. [1, 13, 14] for numerical approaches, [5, 6, 12,

15] for analytical and qualitative results, or [7] for applications). Here, we consider the alternative splitting

$$J_i(u_1, u_2) = u_i \nabla(u_1 + u_2) + u_i \mathbf{q}, \tag{6}$$

which brings problem (5) to the form of the Busenberg and Travis model [4]. Despite the simplicity of this definition, there is no general proof of existence of solutions for problem (5) with flows given by (6); some partial results may be found in [2, 3] (for the cell-growth contact-inhibition) and [9, 10] (for other specific situations).

Definition of the Growth-Competition Terms. In order to make the PDE versions that generalize the non-differentiation and differentiation ODE problems (2) and (3), we consider two different sets of Lotka-Volterra terms:

$$f_i(u_1, u_2) = u_i(\alpha - \beta(u_1 + u_2)), \qquad f_i(u_1, u_2) = u_i\big(\alpha_i - \big(\beta_{i1} u_1 + \beta_{i2} u_2\big)\big), \tag{7}$$

for $i = 1, 2$. We shall refer to problem (5) with flows given by (6) and with growth-competition terms given by (7_1) or (7_2) as to *problems (ND) or (D)*, respectively.

It is important to notice that the existence of solution of (ND) was proven in [9] even in the multi-dimensional case. This is shown making use of a nonlinear parabolic regularization of problem (ND): Its solutions (u_1^δ, u_2^δ) are such that they only converges weakly in $L^\infty(Q_T)$ to the solution (u_1, u_2) of the limit problem; however, the global density $u_1^\delta + u_2^\delta$ converges strongly in $L^2(Q_T)$ to $u_1 + u_2$, so that weak and strong convergences compensate and allow us to deduce the existence of a solution of problem (ND). On the contrary, in case of problem (D) there is no compensation and we cannot handle it directly with this technique. In order to use a different approach, we have in mind the strategy of Bertsch et al. in [2] for the one-dimensional case when $q = 0$: They introduce a parabolic-hyperbolic auxiliary problem (see $\mathrm{P}^B_{\delta_B}$ in Sect. 2), whose solution u, $0 \le r \le 1$ allow them to build a solution of problem (D) with the form $u_1 = ru, u_2 = (1 - r)u$. In [11], we propose an alternative proof of the existence of solutions of problem (D): We use the same direct parabolic regularization of problem (ND) as in [9], but we take advantage of the techniques employed in [2] to obtain the strong convergence of each component of the regularized problem; more explicitly, the following result is proven in [11].

Theorem 1 *Let $\Omega \subset \mathbb{R}$ be a bounded interval, $T > t^* > 0$ fixed, and $q \in C^{0,1}(\bar{Q}_{(t^*,T)})$ with $q(t, \cdot) = 0$ on $\partial\Omega$ for all $t \in [t^*, T]$. Assume that u_{10}, $u_{20} \in BV(\Omega)$ are non-negative, with $u_0 := u_{10} + u_{20} > 0$ in $\bar{\Omega}$, $u_0 \in C^{0,1}(\bar{\Omega})$ and $\partial_x u_0 = 0$ on $\partial\Omega$. Then, there exists $(u_1, u_2) \in (L^\infty(t^*, T; BV(\Omega)) \cap BV(t^*, T; L^1(\Omega)))^2$ such that $u := u_1 + u_2 \in C^{0,1}(\bar{Q}_{(t^*,T)}) \cap L^2(t^*, T; H^2(\Omega))$, and which defines a weak solution of problem (D) in the following sense: for $i = 1, 2$, $u_i \ge 0$ a.e. in $Q_{(t^*,T)}$ and*

$$\int_{Q_{(t^*,T)}} u_i \, \partial_t \varphi = \int_{Q_{(t^*,T)}} \big(u_i \partial_x u + u_i q\big) \, \partial_x \varphi - \int_{Q_{(t^*,T)}} f_i(u_1, u_2) \, \varphi - \int_\Omega u_{i0} \, \varphi(t^*, \cdot)$$

for all $\varphi \in C^1(t^*, T; H^1(\Omega))$ with $\varphi(T, \cdot) = 0$ in Ω.

2 Approximated Problems and Discretization

The second aim of this article is to numerically investigate and compare the resulting Finite Element schemes of each regularizing approach.

- The approximations of problem (D) constructed via the regularization scheme used to prove Theorem 1 are the following:

$$\mathrm{P}_\delta \begin{cases} \partial_t u_i - \partial_x\big(u_i\partial_x(u_1+u_2)+u_i q\big) - \frac{\delta}{2}\partial_{xx}(u_i(u_1+u_2)) = f_i(u_1,u_2) \ \text{ in } Q_{(t^*,T)}, \\ \partial_x u_i = 0 \text{ on } \Gamma_{(t^*,T)}, \ u_i(t^*,\cdot) = u_{i0}^\delta \ \text{ in } \Omega, \end{cases}$$

  for $\delta > 0$, and some non-negative $u_{i0}^\delta \in C^1(\bar{\Omega})$ such that $u_{i0}^\delta \to u_{i0}$ strongly in $BV(\Omega)$, for $i = 1, 2$.
- The approximations of the problem introduced in [2] are the following:

$$\mathrm{P}_{\delta_B}^B \begin{cases} \partial_t u - \partial_x(u(\partial_x u + q)) = F_1(u,r) \ \text{ in } Q_{(t^*,T)}, \\ \partial_t r - (\partial_x u + q)\partial_x r - \delta_B \partial_{xx} r = F_2(u,r) \ \text{ in } Q_{(t^*,T)}, \\ \partial_x u = \partial_x r = 0 \text{ on } \Gamma_{(t^*,T)}, \ u(t^*,\cdot) = u_0, \ r(t^*,\cdot) = r_0^{\delta_B} \text{ in } \Omega, \end{cases}$$

  for $\delta_B > 0$, $r_0^{\delta_B} \in C^1(\bar{\Omega})$ such that $r_0^{\delta_B} \to r_0$ strongly in $BV(\Omega)$, and where we denote $F_1(u,r) = (f_1+f_2)(u_1,u_2)$, $F_2(u,r) = (1-r)\frac{f_1(u_1,u_2)}{u} - r\frac{f_2(u_1,u_2)}{u}$ with $u_1 = ru, u_2 = (1-r)u$.

For the *numerical discretization* of problems P_δ and $\mathrm{P}_{\delta_B}^B$, we consider a fully discrete approximation using finite elements in space and backward finite differences in time. Let us notice that the proof of the convergence of the numerical scheme for problem P_δ may be found in [9]. On the other hand, according to [2], for any $\delta_B > 0$ there exists a regular solution to problem $\mathrm{P}_{\delta_B}^B$; further, when $\delta_B \to 0$, it converges strongly in $L^1(Q_{(t^*,T)})$ to some (u,r) such that $u_1 = ru, u_2 = (1-r)u$ solves problem (D).

More precisely, for the numerical experiments, we consider a quasi-uniform mesh on the interval Ω, $\{\mathcal{T}_h\}_h$, with h representing step size, and take $\mathbb{P}_1$-finite elements. Besides, for the time discretization, we fix $t^* = 0$ and define a uniform partition of $[0,T]$ of time step τ. This strategy leads to nonlinear algebraic problems, which we solve approximately at each time step with a fixed point method.

For the sake of brevity, here we focus our attention into an *invasion* problem, which arises in Tumor theory (cf. [2]) and where an initial small perturbation (tumor) inside a healthy tissue evolves in time keeping a sharp interface. More precisely, we take the initial data $u_{10}(x) = 0.22\exp(-(x-0.25)^2/0.001)$, $u_{20}(x) = 0.45 - u_{10}(x)$ for $x \in \Omega$, and the Lotka-Volterra coefficients $\alpha_i = 1$, $\beta_{ij} = i$ for $i,j = 1,2$. For the numerical simulations, we consider the spatial domain $\Omega = (-2,2)$ and take several choices of the mesh size, h; only the experiments corresponding to $h = 0.04$ (101 nodes), $h = 0.013$ (301 nodes), and $h = 0.008$ (501 nodes) are shown in Fig. 1. Besides, the time step is empirically chosen as $\tau = 10^{-3}$, whereas the regularization parameters are taken as $\delta = h^2$, $\delta_B = 2h^2$.

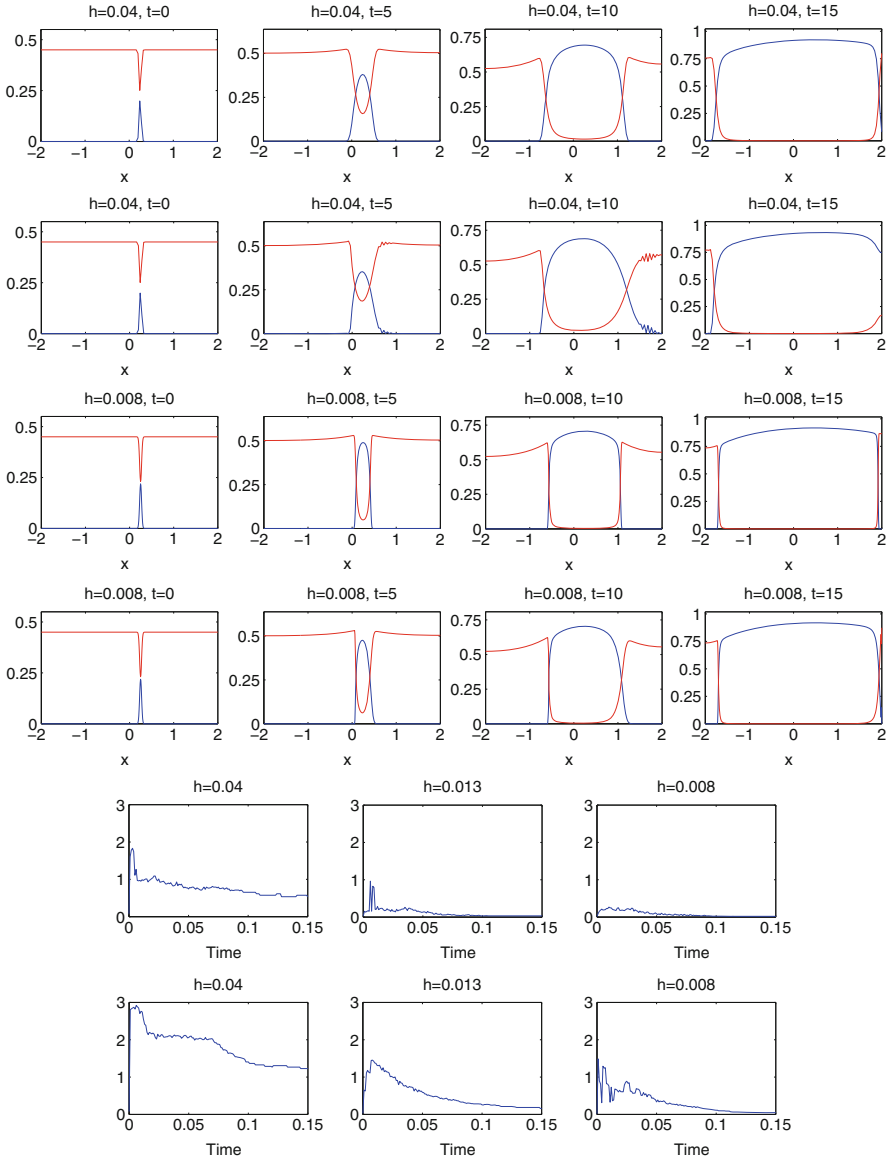


Fig. 1 *Top block*: Odd rows correspond to the approximated solution of P_δ for several time slices, t, while even rows correspond to the solution of $P^B_{\delta_B}$. Rows correspond to different mesh sizes, captured by parameter h. Mind the different vertical scales among time slices. *Bottom block*: Each row corresponds to the oscillation measure $\mathrm{osc}(u)(t)$ of the solutions approximated by P_δ and $P^B_{\delta_B}$, respectively. Columns correspond to different mesh sizes, captured by parameter h

To avoid negative values of the discrete solutions, we use the functions λ_ε and Λ_ε defined as in [11] setting $\varepsilon = 10^{-10}$. Finally, we fix the tolerance for the fixed point stopping criterion, observing good convergence properties of the fixed point algorithm and reaching the prescribed tolerance in less than ten iterations.

For the example at hand, as well as those investigated in [11], when we compare the approximations of the solutions of problem (D) either by the scheme derived from the regularized parabolic-hyperbolic formulation $\mathrm{P}^B_{\delta_B}$ of [2], or by the direct viscosity approximation introduced in this article, P_δ, we come up with the following conclusion: The sequence of solutions of problem P_δ is more robust against instabilities produced in discontinuity points than that given by the sequence corresponding to $\mathrm{P}^B_{\delta_B}$. To clarify that the solutions corresponding to problem $\mathrm{P}^B_{\delta_B}$ produce more oscillations than those of problem P_δ, we approximate the amount of zero-crossings of $\partial_x u(t,\cdot)$ in Ω by $\mathrm{osc}(u)(t) = h\sum|\Delta(\mathrm{sign}(\Delta u(t,\cdot)))|$, where Δ stands for the difference between two consecutive node values, and the sum runs over the spatial nodes. In particular, this allows us to verify that the oscillations of the solutions of P_δ attenuate when t increases, whereas those corresponding to solutions of $\mathrm{P}^B_{\delta_B}$ are always above some threshold value. With respect to the mesh size, the size of oscillations diminish for both approximations when h decreases. Indeed, for the experiment at hand, in Fig. 1 we may check that for the finer mesh (third and fourth rows), both approximations give similar results. On the contrary, for the coarsest mesh and for the printed resolution, instabilities in the approximation obtained from $\mathrm{P}^B_{\delta_B}$ are present (second row), whereas they are not visible in the corresponding approximation obtained from P_δ (first row). These graphical evidences are confirmed through the oscillation measure $\mathrm{osc}(u)(t)$ plotted in Fig. 1 (bottom block).

3 Conclusions

We have generalized the dynamical system model for the splitting-differentiation process of populations [17] to the space dependent situation using the cross-diffusion PDE problem of Busenberg and Travis, in which the segregation effects are stronger (cf. [9]). For this model, we have shown the existence of solutions using a direct parabolic regularization of the problem, meanwhile the previous proof in [2] is conceptually more complicated because it involves a change of unknowns rendering the problem to a parabolic-hyperbolic formulation.

Moreover, it is difficult to ensure, but our approach also seems to select a unique *natural* solution of the problem as a limit of vanishing viscosity solutions of the regularized parabolic problems, tackling thus the non-uniqueness issue of the parabolic-hyperbolic formulation. Moreover, our numerical results show that our approach is more stable in the tricky regions where the solutions exhibit discontinuities. Such experiments also show that the positivity of the initial total mass is not a necessary condition of the existence theorems. This condition, as well as the generalization to the multi-dimensional case, are to be studied in future.

References

1. Berres, S., Ruiz-Baier, R.: A fully adaptive numerical approximation for a two-dimensional epidemic model with nonlinear cross-diffusion. Nonlinear Anal. Real World Appl. **12**, 2888–2903 (2011)
2. Bertsch, M., Dal Passo, R., Mimura, M.: A free boundary problem arising in a simplified tumour growth model of contact inhibition. Interfaces Free Bound. **12**, 235–250 (2010)
3. Bertsch, M., Hilhorst, D., Izuhara, H., Mimura, M.: A nonlinear parabolic-hyperbolic system for contact inhibition of cell-growth. Differ. Equ. Appl. **4**(1), 137–157 (2012)
4. Busenberg, S.N., Travis, C.C.: Epidemic models with spatial spread due to population migration. J. Math. Biol. **16**, 181–198 (1983)
5. Chen, L., Jüngel, A.: Analysis of a multidimensional parabolic population model with strong cross-diffusion. SIAM J. Math. Anal. **36**(1), 301–322 (2004)
6. Desvillettes, L., Lepoutre, T., Moussa, A.: Entropy, duality, and cross diffusion. SIAM J. Math. Anal. **46**, 820–853 (2014)
7. Galiano, G.: Modelling spatial adaptation of populations by a time non-local convection cross-diffusion evolution problem. Appl. Math. Comput. **218**(8), 4587–4594 (2011)
8. Galiano, G.: On a cross-diffusion population model deduced from mutation and splitting of a single species. Comput. Math. Appl. **64**, 1927–1936 (2012)
9. Galiano, G., Selgas, V.: On a cross-diffusion segregation problem arising from a model of interacting particles. Nonlinear Anal. Real World App. **18**, 34–49 (2014)
10. Galiano, G., Selgas, V.: Deterministic particle method approximation of a contact inhibition cross-diffusion problem. Appl. Numer. Math. **95**, 229–237 (2015)
11. Galiano, G., Selgas, V.: Analysis of a splitting-differentiation population model leading to cross-diffusion. Comput. Math. Appl. **70**(12), 2933–2945 (2015)
12. Galiano, G., Velasco, J.: Competing through altering the environment: a cross-diffusion population model coupled to transport-Darcy flow equations. Nonlinear Anal. Real World App. **12**(5), 2826–2838 (2011)
13. Galiano, G., Garzón, M.L., Jüngel, A.: Semi-discretization in time and numerical convergence of solutions of a nonlinear cross-diffusion population model. Numer. Math. **93**(4), 655–673 (2003)
14. Gambino, G., Lombardo, M.C., Sammartino, M.: A velocity-diffusion method for a Lotka-Volterra system with nonlinear cross and self-diffusion. Appl. Numer. Math. **59**, 1059–1074 (2009)
15. Gambino, G., Lombardo, M.C., Sammartino, M.: Pattern formation driven by cross-diffusion in a 2D domain. Nonlinear Anal.: Real World Appl. **14**, 1755–1779 (2013)
16. Okubo, A.: Diffusion and Ecological Problems. Springer, New York (2002)
17. Sánchez-Palencia, E.: Competition of subspecies and structural stability. Survival of the best adapted or coexistence? In: International Congress on Nonlinear Models in Partial Differential Equations, Toledo, Spain (2011)
18. Shigesada, N., Kawasaki, K., Teramoto, E.: Spatial segregation of interacting species. J. Theor. Biol. **79**, 83–99 (1979)

A Discrete Cross-Diffusion Model for Image Restoration

Adérito Araújo, Silvia Barbeiro, Eduardo Cuesta, and Ángel Durán

Abstract In this paper a fully discrete cross-diffusion model for image restoration is introduced. The image is represented by a two-component vector field, and the restoration process is governed by a nonlinear cross-diffusion difference system. We explore numerically the potentialities of using the nonlinear cross-diffusion approach as an image filter, in particular as a preprocessing step for image segmentation.

1 Introduction

The use of differential problems of cross-diffusion type as mathematical models for image filtering processes has been recently proposed in [1, 2], where the continuous linear and nonlinear cases are analyzed. These systems are usually considered in the study of various physical phenomena (especially in models on population dynamics) [12]; their application to image processing problems is motivated by the idea of distributing the properties of the image in two real-valued functions and generalizes the complex diffusion models proposed by Gilboa et al. [10], as well as some other previous works, [11] (see also [3] and references therein).

While the approach in [1, 2] was devoted to the continuous model, in the present paper a discrete cross-diffusion formulation is introduced and its behaviour is illustrated numerically. The experiments are focused on the application of the model as a preprocessing step of the segmentation of Optical Coherence Tomography (OCT) images [9]. This technique, with important applications in ophthalmology (among other fields), is based on synthesizing cross-section images from depth resolution scanning. On the other hand, segmentation of anatomical and pathological structures in OCT images is crucial for the diagnosis and study of ocular diseases, [6, 8]. Results obtained in [4] suggest some advantages of using a complex diffusion filter

A. Araújo • S. Barbeiro
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
e-mail: alma@mat.uc.pt; silvia@mat.uc.pt

E. Cuesta (✉) • Á. Durán
Department of Applied Mathematics, University of Valladolid, Valladolid, Spain
e-mail: eduardo@mat.uva.es; angel@mac.uva.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_63

before the OCT retina layer segmentation in order to reduce the speckle noise suffered by this kind of images [9]. The discrete model introduced in the present paper generalizes the complex diffusion approach, and the goal of the experiments is to discuss the performance of cross-diffusion filters as a preprocessing step for image segmentation.

The paper is structured as follows. In Sect. 2 the discrete model is introduced, and some properties are discussed. In Sect. 3 some schemes corresponding to different choices of the cross-diffusion matrix are compared as nonlinear filters in two numerical experiments. The main conclusions are outlined in Sect. 4.

2 A Discrete Cross-Diffusion Model

2.1 Pixel Mesh, Initial Distribution and Cross-Diffusion Matrix

Some elements of the model are here introduced. We consider an interval $\Omega = (a_1, b_1) \times (a_2, b_2) \subset \mathbb{R}^2$, and denote its boundary by Γ. The domain $\overline{\Omega} := \Omega \bigcup \Gamma$ is discretized by points $z_{\mathbf{j}} = (x_{j_1}, x_{j_2})$, $\mathbf{j} = (j_1, j_2)$ with

$$x_{j_1} = a_1 + h_1 j_1, \quad x_{j_2} = a_2 + h_2 j_2, \quad j_k = 0, 1, \ldots, N_k, \quad h_k = \frac{b_k - a_k}{N_k}, \quad k = 1, 2,$$

for two given integers $N_1, N_2 \geq 1$. This first mesh on Ω is denoted by $\Omega_{\mathbf{h}}$, $\mathbf{h} = (h_1, h_2)$. (The discretized boundary $\Gamma_{\mathbf{h}}$ is generated in the same way.) For mathematical purposes we define a second mesh, denoted by $\Omega_{\mathbf{h}/2}$, consisting of the midpoints $Z_{\mathbf{j} \pm (\mathbf{1/2})\mathbf{e_k}}$, $k = 1, 2$, where $\{\mathbf{e}_1, \mathbf{e}_2\}$ is the canonical basis. The corresponding stencil is shown in Fig. 1.

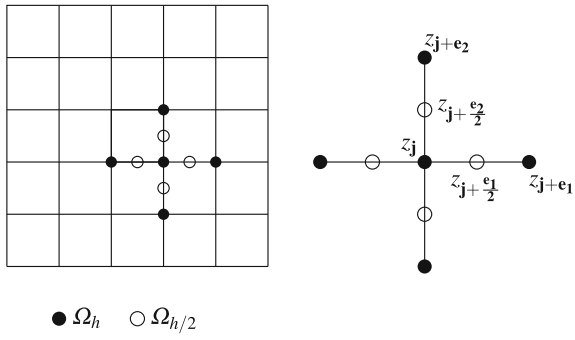


Fig. 1 Pixel mesh on Ω

Consider a vector map $\mathbf{Z} = (Z_{\mathbf{j}})_{\mathbf{j}}$ defined on $\Omega_{\mathbf{h}} \cup \Omega_{\mathbf{h}/2}$. Associated to the mesh the following operators, that will be used below, are defined:

$$\delta_k^+ Z_{\mathbf{j}} := \frac{Z_{\mathbf{j}+\mathbf{e_k}} - Z_{\mathbf{j}}}{h_k}, \quad \delta_k^- Z_{\mathbf{j}} := \frac{Z_{\mathbf{j}} - Z_{\mathbf{j}-\mathbf{e_k}}}{h_k}, \quad \delta_k Z_{\mathbf{j}} := \frac{Z_{\mathbf{j}+(\mathbf{1/2})\mathbf{e_k}} - Z_{\mathbf{j}-(\mathbf{1/2})\mathbf{e_k}}}{h_k},$$

for $k = 1, 2$.

In order to formulate the discrete cross-diffusion restoration problem, let $u_0 : \Omega_{\mathbf{h}} \to \mathbb{R}$ be a discrete real-valued function standing for the grey level values of a noisy image on $\Omega_{\mathbf{h}}$ to be restored. From u_0, an initial distribution $\mathbf{W}^0 = (U^0, V^0)$, for the cross-diffusion, is required. This is given by two real-valued functions $U^0, V^0 : \Omega_{\mathbf{h}} \to \mathbb{R}$, that can be selected following different criteria. A simple choice, that will considered in the experiments below, consists of taking $U^0 = u_0$, and V^0 as the zero array. A more detailed discussion on the initial data can be seen in [1, 2].

Finally, if $\mathscr{M}_{2\times 2}(\mathbb{R})$ denotes the space of 2×2 real matrices, then we consider a mapping $D : \mathbb{R}^2 \to \mathscr{M}_{2\times 2}(\mathbb{R})$,

$$D(\mathbf{W}) := \begin{pmatrix} D_{11}(\mathbf{W}) & D_{12}(\mathbf{W}) \\ D_{21}(\mathbf{W}) & D_{22}(\mathbf{W}) \end{pmatrix}, \quad \mathbf{W} = (U, V) \in \mathbb{R}^2,$$

which is uniformly positive definite, and such that its entries $D_{ij} : \mathbb{R}^2 \to \mathbb{R}$, $i, j = 1, 2$, are bounded, and globally Lipschitz. The mapping D will define the cross-diffusion matrix coefficient of the model, [2].

2.2 Fully Discrete Model

From the values of the initial distribution $\mathbf{W}^0 = (U^0_{\mathbf{j}}, V^0_{\mathbf{j}})$, where $U^0_{\mathbf{j}}, V^0_{\mathbf{j}} : \Omega_{\mathbf{h}} \cup \Omega_{\mathbf{h}/2} \to \mathbb{R}^{N_1 \times N_2}$, the discrete model computes the restored image $\mathbf{W}^m = (U^m_{\mathbf{j}}, V^m_{\mathbf{j}})$ at several stages $m = 1, 2, \ldots$ If Δt stands for the time computed between $\mathbf{W}^m$ and $\mathbf{W}^{m+1}$ (which is assumed to be independent of m), then the restored image $\mathbf{W}^{m+1} = (U^{m+1}_{\mathbf{j}}, V^{m+1}_{\mathbf{j}})$ at level $m + 1$ is given by the system

$$\frac{U^{m+1}_{\mathbf{j}} - U^m_{\mathbf{j}}}{\Delta t} = \sum_{k=1}^{2} \delta_k \left(D_{11}(\mathbf{W}^m) \delta_k U^m_{\mathbf{j}} + D_{12}(\mathbf{W}^m) \delta_k V^m_{\mathbf{j}} \right), \tag{1}$$

$$\frac{V^{m+1}_{\mathbf{j}} - V^m_{\mathbf{j}}}{\Delta t} = \sum_{k=1}^{2} \delta_k \left(D_{21}(\mathbf{W}^m) \delta_k U^m_{\mathbf{j}} + D_{22}(\mathbf{W}^m) \delta_k V^m_{\mathbf{j}} \right), \tag{2}$$

along with boundary conditions on Γ_h of the form,

$$(\delta_k^+ + \delta_k^-) U^m_{\mathbf{j}} = (\delta_k^+ + \delta_k^-) V^m_{\mathbf{j}} = 0, \quad k = 1, 2. \tag{3}$$

In (1)–(2) the action of δ_k on the cross-diffusion matrix $D(\mathbf{W}^m)$, $m = 0, 1, \ldots$, is defined as follows: If $\mathbf{Z} = (\mathbf{Z_j})_\mathbf{j}$, $\mathbf{Z_j} = (U_\mathbf{j}, V_\mathbf{j}) \in \mathbb{R}^2$,

$$D(\mathbf{Z})_{\mathbf{j}+(\mathbf{1}/\mathbf{2})\mathbf{e_k}} := D\left(\frac{\mathbf{Z_j} + \mathbf{Z_{j+e_k}}}{2}\right).$$

The formulation of (1)–(3) for the particular case of complex diffusion was proposed and analyzed in [3]. As a scheme of approximation to the continuous problem, some stability condition was required. From this point of view, a similar restriction on Δt and $\mathbf{h}$ for (1)–(3) can be derived. However, the one obtained in [3] for complex diffusion was enough for the experiments performed in Sect. 3 to be stable.

3 Some Numerical Examples

Some experiments on the behaviour of (1)–(3) as nonlinear filter are here presented. As mentioned in the Introduction, the main goal of the paper is to study the use of cross-diffusion to reduce speckle noise in OCT images and to compare with the particular case of the complex diffusion in the visual assessment and the structure segmentation. To this end, some tests were conducted in the real human eye fundus OCT data using the high-definition spectral domain Cirrus OCT (Carl Zeiss Meditec, Dublin, CA, USA), courtesy of IBILI—Institute for Biomedical Imaging and Life Sciences of the University of Coimbra, Portugal. In particular, several high resolution B-scan images of 1024×1024 pixels were considered.

The one shown in Fig. 2 was taken as a grey scale image u_0, from which the initial distribution $\mathbf{W}^0 = (U^0, V^0)$, $U^0 = u_0$, $V^0 = 0$, is defined. The formulation (1)–(3) was run for two choices of the cross-diffusion matrix, namely

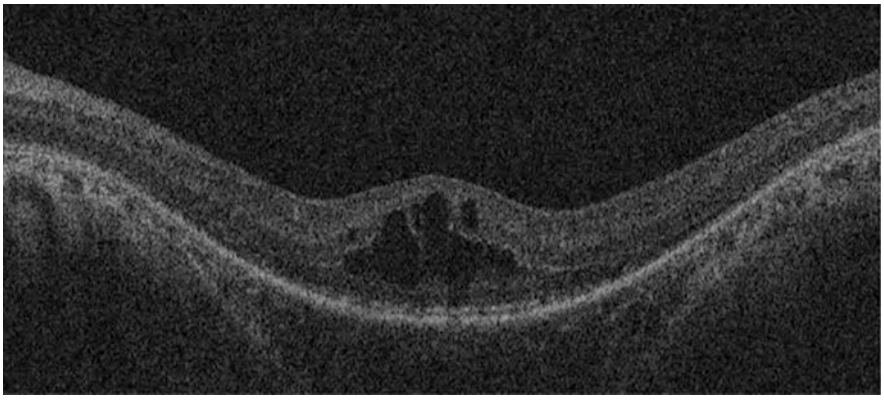


Fig. 2 B-scans of the human retina, courtesy of IBILI—Institute for Biomedical Imaging and Life Sciences of the University of Coimbra, Portugal

$$D(U,V) = g(V)\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}, \quad \text{and} \quad D(U,V) = g(V)\begin{pmatrix}\cos\theta & 9\\ 0.1 & \cos\theta\end{pmatrix}, \tag{4}$$

where,

$$g(V) = \frac{1}{1+\left(\frac{V}{k\theta}\right)^2}, \quad \text{for} \quad \theta = \pi/30, \;\; k = 10. \tag{5}$$

This leads to two different models, denoted by NCDF1, and NCDF2 respectively. Note that NCDF1 corresponds to a complex diffusion case, written as a cross-diffusion system [1, 2], with a diffusion coefficient of the form $e^{i\theta}g(v)$, v being the imaginary part of the image, and g given by (5). The choice of (4) and (5) is in some sense exploratory, but motivated by the results in [1, 2] on the continuous problem, and some experimental studies in [10] (and references therein) for complex diffusion.

A first experiment compares the evolution of the filtering process given by NCDF1 and NCDF2. This is illustrated in Fig. 3. For both systems, the code (1)–(3) is run with $\Delta t = 0.05$, and up to $m = 500$. Figure 3 represents the behaviour of the blur effect given in each case. This was measured by using a no-reference perceptual blur metric proposed in [7], based on the discrimination between levels of blur on the same image. The metric ranges from 0 (the best quality blur perception), to 1 (the worse one). According to the results, NCDF2 improves the behaviour of NCDF1 reducing the overdiffusion by approximately half at the end of the evolution. The slower increment of the metric in the case of NCDF2 suggests a bigger strength against long time filtering processes.

The second experiment shows how the efficiency of the filtering algorithm (1)–(3) allows an accurate segmentation and clustering of the filtered image. The

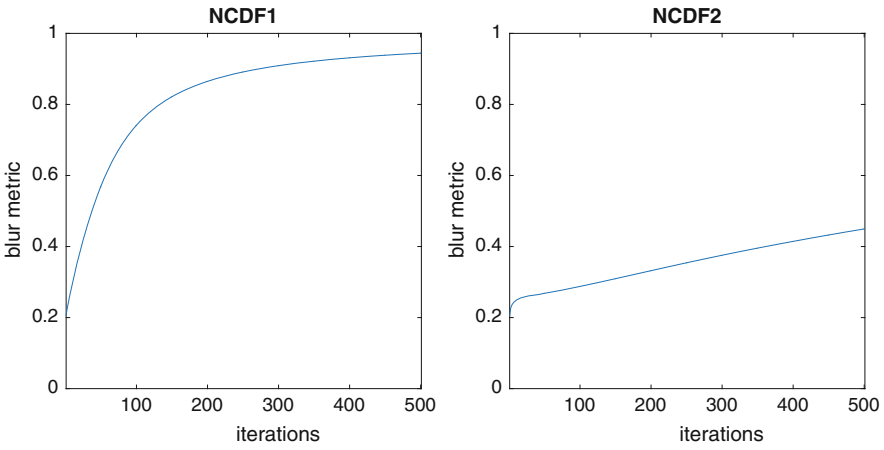


Fig. 3 Blur evolution of the denoised images obtained by NCDF1 and NCDF2

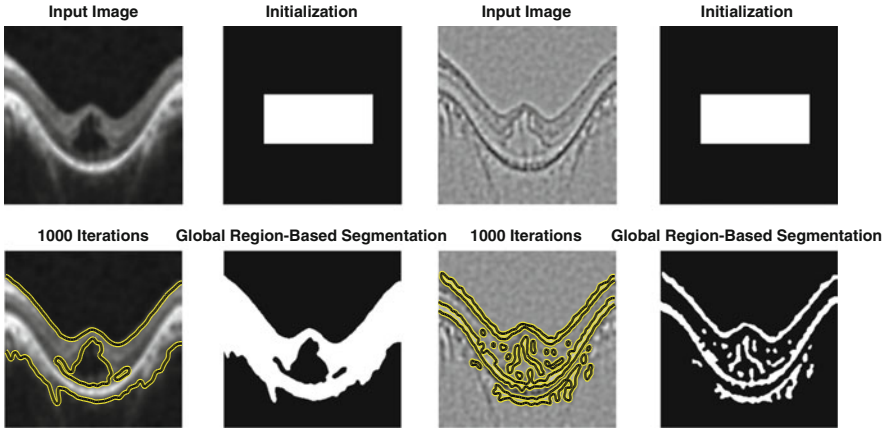


Fig. 4 NCDF1 algorithm. *First row from left*: First component of the filtered image; Initialization of segmentation algorithm; Second component of the filtered image; Initialization of segmentation algorithm. *Second row from left*: First component of the filtered image after segmentation; First component after clustering; Second component of the filtered image after segmentation; Second component after clustering

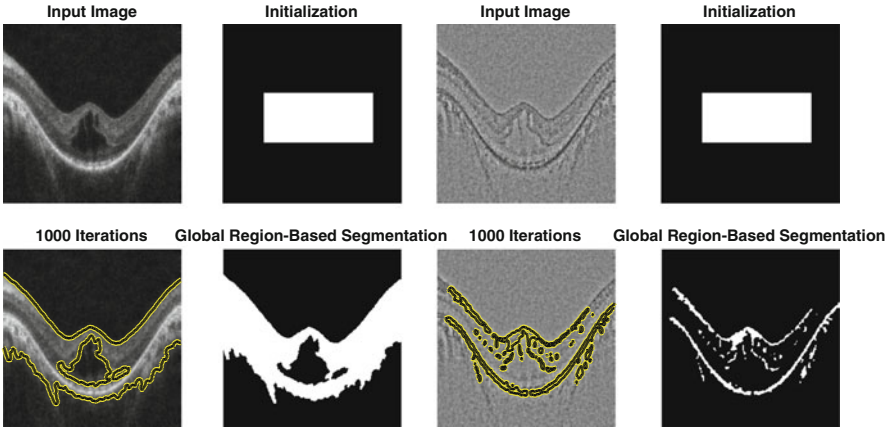


Fig. 5 As Fig. 4, here for NCDF2

image in Fig. 2 was filtered by NCDF1 and NCDF2 with $\Delta t = 0.05$ and up to $m = 1000$ units of time, and each component of the resulting image was segmented by the algorithm proposed in [5]. From Figs. 4–5 we observe that segmentation of the first component $(U_j^N)_j$ for the identification of the vitreous/internal limiting membrane (ILM) (first layer) and the Bruch's membrane (BrM)/choroid (last layer) is accurately performed. On the other hand, that segmentation of $(V_j^N)_j$ identifies the inner layers, in particular the retinal nerve fiber layer (RNFL), and the retinal pigment epithelium (RPE).

Commercial OCT systems are equipped with segmentation software mainly targeted to measure the RNFL and the total retinal thicknesses [5]. The numerical tests presented in Figs. 4–5 suggest the potential of cross-diffusion filtering for medical image segmentation.

4 Concluding Remarks

In the present paper a fully discrete model for image restoration, based on cross-diffusion is introduced. The model is characterized by a distribution of the image in two arrays of values on the pixel domain, and some evolutionary process governed by a cross-diffusion system of difference equations. Two numerical examples, related to the problem of reducing the speckle noise in OCT images, reveal that suitable choices of the cross-diffusion matrix improve the performance of the complex diffusion approach for these problems.

Acknowledgements This work was supported by Spanish Ministerio de Economía y Competitividad under the Research Grant MTM2014-54710-P. A. Araújo and S. Barbeiro were also supported by the Centre for Mathematics of the University of Coimbra—UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MCTES and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

1. Araújo, A., Barbeiro, S., Cuesta, E., Durán, A.: Cross-diffusion systems for image processing: I. The linear case. Preprint available at http://arxiv.org/abs/1605.02923
2. Araújo, A., Barbeiro, S., Cuesta, E., Durán, A.: Cross-diffusion systems for image processing: II. The nonlinear case. Preprint available at http://arxiv.org/abs/1605.03721
3. Araújo, A., Barbeiro, S., Serranho, P.: Convergence of finite difference schemes for nonlinear complex reaction-diffusion processes. SIAM J. Numer. Anal. **53**(1), 228–250 (2015)
4. Bernardes, R., Maduro, C., Serranho, P., Araújo, A., Barbeiro, S., Cunha-Vaz, J.: Improved adaptive complex diffusion despeckling filter. Opt. Express **18**(23), 24048–24059 (2010)
5. Chiu, S.J., Li, X.T., Nicholas, P., Toth, C.A., Izatt, C.A., Farsiu, S.: Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. Opt. Express **18**(18), 19413–19428 (2010)
6. Ciulla, T., Amador, A., Zinman, B.: Diabetic retinopathy and diabetic macular edema. Diabetes Care **26**(9), 2653–2664 (2003)
7. Crété-Roffet, F., Dolmiere, T., Ladret, P., Nicolas, M.: The blur effect: perception and estimation with a new no-reference perceptual blur metric. In: Proceedings of SPIE on Human Vision and Electronic Imaging XII, San Jose, California, USA. 11 pp. (2007)
8. Cunha-Vaz, J., Bernardes, R.: Nonproliferative retinopathy in diabetes type 2. Initial stages and characterization of phenotypes. Prog. Retin. Eye Res. **24**(3), 355–377 (2005)
9. Fercher, A.F., Drexler, W., Hitzenberger, C.K., Lasser, T.: Optical coherence tomography-principles and applications. Rep. Prog. Phys. **63**2, 239–303 (2003)

10. Gilboa, G., Sochen, N. A., Zeevi, Y. Y.: Image enhancement and denoising by complex diffusion processes. IEEE Trans. Pattern Anal. Mach. Intell. **26**(8), 1020–1036 (2004)
11. Lorenz, D.A., Bredies, K., Zeevi, Y.Y.: Nonlinear complex and cross diffusion. Unpublished report, University of Bremen (2006)
12. Ni, W.-M.: Diffusion, cross-diffusion and their spike-layer steady states. Not. Am. Math. Soc. **45**(1), 9–18 (1998)

Consensus-Based Global Optimization

Claudia Totzeck

Abstract We discuss some algorithms for global optimization and opinion formation and their relation to consensus-based optimization (CBO). The proposed CBO algorithm allows to pass to the mean-field limit, resulting in a Fokker-Plank type equation with non-linear, non-local and degenerate drift and diffusion term. We shed some light on the prospects of justifying the efficacy of the CBO algorithm on the PDE level.

1 Introduction

The global optimization of target functions is a common problem in many applications and still a challenge from the mathematical point of view [14]. Most of the approaches used to achieve sufficiently good approximations of the global optimum are *metaheuristics*, which are often inspired by nature [18]. For example, swarm intelligence (SI) algorithms, such as *particle swarm optimization* (PSO) (cf. [12]), *ant colony optimization* (ACO) [15] or *artificial bee colony* (ABC) optimization [11] typically consist of a population of agents communicating with each other and the environment. This leads to an 'intelligent' collective behaviour, that exceeds the knowledge of the individual agents. In fact, the basic setting is a population of agents exploring the state space of the target function while having access to the function values at every instance of time. The agents' motion is influenced by the individual steering and interaction with the collective. A common feature is the information on the global best and local best information stored along the trajectories which is used in the dynamic. One of the challenges is the balancing of weights between the steering and the ability to explore that is assigned to the agents' behaviour. In the literature many of the algorithms can be found mainly differing in memory effects, stochasticity or time discretization.

Moreover, individual-based models are widely used for studies of collective behaviour effects in applications of mathematical biology, swarming, crowd

C. Totzeck (✉)
TU Kaiserslautern, Erwin-Schrödinger Str. 1, 67663 Kaiserslautern, Germany
e-mail: totzeck@mathematik.uni-kl.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_64

dynamics or opinion dynamics (see [16]) for an overview). The three common formations of opinion dynamics are either consensus, polarisation or fragmentation. To achieve a thorough understanding of such phenomena, a rigorous formulation as mathematical model, which describes the evolution of opinions in the population under investigation, is needed (cf. [7]). For a large number of individuals a thorough study of individual-based models becomes exceedingly difficult, if not impossible. In this case it is popular to employ techniques from kinetic theory to approximate the models by means of partial differential equations (PDE), which describe the evolution of the agents in terms of a probability density (see for example [2, 9] for references in the mean-field context or [1, 3] for evolutionary differential games). With the help of PDE methods the investigation of stationary states, ergodicity and pattern formation is often easier than on the discrete particle level.

The *consensus-based optimization* (CBO) algorithm [4, 17] we discuss in the following can be classified as an stochastic SI model that is strongly related to consensus formation models. The agent-based global optimization strategy is efficient and robust and it allows for passing to the mean-field limit. The resulting Fokker-Planck PDE is non-standard as its drift and diffusion terms are non-local, non-linear and degenerate. A rigorous analysis of this PDE allows to partially justify the efficiency of the CBO algorithm.

2 Models

In the following we give some examples for Global Optimization Algorithms and Consensus Formation approaches that are related to CBO. Then, we discuss the dynamic of the CBO algorithm.

2.1 Examples for Global Optimization Algorithms

The general goal of global optimization algorithms is the following:

For given function $f\colon \mathbb{R}^d \to \mathbb{R}$ *find* $\overline{x} = \operatorname{argmin}_{x\in\Omega} f(x)$, *where* $\Omega \subset \mathbb{R}^d$ *and* d *denotes the dimension of the state space.*

The metaheuristic algorithms in the context of global optimization can be classified with the help of the following general properties:

- allowing for irregular target functions,
- usage of gradient information,
- guarantee of finding global minimum,
- agent-based / collective behaviour,
- resilience to finding local minima,
- incorporating stochasticity.

In the following we shortly discuss the well-known algorithms called Simulated Annealing and Particle Swarm Optimization.

Simulated Annealing [13]: the particle dynamics are given by

$$dX_t = -\nabla f(X_t)dt + (1/\beta_t)dW_t,$$

where β_t is the optimal cooling schedule. It uses gradient information, i.e., the target function therefore must be sufficiently regular. In addition, a stochastic influence enters the system in form of the Brownian motion. In certain settings, convergence proofs for the simulated annealing procedure are available. Those are, for example, based on Gibbs distributions [8].

Particle Swarm Optimization (PSO) [6]: in this case the evolution follows the ODE system

$$\begin{aligned}
&\frac{d}{dt}X = V,\\
&\frac{d}{dt}V = -\alpha V + \beta(X_{lb} - X) + \gamma(X_{gb}T - X),\\
&\frac{d}{dt}X_{lb} = \delta(X - X_{lb})[I_n + \mathrm{diag}(sgn(f(X_{lb}) - f(X)))],\\
&X_{gb} = X_{lb}Q_j \qquad\qquad \text{where} \quad j = \mathrm{argmin} f(X_{lb_i}),
\end{aligned}$$

where T is a row vector having all entries equal to 1, and Q_i is a column vector having all elements equal to zero except the i^{th} element that equals 1. Further, α, β, γ and δ are parameters and X and V are matrices containing the position and velocity information of the agents. X_{lb} and X_{gb} denote local and global best information collected along the past function evaluations. Altogether, it is a deterministic agent-based model of second order which is using no gradient information. To the author's knowledge there is no convergence result available and it may happen that the algorithm is getting caught in a local minimum.

2.2 Examples for Consensus Formation Algorithms

In this class of algorithms the main features are:

- involving multiple agents,
- adaption of other opinions,
- rejection of other opinions,
- influence on neighbors.

In the following we introduce Abelson's model and the model by Hegselmann-Krause as representative of this class.

Abelson's (repeated averaging) model: the evolution of the agents is given by

$$X^i_{t+1} = \sum_j a_{ij}(X^i_t - X^j_t) = (AX_t)^i,$$

where a_{ij} is the rate of influence of X_t^i on X_t^j and $A = (a_{ij})_{i=1,\dots,N,j=1,\dots,N}$. The model involves multiple agents. If other features apply depends on the structure of the matrix A. Convergence proofs based on Markov chain theory can be found in [5] and the references therein.

Hegselmann-Krause (bounded confidence) model: in this case the dynamics are defined as

$$X_{t+1}^i = \frac{1}{|N_t^i|} \sum_{j \in N_t^i} X_t^j, \qquad N_t^i = \{j : |X_t^i - X_t^j| \leq \epsilon\}.$$

In this deterministic model, agents interact with their neighbors in a region given by the parameter $\epsilon > 0$. For this framework a convergence result can be found in [10].

2.3 Consensus-Based Global Optimization

The CBO algorithm [17] is a first-order minimization algorithm with smooth deterministic and multiplicative stochastic forcing. It combines advantages of the aforementioned global optimization and consensus formation algorithms. The target function f that is supposed to be minimized is reflected in the dynamic by means of the weighted average v_f. Altogether, the particle dynamic for N agents is given by

$$dX_t^i = -\lambda(X_t^i - v_f)H^\epsilon(f(X_t^i) - f(v_t))dt + \sqrt{2}\sigma|X_t^i - v_f|dW_t^i,$$

where $\lambda > 0$ is a drift parameter and σ is a parameter which allows to adjust the stochastic influence and $i = 1, \dots, N$. The mean v_f is given by

$$v_f = \frac{1}{\sum_i \omega_f^\alpha(X_t^i)} \sum_i X_t^i \omega_f^\alpha(X_t^i).$$

It is a weighted average that stresses positions with small function values. The system is supplemented with initial data $X_t^i(0) = X_0^i \in \mathbb{R}^d$, $i = 1, \dots, N$. The weight function ω_f^α is given as a power of the reciprocal function value. We use

$$\omega_f^\alpha(x) = \exp(-\alpha f(x)), \quad \alpha > 0,$$

as proposed in [17]. Further, $H^\epsilon : \mathbb{R} \to \mathbb{R}$ denotes a smooth regularization of the Heaviside function and W_t^i is the i^{th} component of an N-dimensional Brownian motion. Note that the positions corresponding to large function values are neglected by the weight if $\alpha \gg 0$.

This model combines several features of the above examples. In fact, we have a deterministic drift in the direction of the weighted average v_f and a stochastic term which is scaled with the distance of the agents to v_f. The stochastic influence

allows each agent to explore the state space. On the other hand, the drift works as a confining potential which forces the particles to concentrate at v_f. In the case that all particles are located at the position of v_f the right-hand side of the SDE equals zero. Thus, the dynamical system reaches a stationary state and the concentration cannot be reversed. The weighted average is the candidate for the global minimizer, since the weight function ω_f^α stresses agents with small target function values and neglects the positions of agents having large function values. In fact, the weight function is of Gibbs type. Thanks to v_f there is no need to memorize global and local best information of past evaluations, therefore allowing to pass formally to the mean-field limit.

3 Mean-Field Equation and Convergence Result

Assuming the propagation of chaos property, we can formally derive the so-called *Mc-Kean nonlinear process*

$$dX_t = -\lambda(X_t - v_f[\rho_t])\, H^\epsilon(f(X_t) - f(v_f[\rho_t]))\, dt + \sqrt{2}\sigma |X_t - v_f[\rho_t]| dW_t \tag{1a}$$

where the weighted average is given by

$$v_f[\rho_t] = \frac{1}{\int_{\mathbb{R}^d} \omega_f^\alpha d\rho_t} \int_{\mathbb{R}^d} x\, \omega_f^\alpha d\rho_t, \qquad \rho_t = \mathrm{law}(X_t). \tag{1b}$$

This may be equivalently expressed as the Fokker–Planck equation

$$\partial_t \rho_t = \Delta(\kappa[\rho_t]\rho_t) + \mathrm{div}(\mu[\rho_t]\rho_t), \tag{2}$$

with

$$\kappa[\rho_t](x) = \sigma^2 |x - v_f[\rho_t]|^2, \qquad \mu[\rho_t](x) = -\lambda(x - v_f[\rho_t]) H^\epsilon(f(x) - f(v_f[\rho_t])),$$

which describes the evolution of the law corresponding to the Mc-Kean nonlinear process $\{X_t \in \mathbb{R}^d \,|\, t \geq 0\}$. We note that the presence of v_f makes the Fokker–Planck equation nonlinear and nonlocal in both the convection and diffusion part. This is nonstandard in the literature and raises several analytical and numerical questions. A further discussion and the rigorous arguments of the mean-field limit can be found in [4]. Additionally, the manuscript contains a rigorous investigation of the convergence behavior. In particular, the proof of the following proposition is provided.

Proposition 1 *Let $f \in \mathscr{C}^2(\mathbb{R}^d)$ satisfy the following conditions:*

(i) f may be expressed as the sum of two functions

$$f(x) = g(x) + h(x),$$

where $g \in \mathscr{C}^2(\mathbb{R}^d)$ is globally convex, i.e.,

$$\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq c_g |x - y|^2 \qquad \forall\, x, y \in \mathbb{R}^d,$$

for some constant $c_g > 0$, and $h \in \mathscr{C}^2(\mathbb{R}^d)$ satisfies $\|\nabla h\|_{Lip} =: c_h < c_g$.

(ii) There exist constants $c_0, c_f > 0$, such that

$$\Delta f \leq c_0 + c_f \, |\nabla f|^2 \quad \text{in } \mathbb{R}^d.$$

Then there exist α and $\lambda > 0$ such that we obtain uniform consensus for ρ_t as $t \to \infty$.

Numerical results comparing the behaviour of the mean-field equation and its particle counterpart are given in [17]. Figure 1 is taken from this reference. Note

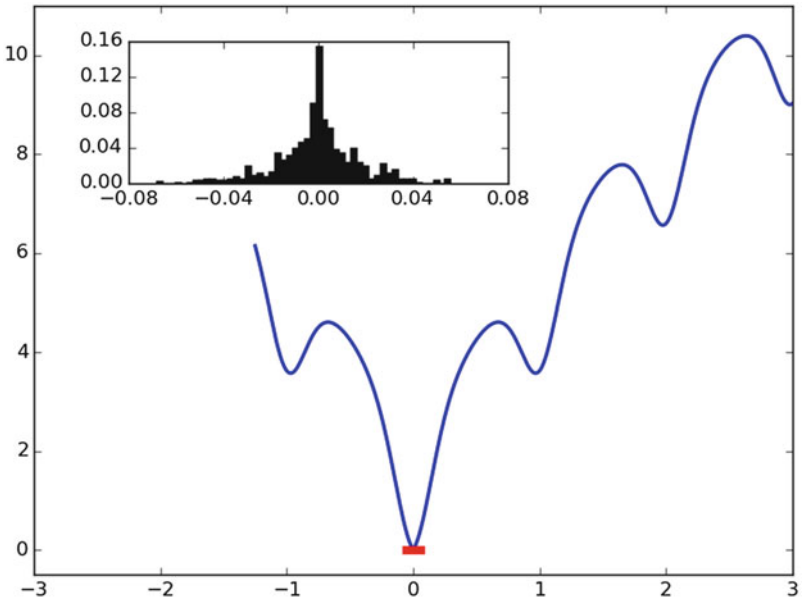


Fig. 1 Global optimization using CBO for the Ackley benchmark function. The support, i.e. $\{x \in \mathbb{R} : \rho(x) > 10^{-6}\}$ at time $T = 80$, of the mean-field solution in depicted in red. The histogram in the upper left corner shows the position of v_f at $T = 80$ for 500 samples. The time interval is $[0, 80]$, with a step-size $dt = 0.1$, $\alpha = 40$, $\sigma = 0.7$, $N = 50$. The details on the numerical schemes are given in [17]

that the positions depicted in the histogram are contained in the support of the mean-field solution as expected from the theory. As the global minimum is located at the center of the support, we expect the algorithm to be reliable in finding the global minimum of the benchmark. This illustrates the theoretical results. As the grid for mean-field computations in high dimensions exceeds usual memory capacities, the CBO algorithm is tested in [17] on a 20-dimensional state space only on the particle level. Again, the results show that the algorithms performs very reliable. Altogether, the results are promising and encourage the study of even more challenging global optimization problems with the help of the CBO algorithm.

Acknowledgements The author thanks J. A. Carrillo, Y. P. Choi, R. Pinnau, O. Tse and S. Martin for their constructive comments and suggestions. Moreover, she acknowledges financial support from the Nachwuchsring of the University of Kaiserslautern.

References

1. Ajmone Marsan, G., Bellomo, N., Gibelli, L.: Stochastic evolutionary differential games toward a systems theory of behavioral social dynamics. Math. Models Methods Appl. Sci. **26**, 1051–1093 (2016)
2. Bolley, F., Cañizo, J.A., Carrillo, J.A.: Stochastic mean-field limit: non-lipschitz forces and swarming. Math. Models Methods Appl. Sci. **21**, 2179–2210 (2011)
3. Boudin, L., Salvarani, F.: A kinetic approach to the study of opinion formation. ESAIM: Math. Modelling Numer. Anal. **43**, 507–522 (2009)
4. Carrillo, J.A., Choi, Y.P., Totzeck, C., Tse O.: An analytical framework for a consensus-based global optimization method. Preprint at https://arxiv.org/abs/1602.00220 (2016)
5. Chatterjee, S., Seneta, E.: Towards consensus: some convergence theorems on repeated averaging. J. Appl. Probab. **14**, 89–97 (1977)
6. Emara, H.M., Fattah, H.A.: Continuous swarm optimization technique with stability analysis. In: Proceedings of the American Control Conference, vol. 3, pp. 2811–2817 (2004)
7. French, P. Jr.: A formal theory of social power. Psychol. Rev. **63**, 181–194 (1956)
8. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
9. Ha, S.Y., Tadmor, E.: From particle to kinetic and hydrodynamic descriptions of flocking. Kinet. Relat. Models **1**, 415–435 (2008)
10. Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence models, analysis, and simulation. J. Artif. Soc. Soc. Simul. **5**(3), 1–33 (2002)
11. Karaboga, D., Gorkemli, B., Ozturk, C., Karaboga, N.: A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artif. Intell. Rev. **41**, 21–57 (2014)
12. Kennedy, J.: Particle swarm optimization. In: Encyclopedia of Machine Learning, pp. 760–766. Springer, Berlin (2011)
13. Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M.P.: Optimization by simulated annealing, Science **220**, pp. 671–680 (1983)
14. Locatelli, M., Schoen, F.: Global Optimization. Theory, Algorithms, and Applications. SIAM Series on Optimization. SIAM, Philadelphia (2013)
15. Mohan, B.C., Baskaran, R.: A survey: ant colony optimization based recent research and implementation on several engineering domain. Expert Syst. Appl. **39** 4618–4627 (2012)

16. Naldi, G., Pareschi, L., Toscani, G.: Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences. Springer Science & Businsess Media, New York (2010)
17. Pinnau, R., Totzeck, C., Tse, O., Martin, S.: A consensus-based model for global optimization and its mean-field limit. Math. Models Methods Appl. Sci. **27**, 183 (2017)
18. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. In: Swarm Intelligence. Springer, Berlin (2007)

Minisymposium: Return of Experience from Study Groups

Georges-Henri Cottet and Agnieszka Jurlewicz

Description

Study Groups with Industry are now a well established tool throughout Europe (and beyond) to promote industrial mathematics. This minisymposium gave the opportunity to PhD students who had recently participated in such Study Groups to report on their experience, present results obtained during the Study Group, and indicate the follow up of their work. This minisymposium also showed the diversity of topics and applications addressed in Study Groups. It gave visibility to the PhD students and advertised about Study Groups towards all public (students, faculty and industrialists) attending the ECMI conference.

The minisymposium featured the following talks:

- M. Kresoja, S. Rackov. University of Novi Sad. *Modelling Credit Scorecard with Time-Dependent Variables.*
  In order to meet regulatory requirement in the most efficient manner, all major banks are using Internal Rating Based (IRB) approach. As a result, expected loss that banks are exposed to, can be minimized and available funds for further investments can be increased, which all lead to profit maximization. Good credit risk management under the IRB approach includes optimization and development of three main risk drivers: probability of default (PD), exposure at default (EAD) and loss given default (LGD). In this research we focused on modelling the

G.-H. Cottet (✉)
Université Grenoble Alpes and CNRS, Grenoble, France
e-mail: Georges-henri.Cottet@univ-grenoble-alpes.fr

A. Jurlewicz
Wroclaw University of Technology, Wroclaw, Poland
e-mail: Agnieszka.Jurlewicz@pwr.edu.pl

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_65

probability of default of a borrower and proposed two approaches: survival analysis and machine learning algorithms.
Using real data set provided by one Serbian bank, we had implemented proposed solutions and compared them with the classical logistic regression. The overall results showed that both approaches outperform the logistic regression. Method with the most superior performance was Support Vector Machine. The performances of the models were evaluated by Gini coefficient and KS statistics. Time dependent variables with highest discriminatory power were inflation index and exchange rate.

- C. Courtès, G. Dusson, R. Hatchi, R. Molina, A. Thomas. University Paris XI (France). *Compressed sensing applied to radar imaging.*
For aircraft industry in which Airbus Group evolves, identifying features of a plane or of an helicopter which is flying over a given area is a crucial point. Those features are mathematically modelled by a sparse function g with only a few dominant point-like scattering centers localized on strategic positions of the airplane. In practice, companies do not have access to the sparse function but only to measurements made by radars. An important issue is then to reconstruct satisfactory signal from those measurements.
The compressed sensing is based on two steps: acquisition of directly compressed data and a nonlinear reconstruction. The first stage consists in obtaining random measurements (not matching initial signal structure at all), while the second phase performs the sparse recovery via a convex l1-minimization. Applying this algorithm to the radar imaging processing was the aim of this talk.
- M. Cruz, I. C. Lopes, A. Mota e Silva. Polytecnic of Porto (Portugal). *Packing and shipping cardboard tubes.*
SpiralPack—Manipulados de Papel, S.A. is a Portuguese company specializing in the production of tubes, angles, multipurpose packaging and cardboard formwork, which supplies their products to several sectors and is one of the main Iberian peninsula players in the production of standard and high performance cardboard tubes. In the context of tube manufacturing, there are certain processes that SpiralPack would like to improve. With a production totalling almost 17.5 million tubes/year, arising from more than 1500 different specifications corresponding to almost 100 tubes with different diameters, an important part of Spiralpack resources is allocated to packing and shipping processes. The company attended the 101st European Study Groups with Industry (ESGI), held in Faculdade de Ciencias da Universidade Nova de Lisboa to address the following questions: (1) Given an order for a certain tube specification possibly with a grouping request, what is the maximum number of tubes that can be packed inside a given container (usually the truck space) and how should they be positioned? (2) Given several pallets of tubes, what is the most efficient way to arrange them in a container? (3) Are there more efficient ways to group and pack tubes than the ones currently used? In this talk, we addressed briefly the options, research and developments made during and after the 5-day 101st ESGI.
- A. Levitt. University Paris Dauphine (France). *SEME 2015: Trajectory optimization for bathymetric navigation.*

This study group project took place in January 2015, and involved six PhD students and postdocs. The subject, provided by iXBlue, was bathymetric navigation. A submarine can measure its elevation with respect to the seafloor, and compare it to known maps to improve its localization. The question submitted to us was to find paths that allowed for a good positioning of the submarine in this way.

We investigated a number of different formalizations of this problem (a graph modelization, a probabilistic approach, and a formulation as an optimal transport problem), and proposed a solution that yielded a non-trivial but plausible behaviour (follow the ridges of the map).

Packing and Shipping Cardboard Tubes

Isabel Cristina Lopes and Manuel Bravo Cruz

Abstract Spiralpack - Manipulados de Papel, S.A. is one of the main Iberian peninsula players in the production of standard and high performance cardboard tubes. This company attended at 101st European Study Groups with Industry (ESGI) to address the following questions concerning their packing and shipping processes: Given an order for a certain tube specification, possibly with a grouping request, what is the maximum number of tubes that can be packed inside a given container (usually the truck space) and how should they be positioned? Given several pallets of tubes, what is the most efficient way to arrange them in a container? In this work we show an industrial mathematics approach to these challenges, as well as some insight on the software developed to help Spiralpack addressing those questions.

1 Problem Description

Spiralpack - Manipulados de Papel, S.A. is a Portuguese company specializing in the production of tubes, angles, multipurpose packaging and cardboard formwork. In the context of tube manufacturing, there are certain processes that Spiralpack would like to improve. Producing almost 17.5 million tubes/year, with more than 1500 different specifications and 100 different diameters, an important part of Spiralpack resources is allocated to packing and shipping processes. Currently, tubes with the same dimensions are grouped in two different ways: in a honeycomb shape or rectangular shape, either vertically or horizontally. The way the tubes are packed may depend on the client specification, the tube size, and the transportation vehicle dimensions. After the tubes are grouped, they are placed on pallets either in vertical or horizontal orientation and, in some cases, placed in rectangular cardboard boxes. Afterwards, the pallets and boxes must be packed in a container for transportation. In the context of 101st European Study Groups with Industry (ESGI), the company

I.C. Lopes (✉)
LEMA/ISCAP/IPP, Rua Jaime Lopes Amorim, 4465-004 S.Mamede Infesta, Portugal
e-mail: cristinalopes@iscap.ipp.pt

M.B. Cruz
LEMA/ISEP/IPP, Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal
e-mail: mbc@isep.ipp.pt

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_66

challenged the participants with two main questions: First, given an order for a certain tube specification, what is the maximum number of tubes that can be packed inside a given container (usually the truck space) and how should they be positioned? Second, given several pallets of tubes, Spiralpack would like to know an efficient way to arrange them in a container.

2 Palletizing Cardboard Tubes

The problem of arranging the tubes in vertical orientation into one pallet is equivalent to the problem of packing circles of equal diameter d in a rectangle of size $l \times b$. This, as accordingly with Spiralpack representatives, all the tubes packed in one pallet have the same dimensions and are placed with the same orientation. If the vertical orientation is used, $l \times b$ corresponds to the pallet width and length. For horizontal arrangements of the tubes in the pallet, the dimensions of the rectangle are given by one of the pallets dimensions and the maximum height of the stack. The problem for packing circles in a rectangle is a classical one in mathematics (e.g. [5, 6]). However, in the context of this work, the solution must be build upon uniform and simple configurations, as the packing process in Spiralpack is handmade. As so, we selected two uniform configurations that give some of the densest packing on arranging the circles: a *rectangular/square packing arrangement*, where the centres of the circles are horizontally and vertically aligned, and an *hexagonal packing arrangement*, where the centres of the circles are aligned in one dimension but vertically offset in the other.

As illustrated in Fig. 1, the dimension is compacted by a factor of $\frac{\sqrt{3}}{2} \approx 0.866$ for each row stacked vertically offset except the first. The same figure illustrates that for an hexagonal circle packing with k **rows** of n **circles** each of **diameter** d,

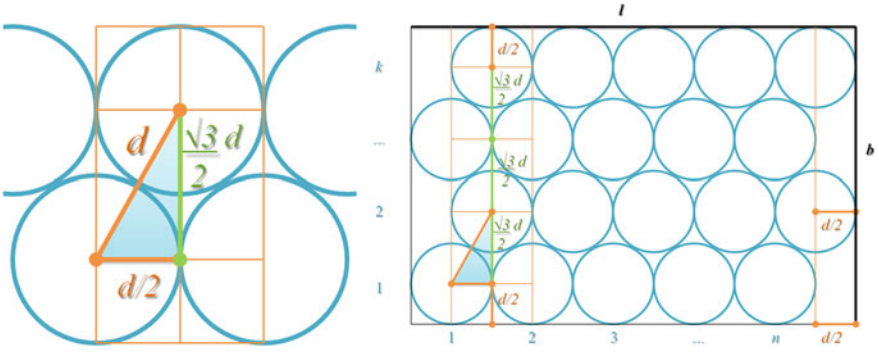


Fig. 1 Measuring the compactness of the hexagonal packing arrangement *(left)* and the dimensions of the hexagonal packing arrangement *(right)*

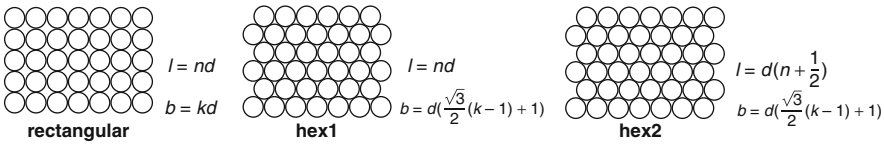


Fig. 2 Types of circle packing arrangement considered

the **vertical dimension** b has width,

$$b = \frac{d}{2} + (k-1) \times \frac{\sqrt{3}}{2} d + \frac{d}{2} = d\left(\frac{\sqrt{3}}{2}(k-1) + 1\right)$$

and **the horizontal dimension** l has length,

$$l = n \times d + \frac{d}{2} = d\left(n + \frac{1}{2}\right).$$

Depending on the rectangle dimensions and diameter of the circle, there may not be enough room for the last column of circles in the hexagonal packing arrangement. In this case, that we call **hex1**, the length is only $l = nd$, while the width b is the same as in the previous arrangement of the circles denoted as **hex2**. It is simple to demonstrate that the **hex2** and the **rectangular** arrangement comprises $k \times n$ circles packed, while the **hex1** arrangement packs $k/2 \times (2n-1)$ or $(k-1)/2 \times (2n-1) + n$ depending on k being even or odd, resp. (Fig. 2).

There is also another factor to consider when palletizing, which is the relative position in which the tubes are put in the pallets. Spiralpack uses tubes of broad range of diameters and thickness, and most of them can be placed in the pallets using the three natural positions, that is, with their length always parallel or perpendicular to one of the pallet dimensions. Note that for some of the combinations of diameter, length and thickness, not all of these positions are allowed. For example, tubes with length superior to the container height can only be placed in horizontal position and they need two or more EUR-pallets side by side to support them. However, generalizing the procedures considered on this work, namely, considering pallets with dimensions that are multiples of EUR-1/EUR-3 pallets, easily addresses the problem of tubes wider than the euro-palette dimensions or container height.

Once the best circle packing and position is chosen for a given reference, tubes can be stacked, layer over layer, as long as the maximum height is not exceeded. The previous procedure may be repeated to take advantage of the remaining space, resulting on a new layer with different orientation, as may be seen on Fig. 3.

Once the palletization patterns are defined for each of the possible orientation, the number of tubes per pallet and the dimensions of the occupied space become known. Many times, is important to know which is the most effective way to transport some fixed number of tubes. This may be accomplished by several ways, but usually is

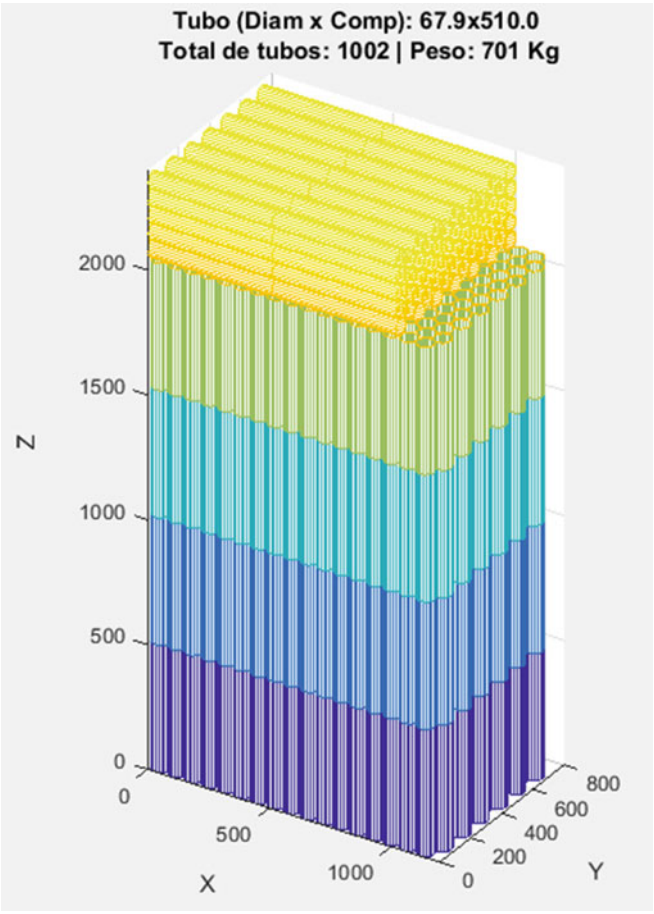


Fig. 3 Layout for packing 67.9 × 510 mm tubes in a EUR-3-pallet

important to occupy the least space possible. Occupied space can be measured as area in the truck floor, or can be considered in three dimensions if the pallets could be stacked over each other. After knowing which are the different possibilities to fill the pallets with a given type of tube, the previous question may be addressed using an integer programming model as follows. Consider i a possible palletization pattern, obtained using the expressions described in the previous section. Let n_i be the number of tubes contained in palletization pattern i, a_i its bottom area and v_i its volume. Let the integer variables x_i represent the number of pallets of type i to be packed in the truck, in order to meet a certain demand n. The objective function, (1), may be settled as the minimization of the total area occupied in the truck floor ($c_i = a_i$), but it can also be replaced with the minimization of the total volume occupied by the pallets $c_i = v_i$. This was also embedded on the software provided to Spiralpack, as the problem dimension is small (i usually less than 10) and it

gives an additional outcome that may be useful to Spiralpack packing optimization process.

$$\min \sum_i c_i x_i, \quad \text{s.t.} \sum_i n_i x_i \geq n, \quad x_i \in \{0, 1, \ldots, n\} \tag{1}$$

3 Container Loading

Considering the second question posed by Spiralpack, *"Given several pallets of tubes, what is the most efficient way to arrange them in a container?"*, we propose an heuristic to decide, given a set of orders, the layout and location of pallets inside the truck. According to the company representative, the objective here is to minimize the empty space in the delivers of several pallets of cardboard tubes to their clients address. This question may be tackled through an operational research point of view, in particular, within the three-dimensional container loading problem (CLP), which is a well known industrial problem that deals with the optimal utilization of container space for distribution of goods. A survey on this type of problems may be found, for example, in [8] or [2], among others. For sake of brevity, we use the CLP classification according to [7]. As so, we may classify the container loading problem in four variants: the Strip Packing Problem (SPP), the Bin Packing Problem (BPP), the Multiple Container Loading Problem (MCLP) and the Knapsack Loading Problem (KLP). According to [3], SPP consists on a container with two of its dimensions fixed and the third dimension is a variable of the problem. Thus, the problem consists of deciding how to load all boxes of different sizes inside the container, so that the variable dimension is minimized. In BPP, there are multiple containers (or bins) of the same fixed sizes and costs, and the problem consists of deciding how to load all boxes, so that the total number of used containers is minimized. Unlike BPP, in MCLP the containers do not necessarily have the same sizes and costs, and the problem consists of deciding how to load all the boxes, so that the total cost of the chosen subset of containers to be loaded is minimized. At last, on KLP, there is only one container of a fixed size, each box has an associated size and value, and the objective is to load a subset of boxes inside the container so that the total value of the loaded boxes is maximized. Obviously, if each box value is proportional to the volume of the box, then the problem consists of maximizing the total volume of the loaded boxes or, equivalently, minimizing the empty spaces left in the container. It is known that the one-dimensional knapsack problem is NP-hard [4] and, as so, also the three-dimensional variant is NP. This leads several authors to use heuristics in order to find a feasible solution for the CLP [1].

There were some practical considerations that we took into account, as they are very important for Spiralpack current procedures. According to the company representative, the purposed solution should obey the following conditions: (i) It should be a LIFO (Last In, First Out) system to assure a proper balance of the cargo inside the container, and because the first clients to serve in a given route must have its cargo on the back of the truck, in such a way that the remaining pallets will form a compact set with no holes between them; (ii) It should consider that some pallets may be stackable, others may not. Obviously this characteristic depends on the resistance of the tubes inside the pallet that is checked when the order is palletized; (iii) For each client there is a maximum height that the overlapped pallets cannot surpass; (iv) For stability reasons, even if a pallet is stackable, it is not possible to put it below another one with greater width. Also, the cargo inside the container must be properly balanced in respect to the longitudinal axis.

Beside those specifications, there are issues regarding the type of trucks that Spiralpack uses to deliver the orders. First, all their trucks allow side loading and, with exception for the last row all the pallets must be removed by the side. This implies that the orientation of pallets inside the truck is fixed. Second, both types of pallets used by Spiralpack have a width that is half of the container width. That means that there is no need for optimization on that dimension. As stated before, there are several references in the literature that treat this (NP-hard) problem, many of them by mean of heuristics. However, none of them fits all the constrains that are related with the Spiralpack side loading problem. As consequence, and being on a short 5-day ESGI approach, we decided to tackle this problem with a simple heuristic. The two-step approach proposed firstly does a vertical groupage of pallets. After this phase, we get a kind of SPP with some additional constrains, as the cargo must be properly balanced with respect to the longitudinal axis. Consequently, we developed a second algorithm that performs a subsequent longitudinal groupage taking into account the problem constrains. This procedure is explained on the following lines.

Let Lc, Hc be the current available length and height of the container, C_i, $i = 1, \ldots, n$ be the set of clients to serve in one journey, sorted by their relative position on the journey routing. For a given client C_i, define H_i as the maximum height allowed for that client ($H_i < Hc$) and for a pallet p_{ij} define its length by l_{ij}, its height by h_{ij} and if it is stackable by the binary variable s_{ij}. Define also the available length on the container as $AL = (AL_l, AL_r)$, where AL represents the length of the available empty space on the left and right side of the container. Then the heuristic developed, may be described in two steps: on a first iteration do some vertical consolidation (Algorithm 1) and then distribute the new consolidated pallets in the container, using an algorithm that balances the cargo concerning the both sides of the container (Algorithm 2).

Algorithm 1 Vertical consolidation

procedure SP_Vert_Cons(C, H, l, h, s)
Initialize $Group(i,:) = [1 : length(h_{ij}]$
for $i = 1 : length(C_i)$
  for every p_{ij} such that $s_{ij} == 0$
    find($\min_{k:s_{ik}==1} (H_i - h_{ik} - h_{ij} > 0 \& l_{ik} > l_{ij})$)
    if $\sim isempty(k)$
      update $h_{ik} = h_{ik} + h_{ij}$ & $s_{ik} = 0$ & $Group(i,k) = [Group(i,k), Group(i,j)]$
      delete $l_{ij}, h_{ij}, s_{ij}, Group(i,j)$
    end if
  end for
  for every p_{ij} such that $s_{ij} == 1$
    while $\sim isempty(find(\min_{k:s_{ik}==1} H_i - h_{ik} - h_{ij} > 0 \& l_{ik} > l_{ij}))$
      update $h_{ik} = h_{ik} + h_{ij}$ & $s_{ik} = 0$ & $Group(i,k) = [Group(i,k), Group(i,j)]$
      delete $l_{ij}, h_{ij}, s_{ij}, Group(i,j)$
    end while
  end for
end for
sort(l) in a non-increasing way and **update** Group,l,h.

Return Group, l, h.

Algorithm 2 Horizontal distribution

procedure SP_Hor_Dist(C, l, AL)
Initialize $T(i,j) = zeros(length(C_i), max(length(p_{ij})))$ t=1;
for $i = 1 : length(C_i)$
  for $j = 1 : length(p_{ij})$
    if $l_{ij} < max(AL)$
      if $AL(1) \leq AL(2)$
        update $AL(2) = AL(2) - l_{ij}$;
        $side(i,j) =$'R';
        $T(i,j) = t$;
      else
        update $AL(1) = AL(1) - l_{ij}$;
        $side(i,j) =$'L';
        $T(i,j) = t$;
      end if
    else
      Display 'Container is full: Adding another container'
      t=t+1;
      AL=[Lc,Lc];
      update $AL(2) = AL(2) - l_{ij}$;
        if $AL(2) < 0$
          break
        end if
      $side(i,j) =$'R';
      $T(i,j) = t$;
    end if
  end for
end for

Return T, side.

This 2-step procedure ensures that all the Spiralpack needs pointed earlier are fulfilled. In fact Algorithm 1, together with the assumption that the clients are sorted by their relative position on the journey routing, guarantee that points 1–3 are met as well as the first part of the fourth point. The Algorithm 2 ensures that the grouped pallets are placed in such a way that the client pallets are reasonably distributed by the both sides of the truck. There should be noted that we considered the length of the pallet as the dimension of choice for the balance finding. Nevertheless, this algorithm may be adapted for considering the balance in terms of the pallets weight.

4 Conclusions

This ESGI problem was very challenging, and since we had only 5 days to solve it, we had to decide for an approach that seemed to us to be more profitable in terms of the company's objectives. The process was afterwards audited by Spiralpack, and it was possible to proceed the initial work done at the 5-days ESGI. Spiralpack contracted a subsequent software development based on these initial ideas that resulted on a software that is now available to the company. The problem was successfully addressed, and modelled in a clever way that allowed to solve it recurring only to low complexity algorithms, involving very few computational resources. As so, the result is an easy to use software, with very fast computation times, that helps Spiralpack improve their efficiency and planning times. Finally, it should be highlighted that all this work was made entirely by mathematicians.

Acknowledgements The authors would like to thank Engineering Mathematical Laboratory (LEMA) and Spiralpack for supporting this work. The authors would like to thank all the participants of the 101st European Study Group for contributions on this subject, namely, Adérito Araújo, Ana Ribeiro, Fábio Chalub, Fernando Pestana da Costa, Luís Trabucho and Michael Grinfeld.

References

1. Bischoff, E., Ratcliff, M.: Issues in the development of approaches to container loading. OMEGA Int. J. Manag. Sci. **23**(4), 377–390 (1995)
2. Bortfeldt, D., Wäscher, G.: Constrains in container loading problem—a state-of-the-art review. Eur. J. Oper. Res. **229**(1), 1–20 (2013)
3. Junqueira, L., Morabito, R., Yamashita, D.: Three-dimensional container loading models with cargo stability and load bearing constraints. Comput. Oper. Res. **39**, 74–85 (2012)
4. Kellerer, H., Pferschy, U., Pisinger, D.: Knapsack Problems. Springer, Berlin (2004)
5. López, C., Beasley, J.E.: A heuristic for the circle packing problem with a variety of containers. Eur. J. Oper. Res. **214**(3), 512–525 (2011)
6. Maranas, C., Floudas, C., Pardalos, P.: New results in the packing of equal circles in a square. Discret. Math. **142**(1), 287–293 (1995)
7. Pisinger, D.: Heuristics for the container loading problem. Eur. J. Oper. Res. **141**(2), 382–392 (2002)
8. Wäscher, G., Haubner, H., Schumann, H.: An improved typology of cutting and packing problems. Eur. J. Oper. Res. **183**(3), 1109–1130 (2007)

Minisymposium: Simulation and Optimization of Water and Gas Networks

Gerd Steinebach, Tim Jax, and Lisa Wagner

Description

The minisymposium dealt with topics regarding aspects of modelling, simulating and optimizing flow behaviour within water and gas networks. These networks can consist of open channels and/or closed pipes. Corresponding examples are systems of rivers, gas and water supply or sewerage. The flow in every single network element is usually modelled by hyperbolic conservation laws or related simplifications. In addition, single flow reaches given must be supplemented by appropriate coupling and boundary conditions. This approach finally leads to PDAEs (partial differential algebraic equations) and requires very robust and efficient numerical methods for their solution. Also, optimizing network operation with respect to security of supply or energy consumption is of importance. Suitable methods to realize these aspects are an active field of research. Some of them were introduced in this minisymposium by including the following talks:

- T. Clees, K. Cassirer, B. Klaassen, I. Nikitin, L. Nikitina. Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI (Germany). *A new NLP analyzer and solver for gas networks with compressor stations.*

G. Steinebach (✉) • T. Jax
Fachbereich Elektrotechnik, Maschinenbau, Technikjournalismus, Hochschule Bonn-Rhein-Sieg, Grantham-Allee 20, 53757 Sankt Augustin, Germany
e-mail: Gerd.Steinebach@h-brs.de; Tim.Jax@h-brs.de

L. Wagner
Department of Mathematics, Numerical Analysis and Scientific Computing, Technische Universität Darmstadt, Dolivostr. 15, 64293 Darmstadt, Germany
e-mail: Wagner@mathematik.tu-darmstadt.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_67

- A. Bermúdez, X. López and M.E Vázquez-Cendón. Universidade de Santiago de Compostela (Spain). *Finite volume schemes for solving transient models of gas transportation networks.*
- Y. Lu, J. Mohring, N. Marheineke. Fraunhofer-Institut für Techno- und Wirtschaftsmathematik ITWM (Germany). *Matrix interpolation for parametric MOR of gas pipeline networks.*
- G. Steinebach. Hochschule Bonn-Rhein-Sieg (Germany). *TWaveSim - A simulator for water works and supply networks.*
- L. Wagner. Technische Universität Darmstadt (Germany). *Second order implicit schemes for solving balance laws with applications to water supply networks.*
- B. Liljegren-Sailer, N. Marheineke. Friedrich-Alexander-Universität Erlangen-Nürnberg (Germany). *Structure preserving MOR for gas networks with active elements.*
- T. Jax. Hochschule Bonn-Rhein-Sieg (Germany). *Generalized ROW-Type Methods for Simulating Water-Supply Networks.*

Stability-Preserving Interpolation Strategy for Parametric MOR of Gas Pipeline-Networks

Yi Lu, Nicole Marheineke, and Jan Mohring

Abstract Optimization and control of large transient gas networks require the fast simulation of the underlying parametric partial differential algebraic systems. Surrogate modeling techniques based on linearization around specific stationary states, spatial semi-discretization and model order reduction allow for the set-up of parametric reduced order models that can act as basis sample to cover a wide parameter range by means of matrix interpolations. However, the interpolated models are often not stable. In this paper, we develop a stability-preserving interpolation method.

1 Introduction

For the efficient simulation and optimization of large transient gas networks parametric reduced order models (pROMs) play a crucial role. They might be obtained from the network model by linearization around specific stationary states wrt. the parameters of interest (e.g., boundary pressure, temperature), spatial semi-discretization and balanced truncation [4, 5]. Treating the pROMs as basis, a wide parameter range can be easily evaluated using matrix interpolation techniques [1, 2]. Since the interpolated models often suffer from instabilities [5], we propose a stability-preserving interpolation strategy in this paper. The basis idea is reformulating the parametric full order models (pFOMS) of differential algebraic systems originally given in descriptor form in the standard state-space form by help of singular value decomposition. Thereby, the system index is reduced while the finite eigenvalues are preserved. To the resulting systems of ordinary differential equations classical techniques from model order reduction and matrix interpolation can be applied, [1, 6].

Y. Lu • N. Marheineke (✉)
FAU Erlangen-Nürnberg, Lehrstuhl Angewandte Mathematik I, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: yi.lu@math.fau.de; marheineke@math.fau.de

J. Mohring
Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM), Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
e-mail: jan.mohring@itwm.fraunhofer.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_68

2 State-Space Form via Singular Value Decomposition

Proceeding from a general linear time invariant system (LTIS) of differential algebraic equations, we transform it into a system of ordinary differential equations (standard state-space form) by help of singular value decomposition (SVD). This transformation goes with an index reduction. Consider

$$\Sigma: \quad \mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^p$ account for the states, inputs and outputs, respectively. The matrix $\mathbf{E} \in \mathbb{R}^{n,n}$ is singular, the other matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ have corresponding dimensions.

Inserting the SVD $\mathbf{E} = \begin{bmatrix}\mathbf{W}_1 & \mathbf{W}_2\end{bmatrix} \mathrm{diag}(\boldsymbol{\Sigma}_\mathbf{E}, \mathbf{0}) \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}$ into (1) yields

$$\boldsymbol{\Sigma}_\mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_1(t) = \mathbf{A}_{11}\mathbf{x}_1(t) + \mathbf{A}_{12}\mathbf{x}_2(t) + \mathbf{B}_1\mathbf{u}(t), \tag{2a}$$

$$\mathbf{0} = \mathbf{A}_{21}\mathbf{x}_1(t) + \mathbf{A}_{22}\mathbf{x}_2(t) + \mathbf{B}_2\mathbf{u}(t), \tag{2b}$$

$$\mathbf{y}(t) = \mathbf{C}_1\mathbf{x}_1(t) + \mathbf{C}_2\mathbf{x}_2(t) + \mathbf{D}\mathbf{u}(t), \tag{2c}$$

where $\mathbf{x}_i = \mathbf{V}_i^\mathrm{T}\mathbf{x}$, $\mathbf{A}_{ij} = \mathbf{W}_i^\mathrm{T}\mathbf{A}\mathbf{V}_j$, $\mathbf{B}_i = \mathbf{W}_i^\mathrm{T}\mathbf{B}$, and $\mathbf{C}_i = \mathbf{C}\mathbf{V}_i$ with $i, j = 1, 2$.

In case that the system (1) is of index-1, $\mathbf{A}_{22}$ is regular. Consequently, we obtain an explicit form for $\mathbf{x}_2$ from (2b), i.e., $\mathbf{x}_2(t) = -\mathbf{A}_{22}^{-1}(\mathbf{A}_{21}\mathbf{x}_1(t) + \mathbf{B}_2\mathbf{u}(t))$, and can eliminate the variable from the equations, which implies the smaller system of ordinary differential equations

$$\boldsymbol{\Sigma}_\mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_1(t) = \mathbf{A}^s\mathbf{x}_1(t) + \mathbf{B}^s\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{C}^s\mathbf{x}_1(t) + \mathbf{D}^s\mathbf{u}(t), \tag{3}$$

where $\mathbf{A}^s = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$, $\mathbf{B}^s = \mathbf{B}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$, $\mathbf{C}^s = \mathbf{C}_1 - \mathbf{C}_2\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ and $\mathbf{D}^s = \mathbf{D} - \mathbf{C}_2\mathbf{A}_{22}^{-1}\mathbf{B}_2$. The state-space representation (3) can be viewed as the result from applying the projections $\mathbf{W}^s = \mathbf{W}_1$ and $\mathbf{V}^s = \mathbf{V}_1$ on the LTIS

$$\mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \tilde{\mathbf{A}}\mathbf{x}(t) + \tilde{\mathbf{B}}\mathbf{u}(t), \qquad \mathbf{y}(t) = \tilde{\mathbf{C}}\mathbf{x}(t) + \tilde{\mathbf{D}}\mathbf{u}(t), \tag{4}$$

with $\tilde{\mathbf{A}} = (\mathbf{I} - \mathbf{A}\boldsymbol{\Pi})\mathbf{A}$, $\tilde{\mathbf{B}} = (\mathbf{I} - \mathbf{A}\boldsymbol{\Pi})\mathbf{B}$, $\tilde{\mathbf{C}} = \mathbf{C}(\mathbf{I} - \boldsymbol{\Pi}\mathbf{A})$ and $\tilde{\mathbf{D}} = \mathbf{D} - \mathbf{C}\boldsymbol{\Pi}\mathbf{B}$, where $\boldsymbol{\Pi} = \mathbf{V}_2(\mathbf{W}_2^\mathrm{T}\mathbf{A}\mathbf{V}_2)^{-1}\mathbf{W}_2^\mathrm{T}$. Note that (1), (2) and (4) are equivalent.

Lemma 1 *Assume that* (1) *is a LTIS of index-1. Then the eigenvalues of the standard state-space formula* (3) *coincide with the finite eigenvalues of* (1).

In case that (1) is of index-2 or higher, $\mathbf{A}_{22}$ is singular. Then reformulating the system in a state-space form requires further singular value decompositions (among others of $\mathbf{A}_{22}$) and substitutions. For the technical details we refer to [4].

3 Stability-Preserving Matrix Interpolation

The standard state-space formulation is the basis for our stability-preserving matrix interpolation strategy. Consider a sample of pFOMs Σ_k of the form (1) wrt. different parameter settings $\mathbf{p}_k \in \mathscr{P}$, $k = 1, \ldots, N$. We compute the pROMS $\Sigma_{r,k}$ of order $r \ll n$ by applying the described state-space transformation ($\mathbf{W}_k^s$, $\mathbf{V}_k^s$) and classical balance truncation [6] for model order reduction ($\mathbf{W}_{r,k}$, $\mathbf{V}_{r,k}$) – i.e., by using the (combined) projections $\mathbf{W}_k = \mathbf{W}_k^s \mathbf{W}_{r,k}$ and $\mathbf{V}_k = \mathbf{V}_k^s \mathbf{V}_{r,k}$,

$$\Sigma_{r,k}: \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_{r,k}(t) = \mathbf{A}_{r,k}\mathbf{x}_{r,k}(t) + \mathbf{B}_{r,k}\mathbf{u}(t), \qquad \mathbf{y}_{r,k}(t) = \mathbf{C}_{r,k}\mathbf{x}_{r,k}(t) + \mathbf{D}_{r,k}\mathbf{u}(t). \tag{5}$$

The matrix interpolation of the pROMs requires a number of further preparatory steps that are briefly presented in the following: the reduced parametric states are transformed into a common subspace in the spirit of [1, 2]. The resulting model systems are rewritten in modal form [3] where the eigenmodes have to be rearranged wrt. a given sequence.

Transformation of Local pROMs into Common Subspaces The states of the pFOMs are recombined during the index reduction due to SVD. During the model order reduction the states are again recombined as they are rearranged according to the Hankel singular values (HSVs) in decreasing order. The states related to the small pHSVs are truncated until the local reduced systems $\Sigma_{r,k}$ have the same order r. Thus, the projections $\mathbf{W}_k$, $\mathbf{V}_k$ for (5) usually span different rank-r subspaces in $\mathbb{R}^n$. Consequently, the states $\mathbf{x}_{r,k}$ have in general no common physical interpretation. Therefore, it is necessary to represent the local systems in a common rank-r subspace. We follow the procedure by [1, 2] where the common projection matrix (basis vectors of the common subspace) is assembled from the local ones. Choosing an appropriate state transformation $\mathbf{T}_k$ that allows for permutation, rotation and length distortion of the basis vectors (for details see [5]), we obtain the local pROM $\bar{\Sigma}_{r,k}$ in the common subspace with the matrices $\bar{\mathbf{A}}_{r,k} = (\mathbf{T}_k)^{-1}\mathbf{A}_{r,k}\mathbf{T}_k$, $\bar{\mathbf{B}}_{r,k} = (\mathbf{T}_k)^{-1}\mathbf{A}_{r,k}$ and $\bar{\mathbf{C}}_{r,k} = \mathbf{C}_{r,k}\mathbf{T}_k$.

Presentation of Local pROMs in Modal Form Consider a parameter setting $\mathbf{p} \notin \{\mathbf{p}_1, \ldots \mathbf{p}_N\}$. A matrix interpolation at $\mathbf{p}$ with positive weights ω_k

$$\Sigma_r(\mathbf{p}): \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_r(t) = \mathbf{A}_r(\mathbf{p})\mathbf{x}_r(t) + \mathbf{B}_r(\mathbf{p})\mathbf{u}(t), \quad \mathbf{y}_r(t) = \mathbf{C}_r(\mathbf{p})\mathbf{x}_r(t) + \mathbf{D}_r(\mathbf{p})\mathbf{u}(t) \tag{6a}$$

$$\mathbf{M}_r(\mathbf{p}) = \sum_{k=1}^{N} \omega_k(\mathbf{p})\, \mathbf{M}_{r,k}, \quad \omega_k(\mathbf{p}) > 0 \tag{6b}$$

that is based on the local systems $\bar{\Sigma}_{r,k}$, i.e., $\mathbf{M} \in \{\bar{\mathbf{A}}, \bar{\mathbf{B}}, \bar{\mathbf{C}}, \mathbf{D}\}$, may carry in instabilities since the interpolated matrix $\mathbf{A}_r$ may contain eigenvalues with non-negative real parts, although all $\bar{\Sigma}_{r,k}$ are stable. This drawback can be overcome, if all $\bar{\mathbf{A}}_{r,k}$ are transformed into the modal form by applying respective transformations $\mathbf{S}_k$ with the eigenvectors as column vectors. The modal form $\breve{\mathbf{A}}_{r,k} = (\mathbf{S}_k)^{-1}\bar{\mathbf{A}}_{r,k}\mathbf{S}_k$ is the real block-diagonal Schur form with 1×1 and 2×2-blocks on the diagonal where the 1×1 and 2×2-blocks contain real and complex-conjugate eigenvalues, respectively. Interpolating $\breve{\mathbf{A}}_{r,k}$ is the same as interpolating eigenvalues. The resulting eigenvalues at $\mathbf{p}$ have negative real parts due to the condition $\omega_k(\mathbf{p}) > 0$ in (6). The local pROM $\breve{\Sigma}_{r,k}$ in modal form consists additionally of $\breve{\mathbf{B}}_{r,k} = (\mathbf{S}_k)^{-1}\bar{\mathbf{B}}_{r,k}$ and $\breve{\mathbf{C}}_{r,k} = \bar{\mathbf{C}}_{r,k}\mathbf{S}_k$.

Reordering of Local Eigenmodes Consider the pROMs $\breve{\Sigma}_{r,k}$ in modal form. Since the (Schur) blocks of the system matrix $\breve{\mathbf{A}}_{r,k}$ are given in any order, it is necessary to reorder them wrt. a prescribed referential sequence of eigenvalues at $\mathbf{p}_{k_0}$, $k_0 \in \{1, \ldots, N\}$ before performing the interpolation (6).

Definition 1 (Correlation, MAC-Value) Let $\mathscr{I} = \{1, \ldots, m\}$ be a index set. Consider $\bar{\mathbf{V}}, \mathbf{V} \in \mathbb{R}^{m \times m}$, where $\bar{\mathbf{v}}_{i'}$ and $\mathbf{v}_i$ denotes the i'-th and i-th column vectors in $\bar{\mathbf{V}}$ and $\mathbf{V}$, respectively. Then, $\mathbf{v}_i$ is said to be correlated to $\bar{\mathbf{v}}_{i'}$, if and only if

$$i' = \underset{j \in \mathscr{I}}{\arg\max} \left|\mathrm{corr}(\mathbf{v}_i, \bar{\mathbf{v}}_j)\right|, \qquad i = \underset{j \in \mathscr{I}}{\arg\max} \left|\mathrm{corr}(\mathbf{v}_j, \bar{\mathbf{v}}_{i'})\right|,$$

with the MAC-value given by $\mathrm{corr}(\mathbf{w}, \mathbf{v}) = \mathbf{w}^{\mathrm{T}}\mathbf{v}/(\|\mathbf{w}\|_2\|\mathbf{v}\|_2)$ for $\mathbf{w}, \mathbf{v} \in \mathbb{R}^m$.
The mapping between the eigenvalues of the system matrix at $\mathbf{p}_k$, $k \in \{1, \ldots, N\} \setminus \{k_0\}$ and the ones at $\mathbf{p}_{k_0}$ can be determined by the MAC-values of the corresponding eigenvectors (cf. Definition 1). Note that if an eigenvector at $\mathbf{p}_k$ is not correlated to any eigenvector at $\mathbf{p}_{k_0}$, there exists always a second uncorrelated eigenvector. This occurs when the dynamics of the pROM $\Sigma_r(\mathbf{p})$ changes very rapidly in a surrounding of $\mathbf{p}_k$ or $\mathbf{p}_{k_0}$ and is known as mode veering phenomena [7]. In such a case, a new interpolant (pROM) associated to a parameter setting between $\mathbf{p}_k$ and $\mathbf{p}_{k_0}$ is desired. Presupposing that all eigenvectors are correlated we perform a reordering. Assume that the eigenmode $(\lambda_i, \mathbf{s}_i)_k$ is correlated to $(\lambda_{i'}, \mathbf{s}_{i'})_{k_0}$, $i, i' = 1, \ldots, r$, then the i-th eigenvector at $\mathbf{p}_k$ is exchanged with the i'-th one and its orientation is adjusted according to $\mathbf{s}_{i,k} = \mathrm{sign}(\mathrm{corr}(\mathbf{s}_{i,k}, \mathbf{s}_{i',k_0}))$. The prescribed reordering procedure can be expressed in terms of a permutation matrix $\mathbf{U}_{k,k_0}$, yielding the pROMs $\hat{\Sigma}_{r,k}$ in reordered form with matrices $\hat{\mathbf{A}}_{r,k} = (\mathbf{U}_{k,k_0})^{\mathrm{T}}\breve{\mathbf{A}}_{r,k}\mathbf{U}_{k,k_0}$, $\hat{\mathbf{B}}_{r,k} = (\mathbf{U}_{k,k_0})^{\mathrm{T}}\breve{\mathbf{B}}_{r,k}$ and $\hat{\mathbf{C}}_{r,k} = \breve{\mathbf{C}}_{r,k}\mathbf{U}_{k,k_0}$. Certainly, $\hat{\mathbf{A}}_{r,k}$ is still in modal form. Moreover, this property is inherited to the interpolated matrix $\mathbf{A}_r$ in (6) which is crucial for our stability-preserving interpolation strategy. An improvement for reordering complex eigenmodes can be found in [4].

Remark 1 The interpolation can also be performed in an appropriate tangent matrix-manifold [1, 5], i.e.,

$$\mathbf{A}_r(\mathbf{p}) = \exp\left(\sum_{k=1}^{N} \omega_k(\mathbf{p}) \ln\left(\hat{\mathbf{A}}_{r,k}\hat{\mathbf{A}}_{r,k_0}^{-1}\right)\right)\hat{\mathbf{A}}_{r,k_0}, \quad \omega_k(\mathbf{p}) > 0. \tag{7}$$

Since $\hat{\mathbf{A}}_{r,k}$ is diagonal with negative entries, (7) ensures that $\mathbf{A}_r$ inherits this property. This implies the stability of the respectively interpolated system at $\mathbf{p}$.

4 Application to Gas Pipeline-Networks

Network Modeling A network of pipelines is described as a directed graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ where the edges are represented by the pipes $e \in \mathscr{E}$. The set of vertices $\mathscr{V}$ consists of sources $\mathscr{V}_{\text{in}}$, sinks $\mathscr{V}_{\text{out}}$ and branching (neutral) nodes $\mathscr{V}_{\text{neu}}$. We model the gas dynamics in a pipe e by the one-dimensional isothermal transient Euler equations in terms of flow rate and pressure $q_e, p_e : [0, L_e] \times [t_0, t_{end}] \to \mathbb{R}$ with pipe length L_e, diameter D_e, roughness parameter κ_e, temperature T and specific gas constant R_s. The gas compressibility $z(p_e, T)$ and friction $\lambda(q_e)$ are empirically given by the AGA and Chen formulas, respectively (see [5] for details). At the branching nodes, mass conservation—known as first Kirchhoff law—and pressure equality (via auxiliary variables p) are imposed as coupling conditions. As boundary conditions we prescribe the flow rate $q(v, t) = f_v(t)$ at $v \in \mathscr{V}_{out}$ and the pressure $p(v, t) = f_v(t)$ at $v \in \mathscr{V}_{in}$. The system is supplemented with consistent initial conditions obtained from solving the stationary problem with the boundary conditions evaluated at time $t_0 = 0$.

We formulate the network model of partial differential algebraic equations as LTIS in descriptor form (1) by help of linearization and spatial discretization. Expanding the nonlinear network model around a stationary state that is specified by a certain parameter setting $\mathbf{p} \in \mathscr{P}$, we obtain a stationary subsystem and a linear transient (correction) subsystem (with initial zero conditions) in first order. Certainly, the coefficients in the linear subsystem depend not only on the stationary state but also on the model parameters. As spatial discretization we use a conservative first-order finite-volume method on a staggered grid to obtain small discretization stencils. The resulting parameter-dependent LTIS $\Sigma(\mathbf{p})$ is of index-2 and treated as pFOM (cf. Sect. 3). The boundary values are the inputs $\mathbf{u}$, whereas the outputs $\mathbf{y}$ are taken here exemplarily as the flow rates at the sources and the pressures at the sinks.

Numerical Results As benchmark for our stability-preserving interpolation strategy we consider the scenario investigated in [5]. In that work the temperature as well as the boundary pressure specified the parameter settings and thus the underlying samples for the interpolation. While the interpolation results were very

robust in case of temperature variations, changes in the pressure caused instabilities in the interpolated pROMs. Therefore, we focus here exclusively on the parametric dependence with respect to the boundary pressure.

The test network has the topology *Fork* with one source, one neutral node and three sinks over the time horizon $[0, t_{end}]$, $t_{end} = 20$ [h]. The four pipes $e_1, \ldots, e_4$ have different lengths $L_{e_{1,\ldots,4}} = (16, 45, 7, 38)$ [km], but the same diameter $D_e = 1$ [m] and roughness parameter $\kappa_e = 5 \times 10^{-5}$ [m]. The last one enters with the dynamical gas viscosity $\mu = 10^{-5}$ [kg/(ms)] in the Chen formula for the friction λ. The temperature is fixed $T = 273$ [K], and the specific gas constant is $R_s = 448$ [J/(kg K)]. The boundary conditions are given by $q(v,t) = 200$ [kg/s] at $v \in \mathscr{V}_{\text{out}}$ and $p(v,t) = p_0 + 0.5(1.05p_0 - p_0)(1 - \cos \pi t/t_{end})$ [bar] at $v \in \mathscr{V}_{\text{in}}$ with boundary pressure p_0. For the spatial discretization of the pipes the grid sizes $h_{e_{1,\ldots,4}} = (6.4, 15, 2.33, 12.67)$ [km] are used. Hence, the resulting FOMs are of order $n = 35$, we choose ROMs of order $r = 5$ that we compute with balance truncation. We solve the LTIS by means of the MATLAB routine ode15s (using the default values) and perform a linear interpolation. For the numerical handling an appropriate scaling of the underlying quantities might be advantageous. Note that the stability-preserving property of our proposed interpolation strategy is independent of model reduction technique and interpolation order. The choice of balance truncation and linear interpolation is here only because of simplicity. In fact it is accountable for the quantitatively high errors in the numerical approximation (Fig. 1). The errors might be improved by more sophisticated methods but this goes beyond the topic of this paper.

The set $\mathscr{P}$ of the parameter settings is determined by the boundary pressure p_0

$$\mathscr{P} = \{\mathbf{p} = p_0 \,|\, p_0 \in [45, 65]\}.$$

In [5] standard interpolation techniques gave only stable results in the interval $p_0 \in [55, 65]$ and failed outside. Our proposed interpolation strategy overcomes this limitation and can handle the extended regime. As parameter sample we consider

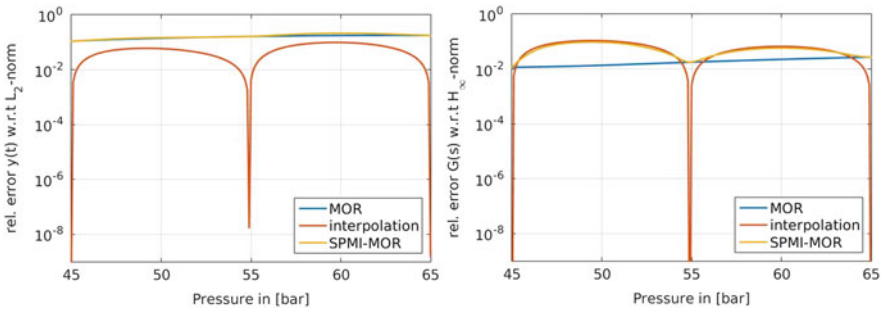


Fig. 1 *Left:* relative $\mathscr{L}^2(0, t_{end})$-error for the outputs; *right:* relative $\mathbb{H}_\infty$-error for the transfer functions. BT-MOR (error between FOM and ROM), interpolation (error between computed and interpolated ROM), SPMI-MOR (error between FOM and interpolated ROM)

$\mathbf{p}_k$, $k = 1, 2, 3$ given by $p_0 = 45, 54.9, 65$ [bar]. The approximation quality of our stability-preserving interpolation strategy is assessed in terms of the relative $\mathscr{L}^2(0, t_{end})$-error of the outputs and the relative $\mathbb{H}_\infty$-error of the transfer functions over $p_0 \in [45, 65]$. Figure 1 shows particularly the individual errors due to the model order reduction (error between FOM and ROM) and due to the interpolation (error between computed ROM and interpolated ROM) as well as the cumulative error (error between FOM and interpolated ROM). As expected the errors in the outputs are dominated by MOR. To improve the results concerning the transfer function, higher order interpolations are in the focus of recent investigations. Concerning the computational performance, the combination of MOR and our interpolation strategy is superior to MOR—similarly as for standard interpolation techniques. On first glance, the SVD to obtain the standard state-space form might seem to be computationally expensive, but the decomposition of a differential algebraic system into proper and improper parts for MOR is of similar costs.

5 Conclusion

In this paper we proposed a stability-preserving interpolation method and showed its applicability for parametric MOR of gas pipeline networks. The method overcomes the restriction being faced by standard interpolation techniques in literature. Future work deals with the extension and analysis of our method in cases of high-dimensional parameter spaces.

Acknowledgements The German CRC TRR 154 *Mathematical Modelling, Simulation and Optimization using the Example of Gas Networks* is acknowledged.

References

1. Amsallem, D., Farhat, C.: An online method for interpolating linear parametric reduced-order models. SIAM J. Sci. Comput. **33**(5), 2169–2198 (2011)
2. Geuss, M., Panzer, K., Lohmann, B.: On parametric model order reduction by matrix interpolation. In: European Control Conference (2013)
3. Kailath, T.: Linear Systems. Prentice Hall, Englewood Cliffs (1980)
4. Lu, Y.: Gramian projection-based interpolation strategy for parametric MOR of gas pipeline-networks. Ph.D. thesis, FAU Erlangen-Nürnberg (2017)
5. Lu, Y., Marheineke, N., Mohring, J.: Interpolation strategy for BT-based parametric MOR for gas pipeline-networks. In: MS&A – Modeling, Simulation and Application. Springer, Heidelberg (2017)
6. Moore, B.C.: Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Autom. Control **26**(2), 17–32 (1981)
7. Morand, H.J.-P., Ohayon R.: Fluid-Structure Interaction. Wiley, New York (1995)

A Structure-Preserving Model Order Reduction Approach for Space-Discrete Gas Networks with Active Elements

Björn Liljegren-Sailer and Nicole Marheineke

Abstract Aiming for an efficient simulation of gas networks with active elements a structure-preserving model order reduction (MOR) approach is presented. Gas networks can be modeled by partial differential algebraic equations. We identify connected pipe subnetworks that we discretize in space and explore with index and decoupling concepts for differential algebraic equations. For the arising input-output system we derive explicit decoupled representations of the strictly proper part and the polynomial part, only depending on the topology. The proper part is characterized by a port-Hamiltonian form that allows for the development of reduced models that preserve passivity, stability and locally mass. The approach is exemplarily used for an open-loop MOR on a network with a nonlinear active element.

1 Gas Pipe Network Modeling and Discretization

Gas networks consist of pipes and active elements, such as compressor stations and valves. They are modeled as coupled systems of nonlinear partial differential and algebraic equations. Especially, the space discretization of the pipes might lead to large dimensional systems whose efficient simulation and optimization motivate the use of MOR. In the following we exploit a structure-preserving MOR approach for connected pipe subnetworks that is based on an appropriate space discretization, index and decoupling concepts and a port-Hamiltonian formulation.

Pipe Network Model The topology of a pipe network can be modeled by a graph consisting of directed edges $\mathscr{E}$ connecting nodes $\mathscr{N}$ which we distinguish in those without and with pressure (boundary) conditions, $\mathscr{N}_J$ and $\mathscr{N}_S$, respectively. On each edge/pipe e the gas dynamics is described by simplified 1d Euler equations with

B. Liljegren-Sailer (✉) • N. Marheineke
Department Mathematik, FAU Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany
e-mail: sailer@math.fau.de; marheineke@math.fau.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_69

friction—in terms of pressure and mass flow $p^e, q^e : [x_L^e, x_R^e] \times \mathbb{R}^+ \to \mathbb{R}$ satisfying

$$\partial_t p^e(x,t) = -\frac{c^2}{a}\partial_x q^e(x,t), \qquad \partial_t q^e(x,t) = -a\partial_x p^e(x,t) - \frac{\lambda c^2}{2aD}\frac{q^e(x,t)|q^e(x,t)|}{p^e(x,t)}$$

with area a, and diameter D, and a possibly nonlinear friction factor λ, [3]. In particular, we presuppose an ideal gas such that the gas density ρ is expressed by $\rho = c^{-2}p$ with speed of sound c. Mass conservation and pressure continuity at the nodes are ensured by the coupling and boundary conditions, i.e.,

$$\sum\nolimits_{e\in\mathscr{E}(v)} n^e(v)q^e(v) = q^v_{dem}, \quad v \in \mathscr{N}$$

$$p^e(v) = p^{e'}(v), \quad v \in \mathscr{N}_J, \qquad\qquad p^e(v) = p^v_{sup}, \quad v \in \mathscr{N}_S,$$

where $\mathscr{E}(v)$ is the set of edges incident to the node v, and $n^e(v) \in \{1,-1,0\}$ depending on whether the pipe e starts or ends at the node v or none of both. The functions p^v_{sup} and q^v_{dem} prescribe pressure (supply) and flow (demand) at the boundary.

Space Discretization Introducing $p^e_{L/R}(t) \approx p^e(x_{L/R},t)$, $q^e_{L/R}(t) \approx q^e(x_{L/R},t)$, approximating pressure and mass flow by linear functions on $[x_L^e, x_R^e]$, $\Delta_x^e = x_R^e - x_L^e$, and applying a central stencil for the friction term, we deal with the following two-point space discretization for a pipe

$$\partial_t \frac{p^e_L + p^e_R}{2} = -\frac{c^2}{a}\frac{q^e_R - q^e_L}{\Delta^e_x}, \quad \partial_t \frac{q^e_L + q^e_R}{2} = -a\frac{p^e_R - p^e_L}{\Delta^e_x} - \frac{\lambda c^2}{4aD}\frac{(q^e_L + q^e_R)|(q^e_L + q^e_R)|}{(p^e_L + p^e_R)}.$$

Finer spatial meshes can be certainly obtained straightforward by subdividing a long pipe into several short ones. Note that the first equation represents the space-discrete mass conservation: apart from a constant scaling the left side equals the integral over the density, and the right side equals the mass flux over the boundaries of $[x_L^e, x_R^e]$.

The overall network topology can be characterized by the incidence matrix $\mathbf{A}_{all} \in \mathbb{R}^{|\mathscr{N}|\times|\mathscr{E}|}$, $[\mathbf{A}_{all}]_{ij} = n^{e_j}(v_i)$. Without loss of generality we compose $\mathbf{A}_{all}$ of the respective matrices $\mathbf{A}_k \in \mathbb{R}^{|\mathscr{N}_k|\times|\mathscr{E}|}$ associated to the nodes in $\mathscr{N}_k$, $k \in \{J,S\}$. Let $\mathbf{p}_k(t) \in \mathbb{R}^{|\mathscr{N}_k|}$ be the vector of node pressures belonging to $\mathscr{N}_k$, $k \in \{J,S\}$. Moreover, let the vectors of edge flows $\mathbf{q}_+(t), \mathbf{q}_-(t) \in \mathbb{R}^{|\mathscr{E}|}$ be entrywisely defined by $[\mathbf{q}_\pm(t)]_j = (q_R^{e_j} \pm q_L^{e_j})/2$. Collecting all states in $\mathbf{x}$ and all boundary conditions (inputs) in $\mathbf{u}$, our space-discretized pipe network model can be written as

$$\mathbf{E}\frac{d}{dt}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{x} = \left(\mathbf{p}_J, \mathbf{p}_S, \mathbf{q}_+, \mathbf{q}_-\right)^T, \quad \mathbf{u} = \left(\mathbf{p}_{sup}, \mathbf{q}_{dem}\right)^T \quad (1)$$

with

$$\mathbf{E} = \begin{pmatrix} |\mathbf{A}_J^T|/2 & |\mathbf{A}_S^T|/2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{D}_2 \\ -\mathbf{D}_1 {\mathbf{A}_J}^T & -\mathbf{D}_1 \mathbf{A}_S^T & -\mathbf{D}_F & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -|\mathbf{A}_J| & -\mathbf{A}_J \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_J \end{pmatrix}$$

and identity matrix $\mathbf{I}$. Here, $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_F$ are positive definite diagonal matrices, where $\mathbf{D}_F$ particularly represents the friction. Note that nonlinear friction models result in state dependent matrices $\tilde{\mathbf{D}}_F(\mathbf{x})$. In view of MOR in Sect. 2, we use here a linearization around the initial value, i.e., $\mathbf{D}_F = \tilde{\mathbf{D}}_F(\mathbf{x}(0))$, however the subsequent analysis and decoupling could analogously be performed for a nonlinear model. The matrix $\mathbf{Q}_J$ is sparse with some one-entries and represents the flow boundary conditions. The absolute value of a matrix is taken elementwisely.

General Input-Output System, Port-Hamiltonian Form Consider an input-output system in state representation of the following form

$$\mathbf{E}\frac{d}{dt}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t). \tag{2}$$

The same mapping from the input $\mathbf{u}$ to the output $\mathbf{y}$ can be described by different state representations, but its characterization in the frequency space, i.e., the transfer function $\mathbf{G}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$, if it exists, is unique.

Definition 1 A transfer function $\mathbf{G}_{sp}$ with $\lim_{s\to\infty} \mathbf{G}_{sp}(s) = \mathbf{0}$ is called strictly proper. Any transfer function $\mathbf{G}$ can be additively decomposed into its strictly proper part and a remainder, the polynomial part $\mathbf{G}_{pol}$, i.e., $\mathbf{G} = \mathbf{G}_{sp} + \mathbf{G}_{pol}$.
An input-output system with a representation (2) with regular $\mathbf{E}$ is strictly proper.

Definition 2 An input-output system is called passive, if $\mathbf{u}$ and $\mathbf{y}$ have the same dimension and for all $\mathbf{u}$ it holds $0 \le \mathbf{u}^T(t)\mathbf{y}(t)$ for all t.
In the following we focus on a subclass of port-Hamiltonian systems. Port-Hamiltonian systems inherit passivity and stability, for a general overview see [5].

Lemma 1 ([5]) *An input-output system in the form*

$$\mathbf{M}\frac{d}{dt}\mathbf{e}(t) = [\mathbf{J} - \mathbf{R}]\,\mathbf{e}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{B}^T\mathbf{e}(t) \tag{3}$$

with $\mathbf{M} = \mathbf{M}^T$, $\mathbf{J} = -\mathbf{J}^T$, $\mathbf{R} = \mathbf{R}^T$, $\mathbf{M} > 0$, $\mathbf{R} \ge 0$ *is stable and passive. It is a regular linear port-Hamiltonian system in co-energy form.*

Decoupled Formulation for Space-Discrete Pipe Network For the pipe network we consider an input-output system (2) consisting of the state equation (1) and its power conjugated output $\mathbf{y} = -(\mathbf{q}_{sup}, \mathbf{p}_{dem})^T$. Then $\mathbf{y}^T\mathbf{u}$ corresponds to the product of pressure and mass flow at the boundaries and can be interpreted as the supplied

power of the pipe system. Note that all these boundary quantities are required for the interconnection of a pipe (sub-)network with other (active) elements (cf. Sect. 3). Proceeding from (1) and using some refactorizations it can be shown that the strictly proper part of the system has a state representation as in (3), Lemma 1 with

$$\mathbf{e}(t) = \begin{pmatrix} \mathbf{p}_J(t) + \tilde{\mathbf{M}}\mathbf{p}_S(t) \\ \mathbf{q}_+(t) \end{pmatrix}, \qquad \mathbf{M} = \begin{pmatrix} \left(|\mathbf{A}_J|\mathbf{D}_1|\mathbf{A}_J|^T\right)/4 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2^{-1} \end{pmatrix}, \tag{4}$$

$$\mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{A}_J \\ -\mathbf{A}_J^T & \mathbf{0} \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{D}_2^{-1}\mathbf{D}_F \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{0} & -\mathbf{Q}_J \\ -\left(\mathbf{A}_S^T - \mathbf{A}_J^T\tilde{\mathbf{M}}\right) & \mathbf{0} \end{pmatrix},$$

where $\tilde{\mathbf{M}} = \left(|\mathbf{A}|_J\mathbf{D}_1|\mathbf{A}_J|^T\right)^{-1}|\mathbf{A}_J|\mathbf{D}_1^{-1}|\mathbf{A}_S|^T$.

Similarly, a topologically based representation of the polynomial part can be given. It is linear (regardless of the friction model) and also passive, [3]. Additionally, the tractability index of the system, which is 2, can be easily concluded, cf. also [2].

2 Model Order Reduction Framework

Starting from an input-output system of the form (2), the idea of projection based MOR is to find appropriate basis matrices $\mathbf{V}, \mathbf{W} : \mathbb{R}^{N\times n}$, $n \ll N$, both full column rank, such that the reduced low-dimensional model

$$\mathbf{W}^T\mathbf{E}\mathbf{V}\frac{d}{dt}\mathbf{x}_r(t) = \mathbf{W}^T\mathbf{A}\mathbf{V}\mathbf{x}_r(t) + \mathbf{W}^T\mathbf{B}\mathbf{u}(t), \qquad \tilde{\mathbf{y}}(t) = \mathbf{C}\mathbf{V}\mathbf{x}_r(t)$$

with transfer function $\mathbf{G}_r$ is a good approximation of the full-order model. We use here a Pade-type approximation constructed by Krylov-subspace methods [1].

Definition 3 For appropriately dimensioned matrices $\mathbf{S}$ and $\mathbf{R}$, the q-th Krylov-subspace is defined as $\mathscr{K}_q(\mathbf{S},\mathbf{R}) = \mathit{range}\left(\left[\mathbf{R}, \mathbf{S}\mathbf{R}, \mathbf{S}^2\mathbf{R}, \ldots, \mathbf{S}^{q-1}\mathbf{R}\right]\right)$.

Theorem 1 ([1]) *Let $s_0 \in \mathbb{C}$ be fixed. A reduced model constructed by the projection ansatz with $\mathbf{W} = \mathbf{V}$, $det(s_0\mathbf{E} - \mathbf{A}) \neq 0$, $det(\mathbf{V}^T(s_0\mathbf{E} - \mathbf{A})\mathbf{V}) \neq 0$ and*

$$\mathscr{K}_q\left(\left[s_0\mathbf{E} - \mathbf{A}\right]^{-1}\mathbf{E}, \left[s_0\mathbf{E} - \mathbf{A}\right]^{-1}\mathbf{B}\right) \subseteq \mathit{range}(\mathbf{V})$$

is a so-called Pade-type approximation with $(\mathbf{G} - \mathbf{G}_r)(s) = \mathscr{O}\left((s - s_0)^q\right)$.

Let $blkdiag(\mathbf{A}_1, \mathbf{A}_2)$[1] abbreviate the block-diagonal matrix with blocks $\mathbf{A}_1, \mathbf{A}_2$.

[1] $blkdiag(\mathbf{A}_1, \mathbf{A}_2) = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}$.

We distinguish three levels of structure-preserving model reduction methods:

Theorem 2 *A projection-based MOR for the strictly proper part of the space-discrete pipe network* (3)–(4) *preserves*

1. *... passivity and stability, if a Galerkin-type projection, i.e.,* $\mathbf{V} = \mathbf{W}$*, is used.*
2. *... additionally*[2] *the block-structure, if additionally* $\mathbf{V} = blkdiag(\mathbf{V}_1, \mathbf{V}_2)$.
3. *... additionally the local mass conservation-property, if additionally for a factorization (e.g., Cholesky factorization) of* $\mathbf{M}_{11} = \mathbf{L}\mathbf{L}^T$*:*
   $\mathbf{V}_1 = \mathbf{L}^{-1}\tilde{\mathbf{V}}_1$, $range(\tilde{\mathbf{V}}_1) \supseteq range(\mathbf{L}^{-1}\left([\mathbf{J}_{12}\mathbf{V}_2]\ \mathbf{B}_1\right))$, $\tilde{\mathbf{V}}_1$ *orthogonal.*

Proof To 1. and 2.: Since the port-Hamiltonian form (3) of Lemma 1 is kept under the projection, passivity and stability are inherited [5]—analogously the block structure.

To 3.: It needs to be shown that the prolongation of the reduced pressure $\mathbf{p}_r$ onto the original grid fulfills a space discrete conservation law, which is mass conservation along the edges, and respectively at the nodes, i.e.,

$$\partial_t \frac{\tilde{p}_L^e + \tilde{p}_R^e}{2} = -\frac{c^2}{a}\frac{\tilde{q}_R^e - \tilde{q}_L^e}{\Delta_x^e} \text{ for } e \in \mathscr{E}, \quad \sum\nolimits_{e\in\mathscr{E}(v)} n^e(v)\tilde{q}^e(v) = q_{dem}^v, \quad v \in \mathscr{N}.$$

Hereby, $\tilde{\mathbf{p}} = \hat{\mathbf{V}}\mathbf{p}_r$ for a prolongation $\hat{\mathbf{V}}$, and $\tilde{q}_{R,L}^e$ can be interpreted as a mass flux over the boundaries of $[x_L^e, x_R^e]$, if the mass conservation at the nodes holds. The latter can be deduced from the fact that only strictly proper parts are reduced, i.e., purely algebraic relations are kept in the reduced model. Therefore, it only remains to show that no projection error is introduced in the strictly proper part of the mass conservation along the edges. With some basic calculations, however, it can be shown that the reduction as suggested in 3. with $\mathbf{V} = blkdiag(\mathbf{V}_1, \mathbf{V}_2)$ leads (up to initial value projection errors) to equivalent models as a reduction with the basis $blkdiag(\mathbf{I}, \mathbf{V}_2)$, which leads to 'half-reduced models'. This finishes the proof. Summarizing, the reduction of point 3 can be read as follows: only the mass flow is directly reduced, whereby the mass conservation at the nodes is not violated for the prolongated reduced mass flows $\tilde{q}_{R,L}^e$. Then the thereby induced low-order pressure and mass conservation-equation on the edges are used.

3 Simulation of Network with Active Element and Discussion

As test case for our MOR approach we consider a gas network of two pipe subnets with a compressor, see Fig. 1. The boundary conditions of the closed-loop system are chosen to be a pressure condition at the left end and a mass flow condition at

[2] An enhanced Pade-approximation condition for the block-structure preserving reduction methods with the energy-conjugated output and $s_0 \in \mathbb{R}$ can be shown with the help of [1], which justifies the use of possibly enlarged block bases.

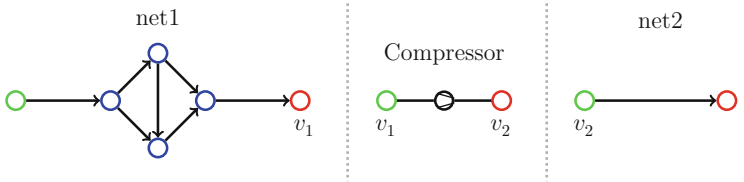


Fig. 1 Exemplary gas network consisting of two pipe subnets and a compressor. All pipes have circular cross-sections with diameter $D \in [1, 1.3]$ m. Both subnets together comprise about 290 km of pipe length: the pipes of net 1 are between 14 and 40 km, and the pipe of net 2 is 100 km long. Nodes with pressure and mass flow boundary conditions are marked in *green* and *red*, respectively

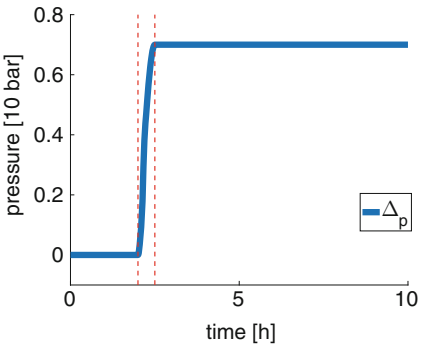


Fig. 2 Pressure-difference control Δ_p the compressor is driven with. It increases smoothly from 0 to 7 bar within 0.5 h

the right end, both constant in time. We initialize the system with the corresponding steady state. The transient simulation is done with a simplified linear friction model $(\lambda|q|/p)(x,t)q(x,t) \approx (\lambda|q|/p)(x,0)q(x,t)$ (cf. Sect. 1).

Compressor and Network Parameters Let p_{v_i}, q_{v_i} be the pressure and mass flow at node v_i, $i = 1, 2$ in Fig. 1. The compressor is driven by a pressure-difference control Δ_p, which implies a nonlinear flow consumption, modeled as in [4] with compressor-specific constant c_c, here $c_c = 3.93$, and isentropic coefficient $\gamma = 1.25$,

$$p_{v_2} - p_{v_1} = \Delta_p, \qquad q_{v_2} - q_{v_1} = -c_c \left(\left(\frac{p_{v_2}}{p_{v_1}} \right)^{\frac{\gamma-1}{\gamma}} - 1 \right) q_{v_1}. \tag{5}$$

The applied control Δ_p is sketched in Fig. 2. We use $c = 340$ m/s as speed of sound and model the friction factor λ with the Chen-Formula [4] giving $\lambda \in [0.008, 0.0095]$. This means that all pipes are 'technically rough'. The spatial discretization of the pipes is chosen as uniform with a step size of 0.5 km.

Coupling of Pipe Subnetworks The compressor model (5) can be formulated as an input-output mapping $\mathbf{y}_c = \mathbf{F}_c(t, \mathbf{u}_c)$ with $\mathbf{u}_c = (-q_{v_1}, p_{v_2})^T$ and $\mathbf{y}_c = (p_{v_1}, q_{v_2})^T$. We use an input-output coupling to interconnect the pipe subsystems to a closed-loop system. The coupling is realized by choosing the respective outputs of the compressor as inputs for net i at v_i, and conversely the inputs of the compressor as outputs of the net i at v_i, $i = 1, 2$ (i.e., power-conjugated output), see Fig. 1.

Numerical Results and Discussion The suggested mass conservative MOR with expansion frequency $s_0 = 0$ is applied for an open-loop reduction of the two separate pipe subnetworks. To avoid projection errors in the initial values, we use the superposition principle to decompose these systems into a non-dynamic part and a homogeneous part, which is actually reduced. Figure 3 shows a comparison of the reduced and direct simulation results. High fidelity, together with a significant order reduction can be observed. The computations are particularly performed with Matlab solver `ode15i.m` with default settings.

Concluding, standard MOR methods yield usually poor results, depending on the parameters. This is in accordance with the well-known observation that discretizations of hyperbolic equations that do not conserve the most relevant physics of the underlying model in some discrete sense may lead to unreasonable results, e.g., instabilities. Our approach overcomes this limitation: our results demonstrate the applicability of MOR for gas network simulations within practically relevant

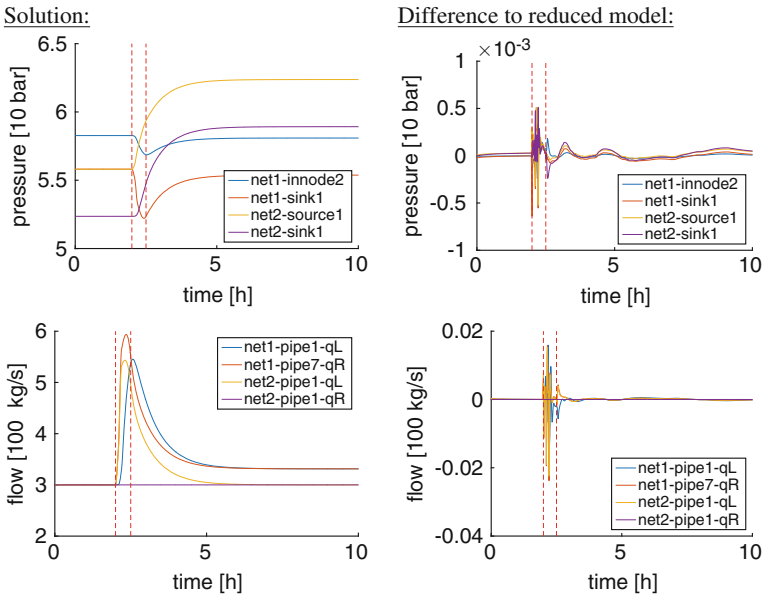


Fig. 3 Simulation results. *Left*: States over time at depicted nodes of the networks using a direct simulation with a strictly proper parts of the pipe nets of order (766 + 400). *Right*: Deviance of the direct and the reduced results. The reduced system has strictly proper parts of order (8 + 8)

parameter regimes, here for long ‘technically rough’ pipes. The key point is the use of structure preservation, and preliminary, an appropriate representation in terms of the decoupled formulation of Sect. 1.

Acknowledgements The funding by DFG CRC/Transregio 154, project C02 is acknowledged.

References

1. Freund, R.W.: On Pade-type model order reduction of j-Hermitian linear dynamical systems. Linear Algebra Appl. **429**(10), 2451–2464 (2008)
2. Grundel, S., Jansen, L., Hornung, N., Clees, T., Tischendorf, C., Benner, P.: Model order reduction of differential algebraic equations arising from the simulation of gas transport networks. In: Progress in Differential-Algebraic Equations, pp. 183–205. Springer, Berlin (2014)
3. Huck, C., Liljegren-Sailer, B., Marheineke, N., Tischendorf, C.: Discretization and MOR for PDAEs describing gas transport in networks (2016, Preprint)
4. Kolb, O.: Simulation and optimization of gas and water supply networks. Ph.D. thesis, TU Darmstadt (2011)
5. van der Schaft, A., Polyuga, R.V.: Structure-preserving model reduction of complex physical systems. In: Proceedings of the 48th IEEE Conference on Decision and Control (CDC), pp. 4322–4327. IEEE, Shanghai (2009)

Generalized ROW-Type Methods for Simulating Water Supply Networks

Tim Jax and Gerd Steinebach

Abstract In order to help water suppliers estimating and improving effectiveness of their facilities with respect to economy and ecology, simulating real pipe networks is of increasing importance. Corresponding models take into account relevant processes such as procurement and preparation of drinking water as well as its subsequent distribution. Resulting mathematical systems prove to be demanding and require research in advanced efficient and robust numerical methods that in particular allow for fast computation. In this context, generalized Rosenbrock-Wanner type methods introduced by Jax and Steinebach (J Comput Appl Math 316:213–228, 2017) seem to be an useful tool to solve arising differential-algebraic equations.

In this article, we investigate pros and cons when exploiting properties of these methods for computing problems that represent typical characteristics of pressurized flows given in water supply networks. Results are promising for tests with proper Jacobian structures. But they also motivate research in enhanced schemes.

1 Introduction

Simulating fluids in pipe networks is relevant for many applications. It allows to determine pressure distribution and flow velocities by solving hyperbolic conservation laws of mass and momentum such as the water hammer equations [1]:

$$h_t + \frac{c^2}{gA} q_x = 0 \tag{1a}$$

$$q_t + gAh_x = -\lambda \frac{q|q|}{2DA} \tag{1b}$$

T. Jax (✉) • G. Steinebach
Hochschule Bonn-Rhein-Sieg, Grantham-Allee 20, 53757 Sankt Augustin, Germany
e-mail: tim.jax@h-brs.de; gerd.steinebach@h-brs.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_70

Here, state-variables h and q denote piezometric head and volume flow. A and D represent a pipe's cross-sectional area and diameter. Parameter g is the gravitational constant, c is the sound velocity and λ corresponds to a friction coefficient.

Today, especially reproducing flow in networks of drinking water supply is important. Simulations help water suppliers to estimate efficiency of their facilities with respect to economical and ecological aspects. For example, they enable to determine optimized plans for operating controllable components such as pumps and valves in order to safe energy and resources. Realizing these plans via an automatic decision support system is currently pursued in the EWave research project that is funded by the German Federal Ministry of Education and Research.

However, simulating water supply systems is demanding. Models need to consider the process chain ranging from procurement and preparation of drinking water in water works to its subsequent distribution in networks. This requires to describe components mathematically, finally leading to a large system of partial differential-algebraic equations (PDAEs) that must be solved by appropriate numerical schemes. In this context, especially aspects of computing times are important. To apply the EWave decision support system successfully times must be minimized to refresh simulation data based on current flow conditions at frequent intervals and to provide staff in water works with resulting optimized operation plans in time.

A key to reduce computing times are improved schemes for solving differential-algebraic equations (DAEs). DAEs arise after spatial semi-discretization applied to the PDAE system using the method of lines. So far, solving them in the EWave project was based on semi-implicit Rosenbrock-Wanner (ROW) methods that proved to be efficient and robust simulating water supply systems [5]. However, a drawback is given by their effort computing exact Jacobians with every time-step.

In [2] generalized ROW-type (GROW) methods for DAEs were defined that allow for arbitrary Jacobian approximations with respect to differential parts. This enables to safe effort by keeping Jacobian entries constant for several time-steps, i.e. applying time-lagged information, or setting Jacobian entries to zero, i.e. using partial explicit integration. Below, we analyze efficiency of these methods solving tests that reproduce characteristics of fluids in water supply. Results will help to estimate if their use might improve computations in the EWave decision support system. In this context, we also focus on effects of different Jacobian structures.

The article is organized as follows: Sect. 2 describes generalized ROW-type methods. Section 3 regards test problems. Section 4 gives a conclusion and outlook.

2 Generalized ROW-Type Methods

DAEs that arise when simulating flow behavior in water supply networks are semi-explicit. They can be expressed by

$$y'(t) = f(y,z), \qquad f = f_N + f_S, \qquad y(t_0) = y_0 \tag{2a}$$

$$0 = g(y,z), \qquad z(t_0) = z_0 \tag{2b}$$

where (2a) and (2b) represent the differential and algebraic parts with $y : \mathbb{R} \to \mathbb{R}^{n_1}$, $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1}$ and $z : \mathbb{R} \to \mathbb{R}^{n_2}$, $g : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_2}$. Function f allows for additive splitting into non-stiff components f_N and stiff components f_S. We assume consistent initial values, i.e. $g(y_0, z_0) = 0$, and index-1 formulation, i.e. g_z is regular.

For solving (2) we use a generalized ROW-type method as defined in [2]:

$$y_1 = y_0 + \sum_{i=1}^{s} b_i k_i \quad , \quad z_1 = z_0 + \sum_{i=1}^{s} b_i k_i^{alg} \tag{3a}$$

$$\begin{pmatrix} k_i \\ 0 \end{pmatrix} = h \begin{pmatrix} f(v_i, w_i) \\ g(v_i, w_i) \end{pmatrix} + h \begin{bmatrix} A_y & A_z \\ (g_y)_0 & (g_z)_0 \end{bmatrix} \sum_{j=1}^{i} \gamma_{ij} \begin{pmatrix} k_j \\ k_j^{alg} \end{pmatrix} \tag{3b}$$

$$v_i = y_0 + \sum_{j=1}^{i-1} \alpha_{ij} k_j \quad , \quad w_i = z_0 + \sum_{j=1}^{i-1} \alpha_{ij} k_j^{alg}. \tag{3c}$$

Here, y_1 and z_1 are numerical solutions to given differential and algebraic parts after a step with step-size h. Parameters k_i and k_i^{alg} are stage-values. Weights b_i and coefficients α_{ij}, γ_{ij} with $i = 1, \ldots, s$, $j = 1, \ldots, i$ are real numbers with $\alpha_{ii} = 0$ and $\gamma_{ii} = \gamma$. Parameter s denotes the number of internal stages.

Characteristic properties of this generalized method result from the special Jacobian expression given in (3b). For differential parts, it considers arbitrary entries A_y and A_z. For algebraic parts, it considers exact derivatives $(g_y)_0$ and $(g_z)_0$ computed at (y_0, z_0). Hence, choosing A_y and A_z appropriately, different schemes from literature and strategies that reduce effort computing the Jacobian can be realized.

Regarding realization of schemes known from literature, (3) leads to standard ROW methods for DAEs derived in [4] when $A_y = (f_y)_0$ and $A_z = (f_z)_0$. A detailed list of further schemes covered is given in [2]. They are generally realized setting entries of the exact Jacobian's differential parts to zero, i.e. exploiting partial explicit integration by underlying Runge-Kutta methods.

Combined with sparse Jacobian structures, this might allow to save effort already. However, partial explicit integration requires to know about given stiff and non-stiff components. Besides, (3) enables to keep Jacobian entries with respect to differential parts constant for several time-steps, i.e. $A_y = f_y + \mathcal{O}(h)$ and $A_z = f_z + \mathcal{O}(h)$. Without having to know about stiff and non-stiff components, this strategy is assumed to save effort of Jacobian computations as for most DAEs $n_1 \gg n_2$.

Order conditions are derived by an adapted version of Butcher's theory of rooted trees [2]. An appropriate set of coefficients satisfying conditions up to order three is ROS34PRW [3]. For further details we refer to [2].

3 Water Hammer Simulation

In order to analyze performance of generalized ROW-type methods simulating water supply networks we compute water hammer effects within pipe elements. A water hammer is a discontinuous shock wave caused by sudden changes in flow behavior.

For realizing the water hammer test we regard inviscid flow within an uniform horizontal pipe segment that measures $D = 1$ in diameter and $L = 20$ in length. Initial conditions assume $u_0 = 1$ for flow velocity and $p_0 = 1.1E5$ for pressure distribution along the entire pipe segment. Rapid changes in flow behavior to generate the shock wave are given by appropriate boundary conditions at the segment's lower bound (Fig. 1).

We solve the DAE system that results after spatial semi-discretization within time interval $t = [0, 0.01]$. To determine the gain on efficiency that is expected to be given by generalized ROW-type methods with special Jacobian approximations, time-integration is realized via ROS34PRW [3]. ROS34PRW will be applied using two variations: The first corresponds to a standard ROW method for DAEs computing exact Jacobians with every successful time-step (i.e. $A_y = (f_y)_0$ and $A_z = (f_z)_0$). The second represents properties of a generalized ROW-type method using time-lagged Jacobians by updating differential parts just with every tenth successful step (i.e. $A_y = f_y + \mathcal{O}(h)$ and $A_z = f_z + \mathcal{O}(h)$).

The test problem enables to apply sparse Jacobian structures that correspond to a diagonal band-matrix. Providing sparse structures significantly reduces computing times as it decreases the number of function calls. However, given information is not always detailed enough to determine sparse Jacobian patterns. In order to analyze how different sparse structures might affect performance we will regard computations with three different Jacobian patterns: The first is a completely sparse structure for differential and algebraic parts (sparse). The second reduces sparse

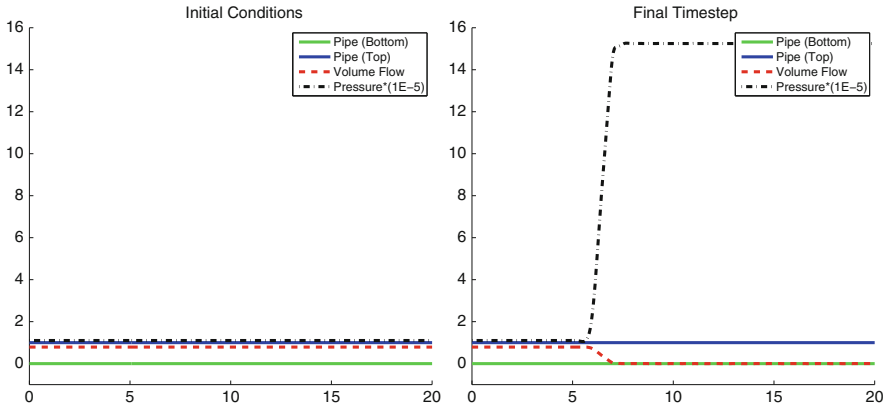


Fig. 1 Initial conditions and final profile simulating the water hammer

structures to algebraic parts, assuming structures of the differential parts to be non-sparse (sparse alg.). The third considers a completely non-sparse structure (full).

In order to show further effects of different Jacobian structures, we will regard the water hammer problem in two variations: The first assumes the segment to be represented by one single pipe. The second considers the segment to consist of two coupled pipes. Coupling conditions are realized by additional algebraic constraints. As a consequence, Jacobian structures will slightly differ for both test cases.

To determine efficiencies for the different integration strategies and test cases, we will regard errors against CPU-times and errors against total number of function evaluations. Errors are evaluated comparing numerical solutions and numerical reference solutions using a weighted quadratic error computation. Numerical solutions are given by different settings for absolute tolerances (*atol*) and relative tolerances (*rtol*) of an integrated step-size control. They range from $tol = 10^{-2}$ to $tol = 10^{-6}$, where $tol = atol = rtol$. Numerical reference solutions were computed by standard MATLAB integrator ode15s using $tol = 10^{-12}$.

3.1 Test Case 1: One Single Pipe Element

We consider the segment to be given by one single pipe element. Choosing 400 cells for spatial discretization, dimensions of the DAE system are characterized by $n_1 = 800$ differential equations and $n_2 = 4$ algebraic constraints.

Results regarding efficiencies of time-integration providing different Jacobian patterns are given in Fig. 2. Comparing performances with respect to CPU-times, applying properties of generalized ROW-type methods by keeping the Jacobian with respect to its differential parts constant for several time-steps proves to be generally

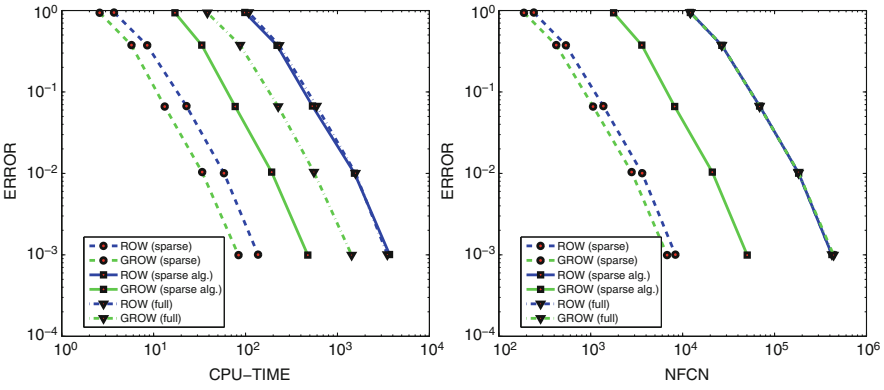


Fig. 2 Results regarding one single pipe element (*Left*: Graphs for ROW (sparse alg.) and ROW (full) coincide; *Right*: Graphs for ROW (sparse alg.), ROW (full) and GROW (full) coincide)

Table 1 Minimum number of function calls required for computing the Jacobian

Approximation		Test case 1			Test case 2		
A_y	A_z	Sparse	Sparse alg.	Full	Sparse	Sparse alg.	Full
$(f_y)_0$	$(f_z)_0$	11	804	804	11	808	808
$f_y + \mathcal{O}(h)$	$f_z + \mathcal{O}(h)$	7	7	804	11	11	808

advantageous. This is especially due to operations that can be saved in function calls required for Jacobian computation when reducing updates to algebraic parts.

Besides, using a time-lagged Jacobian benefits a reduced number of total function calls regarding sparse patterns as the number of calls required to compute a Jacobian is decreased when keeping its differential parts. The resulting number of function calls for a single Jacobian computation is listed in Table 1. It shows the minimum number of calls required, i.e. for the case $A_y = f_y + \mathcal{O}(h)$ and $A_z = f_z + \mathcal{O}(h)$ just the number of calls for updates reduced to algebraic parts is given. In steps with full Jacobian updates it equals the number for the case $A_y = (f_y)_0$ and $A_z = (f_z)_0$.

The gain on efficiency is most relevant where sparse Jacobian structures are provided just with respect to algebraic parts. However, gain proves to be less significant when providing a complete sparse structure. In fact, in this case the Jacobian pattern becomes this sparse that there is just little difference whether Jacobian computations are reduced to given algebraic parts for most time-steps or not.

3.2 Test Case 2: Two Coupled Pipe Elements

The same problem is realized assuming the segment to consist of two coupled pipe elements. Both pipes are discretized by 200 cells each. Consequently, dimensions of the DAE system are characterized by $n_1 = 800$ differential equations and $n_2 = 8$ algebraic constraints. So, while keeping all parameters as previously described, the only renewal in this test is a slightly increased number of algebraic constraints.

Efficiencies for different Jacobian patterns are given in Fig. 3. As in the previous test case, exploiting properties of generalized ROW-type methods by time-lagged Jacobians with respect to differential parts is generally advantageous regarding CPU-times. Again, this is notably due to a decrease of operations required for function calls that compute Jacobians reduced to algebraic constraints.

However, there is a significant difference regarding the total number of function evaluations: Although problem formulation is pretty much the same as in the previous test case, the number of function calls that can be saved is negatively affected. Especially regarding completely sparse Jacobian structures it is not decreased when applying $A_y = f_y + \mathcal{O}(h)$ and $A_z = f_z + \mathcal{O}(h)$. Details are listed in Table 1. This is due to the additional algebraic constraints for realizing the mutual coupling of given pipe elements. Despite the number of additional conditions is small, they

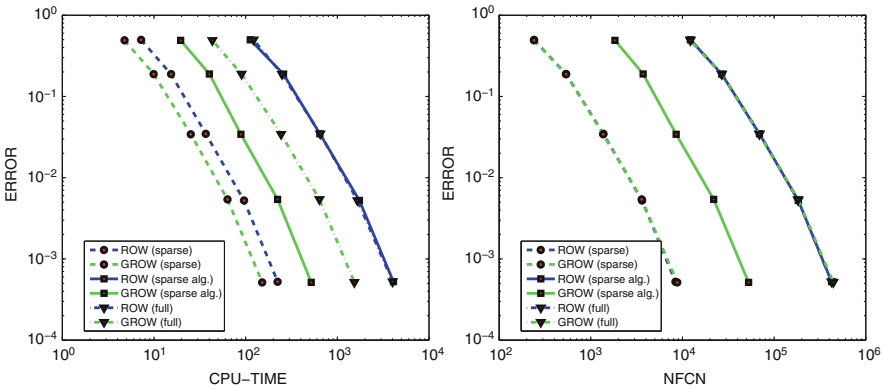


Fig. 3 Results regarding two coupled pipe elements (*Left*: Graphs for ROW (sparse alg.) and ROW (full) coincide; *Right*: Graphs for ROW (sparse) and GROW (sparse) coincide, Graphs for ROW (sparse alg.), ROW (full) and GROW (full) coincide)

significantly influence the Jacobian structure. As a consequence, the number of function calls cannot be further decreased regarding time-lagged Jacobians with respect to differential parts. It is reduced to a minimum already by providing sparse Jacobian structures.

4 Conclusion and Outlook

We applied a generalized ROW-type method to the water hammer test problem that represents characteristics of pressurized flows in networks for water supply. In this context, a time-lagged Jacobian with respect to differential parts of the arising DAE system was considered and compared to a standard ROW method for varying Jacobian structures. Results were advantageous as the time-lagged Jacobians generally reduced computing times. Hence, exploiting properties of the generalized ROW-type scheme is promising to improve computations within the EWave decision support system.

However, the generalized ROW-type method considered is restricted to arbitrary Jacobian approximations with respect to differential parts. Thus, properties of time-lagged Jacobians proved to be dependent on provided Jacobian structures. As a consequence, the number of function calls that can be saved by time-lagged Jacobians might be little, especially when using extended sparse Jacobian patterns. In order to overcome this problem and to exploit potential of time-lagged Jacobians even more, generalized schemes that allow for additional approximations with respect to algebraic parts are required. Realizing these schemes is currently in progress.

Acknowledgements We thank members of the Chair of Applied Mathematics/Numerical Analysis at Bergische Universität Wuppertal for all their fruitful discussions. Also, we gratefully acknowledge that the EWave research project is funded by the German Federal Ministry of Education and Research (BMBF) under grant number 02WER1323E. The corresponding author is indebted to the Hochchschule Bonn-Rhein-Sieg Graduate-Institute for financial support by Ph.D. scholarship.

References

1. Abreu, J., Cabrera, E., Izquierdo, J., García-Serra, J.: Flow modeling in pressurized systems revisited. J. Hydraul. Eng. **125**, 1154–1169 (1999)
2. Jax, T., Steinebach, G.: Generalized ROW-type methods for solving semi-explicit DAEs of index-1. J. Comput. Appl. Math. **316**, 213–228 (2017)
3. Rang, J.: An analysis of the Prothero-Robinson example for constructing new DIRK and ROW methods. J. Comput. Appl. Math. **262**, 105–114 (2014)
4. Roche, M.: Rosenbrock methods for differential algebraic equations. Numer. Math. **52**, 45–63 (1988)
5. 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)

Minisymposium: Spacetime Models of Gravity in Space Geolocation and Acoustics

Jose M. Gambi, Michael M. Tung, Emilio Defez, and Manuel Carretero

Description

The geometrization of gravity has become one of the cornerstones of modern science having an impact on the industrial progress connected to many activities of daily life. In fact, in the last decades substantial research has been invested into post-Newtonian corrections for high-precision space geodesy and navigation [1–3, 12], as well as into the design of analogue models of gravity by making use of advanced optical and acoustic metamaterials (see e.g. [4, 5, 13]). Other present industrial procedures requiring very accurate timing show the need of innovative development of computationally efficient space-time models for use in space. In particular, these models become important in geolocation of passive radiotransmitters in space and to improve active space debris removal [8, 14, 15]. Moreover, acoustic metamaterials—artificially produced materials with exceptional properties not found in nature—provide the engineer with tools to fabricate acoustic devices with highly unusual features. However, the mathematical modelling of acoustic phenomena with curved background spacetimes not only poses challenges in engineering, but may also raise fundamental questions beyond their possible verification in the laboratory environment. This minisymposium was a continuation of the ECMI 2014 Minisymposium "Spacetime Models of Gravity in Space Geolocation and Acoustics", and it was dedicated to the last three topics just mentioned. The session started with a contribution aimed to determine velocities and frequencies of emission of passive radiotransmitters in space by FDOA. The work

J.M. Gambi (✉) • M. Carretero
Gregorio Millán Institute, Univ. Carlos III de Madrid, 28911 Madrid, Spain
e-mail: gambi@math.uc3m.es; manili@math.uc3m.es

M.M. Tung • E. Defez
Instituto de Matemática Multidisciplinar, Univ. Politècnica de València, 46022 Valencia, Spain
e-mail: mtung@mat.upv.es; edefez@mat.upv.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_71

was in line with previous works carried out at the University Carlos III of Madrid [9–11] with the collaboration of the Instituto de Matemática Multidisciplinar of the Universitat Politècnica de València. The speaker and title of the talk were

Javier Clares. Universidad Carlos III de Madrid (Spain). *FDOA determination of velocities and emission frequencies of passive radiotransmitters in space.*

The following two contributions form the last step in a series of recent works [6, 7]. This series aimed to help increase the accuracy of the space based APT systems endowed with very narrow laser beams e.g. for active space debris removal. In particular, the second contribution showed the results of several experiments obtained by numerically solving the equations of the first contribution, and was the result of a cooperation between Gregorio Millan Institute from the University Carlos III of Madrid and the Institute for Analysis and Scientific Computing from Vienna University of Technology. The speaker and title of the first talk were

Jose M. Gambi. Universidad Carlos III de Madrid (Spain). *Non-linear post-Newtonian equations for the motion of designated targets with respect to space based APT laser systems.*

and the speaker and title of the second talk were

Maria L. García del Pino. Universidad Carlos III de Madrid (Spain). *Post-Newtonian corrections to the Newtonian predictions for the motion of designated targets with respect to space based APT laser systems.*

The final contribution was dedicated to the emerging field of transformation acoustic. Here, in particular the focus was on modelling acoustic wave propagation in 2D spaces of constant curvature with the help of a variational principle. The contribution on this topic was led by the Universitat Politècnica de València [16–18], and so far, one previous work includes the cooperation of members of the Institute for Analysis and Scientific Computing from Vienna University of Technology [19]. The speaker and title of the talk were

Michael M. Tung. Universitat Politècnica de València (Spain). *Acoustics in 2D spaces of constant curvature.*

References

1. Ashby, N.: Relativity in the global positioning system. Living Rev. Relativ. **6**(1), 1–42 (2003)
2. Bahder, T.B.: Navigation in curved space-time. Am. J. Phys. **69**, 315–321 (2001)
3. Bahder, T.B.: Clock Synchronization and Navigation in the Vicinity of the Earth. Nova Science Publishers Inc., New York (2009)
4. Chen, H.Y., Chan, C.T.: Acoustic cloaking and transformation acoustics. J. Phys. D **43**, 113001 (2010)
5. Cummer, S.A.: Transformation acoustics. In: Craster, R.V., Guenneau, S. (eds.) Acoustic Metamaterials. Springer Series in Materials Science, vol. 166, pp. 197–218. Springer, Berlin (2013)

6. Gambi, J.M., Garcia del Pino, M.L.: A satellite-to-satellite laser tracking solution within the post-Newtonian model of the earth outer space. In: Progress in Industrial Mathematics at ECMI 2012. Mathematics in Industry, vol. 19, pp. 347–352. Springer, Berlin (2014)
7. Gambi, J.M., Garcia del Pino, M.L.: Post-Newtonian equations of motion for inertial guided space APT systems. WSEAS Trans. Math. **14**, 256–264 (2015)
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. 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. Mathematics in Industry, vol. 17, pp. 489–495. Springer, Berlin (2012)
10. Gambi, J.M., Clares, J., Garcia del Pino, M.L.: FDOA post-Newtonian equations for the location of passive emitters placed in the vicinity of the earth. Aerosp. Sci. Technol. **46**, 137–145 (2015)
11. Gambi, J.M., Rodriguez-Teijeiro, M.C., Garcia del Pino, M.L.: Newtonian and post-Newtonian passive geolocation by TDOA. Aerosp. Sci. Technol. **51**, 18–25 (2016)
12. Moritz, H., Hofmann-Wellenhof, B.: Geometry, Relativity, Geodesy. H. Wichmann Verlag GmbH, Karlsruhe (1993)
13. Norris, A.N.: Acoustic metafluids. J. Acoust. Soc. Am. **125**, 839–849 (2009)
14. Schmitz, M., Fasoulas, S., Utzmann, J.: Performance model for space-based laser debris sweepers. Acta Astronaut. **115**, 376–386 (2015)
15. Shuangyan, S., Xing, J., Hao, C.: Cleaning space debris with a space-based laser system. Chin. J. Aeronaut. **24**(4), 805–811 (2014)
16. Tung, M.M.: A fundamental Lagrangian approach to transformation acoustics and spherical spacetime cloaking. Europhys. Lett. **98**, 34002–34006 (2012)
17. Tung, M.M.: Modelling metamaterial acoustics on spacetime manifolds. In: Jódar, L., Acedo, L., Cortés, J.C., Pedroche, F. (eds.) Modelling for Engineering and Human Behaviour 2012, pp. 235–339. Universidad Politécnica de Valencia (2012)
18. 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. 19, pp. 335–340. Springer, Berlin (2014)
19. Tung, M.M., Weinmüller, E.B.: Gravitational frequency shifts in transformation acoustics. Europhys. Lett. **101**, 54006–54011 (2013)

FDOA Determination of Velocities and Emission Frequencies of Passive Radiotransmitters in Space

Jose M. Gambi, Michael M. Tung, Maria L. García del Pino, and Javier Clares

Abstract Two systems of FDOA equations are introduced to determine in real time the velocities of passive, i.e. non-cooperative, radiotransmitters at the emission instants of the signals, together with the frequencies of emission. The systems correspond to the Newtonian and post-Newtonian frameworks of the Earth surrounding space. The transmitters may be located on the Earth surface or in space. Each system yields accurate unique solutions at the corresponding level of approximation, provided that the locations are determined by appropriated TDOA measurements. The systems are derived by means of Synge's world functions for the vicinity of the Earth, since it allows to clearly identify the post-Newtonian corrections due to the Earth tidal potential and to the gravitational signals' time delays.

1 Introduction

The two systems of Frequency Difference of Arrival (FDOA) equations introduced in this paper are designed for a cluster of satellites to determine in real time the velocities of passive, i.e. non-cooperative, radiotransmitters at the emission instants of the signals, together with the frequencies of emission. The equations correspond to the Newtonian and post-Newtonian approximations to the Schwarzschild field for

J.M. Gambi (✉)
Gregorio Millán Institute, Univ. Carlos III de Madrid, 28911 Madrid, Spain
e-mail: gambi@math.uc3m.es

M.M. Tung
Instituto de Matemática Multidisciplinar, Univ. Politècnica de València, 46022 Valencia, Spain
e-mail: mtung@mat.upv.es

M.L. García del Pino
Department of Mathematics, I.E.S. Alpajes, 28300 Aranjuez, Madrid, Spain
e-mail: lgarciadelpino@educa.madrid.org

J. Clares
Department of Physics, Univ. Carlos III de Madrid, 28911 Madrid, Spain
e-mail: fclares@fis.uc3m.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_72

the Earth surrounding space and are referenced to the Earth Centered Inertial (ECI) reference frame. The radiotransmitters may be placed on the Earth surface or in orbit about the Earth, and the determinations can be carried on while the positions of the radiotransmitters are under determination.

The equations are consistent with other equations in previous works related to Geolocation [3–7]. In fact, they are derived by means of the same tools, which are the Newtonian and post-Newtonian versions of Synge's world function for the Earth surrounding space [9].

Synge's world function is used not only because it has proved very efficient in Geolocation, but also in Navigation [1]. The reason to use it on this occasion is that it allows a clear identification of the post-Newtonian corrections that correspond to the Earth tidal potential and to the time delay of the gravitational signals. In fact, it is so efficient in this and other aspects that it is presently considered a universal tool [2, 8, 10].

As a consequence we have that each system of FDOA equations yields accurate unique solutions at the desired level of approximation, provided that the locations have been determined by means of appropriated Time Difference of Arrival (TDOA) measurements.

To facilitate the reading of the following Sections, the most simple world function, the function for the 3-D Euclidean space, is first introduced in Sect. 2 as prototype of the two functions considered in this work. Then, by means of the first derivatives of these functions, the FDOA systems are introduced in Sect. 3. The suitability of these systems, together with their main future prospects, are finally discussed.

2 Synge's World Functions

To begin with, let us assume that P_1, P_2 are the positions, not necessarily at the same instant, of two objects in orbit about the Earth. Assuming that the 3-D space about the Earth is Euclidean, we have that Synge's world function, $\Omega(P_1, P_2)$, assigns to P_1, P_2 half the square of the length of the straight line segment that joins P_1 and P_2; therefore, if $x^{\alpha_1}, x^{\alpha_2}$ are the ECI coordinates of P_1 and P_2 at t^1 and t^2 respectively,

$$\Omega(P_1, P_2) = \frac{1}{2}\delta_{\alpha\beta}\Delta x^{\alpha}\Delta x^{\beta} = \frac{1}{2}[(\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2], \tag{1}$$

where $\Delta x^{\alpha} = x^{\alpha_2} - x^{\alpha_1}$ ($\alpha = 1, 2, 3$) (Fig. 1, left).

Let us now assume that P_1, P_2 are the corresponding events in the space-time evolution of those two objects about the Earth. Then the space-time ECI coordinates of P_1 and P_2 are $x^{i_1} \equiv (x^{\alpha_1}, t^1)$ and $x^{i_2} \equiv (x^{\alpha_2}, t^2)$ ($i = 1, 2, 3, 4$), so that $x^{4_1} = t^1$ and $x^{4_2} = t^2$. Thus we have, as immediate generalization of (1), that if the 4-D space-time about the Earth is assumed to be flat, the world function for P_1, P_2 is

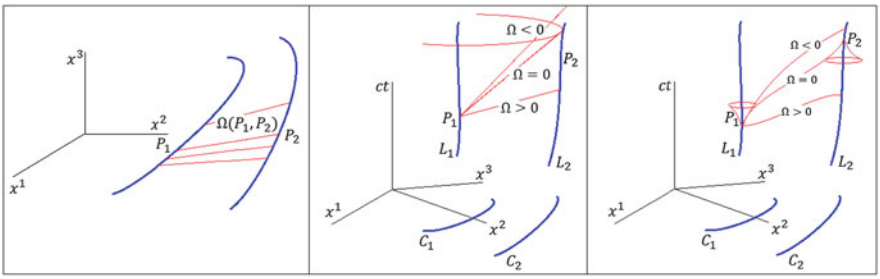


Fig. 1 The world function: *left*, for E_3; *center*, for the 4-D flat space-time; *right*, for general space-times

given by

$$\Omega(P_1, P_2) = \frac{1}{2}\left[\delta_{\alpha\beta}\Delta x^\alpha \Delta x^\beta - (\Delta t)^2\right] = \frac{1}{2}\eta_{ij}\Delta x^i \Delta x^j, \tag{2}$$

where $\eta_{ij} = diag(1, 1, 1, -1)$; $\Delta x^i = x^{i_2} - x^{i_1}$ and $\Delta t = x^{4_2} - x^{4_1} = t^2 - t^1$. (Note that both the speed of light, c, and the universal gravitational constant, G, are made here equal to 1. Hence, in what follows all the magnitudes will be given in seconds, as in [9]). The first world function considered in this work is (2) (Fig. 1, center).

In general, whatever the structure of the space-time about the Earth is considered to be correct (signature $+, +, +, -$), the world function is, analogously to (1)–(2), half the square of the measure of the geodesic that joins any two events in that space-time; therefore, the world function is a line integral, which, in case the space-time is nearly flat, reads

$$\Omega(P_1, P_2) = \frac{1}{2}\eta_{ij}\Delta x^i \Delta x^j + \frac{1}{2}\Delta x^i \Delta x^j \int_C \gamma_{ij} d\omega + \mathcal{O}(\varepsilon^3), \tag{3}$$

since the metric field about the Earth, $g_{ij}(x^k)$, is of the form $g_{ij}(x^k) = \eta_{ij} + \gamma_{ij}(x^k)$, with $\gamma_{ij}(x^k) = \mathcal{O}(\varepsilon^2)$, where $\varepsilon^2 = \mathcal{O}(U) \ll 1$. ($U$ is the Earth gravitational potential, and C, the straight line segment that joins P_1 and P_2 – Fig. 1, right).

If m is the mass of the Earth (measured in seconds), then it is $\gamma_{\alpha\beta} = 2mx^\alpha x^\beta / r^3$ and $\gamma_{44} = 2m/r$, with $r^2 = x^\alpha x^\alpha$ for the post-Newtonian approximation to the Schwarzschild field about the Earth. Hence from (3) we have that the corresponding world function is [5]

$$\Omega(P_1, P_2) = \frac{1}{2}\left[\Delta x^\alpha \Delta x^\alpha - (\Delta t)^2\right]$$

$$+ m|\Delta x^\alpha|\left[\log\frac{r_2 + d_2}{r_1 + d_1} + \frac{d_1}{r_1} - \frac{d_2}{r_2}\right] + \frac{m(\Delta t)^2}{|\Delta x^\alpha|}\log\frac{r_2 + d_2}{r_1 + d_1} + \mathcal{O}(\varepsilon^3), \tag{4}$$

where $\Delta x^\alpha = x^{\alpha_2} - x^{\alpha_1}$, $|\Delta x^\alpha|^2 = \Delta x^\alpha \Delta x^\alpha$ and $\Delta t = t^2 - t^1$; $r_1^2 = x^{\alpha_1} x^{\alpha_1}$, $r_2^2 = x^{\alpha_2} x^{\alpha_2}$, and $(d_1)^2 = r_1^2 - d^2$, $(d_2)^2 = r_2^2 - d^2$, where d is the Euclidean distance from the ECI center to the straight line passing through P_1 and P_2. Thus, the second world function considered in this work is (4).

3 The FDOA Systems

The most interesting properties of $\Omega(P_1, P_2)$ with respect to the frequency shift formulae arise from the computation of its two gradients, which are the partial derivatives with respect to P_1 and P_2. In fact, with the coordinates assigned to P_1 and P_2 previous to (1), we have from (1) that $-\Omega^{\alpha_1}(P_1, P_2) = -\Omega_{\alpha_1}(P_1, P_2) = \partial(\Omega(P_1, P_2))/\partial x^{\alpha_1} = \Delta x^\alpha$, which is the relative position of P_2 with respect to P_1, and analogously, $-\Omega^{\alpha_2}(P_1, P_2) = -\Omega_{\alpha_2}(P_1, P_2) = \partial(\Omega(P_1, P_2))/\partial x^{\alpha_2} = -\Delta x^\alpha$, which is the relative position of P_1 with respect to P_2 (Fig. 2, left).

Analogously, the first derivatives of (2) are

$$\Omega_{\alpha_1} = -\Delta x^\alpha, \quad \Omega_{4_1} = \Delta t, \quad \Omega_{\alpha_2} = \Delta x^\alpha, \quad \Omega_{4_2} = -\Delta t, \tag{5}$$

and for (4), we have

$$\Omega_{\alpha_1} = -\Delta x^\alpha - m\frac{\Delta x^\alpha}{|\Delta x^\delta|}\left[\log\frac{r_2 + d_2}{r_1 + d_1} + \frac{d_1}{r_1} - \frac{d_2}{r_2}\right]$$
$$-m|\Delta x^\alpha|\frac{d_1}{r_1^3}x^{\alpha_1} + m(\Delta t)^2\frac{\Delta x^\alpha}{|\Delta x^\delta|^3}\log\frac{r_2 + d_2}{r_1 + d_1} - \frac{m(\Delta t)^2}{|\Delta x^\delta|}\frac{x^{\alpha_1}}{r_1 d_1} + \mathcal{O}(\varepsilon^3),$$
$$\Omega_{\alpha_2} = \Delta x^\alpha + m\frac{\Delta x^\alpha}{|\Delta x^\delta|}\left[\log\frac{r_2 + d_2}{r_1 + d_1} + \frac{d_1}{r_1} - \frac{d_2}{r_2}\right]$$
$$+m|\Delta x^\alpha|\frac{d_2}{r_2^3}x^{\alpha_2} - m(\Delta t)^2\frac{\Delta x^\alpha}{|\Delta x^\delta|^3}\log\frac{r_2 + d_2}{r_1 + d_1} + \frac{m(\Delta t)^2}{|\Delta x^\delta|}\frac{x^{\alpha_2}}{r_2 d_2} + \mathcal{O}(\varepsilon^3),$$
$$\Omega_{4_1} = -\Omega_{4_2} = \Delta t - 2m\frac{\Delta t}{|\Delta x^\delta|}\log\frac{r_2 + d_2}{r_1 + d_1} + \mathcal{O}(\varepsilon^3), \tag{6}$$

where $|\Delta x^\delta|^2 = \Delta x^\delta \Delta x^\delta$.

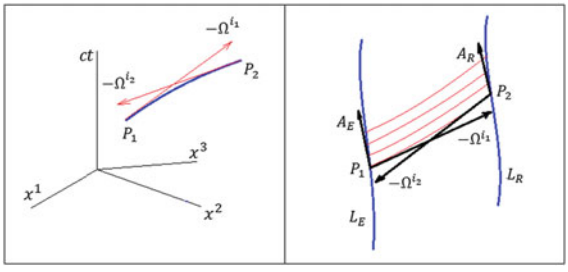


Fig. 2 *Left*: general first derivatives of $\Omega(P_1, P_2)$ (in Euclidean space, the arrow tips coincide with P_1, P_2). *Right*: derivatives of $\Omega(P_1, P_2)$ for radio-signals ($\Omega(P_1, P_2) = 0$)

Let us assume now that P_1 and P_2 are the emission and reception events of a radio-signal, say E and R, respectively. In that case, the pseudo-range from E to R and the travel time of the signal are drawn from the fundamental equation $\Omega(P_1, P_2) = 0$ [3]. Further, if the emission and reception frequencies are denoted by f_E and f_R, the frequency shift formula for (2) and (4) becomes [9]:

$$\frac{f_E - f_R}{f_E} = \frac{\Omega_{i_E} A_E^i + \Omega_{i_R} A_R^i}{\Omega_{j_E} A_E^j}, \tag{7}$$

where A_E^i, A_R^i are the unit tangent vectors of the space-time evolutions L_E, L_R of the emitter and the receiver at E and R respectively, that is to say, the 4-velocities of the emitter and the receiver, respectively (to be computed at the respective order of approximation) (Fig. 2, right).

Since, according to the coordinates assigned to P_1 and P_2, L_E and L_R are given by $x^{i_1}(s_E)$ and $x^{i_2}(s_R)$, where s_E and s_R are the times recorded by the emitter and the receiver respectively, we have within the classical framework:

$$A_E^\alpha = v_E^\alpha, \quad A_E^4 = (ds_E/dt)^{-1} = 1, \quad A_R^\alpha = v_R^\alpha, \quad A_R^4 = (ds_R/dt)^{-1} = 1, \tag{8}$$

where v_E^α, v_R^α are the velocities of E and R at $x^{4_1}(s_E) \equiv t_E$ and $x^{4_2}(s_R) \equiv t_R$, respectively. And for the post-Newtonian framework:

$$A_E^\alpha = v_E^\alpha + \mathscr{O}(\varepsilon^3), \quad A_E^4 = \left(\frac{ds_E}{dt}\right)^{-1} = 1 + \frac{m}{r_E} + \frac{1}{2}(v_E)^2 + \mathscr{O}(\varepsilon^3),$$

$$A_R^\alpha = v_R^\alpha + \mathscr{O}(\varepsilon^3), \quad A_R^4 = \left(\frac{ds_R}{dt}\right)^{-1} = 1 + \frac{m}{r_R} + \frac{1}{2}(v_R)^2 + \mathscr{O}(\varepsilon^3), \tag{9}$$

where $(v_E)^2 = v_E^\alpha v_E^\alpha$ and $(v_R)^2 = v_R^\alpha v_R^\alpha$. Hence, while by substituting from (5), (8) into (7) we have the familiar formula

$$f_R = f_E \Big\{ 1 - \frac{\Delta x^\alpha}{|\Delta x^\delta|}(v_R^\alpha - v_E^\alpha) \Big\}, \tag{10}$$

from (6) and (9) we have instead [5]

$$f_R = f_E \Big\{ 1 - \frac{\Delta x^\alpha}{|\Delta x^\delta|}(v_R^\alpha - v_E^\alpha)$$

$$+ \frac{m}{r_R} - \frac{m}{r_E} + \frac{1}{2}\big[(v_R)^2 - (v_E)^2\big] - \frac{\Delta x^\alpha}{|\Delta x^\delta|}(v_R^\alpha - v_E^\alpha)\frac{\Delta x^\beta}{|\Delta x^\delta|} v_E^\beta \Big\} + \mathscr{O}(\varepsilon^3). \tag{11}$$

Let us finally assume that S_1 and S_2 are two satellites involved in the geolocation of a radiotransmitter. If f_{S_1}, f_{S_2} are the reception frequencies of the signal at S_1, S_2,

we have from (10) and (11), respectively,

$$F_{12} = f_{S_1} - f_{S_2} = f_E\Big\{u^\alpha_{(2)}(v^\alpha_{S_2} - v^\alpha_E) - u^\alpha_{(1)}(v^\alpha_{S_1} - v^\alpha_E)\Big\}, \tag{12}$$

$$\begin{aligned} F_{12} = f_{S_1} - f_{S_2} = f_E\Big\{&\big[u^\alpha_{(2)}(v^\alpha_{S_2} - v^\alpha_E)\big]\big[1 + u^\alpha_{(2)}v^\alpha_E\big] - \big[u^\alpha_{(1)}(v^\alpha_{S_1} - v^\alpha_E)\big]\big[1 + u^\alpha_{(1)}v^\alpha_E\big] \\ &+ \frac{1}{2}\big[(v_{S_1})^2 - (v_{S_2})^2\big] + \Big[\frac{m}{r_{S_1}} - \frac{m}{r_{S_2}}\Big]\Big\} + \mathscr{O}(\varepsilon^3), \end{aligned} \tag{13}$$

where $u^\alpha_{(i)} = (x^\alpha_{S_i} - x^\alpha_E)/|x^\delta_{S_i} - x^\delta_E|$, so that $u^\alpha_{(i)}$ are data to be provided by TDOA measurements $(i = 1, 2)$ [6].

Thus, we can conclude from (12) that any FDOA system involving five satellites, say $S_1, \ldots, S_5$, is within the Classical framework of the form $A\ V = W$, such as

$$\begin{pmatrix} u^1_{(1)} - u^1_{(2)}, \ldots, u^3_{(1)} - u^3_{(2)}, -F_{12} \\ u^1_{(1)} - u^1_{(3)}, \ldots, u^3_{(1)} - u^3_{(3)}, -F_{13} \\ u^1_{(4)} - u^1_{(2)}, \ldots, u^3_{(4)} - u^3_{(2)}, -F_{42} \\ u^1_{(2)} - u^1_{(5)}, \ldots, u^3_{(2)} - u^3_{(5)}, -F_{25} \end{pmatrix} \begin{pmatrix} v^1_E \\ v^2_E \\ v^3_E \\ f_E^{-1} \end{pmatrix} = \begin{pmatrix} u^\alpha_{(1)}v^\alpha_{S_1} - u^\alpha_{(2)}v^\alpha_{S_2} \\ u^\alpha_{(1)}v^\alpha_{S_1} - u^\alpha_{(3)}v^\alpha_{S_3} \\ u^\alpha_{(4)}v^\alpha_{S_4} - u^\alpha_{(2)}v^\alpha_{S_2} \\ u^\alpha_{(2)}v^\alpha_{S_2} - u^\alpha_{(5)}v^\alpha_{S_5} \end{pmatrix} \tag{14}$$

Analogously, we have from (13) that all the FDOA systems within the Post-Newtonian framework are of the form $\overline{A}\ \overline{V} = \overline{W}$, such as, to be compared with (14),

$$\overline{A} = \begin{pmatrix} u^1_{(1)}(1 - C_1) - u^1_{(2)}(1 - C_2), \ldots, u^3_{(1)}(1 - C_1) - u^3_{(2)}(1 - C_2), -F_{12} \\ u^1_{(1)}(1 - C_1) - u^1_{(3)}(1 - C_3), \ldots, u^3_{(1)}(1 - C_1) - u^3_{(3)}(1 - C_3), -F_{13} \\ u^1_{(4)}(1 - C_4) - u^1_{(2)}(1 - C_2), \ldots, u^3_{(4)}(1 - C_4) - u^3_{(2)}(1 - C_2), -F_{42} \\ u^1_{(2)}(1 - C_2) - u^1_{(5)}(1 - C_5), \ldots, u^3_{(2)}(1 - C_2) - u^3_{(5)}(1 - C_5), -F_{25} \end{pmatrix}$$

$$\overline{V} = \begin{pmatrix} \overline{v}^1_E \\ \overline{v}^2_E \\ \overline{v}^3_E \\ (\overline{f}_E)^{-1} \end{pmatrix} \qquad \overline{W} = \begin{pmatrix} C_1 - C_2 - \frac{1}{2}H_{12} - M_{12} - (u^\alpha_{(1)}v^\alpha_E)^2 + (u^\alpha_{(2)}v^\alpha_E)^2 \\ C_1 - C_3 - \frac{1}{2}H_{13} - M_{13} - (u^\alpha_{(1)}v^\alpha_E)^2 + (u^\alpha_{(3)}v^\alpha_E)^2 \\ C_4 - C_2 - \frac{1}{2}H_{42} - M_{42} - (u^\alpha_{(4)}v^\alpha_E)^2 + (u^\alpha_{(2)}v^\alpha_E)^2 \\ C_2 - C_5 - \frac{1}{2}H_{25} - M_{25} - (u^\alpha_{(2)}v^\alpha_E)^2 + (u^\alpha_{(5)}v^\alpha_E)^2 \end{pmatrix} \tag{15}$$

where $C_i = u^\gamma_{(i)}v^\gamma_{S_i}$, $H_{ij} = \big[(v_{S_i})^2 - (v_{S_j})^2\big]$, $M_{ij} = \big[m/r_{S_i} - m/r_{S_j}\big]$ $(i, j = 1, \ldots, 5)$.

Let us note with respect to these systems that a way to guarantee unique solutions is to use the same satellites used for the TDOA measurements. In addition, since $u^{\alpha}_{(i)}$ are to be provided by these satellites, the systems may be fed in real time. Finally, since the magnitude of the corrections provided by (13) are significant [5], it is expected that for objects orbiting the Earth, the magnitude of the corrections provided by the systems of type (15) will be also significant.

References

1. Bahder, T.B.: Navigation in curved space-time. Am. J. Phys. **69**, 315–321 (2001)
2. Clarke, C.J.S.: Synge world function. In: Encyclopedia of Mathematics. Springer, Berlin (2011). http://www.encyclopediaofmath.org/index.php?title=Synge_world_function&oldid=18235
3. Gambi, J.M., Rodriguez-Teijeiro, M.C., García 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)
4. Gambi, J.M., Rodriguez-Teijeiro, M.C., García del Pino, M.L.: The post-Newtonian geolocation problem by TDOA. In: M. Günter, A. Bartel, M. Brunk, S. Schöps, M. Striebel (eds.) Progress in Industrial Mathematics at ECMI 2010. Mathematics in Industry, vol. 17, pp. 489–495. Springer, Berlin (2012)
5. Gambi, J.M., Clares, J., García del Pino, M.L.: FDOA post-Newtonian equations for the location of passive emitters placed in the vicinity of the earth. Aerosp. Sci. Technol. **46**, 137–145 (2015)
6. Gambi, J.M., Rodriguez-Teijeiro, M.C., García del Pino, M.L.: Newtonian and post-Newtonian passive geolocation by TDOA. Aerosp. Sci. Technol. **51**, 18–25 (2016)
7. Gambi, J.M., Clares, J., Rodriguez-Teijeiro, M.C.: Post-Newtonian geolocation of passive radio transmitters by TDOA and FDOA. In: G.R. Russo, V. Capasso, G. Nicosia, V. Romano (eds.) Progress in Industrial Mathematics at ECMI 2014. Mathematics in Industry, vol. 21. Springer, Berlin (2016)
8. Hazewinkel, M.: Synge world function. In: Hazewinkel, M. (ed.) Encyclopaedia of Mathematics: Supplement, vol. 1, pp. 464–465. Kluwer Academic Pub., Dordrecht (2012)
9. Synge, J.L.: Relativity: The General Theory, chap. 2. North-Holland, Amsterdam (1960)
10. Teyssandier, P., Le Poncin-Lafitte, C., Linet, B.: A universal tool for determining the time delay and the frequency shift of light: Synge's world function. In: H. Dittus, C. Lämmerzahl, S.G. Turyshev (eds.) Lasers, Clocks and Drag-Free Control, pp. 153–180. Springer, Berlin (2008)

Non-linear Post-Newtonian Equations for the Motion of Designated Targets with Respect to Space Based APT Laser Systems

Jose M. Gambi, Maria L. García del Pino, and Maria C. Rodríguez-Teijeiro

Abstract The equations introduced in this paper are aimed to gain accuracy in the determination of the motion of middle size space objects with respect to space based APT laser systems; therefore, they can be used for these systems to engage cataloged space debris objects whose size ranges between 1 and 10 cm. The equations are derived under the assumption that the framework of the Earth surrounding space is post-Newtonian and, unlike the standard p-N equations, they are valid for distant targets. Further, their time validity is also substantially larger than that of the standard equations. The reason is that they include non-linear terms that model the Earth tidal potential along the lines joining the systems and the targets. The equations are derived in local Cartesian orbital coordinates; therefore, they are primarily adapted for use with inertial-guided systems.

1 Introduction

The increasing presence of middle size debris objects, above all in low orbits about the Earth, represents a serious threat of collision with satellites on active duty. This scenario set us before the aim of ablating these objects by means of space-based Acquisition, Pointing and Tracking (APT) systems endowed with very narrow laser beams. This is a challenge, since, according to the catalogue of the US Space Surveillance Network (SSN), the average size of these objects ranges between 1 and 10 cm [9]. Hence the main issue previous to the implementation of the pointing

J.M. Gambi (✉)
Gregorio Millán Institute, Univ. Carlos III de Madrid, 28911 Madrid, Spain
e-mail: gambi@math.uc3m.es

M.L. García del Pino
Department of Mathematics, I.E.S. Alpajes, 28300 Aranjuez, Madrid, Spain
e-mail: lgarciadelpino@educa.madrid.org

M.C. Rodríguez-Teijeiro
UNED-Madrid, Madrid, Spain
e-mail: mdcrodriguez@madrid.uned.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_73

directions is to accurately predict the motion of these targets with respect to the systems.

The equations of motion introduced in this paper are aimed to increase this accuracy. They are derived under the assumption that the structure of the Earth surrounding space is post-Newtonian, keeping thus consistency with previous works in Space Geodesy, Positioning, Navigation and Geolocation (see e.g. [1, 2, 5–8]). The equations are derived from Synge's local metric in Cartesian coordinates [11], and are non-linear. They are, in fact, the last step ahead up to the date of a sequence of equations that generalize the familiar equations of the geodesic deviation for the weak approximation to the exterior Schwarzschild field [10]. Thus, unlike these equations, whose coefficients are the values of the tidal potential at the positions of the systems, the coefficients of the equations introduced in this paper are integrals of this potential, and of its spatial derivatives, taken along the straight line segments that, while orbiting, join systems and targets. As a consequence, their time validity is larger than that of the standard p-N equations and are not only valid for targets close to the systems, but also for distant targets. The equations are formulated in local Cartesian coordinates; therefore, they are primarily adapted for use with inertial-guided APT systems.

The structure of the model considered and the approximations made to derive the equations are shown in Sect. 2; the equations are introduced in Sect. 3, and finally, an analysis on the suitability and potential advantages of these equations is proposed at the end of this Sect. 3.

2 Model Structure and Approximations

We assume that the structure of the space-time about the Earth is that of the post-Newtonian weak approximation to the exterior Schwarzschild field. In terms of a dimensionless parameter ε satisfying $\varepsilon = \mathcal{O}(m/r) = \mathcal{O}(v^2/c^2)$ this assumption means that the geometry of the space-time about the Earth in Earth Centered Inertial (ECI) coordinates $x^i(x^\alpha, t)$ is given by

$$g_{\alpha\beta} = \delta_{\alpha\beta} + \frac{2mx^\alpha x^\beta}{r^3} + \mathcal{O}(\varepsilon^2), \quad g_{\alpha 4} = \mathcal{O}(\varepsilon^3), \quad g_{44} = -1 + \frac{2m}{r} + \mathcal{O}(\varepsilon^2), \qquad (1)$$

where m is the mass of the Earth, v is the characteristic speed of the objects of interest in orbit about the Earth, and $r^2 = x^\alpha x^\alpha$.[1] (Latin indices range from 1 to 4, and Greek, from 1 to 3.)

[1] We make the speed of light, c, and the universal gravitational constant, G, equal to 1, so that all magnitudes with dimension are given in seconds. Thus, m, the mass of the Earth, which is 1.479×10^{-11} approx., and the semi-major axes of the debris objects e.g. in LEO orbits, which range between 2.258×10^{-2} and 2.792×10^{-2} approx., give that ε is of the order of 10^{-9}.

Let us denote the APT system by S, and the target by D. Taking into account that during the tracking action of S the relative speed of D with respect to S is small as compared to v^μ (the ECI velocity of S), we are entitled to assume without loosing consistency with the works cited above that the relationship between the proper times of S and D, s and s' respectively, is given by $ds'/ds = 1$ approx. Further, if the space-time trajectory of S, say L, is given by $x^i(s) \equiv (x^\alpha(s), t(s))$, we have that the unit tangent vector to L, $\lambda^i_{(4)}(s)$, is given by

$$\lambda^\mu_{(4)}(s) = v^\mu, \quad \lambda^4_{(4)}(s) = 1 + \frac{m}{r} + \frac{1}{2}(v)^2, \tag{2}$$

where $(v)^2 = v^\mu v^\mu$. Now we have, as usual, that the conditions of orthogonality with respect to the metric under consideration—(1) in this case—give the components of the canonical inertial-guided system, $\lambda^i_{(\alpha)}(s)$, co-moving with S. Thus, in our case

$$\lambda^\mu_{(\alpha)}(s) = \delta^\mu_\alpha, \quad \lambda^4_{(\alpha)}(s) = v^\alpha, \tag{3}$$

so that, according to (2), $\lambda^4_{(4)}(= dt/ds)$ gives the post-Newtonian relationship between the ECI time coordinate and the proper time of S used in Navigation by GPS, and according to (3), the local space coordinates of D with respect to S are Cartesian coordinates of D with respect to S. Let us denote these coordinates by $X^{(\alpha)}$ (the fourth coordinate of D with respect to S, $X^{(4)}$, is obviously s).

Since the classical post-Newtonian space-time as seen by S is given by [10]

$$\begin{aligned} g_{(\alpha\beta)} &= \delta_{\alpha\beta} - \frac{1}{3} R_{(\alpha\mu\beta\nu)}|_S X^{(\mu)} X^{(\nu)} + \mathcal{O}(\varepsilon^2), \\ g_{(\alpha 4)} &= \mathcal{O}(\varepsilon^3), \\ g_{(44)} &= -1 - R_{(4\mu 4\nu)}|_S X^{(\mu)} X^{(\nu)} + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{4}$$

where $R_{(ijkl)}|_S$ is the Riemann tensor evaluated at S with the metric (1), the equations for the relative motion derived from this metric are appropriated only for objects D orbiting close to S. In fact, according to the geodesic principle, these equations are

$$\frac{d^2 X_{(\alpha)}}{ds^2} = -R_{(\alpha 4\beta 4)}|_S X^{(\beta)}, \tag{5}$$

so that $R_{(\alpha 4\beta 4)}$ is evaluated at S. Therefore, to get equations that can also be used when these same objects, or others, are orbiting far from S, it is necessary to substitute the metric in (4) by some other suitable for wide spaces about S, and that reduces to (4) close to S.

This need is fulfilled by Synge's metric in a nice way, since this metric accounts for the value of $R_{(ijkl)}$ all along the straight line segments that, while S and D are orbiting the Earth, join S and D. Now, since these values reduces in the most simple

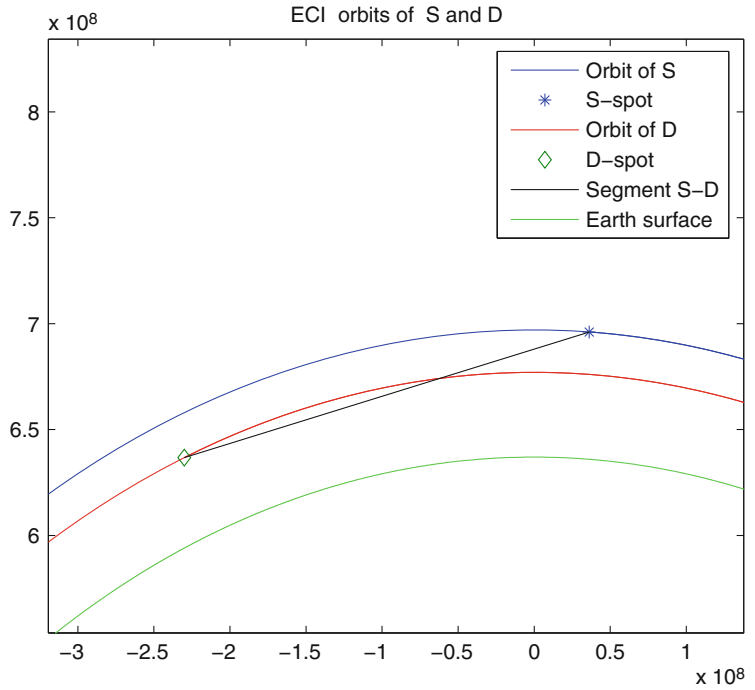


Fig. 1 This figure shows an enlarged snapshot of a simulation that involves a debris object in circular orbit, 400 km above the Earth surface on the equatorial plane, and an APT system 200 km above the object at perigee. The figure shows the segment S-D considered in (7)

way to the value of $R_{(ijkl)}$ at S when the objects D are very close to S, we have here the reason why this metric results to be the best option (Fig. 1).

If x^{α} and $x^{\alpha'}$ are the space ECI coordinates of S and D at s respectively, and $(X^{(\alpha)}, X^{(4)})$ are, as before, the coordinates of D with respect to S and $\lambda^{\mu}_{(\alpha)}$ at s, Synge's metric at D reads [11]

$$\begin{aligned} \bar{g}_{(\alpha\beta)} &= \delta_{\alpha\beta} + 2\bar{h}_{(\alpha\beta)} + \mathscr{O}(\varepsilon^2), \\ \bar{g}_{(\alpha 4)} &= \mathscr{O}(\varepsilon^3), \\ \bar{g}_{(44)} &= -1 + 2\bar{h}_{(44)} + \mathscr{O}(\varepsilon^2), \end{aligned} \tag{6}$$

where

$$\bar{h}_{(\alpha\beta)} = \frac{3}{2} X^{(\mu)} X^{(\nu)} \int_0^1 (1-u) u S_{(\alpha\beta\mu\nu)} du,$$

$$\bar{h}_{(44)} = \frac{3}{2} X^{(\mu)} X^{(\nu)} \int_0^1 (1-u) S_{(44\mu\nu)} du, \tag{7}$$

the integrals being taken, as mentioned above, along the segments $\bar{x}^\alpha(u) = (1-u)x^\alpha + ux^{\alpha'}$ $(0 \le u \le 1)$. We note that $S_{(abcd)}$ in (7) is the symmetrized Riemann tensor at $\bar{x}^\alpha(u)$ evaluated with the metric (1). Therefore, the value of $S_{(abcd)}$ to be introduced in (7) is

$$S_{(abcd)}(\bar{x}^\mu(u)) = -\frac{1}{3}\big(R_{(acbd)} + R_{(adbc)}\big)\big(\bar{x}^\mu(u)\big), \tag{8}$$

where $R_{(acbd)}$ is the Riemann tensor at $\bar{x}^\alpha(u)$ evaluated with (1). In particular, the components of $R_{(acbd)}$ at $\bar{x}^\alpha(u)$ needed to derive the equations of motion are

$$R_{(\alpha 4\gamma 4)}(\bar{x}^\mu(u)) = -m\left(\frac{3\bar{x}^\alpha(u)\bar{x}^\gamma(u)}{\bar{r}(u)^5} - \frac{\delta_{\alpha\gamma}}{\bar{r}(u)^3}\right), \tag{9}$$

where $\bar{r}(u)^2 = \bar{x}^\mu(u)\bar{x}^\mu(u)$ (this will be shown in Sect. 3). Finally, we note that $R_{(\alpha 4\gamma 4)}$ are $\mathcal{O}(\varepsilon)$.

3 The Equations of Motion

According to Synge, we could undoubtedly worry out the equations of a geodesic from (6). However, this would involve differentiating (6). For this reason, we follow Synge's procedure to model the motion of D with respect to S [11].

Let L be, as above, the space-time evolution of S, and L', the evolution of D. L and L' are geodesics. Let us consider for each event P' of L' the event P of L which is simultaneous to P' as measured by the clock carried on by S. Let us also assume that the coordinates of D with respect to S at s are $X^{(i)} \equiv (X^{(\alpha)}, X^{(4)})$, so that $X^{(4)} = s$ (as in Sect. 2). Then we have that the coordinates in (6) of P are $(0, s)$. Let us finally assume, according to Sect. 2, that the ECI coordinates of P and P' are $x^i \equiv (x^\alpha, x^4)$ and $x^{i'} \equiv (x^{\alpha'}, x^{4'})$, so that x^4 and $x^{4'}$ are t and t' respectively.

Since L' is determined by a point on it and a direction, the plan to find the equations of L' is to write $X^{(\alpha)}(s)$ in terms of the length of the straight-line segment $\Gamma(\bar{x}^i(u))$ that joins P and P'. Now, as P and P' run along L and L', this length is function of x^i and $x^{i'}$, and so the goal is to obtain the second derivative of $X^{(\alpha)}$ with respect to s in terms of the derivatives of such function. Thus, with $\Omega(x^i, x^{i'})$ defined by

$$\Omega(x^i, x^{i'}) = \frac{1}{2}\int_P^{P'} g_{ij}\frac{d\bar{x}^i}{du}\frac{d\bar{x}^j}{du}du, \tag{10}$$

which is half the square of the length of Γ, we have

$$X^{(\alpha)} = X_{(\alpha)} = -\Omega_i \lambda^i_{(\alpha)}, \tag{11}$$

with

$$\Omega_i \lambda^i_{(4)} = 0, \tag{12}$$

where g_{ij} are as in (1); $\lambda^i_{(\alpha)}$ are as in (3), and $\Omega_i = \partial\Omega(x^i, x^{i'})/\partial x^i$. (Note that (11) holds because Ω_i is the gradient of Ω at P. Note also that (12) is the condition for P and P' to be simultaneous for S.)

The condition for L to be geodesic is $d\lambda^i_{(\alpha)}/ds = 0$. Then, differentiation of (11) gives, by this condition and by (12),

$$\frac{dX_{(\alpha)}}{ds} = -\Omega_{ij}\lambda^i_{(\alpha)}\lambda^j_{(4)} - \Omega_{ij'}\lambda^i_{(\alpha)}\lambda^{j'}_{(4)}, \tag{13}$$

where $\Omega_{ij} = \partial^2\Omega/\partial x^j\partial x^i$; $\Omega_{ij'} = \partial^2\Omega/\partial x^{j'}\partial x^i$, and $\lambda^{j'}_{(4)} = dx^{i'}/ds$, the unit tangent vector to L'. Therefore, the second differentiation gives

$$\begin{aligned}\frac{d^2X_{(\alpha)}}{ds^2} &= -\Omega_{ijk}\lambda^i_{(\alpha)}\lambda^j_{(4)}\lambda^k_{(4)} - \Omega_{ijk'}\lambda^i_{(\alpha)}\lambda^j_{(4)}\lambda^{k'}_{(4)} \\ &\quad - \Omega_{ij'k}\lambda^i_{(\alpha)}\lambda^{j'}_{(4)}\lambda^k_{(4)} - \Omega_{ij'k'}\lambda^i_{(\alpha)}\lambda^{j'}_{(4)}\lambda^{k'}_{(4)},\end{aligned} \tag{14}$$

where $\Omega_{ijk} = \partial^3\Omega/\partial x^k\partial x^j\partial x^i$; $\Omega_{ijk'} = \partial^3\Omega/\partial x^{k'}\partial x^j\partial x^i$; $\Omega_{ij'k} = \partial^3\Omega/\partial x^k\partial x^{j'}\partial x^i$, and $\Omega_{ij'k'} = \partial^3\Omega/\partial x^{k'}\partial x^{j'}\partial x^i$.

Now, by (2), $\lambda^\mu_{(4)} = v^\mu$ and $\lambda^4_{(4)} = 1 + \mathcal{O}(\varepsilon)$, and analogously, $\lambda^{\mu'}_{(4)} = v^{\mu'}$ and $\lambda^{4'}_{(4)} = 1 + \mathcal{O}(\varepsilon)$, where $v^{\mu'}$ is the ECI velocity of D at s; hence (14) becomes

$$\frac{d^2X_{(\alpha)}}{ds^2} = -\Omega_{(\alpha 44)} - \Omega_{(\alpha 44')} - \Omega_{(\alpha 4'4)} - \Omega_{(\alpha 4'4')}. \tag{15}$$

The final step is to compute the third derivatives of Ω for them to be introduced in (15). A straightforward calculation from (10) gives

$$\Omega_{(\alpha 44)} = -\Omega_{(\alpha 44')} = -\Omega_{(\alpha 4'4)} = -X^{(\gamma)}\int_0^1 (1-u)^2 R_{(\alpha\gamma 4)}du,$$

$$\Omega_{(\alpha 4'4')} = 2X^{(\gamma)}\int_0^1 u^2 R_{(\alpha\gamma 4)}du - X^{(\mu)}X^{(\nu)}\int_0^1 (1-u)u^2\frac{\partial R_{(\mu 4\nu 4)}}{\partial x^\alpha}du, \tag{16}$$

and so we have that the equations are

$$\frac{d^2X_{(\alpha)}}{ds^2} = -X^{(\gamma)}\int_0^1 (1-2u+3u^2)R_{(\alpha\gamma 4)}du + X^{(\mu)}X^{(\nu)}\int_0^1 (1-u)u^2\frac{\partial R_{(\mu 4\nu 4)}}{\partial x^\alpha}du, \tag{17}$$

or, more neatly, thus showing the linear and non-linear terms,

$$\begin{aligned}\frac{d^2X_{(1)}}{ds^2} &= A_{11}X^{(1)} + A_{12}X^{(2)} + A_{13}X^{(3)} \\ &+ \bar{A}_{11}(X^{(1)})^2 + \bar{A}_{12}X^{(1)}X^{(2)} + \bar{A}_{13}X^{(1)}X^{(3)} \\ &+ \bar{A}_{21}X^{(2)}X^{(1)} + \bar{A}_{22}(X^{(2)})^2 + \bar{A}_{23}X^{(2)}X^{(3)} \\ &+ \bar{A}_{31}X^{(3)}X^{(1)} + \bar{A}_{32}X^{(3)}X^{(2)} + \bar{A}_{33}(X^{(3)})^2,\end{aligned}$$

$$\begin{aligned}\frac{d^2X_{(2)}}{ds^2} &= A_{21}X^{(1)} + A_{22}X^{(2)} + A_{23}X^{(3)} \\ &+ \bar{B}_{11}(X^{(1)})^2 + \bar{B}_{12}X^{(1)}X^{(2)} + \bar{B}_{13}X^{(1)}X^{(3)} \\ &+ \bar{B}_{21}X^{(2)}X^{(1)} + \bar{B}_{22}(X^{(2)})^2 + \bar{B}_{23}X^{(2)}X^{(3)} \\ &+ \bar{B}_{31}X^{(3)}X^{(1)} + \bar{B}_{32}X^{(3)}X^{(2)} + \bar{B}_{33}(X^{(3)})^2,\end{aligned}$$

$$\begin{aligned}\frac{d^2X_{(3)}}{ds^2} &= A_{31}X^{(1)} + A_{32}X^{(2)} + A_{33}X^{(3)} \\ &+ \bar{C}_{11}(X^{(1)})^2 + \bar{C}_{12}X^{(1)}X^{(2)} + \bar{C}_{13}X^{(1)}X^{(3)} \\ &+ \bar{C}_{21}X^{(2)}X^{(1)} + \bar{C}_{22}(X^{(2)})^2 + \bar{C}_{23}X^{(2)}X^{(3)} \\ &+ \bar{C}_{31}X^{(3)}X^{(1)} + \bar{C}_{32}X^{(3)}X^{(2)} + \bar{C}_{33}(X^{(3)})^2,\end{aligned} \tag{18}$$

where

$$\begin{aligned}A_{\alpha\gamma} &= -\int_0^1 (1 - 2u + 3u^2)R_{(\alpha 4\gamma 4)}du, \\ \bar{A}_{\mu\nu} &= \int_0^1 (1-u)u^2\frac{\partial R_{(\mu 4\nu 4)}}{\partial x^1}du, \\ \bar{B}_{\mu\nu} &= \int_0^1 (1-u)u^2\frac{\partial R_{(\mu 4\nu 4)}}{\partial x^2}du, \\ \bar{C}_{\mu\nu} &= \int_0^1 (1-u)u^2\frac{\partial R_{(\mu 4\nu 4)}}{\partial x^3}du,\end{aligned} \tag{19}$$

with $R_{(\mu 4\nu 4)}$ computed from (9).

To analyze the suitability of these equations so as, in particular, the role played by the non-linear terms, there are at disposal three additional systems of differential equations for the relative motions: the known system made from the difference of ECI Newtonian equations (see e.g. [3]); the analogous from the ECI p-N equations

(see e.g. [10]), and the linear equations contained in (17) (see [4]). An analysis involving LEO, MEO and HEO debris objects is made by means of the numerical experiments shown in the ECMI 2016 contribution *Post-Newtonian Corrections to the Newtonian Predictions for the Motion of Designated Targets with respect to Space Based APT Laser Systems*.

References

1. Ashby, N.: Relativity in the global positioning system. Living Rev. Relativ. **6**, 1–42 (2003)
2. Bahder, T.B.: Clock Synchronization and Navigation in the Vicinity of the Earth. Nova Science, New York (2009)
3. Curtis, H.D.: Orbital Mechanics for Engineering Students, 3rd edn., pp. 367–396. Elsevier, Oxford (2014)
4. Gambi, J.M., García del Pino, M.L.: Post-Newtonian Equations of Motion for Inertial Guided Space APT Systems. WSEAS Trans. Math. **14**, 256–264 (2015). Available in http://www.wseas.org/multimedia/journals/mathematics/2015/a4872609-118.pdf
5. Gambi, J.M., Rodriguez-Teijeiro, M.C., García 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)
6. Gambi, J.M., Clares, J., García del Pino, M.L.: FDOA post-Newtonian equations for the location of passive emitters placed in the vicinity of the Earth. Aerosp. Sci. Technol. **46**, 137–145 (2015)
7. Gambi, J.M., Rodriguez-Teijeiro, M.C., García del Pino, M.L.: Newtonian and post-Newtonian passive geolocation by TDOA. Aerosp. Sci. Technol. **51**, 18–25 (2016)
8. Moritz, H., Hofmann-Wellenhof, B.: Geometry, Relativity, Geodesy. H. Wichmann Verlag GmbH, Berlin (1993)
9. Schmitz, M., Fasoulas, S., Utzmann, J.: Performance model for space-based laser debris sweepers. Acta Astronaut. **115**, 376–386 (2015)
10. Soffel, M.H.: Relativity in Astrometry, Celestial Mechanics and Geodesy. Springer, Berlin (1989)
11. Synge, J.L.: Relativity: The General Theory, pp. 87–95. North-Holland, Amsterdam (1960)

Post-Newtonian Corrections to the Newtonian Predictions for the Motion of Designated Targets with Respect to Space Based APT Laser Systems

Jose M. Gambi, Maria L. García del Pino, Jürgen Gschwindl, and Ewa B. Weinmüller

Abstract Numerical experiments are carried out to validate the short/long term differences between the solutions of the Newtonian equations for the relative motion of middle size targets in space with respect to space based APT systems and the respective solutions of a system of non-linear post-Newtonian equations. This system has been introduced in the ECMI 2016 contribution *Non-linear post-Newtonian equations for the motion of designated targets with respect to space based APT laser systems*. Two auxiliary systems of post-Newtonian equations are used to carry out the validation. The simulations are made under the following assumptions: (i) the structures of the Earth surrounding space are respectively the Euclidean and that of the post-Newtonian approximation to the exterior Schwarzschild field; (ii) the targets are on equatorial circular orbits, and (iii) the APT systems are ECI oriented inertial-guided systems placed onboard HEO, MEO and LEO satellites in equatorial orbits about the Earth. The APT systems have initially been placed at short and successively increasing distances from the targets.

1 Introduction

Numerical solutions of the Newtonian equations for the Earth Centered Inertial (ECI) relative motions of middle size targets with respect to space based

J.M. Gambi (✉)
Gregorio Millán Institute, Univ. Carlos III de Madrid, 28911 Madrid, Spain
e-mail: gambi@math.uc3m.es

M.L. García del Pino
Department of Mathematics, I. E. S. Alpajes, 28300 Aranjuez, Madrid, Spain
e-mail: lgarciadelpino@educa.madrid.org

J. Gschwindl • E.B. Weinmüller
Institute for Analysis and Scientific Computing, Vienna University of Technology, A-1040 Wien, Austria
e-mail: j.gschwindl@gmx.at; e.weinmueller@tuwien.ac.at

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_74

laser Acquisition, Pointing and Tracking (APT) systems are used to estimate the short/long term differences between the predictions provided by these equations and those provided by three systems of post-Newtonian equations.

The first post-Newtonian system describes the relative motions as differences of the ECI post-Newtonian equations for the targets and systems in the Earth Schwarzschild field, thus resembling the procedure followed to derive the Newtonian equations [6]; the second, by means of the linear approximation to Synge's equations for the relative motion in Fermi coordinates derived by Gambi et al. [1], and the third, by means of the equations introduced in the contribution *Non-linear post-Newtonian equations for the motion of designated targets with respect to space based APT laser systems* presented in this ECMI 2016 Minisymposium.

The estimations are made under the assumption that the structures of space-time about the Earth are respectively those of the Newtonian and of the post-Newtonian approximation to the exterior Schwarzschild field. The APT systems are assumed to be ECI oriented inertial-guided systems that, from the ranging and angle measurements made, finally assign Cartesian coordinates $(X_{(\alpha)}, t) = (x^{\alpha_2} - x^{\alpha_1}, t)$, where (x^{α_2}, t) and (x^{α_1}, t) $(\alpha = 1, 2, 3)$ are ECI coordinates of the targets and the systems, respectively. To consider simulations resembling actual configurations, the APT systems are assumed to be onboard GEO satellites, and on LEO and MEO satellites in equatorial orbits about the Earth; the targets are assumed to be on equatorial circular orbits, and finally, the systems are assumed to be initially placed at short, and successively increasing distances from the targets.

The computations here carried out for each configuration and the magnitudes analyzed correspond to (i) the ECI Newtonian and post-Newtonian orbits of S and T (S stands for the APT systems, and T, for the targets); (ii) the Newtonian and post-Newtonian relative orbits of T with respect to S on the basis of the three different systems of equations mentioned above; (iii) the ranges from the ECI Newtonian to the three ECI post-Newtonian positions of S; (iv) the ranges from the ECI Newtonian to the three ECI post-Newtonian positions of T; (v) the differences of the Newtonian and post-Newtonian distances from the ECI center to S; (vi) the differences of the Newtonian and post-Newtonian distances from the ECI center to T; (vii) the distances from the Newtonian to the post-Newtonian linear and non-linear relative positions of T with respect to S, and finally, (viii) the distances from the ECI post-Newtonian position of T to the position of T obtained by adding the ECI post-Newtonian position of S to the relative post-Newtonian position of T with respect to S derived from the non-linear equations.

2 The Systems of Equations

The notations used in the papers cited above were unified in this work by making $(x^1, x^2) = (x, y)$, $(X_{(1)}, X_{(2)}) = (x^{1_2} - x^{1_1}, x^{2_2} - x^{2_1}) = (X, Y)$, $(x_1, y_1) = (x^{1_1}, x^{2_1})$ and $(x_2, y_2) = (x^{1_2}, x^{2_2})$ throughout. The ECI and relative Newtonian equations whose solutions are considered as reference of the subsequent results are the known

classical equations (see e.g. [2])

$$\frac{d^2x}{dt^2} = \frac{-mx}{r^3}, \quad \frac{d^2y}{dt^2} = \frac{-my}{r^3}, \tag{1}$$

$$\frac{d^2X}{dt^2} = -m\left(\frac{x_2}{r_2^3} - \frac{x_1}{r_1^3}\right), \quad \frac{d^2Y}{dt^2} = -m\left(\frac{y_2}{r_2^3} - \frac{y_1}{r_1^3}\right), \tag{2}$$

where $r^2 = (x)^2 + (y)^2$, $r_1^2 = (x_1)^2 + (y_1)^2$ and $r_2^2 = (x_2)^2 + (y_2)^2$.

The first system of relative post-Newtonian equations are derived from the difference between the equations of the geodesics for S and T in the weak approximation to the Earth Schwarzchild field [7]. After a straightforward calculation, we have that the equations of the geodesics are [6]

$$\frac{d^2x}{dt^2} = \frac{-m}{r^3}\left[1 - \frac{2m}{r} + \left(1 - \frac{3(x)^2}{r^2}\right)\left(\frac{dx}{dt}\right)^2 - \frac{6xy}{r^2}\frac{dx}{dt}\frac{dy}{dt} + \left(1 - \frac{3(y)^2}{r^2}\right)\left(\frac{dy}{dt}\right)^2\right]x,$$

$$\frac{d^2y}{dt^2} = \frac{-m}{r^3}\left[1 - \frac{2m}{r} + \left(1 - \frac{3(x)^2}{r^2}\right)\left(\frac{dx}{dt}\right)^2 - \frac{6xy}{r^2}\frac{dx}{dt}\frac{dy}{dt} + \left(1 - \frac{3(y)^2}{r^2}\right)\left(\frac{dy}{dt}\right)^2\right]y, \tag{3}$$

so that

$$\frac{d^2X}{dt^2} = \frac{d^2x_2}{dt^2} - \frac{d^2x_1}{dt^2}, \quad \frac{d^2Y}{dt^2} = \frac{d^2y_2}{dt^2} - \frac{d^2y_1}{dt^2}. \tag{4}$$

The linear relative equations are [1]

$$\frac{d^2X}{dt^2} = -mX\int_0^1\left(1 - \frac{3\left[x_1(1-u) + x_2u\right]^2}{r(u)^2}\right)\frac{1 - 2u + 3u^2}{r(u)^3}du$$

$$+mY\int_0^1\left(\frac{3x_1y_1(1-u)^2 + 3(1-u)u(x_1y_2 + y_1x_2) + 3x_2y_2u^2}{r(u)^2}\right)\frac{1 - 2u + 3u^2}{r(u)^3}du,$$

$$\frac{d^2Y}{dt^2} = -mY\int_0^1\left(1 - \frac{3\left[y_1(1-u) + y_2u\right]^2}{r(u)^2}\right)\frac{1 - 2u + 3u^2}{r(u)^3}du$$

$$+mX\int_0^1\left(\frac{3x_1y_1(1-u)^2 + 3(1-u)u(x_1y_2 + y_1x_2) + 3x_2y_2u^2}{r(u)^2}\right)\frac{1 - 2u + 3u^2}{r(u)^3}du, \tag{5}$$

where $r(u)^2 = \left[x_1(1-u) + x_2u\right]^2 + \left[y_1(1-u) + y_2u\right]^2$.

Finally, the relative non-linear equations are

$$\frac{d^2X}{dt^2} = -mX\int_0^1 \Big(1 - \frac{3\big[x_1\,(1-u) + x_2u\big]^2}{r(u)^2}\Big)\frac{1-2u+3u^2}{r(u)^3}du$$

$$+mY\int_0^1 \Big(\frac{3x_1y_1\,(1-u)^2 + 3\,(1-u)\,u\,(x_1y_2 + y_1x_2) + 3x_2y_2u^2}{r(u)^2}\Big)\frac{1-2u+3u^2}{r(u)^3}du$$

$$-9mX^2\int_0^1 \Big(1 - \frac{5\big[x_1\,(1-u) + x_2u\big]^2}{3r(u)^2}\Big)\frac{x_1(1-u)^2u + x_2(1-u)u^3}{r(u)^5}du$$

$$-6mXY\int_0^1 \Big(1 - \frac{5\big[x_1\,(1-u) + x_2u\big]^2}{r(u)^2}\Big)\frac{y_1(1-u)^2u + y_2(1-u)u^3}{r(u)^5}du$$

$$-3mY^2\int_0^1 \Big(1 - \frac{5\big[y_1\,(1-u) + y_2u\big]^2}{r(u)^2}\Big)\frac{x_1(1-u)^2u + x_2(1-u)u^3}{r(u)^5}du,$$

$$\frac{d^2Y}{dt^2} = -mY\int_0^1 \Big(1 - \frac{3\big[y_1\,(1-u) + y_2u\big]^2}{r(u)^2}\Big)\frac{1-2u+3u^2}{r(u)^3}du$$

$$+mX\int_0^1 \Big(\frac{3x_1y_1\,(1-u)^2 + 3\,(1-u)\,u\,(x_1y_2 + y_1x_2) + 3x_2y_2u^2}{r(u)^2}\Big)\frac{1-2u+3u^2}{r(u)^3}du$$

$$-3mX^2\int_0^1 \Big(1 - \frac{5\big[x_1\,(1-u) + x_2u\big]^2}{r(u)^7}\Big)\frac{y_1(1-u)^2u + y_2(1-u)u^3}{r(u)^5}du$$

$$-6mXY\int_0^1 \Big(1 - \frac{5\big[y_1\,(1-u) + y_2u\big]^2}{r(u)^2}\Big)\frac{x_1(1-u)^2u + x_2(1-u)u^3}{r(u)^5}du$$

$$-9mY^2\int_0^1 \Big(1 - \frac{5\big[y_1\,(1-u) + y_2u\big]^2}{3r(u)^2}\Big)\frac{y_1(1-u)^2u + y_2(1-u)u^3}{r(u)^5}du, \qquad (6)$$

which are deduced by means of simple computations from the two first equations introduced in the contribution *Non-linear post-Newtonian equations for the motion of designated targets with respect to space based APT laser systems* mentioned above.

3 Numerical Simulations

In searching for patterns from the data produced by the numerical simulations made, each experiment was included into one of the three groups corresponding to the LEO-MEO-GEO classification. In turn, each group was split into two subgroups according to the ratio of the orbital radii of S and T, r_S and r_T ($r_S/r_T > 1, r_S/r_T < 1$).

Finally, the simulations in each of the resultant subgroups were ordered according to the initial distance from S to T. As a result, we found two family of patterns associated to the ratio r_S/r_T.

The following figures illustrate the results; they correspond to a rather standard S-T configuration for which $r_S/r_T > 1$ (see e.g. [3–5]): T is in LEO, 631 km above the Earth, and S is initially placed 200 km above T. (Note that these are approximated data. In fact, for the sake of accuracy, the equations were computed after having made $c = G = 1$, so that both data and results were given, resp. obtained, in seconds, and the accuracy achieved was 10^{-10}. We note finally that the results shown in these figures are in cm.)

Figure 1, left shows the ECI Newtonian orbits of T (red) and S (blue) derived from (1). The green line represents the Earth surface, and the black solid line, the distance from S to T after 2 h orbiting. On the right, it is shown the corresponding relative orbit of T with respect to S, derived from (2). The results are in cm, as was said previously.

Figure 2 shows the distances, D, from the ECI Newtonian to the ECI post-Newtonian positions of T, and the differences, d, between the respective orbital radius of T. The orbital time is again 2 h and the results appear in cm to facilitate the physical interpretation of the graphs.

The Newtonian and the three post-Newtonian relative orbits of T with respect to S are plotted in Fig. 3, left. The orbit showing a clear deviation (the green trajectory) corresponds to the solution of (5), which are the linear equations. Since all the experiments carried out lead to similar (time-increasing) deviations from the early stages of the integrations, system (5) is disregarded as candidate to account for the post-Newtonian corrections. Since the solutions of (2) and of (4), (6), i.e., of the other two post-Newtonian equations, are undistinguishable from each other in this graphic (they are represented by the red line) and in all similar graphics, deep analysis of these solutions were necessary.

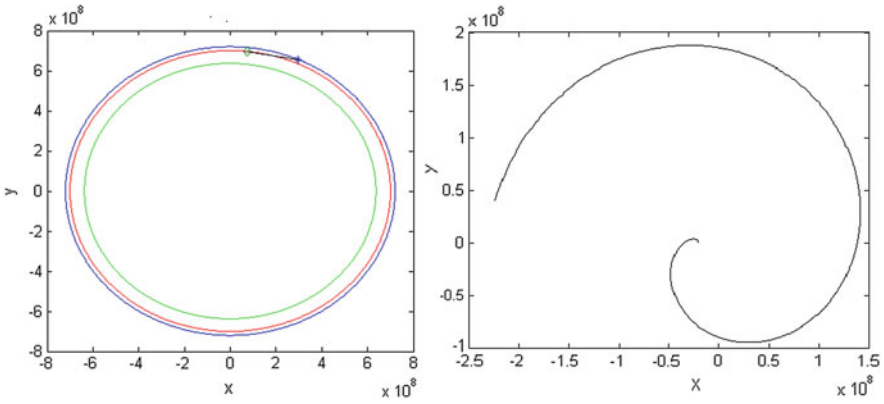


Fig. 1 ECI Newtonian orbits of S and T and relative Newtonian orbit of T with respect to S

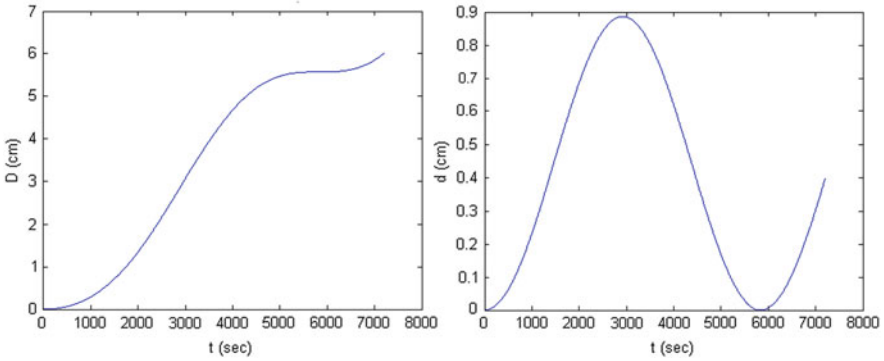


Fig. 2 Distances from the ECI Newtonian to the ECI post-Newtonian positions of T and differences between the orbital radius

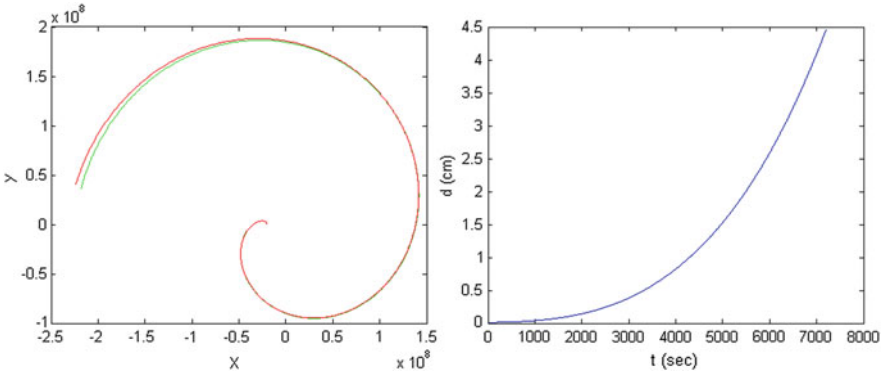


Fig. 3 Relative orbits of T w.r.t. S and distances from the p-N positions of T to the p-N positions of S plus the solutions of (6)

From the order of the differences of the respective solutions we found that the size of the differences between the solutions of (2) and (4) are not large enough for the curved structure of the space-time about the Earth, which is given by the Riemann tensor, manifests throughout them. In fact, as Fig. 3 on the right shows, the distances from the positions of T derived from (3) to the positions derived by adding the positions of S, also obtained with (3), to the solution of (6) are the only corrections that satisfy this requirement. Therefore, we can conclude that (4) must also be disregarded, since, unlike for the role played by Eq. (2) with respect to (1), Eq. (4) cannot play the same role with respect to Eq. (3); or shortly said (4) cannot yield relative motions within the post-Newtonian framework.

Table 1 shows the size of some differences between the solutions of (2) and (6). They are depicted to illustrate the magnitude of the post-Newtonian corrections corresponding to this example. We note to this respect that orbiting along 2 h (the time considered here) is equivalent to 1.23 revolutions of T about the Earth, approximately.

Table 1 Post-Newtonian corrections from Eq. (6)[a]

Time	Correction	Time	Correction
4555.102	1.03	6024.490	2.43
4848.979	1.26	6318.367	2.78
5142.857	1.52	6612.245	3.16
5436.735	1.80	6906.122	3.58
5730.612	2.10	7200	4.03

[a] Time in s, and corrections, in cm

4 Conclusions

From the simulations carried out in this work, we conclude that for effective tracking of middle size targets, the non-linear system (6) is the appropriate for an APT systems to correct the Newtonian predictions used at practically any instant and any *S*-*T* distance, particularly at the late stages of the tracking. The reason is that the corrections due to the non-linear terms included in these equations make the size of the total post-Newtonian corrections to fit very neatly with the size of the gravitational corrections corresponding to the post-Newtonian approximation. In fact, the underlying curved structure of space-time accounted for this approximation becomes apparent through this system of equations. Further, the corrections become measurable shortly after the initial integration instants. Therefore, they can be taken into consideration to increase the pointing accuracy of the APT systems, since the size of these corrections may be larger than the size of the targets in the scenarios considered.

References

1. Gambi, J.M., García del Pino, M.L.: Post-Newtonian equations of motion for inertial guided space APT systems. WSEAS Trans. Math. **14**, 256–264 (2015). Available in http://www.wseas.org/multimedia/journals/mathematics/2015/a4872609-118.pdf
2. Montenbruck, O., Gill, E.S.: Satellite Orbits. Springer, Berlin (2000)
3. Phipps, C.R.: L'ADROIT – a spaceborne ultraviolet laser system for space debris cleaning. Acta Astronaut. **104**, 243–255 (2015)
4. Schmitz, M., Fasoulas, S., Utzmann, J.: Performance model for space-based laser debris sweepers. Acta Astronaut. **115**, 376–386 (2015)
5. Shuangyan, S., Xing, J., Hao, C.: Cleaning space debris with a space-based laser system. Chin. J. Aeronaut. **24**(4), 805–811 (2014)
6. Soffel, M.H.: Relativity in Astrometry, Celestial Mechanics and Geodesy. Springer, Berlin (1989)
7. Synge, J.L.: Relativity: The General Theory, pp. 91–95. North-Holland, Amsterdam (1960)

Acoustics in 2D Spaces of Constant Curvature

Michael M. Tung, José M. Gambi, and María L. García del Pino

Abstract Maximally symmetric spaces play a vital role in modelling various physical phenomena. The simplest representative is the 2-sphere $\mathbb{S}^2$, having constant positive curvature. By embedding it into $(2 + 1)$D spacetime with Lorentzian signature it becomes the prototype of homogeneous and isotropic spacetime of constant curvature with constant scale factor: the Einstein cylinder $\mathbb{R} \times \mathbb{S}^2$. This work outlines a variational approach on how to model acoustic wave propagation on this particular curved spacetime. On the Einstein cylinder, the analytical solutions of the wave equation for the acoustic potential are shown to reduce to solutions of a differential equation of Sturm-Liouville type and simple harmonic time and angular dependence. Moreover, we discuss the implementation of such an underlying curved spacetime within an acoustic metamaterial—an artificially engineered material with remarkable properties exceeding the possibilities found in nature.

1 Introduction

The principal aim of metamaterial research is the theoretical design and conception of artificial materials followed by its industrial engineering. Metamaterials possess remarkable properties which by far exceed the ones found in nature. This makes them not only attractive for fundamental research but they will certainly provide useful future applications in all sectors of living.

The theoretical framework for designing advanced devices with acoustic metamaterials belongs to so-called analogue models of gravity [18]. Its idea is to model physical phenomena as close as possible to general relativity without giving up too much of its underlying differential-geometric structure. Obviously classical

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

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_75

acoustics is not a relativistic theory—in acoustics there is no such equivalent as a constant speed of light, independent of the observer's reference frame—but certain features can be adapted and carried over with the prime objective to emulate acoustic wave propagation in curved spacetime.

By implementing specific spacetimes in acoustics, transformation acoustics [2, 4, 11, 15] can be used, for example, to design perfect acoustic lenses with unlimited resolution (see e.g. [17]), to construct ships and submarines invisible to sonar detection, to improve concert halls, and to devise applications involving acoustic cloaking [2, 5, 14].

Spaces of constant curvature are maximally symmetric spaces (see e.g.[19]) which explains their importance in a variety of important applications in physics and engineering, such as e.g. the description of uncharged, perfect relativistic fluids [8] and standard cosmological models [6]. Moreover, in the past years, quantum mechanical phenomena in spaces of constant curvature have attracted the focus of intense investigation [12], raising critical fundamental questions beyond their possible experimental verification. However, the simulation of acoustic phenomena [4] in such spaces has so far been vastly neglected.

Among the most significant spaces of constant curvature are the n-spheres embedded in $(n + 1)$-dimensional Euclidean space, denoted by $\mathbb{S}^n(K)$, where $K > 0$ is the curvature. The corresponding m-dimensional spacetime consists of the Lorentzian manifold $\mathbb{R} \times \mathbb{S}^{m-1}(K)$. It is the simplest non-Euclidean geometry with elliptic geometry, that is, a homogeneous and isotropic spacetime of Robertson-Walker type with positive curvature and constant scale factor. In this sense, it may be considered as the counterpart of flat Minkowski space for the sphere.

In 2012, we proposed a general framework for transformation acoustics [14] using a variational principle for the acoustic potential in order to model acoustic wave propagation with a curved background space. We also obtained the general constitutive equations, linking a chosen spacetime with the relevant physical parameters for the acoustic metamaterial. So far several applications of interest already have been studied [14–17]. Here we shall examine and implement in transformation acoustics the static spacetime $\mathbb{R}\times\mathbb{S}^2$—sometimes called the *Einstein cylinder* [3].

In the following, we demonstrate how this approach yields a partial differential equation for the acoustic potential, which leads to a classical Sturm-Liouville problem for the radial isotropic coordinates that can be tackled analytically. This will make further investigations and expected wave predictions possible. We also comment on the design and implementation of such spacetime with suitable acoustic metadevices.

2 Variational Principle and Constitutive Equations

Hamilton's or Fermat's variational principle are powerful methods in classical mechanics and optics to describe in a very concise manner the laws which govern the physical phenomena for these respective fields. In this formalism the solutions

which extremize the postulated action integral yield the equations of motion that completely determine the system in question. Variational principles owe much of their elegance due to their coordinate-frame independence and because symmetry properties are easily revealed by Noether's theorem. Moreover, in this approach a physical law manifests itself in a self-adjoint differential operator acting on the physical fields in question [9]. As a consequence the equations of motion are typically separable partial differential equations which include a Sturm-Liouville problem for one of the variables (see also [7]). This dramatically simplifies the analytical or semi-analytical treatment of these solutions.

Hamilton's variational principle in transformation acoustics for a smooth spacetime M (with Lorentzian metric $\mathbf{g}$ so that $g = \det \mathbf{g} < 0$) only requires to postulate the explicit form of the Lagrangian density function $\mathscr{L} : M \times TP \to \mathbb{R}$, where P is the ambient space defined by the acoustic potential $\phi : M \to \mathbb{R}$. In a fixed laboratory frame the gradient[1] satisfies [14]

$$\phi_{;\mu} = \phi_{,\mu} = \begin{pmatrix} p/c\varrho_0 \\ -\mathbf{v} \end{pmatrix}, \tag{1}$$

where in the acoustic metamaterial $\mathbf{v}$ denotes the local fluid velocity, p the acoustic pressure, and ϱ_0 its density. As usual, $c > 0$ is the time-independent acoustic wave speed[10].

In general, the Lagrangian may be a function of x^μ, ϕ, and $\phi_{,\mu}$ as well as higher-order derivatives. However, physical symmetry constraints as energy-momentum conservation, locality and free-wave propagation restrict the acoustic Lagrangian to take the following simplest possible form [14], containing only a covariant kinetic term:

$$\mathscr{L}(\phi_{,\mu}) = \tfrac{1}{2}\sqrt{-g}\, g^{\mu\nu}\phi_{,\mu}\phi_{,\nu}. \tag{2}$$

Thus, the associated action integral is given by

$$\mathscr{A}[\phi] = \int d\mathrm{vol}_g \mathscr{L}(\phi_{,\mu}) \tag{3}$$

and its functional derivative with respect to the field variable ϕ must vanish [14]:

$$\frac{\delta}{\delta\phi}\mathscr{A}[\phi] = 0 \quad \Rightarrow \quad \frac{\delta}{\delta\phi}\int_\Omega d\mathrm{vol}_g\, g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} = 0. \tag{4}$$

[1] Greek tensor indices indicate the full range of spacetime values, whereas Latin will only refer to the spatial values. Comma and semicolon are standard notation for partial and covariant derivatives, respectively. For scalars, partial and covariant derivative are identical.

The integration Ω domain is a bounded, closed set of spacetime and $d\mathrm{vol}_g = \sqrt{-g}\, dx^0 \wedge \ldots \wedge dx^3$ is the invariant volume element (see e.g. [13, p. 14]). Equation (4) produces the Euler-Lagrange equation for the acoustic potential, which is sufficient to fully describe the dynamics of the acoustic system with underlying spacetime metric $\mathbf{g}$.

The constitutive equations describe how to precisely implement the acoustic system in the laboratory—this is *physical space* and denoted by unbarred quantities. *Virtual space*, on the other hand, is the space with known acoustic wave propagation and denoted by barred quantities. The relation of mass-density tensor $\boldsymbol{\rho} = \rho_{ij}\, dx^i \otimes dx^j$ and bulk modulus κ in both transformation spaces are determined by [14]

$$\kappa = \frac{\sqrt{-g}}{\sqrt{-\bar{g}}}\, \bar{\kappa}, \qquad \rho_0 \rho^{ij} = \frac{\sqrt{-\bar{g}}}{\sqrt{-g}}\, \bar{g}^{ij}, \tag{5}$$

where without loss of generality $\bar{\rho}/\rho_0 \equiv 1$. Moreover, for simplicity $\bar{\kappa}$ may be set to unity.

2.1 The Einstein Cylinder and Its Acoustic Implementation

The spacetime line element for the *Einstein cylinder* $\mathbb{R} \times \mathbb{S}^2$ may be written in terms of the solid angle $d\Omega^2 = d\vartheta^2 + \sin^2 \vartheta\, d\varphi^2$, with constant $a > 0$, so that

$$ds^2 = -c^2 dt^2 + a^2 d\Omega^2. \tag{6}$$

It will be convenient to use isotropic radial coordinates, taking $r = a \sin \vartheta$, in which the line element takes the form:

$$ds^2 = -(c\, dt) \otimes (c\, dt) + \frac{dr}{\sqrt{1 - r^2/a^2}} \otimes \frac{dr}{\sqrt{1 - r^2/a^2}} + (r d\varphi) \otimes (r d\varphi). \tag{7}$$

Then, we may identify the nonholonomic (noncoordinate) basis 1-forms as $\theta^0 = c\, dt$, $\theta^1 = dr/\sqrt{1 - r^2/a^2}$, and $\theta^2 = r\, d\varphi$. Cartan's structure equations readily yield the only independent curvature 2-form in this frame as $\hat{\Omega}^1{}_2 = \hat{R}^1{}_{212}\, \theta^1 \wedge \theta^2$, where the only independent component of the Riemann tensor is $\hat{R}^1{}_{212} = 1/a^2 = -\hat{R}^2{}_{112}$. All other components vanish.

From this the only non-zero components of the Ricci tensor in the nonholonomic frame are computed as $\hat{R}_{11} = \hat{R}_{22} = 1/a^2$. Therefore the associated Einstein tensor also has only the following non-zero components

$$G_{00} = \hat{G}_{00} = \hat{R}_{00} - \frac{1}{a^2}\, \eta_{00} = \frac{1}{a^2}, \tag{8}$$

a result which equally holds for both frames, nonholonomic or coordinate frame. Comparing $G_{\mu\nu}$ with the stress-energy tensor of a perfect fluid (being shear-free and isotropic), we conclude that an observer falling along a geodesic in spacetime $\mathbb{R}\times\mathbb{S}^2$ is pressure-free and only has constant mass-energy density $\rho_0 = c^2/8\pi G a^2$, where here c is the speed of light and G is the gravitational constant. It is just this mass-energy distribution which generates $\mathbb{R}\times\mathbb{S}^2$.

For the implementation of the equivalent acoustic space and its relevant physical parameters, we only require the components of metric $\mathbf{g}$ implicit in Eq. (7), namely

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & \dfrac{1}{1-r^2/a^2} & 0 \\ 0 & 0 & r^2 \end{pmatrix}. \tag{9}$$

Substituting this metric into the constitutive equations, Eq. (5), immediately gives the prescription

$$\kappa = \frac{1}{\sqrt{1-r^2/a^2}}, \qquad \rho_0\,\rho^{ij} = \sqrt{1-r^2/a^2}\begin{pmatrix} 1 & 0 \\ 0 & 1/r^2 \end{pmatrix}. \qquad 0<r<a. \tag{10}$$

2.2 Acoustic Wave Propagation on the Einstein Cylinder

The variational principle, Eq. (4), with underlying metric, Eq. (9), generates the acoustic wave equation as geodesics for field ϕ on the Einstein cylinder $\mathbb{R}\times\mathbb{S}^2$. The associated Euler-Lagrange equation is

$$\Delta_{\mathbb{R}\times\mathbb{S}^2}\phi = -\frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} + \frac{\sqrt{1-r^2/a^2}}{r}\frac{\partial}{\partial r}\left(r\sqrt{1-r^2/a^2}\,\frac{\partial\phi}{\partial r}\right) + \frac{\partial^2\phi}{\partial\varphi^2} = 0, \tag{11}$$

where $\Delta_{\mathbb{R}\times\mathbb{S}^2}$ is the Laplace-Beltrami operator on manifold $\mathbb{R}\times\mathbb{S}^2$. For the derivation in the general case with pseudo-Riemannian manifold M see [14].

The solution $\phi(t,r,\varphi)$ for wave equation, Eq. (11), displays a harmonic dependence in the time variable t and for azimuthal angle φ. All of the non-trivial behaviour is contained in the radial dependence, as expected. Employing the standard technique, separation of variables [7] yields a solution of the general form

$$\phi(t,r,\varphi) = \phi_1(r)\left[A\cos\left(\sqrt{\lambda}ct\right) + B\sin\left(\sqrt{\lambda}ct\right)\right]\left[C\cos\left(\sqrt{\mu}\,\varphi\right) + D\sin\left(\sqrt{\mu}\,\varphi\right)\right] \tag{12}$$

$$\text{with}\quad \left(1-\frac{r^2}{a^2}\right)\phi_1'' + \left(\frac{1}{r} - \frac{2r}{a^2}\right)\phi_1' + (\lambda-\mu)\phi_1 = 0, \tag{13}$$

where $\phi_1 = \phi_1(r)$ is the radial function, $\lambda, \mu > 0$ are the harmonic eigenvalues, and $A, \ldots, D$ are integration constants. The general analytic solutions of Eq. (13) can be expressed in terms of a hypergeometric series and Meijer-G functions. (see e.g. [1]). In practice, however, a semi-numerical approach is advisable, where Eq. (13) is assessed numerically. Of significant interest are also certain limiting cases; e.g. note that for $r \to a$ the r-dependent solution approaches $\phi_1(r) \sim e^{a(\lambda-\mu)r}$.

3 Conclusions

The acoustic analogue of the Einstein cylinder is a particularly intriguing model for transformation physics as it was the first and simplest cosmological model to be formulated within Einstein's geometric gravity.

Whereas in gravity the Einstein-cylinder world requires space to be filled with a uniform static mass-energy distribution, we have shown that its acoustic pendant requires for its metamaterial implementation a bulk modulus and isotropic density which display a specific radial dependence, viz. Eq. (5).

We have also outlined how a covariant variational principle gives rise to the wave equation for the acoustic potential which provides the complete description of acoustic free-wave phenomena with the underlying spacetime of an Einstein cylinder. Although fully analytic solutions are available in terms of hypergeometric series and Meijer-G functions, it might be more practicable to numerically approximate the radial dependence of the potential via the derived second-order linear differential equation of Sturm-Liouville type. Time and azimuthal angular dependence are harmonic and pose no difficulties.

The variational spacetime approach to transformation acoustics supplies a powerful tool for the study and design of acoustic metadevices. It may help to open up new research pathways in this field, overcoming challenges in the engineering of acoustic phenomena with curved background spacetimes.

Acknowledgements M. M. T. wishes to thank the Spanish *Ministerio de Economía y Competitividad* and the European Regional Development Fund (ERDF) for financial support under grant TIN2014-59294-P.

References

1. Beals, R., Szmigielski, J.: Meijer G-functions: a gentle introduction. Not. Am. Math. Soc. **60**(7), 866–872 (2013)
2. Chen, H.Y., Chan, C.T.: Acoustic cloaking and transformation acoustics. J. Phys. D **43**(11), 113001 (2010)
3. Choquet-Bruhat, Y., Damour, T.: Introduction to General Relativity, Black Holes, and Cosmology. Oxford University Press, Oxford (2015)

4. Cummer, S.A.: Transformation acoustics. In: Craster, V.R., Guenneau, S. (eds.) Acoustic Metamaterials: Negative Refraction, Imaging, Lensing and Cloaking, pp. 197–218. Springer Netherlands, Dordrecht (2013)
5. Cummer, S.A., Schurig, D.: One path to acoustic cloaking. New J. Phys. **9**(3), 45–52 (2007)
6. Islam, J.N.: An Introduction to Mathematical Cosmology. Cambridge University Press, Cambridge (2001)
7. Kalnins, E.G.: Separation of Variables for Riemannian Spaces of Constant Curvature. Pitman Monographs and Surveys in Pure and Applied Mathematics. Longman Scientific & Technical, New York (1986)
8. Kuchowicz, B.: Conformally flat space-time of spherical symmetry in isotropic coordinates. Int. J. Theor. Phys. **7**(4), 259–262 (1973)
9. Lanczos, C.: The Variational Principles of Mechanics. Dover Publications, New York (1970)
10. Mechel, F.P.: Formulas of Acoustics. Springer, Berlin (2002)
11. Norris, A.N.: Acoustic metafluids. J. Acoust. Soc. Am. **125**(2), 839–849 (2009)
12. Redkov, V.M., Ovsiyuk, E.M.: Quantum mechanics in spaces of constant curvature. In: Contemporary Fundamental Physics. Nova Science, New York (2012)
13. Rosenberg, S.: The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds. London Mathematical Society Student Text, vol. 31. Cambridge University Press, Cambridge (1997)
14. Tung, M.M.: A fundamental Lagrangian approach to transformation acoustics and spherical spacetime cloaking. Europhys. Lett. **98**, 34002–34006 (2012)
15. 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. 19. Springer, Berlin (2014)
16. Tung, M.M., Weinmüller, E.B.: Gravitational frequency shifts in transformation acoustics. Europhys. Lett. **101**, 54006–54011 (2013)
17. Tung, M.M., Gambi, J.M., García del Pino, M.L.: Maxwell's fish-eye in (2+1)D spacetime acoustics. In: Russo, G.R., Capasso, V., Nicosia, G., Romano, V. (eds.) Progress in Industrial Mathematics at ECMI 2014. Mathematics in Industry, vol. 22. Springer, Berlin (2016)
18. Visser, M., Barceló, C., Liberati, S.: Analogue models of and for gravity. Gen. Rel. Grav. **34**, 1719–1734 (2002)
19. Wolf, J.A.: Spaces of Constant Curvature. American Mathematical Society, Providence, Rhode Island (2011)

Minisymposium: Stochastic PDEs and Uncertainty Quantification with Applications in Engineering

Clemens Heitzinger and Hermann Matthies

Description

Stochastic partial differential equations are becoming more and more important to model uncertainties, noise, fluctuations, process variations, material properties etc. in various applications.

Stochastic PDEs subsume the usability of PDEs and extend the range of applications that can be modeled. In short, higher-order moments of the solutions can be modeled. The emergence and growth of the field of uncertainty quantification in recent years shows that there is profound interest in these questions in many different application areas.

Modeling and simulation based on stochastic PDEs poses many new and exciting problems regarding the analysis of the solutions, the development of efficient numerical algorithms, optimal-control problems, optimization, and parameter estimation.

This minisymposium brought together researchers in stochastic PDEs and offered an opportunity to discuss the latest developments. Common themes were how uncertain input in engineering problems can be included in mathematical models, how randomness propagates through a PDE model, and how optimization problems can eventually be solved in view of the randomness. Both theoretical aspects such as existence, uniqueness, and regularity of the solutions as well as numerical methods and approaches were covered.

C. Heitzinger (✉)
Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria
e-mail: Clemens.Heitzinger@TUWien.ac.at

H. Matthies
Institut für Wissenschaftliches Rechnen, Technische Universität Braunschweig, Mühlenpfordtstraße 23, 38106 Braunschweig, Germany
e-mail: wire@tu-bs.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_76

We expect that discussing the similarities of and the differences between leading model equations such as elliptic equations, parabolic equations, the Schrödinger equation, the Boltzmann equation, and the Maxwell equations was fruitful and provided guidance for future research with applications in engineering.

The talks included in the minisymposium were the following:

- Jianyu Li, Roger Ghanem. Tianjin University of Science and Technology, Tianjin, China; University of Southern California. *Numerical solution method for stochastic variational inequalities with polynomial chaos decompositions.*
- José A. Morales Escalante, Irene Gamba. ICES, UT Austin. *Reflective boundary conditions in discontinuous Galerkin methods for Boltzmann-Poisson models of electron transport in semiconductors and zero flux condition for general mixed reflection.*
- Marc Dambrine, Charles Dapogny, Helmut Harbrecht. Universität Basel. *Shape optimization for quadratic functional and states with random right-hand sides.*
- Zeger Bontinck, Oliver Lass, Sebastian Schöps. TU Darmstadt. *Robust optimization of the size of permanent magnets in a synchronous machine using deterministic and stochastic approaches.*
- P. Benner, Akwuzm Onwunta, M. Stol. Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg. *Fast solvers for optimal control problems constrained by PDEs with uncertain inputs.*

Uncertainty Quantification for a Permanent Magnet Synchronous Machine with Dynamic Rotor Eccentricity

Zeger Bontinck, Oliver Lass, Herbert De Gersem, and Sebastian Schöps

Abstract The influence of dynamic eccentricity on the harmonic spectrum of the torque of a permanent magnet synchronous machine is studied. The spectrum is calculated by an energy balance method. Uncertainty quantification is applied by using generalized Polynomial Chaos and Monte Carlo. It is found that the displacement of the rotor impacts the spectrum of the torque the most.

1 Introduction

During mass production of permanent magnet synchronous machines (PMSMs) small deviations on the machine are introduced. These deviations may lead to underperformance. Therefore the field of uncertain quantification is gaining more and more interest for the machine simulation. The rotor position is a typical uncertainty in an electrical machine design. It is known that rotor eccentricity increases the cogging torque (see e.g. [3]). In the literature different types of eccentricity are identified. Static eccentricity means that the rotor is shifted out of the center of the stator but still rotates around the center of the rotor. In the case of dynamic eccentricity, the rotor rotates around the center of the stator. Also more complex eccentricities like e.g. an inclined rotor shaft are possible (see e.g. [2]).

Z. Bontinck (✉) • S. Schöps
Graduate School of Computational Engineering, Technische Universität Darmstadt, Dolivostr. 15, 64293 Darmstadt, Germany
e-mail: bontinck@gsc.tu-darmstadt.de; schoeps@gsc.tu-darmstadt.de

O. Lass
Department of Mathematics, Chair of Nonlinear Optimization, Technische Universität Darmstadt, Dolivostr. 15, 64293 Darmstad, Germany
e-mail: lass@mathematik.tu-darmstadt.de

H. De Gersem
Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, Schlossgartenstr. 8, 64289 Darmstadt, Germany
e-mail: degersem@temf.tu-darmstadt.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_77

In this paper the influence of dynamic eccentricity on the torque and its total harmonic distortion will be studied. In [9] an overview and comparison of different methods for torque calculation can be found. The most used method relies on the Maxwell stress tensor. This method, however, is very sensitive to the applied space discretization and to the location in the airgap where the quantities are evaluated. As a consequence big numerical deviations are easily introduced [7, 14]. Another way to calculate the torque is to use the energy balance method, which relies on the principle of conservation of energy (see e.g. [13]).

The introduction of uncertainties in the PMSM model is the first step in a bigger procedure in which the magnets of the machine are subjected to robust optimization. This can be performed by analyzing the worst-case scenario or by taking into account the standard deviation. In both cases, an accurate and reliable representations of rotor eccentricity is needed. Moreover, a fast simulation technique for calculating the according statistic moment of the harmonics of the torque is indispensable.

The paper is structured as follows: in Sect. 2 the PMSM is introduced. It is explained how the torque is calculated and how the eccentricity is modeled. In Sect. 3 the results are presented and discussed and finally conclusions are drawn.

2 Problem Description

The PMSM has six slots per pole and a double layered winding scheme with two slots per pole per phase (Fig. 1a). The machine has length $\ell_z = 10\,\text{mm}$. The machine is constructed of laminated steel with a relative permeability $\mu_r = 500$. Let Ω_s

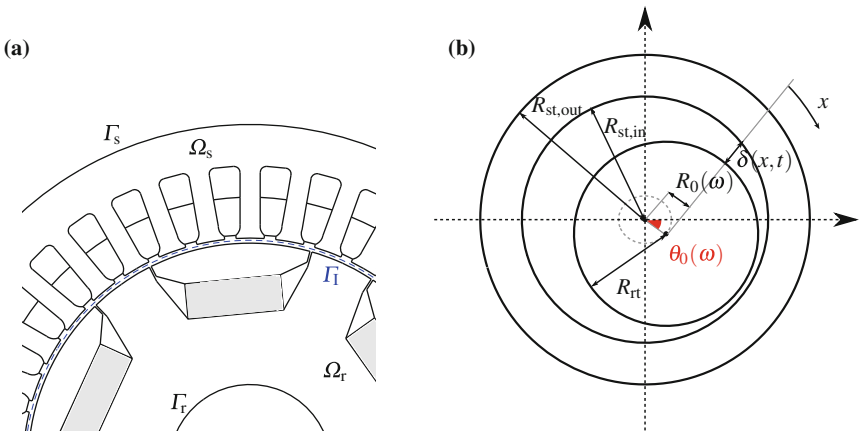


Fig. 1 Crosssectional view of the PMSM submitted to uncertainty quantification. (**a**) Detailed view of the PMSM. The *blue dashed line* indicates the location of the interface. The magnets are marked in *gray*. (**b**) Schematic representation of the full machine with rotor eccentricity (figure adapted from [1])

and Ω_r depict the stator and rotor domain respectively. Define $\Gamma_\mathrm{I} = \partial\Omega_\mathrm{s} \cap \partial\Omega_\mathrm{r}$, $\Gamma_\mathrm{s} = \partial\Omega_\mathrm{s} \setminus \Gamma_\mathrm{I}$ and $\Gamma_\mathrm{r} = \partial\Omega_\mathrm{r} \setminus \Gamma_\mathrm{I}$. To calculate the torque of the machine the magnetostatic approximation of the Maxwell's equations has to be solved for both domains. This implies that the eddy and displacement currents are neglected and one obtains the semi-elliptic partial differential equations

$$\nabla \times (\nu \nabla \times \mathbf{A}_\mathrm{s}(t)) = \mathbf{J}_\mathrm{src}(t), \quad \text{on } \Omega_\mathrm{s} \tag{1a}$$

$$\nabla \times (\nu \nabla \times \mathbf{A}_\mathrm{r}(t)) = -\nabla \times (\nu \mathbf{B}_\mathrm{rem}), \quad \text{on } \Omega_\mathrm{r} \tag{1b}$$

with Dirichlet boundary conditions $\mathbf{A}_\mathrm{s}|_{\Gamma_\mathrm{s}} = 0$ and $\mathbf{A}_\mathrm{r}|_{\Gamma_\mathrm{r}} = 0$. At the interface it holds that

$$\nu \mathbf{n} \times \mathbf{A}_\mathrm{s}|_{\Gamma_\mathrm{I}}(\mathbf{Y}(t)) - \nu \mathbf{n} \times \mathbf{A}_\mathrm{r}|_{\Gamma_\mathrm{I}}(\mathbf{Y}(t)) = 0, \tag{1c}$$

where $\mathbf{Y}(t)$ describes how the interfaces are connected at every time step t. The reluctivity is depicted by a scalar ν since only linear isotropic materials are considered. $\mathbf{A}(t)$ is the magnetic vector potential, $\mathbf{J}_\mathrm{src}(t)$ represents the imposed source current density, which is related to the applied currents in the coils, and $\mathbf{B}_\mathrm{rem}$ the remanence of the permanent magnets. The applied current density is aligned with the z-direction, whereas the remanence is in the xy-plane. It is generally accepted that machines are adequately modeled in 2D, meaning that the magnetic field has no z-component: $\mathbf{B}(t) = (B_x(t), B_y(t), 0)$. Since $\mathbf{B}(t) = \nabla \times \mathbf{A}(t)$, one can write $\mathbf{A}(t) = (0, 0, A_z(t))$ and also $\mathbf{J}$ has only a z-component. Discretizing $\mathbf{A}_\mathrm{s}$ by linear edge shape functions $\mathbf{w}_i(x, y)$ one approximates

$$\mathbf{A}_\mathrm{s} \approx \sum_{i=1}^{N} a_i^{(\mathrm{s})} \mathbf{w}_i(x, y) = \sum_{i=1}^{N} a_i^{(\mathrm{s})} \frac{N_i(x, y)}{\ell_z} \mathbf{e}_z, \tag{2}$$

where $N_i(x, y)$ depicts the nodal finite elements which are associated with the triangulation of the machine's cross-section and $\mathbf{e}_z$ is the unit vector in z-direction. Using the Ritz-Galerkin approach one finds the discretized form

$$\mathbf{K}_\nu^{(\mathrm{s})} \mathbf{a}^{(\mathrm{s})}(t) + \mathbf{j}_\mathrm{int}^{(\mathrm{s})} = \mathbf{j}_\mathrm{src}(t) \quad \text{and} \quad \mathbf{K}_\nu^{(\mathrm{r})} \mathbf{a}^{(\mathrm{r})}(t) + \mathbf{j}_\mathrm{int}^{(\mathrm{r})} = \mathbf{j}_\mathrm{pm}, \tag{3}$$

where $\mathbf{K}_\nu^{(\mathrm{s})}$ and $\mathbf{K}_\nu^{(\mathrm{r})}$ are the finite element system matrices for the stator and rotor, respectively, $\mathbf{a}^{(\mathrm{s})}$ and $\mathbf{a}^{(\mathrm{r})}$ depict the degrees of freedom (DoFs) and $\mathbf{j}_\mathrm{src}$ ($\mathbf{j}_\mathrm{pm}$) are the discretized versions of the current densities (permanent magnets). $\mathbf{j}_\mathrm{int}^{(\mathrm{s})}$ and $\mathbf{j}_\mathrm{int}^{(\mathrm{r})}$ are the discretized versions of the conditions at Γ_I.

A time-stepping technique based on [8] is implemented for the rotation of the machine. In the airgap a contour is defined which splits the full domain in two parts: a fixed outer domain connected to the stator Ω_s and an inner domain connected to the rotor Ω_r, where the mesh will be rotated. At the contour the nodes are distributed equidistantly. The time step is chosen so that the inner domain is rotated by exactly three nodes between successive time steps. The connections between the nodes on

the contour and the ones in the airgap are updated at every time step. One can write $\mathbf{j}_{\mathrm{int}}^{(s)} = \mathbf{P}_s\lambda$ and $\mathbf{j}_{\mathrm{int}}^{(r)} = \mathbf{P}_r(t)\lambda$ with λ the Lagrange multiplier and projectors $\mathbf{P}_{\mathrm{r}}$ and $\mathbf{P}_{\mathrm{s}} \in \{-1, 0, 1\}^N$. This yields

$$\underbrace{\begin{bmatrix} \mathbf{K}_\nu^{(\mathrm{s})} & 0 & \mathbf{P}_{\mathrm{s}} \\ 0 & \mathbf{K}_\nu^{(\mathrm{r})} & \mathbf{P}_{\mathrm{r}}(t) \\ \mathbf{P}_{\mathrm{s}}{}^{\top} & \mathbf{P}_{\mathrm{r}}{}^{\top}(t) & 0 \end{bmatrix}}_{\mathbf{K}_\nu(t)} \begin{bmatrix} \mathbf{a}^{(\mathrm{s})} \\ \mathbf{a}^{(\mathrm{r})} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{j}_{\mathrm{src}}(t) \\ \mathbf{j}_{\mathrm{pm}} \\ 0 \end{bmatrix} \tag{4}$$

with $\mathbf{K}_\nu(t)$ the finite element system matrix [10]. Solving for $\mathbf{a}^{(\mathrm{s})}$ and $\mathbf{a}^{(\mathrm{r})}$ enables the calculation of the torque by using an energy balance method. The change in the stored magnetic energy can be written as

$$\frac{\mathrm{d}W_{\mathrm{mag}}(t)}{\mathrm{d}t} = P_{\mathrm{e}}(t) - P_{\mathrm{l}}(t) - P_{\mathrm{m}}(t), \tag{5}$$

with P_{e} the electrical energy, the mechanical energy P_{m} and the losses P_{l}. To calculate the torque a time-averaging approach is used, i.e. integrating Eq. (5) over one period T. Due to the conservation of energy $W_{\mathrm{mag}}(t) = W_{\mathrm{mag}}(t+T)$ and one finds that $\bar{P}_{\mathrm{m}} = \bar{P}_{\mathrm{e}} - \bar{P}_{\mathrm{l}}$. The total electrical energy is given by

$$\bar{P}_{\mathrm{e}} = \frac{1}{T}\int_0^T P_{\mathrm{e}}(t)\,\mathrm{d}t = \frac{1}{T}\int_0^T \mathbf{u}_{\mathrm{str}}^{\top}(t)\mathbf{i}_{\mathrm{str}}(t)\,\mathrm{d}t, \tag{6}$$

where $\mathbf{i}_{\mathrm{str}}$ depicts the current in the coils, such that $\mathbf{j}_{\mathrm{src}} = \mathbf{X}_{\mathrm{str}}\mathbf{i}_{\mathrm{str}}$, with $\mathbf{X}_{\mathrm{str}}$ the winding function [12]. The voltages of the stranded conductors $\mathbf{u}_{\mathrm{str}}$ are calculated from the solutions of Eq. (4) by

$$\mathbf{u}_{\mathrm{str}}(t) = \mathbf{R}_{\mathrm{str}}\mathbf{i}_{\mathrm{str}}(t) + \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{X}_{\mathrm{str}}^{\top}\mathbf{a}(t)), \tag{7}$$

with $\mathbf{R}_{\mathrm{str}}$, the DC resistance of coils. The losses are given by

$$\bar{P}_{\mathrm{l}} = \frac{1}{T}\int_0^T P_{\mathrm{l}}(t)\,\mathrm{d}t = \frac{1}{T}\int_0^T \mathbf{i}_{\mathrm{str}}^{\top}(t)\mathbf{R}_{\mathrm{str}}\mathbf{i}_{\mathrm{str}}(t)\,\mathrm{d}t. \tag{8}$$

For the time-averaged mechanical energy one finds

$$\bar{P}_{\mathrm{m}} = \frac{1}{T}\int_0^T P_{\mathrm{m}}(t)\,\mathrm{d}t = \frac{1}{T}\int_0^T \left(P_{\mathrm{e}}(t) - P_{\mathrm{l}}(t)\right)\,\mathrm{d}t \tag{9}$$

leading to time-averaged torque

$$\bar{\tau}_0 = \frac{1}{T}\int_0^T \tau_{\mathrm{m}}(t)\,\mathrm{d}t := \frac{1}{T\omega_{\mathrm{m}}}\int_0^T P_{\mathrm{m}}(t)\,\mathrm{d}t. \tag{10}$$

The mechanical angular frequency is $\omega_\mathrm{m} = \omega_\mathrm{e}/N_p$, with ω_e the electric angular frequency and $N_p = 6$ the pole pair number. To calculate the higher harmonics of the torque, Eq. (5) is solved, meaning

$$\tau_\mathrm{m}(t) = \frac{1}{\omega_\mathrm{m}} \left(P_\mathrm{e}(t) - P_\mathrm{l}(t) - \frac{\mathrm{d}W_\mathrm{mag}(t)}{\mathrm{d}t} \right), \tag{11}$$

with $W_\mathrm{mag} = 1/2\mathbf{a}^\top \mathbf{K}_\nu \mathbf{a}$, where $\mathbf{a}$ depicts all DoFs. Taking a Fourier transform of τ_m one can then define the total harmonic distortion (THD) as

$$\mathrm{THD} = \frac{\sum_{i=1}^{n} \tau_i}{\tau_0}, \tag{12}$$

where τ_i depicts the harmonics of the torque.

The eccentricity of the rotor is described by two parameters: $R_0(\omega)$, which defines the initial displacement from the central position, and $\theta_0(\omega)$, which depicts the direction of its displacement (see Fig. 1b). The stochastic nature of a quantity is indicated by ω. It is assumed that $R_0, \theta_0 \in L^2(\Theta, \Sigma, P)$ are independent random variables from the probability space (Θ, Σ, P) and that R_0 is normal distributed and θ_0 uniformly distributed, $R_0 \sim \mathcal{N}(0, \sigma_{R_0}^2)$ and $\theta_0 \sim \mathcal{U}(0, \pi)$, with $\sigma_{R_0} = 0.4/3\,\mathrm{mm}$, which corresponds to an eccentricity of 13%.

Due to the chosen approach for modeling the rotation, the eccentricity is implicitly considered according to the dynamic model of [6], meaning that the airgap width $\delta(x, t)$ (see Fig. 1b) can be expressed as a function of the arc x, such that

$$\delta(x, t) = \delta_\mathrm{m} \left[1 - \varepsilon_\mathrm{m}(\omega) \cos(\alpha(x, t, \omega)) \right], \tag{13}$$

with the eccentricity $\varepsilon_\mathrm{m}(\omega) = R_0(\omega)/\delta_\mathrm{m} = R_0(\omega)/(R_\mathrm{st,in} - R_\mathrm{rt})$, where δ_m is the mean mechanical airgap width. The angle α is given by

$$\alpha(x, t, \omega) = x - \omega_\mathrm{m} t - \theta_0(\omega). \tag{14}$$

Figure 2 illustrates the influence of the eccentricity on the spectrum of the torque. In Fig. 2a the spectra for a nominal machine and for a machine with $\varepsilon_\mathrm{m} = 50\%$ eccentricity are shown. The 36th harmonic, which is related to the cogging torque, is increased due to the eccentricity. To study the full effect of eccentricity on the spectrum, the THD is calculated. One sees in Fig. 2b that this quantity is increasing with a higher eccentricity. This implies that an eccentric machine suffers from increased undesired effects, e.g. [5].

Let $Z(\omega) = f(R_0(\omega), \theta_0(\omega))$ be a discrete random variable then the expectation value and the variance can be approximated by

$$\mu_Z = \mathbb{E}[Z] \approx \sum_{j=1}^{N} w_j f_j \quad \text{and} \quad \sigma_Z^2 = \mathbb{V}[Z] \approx \sum_{j=1}^{N} w_j \left(f_j - \mu_z \right)^2, \tag{15}$$

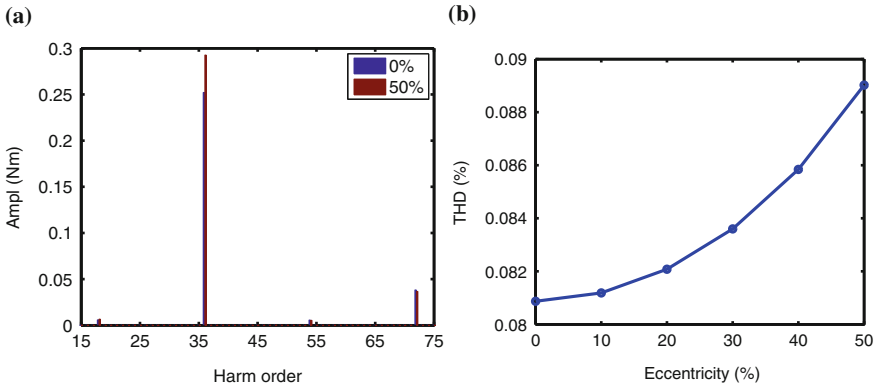


Fig. 2 Study of the influence of the eccentricity on the torque spectrum. (**a**) Part of the spectrum for a nominal machine and for a machine with $\varepsilon = 50\%$. (**b**) The total harmonic distortion as a function of the rotor eccentricity

respectively, where $f_j = f(R_0^{(j)}, \theta_0^{(j)})$ denotes a random outcome. The choice of weights w_j and samples f_j depends on the method. In the Monte Carlo (MC) approach the samples are chosen randomly and have an equal weight $1/N$. For collocation based on generalized Polynomial Chaos (gPC) they are determined according to quadrature rules. The gPC bases for normal and uniform distributions are constructed from Hermite polynomials and Legendre polynomials, respectively. In this paper we use both methods for the uncertainty quantification. With aid of the MC sampling a variance based sensitivity analysis is performed based on [11]. However, it is known, that for a low number of random variables gPC converges faster [15] than Monte Carlo (see e.g. [4]). To avoid numerical noise, remeshing for every sample j is avoided. Instead, the original mesh is mapped to a new one as in [2].

3 Results

The nominal machine, i.e. $R_0 = 0\,\text{mm}$ and $\theta_0 = 0$ has a torque $\tau_0 = 4.060\,\text{Nm}$ and a THD $= 0.809\%$. The UQ was performed by performed by using gPC with 5×5 tensor product grid and by using $N_{\text{MC}} = 3200$ MC samples for the uncertain input parameters. For μ_τ a value of 4.066 Nm is retrieved by both approaches. The found standard deviation is $8.1 \cdot 10^{-3}\,\text{Nm}$. For the THD one finds and $\sigma_{\text{THD}} = 7.8 \cdot 10^{-4}\%$. The error ϵ_{MC} on μ for MC is estimated by $\sigma_Z/\sqrt{N_{\text{MC}}}$. For the $\sigma_Z/\sqrt{N_{\text{MC}}}$. For the torque one finds $\epsilon_{\text{MC}} = 1.5 \cdot 10^{-4}\,\text{Nm}$ and for the THD $\epsilon_{\text{MC}} = 0.14 \cdot 10^{-4}\%$. The sensitivity of $R_0 \approx 1$, where the sensitivity of $\theta_0 < 10^{-6}$.

4 Conclusions

Uncertainty quantification was performed to determine the effects of rotor eccentricity on the torque and on the total harmonic distorting on the torque. It is found that the displacement, rather than the direction of the displacement influences these quantities the most.

Acknowledgements This work is supported by the German BMBF in the context of the SIMUROM project (grant number 05M2013), by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at TU Darmstadt.

References

1. Belmans, R., Vandenput, A., Geysen, W.: Calculation of the flux density and the unbalanced pull in two pole induction machines. Electr. Eng. **70**, 151–161 (1987)
2. Bontinck, Z., De Gersem, H., Schöps, S.: Response surface models for the uncertainty quantification of eccentric permanent magnet synchronous machines. IEEE Trans. Magn. **52**, (2016)
3. Coenen, I., Herranz Garcia, M., Hameyer, K.: Influence and evaluation of non-ideal manufacturing process on the cogging torque of a permanent magnet excited synchronous machine. COMPEL **30**, 876–884 (2011)
4. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
5. Dorrell, D.G., Chindurza, I., Cossar, C.: Effects of rotor eccentricity on torque in switched reluctance machines. IEEE Trans. Magn. **41**, 3961–3963 (2005)
6. Frohne, H.: Über den einseitigen magnetischen Zug in Drehfeldmaschinen. Electr. Eng. **51**, 300–308 (1968)
7. Mizia, J., Adamiak, K., Eastham, A.R., Dawson, G.E.: Finite element force calculation: comparison of methods for electric machines. IEEE Trans. Magn. **24**, 447–450 (1988)
8. Preston, T., Reece, A., Sangha, P.: Induction motor analysis by time-stepping techniques. IEEE Trans. Magn. **24**, 471–474 (1988)
9. Sadowski, N., Lefevre, Y., Lajoie-Mazenc, M., Cros, J.: Finite element torque calculation in electrical machines while considering the movement. IEEE Trans. Magn. **28**, 1410–1413 (1992)
10. Salon, S.J.: Finite Element Analysis of Electrical Machines. Kluwer, Boston (1995)
11. Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S.: Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity check. Comput. Phys. Commun. **181**, 259–270 (2010)
12. Schöps, S., De Gersem, H., Weiland, T.: Winding functions in transient magnetoquasistatic field-circuit coupled simulations. COMPEL **32**, 2063–2083 (2013)
13. Silwal, B., Rasilo, P., Perkkiö, L., Oksman, M., Hannukainen, A., Eirola, T., Arkkio, A.: Computation of torque of an electrical machine with different types of finite element mesh in the air gap. IEEE Trans. Magn. **50**, 1–9 (2014)
14. Tärnhuvud, T., Reichert, K.: Accuracy problems of force and torque calculations in FE-systems. IEEE Trans. Magn. **24**, 443–446 (1988)
15. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)

Minisymposium: The Treatment of Singularities and Defects in Industrial Applications

María Aguareles and Marco Antonio Fontelos

Description

Very often industries are interested in removing, and therefore understanding, defects and singularities arising in materials and fluids. In other instances, it is the focusing inherent to singularities and their potential use to manufacture small things what makes them interesting for industrial purposes. This minisymposium gathered researchers who deal with singularities and defects that appear in industrial applications like, for instance, in problems of electro-wetting, superconductivity or dislocations. Such defects are usually undesirable and it becomes crucial to control its origin and evolution. In this sense, many of the topics that were covered in these sessions were at the latest cutting edge. Also, the topic itself was transversal since its industrial motivation ranges from so different fields like fluid dynamics to the allocation of dislocations or superconductivity. Finally, the mathematics involved are usually far from trivial, involving strong nonlinear effects and multiple length and time scales.

The talks included in the minisymposium were the following:

- Michael Dallaston. Imperial College (UK). *On thin-film rupture with a general disjoining pressure.*
- Lia Bronsard. McMaster University (Canada). *On the mathematical study of defects in liquid crystals.*

M. Aguareles (✉)
IMAE, Universitat de Girona, Girona, Spain
e-mail: maria.aguareles@udg.edu

M.A. Fontelos
ICMAT, Consejo Superior de Investigaciones Científicas, Madrid, Spain
e-mail: marco.fontelos@icmat.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_78

- Hyeonjeong Kim. University of Heidelberg (Germany). *Capillary oscillations at the surface of a fluid in various geometries.*
- María Garzón. Universidad de Oviedo (Spain). *Computing through singularities in potential flow with applications to electrohydrodynamic problems.*
- David N. Sibley. Department of Mathematical Sciences, Loughborough University, Loughborough (UK). *Uncovering the Nanoscale physics at the moving contact line within a statistical mechanics of fluids framework.*

Computing Through Singularities in Potential Flow with Applications to Electrohydrodynamic Problems

Maria Garzon, James A. Sethian, Len J. Gray, and August Johansson

Abstract Many interesting fluid interface problems, such as wave propagation and breaking, droplet and bubble break-up, electro-jetting, rain drops, etc. can be modeled using the assumption of potential flow. The main challenge, both theoretically and computationally, is due to the presence of singularities in the mathematical models. In all the above mentioned problems, an interface needs to be advanced by a velocity determined by the solution of a surface partial differential equation posed on this moving boundary. By using a level set framework, the two surface equations of the Lagrangian formulation can be implicitly embedded in PDEs posed on one higher dimension fixed domain. The advantage of this approach is that it seamlessly allows breakup or merging of the fluid domain and therefore provide a robust algorithm to compute through these singular events. Numerical results of a solitary wave breaking, the Rayleigh-Taylor instability of a fluid column, droplets and bubbles breaking-up and the electrical droplet distortion and subsequent jet emission can be obtained using this levelset/extended potential model.

1 The Level Set/Extended Potential Flow Model

Let $\Omega_1(t)$ be a fluid domain immersed in an infinite exterior fluid $\Omega_2(t)$, $\Gamma_t = \Gamma(t)$ be the free boundary separating both domains, and Ω_D be a fixed domain that should contain the free boundary for all $t \in [0, T]$. The level set/extended potential flow

M. Garzon (✉)
Department of Mathematics, Universidad de Oviedo, Calvo Sotelo s/n, 33007 Oviedo, Spain
e-mail: maria.garzon.martin@gmail.com

J.A. Sethian
University of California, Berkeley, CA, USA
e-mail: sethian@math.berkeley.edu

L.J. Gray
Bergen Software Services International, Bergen, Norway
e-mail: len@bssi-it.com

A. Johansson
Simula Research Laboratory, Fornebu, Norway
e-mail: august@simula.no

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_79

model, [4, 5], may be then written as:

$$\mathbf{u} = \nabla\phi \ \text{ in } \ \Omega_1(t) \tag{1}$$

$$\Delta\phi = 0 \ \text{ in } \ \Omega_1(t) \tag{2}$$

$$\Psi_t + \mathbf{u}_{\text{ext}} \cdot \nabla\Psi = 0 \text{ in } \Omega_D \tag{3}$$

$$G_t + \mathbf{u}_{\text{ext}} \cdot \nabla G = f_{\text{ext}} \text{ in } \Omega_D. \tag{4}$$

Here, $\phi = \phi(x,t)$ is the velocity potential, $\mathbf{u} = \mathbf{u}(x,t)$ the velocity field, $\Psi = \Psi(x,t)$ the level set function, $G = G(x,t)$ the extended potential function, $f = f(x,t)$ accounts for the surface forces, and the subscript "ext" refers to the extended quantities off the front into Ω_D. This hydrodynamic problem can be coupled with any other exterior problem on $\Omega_2(t)$. In particular, assuming a uniform electric field $\mathbf{E} = \mathbf{E}(x,t)$ in $\Omega_2(t)$, acting in the direction of the z axis and $\mathbf{E} = 0$ in $\Omega_1(t)$ (perfect conductor fluid) then:

$$\mathbf{E} = -\nabla U \ \text{ in } \ \Omega_2(t) \tag{5}$$

$$\Delta U = 0 \ \text{ in } \ \Omega_2(t) \tag{6}$$

$$U = U_0 \ \text{ on } \Gamma_t \tag{7}$$

$$U = -E_\infty z \ \text{ at the far field,} \tag{8}$$

where $U = U(x,t)$ is the electric potential and E_∞ is the electric field intensity.

2 Numerical Approximation

The semidiscretization in time of the model equations is:

$$\mathbf{u}^n = \nabla\phi^n \ \text{ in } \ \Omega_1(t_n) \tag{9}$$

$$\Delta\phi^n(r,z) = 0 \ \text{ in } \ \Omega_1(t_n) \tag{10}$$

$$\frac{\Psi^{n+1} - \Psi^n}{\Delta t} = -\mathbf{u}^n_{\text{ext}} \cdot \nabla\Psi^n \ \text{ in } \ \Omega_D \tag{11}$$

$$\frac{G^{n+1} - G^n}{\Delta t} = -\mathbf{u}^n_{\text{ext}} \cdot \nabla G^n + f^n_{\text{ext}} \ \text{ in } \ \Omega_D, \tag{12}$$

$$\Delta U^n(r,z) = 0 \ \text{ in } \ \Omega_2(t_n) \tag{13}$$

where a first order explicit scheme has been applied. For the space discretization of Eqns. (11) and (12) a first order upwind scheme has been be used. Higher order upwind schemes could also be implemented. The approximation of the Laplace equations (10) and (13) is crucial in this numerical method, as it provides the velocity to advance the free boundary and also the velocity potential evolution

within this front. We have coupled the following solvers for the interior and exterior Laplace equations:

- For 2D and 3D axisymmetric geometries a Galerkin boundary integral solution is established, where the boundary element method with linear elements have been used to approximate the integral equations, see [7, 9].
- For the fully 3D approximation a non conforming Nitsche finite element method has been used together with stabilization techniques of the bilinear forms, as the jump stabilization or the ghost penalty stabilization, see [1, 2, 12].

3 Physical Scenarios

Equations (1)–(8) can be used to model various interesting phenomena. Depending upon the geometrical characteristics of the physical process and the forces acting on the free surface it is possible to obtain numerical results in the following settings:

1. **Wave breaking**:
   A solitary wave propagating over slopping bottoms will breakup following different patterns which depend upon the slope of the bottom, see [10]. To model wave breaking over slopping bottoms only the interior fluid is considered to be dynamically active. This means we have a pure hydrodynamic problem, *i.e.* only Eqs. (1)–(4), and the fluid domain is a 2D vertical section. The relevant forces acting in the free boundary are inertia and gravity. Therefore the driving term f is:
   $$f = -gz + \frac{1}{2}|\mathbf{u}|^2,$$
   where z is the free surface vertical coordinate and g the gravity acceleration.
2. **The Rayleigh-Taylor instability**:
   When a sinusoidal perturbation is induced in an infinite liquid column it grows in time, and eventually, if the wave number is less than 1, the liquid will break up giving rise to a cascade of drop formation and re-connections, see for example [3]. For this problem we developed a 3D axi-symmetric approximation, only the interior fluid is active. In this case, as gravity forces are negligible compared to surface tension forces we have:
   $$f = \frac{1}{2}|\mathbf{u}|^2 - \frac{\gamma}{\rho}\kappa,$$
   where γ is the surface tension coefficient, ρ the fluid density and κ is twice the mean curvature of the free surface.
3. **Two fluid system**:
   Pinch-off patterns of water droplets and air bubbles are perfectly distinguishable. Water droplets surface overturns before pinching whereas air bubbles exhibit a symmetric cone shape before breaking up. The power law behaviors near the singular times are also distinct. To simulate water droplets in air or air bubbles in

water we considered a two inviscid fluid system. An interior fluid of density ρ_1 surrounded by an exterior fluid of density ρ_2. Let be $D = \frac{\rho_2}{\rho_1}$ the ratio between the fluid densities. Here we have to solve two Laplace equations in both domains to get the interior and exterior potentials, ϕ_1 and ϕ_2 respectively. This is done through a boundary integral formulation. By setting $\phi_D = \phi_1 - D\phi_2$ it is possible to write a material derivative for ϕ_D on Γ_t:

$$\frac{\partial \phi_D}{\partial t} + \mathbf{u}_1 \cdot \nabla \phi_D = f \quad \text{on} \quad \Gamma_t(\mathbf{s})$$

being the forcing term in this case:

$$f = \mathbf{u}_1 \cdot (\frac{1}{2}\mathbf{u}_1 - D\mathbf{u}_2) + \frac{D}{2}|\mathbf{u_2}|^2 - \frac{\gamma}{\rho_2}\kappa$$

The only parameter left in the non-dimensional model is the fluid density ratio $D = \frac{\rho_2}{\rho_1}$, and simulations of the breaking up transition patterns from air bubbles to water droplets have been computed. All the details regarding the mathematical formulation can be found in [6].

4. **Electrical droplet distortion and burst**:
A droplet subject to a uniform electric field in the direction of the symmetry axis will oscillate with a frequency that depends on the electric field intensity E_∞, see [11]. Above a certain critical value the droplet will destabilize and filament discharge at the droplet end points will take place. In this case the full electro-hydrodynamic model, Eqs. (1)–(8), has to be used and the electrical stresses on the free surface have to be included in the forcing term. Therefore:

$$f = \frac{1}{2}|\mathbf{u}|^2 - \frac{\gamma}{\rho}\kappa + \frac{\epsilon}{2\rho}E_n^2.$$

Here ϵ is the electric permittivity of the exterior dielectric fluid and E_n is the gradient of the electric potential in the normal direction to Γ_t.

4 Numerical Results

As mentioned in the above section, several physical scenarios can be simulated using the assumptions and the numerical method summarized here. In the case of pure hydrodynamic problems, Eqs. (1)–(4), results for the wave breaking phenomena in a 2D geometry have been presented in [4], where splitting of the fluid domain was not considered.

The first simulation involving computations through singular events was presented in [5], where the pinch-off of an infinite fluid jet and subsequent cascade of drop formation was reproduced in a seamless 3D axi-symmetric computation. In Fig. 1 we present the comparison of the satellite break up simulation with laboratory photographs.

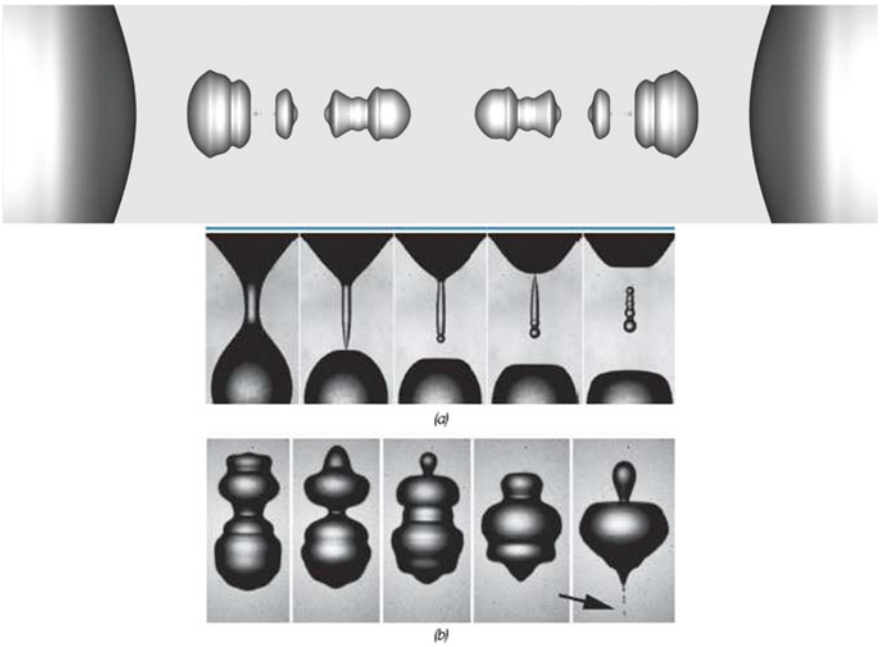


Fig. 1 Satellite drop breaking up, Computed profiles (*top*) and Laboratory photographs, [13] (**a**) Satellite formation during the pinch-off of a drop from a 5 mm nozzle. (**b**) Sequence shows the ejection of 30 mm micro-droplets which emerge at 10 m/s

The interaction of two inviscid fluids of different densities was studied in [6]. The only parameter in the non-dimensional model is the fluid density ratio and simulations of the breaking up transition patterns from air bubbles to water droplets have been computed.

When electrical forces acting on the free surface are also considered, Eqs. (1)–(8), the flow gets even more interesting: a charged water droplet will elongate until Taylor cones are formed, from which fine filaments will be ejected from both droplet tips. As soon as the droplet losses enough charge, it will recoil and oscillate back to equilibrium. In Fig. 2 we show also a comparison between computed profiles on top and Laboratory experiments on bottom at corresponding times. See details in [8].

Finally we also include some numerical results corresponding to a complete 3D calculation of a thin fluid film, which is destabilized by applying a potential difference between a bottom and top electrodes. The top electrode has a protruding parallelepiped that will cause the formation of a jet and eventually the so called tip streaming. In Fig. 3 we show, for example, the iso-surfaces of the electric potential in Ω_2 and Fig. 4 a zoom of the tip streaming (droplet detachment) is depicted. All the mathematical details and numerical examples performed with this fully 3-Dimensional algorithm can be found in [12].

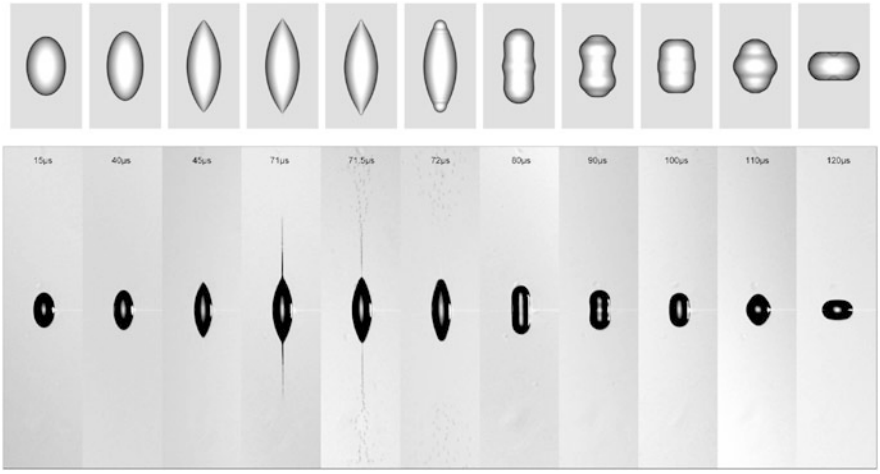


Fig. 2 Laboratory snapshots at indicated times of the evolution of a surface charged super-cooled water droplet, Reprinted figure with permission from E. Giglio, D. Duft and T. Leisner, Phys. Rev. E, 77, 036319 (2008). Copyright(2008) by the American Physical Society (*bottom*); and computed profiles at times 80, 101.2, 108.1, 108.5, 109.8, 112.1, 124.2, 133.4, 138, 142, 154.1 μs (*top*)

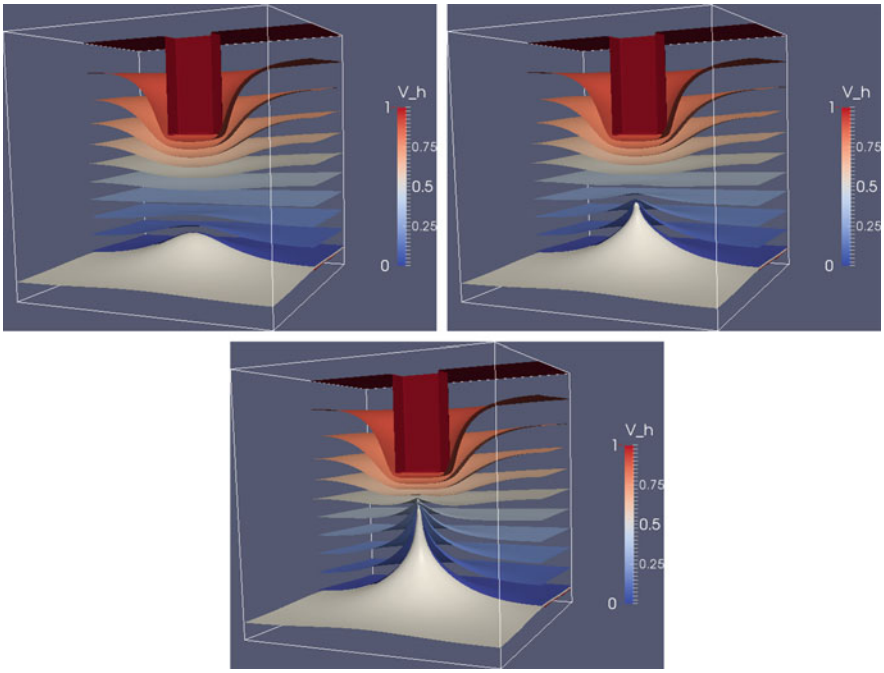


Fig. 3 Tip streaming: Free surface in *white*, and iso-surfaces of the electric potential at $t = 7.25$, $t = 7.95$ and $t = 8.5625$

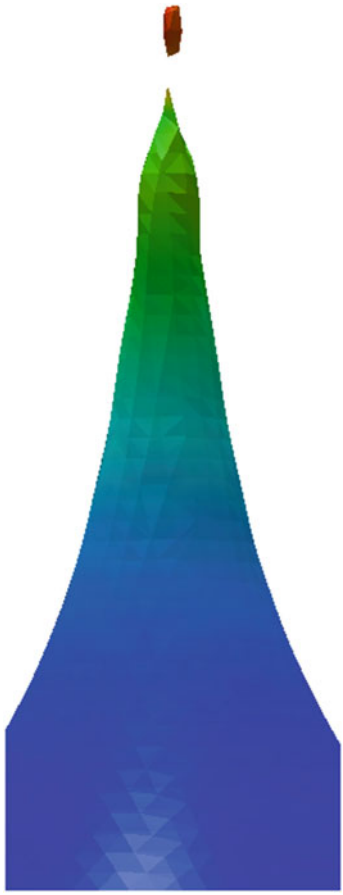


Fig. 4 Tip streaming: Zoom of the tip at $t = 8.5625$. *Colors* represent the velocity field

Acknowledgements This work was supported by the U.S. Department of Energy, under contract Number DE-AC02-05CH11231, the Spanish Ministry of Science and Innovation, Project Number MTM2013-43671-P. The third author was also supported by the Research Council of Norway through a Centers of Excellence grant to the Center for Biomedical Computing at Simula Research Laboratory, Project Number 179578, as well as through the FRIPRO program at Simula Research Laboratory, project number 251237.

References

1. Burman, E.: Ghost penalty. C. R. Math. **348**, 1217–120 (2010)
2. Burman, E., Hansbo, P.: Fictitious domain finite element methods using cut elements: II. A stabilized Nitsche method. Appl. Numer. Math. **62**(4), 328–341 (2012)
3. Eggers, J.: Nonlinear dynamics and breakup of free surface flows. Rev. Mod. Phys. **69**(3), 865–929 (1967)

4. Garzon, M., Adalsteinsson, D., Gray, L.J., Sethian, J.A.: A coupled level set-boundary integral method for moving boundary simulations. Interfaces Free Bound. **7**, 277–302 (2005)
5. Garzon, M., Gray, L.J., Sethian, J.A.: Numerical simulation of non-viscous liquid pinch-off using a coupled levelset-boundary integral method. J. Comput. Phys. **228**, 6079–6106 (2009)
6. Garzon, M., Gray, L.J., Sethian, J.A.: Simulation of the droplet-to-bubble transition in a two-fluid system. Phys. Rev. E. **83**, 046318 (2011)
7. Garzon, M., Gray, L.J., Sethian, J.A.: Axisymmetric boundary integral formulation for a two-fluid system. Int. J. Numer. Methods Fluids **69**, 1124–1134 (2012)
8. Garzon, M., Gray, L.J., Sethian, J.A.: Numerical simulations of electrostatically driven jets from nonviscous droplets. Phys. Rev. E. **89**, 033011 (2014)
9. Gray, L.J., Garzon, M., Mantic, V., Graciani, E.: Galerkin boundary integral analysis Iof the axi-symmetric Laplace equation. Int. J. Numer. Methods Eng. **66**, 2014–2034 (2006)
10. Grilli, S.T., Guyenne, P., Dias, F.: A fully nonlinear model for three dimensional overturning waves over arbitrary bottom. Int. J. Numer Methods Fluids **35**, 828–867 (2001)
11. Grimm, R.L., Beauchamp, J.L.: Dynamics of field induced droplet ionization: time-resolved studies of distortion, jetting and progeny formation from charged and neutral methanol droplets exposed to strong electric fields. J. Phys. Chem. B **109**, 8244–8250 (2005)
12. Johansson, A., Garzon, M., Sethian, J.A.: A three-dimensional coupled Nitsche and level set method for electrohydrodynamic potential flows in moving domains. J. Comput. Phys. **309**, 88–111 (2016)
13. Thoroddsen, S.T.: Micro-droplets and micro-bubbles imaging motion at small scales. Nus. Eng. Res. News **22**(1) (2007)

Minisymposium: 8 Years of East African Technomathematics

Matti Heiliö and Matylda Jabłońska-Sabuka

Description

Lappeenranta University of Technology, with the help of funding from the Finnish Ministry of Foreign Affairs, has now for eight consecutive years supported the development of applied and industrial mathematics in the East African region. The collaboration started from North-South-South exchange projects where students from the South would get the chance to visit LUT for study periods from 5 to 9 months. Over the last 3 years our activities have also extended into revision of the Applied Mathematics curricula in the East African universities, as well as establishment of contact between local industries and academia to enhance research collaboration and applicability. Several staff visits of 1–5 weeks and intensive courses have been organised.

This minisymposium gave an overview of the collaboration, the success stories and challenges we have faced on the way. These included:

- the history of how the collaboration was initialized and carried on,
- the inspirations gained from the activities organized in partner African countries (like the training of instructors for Modeling Weeks),
- promotion of mathematical sciences in East African region, i.e. through AIMS Tanzania,
- personal student/teacher point of view on the collaboration,
- and tangible capacity building outputs, like creation of new study programmes.

M. Heiliö (✉) • M. Jabłońska-Sabuka
Department of Computational Engineering, Lappeenranta University of Technology, Lappeenranta, Finland
e-mail: matti.heilio@lut.fi; matylda.jablonska-sabuka@lut.fi

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_80

The talks included in the minisymposium were the following:

- Matti Heiliö, Lappeenranta University of Technology (Finland). *Building Applied Mathematics Knowledge Base in East Africa.*
- Tuomo Kauranne, Lappeenranta University of Technology (Finland). *Win-Win-Win projects: How to run African and European industrial collaboration projects in Technomathematics.*
- Godwin Kakuba, Makerere University (Uganda). *Modelling weeks: the European and Tanzanian experience from a student's/instructor's point of view.*
- Matylda Jabłońska-Sabuka, Lappeenranta University of Technology (Finland). *The training of instructors for Mathematical Modeling Weeks – a new concept born from LUT-East Africa collaboration.*
- Wilson Mahera Charles, African Institute of Mathematical Sciences (Tanzania). *AIMS - Promoting the Mathematical Sciences and their Applications.*

Building Applied Mathematics Knowledge Base in East Africa

Matti Heiliö, Matylda Jabłońska-Sabuka, and Godwin Kakuba

Abstract There is vast demand in Africa for technological development including modernization of higher education. Reforms in industrial processes through engineering skills are pivotal for the environmental concern and goals of sustainable development. Lappeenranta University of Technology has actively contributed to the spread of Industrial Mathematics in East African region over the past decade through development projects financed by the Finnish Ministry of Foreign Affairs. In this article, we summarize these projects and present their achievements. The story of European Consortium for Mathematics in Industry (ECMI) and the accumulated experience over 25 years have been the encouragement and inspiration for our initiatives. They were focused on Applied Mathematics curriculum development in Partner countries, and on organization of ECMI-style practical workshops like modeling weeks. There is obvious demand to broaden the cooperation between Africa and the European applied mathematics community.

1 Background

There is vast demand in Africa for technological development including modernization of higher education. Reforms in industrial processes, communal and regional networks through engineering skills are pivotal for the environmental concern and goals of sustainable development. The repertoire of skills and knowledge that ECMI represents could provide a welcome and needed components in the reform of the rising continent. McKinsey Global Institute (MGI) recent report said: "Africa's economic pulse has quickened, infusing the continent with a new commercial vibrancy. Many of Africa's economies face serious challenges. Yet the continent is among the world's most rapidly growing economic regions." [5]

M. Heiliö • M. Jabłońska-Sabuka (✉)
Department of Computational Engineering, Lappeenranta University of Technology, Lappeenranta, Finland
e-mail: matti.heilio@lut.fi; matylda.jablonska-sabuka@lut.fi

G. Kakuba
Department of Mathematics, Makerere University, Kampala, Uganda
e-mail: godwin.a.kakuba@gmail.com

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_81

Development of science education and teachers training are important factors in building the foundations for industrial activity and welfare production. Africa needs knowledge-based industries, educating administrators for information society, skills needed in information based professions.

2 Development Projects for East African Mathematics

Motivated by these factors and enabled by valuable personal contacts Lappeenranta University of Technology (LUT) initiated collaboration with mathematics departments of a number of African universities a decade ago. We have coordinated two projects involving eight universities in East African region: (1) East Africa Technomathematics 2007–2015 and (2) Mathematics education and working life relevance in East Africa 2013–2015 [2]. The aim was to transfer to African colleagues the ideas and encouragement that ECMI stands for, to promote awareness about real world impact of mathematics. The development projects were funded by the Finnish Ministry of Foreign Affairs and administered by Finnish Center for International Mobility (CIMO).

The East African region is experiencing rapid changes in demographic profiles characterized by growth in population, urbanization, and income. Challenging problems of Africa persist: civil infrastructure, food production and agricultural practices, logistic chains, public health and sanitation, environmental hazards, energy scarcity. The region needs reform in higher education that would alleviate the shortage of scientific and technological experts. A crucial element in upgrading the education for technological and science based professions is to strengthen computational skills, adequate use of data, analyzing systems by computational models. The key to address these challenges and exploiting opportunities is competences in managing data and making valid inferences for resource planning. Science and Technology rely on good background in mathematics education, which plays an important role in the development of any nation.

For such purpose we have initiated in last decade cooperation with a number of universities in the East African region to build capacity in modernizing the applications-oriented mathematics education. A weakness in math education in third world countries is often the missing touch with applications. The educational culture fails to bring forward how to apply mathematics to the benefit to society. The aim of our project is to enrich the higher education culture by illuminating the real world context, the possibilities of interaction with stakeholders in society. The means include revising curricula, updating educational approaches and promoting collaboration between mathematics, industry and working life in general. The activities are meant also to facilitate “within East-Africa” network effects and encourage female students to seek careers in technology and science.

The project plan has included student exchange on Masters/doctoral studies level, staff visits, intensive courses including three modeling weeks, workshops on curriculum questions, university-industry interaction, challenges on teacher

training. Partners are universities in Tanzania, Rwanda, Uganda, Kenya and Ethiopia and three universities from Finland. We wish to continue the mission and to establish, broaden and utilize the aforementioned links and initiated reforms in educational practices. An asset for our work is the now existing North-South-South network. Some curriculum reforms are underway and the well-received pilot-versions of educational initiatives have been: modeling week, scientific computing culture, introduction of some new course modules.

The project has helped to increase Masters of Science (MSc) degree production locally. Also a number of MSc and PhD degrees have been generated as a direct output from the exchange. We have promoted curriculum modifications, Matlab skills, data assimilation methods, weather models, applied statistics, Markov chain Monte Carlo (MCMC) methods. An important feature has been introduction of modeling week concept, a novel idea in the region. Several students, who have graduated as the result of the cooperation, are now employed as staff members of their departments, or gained professional positions in financial or governmental institutions. The latter also proves that the society is ready to think of mathematics graduates not only as future math teachers, but also skilled professionals in various areas.

Start-ups of local innovative enterprises would be crucial and education of science-based professions will supply the crucial workforce to take initiative for indigenous growth of forward looking small and medium enterprises (SMEs). Regarding industrial mathematics education the immediate challenge would be to get mathematics departments linked with other schools so that we would see increasing number of student projects, company placements and internships, summer jobs and MSc thesis projects coming from actual applied industrial areas.

3 Research Training, Inverse Problems Africa (IPA) and African Institute for Mathematical Sciences (AIMS)

One of the Finnish partners in the project is the Sodankylä Geophysical Observatory (SGO) of University of Oulu, who for a long time have collaborated with Bahir Dar University (BDU) in Ethiopia in the field of atmospheric science and analysis of radar data. Student exchange and staff visits have been carried out. The SGO/BDU collaboration addresses key challenges on inverse problems and data science guided by applications arising in near-space and atmospheric remote sensing. The project organized Bahir Dar Winter School in 2013 where series of lectures on applied mathematics and space physics were given and additional sessions on various topics like riometry radar installation were carried out. In addition, we organised a research training school and 2nd Inverse Problems Africa workshop in 2015.

Two MSc theses on weather models have been supervised in collaboration between LUT, University of Dar es Salaam (UDSM) and the Tanzanian Meteorological Agency (TMA). Other research partners include Ifakara Health Institute (IHI) in Tanzania. Several MSc theses and one ongoing PhD on malaria transmission and

mosquito dynamics have resulted. Ifakara is best-known for its malaria research, hosting the most advanced malaria vector research group in Africa [3].

An important recent step in the development of applied mathematics research base in East Africa is the opening of AIMS Tanzania in 2013. This institution will strengthen the formation of mathematics knowledge base in the region. We envision increasing future collaboration with AIMS Tanzania and (UDSM). AIMS is an all-African network of centres for tertiary mathematical education and research, founded in order to facilitate initiatives in education, research and technology [1]. The objective is expressed as "to enable the continent's youth to shape the continent's future through Science Technology Mathematics (STEM) education". AIMS offers a structured Masters in mathematical sciences that could be a potential double-degree partner with some ECMI institutions.

4 Modeling Weeks in East Africa

On European ground, one of the most powerful tools for spreading the concept of industrial mathematics among students and academicians are study groups and modeling weeks. Therefore, we have decided to incept also the modeling week idea onto the African environment.

Three modeling weeks have been organized within our projects. Students from five African countries have participated in the events. As project instructors we have used local staff, PhD students and postdocs (initially the more experienced staff from Finland, and later the local African representatives).

4.1 East African Modeling Week 2014 in Dar es Salaam

The modeling week in 2014 organized in Dar es Salaam was the third one in the course of the overall collaboration. The event hosted the total of 35 students, that is 25 local participants from University of Dar Es Salaam and 10 students from other East African partner universities, that is National University of Rwanda, University of Dodoma, University of Nairobi, Makerere University and Bahir Dar University. Moreover, all the four problems presented and solved during the week were brought by East African instructors. The groups worked on the following projects:

Modeling Groundwater Flow and Rain The students were to consider a long strip of land between two parallel canals of different water levels. Above a certain level, the ground is porous like sand, allowing water to flow very slowly from the canal with the higher level to the other one. Below the porous top layer, the ground is semipermeable like clay, and the water disappears through it at a rate proportional to the local hydrostatic pressure. From above there is a vertical influx of precipitation, i.e., rain. The aim of the students' project was to model the groundwater level in a

situation of steady state for moderate to small amounts of rain when one of the canal water levels is below semipermeable layer. When being presented the problem, the students were not given any mathematical hints which made it a perfect Modeling Week project. At the end of the week they have reached the final solution.

Survival Data Analysis Using R Statistical Software Initially, using health and demographic surveillance data collected by KEMRI-Wellcome Trust, Kilifi-Kenya, interactions between RSV and coronavirus were analyzed for. The result of the analysis was the time to infection with RSV since previous coronavirus infection (for cases) and time to infection with RSV since the start of the study (for controls). The aim of students' work at the modeling weeks was to compare hazard rates in the two groups of survival data. When being presented the problem, the students were not given any mathematical hints which made it a perfect modeling week project. By the end of the week they have identified and studied various statistical relationships within the data set.

Approximation Techniques on One Dimensional Nonlinear Heat Transfer in Singular Fins In this problem, students were asked to model the non-linear heat transfer equation in one dimensional triangular fin and use some approximation techniques to solve this problem. The project was not a typical modeling week task as most of the mathematical background was revealed to the students together with problem definition and the students were asked to improved the existing methodology.

Optimal Control in Prey-Predator Models In this project the students were asked to formulate a mathematical model to control the threatened prey-predator system and then perform numerical simulations to investigate the effect of control strategies for the threats in the system. The weakness of this project from modeling challenge perspective was that the mathematical base of the problem was given to the students right from the start and they were only asked to extend the model and perform numerical simulations, which the students easily attained at the end of the week.

4.2 Training of Instructors

Since the modeling week idea was new to most of our colleagues in East Africa, we decided to extend the classical event by preceding it with a training of instructors. The idea was to familiarize the group instructors with the pedagogical aspects of the modeling week. The following points have been discussed with the future supervisors: (i) the history of the ECMI modeling weeks, (ii) what is a good problem for a modeling week, (iii) the benefits from participation for both students and instructors, (iv) how to lead and mentor the students through the course of the week, keeping in mind that it should be the students who do the actual modeling and solution, and not the instructor, (v) how to break down the problem into subproblems to help the group manage resources and time.

As a result, each of the modeling problems brought by our East African colleagues to the event was pre-discussed so that the instructors would know how to lead the groups. The discussions also lead to improvement of presentations of the heat transfer project and the prey-predator project to make them more suitable from mathematical modeling challenge perspective. Moreover, each of the new instructors was paired with another colleague who already had such an experience from ECMI modeling weeks (whether as a student or an instructor).

In the course of the project we have also realized that training of instructors before modeling weeks would be much in place also during the ECMI events, where often inexperienced doctoral students take the challenge of being group instructors and they rarely know the right way to mentor and "tease" the group towards finding the solution.

4.3 African Instructor Experience Reported by Dr. Godwin Kakuba

"Having done my postgraduate education from highly industrialized western Europe, I had come to appreciate the importance of mathematics in solving real life problems. As a PhD student, I attended several workshops and conferences in applied mathematics, including the ECMI Modeling Week 2008 in Europe and I always wished we could have the same back home in East Africa. In August 2014, I was presented a very unique opportunity, to participate as an instructor at the East African Modeling Week in Dar es Salaam, Tanzania. I realized that, as an instructor, you have to pose a good challenge problem to the students but then also be able to guide the students to draw up their own solutions. I was able to achieve it all thanks to the training of instructors organized before the modeling week.

"I presented a problem of groundwater modeling with rain. The problem was a localization of a similar problem in [4] to a home situation of encroachment on wetlands, where in one particular wetland in Kampala, a highway has been constructed on one side and there is land filling for human settlement on the other side. On both sides canals have been constructed to reduce on the resulting floods especially in the wet seasons. Students were not given any clues on the solution. Several sessions of brainstorming and discussions ensued and it was very interesting that by the end of the week the students had a mathematical model for the problem as well as the solutions. This was testimony that there is potential for Industrial Mathematics in East Africa and what is lacking is more of such opportunities to students and academia in general.

"In addition to the training of instructors, I was also paired with a colleague from Finland as an instructor and he was very resourceful to the whole group. This was a good illustration of team work and collaboration among academic experts, a trait we have to develop more in East Africa. As instructors, we manage to present problems from industry. But, with more of these workshops, we may also have experts from

industry present their problems and thus develop and strengthen the bond between academia and industry.

"East Africa has for some time now enjoyed relative political stability, made progress on economic development and there are clear political efforts to advance science-based industrial growth. The Technomathematics projects in East Africa spearheaded by LUT could not be more timely."

5 Conclusions

The story of ECMI and the accumulated experience over 25 years have been the encouragement and inspiration for our projects. Africa could benefit a lot from broadened and deepened collaboration with ECMI. The challenge in developing real world applied mathematics in Africa is big and diverse. Areas of fruitful collaboration include: (i) Curriculum development, (ii) Awareness of career opportunities in STEM, also among girls, (iii) Promoting academia-industry-society interaction, (iv) Connection to school mathematics/teachers.

We encourage ECMI members to build partnerships to promote these goals. There is obvious demand to broaden the cooperation between Africa and the European applied mathematics community. The skills and knowledge reservoir that ECMI contains could have an impact on the development of technology, applied sciences and public governance. Objectives would be increased skills in modeling, information and communication technology tools and quantitative methods in engineering, agriculture, environmental issues, energy sector to name a few. To influence the skills and aptitude of the next generation of mathematics teachers is another big task.

Acknowledgements The work is supported by Finnish Center for International Mobility (CIMO). ECMI sponsored four participants from Africa to attend ECMI 2016 Conference.

References

1. AIMS global website. http://www.nexteinstein.org, Cited 15 Jan 2017
2. HEI ICI Mathematics Education and Working Life Relevance in East Africa - project website. https://mafy.lut.fi/HEI-ICI, Cited 15 Jan 2017
3. Ifakara Health Institute website. http://www.ihi.or.tz, Cited 15 Jan 2017
4. Mattheij, R.M.M., Rienstra, S.W., Ten Thije Boonkkamp, J.H.M.: Partial Differential Equations: Modelling, Analysis, Computation, pp. 540–546. SIAM, Philadelphia (2005)
5. Roxburgh, C., Dörr, N., Leke, A., Tazi-Riffi, A., van Wamelen, A., Lund, S., Chironga, M., Alatovik, T., Atkins, C., Terfous, N., Zeino-Mahmalat, T., McKinsey Global Institute: Lions on the Move: The Progress and Potential of African Economies, pp. 1–2. McKinsey and Company (2010)

Minisymposium: 10 Years of Portuguese Study Groups with Industry

Manuel Cruz, Pedro Freitas, and João Nuno Tavares

Description

This minisymposium provided an overview of the implementation and evolution of European Study Groups (ESGI) in Portugal, describing our experience and some industrial problems dealt with, the challenges that had to be overcome and examples of successful and less successful stories.

The talks included in the minisymposium are the following:

- J. Orestes Cerdeira. Dep. de Matemática and Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Portugal. *Warehouse storing and collecting of parts: a challenge addressed on the* 65th *European Study Group with Industry.*
- Eliana Costa e Silva. CIICESI, ESTGF, Polytechnic Institute of Porto, Portugal. *Scheduling aircraft engines repair process: a challenge addressed on the* 86th *European Study Group with Industry.*
- A. Ismael F. Vaz. Department of Production and Systems, School of Engineering, University of Minho, Portugal. *A scheduling application to a molding injection machine: a challenge addressed on the* 109th *European Study Group with Industry.*

M. Cruz (✉)
LEMA, Polytechnic of Porto, Porto, Portugal
e-mail: mbc@isep.ipp.pt

P. Freitas
Group of Mathematical Physics, University of Lisbon, Lisbon, Portugal
e-mail: psfreitas@fc.ul.pt

J.N. Tavares
CMUP, University of Porto, Porto, Portugal
e-mail: jntavar@fc.up.pt

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_82

- Ana Moura. LEMA/CMUP, Neoturf, Portugal. *A routing/assignment problem in garden maintenance services: a challenge addressed on the* 86th *European Study Group with Industry.*
- Adérito Araújo. CMUC, Department of Mathematics, University of Coimbra, Portugal. *A characterisation of ESGIs problems.*

The first talk, presented by Jorge Orestes Cerdeira, described a problem submitted by Grohe, a sanitary fittings manufacturer. The original problem, presented at the 65th ESGI, stated that the company would like to reduce the costs resulting from storing and collecting items in the warehouse. Storage and collecting operations are accomplished by forklifts that operate inside the corridors. In this talk, it was shown how to convert the problem of determining optimal corridor tasks into an assignment problem [2], as well as that the problem of assigning forklifts to corridors and settling how each forklift should visit the corridors assigned to it can be formulated as a capacitated vehicle routing problem [12].

The second talk addressed a real world scheduling problem, presented at the 86th ESGI, concerning the repair process of aircraft engines by TAP - Maintenance and Engineering (TAP-ME), which is the maintenance, repair and overhaul organization of TAP Portugal, Portugal's leading airline. The company's Engine Shop Production Planning Department aimed to have a mathematical model for the engines repair process that would determine the optimal sequencing of tasks within the workstations, in order to minimize total weighted tardiness, while assigning relative priorities to different clients. Because of its similarities with the classical jobshop problem [9] the team, here represented by Eliana Costa e Silva, developed a mixed integer programming model for the specific issues of the TAP engine repair process. Computational experiment on a real instance provided by TAP-ME from a regular working week was successful in finding an optimal solution. Results of the model using benchmarking instances available in the literature [8] were presented.

In the last talk of the first day, Ismael Vaz presented in detail the challenge of optimizing the scheduling of a shoes injection moulding machine. The solution proposed for this problem at the 109th ESGI, modeled it as an integer mathematical optimization problem [10, 11]. The model was coded in AMPL [1], allowing state-of-the-art integer programming solvers to be used and tested. The Gurobi [7] solver, available through the NEOS [6] server platform, was used to provide a solution for a particular shoes manufacturing order. The numerical results for two types of optimization problems by considering different sets of constraints were discussed.

The second session of the 10 years of Portuguese Study Groups with Industry minisymposium started with a talk about the challenge posed by Neoturf to the 2012 Portuguese ESGI from both the academic and industrial perspectives. The aim of this challenge was to define a procedure for scheduling and routing efficiently its clients of garden maintenance services, over a wide geographic region. The total distance travelled then weighs heavily on the company costs and the original statement of the problem focused on a reduction of these costs, while satisfying agreements with its clients. In this talk the authors presented a mixed integer linear programming formulation for the problem [3], discussed the limitations

on the size of instances that can be solved to guarantee optimality, presented a modification of the Clarke and Wright heuristic [4] for the vehicle routing with time windows [5, 13], and reported some results obtained with Neoturf data.

This minisymposium ended with a talk presented by Adérito Araújo, overviewing the first 10 years of European Study Groups in Portugal, focusing on what the authors believe to be the main issues arising in this and similar initiatives. After a brief description of the very beginnings of Portuguese ESGIs, which they hope to be of interest for anyone deciding to embark on such an enterprise, the different types of problems posed by industry in the Portuguese series were discussed. Finally, a characterisation of what the authors perceive as the main problems arising in the relationship between the industrial and the academic communities in Portugal was presented. In spite of specific idiosyncrasies, we believe that there is much to be learnt from sharing these and similar experiences, and hope that this was of interest to ESGI organisers in other countries.

References

1. AMPL Optimization LLC: AMPL: A Modeling Language for Mathematical Programming (2012). Available from: http://www.ampl.com. Cited 10 Feb 2017
2. Burkard, R., Dell'Amico, M., Martello, S.: Assignment Problems. Society for Industrial and Applied Mathematics, Philadelphia (2009)
3. Cerdeira, J.O., Cruz, M., Moura, A.: A Routing/Assignment Problem in Garden Maintenance Services. In: Almeida, J.P., Oliveira, J.F., Pinto, A. (eds.) Operational Research, IO 2013 - XVI Congress of APDIO, Bragança, June 3–5, 2013. CIM Series in Mathematical Sciences, vol. 4, pp. 145–155, Springer, Berlin (2015)
4. Clarke, G.,Wright, J.W.: Scheduling of vehicles from a central depot to a number of delivery points. Oper. Res. **169**(4), 568–581 (1964)
5. Desaulniers, G., Madsen, O.B.G., Ropke, S.: Vehicle routing problems with time windows. In: Toth, P., Vigo, D. (eds.) Vehicle Routing: Problems, Methods, and Applications, 2nd edn. MOS-SIAM Series on Optimization, pp. 119–159, SIAM, Philadelphia (2014)
6. Dolan, E.: The NEOS Server 4.0 Administrative Guide. Mathematics and Computer Science Division, Argonne National Laboratory (2001)
7. Gurobi Optimization, Inc.: Gurobi Optimizer Reference Manual, 2015. Available from: http://www.gurobi.com. Cited 10 Feb 2017
8. Hurink, J., Jurisch, B., Thole, M.: Tabu search for the job-shop scheduling problem with multi-purpose machines. Oper. Res. Spektrum **15**(4), 205–215 (1994)
9. Manne, A.: On the job shop scheduling problem. Oper. Res. **8**, 219–223 (1960)
10. Nash, A., Sofer, A.: Linear and Nonlinear Programming. McGraw-Hill, New York (1996)
11. Pinedo, L.M.: Planning and Scheduling in Manufacturing Services. Springer, New York (2005)
12. Ralphs, T.K., Kopman, L., Pulleyblank, W.R., Trotter, L.E.: On the capacitated vehicle routing problem. Math. Program. **94**, 343–359 (2003)
13. Toth, P., Vigo, D.: The Vehicle Routing Problem. SIAM Monographs on Discrete Mathematics and Applications, Philadelphia (2002)

A Scheduling Application to a Molding Injection Machine: A Challenge Addressed on the 109th European Study Group with Industry

Isabel Cristina Lopes, Sofia O. Lopes, Rui M.S. Pereira, Senhorinha Teixeira, and A. Ismael F. Vaz

Abstract This paper addresses a scheduling optimization problem applied to a molding injection machine. This optimization problem was posed as a challenge to the mathematical community present at the 109th European Study Group with Industry (ESGI), held in Portugal in 2015. We propose a mathematical model for the scheduling optimization problem, which was coded in a widely known modeling language. The model is validated through a set of numerical results obtained with a state-of-the-art solver.

1 Introduction

In this paper, we describe in detail the challenge of optimizing a shoes injection moulding machine, the used approach, and we discuss some numerical results with some real word and academic examples. The challenge corresponds to a scheduling optimization problem [6], which was modeled as an integer mathematical optimization problem [2, 5]. The model was coded in AMPL [1], allowing state-of-the-art integer programming solvers to be used and tested. The Gurobi [4] solver, available through the NEOS [3] server platform, was used to provide a solution for a particular shoes manufacturing order. We discuss numerical results for two types of optimization problems by considering different sets of constraints.

I.C. Lopes
Instituto Politécnico do Porto, Porto, Portugal
e-mail: cristinalopes@eu.ipp.pt

S.O. Lopes • R.M.S. Pereira
Center of Mathematics, School of Science, University of Minho, Guimarães, Portugal
e-mail: sofialopes@math.uminho.pt; rmp@math.uminho.pt

S. Teixeira • A.I.F. Vaz (✉)
School of Engineering, University of Minho, Guimarães, Portugal
e-mail: st@dps.uminho.pt; aivaz@dps.uminho.pt

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_83

The proposed challenge, which is known in the literature as a scheduling optimization problem, may be addressed by many operational research techniques, like integer programming and heuristics based on graphs theory. We chose to address this problem by using integer programming, because it is easier to implement and there are plenty of freeware platforms to help us on this task.

2 The Challenge

The challenge was submitted to the 109th European Study Group with Industry (ESGI), which took place from May 11th to May 15th, 2015, at the Department of Production and Systems of the School of Engineering, University of Minho. In this type of events, industries are requested to pose mathematical challenges to a set of experienced mathematical researchers, which dedicate a full working week in providing a solution, or avenues to get a solution, to the posed challenges.

The herein addressed challenge was submitted by a shoes company consisting in the scheduling of a shoe injection moulding machine, where an optimal shoe size scheduling is to be found, subject to some operational constraints. The shoes injection moulding machine is a circular machine that injects PVC or rubber in a mould for making the shoe soles. See Fig. 1 for the machine scheme.

The injection machine has the following characteristics, which should be accounted in the model: the machine has 14 spots to place shoe moulds (meaning

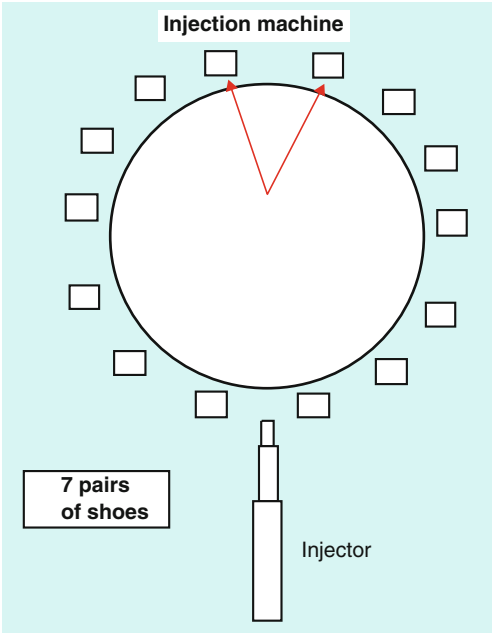


Fig. 1 Scheme of the moulding machine

Table 1 Company provided manufacturing orders and existing moulds

Shoe size	35	36	37	38	39	40	41	42	43	44	45	46	47
Number of moulds	1	1	1	2	2	2	1	1	1	1	1	1	1
Order 1	300	400	1000	1500	1800	1400	1000	800	700	600	300	100	100
Order 2			200	700	400	30	40	100	30				

seven pairs of shoes simultaneous in the machine); each shoe sole injection lasts 20 s; a whole cycle of shoes injection takes $\sim$ 4.67 min ($20 \times 14 = 280$ s); the time needed to change a mould is 30 min (machine is stopped for the changing operation); the time needed to change a shoe model is 4 h (machine is stopped and cleaned); there are 16 moulds available (each mould price is $\sim$3000 euros); stopping the machine costs 0.15 euros per minute.

The injection machine is to be used with the following set of assumptions on the manufacturing orders: orders may include 13 shoe sizes, ranging from size 35 to size 47; manufacturing orders are assumed to have different injection models and are, therefore, treated separately; manufacturing orders are to be delivered as soon as possible, so no deadlines are considered.

Two examples, close to real world manufacturing orders, provided by the company are presented in Table 1. Given a set of manufacturing orders, the main objective is to optimize the use of the shoes injection moulding machine (total manufacturing cost), taking into account the possibility to consider stock operations. Also, there is a limited number of moulds available in the company, e.g. only two moulds are available for the shoe size 39. The number of available moulds is also provided in Table 1.

3 Mathematical Model

This section is used to explain the proposed mathematical model. We start, in the next subsection, to describe the model parameters; Sect. 3.2 introduces the variables considered, which correspond to the scheduling unknowns. Finally, we present the integer linear programming formulation.

3.1 The Model Parameters

Parameters are entities that are fixed in the problem definition, but can be changed in order to fit the model to reality. The considered parameters are: *Cost*—setup cost (cost per minute of stopping the machine); *Stop*—machine mould changing time;

Stock—stock cost per pair of shoes; C_j—the demand of shoes of size j; b_j—number of available moulds for size j ($j = 35, \ldots, 47$).

3.2 The Model Variables

The herein scheduling problem considers time cycles (e.g. the first time cycle corresponds to the first 20 operational seconds). The subscript k identifies the time cycle. The subscript p and j are used to identify the machine position ($p = 1, \ldots, 7$, since 14 pairs of shoes are considered) and the mould size ($j = 35, \ldots, 47$), respectively. The model considers $X_k(j,p)$ binary variables that take the value 1 if a mould of size j is in position p at time cycle k, and binary variables $Y_k(p)$ that identifies if, in cycle k, the mould in the position p changes. Since only integer (in fact binary) variables are considered, the model corresponds to an integer optimization problem.

3.3 The Integer Optimization Model

The herein mathematical model aims at the minimization of the operational costs (costs of changing moulds, shoe models, and stock), subject to operational constraints.

The considered objective function, to be minimized, is defined as

$$\sum_{k,p} Cost.Y_k(p).Stop + Stock.\left(\sum_{j}\left(\sum_{k,p} X_k(j,p) - C_j\right)\right) + \sum_{k,p,j} k.X_k(j,p) \qquad (1)$$

where the first parcel represents moulds changing cost in the middle of the production process, the second parcel represents the stocking cost, and the third parcel is used to minimize the order machine full time (production for high values of k will be penalized).

The first constraint to be considered assures that, for each cycle, the number of moulds used in the machine are less or is equal to 7 (the number of shoe pairs positions available), i.e.

$$\sum_{j}\sum_{p} X_k(j,p) = 7, \quad k = 1, \ldots. \qquad (2)$$

The second constraint assures that the demanded shoe sizes are produced, i.e.

$$\sum_{k}\sum_{p} X_k(j,p) \geq C_j, \quad j = 35, \ldots, 47. \qquad (3)$$

The third constraint ensures the physical existence of the moulds, i.e. the number of moulds in the machine is less or equal than the number of available moulds,

giving

$$\sum_{p} X_k(j,p) \leq b_j, \quad k = 1,\ldots, \quad j = 35,\ldots,47. \tag{4}$$

The next constraint ensures that only one mould fits in each position in each cycle.

$$\sum_{j} X_k(j,p) = 1, \quad k = 1,\ldots, \quad p = 1,\ldots,7. \tag{5}$$

The last constraint is used to set the Y variables to 1 whenever a change in the model occurs. This insures costs, related with changing moulds, are considered in the objective function.

$$Y_k(p) \geq X_k(j,p) - X_{k-1}(j,p), \quad k = 2,\ldots, \quad p = 1,\ldots,7, \quad j = 35,\ldots,47. \tag{6}$$

The mathematical model defined by the objective function (1) subject to constraints (2)–(6) forces all mould machine positions to be used, possibly producing more shoes than demanded, creating stock. Not imposing all mould machine positions to be used can be achieved by changing the equality constraints (2) and (5) to lower or equal inequality constraints.

4 Numerical Results

Numerical results where obtained using the Gurobi [4] solver, which corresponds to a state-of-the-art solver for (among others) integer programming. While Gurobi is a commercial software it can be freely used through the NEOS [3] server.

The parameters considered in the numerical results are: *Cost*=0.15€, *Stop* = 30 min, and *Stock* = 10€, which were provided by the company.

Our first example considers a manufacturing order whose parameters and optimal solution can be seen in Fig. 2 (colors are used for better visibility). In Fig. 2a 'Sizes' are the shoe size, 'Production' is the optimal number of shoes to be produced, 'Demand' is the number of shoes requested in the order, and 'Excess' is the number of shoes to be stocked. The optimal solution was obtained in 18 min of computing time.[1] In the numerical results we assume that every mould machine position is used, allowing stock to be produced.

For a better visibility, we further display in Fig. 2b, a Gantt chart of the first example obtained optimal solution. Each line in the figure corresponds to a machine

[1]Computing time is provided as illustrative, as under NEOS no control is possible about the computer used for the results.

a)

Sizes	Production	Demand	Excess
35	1	1	0
36	6	2	4
37	6	3	3
38	3	1	2
39	6	4	2
40	9	8	1
41	1	1	0
42	0	0	0
43	4	3	1
44	0	0	0
45	0	0	0
46	6	3	3
47	0	0	0

b) Moulds to place in each position in each time cycle

Machine Positions:	1	2	3	4	5	6
1	40	40	40	40	40	40
2	40	40	40	38	38	38
3	37	37	37	37	37	37
4	39	39	39	39	39	39
5	36	36	36	36	36	36
6	46	46	46	46	46	46
7	35	43	43	43	43	41

Time Cycles

Fig. 2 First academic example optimal solution

a)

Sizes	Production	Demand	Excess
35	1	1	0
36	2	2	0
37	3	3	0
38	1	1	0
39	4	4	0
40	8	8	0
41	1	1	0
42	0	0	0
43	3	3	0
44	0	0	0
45	0	0	0
46	3	3	0
47	0	0	0

b) Moulds to place in each position in each time cycle

Machine Positions	1	2	3	4	5	6
1	46	46	46			
2	43	43	43	38		
3	40	40	40	40		
4	41	37	37	37		
5	35	36	36			
6	40	40	40	40		
7	39	39	39	39		

Time Cycles

Fig. 3 Second academic example optimal solution

mould position, where the shoe size is represented, for each time cycle, in a different color. From the figure we can observe that the machine is fully loaded, being some shoe sizes stocked.

The second academic example corresponds to the first example, but not imposing shoes to be produced for stock. Figure 3 provides the obtained numerical results, where the columns have the same meaning as in Fig. 2. When no stock is imposed no excess shoes are produced. Seven seconds of computing time were used to obtain the solution, being observed a drastic decrease in computing time from the first example due to the change in the constraints sign (from equality to inequality—providing a stronger model).

The model was also applied to the manufacturing orders provided by the company (see Table 1). The optimal solution for the first order, where stock is not allowed, was obtained in 31 min of computing time and is depicted in Fig. 4.

The second real order optimal solution is depicted in Fig. 5. Stock is allowed and the solution was obtained in 30 s of computing time.

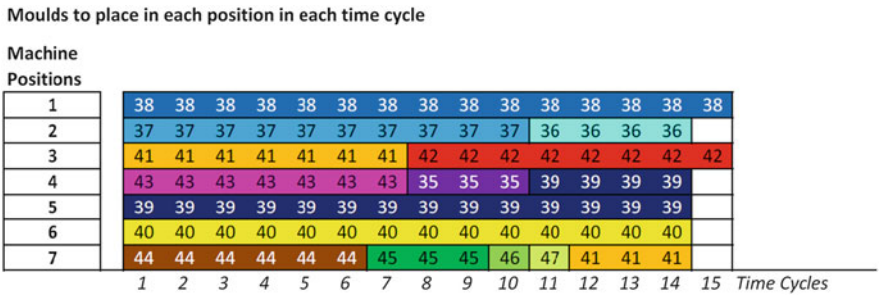


Fig. 4 First real world order optimal solution

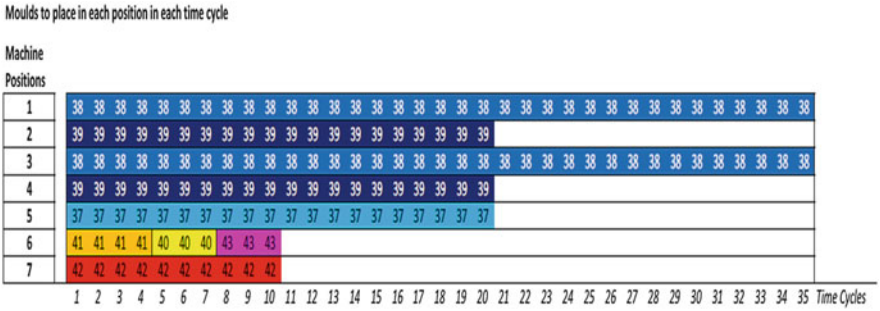


Fig. 5 Second real world order optimal solution

5 Conclusions

A challenge proposed by a shoe company was addressed, which consisted in the scheduling of an injection mould machine. The scheduling problem was modeled as an integer optimization problem and a state-of-the-art optimization solver was used to obtain the numerical results.

The proposed approach has shown to be suitable for the proposed challenge, since it provided an optimal solution to two academic and two real world provided manufacturing orders.

By providing a mathematical modeling of the optimization problem, users may not only obtain information about the optimal scheduling, but also about investments in equipment (e.g. increasing the number of available moulds), if a sensitivity analysis is carried out.

References

1. AMPL Optimization LLC: AMPL: a modeling language for mathematical programming (2012). http://www.ampl.com
2. Conforti, M., Cornuejols, G., Zambelli, G.: Integer Programming. Springer International Publishing, Cham (2014)

3. Dolan, E.: The neos server 4.0 administrative guide. Tech. Rep. ANL/MCS-TM-250, Argonne National Laboratory, Mathematics and Computer Science Division (2001)
4. Gurobi Optimization, Inc.: Gurobi optimizer reference manual (2015). http://www.gurobi.com
5. Nash, S., Sofer, A.: Linear and Nonlinear Programming. McGraw-Hill, New York (1996)
6. Pinedo, L.: Planning and Scheduling in Manufacturing Services. Springer, New York (2005)

Part IV
Contributed Talks

Numerical Simulation of a Li-Ion Cell Using a Thermoelectrochemical Model Including Degradation

David Aller Giráldez, M. Teresa Cao-Rial, Pedro Fontán Muiños, and Jerónimo Rodríguez

Abstract A thermoelectrochemical model for the simulation of Li-ion cells is presented and numerically solved herein. The model, based on Newman's (Newman and Thomas-Alyea, Electrochemical Systems, Wiley, New York, 2004) concentrated solution theory, covers electrochemical, thermal and aging effects. The degradation mechanism considered is the growth, due to a solvent decomposition reaction, of the Solid Electrolyte Interface layer (SEI) in the graphite electrode. Model homogenization is phenomenological but detailed particle-scale models are considered for the diffusion of species within active particles and SEI. The one dimensional thermal model incorporates reversible, irreversible, and ohmic heat generation, as well as the temperature dependence of the various transport, kinetic and mass transfer parameters. The governing equations are semi-discretized in space using finite elements (with FEniCS software package (Alnæs et al., Arch Numer Softw 3(100), 2015)) and integrated in time with implicit Euler method. For each time step, this leads to a set of nonlinear equations that is solved fully coupled through Newton's iterations. This implementation is used to numerically simulate several charge-discharge cycles of a cell at 1C regime with an execution time around 50 times faster than real-time in a standard PC.

D. Aller Giráldez
Repsol Technology Center, Calle Agustín de Betancourt s/n, 28935 Móstoles, Spain
e-mail: david.aller@repsol.com

M.T. Cao-Rial
ITMATI, Campus Vida, 15782 Santiago de Compostela, Spain

Dpto. de Matemáticas, Universidade da Coruña, 15001 A Coruña, Spain
e-mail: teresa.cao@udc.es

P. Fontán Muiños
ITMATI, Campus Vida, 15782 Santiago de Compostela, Spain
e-mail: pedro.fontan@usc.es

J. Rodríguez (✉)
ITMATI, Campus Vida, 15782 Santiago de Compostela, Spain

Dpto. de Matemática Aplicada, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, A Coruña, Spain
e-mail: jeronimo.rodriguez@usc.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_84

1 Introduction

Li-ion batteries are nowadays one of the most advanced rechargeable electrochemical energy storage technologies in terms of both energy density and lifetime. Furthermore, it is a technology deployed at a large scale commercial level with applications that range from consumer electronics to electric vehicles. However, despite the millions of units produced, the basic mechanisms that rule how the properties of a cell will evolve in time under real-life use conditions are not very well understood. As in many other fields, mathematical modeling and numerical simulation of the underlying physics (combined with laboratory experimentation) are fundamental tools to accomplish not only this task but also to improve the battery performance, reduce the charging time and increase its durability. In this regard, almost all works that use this approach are based on the pioneering work of Newman [5]. There are several authors that have applied Newman's theory to produce physics based electrochemical models of Li-ion cells [2, 4, 11]. The coupling of the electrochemical system with the generation and transport of heat has also been addressed [3, 9]. Finally, the aging process of a Li-ion cell has received much less attention. Among all possible aging mechanism that a cell undergoes (see the extensive review by Vetter et al. [10]) the growth of the SEI (Solid Electrolyte Interface) on a graphite anode is the most studied one [6–8].

In this contribution we present and numerically solve a thermoelectrochemical model for the simulation of Li-ion cells. In Sect. 2 the complete mathematical model is presented. In Sect. 3 some numerical results using this approach are shown.

2 Mathematical Model

In this section we will recall the equations that model Li-ion cells based on Newman's concentrated-solution theory [5] and the equations for the SEI growth.

This electrochemical device is composed by the negative electrode that occupies $\Omega_- := [0, L_-]$ and a positive electrode placed at $\Omega_+ := [L_0 + L_-, L]$ (where $L := L_- + L_0 + L_+$) that are separated by the separator located at $\Omega_0 = [L_-, L_- + L_0]$ which is assumed to be an electric insulator (see Fig. 1 for further details). These

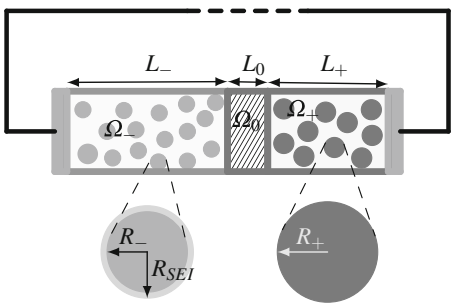


Fig. 1 Sketch of a cell. On the *left* (resp. *right*), the negative electrode Ω_- (resp. positive electrode Ω_+). In the *middle*, the separator Ω_0 which is assumed to be an electric insulator. At the micro-scale level, the electrodes are assumed to be formed by spheres (represented here by *circles*)

components are assumed to be porous and soaked with an electrolyte that occupies the whole domain $\Omega = \Omega_- \cup \Omega_0 \cup \Omega_+$. Notice that some of the components overlap due a homogenization process. The first set of equations, that are defined at the macroscopic level, are the following

$$\left|\begin{array}{rll} \frac{\partial \epsilon_e c_e}{\partial t} - \frac{\partial}{\partial x}\left(D_e^{\text{eff}} \frac{\partial c_e}{\partial x}\right) = \frac{1-t_+}{F} j_{\text{Li}}, & \text{in } \Omega, & (a) \\ \frac{\partial}{\partial x}\left(k^{\text{eff}} \frac{\partial \phi_e}{\partial x}\right) + \frac{\partial}{\partial x}\left(k_D^{\text{eff}} \frac{\partial \ln c_e}{\partial x}\right) = -j_{\text{Li}}, & \text{in } \Omega, & (b) \\ \frac{\partial}{\partial x}\left(\sigma^{\text{eff}} \frac{\partial \phi_s}{\partial x}\right) = j_{\text{Li}}, & \text{in } \Omega_\pm, & (c) \\ \frac{\partial \rho c_p T}{\partial t} - \frac{\partial}{\partial x}\left(k_T^{\text{eff}} \frac{\partial T}{\partial x}\right) = \dot{q}, & \text{in } \Omega, & (d) \end{array}\right. \tag{1}$$

where (see Table 1)

$$\begin{aligned} k_D^{\text{eff}} &= \frac{2RT}{F} k^{\text{eff}} \left(t^+ - 1\right)\left(1 + \frac{\partial \ln f_\pm}{\partial \ln c_e}\right), \\ \dot{q} &= j_{\text{Li}} \eta + j_{\text{Li}} T \frac{\partial U}{\partial T} + \sigma^{\text{eff}} \nabla \phi_s \cdot \nabla \phi_s + k^{\text{eff}} \nabla \phi_e \cdot \nabla \phi_e + k_D^{\text{eff}} \nabla \phi_e \cdot \nabla \ln c_e. \end{aligned} \tag{2}$$

Equation (1)(a) accounts for the time evolution of c_e the Li_+ concentration in the electrolyte by diffusion and migration. The volumetric reaction term involving j_{Li} comes from the reversible (reduction-oxidation) reaction produced at the interface between the active particles (the electrodes) and the electrolyte and it will be discussed later. Equation (1)(b) dictates the behaviour of ϕ_e the electric potential in the electrolyte and represents the charge conservation while (1)(c) is Ohm's law for electric conducting materials and it involves ϕ_s the electric potential in the electrodes. Finally, Eq. (1)(d) is the heat equation that includes the reversible, irreversible and ohmic heat generation terms (see the second equation in (2)).

The boundary conditions involving the variables introduced so far are vanishing normal derivative of c_e and ϕ_e at both ends (i.e. $x \in \{0, L\}$), vanishing normal derivative of ϕ_s at $x \in \{L_-, L_- + L_0\}$ and a Fourier type condition for the temperature, namely

$$k_T^{\text{eff}} \frac{\partial T}{\partial x}\Big|_{x=0} = h_T \left(T - T_{ext}\right)|_{x=0}, \quad -k_T^{\text{eff}} \frac{\partial T}{\partial x}\Big|_{x=L} = h_T \left(T - T_{ext}\right)|_{x=L}. \tag{3}$$

The boundary conditions for ϕ_s at the current collectors (i.e. $x \in \{0, L\}$) depend on the way the cell is operated. On one hand, when the current is applied we set

$$-\sigma^{\text{eff}} \frac{\partial \phi_s}{\partial x}\Big|_{x=0} = \frac{i_{\text{app}}}{S}, \qquad -\sigma^{\text{eff}} \frac{\partial \phi_s}{\partial x}\Big|_{x=L} = \frac{i_{\text{app}}}{S}. \tag{4}$$

On the other hand, when the voltage is applied we impose

$$\phi_s|_{x=L} - \phi_s|_{x=0} = v_{\text{app}}, \qquad \sigma^{\text{eff}} \frac{\partial \phi_s}{\partial x}\Big|_{x=0} = \sigma^{\text{eff}} \frac{\partial \phi_s}{\partial x}\Big|_{x=L}. \tag{5}$$

Table 1 Main symbols and notations

Symbol	Meaning	Units	Symbol	Meaning	Units
c_e	Electrolyte phase Li concentration	$\mathrm{mol\,m^{-3}}$	c_s	Solid phase Li concentration	$\mathrm{mol\,m^{-3}}$
ϕ_e	Electrolyte phase potential	V	ϕ_s	Solid phase potential	V
ϵ_e	Volume fraction of electrolyte	Dimensionless	D_e^{eff}	Effective diffusion coefficient of lithium ions in electrolyte	$\mathrm{m^2\,s^{-1}}$
k^{eff}	Effective ionic conductivity	$\mathrm{S\,m^{-1}}$	k_D^{eff}	Effective diffusional conductivity	$\mathrm{S\,m^{-1}}$
D_s	Diffusion coefficient of Li within the active particle	$\mathrm{m^2\,s^{-1}}$	σ^{eff}	Effective electronic conductivity of solid phase	$\mathrm{S\,m^{-1}}$
t_+	Transference number	Dimensionless	F	Faraday constant	C/mol
T	Temperature	K	k_T^{eff}	Effective thermal conductivity	$\mathrm{W\,m^{-1}\,K^{-1}}$
h_T	Thermal conductivity coefficient with the exterior	$\mathrm{W\,m^{-2}\,K^{-1}}$	R	Universal gas constant	$\mathrm{mol^{-1}\,K^{-1}}$
ρ	Density	$\mathrm{kg\,m^{-3}}$	c_p	Specific heat capacity	$\mathrm{J\,kg^{-1}\,K^{-1}}$
I_{app}	Applied current density	V	v_{app}	Applied voltage	V
A	Current collectors area	$\mathrm{m^2}$	$\alpha_+,\ \alpha_-$	Kinetic constants	Dimensionless
j_{Li}	Transfer current	$\mathrm{A\,m^{-3}}$	$\dot{q}$	Heat generation per unit volume	$\mathrm{W\,m^{-3}}$
i_0	Exchange current density	$\mathrm{A\,m^{-2}}$	a_s	Active surface area to volume ratio	$\mathrm{m^{-1}}$
η	Overpotential	V	U	Open circuit potential	V
R_{SEI}	Active particles with SEI layer radius	m	c_r	Reactant concentration	$\mathrm{mol\,m^{-3}}$
D_r	Diffusion coefficient in SEI layer	$\mathrm{m^2\,s^{-1}}$	ϵ_s	Volume fraction of solid phase	Dimensionless
k_r	SEI reaction rate	$\mathrm{C\,m\,mol^{-1}\,s^{-1}}$	ρ_{SEI}	SEI density	$\mathrm{kg\,m^{-3}}$
M_p	Molecular mass of SEI product	$\mathrm{kg\,mol^{-1}}$	κ_{SEI}	SEI ionic conductivity	$\mathrm{S\,m^{-1}}$
η_r	SEI overpotential	V	U_r	SEI open circuit potential	V
β_r	SEI kinetic constant	Dimensionless	j_r	SEI transfer current	$\mathrm{A\,m^{-3}}$

The other fundamental phenomena in the cell, that is modeled at a different and smaller scale, is the transport of lithium into the active particles by pure diffusion and the reactions produced at their boundaries. In this sense, the porous electrodes are assumed to be formed by small spheres of radius $R_\star$, $\star \in \{-,+\}$ constant in each electrode. The equations, using spherical coordinates and assuming spherical symmetry at the small scale, are the following (where $c_s \equiv c_s(x,r)$ represents the lithium concentration into the active particles)

$$\begin{aligned} &\tfrac{\partial c_s}{\partial t} - \tfrac{1}{r^2}\tfrac{\partial}{\partial r}\left(r^2 D_s \tfrac{\partial c_s}{\partial r}\right) = 0, \quad \text{in } \Omega_{ap}, \quad \left.\tfrac{\partial c_s}{\partial r}\right|_{r=0} = 0, \\ &-D_s \left.\tfrac{\partial c_s}{\partial r}\right|_{r=R_-} = \tfrac{1}{a_s F}\left(j_{\mathrm{Li}} - j_r\right), \quad -D_s \left.\tfrac{\partial c_s}{\partial r}\right|_{r=R_+} = \tfrac{j_{\mathrm{Li}}}{a_s F}. \end{aligned} \tag{6}$$

It is worth noting that the flux at the boundary depends on the reversible (reduction-oxidation) reaction, that involves (neutral) lithium in the electrodes and lithium ions in the electrolyte, produced at the boundary of the active particles. This reaction is modeled by Butler-Volmer equation, namely (j_{Li} is assumed to vanish in Ω_0)

$$j_{\mathrm{Li}} = a_s\, i_0 \left(e^{\frac{\alpha - F}{RT}\eta} - e^{-\frac{\alpha_+ F}{RT}\eta}\right), \quad \eta = \phi_s - \phi_e - U - \frac{Z_{\mathrm{SEI}}}{a_s}\left(j_{\mathrm{Li}} - j_r\right), \tag{7}$$

$$i_0 = k_0\, c_e^{\alpha-}\, (c_{s,\max} - c_s)^{\alpha-}\, c_s^{\alpha+}, \quad a_s = \frac{3\epsilon_s}{R_\star},\ \star \in \{-,+\}, \quad Z_{SEI}(t) = \tfrac{R_{SEI}(t) - R_-}{\kappa_{SEI}}.$$

In the negative electrode this flux also depends on j_r, term that models the aging effects that will be introduced in the next paragraph.

In order to take into account degradation phenomena, we consider the Solid Electrolyte Interface layer in the negative electrode $\Omega_{SEI} = \Omega_- \times [R_-, R_{SEI}]$. In this new domain, the SEI growth is considered to be produced by an irreversible 2-electron side reaction that happens at the interface between the active particle and the SEI. According to [8], the main electrolyte decomposition reaction is,

$$2\,(\mathrm{CH_2O})_2\,\mathrm{CO} + 2\mathrm{e}^- + 2\mathrm{Li}^+ \rightarrow (\mathrm{CH_2OCO_2Li})_2 + \mathrm{H_4C_{2(g)}}. \tag{8}$$

The SEI growth is thus modeled by the following equation, expressed again in spherical coordinates, to be solved at the small scale (where $c_r \equiv c_r(x,r)$ represents the reactant concentration in the SEI)

$$\tfrac{\partial c_r}{\partial t} = \tfrac{1}{r^2}\tfrac{\partial}{\partial r}\left(r^2 D_r \tfrac{\partial c_r}{\partial r} - r^2 \tfrac{\mathrm{d}R_{SEI}}{\mathrm{d}t} c_r\right) \quad \text{in } \Omega_{SEI}, \tag{9}$$

where the radius of the layer at point $x \in \Omega_-$, denoted by $R_{SEI}(x,t)$, is a monotonically increasing magnitude driven by the following ordinary differential equation

$$\tfrac{\mathrm{d}R_{SEI}}{\mathrm{d}t} = -\tfrac{R_-^3 M_p}{6F\epsilon_s R_{SEI}^2 \rho_{SEI}} j_r. \tag{10}$$

This time evolution of the SEI introduces another difficulty in the model since the volume fraction of electrolyte will no longer be constant but it will be given by

$$\epsilon_e(t) = 1 - \epsilon_{\text{other}} - \epsilon_s - \epsilon_{SEI}(t), \qquad \epsilon_{SEI}(t) = \epsilon_s\left(\frac{R_{SEI}^3(t) - R_-^3}{R_-^3}\right). \tag{11}$$

Equation (9) is completed with the following boundary conditions

$$D_r \frac{\partial c_r}{\partial r} - \frac{\mathrm{d}R_{SEI}}{\mathrm{d}t} c_r\Big|_{r=R_-} = \frac{j_r}{a_s F}, \qquad c_r|_{r=R_{SEI}} = 0.5\, c_{r,\text{sln}}, \tag{12}$$

where j_r, which is assumed to vanish in $\Omega \setminus \overline{\Omega_-}$, is given by

$$j_r = -a_s k_r c_r e^{-\frac{\beta_r F}{RT}\eta_r},\ \eta_r = \phi_s - \phi_e - U_r - \frac{Z_{SEI}}{a_s}\left(j_{\text{Li}} - j_r\right). \tag{13}$$

The first boundary condition indicates that the flux of reactant across the interface where the irreversible reaction takes place is proportional to j_r while the second one means that the concentration of reactant at the external boundary of the layer is proportional to $c_{r,\text{sln}}$ the concentration of solvent in the electrolyte (which is assumed to be very high). The equations are completed with the additional constraint $\int_\Omega \phi_e \, \mathrm{d}x = 0$ and initial conditions for c_e, c_s, c_r and R_{SEI}.

This system is semi-discretized in space with finite elements using FEniCS software package [1] and integrated in time with implicit Euler method. For each time step, the nonlinear system of equations is solved with Newton's iterations.

3 Numerical Results

In this section we show some results obtained with the numerical resolution of the previous model using the reference values for the parameters shown in Table 2 (see also [8]).

Table 2 Reference values for the parameters

	N.E.	P.E		N.E.	P.E.		SEI
ϵ_s	0.515	0.615	D_e	7.5e−10	7.5e−10	D_r	6.8e−21
D_s	2.0e−14	1.0e−14	σ	5.764	12.117	k_r	1.3122e−7
t_+	0.363	0.363	L	88.e−6	80.e−6	ρ_{SEI}	1690
h_T	10	10	k_T	5	5	κ_{SEI}	1.e−6
ρ	2500	1500	c_p	700	700	U_r	0.4
$R_\pm$	2.e−6	2.e−6	A	0.0674	0.0674	$\alpha_\star$, β_r	0.5
i_0	36	26	T	298.15	298.15	M_p	0.162

N.E. (resp. P.E.) stands for negative (resp. positive) electrode

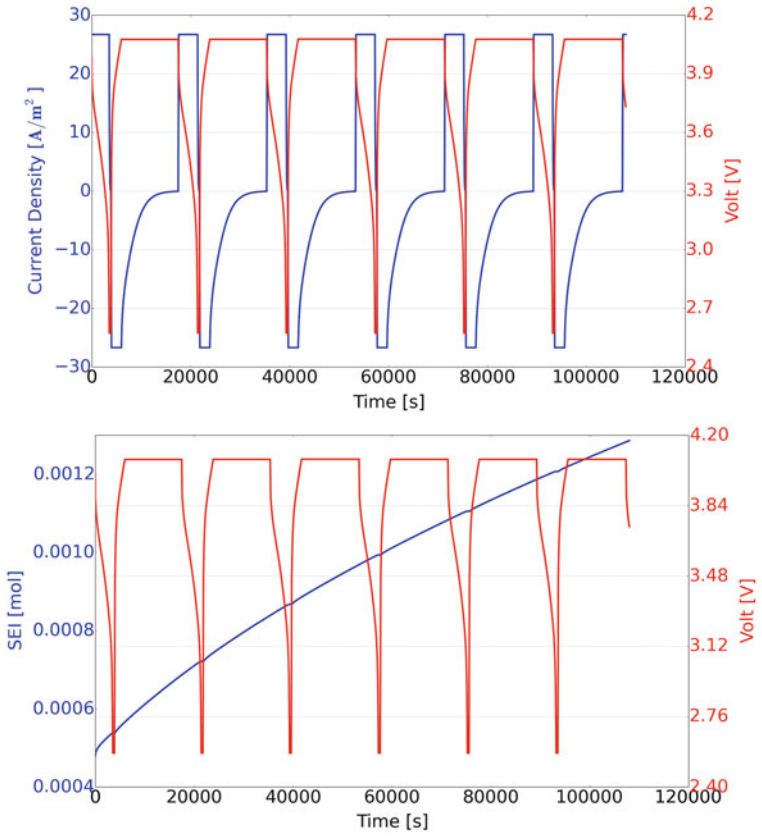


Fig. 2 Voltage, current and SEI growth evolution in time during 30 h cycling

The simulation consists on a battery performing several charge-discharge cycles at 1C regime for 30 h (the computational time on a standard PC is about 50 times faster than real-time). As shown in Fig. 2 on the top, this is done by switching between the boundary conditions (4) and (5). In this case, we start with applied current and, when the voltage reaches some trigger value, the boundary conditions are changed to applied (constant) voltage. When the current reaches a minimum value, the boundary conditions are changed again to applied current. On the bottom we show the evolution of the SEI growth and its correlation with the voltage. It is interesting to note that the smaller the voltage, the slower the SEI growth is.

In Fig. 3 on the top, we show the time evolution of temperature and voltage when considering a thermal conductivity with the exterior of $10[\mathrm{W}/(\mathrm{m}^2\,\mathrm{K})]$. On the bottom we show voltage versus the State of Charge (SOC) while cycling. It is slightly noticeable the loss of capacity, even though the number of applied cycles is

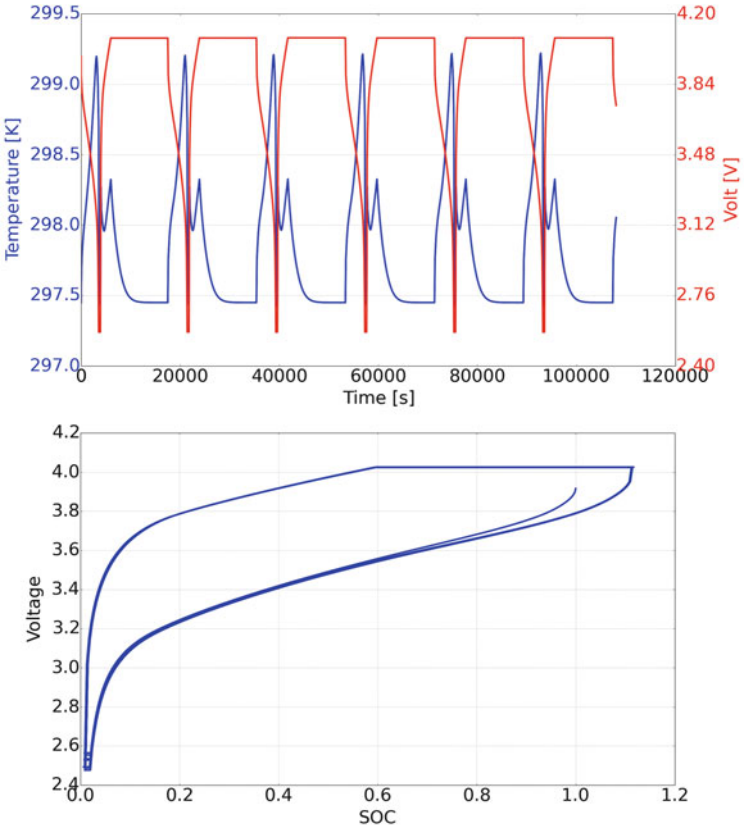


Fig. 3 Temperature and voltage evolution during cycling. Voltage vs. SOC

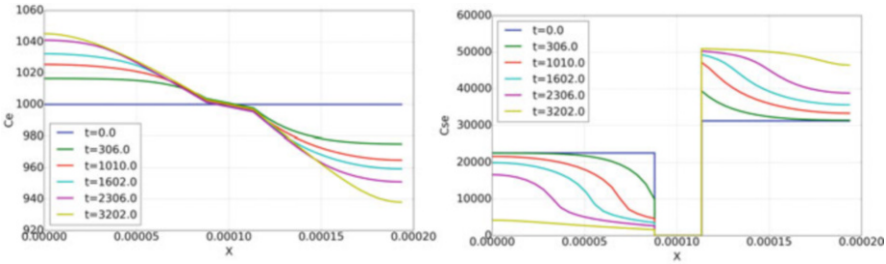


Fig. 4 Snapshots of c_e (*left*) and $c_s(r = R_\star)$, $\star \in \{-, +\}$ (*right*) during the first 1 h discharge

quite small. Finally, in Fig. 4 we show several snapshots of c_e (the Li_+ concentration in the electrolyte (left)) and $c_s(r = R_\star)$ (the Li concentration at the surface of the active particles (right)) during the first discharge. As expected, the amount of lithium decreases in the negative electrode while it increases in the positive.

Acknowledgements This work was funded by *Xunta de Galicia*, *ITMATI* and *Repsol* through the Joined Research Unit IN853*A* 2014/03.

References

1. Alnæs, M.S., et al.: The FEniCS Project Version 1.5. Arch. Numer. Softw. **3**(100) (2015)
2. Fang, W., Kwon, O.J., Wang, C.Y.: Electrochemical-thermal modeling of automotive Li-ion batteries and experimental validation using a three-electrode cell. Int. J. Energy Res. **34**(2), 107–115 (2010)
3. Gu, W.B., Wang, C.Y.: Thermal-electrochemical modeling of battery systems. J. Electrochem. Soc. **147**(8), 2910–2922 (2000)
4. Mazumder, S., Lu, J.: Faster-than-real-time simulation of lithium ion batteries with full spatial and temporal resolution. Int. J. Electrochem. **2013**, 10 pp. (2013)
5. Newman, J., Thomas-Alyea, K.E.: Electrochemical Systems. Wiley, New York (2004)
6. Ning, G., Popov, B.N.: Cycle life modeling of lithium–ion batteries. J. Electrochem. Soc. **151**(10), A1584–A1591 (2004)
7. Ploehn, H.J., Ramadass, P., White, R.E.: Solvent diffusion model for aging of lithium–ion battery cells. J. Electrochem. Soc. **151**(3), A456–A462 (2004)
8. Safari, M., Morcrette, M., Teyssot, A., Delacourt, C.: Multimodal physics-based aging model for life prediction of Li–ion batteries. J. Electrochem. Soc. **156**(3), A145–A153 (2009)
9. Srinivasan, V., Wang, C.Y.: Analysis of electrochemical and thermal behavior of Li-ion cells. J. Electrochem. Soc. **150**(1), A98–A106 (2003)
10. Vetter, J., Novák, P., Wagner, M.R., Veit, C., Möller, K.-C., Besenhard, J.O., Winter, M., Wohlfahrt-Mehrens, M., Vogler, C., Hammouche, A.: Ageing mechanisms in lithium-ion batteries. J. Power Sources **147**(1), 269–281 (2005)
11. Wang, C.Y., Gu, W.B., Liaw, B.Y.: Micro-macroscopic coupled modeling of batteries and fuel cells I. Model development. J. Electrochem. Soc. **145**(10), 3407–3417 (1998)

Numerical Simulation of a Network of Li-Ion Cells Using an Electrochemical Model

David Aller Giráldez, M. Teresa Cao-Rial, Manuel Cremades Buján, Pedro Fontán Muiños, and Jerónimo Rodríguez

Abstract The battery system of an electric vehicle comprises hundreds of battery packs connected in both parallel and series, plus many other electric components required for the correct charge, power and thermal management of the system. Furthermore, each battery pack is a stack of several individual electrochemical cells connected in both parallel and series and thermally coupled. The aim of the present work is the numerical simulation of a pack of Li-ion cells. To this end, the device is modeled as an electrical network where the edges that contain electrochemical cells are handled by means of a Steklov-Poincaré operator associated to a model that covers electrochemical, thermal and aging effects. The full differential-algebraic system issued from the space discretization of the cells and the coupling between edges through Kirchhoff's laws is integrated in time by means of implicit Euler method and Newton's iterations. This procedure allows to decouple the resolution of the electrochemical cells leading to a highly parallelizable and computationally efficient algorithm. This approach is used to numerically simulate a pack of cells.

D. Aller Giráldez
Repsol Technology Center, Calle Agustín de Betancourt s/n, 28935 Móstoles, Spain
e-mail: david.aller@repsol.com

M.T. Cao-Rial
ITMATI, Campus Vida, 15782 Santiago de Compostela, Spain

Dpto. de Matemáticas, Universidade da Coruña, 15001 A Coruña, Spain
e-mail: teresa.cao@udc.es

M. Cremades Buján • P. Fontán Muiños
ITMATI, Campus Vida, 15782 Santiago de Compostela, Spain
e-mail: manuel.cremades@usc.es; pedro.fontan@usc.es

J. Rodríguez (✉)
ITMATI, Campus Vida, 15782 Santiago de Compostela, Spain

Dpto. de Matemática Aplicada, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, A Coruña, Spain
e-mail: jeronimo.rodriguez@usc.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_85

1 Introduction

It is widely agreed that the future of mobility is linked to electrical vehicles although it is not so clear when this transition will take place. It is also commonly agreed that the main technological limitation for the full deployment of the electric vehicle is the energy storage system. The Li-ion battery, although is probably not the best for this application, is one of the best technologies currently available. In this regard, the BMS (Battery Management System) of an electric vehicle plays a major role in the performance and lifetime of the batteries. Improving current BMS technology requires a deep understanding on the interaction of the BMS components with the complex electrochemical systems that are the batteries. Mathematical modeling of the BMS is one approach to achieve this objective. However, almost all published articles on BMS modeling are based on semi-empirical simplified models based on equivalent circuit analogies or other kind of fits, see for example Uddin et al. [7] and Gerver [3]. Nonetheless, there are few authors that have tried to use a full physics based approach for the battery pack level. We would like to point out the works of Guo et al. [4] and Kim et al. [5]. Finally, and to the best knowledge of the authors, the modeling of a full BMS by means of physics-based cell models has not been attempted yet.

In this contribution we model a pack of Li-ion cells as an electrical network where the constitutive laws for the edges that contain cells are based on the 1D-2D electrochemical model in [1]. In Sect. 2 we justify the fact that each cell can be treated as a device of an electrical circuit. In Sect. 3 the basic ideas of electrical network modeling are recalled. Special attention is paid to the edges with electrochemical cells whose constitutive law is handled by means of a Steklov-Poincaré operator associated to the cell model. In Sect. 4 several numerical simulations using this network model are presented.

2 Electrochemical Model

Since the main goal of this contribution is the simulation of several interconnected Li-ion cells we start by briefly describing the model for such device (see [1, 6] for further details). The 1D-2D electrochemical model for a Li-ion cell consists on a set of (at least) five nonlinear and fully coupled partial differential equations that involves many physical variables such as concentrations, electric potentials and temperature. Assuming that there is no thermal coupling between cells, effect that will be disregarded in this work (this difficulty will be tackled in future works), the cell only communicates with the external circuit through ϕ_s (the electric potential on the electrodes) whose main equation is

$$\frac{\partial}{\partial x}\left(\sigma^{eff}\frac{\partial \phi_s}{\partial x}\right) = j_{\mathrm{Li}}. \tag{1}$$

More precisely, this coupling is only produced through its boundary conditions at both ends of the cell, that is, at the current collectors. When a single cell is considered, the most common boundary conditions for ϕ_s at the current collectors are

$$-\sigma^{eff}\frac{\partial\phi_s}{\partial x}\bigg|_{x=0} = \frac{i_{\text{app}}}{S}, \quad (a) \qquad \text{or } \phi_{|x=L} - \phi_{|x=0} = v_{\text{app}}, \quad (b) \tag{2}$$

that will be referred as *applied current* and *applied voltage*.

An important fact on this model is that, regardless what these boundary conditions are, the following compatibility condition needs to be satisfied

$$\sigma^{eff}\frac{\partial\phi_s}{\partial x}\bigg|_{x=0} = \sigma^{eff}\frac{\partial\phi_s}{\partial x}\bigg|_{x=L}. \tag{3}$$

We can thus introduce (S is the area of the current collectors)

$$i := -S\,\sigma^{eff}\frac{\partial\phi_s}{\partial x}\bigg|_{x=0}\left(= -S\,\sigma^{eff}\frac{\partial\phi_s}{\partial x}\bigg|_{x=L}\right), \qquad v := \phi_{|x=L} - \phi_{|x=0}, \tag{4}$$

the current and the voltage through the cell. In this way, the electrochemical cell can be understood as a device that relates the current and the voltage (as any other classical component in an electrical network) through the *rather complex* cell model (see [1] for further details).

3 Network Model

Following Bermúdez et al. [2], the network is modeled as an oriented graph with N nodes and E edges were the electrical devices are located (see Fig. 1). The connectivity of the network can be represented by the *incidence* matrix

$$A_{jk} = \begin{cases} -1 \text{ if the j-th node is the first of the k-th edge,} \\ 1 \text{ if the j-th node is the last of the k-th edge,} \\ 0 \text{ in any other case.} \end{cases} \tag{5}$$

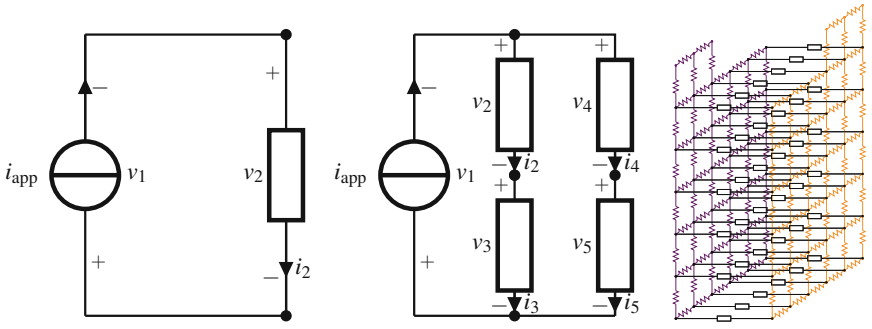


Fig. 1 A single cell, a battery pack and the equivalent circuit for a cubic cell

The unknowns in our network model will be the current through the edges

$$i_k, \quad k = 1, \ldots, E, \tag{6}$$

and the potentials on the nodes

$$u_j, \quad j = 1, \ldots, N. \tag{7}$$

3.1 Kirchhoff's Laws

Currents on edges are constrained by charge conservation on the network. Indeed, the current flowing into a given node must be equal to the current flowing out that node. This constraint can be written as ($\mathbf{i} \in R^E$ is the vector of edge currents)

$$A\mathbf{i} = \mathbf{0} \quad \Leftrightarrow \quad \sum_{k=1}^{E} A_{jk} i_k = 0, \quad j = 1, \ldots, N, \tag{8}$$

and it is often called *Kirchhoff's current law*. Some of these equations are linearly dependent as stated in the following proposition whose proof can be found in [2].

Proposition 1 *If the network is connected then Kirchhoff's current law provides exactly $N - 1$ linearly independent equations.*

It is useful to introduce the voltages on edges defined as the difference between potentials on the corresponding nodes. These relations between voltages and potentials, often called *Kirchhoff's voltage law* can be mathematically written as follows

$$\mathbf{v} = A^T \mathbf{u} \quad \Leftrightarrow \quad v_k = \sum_{j=1}^{N} A_{jk} u_j, \quad k = 1, \ldots, E, \tag{9}$$

where $\mathbf{v} \in R^E$ (resp. $\mathbf{u} \in R^N$) is the vector of edge voltages (resp. the vector of node potentials).

3.2 Constitutive Laws

At each edge, we have an electrical device with a constitutive law relating the current flowing through that edge with the respective voltage, denoted by:

$$f_k(t, i_k, v_k) = 0, \quad k = 1, \ldots, E. \tag{10}$$

For example, for a linear resistor this relation is given by $f_k(t, i_k, v_k) = v_k - R_k i_k$ where R_k is the resistance (see Bermúdez et al. [2] for other classical devices).

The constitutive law for the electrochemical cells can be described by a Steklov-Poincaré operator associated to the cell model. More precisely, we set

$$f_k(t, i_k, v_k) = v_k - \mathscr{V}_k(t, i_k), \tag{11}$$

where $\mathscr{V}_k(t, i_k) := \phi_s(x = L_k, t) - \phi_s(x = 0, t)$ and ϕ_s is the electric potential in the electrodes obtained by solving the cell model with *applied current* boundary conditions with $i_{\text{app}} := i_k$ (see 2(a)). This operator $\mathscr{V}$ is often called the Neumann to Dirichlet operator.

Remark 1 An alternative way to establish the constitutive law for the cell would be

$$f_k(t, i_k, v_k) = i_k - \mathscr{I}_k(t, v_k), \tag{12}$$

where the Dirichlet to Neumann operator $\mathscr{I}_k(t, v_k) := -S_k\, \sigma^{eff}\, \partial\phi_s/\partial x\big|_{x=0}$ and ϕ_s is the electric potential in the electrodes obtained by solving the cell model with *applied voltage* boundary conditions with $v_{\text{app}} := v_k$ (see 2(b)). Nevertheless, in this contribution we will only use the Neumann to Dirichlet operator.

3.3 Edge-Node Formulation

Combining (10) with *Kirchhoff's voltage law* (see (9)) we get

$$f_k(t, i_k, \textstyle\sum_{j=1}^{N} A_{jk} u_j) = 0, \quad k = 1, \ldots, E, \tag{13}$$

equations that we complete with *Kirchhoff's current law* (see (8)) to obtain a system with $N + E$ equations (for the same number of unknowns). However, due to Proposition 1, this system is not uniquely solvable. This is related to the fact that potentials are defined up to an additive constant.

For this reason in practice we consider $N - 1$ linearly independent equations in (8) (for simplicity we will assume that we can take the $N-1$ first) and we impose vanishing mean on the potentials. This leads to the following set of equations

$$\left|\begin{array}{r} f_k(t, i_k, \sum_{j=1}^{N} A_{jk} u_j) = 0,\ k = 1, \ldots, E, \\ \sum_{k=1}^{E} A_{jk} i_k = 0,\ j = 1, \ldots, N-1, \\ \sum_{j=1}^{N} u_j = 0. \end{array}\right. \tag{14}$$

This system is discretized in time using implicit Euler method. For each time step, the nonlinear algebraic system is solved using Newton's iterations.

Remark 2 At each Newton's iteration, we need to evaluate the constitutive laws and their derivatives with respect to the currents and voltages. This operation, that can be computationally expensive for edges containing cells, can be done fully in parallel.

4 Numerical Results

In this section we present numerical simulations for the three networks on Fig. 1. For all those simulations, four charge-discharge cycles are considered, with the same cell model parameters used by Aller et al. [1].

4.1 Simulation of a Single Cell

The cycling process starts with the cell fully charged. A constant current is applied until a minimum voltage is reached. At that time, a constant voltage (equal to this minimum value) is applied until the current drops near zero: the cell is almost fully discharged. The process for charging is analogous. See Fig. 2 on the left (resp. on the right) for the voltage and current evolution (resp. the number of Newton's iterations).

Remark 3 In Aller et al. [1] this cycling process is performed by changing the boundary conditions of the cell model when the cell operational mode changes. It is worth noting that in this work only the cell solver with applied current is used (see Remark 1).

4.2 Simulation of a Battery Pack

For the battery pack of Fig. 1 mid, it can be seen that the current and voltage through the source is duplicated with respect to the case of a single cell (cf. Figs. 2 and 3).

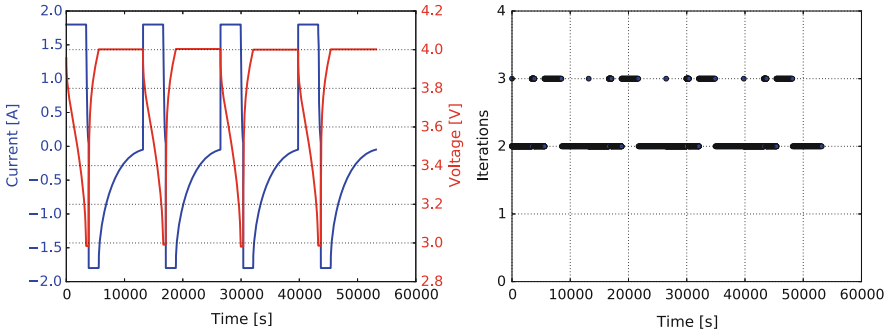


Fig. 2 Currents and voltages (*left*) on the source and number of Newton iterations (*right*)

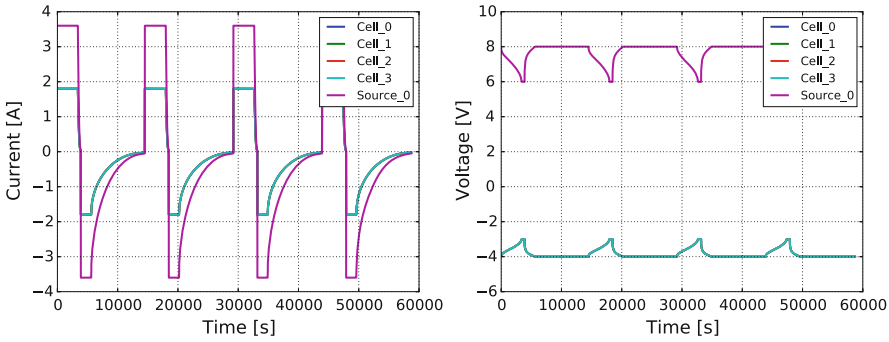


Fig. 3 Currents (*left*) and voltages (*right*) on the circuit edges

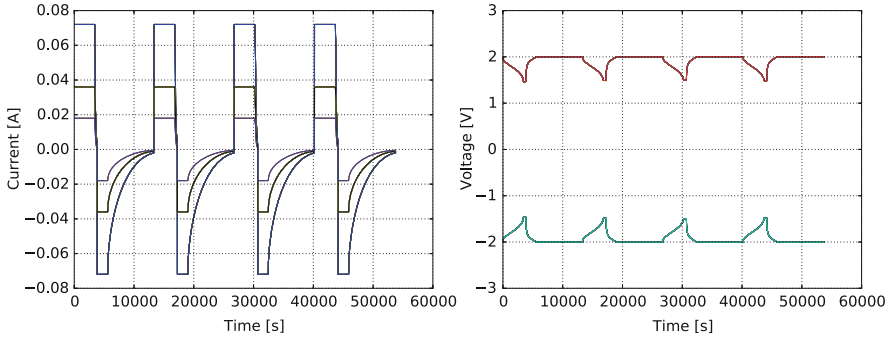


Fig. 4 Currents (*left*) and voltages (*right*) on the 64 Li-ion cells

4.3 Simulation of a Cubic Cell

Assuming that ionic transport is (mainly) perpendicular to the current collectors, we can model a 3D cubic cell interconnecting through the current collectors several 1D cell models. The current collectors are modeled in this case as multiple resistors (see the colored lines in Fig. 1 on the right) whose resistances are computed by means of Poulliet's law from the resistivity, that is an intrinsic property of a given material. In Fig. 4 the results for this configuration are shown.

Acknowledgements This work was funded by *Xunta de Galicia*, *ITMATI* and *Repsol* through the Joined Research Unit IN853*A* 2014/03.

References

1. Aller Giráldez, D., Cao-Rial, M.T., Fontan Muiños, P., Rodríguez, J.: Numerical simulation of a Li-ion cell using a thermoelectrochemical model including degradation. In: Quintela, P., Barral, P., Gómez, D., Pena, F.J., Rodríguez, J., Salgado, P., Vázquez-Méndez, M.E. (eds.), Progress in Industrial Mathematics at ECMI 2016. Mathematics in Industry, vol. 26. Springer, Cham (2017). doi:10.1007/978-3-319-63082-3

2. Bermúdez, A., Gómez, D., Salgado, P.: Mathematical Models and Numerical Simulation in Electromagnetism. Springer, Berlin (2014)
3. Gerver, R.E.: 3D thermal-electrochemical lithium-ion battery computational modeling. The University of Texas at Austin (2009)
4. Guo, M., White, R.E.: Thermal model for lithium ion battery pack with mixed parallel and series configuration. J. Electrochem. Soc. **158**(10), A1166–A1176 (2011)
5. Kim, S.U., Albertus, P., Cook, D., Monroe, C.W., Christensen, J.: Thermoelectrochemical simulations of performance and abuse in 50-Ah automotive cells. J. Power Sources **268**, 625–633 (2014)
6. Newman, J., Thomas-Alyea, K.E.: Electrochemical Systems. Wiley, New York (2004)
7. Uddin, K., Picarelli, A., Lyness, C., Marco. J.: An Acausal Li-Ion battery pack model for automotive applications. Energies **7**, 5675–5700 (2014)

Symplectic Lanczos and Arnoldi Method for Solving Linear Hamiltonian Systems: Preservation of Energy and Other Invariants

Elena Celledoni and Lu Li

Abstract Krylov subspace methods have become popular for the numerical approximation of matrix functions as for example for the numerical solution of large and sparse linear systems of ordinary differential equations. One well known technique is based on the method of Arnoldi which computes an orthonormal basis of the Krylov subspace. However, when applied to Hamiltonian linear systems of ODEs, this method fails to preserve the symplecticity of the solution under numerical discretization, or to preserve energy. In this work we apply the Symplectic Lanczos Method to construct a J-orthogonal basis of the Krylov subspace. This basis is then used to construct a numerical approximation which is energy preserving. The symplectic Lanczos method is widely used to approximate eigenvalues of large and sparse Hamiltonian matrices, but the approach for solving linear Hamiltonian systems is not well known in the literature. We also show that under appropriate additional assumptions on the structure of the linear Hamiltonian system, the Arnoldi method can preserve certain invariants of the system. We finally investigate numerically the energy and global error behaviour for the methods.

1 Krylov Projection Method–Arnoldi Projection Method

Consider a linear Hamiltonian initial value problem of the form

$$\begin{aligned} \dot{y} &= Ay \\ y(0) &= y_0 \quad t > 0, \end{aligned} \tag{1}$$

E. Celledoni • L. Li (✉)
Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
e-mail: elena.celledoni@math.ntnu.no; lu.li@math.ntnu.no

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_86

where $y(t) \in \mathbb{R}^{2m}$, $A \in \mathbb{R}^{2m\times 2m}$ and $JA = H$ is symmetric, and $y_0 \in \mathbb{R}^{2m}$. We denote by J the $2m \times 2m$ matrix

$$J = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}, \tag{2}$$

with I_m the $m \times m$ identity matrix. The energy of system (1) is

$$\mathscr{H}(y) = \frac{1}{2} y^T J A y \equiv \frac{1}{2} {y_0}^T H y_0, \tag{3}$$

and it is preserved along solution trajectories, i.e. $\frac{d\mathscr{H}(y(t))}{dt} = 0$.

Consider the Krylov subspace of dimension n, generated by the matrix A and the vector y_0:

$$\mathscr{K}_n(A, y_0) := \text{span}\{y_0, Ay_0, \cdots, A^{n-1}y_0\}. \tag{4}$$

The basic idea of Krylov projection methods is to build a numerical approximation for the solution of (1) evolving in the Krylov subspace. This is done by solving (projected) linear systems of ordinary differential equations (ODEs) of much lower dimension than the original system.

One well known Krylov projection method is the one based on the Arnoldi algorithm [1, 3, 5]. This algorithm generates an orthonormal basis for $\mathscr{K}_n(A, y_0)$, which is stored in a $2m \times n$ matrix V_n, and an upper Hessenberg $n \times n$ matrix H_n, such that V_n and H_n satisfy

$$\begin{aligned} AV_n = V_n H_n + w_{n+1} e_n^T, \quad w_{n+1} = h_{n+1,n} v_{n+1}, \\ {V_n}^T V_n = I_n, \end{aligned} \tag{5}$$

where v_{n+1} is the last column of V_{n+1} and $h_{n+1,n}$ is the $(n+1, n)$ entry of H_{n+1}. The Arnoldi projection method (APM) is then defined by searching for an approximation $y_A(t) = V_n z(t)$ of $y(t)$, where z is the solution of the following smaller system

$$\begin{aligned} \dot{z} &= H_n z, \\ z(0) &= z_0. \end{aligned} \tag{6}$$

When applied to Hamiltonian systems, the APM fails in general to preserve symplecticity or energy. However, we will see that in special cases, the APM can show a very good behaviour with respect to energy-preservation.

In Fig. 1 we report numerical tests performed with the APM on the problem (1) when A is Hamiltonian and simultaneously skew-symmetric. In the case when A is Hamiltonian and simultaneously skew-symmetric, the APM shows almost preservation of energy and boundedness of the global error over long integration times. This does not occur for general Hamiltonian matrices, as can be confirmed

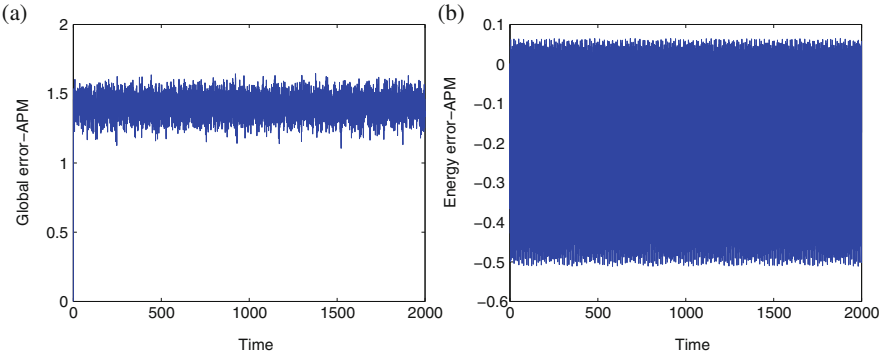


Fig. 1 (**a** and **b**) Block skew-symmetric A. (**a**) We plot the global error $|| \ y(t) - y_A(t) \ ||_2$ versus time, where $y(t)$ is the reference solution. The reference solution is always computed using Cayley transform. (**b**) We plot the energy error, where energy error is $\mathscr{H}(y_A(0)) - \mathscr{H}(y_A(t))$, and $y_A(0) = y_0$

by numerical experiments with JA symmetric block diagonal or symmetric block tridiagonal. We will explain this behaviour in the next section.

For all the experiments including all the figures in the following chapters, we always consider $A \in \mathbb{R}^{200\times 200}$, and we will fix the dimension of the Krylov subspace to be 4; A and y_0 are randomly generated.

1.1 APM Case When JA = AJ

In this section we consider a Hamiltonian matrix A such that A and J commute.

Proposition 1 *Suppose A is a Hamiltonian matrix. Then J and A commute if and only if the matrix A is skew-symmetric.*

So, in this case, the numerical experiments of Fig. 1 show that the global error and the energy error remain bounded over time. One can show that in this case (1) has m preserved first integrals in involution, and APM preserves r modified first integrals.

Proposition 2 *If $JA = AJ$, the Hamiltonian system (1) has m independent first integrals in involution, $\mathscr{H}_k(y) := \frac{1}{2}y^T A^{2k} y$ for $k = 0, 1, \ldots, m-1$, and in involution with the Hamiltonian $\mathscr{H}$.*

Proposition 3 *Applying the Arnoldi projection method to the Hamiltonian system (1), under the assumption $JA = AJ$, the numerical approximation preserves the following first integrals*

$$\mathscr{H}_k^A(y) := \frac{1}{2}y^T V_n (H_n)^{2k} V_n^T y, \quad k = 0, 1, \ldots r,$$

with $r = n/2 - 1$ if n is even and $r = (n-1)/2 - 1$ if n is odd.

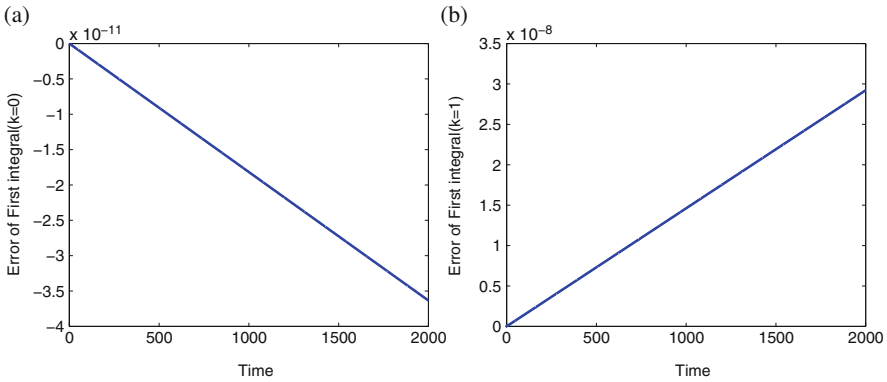


Fig. 2 (**a** and **b**) Block skew-symmetric A. Preservation of the two first integrals by APM for a skew-symmetric and Hamiltonian matrix A. (**a**) Numerical error $\mathscr{H}_0^A(y_A(0)) - \mathscr{H}_0^A(y_A(t))$, and $y_A(0) = y_0$. (**b**) Numerical error $\mathscr{H}_1^A(y_A(0)) - \mathscr{H}_1^A(y_A(t))$

In Fig. 2 we have performed experiments with $n = 4$, under the assumption that $JA = AJ$, APM preserves $r = 2$ first integrals, these are $\mathscr{H}_0^A(y_A) = \frac{1}{2}{y_A}^T y_A$, and $\mathscr{H}_1^A(y_A) = -\frac{1}{2}{y_A}^T V_n {H_n}^2 {V_n}^T y_A$.

The case $JA = AJ$ is not the only case when the energy error for APM appears to be bounded. We will introduce and explain another example in the following subsection. In what follows, we rewrite $JA = H$ in a block form

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{pmatrix}, \tag{7}$$

and then the original Hamiltonian system (1) has the form

$$\begin{aligned} \dot{p} &= -H_{12}^T p - H_{22} q \\ \dot{q} &= H_{11} p + H_{12} q, \end{aligned} \tag{8}$$

where $(p^T, q^T)^T = y$.

1.2 APM Case When $H_{1,2} = 0$, $H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$

If we consider Hamiltonian system (8) with $H_{1,2} = 0$, $H_{1,1} = I$ and an initial vector of the form $y_0 = ({p_0}^T, 0)^T$, we can show that the APM preserves the energy.

In Fig. 3, one can see the good performance of the APM when $H_{1,2} = 0, H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$. The energy of the original system is preserved and the global error is bounded. This can be explained by the following proposition.

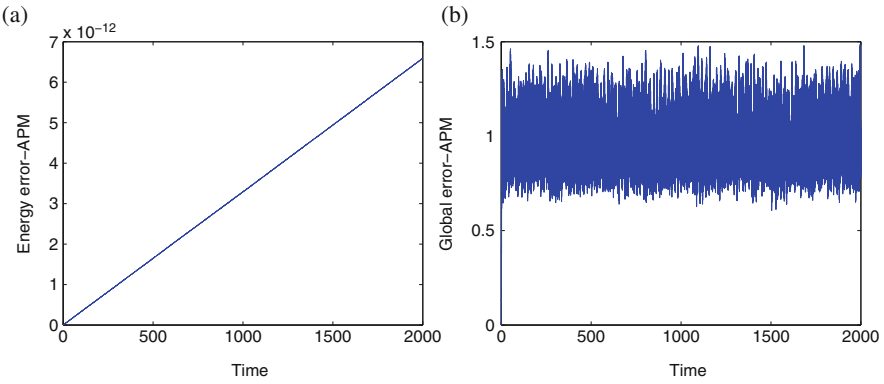


Fig. 3 For both experiments we consider A, such that $H_{1,2} = 0$, $H_{1,1} = I$, and an initial vector $y_0 = ({p_0}^T, 0)^T$. (**a**) Energy error versus time. (**b**) Global error versus time

Proposition 4 *In the case when $H_{1,2} = 0$, $H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$ (and subject to a suitable permutation of the equations), the projected system (6) for the APM is a Hamiltonian system. The energy of the original system (1) is preserved by the numerical solution of the APM.*

Remark 1 The APM applied to the special case when $H_{1,2} = 0$, $H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$ corresponds to performing a structure preserving model reduction in the spirit of [4]. Let V_n be the basis of the Krylov subspace $\mathscr{K}_n(-H_{22}, p_0)$. The approximations to the position q and the momentum p are restricted to evolve on this Krylov subspace and to take the form $V_n\hat{p}$ and $V_n\hat{q}$. Deriving the Euler-Lagrange equations and the corresponding Hamiltonian equations under this assumption, leads to a projected Hamiltonian system which coincides with (6).

2 Symplectic Lanczos Projection Method

Instead of using an orthonormal basis, we consider applying the Symplectic Lanczos algorithm to construct a J-orthogonal basis for the Krylov subspace [2, 6].

Denote by J_{2n} the matrix (2) with $m = n$. Given the Hamiltonian matrix $A \in \mathbb{R}^{2m,2m}$ and the starting vector $y_0 \in \mathbb{R}^{2m}$, the symplectic Lanczos method generates a sequence of $2m \times 2n$ matrices

$$S_{2n} = [v_1, \ldots, v_n, w_1, \ldots w_n], \tag{9}$$

which satisfy

$$AS_{2n} = S_{2n}H_{2n} + r_{n+1}e_{2n}^T, \tag{10}$$

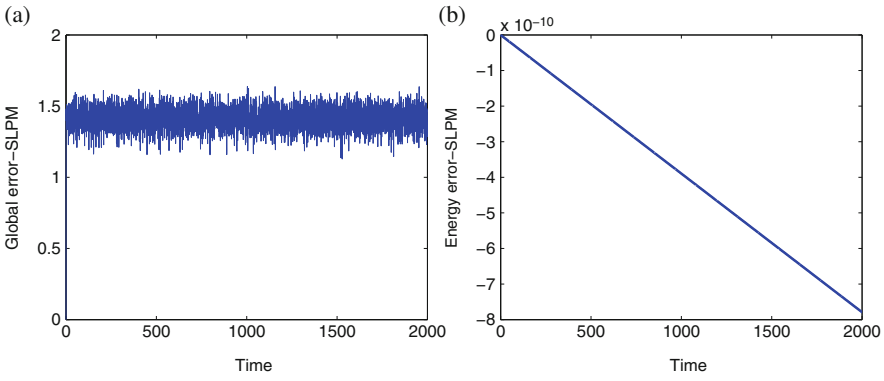


Fig. 4 (**a** and **b**) Block diagonal *JA*. (**a**) We plot the global error $|| \ y(t) - y_S(t) \ ||_2$ versus time. (**b**) We plot the energy error, where energy error is $\mathscr{H}(y_S(0)) - \mathscr{H}(y_S(t))$, and $y_S(0) = y_0$

where H_{2n} is a tridiagonal Hamiltonian $2m \times 2m$ matrix, S_{2n} is a symplectic matrix, i.e.,

$$S_{2n}{}^T J S_{2n} = J_{2n}, \tag{11}$$

and $r_{n+1} = \zeta_{n+1} v_{n+1}$ is J-orthogonal to the columns of S_{2n}, see [6] for details.

The Symplectic Lanczos projection method (which we here denote SLPM) constructs the approximation $y_S = S_{2n} \hat{z}(t)$ of $y(t)$, where $\hat{z}$ is the exact flow for the following smaller system

$$\begin{aligned} \dot{\hat{z}} &= H_{2n} \hat{z}, \\ \hat{z}(0) &= \hat{z}_0. \end{aligned} \tag{12}$$

Proposition 5 *The SLPM method preserves the energy when applied to the system (1) with A a Hamiltonian matrix.*

In Fig. 4, we report the performance of the SLPM on symmetric block diagonal *JA*, similar results have been obtained with a range of different Hamiltonian matrices. As expected, the energy is preserved and the global error remains bounded.

3 Conclusion

In this paper we have reported some experiments with different Krylov subspace methods for linear Hamiltonian systems: a method based on the symplectic Lanczos method, which always preserves the energy; and a projection method based on the classical Arnoldi algorithm. We have seen that the Arnoldi projection method can preserve some modified integrals when applied to special systems (i.e., when

A is Hamiltonian and skew-symmetric), and the APM preserves the energy when applied to some other special Hamiltonian matrices. We have recently obtained other Krylov subspace methods based on Gram Schmidt processes similar to the Arnoldi method, which preserve energy by solving exactly smaller Hamiltonian systems obtained by projection. The properties of all these Krylov methods and an appropriate comparison in terms of performance, stability with respect to propagation of rounding errors, and computational complexity will be discussed in a forthcoming paper.

References

1. Arnoldi, A.E.: The principle of minimized iterations in the solution of the matrix eigenvalue problem. Q. Appl. Math. **9**(1), 17–29 (1951)
2. Benner, P., Faßbender, H., Stoll, M.: A Hamiltonian Krylov–Schur-type method based on the symplectic Lanczos process. Linear Algebra Appl. **435**(3), 578–600 (2011)
3. Gallopulos, E., Saad, Y.: Efficient solution of parabolic equations by Krylov approximation method. SIAM J. Sci. Stat. Comput. **13**, 1236–1264 (1992)
4. Lall, S., Krysl, P., Marsden, J.E.: Structure-preserving model reduction for mechanical systems. Physica D **184**(1), 304–318 (2003)
5. Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia (2003)
6. Watkins, D.S.: On Hamiltonian and symplectic Lanczos processes. Linear Algebra Appl. **385**, 23–45 (2004)

A Self-adapting LPS Solver for Laminar and Turbulent Fluids in Industry and Hydrodynamic Flows

Tomás Chacón Rebollo, Enrique Delgado Ávila, Macarena Gómez Mármol, and Samuele Rubino

Abstract In this work we address the solution of the Navier–Stokes equations (NSE) in turbulent regime. On one side we focus is on the high-order term-by-term stabilization method that has one level, in the sense that it is defined on a single mesh, and in which the projection-stabilized structure of standard LPS methods is replaced by an interpolation-stabilized structure. The interest of LPS methods is that it ensures a self-adapting high accuracy in laminar regions of a turbulent flow, that turns to be of overall optimal high accuracy if the flow is fully laminar. On another side we present a reduced basis Smargorinsky turbulence model, based upon an empirical interpolation of the eddy viscosity term. This method yields dramatical improvements of the computing time, over 1000, for benchmark flows.

1 Introduction

In this paper, we consider two stabilized FE approximations of the unsteady NSE arising from LPS methods. At first we focus on the high-order term-by-term stabilization method introduced in (cf. [5]) for the Oseen equations. This method is a particular type of LPS scheme, which constitutes a low-cost, accurate solver for incompressible flows, despite being only weakly consistent since does not involve the full residual. It differs from the standard LPS methods (cf. [3, 11]) because it uses continuous buffer functions, it does not need enriched FE spaces, it does not need

T. Chacón Rebollo (✉) • E. Delgado Ávila
Departamento EDAN & IMUS, Universidad de Sevilla, C/Tarfia s/n, 41012 Sevilla, Spain
e-mail: chacon@us.es; edelgado1@us.es

M. Gómez Mármol
Departamento EDAN, Universidad de Sevilla, C/Tarfia s/n, 41012 Sevilla, Spain
e-mail: macarena@us.es

S. Rubino
Departamento de Análisis Matemático & IMUS, Universidad de Sevilla, C/Tarfia s/n, 41012 Sevilla, Spain
e-mail: samuele@us.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_87

element-wise projections satisfying suitable orthogonality properties, and it does not need multiple meshes. Commonly to standard LPS methods, the stabilization terms only act on the small scales of the flow, thus ensuring a higher accuracy with respect to more classical stabilization procedures, such as penalty-stabilized methods (cf. [4]).

We present some numerical tests for basic industrial flows, to assess the performance of the proposed LPS method as efficient and accurate solver for the simulation of turbulent incompressible flows arising in industrial and environmental applications.

We also present a promising reduced-basis Smagorinsky turbulence model, based upon the approximation of the eddy diffusion term by means of the empirical interpolation method, and on the approximation of velocity-pressure by a Greedy algorithm built with a specific error estimator. We present some numerical results for benchmark flows that show a dramatic speed-up of computation time, over 1000. The adaptation of this solver to complex flows, now in progress, is of primary interest for analysis and optimal design in fluid mechanics industrial applications.

2 A High-Order LPS Discretization of Evolution NSE

We introduce a numerical approximation for an Initial-Boundary Value Problem (IBVP) for the incompressible evolution NSE. Let $[0, T]$ be the time interval, and Ω a bounded polyhedral domain in $\mathbb{R}^d$, $d = 2$ or 3, with a Lipschitz-continuous boundary $\Gamma = \partial\Omega$. Let $\{\mathscr{T}_h\}_{h>0}$ be a family of affine-equivalent, conforming (i.e., without hanging nodes) and regular triangulations of $\overline{\Omega}$, formed by triangles ($d = 2$) or tetrahedra ($d = 3$).

Given an integer $l \geq 0$ and a mesh cell $K \in \mathscr{T}_h$, denote by $\mathbb{P}_l(K)$ the space of polynomials of degree $\leq l$, defined on K. We consider the following FE spaces for the velocity:

$$\begin{cases} Y_h^l = V_h^l(\Omega) = \{v_h \in C^0(\overline{\Omega}) : v_h|_K \in \mathbb{P}_l(K),\ \forall K \in \mathscr{T}_h\}, \\ \mathbf{Y}_h^l = [Y_h^l]^d = \{\mathbf{v}_h \in [C^0(\overline{\Omega})]^d : \mathbf{v}_h|_K \in [\mathbb{P}_l(K)]^d,\ \forall K \in \mathscr{T}_h\}, \\ \mathbf{X}_h = \mathbf{Y}_h^l \cap \mathbf{H}_0^1(\Omega). \end{cases} \tag{1}$$

We approximate the weak formulation of the unsteady NS equations for by a high-order term-by-term stabilization procedure in space (cf. [5]). To state this discretization, consider a positive integer number N and define $\Delta t = T/N$, $t_n = n\Delta t$, $n = 0, 1, \ldots, N$. We compute the approximations $\mathbf{u}_h^n$, p_h^n to $\mathbf{u}^n = \mathbf{u}(\cdot, t_n)$ and $p^n = p(\cdot, t_n)$ by:

- **Initialization.** Set:$\mathbf{u}_h^0 = \mathbf{u}_{0h}$.

- **Iteration.** For $n = 0, 1, \ldots, N-1$: *Given* $\mathbf{u}_h^n \in \mathbf{X}_h$, *find* $(\mathbf{u}_h^{n+1}, p_h^{n+1}) \in \mathbf{X}_h \times \mathbb{M}_h$ *such that:*

$$\begin{cases} \left(\dfrac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^n}{\Delta t}, \mathbf{v}_h\right)_\Omega + b(\mathbf{u}_h^n, \mathbf{u}_h^{n+1}, \mathbf{v}_h) + a(\mathbf{u}_h^{n+1}, \mathbf{v}_h) \\ -(p_h^{n+1}, \nabla \cdot \mathbf{v}_h)_\Omega + s_{\text{conv}}(\mathbf{u}_h^n, \mathbf{u}_h^{n+1}, \mathbf{v}_h) + s_{\text{div}}(\mathbf{u}_h^{n+1}, \mathbf{v}_h) = \langle \bar{\mathbf{f}}^{n+1}, \mathbf{v}_h \rangle, \\ (\nabla \cdot \mathbf{u}_h^{n+1}, q_h)_\Omega + s_{\text{pres}}(p_h^{n+1}, q_h) = 0, \end{cases} \tag{2}$$

  for any $(\mathbf{v}_h, q_h) \in \mathbf{X}_h \times \mathbb{M}_h$, where $\mathbb{M}_h = Y_h^l \cap L_0^2(\Omega)$, $\bar{\mathbf{f}}^{n+1}$ is the average value of $\mathbf{f}$ in $[t_n, t_{n+1}]$, and $\mathbf{u}_{0h}$ is some stable approximation to $\mathbf{u}_0$ belonging to $\mathbf{X}_h$, e.g., its discrete Stokes projection.

The forms a and b in (2) are given by:

$$b(\mathbf{u}_h^n, \mathbf{u}_h^{n+1}, \mathbf{v}_h) = \frac{1}{2}\left[(\mathbf{u}_h^n \cdot \nabla \mathbf{u}_h^{n+1}, \mathbf{v}_h)_\Omega - (\mathbf{u}_h^n \cdot \nabla \mathbf{v}_h, \mathbf{u}_h^{n+1})_\Omega\right], \tag{3}$$

$$a(\mathbf{u}_h^{n+1}, \mathbf{v}_h) = 2\nu\,(D(\mathbf{u}_h^{n+1}), D(\mathbf{v}_h))_\Omega. \tag{4}$$

where $D(\mathbf{u})$ is the symmetric deformation tensor. On the other hand, the forms s_{conv}, s_{div} and s_{pres} in (2) correspond to a high-order term-by-term stabilized method (cf. [5]), and are given by:

$$s_{\text{conv}}(\mathbf{u}_h^n, \mathbf{u}_h^{n+1}, \mathbf{v}_h) = \sum_{K \in \mathscr{T}_h} \tau_{v,K}(\sigma_h^*(\mathbf{u}_h^n \cdot \nabla \mathbf{u}_h^{n+1}), \sigma_h^*(\mathbf{u}_h^n \cdot \nabla \mathbf{v}_h))_K, \tag{5}$$

$$s_{\text{div}}(\mathbf{u}_h^{n+1}, \mathbf{v}_h) = \sum_{K \in \mathscr{T}_h} \tau_{d,K}(\sigma_h^*(\nabla \cdot \mathbf{u}_h^{n+1}), \sigma_h^*(\nabla \cdot \mathbf{v}_h))_K, \tag{6}$$

$$s_{\text{pres}}(p_h^{n+1}, q_h) = \sum_{K \in \mathscr{T}_h} \tau_{p,K}(\sigma_h^*(\nabla p_h^{n+1}), \sigma_h^*(\nabla q_h))_K. \tag{7}$$

Here, $\tau_{v,K}$, $\tau_{d,K}$ and $\tau_{p,K}$ are stabilization coefficients for convection, divergence and pressure gradient, respectively, and $\sigma_h^* = Id - \sigma_h$, where σ_h is some locally stable projection or interpolation operator from $\mathbf{L}^2(\Omega)$ on the foreground vector-valued space $\mathbf{Y}_h^{l-1}$ (also called "buffer space" in this context). Actually, σ_h is globally stable in $L^2(\Omega)$-norm, due to the regularity of the mesh. We also assume that σ_h satisfies optimal error estimates. In practical implementations, we choose σ_h as a Scott–Zhang-like linear interpolation operator in the space $\mathbf{Y}_h^{l-1}$ (cf. [13]). Actually, if needed, specific stabilizations for convection, divergence and pressure gradient may be used, through different approximation operators.

This method been recently supported by a thorough numerical analysis (stability, convergence, error estimates, asymptotic energy balance) for the nonlinear problem

related to the evolution NSE (cf. [2, 6]), which is to our knowledge unavailable for most turbulence models in the current literature (cf. [1]).

3 Numerical Experiments

In this section, we discuss some numerical results to analyze the basic numerical performances of the proposed model applied to the computation of turbulent flows, both in environmental and industrial applications.

3.1 Plane Mixing Layer (2D)

The mixing layer that forms between two co-flowing fluid streams merging with different velocities is an important model problem for the study of turbulence in free shear layers. Here, we are concerned with the hydrodynamic evolution of an incompressible plane mixing layer. For the setup of our numerical simulations, we chose to follow the guidelines given in [8], where numerical studies of a 2D mixing layer problem at $Re = 10^4$ for a LES with a three-level Variational Multi-Scale (VMS) were performed.

Free-slip boundary conditions are applied at the horizontal boundaries, and periodic boundary conditions are prescribed at the vertical boundaries. The initial velocity field is given by a hyperbolic tangent profile to which we superimposed a white-noise divergence-free perturbation of small amplitude (see [2] for a detailed description of the setting).

Three computational grids are used, consisting of uniform 40^2, 80^2, and 160^2 partitions of the domain. On these meshes, we consider two-dimensional $\mathbb{P}_2$ FE for velocity and pressure. The physical evolution of the flow can be described with the help of Fig. 1. These pictures are the result of a simulation using the proposed LPS method (2) on the finest grid of 160^2 mesh cells. They present the evolution of the vorticity through meaningful non-dimensional instants. This behavior almost corresponds to the one observed by Gravemeier on the finest grid in [8].

3.2 Turbulent Channel Flow (3D)

The proposed test consists of a fluid that flows between two parallel walls driven by an imposed pressure gradient source term which is defined by the friction Reynolds number Re_τ. For the setup of our numerical simulations, we chose to follow the guidelines given by Gravemeier in [7]. As a benchmark, we will use the fine Direct Numerical Simulation (DNS) of Moser et al. [12].

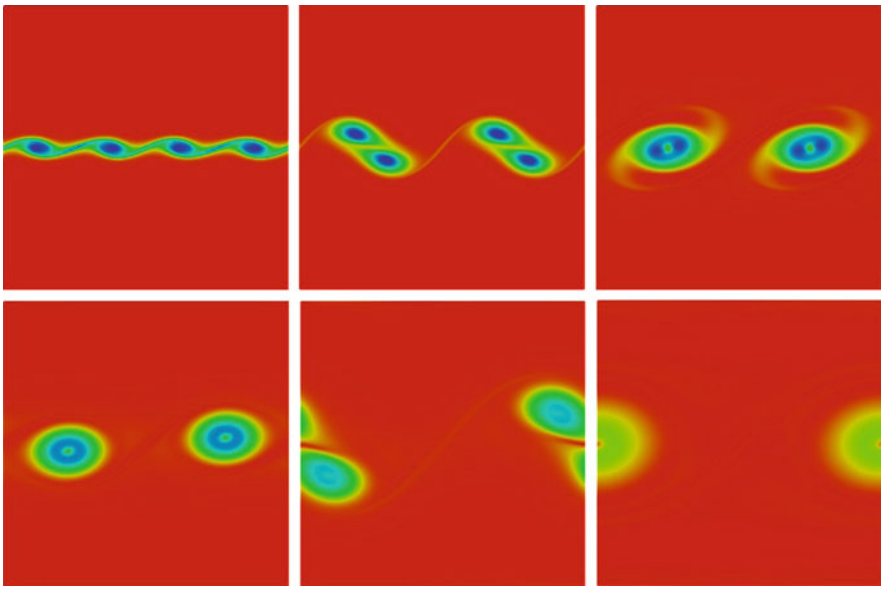


Fig. 1 Example 3.1. Colored vorticity field (*blue*: intense vorticity, *red*: irrotational outer flow) at time units 10, 30, 40, 100, 115, 200 (*left to right*, *top to bottom*)

The boundary conditions are periodic in both the stream-wise and span-wise directions (homogeneous directions). We perform a comparison between the application of the logarithmic wall-law of Prandtl and Von Kármán and no-slip boundary conditions at the walls.

We aim to obtain a good accuracy with a relatively coarse spatial resolution. Our grid consists of a $16 \times 32 \times 16$ partition of the channel, uniform in the homogeneous directions. The distribution of nodes in the wall-normal direction is non-uniform, and obeys the cosine function of Gauss-Lobatto. We use three-dimensional $\mathbb{P}_2$ FE for velocity and pressure.

We consider a semi-implicit Crank-Nicolson scheme for the temporal discretization. It provides a good compromise between accuracy and computational complexity, while keeping the numerical diffusion levels below the sub-grid terms (cf. [10]).

In Fig. 2 (left), we show the mean stream-wise velocity profiles, normalized by the computed wall shear velocity, in wall coordinates. The results show an acceptable agreement with the fine DNS, even with the very coarse basic discretization at hand (almost four times coarser than the DNS one).

Figure 2 (right) displays the normalized r.m.s. values of the stream-wise velocity fluctuations. If we compare with DNS data the LPS method by interpolation tested with no-slip boundary conditions, we observe a noticeable over-prediction, which seems to be corrected by the use of wall laws in the so-called inertial layer, starting from the first interior node.

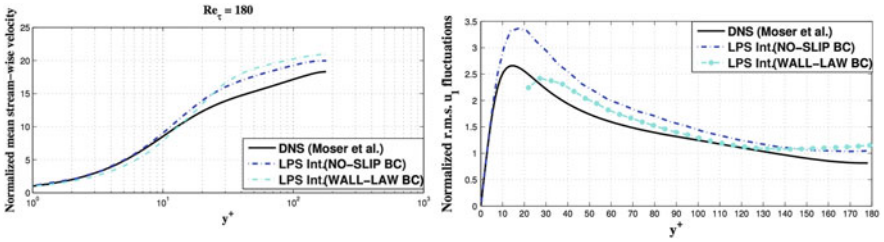


Fig. 2 Example 3.2. Normalized mean stream-wise velocity profiles (*left*) and r.m.s. velocity fluctuations (*right*) in wall coordinates

The numerical studies performed in the present paper indicate that the considered LPS method (improved with wall laws for wall-bounded flows) gives quite good results for both first and second-order statistics with relatively coarse basic discretizations, and thus might be a useful tool in the challenging simulation of turbulent flows, providing reliable numerical results with a comparatively small computational complexity, which is an extremely important feature in the context of realistic applications in Computational Fluid Dynamics (CFD).

4 Smagorinsky Reduced Basis Model

In this section, we present a reduced basis Smagorinsky turbulence model. The idea supporting the reduced basis method is to build a *reduced* basis formed by few number of solutions from the original problem for some values of the parameter, in the *offline* phase. Then the problem is solved by a Galerkin projection onto the space $\mathbf{X}_N \times \mathbb{M}_N$ spanned by the reduced basis, in the *online* phase.

In our case, the parameter that we consider is the Reynolds number μ, assumed to range in a compact interval $\mathscr{D} \subset \mathbb{R}$. The reduced Smagorinsky model is given by:

Given $\mu \in \mathscr{D}$ *find* $(\mathbf{u}_N(\mu), p_N(\mu)) \in \mathbf{X}_N \times \mathbb{M}_N$ *such that:*

$$\begin{cases} \dfrac{1}{\mu}(\nabla \mathbf{u}_N, \nabla \mathbf{v}_N)_\Omega - (p_N, \nabla \cdot \mathbf{v}_N)_\Omega + (\nabla \cdot \mathbf{u}_N, q_N)_\Omega \\ + (\mathbf{u}_N \cdot \nabla \mathbf{u}_N, vN)_\Omega + (\nu_T(\mathbf{u}_N)\nabla \mathbf{u}_N, \nabla \mathbf{v}_N)_\Omega = \langle \mathbf{f}, \mathbf{v}_N \rangle, \end{cases} \tag{8}$$

for any $(\mathbf{v}_N, q_N) \in \mathbf{X}_N \times \mathbb{M}_N$. The eddy viscosity in the Smagorinsky model is defined as $\nu_T(\mathbf{u}_N) = (C_S h_K)^2 \left| \nabla \mathbf{u}_{N|_K}(x) \right|$, where $\left| \cdot \right|$ denotes the Frobenius norm and C_S is the Smagorinsky constant. To linearise this non-linear term, we use the Empirical Interpolation Method (c.f. [9]). This allows a large speed-up in the solution of problem (8).

In the offline phase, to construct the reduced spaces we use the Greedy Algorithm. For the startup of the greedy algorithm we choose an arbitrary parameter value $\mu^1 \in \mathscr{D}$, and compute $(\mathbf{u}_h(\mu^1), p_h(\mu^1))$. The Greedy algorithm recursively

chooses the following snapshot for the reduced space as the value of the parameter that yields the maximum error between the finite element solution and the reduced basis solution. As the computation of the exact error is quite expensive, we have built an a posteriori error bound estimator. In this way, the reduced spaces for velocity and pressure are defined as

$$\mathbb{M}_N = \text{span}\{\xi_k^p := p_h(\mu^k), k = 1, \ldots, N\}, \tag{9}$$

$$\mathbf{X}_N = \text{span}\{\zeta_k^{\mathbf{v}} := \mathbf{u}_h(\mu^k), T_p^{\mu}\xi_k^p, k = 1, \ldots, N\}; \tag{10}$$

Here, T_p^{μ} is the so-called inner pressure *supremizer* operator $T_p^{\mu} : \mathbb{M}_h \to \mathbf{X}_h$, defined as $\left(T_p^{\mu} p_h, \mathbf{v}_h\right) = -(p_h, \nabla \cdot \mathbf{v}_h)_{\Omega}, \forall \mathbf{v}_h \in \mathbf{X}_h$, where $(\cdot, \cdot)$ is a scalar product in the velocity space $\mathbf{H}_0^1(\Omega)$. Adding the supremizers of the pressures to the reduced velocity space $\mathbf{X}_N$ ensures the stability of the discretization of the pressure in problem (8).

4.1 Numerical Results

In this section we present the speedup of the solution of the Navier-Stokes equations provided by the reduced basis problem (8). We consider the backward-facing step problem, with Reynolds number ranging in $\mu \in [50, 450]$. In the offline phase, we compute the FE approximation with Taylor-Hood finite element. The mesh selected for this test is composed by 10842 triangles and 5.703 nodes. The FE steady state solution is computed through a semi-implicit evolution approach.

We start the Greedy algorithm for $\mu = 50$, and we need $N = 17$ snapshots until reaching the Greedy tolerance of $7{\cdot}10^{-5}$ for the a posteriori error bound estimator. In Table 1, we show the results obtained for several values of μ in $\mathscr{D}$. The FE problem has 54.439 degrees of freedom, while the RB problem has 51 degrees of freedom, plus 75 degrees of freedom for the Empirical Interpolation Method. We observe a dramatic reduction of the computational time, over 1.000 for large *Re*, with errors below the Greedy tolerance.

Table 1 Computational time for FE solution and RB online phase, with the speedup and the relative error

Data	$\mu = 56$	$\mu = 132$	$\mu = 236$	$\mu = 320$
T_{FE}	152.7 s	508.7 s	991.9 s	1929.1 s
T_{online}	1.37 s	1.55 s	1.60 s	1.60 s
Speedup	111	326	626	1204
$\|\mathbf{u}_h - \mathbf{u}_N\|/\|\mathbf{u}_h\|$	$1.67 \cdot 10^{-7}$	$2.30 \cdot 10^{-6}$	$2.76 \cdot 10^{-6}$	$5.60 \cdot 10^{-6}$
$\|p_h - p_N\|/\|p_h\|$	$7.92 \cdot 10^{-8}$	$8.57 \cdot 10^{-7}$	$7.13 \cdot 10^{-6}$	$1.98 \cdot 10^{-5}$

Acknowledgements The research of Tomás Chacón Rebollo and Macarena Gómez Mármol has been partially supported by the Spanish Government Project MTM2015-64577-C2-1-R. The research of Samuele Rubino has been partially supported by the Junta de Andalucía Excellence Project P12-FQM-454. Samuele Rubino would also gratefully acknowledge the financial support received from ERC Project H2020-EU.1.1.-639227 during his postdoctoral research involved in this article.

References

1. Ahmed, N., Chacón Rebollo, T., John, V., Rubino, S.: A review of variational multiscale methods for the simulation of turbulent incompressible flows. Arch. Comput. Methods Eng. **24**, 115–164 (2015). doi:10.1007/s11831-015-9161-0
2. Ahmed, N., Chacón Rebollo, T., John, V., Rubino, S.: Analysis of a full space-time discretization of the Navier-Stokes equations by a local projection stabilization method. Preprint 2166, WIAS, Berlin. IMA J. Numer. Anal. **37**, 1437–1467 (2016). doi:10.1093/imanum/drw048
3. Braack, M., Burman, E.: Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. SIAM J. Numer. Anal. **43**(6), 2544–2566 (2006)
4. Chacón Rebollo, T.: A term by term stabilization algorithm for finite element solution of incompressible flow problems. Numer. Math. **79**(2), 283–319 (1998)
5. Chacón Rebollo, T., Gómez Mármol, M., Girault, V., Sánchez Muñoz, I.: A high order term-by-term stabilization solver for incompressible flow problems. IMA J. Numer. Anal. **33**(3), 974–1007 (2013)
6. Chacón Rebollo, T., Gómez Mármol, M., Restelli, M.: Numerical analysis of penalty stabilized finite element discretizations of evolution Navier-Stokes equation. J. Sci. Comput. **61**(1), 1–28 (2014)
7. Gravemeier, V.: Scale-separating operators for variational multiscale large eddy simulation of turbulent flows. J. Comput. Phys. **212**(2), 400–435 (2006)
8. Gravemeier, V., Wall, W.A., Ramm, E.: Large eddy simulation of turbulent incompressible flows by a three-level finite element method. Int. J. Numer. Methods Fluids **48**(10), 1067–1099 (2005)
9. Grelp, M.A., Maday, Y., Nguyen, N.C., Patera, A.T.: Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. ESAIM: Math. Model. Numer. Anal. **41**(3), 575–605 (2007)
10. John, V., Kindl, A.: Numerical studies of finite element variational multiscale methods for turbulent flow simulations. Comput. Methods Appl. Mech. Eng. **199**(13–16), 841–852 (2010)
11. Knobloch, P., Lube, G.: Local projection stabilization for advection-diffusion-reaction problems: one-level vs. two-level approach. Appl. Numer. Math. **59**(12), 2891–2907 (2009)
12. Moser, R., Kim, J., Mansour, N.N.: Direct numerical simulation of turbulent channel flow up to $Re_\tau = 590$. Phys. Fluids **11**(4), 943–945 (1999)
13. Scott, R.L., Zhang, S.: Finite element interpolation of non-smooth functions satisfying boundary conditions. Math. Comput. **54**(190), 483–493 (1990)

Classification of Codimension-One Bifurcations in a Symmetric Laser System

Juancho A. Collera

Abstract We consider a class symmetric laser system, particularly rings of n identical semiconductor lasers coupled bidirectionally. The model is described using the Lang-Kobayashi rate equations where the finite propagation time of the light from one laser to another is reflected by a constant delay time parameter in the laser optical fields. Due to the network structure, the resulting system of delay differential equations has symmetry group $\mathbb{D}_n \times \mathbf{S}^1$. We follow the discussions in Buono and Collera (SIAM J Appl Dyn Syst 14:1868–1898, 2015) for the general case, then give the important example where $n = 4$ to explicitly illustrate the general method which is a classification of codimension-one bifurcations into regular and symmetry-breaking. This particular example complements the recent works on rings with unidirectional coupling (Domogo and Collera, AIP Conf Proc 1787:080002, 2016), and on laser networks with all-to-all coupling (Collera, Mathematical and computational approaches in advancing modern science and engineering. Springer, Cham, 2016). Numerical continuation using DDE-Biftool are carried out to corroborate our classification results.

1 Introduction

In the early eighties, Lang and Kobayashi published a paper which studied the influences of external optical feedback on the properties of a semiconductor laser [9]. In their study, a portion of the laser output is reflected back to the laser cavity from an external reflector. The non-dimensionalized form of the so-called *Lang-Kobayashi (LK) rate equations* was derived by Alsing et al. in [1], where the feedback time is reflected by a delay parameter. The LK rate equations were also used by Erzgräber et al. in [7] in examining the dynamics of two mutually delay-coupled lasers with detuning. In their study, each laser instead receives a feedback light from the other laser. For the case where the detuning parameter

J.A. Collera (✉)
Department of Mathematics and Computer Science, University of the Philippines Baguio, Gov. Pack Road, Baguio City 2600, Philippines
e-mail: jacollera@up.edu.ph

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_88

is zero, the system of equations is equivariant under some group. This symmetry property can then be used to find symmetric solutions and identify symmetry-breaking bifurcations along branches of symmetric solutions [3]. Rings of n identical semiconductor lasers coupled unidirectionally and bidirectionally were also studied from this symmetry point of view [2], including case studies for $n = 3$ and $n = 8$.

In the following sections, we consider a class symmetric laser system, particularly rings of identical semiconductor lasers with bidirectional coupling. We follow the discussions in [2] for the general case, then give the important example where $n = 4$ to explicitly illustrate the general method. This particular example complements the recent works on rings with unidirectional coupling [5], and on laser networks with all-to-all coupling [4].

2 Symmetric Laser Systems

The Lang-Kobayashi rate equations for the symmetric ring of n identical semiconductor lasers with bidirectional coupling is given by the following system

$$\begin{aligned} \dot{E}_j(t) &= (1 + i\alpha)N_j(t)E_j(t) + \kappa e^{-iC_p}[E_{j-1}(t-\tau) + E_{j+1}(t-\tau)], \\ T\dot{N}_j(t) &= P - N_j(t) - (1 + 2N_j(t))|E_j(t)|^2, \end{aligned} \tag{1}$$

for $j = 1, 2, 3, \ldots, n \pmod n$. Here, $E_j(t)$ and $N_j(t)$ are the complex electrical field and the excess carrier number corresponding to the jth laser, respectively, with linewidth enhancement factor α, coupling strength κ, coupling time τ, coupling phase C_p, electron decay rate T, and pump parameter P. The basic solutions of (1) are rotating waves, called *compound laser modes (CLMs)*, and are of the form $E_j(t) = R_j e^{i\omega t + i\sigma_j}$ and $N_j(t) = N_j$, where for all j, we have $\omega, \sigma_j, R_j, N_j \in \mathbb{R}$ with $\sigma_1 = 0$ and $R_j > 0$.

Our task now is to find symmetric solutions to system (1), that is, CLMs fixed by subgroups of the system's symmetry group. We first give some basic notations and definitions. Let $r > 0$ and denote by $\mathscr{C} = C([-r, 0], \mathbb{R}^n)$ the space of continuous functions mapping the interval $[-r, 0]$ to $\mathbb{R}^n$. For $x_t \in \mathscr{C}$, we use the notation $x_t(\tau) = x(t + \tau)$ for $-r \le \tau \le 0$.

Definition 1 For continuously differentiable mapping f, the system $\dot{x}(t) = f(x_t)$ is *equivariant with respect to a group G* if $f(g \cdot x_t) = g \cdot f(x_t)$ for all $g \in G$.
Equivalently, Definition 1 means that if $x(t)$ is a solution to the system $\dot{x}(t) = f(x_t)$, then so thus $g \cdot x(t)$ for all $g \in G$.

Theorem 1 *System (1) is equivariant with respect to the group $\mathbb{D}_n \times S^1$.*

Proof First, we define the action of the diheral group $\mathbb{D}_n = \langle \gamma, \mu \rangle$ where $\gamma^n = 1$ and $\mu^2 = 1$, and the circle group S^1 with elements ϑ, to the state variables $(E_j(t), N_j(t))$ for $j = 1, 2, 3, \ldots, n$ as follows: $\gamma \cdot (E_j(t), N_j(t)) = (E_{j+1}(t), N_{j+1}(t))$,

$\mu \cdot (E_j(t), N_j(t)) = (E_{n-j+1}(t), N_{n-j+1}(t))$, and $\vartheta \cdot (E_j(t), N_j(t)) = (E_j(t)e^{i\vartheta}, N_j(t))$. The order-$n$ element γ of $\mathbb{D}_n$ permutes the lasers in a cyclic manner while the order-two element μ of $\mathbb{D}_n$ reverses the order of the lasers. Moreover, observe that the element ϑ of the circle group S^1 acts only on the complex optical field state variables $E_j(t)$. One can easily check that if $(E_j(t), N_j(t))$ is a solution to (1) then so are $\gamma \cdot (E_j(t), N_j(t))$, $\mu \cdot (E_j(t), N_j(t))$, and $\vartheta \cdot (E_j(t), N_j(t))$. This equivariance property is a consequence of the coupling structure and the rotational symmetry on the complex electric field $E_j(t)$. Thus, system (1) has symmetry group $\mathbb{D}_n \times S^1$. □

We are interested with CLMs are that fixed by subgroups of $\mathbb{D}_n \times S^1$ or the so-called *isotropy subgroups*. To find such CLMs, we need the following from [8].

Definition 2 Let Γ be a group with subgroup H, and let $\vartheta : H \to S^1$ be a homomorphism. We call $H^{\vartheta} = \{(h, \vartheta(h)) \in \Gamma \times S^1 \mid h \in H\}$ a *twisted subgroup* of $\Gamma \times S^1$.

It was shown in [8] that all proper isotropy subgroups of $\Gamma \times S^1$ are twisted subgroups. This result provides us a means to find isotropy subgroups of $\mathbb{D}_n \times S^1$.

For the case where $n = 4$, one such isotropy subgroup of $\mathbb{D}_4 \times S^1$ is the subgroup $\mathbb{D}_4((\gamma, 0), (\mu, 0))$. This isotropy subgroup is generated by the elements $(\gamma, 0)$ and $(\mu, 0)$ of $\mathbb{D}_4 \times S^1$, and is isomorphic to $\mathbb{D}_4$. Notice that CLMs fixed by this subgroup have phase shifts $\sigma_j = \sigma_1 = 0$ for all j. Moreover, these CLMs also have $R_j = R_1$ and $N_j = N_1$ for all j. In general, we call such CLMs as the *fully symmetric CLMs* and has the following form

$$E_j(t) = Re^{i\omega t} \qquad \text{and} \qquad N_j(t) = N, \tag{2}$$

for $j = 1, 2, 3, \ldots, n$. The simpler form of CLMs in (2) allows us to easily compute them even in the general case.

Theorem 2 *Solutions to system (1) of the form given in (2) are obtained by solving for ω in the following transcendental equation*

$$\omega + 2\,\kappa\sqrt{1+\alpha^2}\,\sin(C_p + \omega\tau + \arctan(\alpha)) = 0. \tag{3}$$

Proof First, substitute the ansatz in (2) to system (1) and then split the equation for $E_j(t)$ into real and imaginary parts. These equations yield Eq. (3) after using the identity $\alpha\cos\theta + \sin\theta = \sqrt{1+\alpha^2}\,\sin(\theta + \arctan\alpha)$. We then solve for ω in (3), and then compute for $N = \omega/\left(\alpha + \tan(C_p + \omega\tau)\right)$ and $R = \sqrt{(P-N)\,(1+2N)}$. Each set of values (ω, N, R) gives the desired fully symmetric CLM in (2). □

Example 1 For $n = 4$ and parameters $\alpha = 2.5$, $T = 392$, $P = 0.23$, $\tau = 20$, $\kappa = 0.1$, and $C_p = 10$, solving (3) yields seven fully symmetric CLMs shown in Fig. 1 as dots. A branch of fully symmetric CLMs (ellipse) is also shown in Fig. 1 and is obtained using a numerical continuation tool [6] varying the parameter C_p.

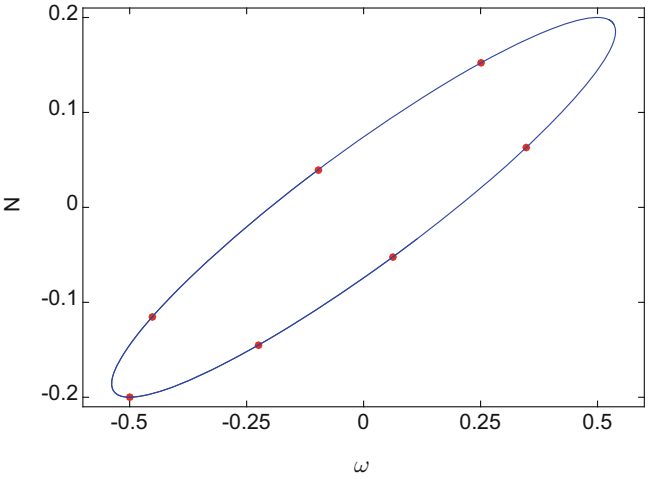


Fig. 1 A branch of fully symmetric CLMs (*ellipse*) obtained by following any of the seven fully symmetric CLMs (shown as *dots*) in DDE-Biftool varying the coupling phase parameter C_p

3 Classification of Codimension-One Bifurcations

This section provides a method of identifying steady-states and Hopf bifurcations along branches of symmetric CLMs, and at the same time classify these codimension-one bifurcations into regular and symmetry-breaking (SB). The identification of SB bifurcations is important as new branches of solutions may emanate from these SB bifurcations. Again, we use the case where $n = 4$ to illustrate our method of classification.

System (1) can be written in polar form using the transformation $E_j(t) = R_j(t)e^{i\phi(t)}$ and $N_j(t) = N_j(t)$ for $j = 1, 2, \ldots, n$ [10]. If we let $X_j(t) = [R_j(t), \phi_j(t), N_j(t)]^T$ and $Y_j(t) = X_j(t-\tau)$ for all j, then the rate equations for the jth laser take the form $\dot{X}_j(t) = f(X_j(t), Y_{j-1}(t), Y_{j+1}(t))$, while the full system can be expressed as $\dot{X}(t) = F(X(t), Y(t))$ with $X(t) = [X_1(t), \ldots, X_n(t)]^T$ and $Y(t) = [Y_1(t), \ldots, Y_n(t)]^T$. The polar form of the fully symmetric CLMs (2) is given by $X^* = [X_1^*, \ldots, X_n^*]^T$ where $X_j^* = [R, \omega t, N]^T$ for all j. The characteristic equation of the linearized system corresponding to the full system about X^* is $\det \Delta(\lambda) = 0$ with $\Delta(\lambda) = \lambda I_{12} - M_1 - e^{-\lambda\tau}M_2$ and where the matrices M_1 and M_2 are obtained from the Jacobian matrix $dF(X^*) = [M_1 \mid M_2]$. For the case where $n = 4$,

$$\Delta(\lambda) = \begin{bmatrix} A & B & 0 & B \\ B & A & B & 0 \\ 0 & B & A & B \\ B & 0 & B & A \end{bmatrix}$$

where $A = \lambda I_3 - \bar{A}$ with $\bar{A} = d_{X_j(t)} f(X_j^*)$ and $B = -e^{-\lambda\tau}\bar{B}$ with $\bar{B} = d_{Y_k(t)} f(X_j^*)$. This matrix $\Delta(\lambda)$ can be expressed in the following block diagonal form [8]

$$\begin{bmatrix} A+2B & 0 & 0 & 0 \\ 0 & A-2B & 0 & 0 \\ 0 & 0 & A & 0 \\ 0 & 0 & 0 & A \end{bmatrix}.$$

The key to this diagonalization technique is examining the action of $L := \Delta(\lambda)$ on the subspaces $V_k = \{[v, \zeta^k v, \zeta^{2k} v, \zeta^{3k} v]^T \mid v \in \mathbb{R}^3\}$, for $k = 0, 1, 2, 3$, where $\zeta = e^{\pi i/2}$. For example, the action of L on $V_0 = \{[v, v, v, v]^T \mid v \in \mathbb{R}^3\}$ is given by $L[v, v, v, v]^T = [(A+2B)v, (A+2B)v, (A+2B)v, (A+2B)v]^T$. Consequently, the eigenvalues coming from $L|_{V_0}$ are those coming from $(A + 2B)$. Since the action of the isotropy subgroup on V_0 is trivial, only regular bifurcations are obtained from the block $(A+2B)$. In contrast, the action of the isotropy subgroup on the remaining subspaces is nontrivial. Thus, symmetry-breaking bifurcations are obtained from the rest of the diagonal blocks, namely $(A - 2B)$ and A.

Theorem 3 *Regular bifurcations are obtained from the diagonal block* $(A + 2B)$ *while symmetry-breaking bifurcations are obtained from the blocks* $(A - 2B)$ *and* A.

We conclude this section by illustrating Theorem 3. Symmetry properties of branches of solutions emanating from SB bifurcations are then examined to corroborate our results.

Example 2 We first compute for determinants $\det(A \pm 2B)$ which is equal to $(\lambda + C_1)[\lambda^2 + 4c\lambda + 4\kappa^2 \mp 4e^{-\lambda\tau}(c\lambda + 2\kappa^2) + 4\kappa^2 e^{-2\lambda\tau}] + RC_2[\lambda + 2\rho(1 \mp e^{-\lambda\tau})]$ and $\det(A) = (\lambda + C_1)(\lambda^2 + 4c\lambda + 4\kappa^2) + RC_2(\lambda + 2\rho)$, where $C_1 = (1 + 2R^2)/T$, $C_2 = 2(1 + 2N)R/T$, $s = \kappa \sin(C_p + \omega\tau)$, $c = \kappa\cos(C_p + \omega\tau)$, and $\rho = -\alpha s + c$. Pitchfork bifurcations are identified by intersecting the graph of Eq. (3) which gives the branch of fully symmetric CLMs with the graph of $\det(M)|_{\lambda=0} = 0$, where $M = (A - 2B)$ or A, which gives the SB bifurcations. Figure 2a shows the plot in the $C_p\omega$-plane of the curves given by (3) and $\det(A - 2B)|_{\lambda=0} = 0$. There are two intersections modulo 2π in C_p. These points are shown in Fig. 3a and are marked with (cross symbol). Meanwhile, in Fig. 2b, we plot of the curves given by (3) and $\det(A)|_{\lambda=0} = 0$. Again, we obtained two intersections modulo 2π in C_p. These points are shown in Fig. 3a with the marker (open diamond). Saddle-node bifurcations are identified by intersecting the graph of Eq. (3) with the graph of $\det(A + 2B)|_{\lambda=0} = 0$ which gives the regular bifurcations. However, the left-hand side of the later equation vanishes. So instead, we look for the intersections of the graph of (3) and the graph of $\frac{d}{d\lambda}\det(A + 2B)|_{\lambda=0} = 0$ as shown in Fig. 2c. There are two common points. These saddle-node bifurcation points are shown in Fig. 3a using the marker (plus symbol). Similarly, regular Hopf bifurcations, shown in Fig. 3b as (asterisk symbol), are identified by the intersections of (3) and $\det(A + 2B)|_{\lambda=i\beta} = 0$, while SB Hopf bifurcations, shown in Fig. 3b as

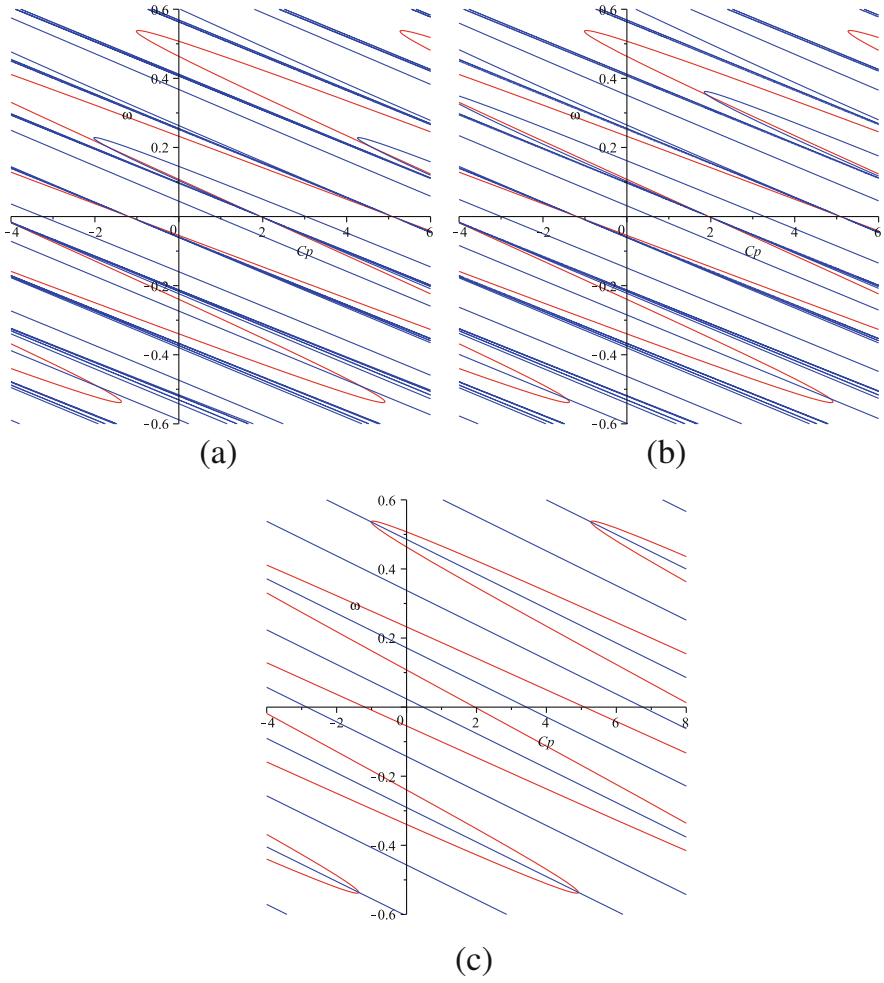


Fig. 2 Plots showing the intersection of the graph of the transcendental equation in (3) with the graph of (**a**) $\det(A-2B)|_{\lambda=0}=0$, (**b**) $\det(A)|_{\lambda=0}=0$, and (**c**) $\frac{d}{d\lambda}\det(A+2B)|_{\lambda=0}=0$

(open circle) (resp. as open square), are identified by the intersections of (3) and $\det(A-2B)|_{\lambda=i\beta}=0$ (resp. $\det(A)|_{\lambda=i\beta}=0$).

Following the pitchfork bifurcations (cross symbol) in DDE-Biftool, again varying the parameter C_p, we obtain a new branch of CLMs shown in Fig. 3a with symmetry group $\mathbb{D}_2(\gamma^2,0)$. Indeed, in this case, symmetry is broken from $\mathbb{D}_4((\gamma,0),(\mu,0))$ of the original branch of fully symmetric CLMs to $\mathbb{D}_2(\gamma^2,0)$ of the new branch of CLMs emanating from the identified pitchfork bifurcations. Now, following each Hopf bifurcation in DDE-Biftool varying C_p gives branches of periodic solutions with different periodic patterns as shown in Fig. 4.

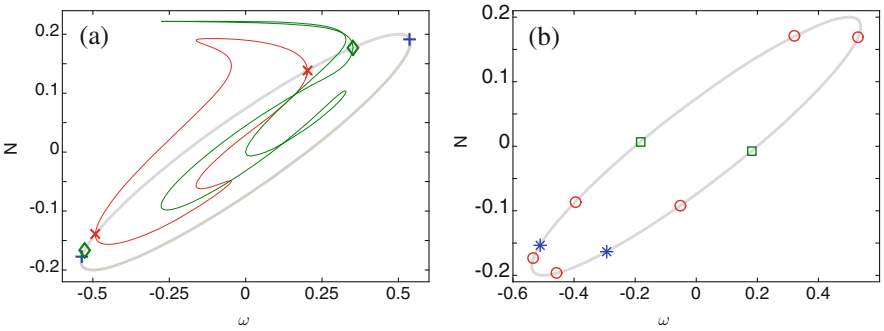


Fig. 3 A branch of CLMs emanates from a pitchfork bifurcation

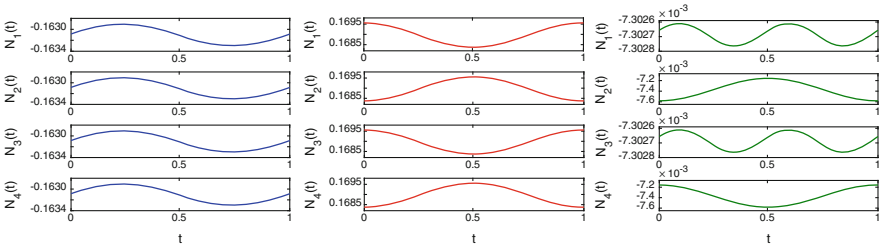


Fig. 4 Time series showing the periodic patterns for branches of periodic solutions emanating from Hopf bifurcations: (*left*) from block $(A + 2B)$ where symmetry is preserved, and (*middle*) from block $(A - 2B)$ and (*right*) from block A where symmetry is broken

4 Conclusions

This paper considered a class of symmetric laser system, particularly rings of semiconductor lasers with bidirectional coupling. The important case where $n = 4$, which complements the recent works on laser networks with unidirectional and all-to-all coupling, was used to illustrate the results of classifying codimension-one bifurcations. New branches of solutions from the SB bifurcations are obtained using numerical continuation whose symmetry corroborate our classification results.

Acknowledgements This work was funded by the UP System Emerging Interdisciplinary Research Program (OVPAA-EIDR-C03-011). The author also acknowledged the support of the UP Baguio through RLCs during the A.Y. 2015–2016.

References

1. Alsing, P.M., Kovanis, V., Gavrielides, A., Erneux, T.: Lang and Kobayashi phase equation. Phys. Rev. A **53**, 4429–4434 (1996)
2. Buono, P.-L., Collera, J.A.: Symmetry-breaking bifurcations in rings of delay-coupled semiconductor lasers. SIAM J. Appl. Dyn. Syst. **14**, 1868–1898 (2015)

3. Collera, J.A.: Symmetry-breaking bifurcations in two mutually delay-coupled lasers. Int. J. Philipp. Sci. Technol. **8**, 17–21 (2015)
4. Collera, J.A.: Symmetry-breaking bifurcations in laser systems with all-to-all coupling. In: Belair, J. et al., (eds.) Mathematical and Computational Approaches in Advancing Modern Science and Engineering, pp. 81–88. Springer, Cham (2016)
5. Domogo, A.A., Collera, J.A.: Classification of codimension-one bifurcations in a tetrad of delay-coupled lasers with feed forward coupling. AIP Conf. Proc. **1787**, 080002 (2016)
6. Engelborghs, K., Luzyanina, T., Samaey, G.: DDE-BIFTOOL v. 2.00: a Matlab package for bifurcation analysis of delay differential equations. Technical report TW-330. Department of Computer Science, K.U. Leuven, Leuven (2001)
7. Erzgräber, H., Krauskopf, B., Lenstra, D.: Compound laser modes of mutually delay-coupled lasers. SIAM J. Appl. Dyn. Syst. **5**, 30–65 (2006)
8. Golubitsky, M., Schaeffer, D., Stewart, I.: Singularities and Groups in Bifurcation Theory Volume II. Springer, New York (1988)
9. Lang, R., Kobayashi, K.: External optical feedback effects on semiconductor injection laser properties. IEEE J. Quantum Electron. **16**, 347–355 (1980)
10. Lunel, S.M.V., Krauskopf, B.: The mathematics of delay equations with an application to the Lang-Kobayashi equations. AIP Conf. Proc. **548**, 66–86 (2000)

Approximating a Special Class of Linear Fourth-Order Ordinary Differential Problems

Emilio Defez, Michael M. Tung, J. Javier Ibáñez, and Jorge Sastre

Abstract Differential matrix models are an essential ingredient of many important scientific and engineering applications. In this work, we propose a procedure to approximate the solutions of special linear fourth-order matrix differential problems of the type $Y^{(4)}(x) = A(x)Y(x) + B(x)$ with higher-order matrix splines. An example is included.

1 Introduction

This work presents a spline method for the approximation of a special class of fourth-order ordinary differential equations of the following form

$$Y^{(4)} = A(x)Y(x) + B(x), \quad A(x) \in \mathbb{C}^{r\times r}, \quad Y(x) \in \mathbb{C}^{r\times q}, \quad x \in [a,b], \tag{1}$$

with initial conditions $Y(a) = Y_a, Y'(a) = Y'_a, Y''(a) = Y''_a$ and $Y^{(3)}(a) = Y^{(3)}_a$, in which the first, second, and third derivatives do not appear explicitly. Routinely, fourth-order ordinary differential equations are transformed to a first-order system of ordinary differential equations with the hindsight that standard numerical methods may be applied. However, this technique comes with an increase of the computational cost due to the increase of size of the problem. On an alternative pathway, integration methods have attracted the attention of several

E. Defez • M.M. Tung (✉)
Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera, s/n, E-46022 Valencia, Spain
e-mail: edefez@imm.upv.es; mtung@mat.upv.es

J.J. Ibáñez
Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València, Camino de Vera, s/n, E-46022 Valencia, Spain
e-mail: jjibanez@dsic.upv.es

J. Sastre
Instituto de Telecomunicaciones y Aplicaciones Multimedia, Universitat Politècnica de València, Camino de Vera, s/n, E-46022 Valencia, Spain
e-mail: jorsasma@iteam.upv.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_89

authors for solving higher-order matrix equations of type (1). These direct methods have demonstrated favourable features in accuracy and speed, see e.g. [2, 4, 6, 7] and references therein. Here we elaborate an algorithm to deal with matrix differential equations of the fourth order (1). Throughout this work we will adopt the conventional notation for norms and matrix splines as in previous publications, see e.g. [1]. We denote by $I_{r\times r}$ and $\|A\|$ the identity matrix of dimension r and any multiplicative norm of the matrix A, respectively.

This paper is organized as follows: Sect. 2 introduces the proposed method and describes all the algorithmic details of the procedure. Section 3 concludes the discussion with a numerical example.

2 Description of the Method

Let us consider the fourth-order matrix problem

$$Y^{(4)}(x) = A(x)Y(x) + B(x), \quad a \le x \le b, \tag{2}$$

where $Y(x) \in \mathbb{C}^{r\times q}$ is the unknown matrix with initial conditions $Y(a) = Y_a, Y'(a) = Y'_a, Y''(a) = Y''_a, Y^{(3)}(a) = Y^{(3)}_a \in \mathbb{C}^{r\times q}$. The matrix-valued functions $A : [a,b] \to \mathbb{C}^{r\times r}$ and $B : [a,b] \to \mathbb{C}^{r\times q}$ are of differentiability class $A, B \in \mathscr{C}^s(I), s \ge 1, I = (a,b)$. We consider the following partition of the interval $[a,b]$:

$$\Delta_{[a,b]} = \{a = x_0 < x_1 < \ldots < x_n = b\}, \; x_k = a + kh, \; k = 0, 1, \ldots, n, \tag{3}$$

where n is a positive integer with the corresponding step size $h = (b-a)/n$. We will construct in each subinterval $[a+kh, a+(k+1)h]$ a matrix spline $S(x)$ of order $m \in \mathbb{N}$ with $4 \le m \le s$, where s is the order of the differentiability class of A and B. This will approximate the solution of problem (2) so that $S(x) \in \mathscr{C}^4([a,b])$. In the first interval $[a, a+h]$, we define the matrix spline as

$$\begin{aligned} S_{|_{[a,a+h]}}(x) &= Y(a) + Y'(a)(x-a) + \frac{1}{2!}Y''(a)(x-a)^2 + \frac{1}{3!}Y^{(3)}(a)(x-a)^3 \\ &\quad + \cdots + \frac{1}{(m-1)!}Y^{(m-1)}(a)(x-a)^{m-1} + \frac{1}{m!}A_0(x-a)^m, \end{aligned} \tag{4}$$

where $A_0 \in \mathbb{C}^{r\times q}$ is a matrix parameter to be determined. It is straightforward to check

$$S_{|_{[a,a+h]}}(a) = Y_a, \quad S'_{|_{[a,a+h]}}(a) = Y'_a, \quad S''_{|_{[a,a+h]}}(a) = Y''_a, \quad S'''_{|_{[a,a+h]}}(a) = Y'''_a,$$

and

$$S^{(4)}_{|_{[a,a+h]}}(a) = Y^{(4)}(a) = A(a)Y(a) + B(a)$$

and therefore the spline (4) satisfies the differential equation (2) at $x = a$.

Next we must obtain the values $Y^{(5)}(a), Y^{(6)}(a), \ldots, Y^{(m-1)}(a)$, and A_0 in order to determine the matrix spline (4). It is straightforward to obtain

$$Y^{(5)}(x) = A'(x)Y(x) + A(x)Y'(x) + B'(x) = g_1\left(x, Y(x), Y'(x)\right), \tag{5}$$

where $g_1 \in \mathscr{C}^{s-1}(I)$. This enables us to evaluate $Y^{(5)}(a)$ as

$$Y^{(5)}(a) = g_1\left(a, Y(a), Y'(a)\right) = g_1\left(a, Y_a, Y'_a\right),$$

using (5). Similarly, we can assume that $A, B \in \mathscr{C}^s(T)$ for $s \geq 2$. Then, the second derivatives of A and B exist and are continuous This gives the sixth derivative $Y^{(6)}(x)$:

$$\begin{aligned} Y^{(6)}(x) &= A''(x)Y(x) + 2A'(x)Y'(x) + A(x)Y''(x) + B''(x) \\ &= g_2\left(x, Y(x), Y'(x), Y''(x)\right) \in \mathscr{C}^{s-2}(I). \end{aligned} \tag{6}$$

Finally we can compute $Y^{(6)}(a) = g_2\left(a, Y(a), Y'(a), Y''(a)\right) = g_2\left(a, Y_a, Y'_a, Y''_a\right)$ using (6). For all higher-order derivatives $Y^{(7)}(x), \ldots, Y^{(m-1)}(x)$ we proceed analogously to obtain

$$\left.\begin{aligned} Y^{(7)}(x) &= g_3\left(x, Y(x), Y'(x), Y''(x), Y'''(x)\right) \in \mathscr{C}^{s-3}(I) \\ &\vdots \\ Y^{(m-1)}(x) &= g_{m-5}\left(x, Y(x), Y'(x), \ldots, Y^{(m-5)}(x)\right) \in \mathscr{C}^{s-(m-5)}(I) \end{aligned}\right\}. \tag{7}$$

These derivatives can be easily computed with standard computer algebra systems. Substituting $x = a$ in (7), one gets $Y^{(7)}(a), \ldots, Y^{(m-1)}(a)$. Now all matrix parameters of the spline which were to be determined are known, except for A_0. To determine A_0, we suppose that (4) is a solution of problem (2) at $x = a + h$, which gives

$$S^{(4)}_{|_{[a,a+h]}}(a+h) = A(a+h)\, S_{|_{[a,a+h]}}(a+h) + B(a+h). \tag{8}$$

Next, we obtain from (8) the matrix equation with only one unknown A_0:

$$\begin{aligned} A_0 = \frac{(m-4)!}{h^{m-4}} \Bigg[& A(a+h)\left(Y(a) + Y'(a)h + \cdots + \frac{h^{m-1}}{(m-1)!}Y^{(m-1)}(a) + \frac{h^m}{m!}A_0\right) \\ & + B(a+h) - Y^{(4)}(a) - Y^{(5)}(a)h - \cdots - \frac{1}{(m-5)!}Y^{(m-1)}(a)h^{m-5}\Bigg]. \end{aligned} \tag{9}$$

Assuming that the matrix equation (9) has only one solution A_0, the matrix spline (4) is totally determined in the interval $[a, a+h]$. In the next interval $[a+h, a+2h]$, the matrix spline takes the form

$$S_{|_{[a+h,a+2h]}}(x) = \sum_{i=0}^{3} \frac{S^{(i)}_{|_{[a,a+h]}}(a+h)}{i!}(x-(a+h))^i + \sum_{j=4}^{m-1} \frac{\overline{Y^{(j)}(a+h)}}{j!}(x-(a+h))^j + \frac{A_1}{m!}(x-(a+h))^m, \tag{10}$$

where

$$\begin{aligned} \overline{Y^{(4)}(a+h)} &= A\,(a+h)\,S_{|_{[a,a+h]}}(a+h) + B\,(a+h)\,, \\ \overline{Y^{(5)}(a+h)} &= g_1\left(a+h, S_{|_{[a,a+h]}}(a+h), S'_{|_{[a,a+h]}}(a+h)\right), \\ &\vdots \\ \overline{Y^{(m-1)}(a+h)} &= g_{m-5}\left(a+h, S_{|_{[a,a+h]}}(a+h), \ldots, S^{(m-5)}_{|_{[a,a+h]}}(a+h)\right). \end{aligned} \tag{11}$$

The matrix spline $S(x)$ defined by (4) and (10) is of differentiability class $\mathscr{C}^4([a, a+2h])$. By construction, spline (10) satisfies the differential equation (2) at $x = a+h$, and all of its coefficients are determined with the exception of $A_1 \in \mathbb{C}^{r\times q}$. The value of A_1 can be found by taking the spline (10) as a solution of (2) at point $x = a+2h$:

$$S^{(4)}_{|_{[a+h,a+2h]}}(a+2h) = A\,(a+2h)\,S_{|_{[a+h,a+2h]}}(a+2h) + B\,(a+2h)\,.$$

An expansion yields the matrix equation with the only unknown A_1:

$$A_1 = \frac{(m-4)!}{h^{m-4}}\left[A\,(a+2h)\left(\sum_{i=0}^{3}\frac{S^{(i)}_{|_{[a,a+h]}}(a+h)}{i!}h^i + \sum_{j=4}^{m-1}\frac{\overline{Y^{(j)}(a+h)}}{j!}h^j + \frac{A_1 h^m}{m!}\right) + B\,(a+2h) - \overline{Y^{(4)}(a+h)} - \overline{Y^{(5)}(a+h)}h - \cdots - \frac{h^{m-5}}{(m-5)!}\overline{Y^{(m-5)}(a+h)}\right]. \tag{12}$$

Assume the matrix equation (12) has only one solution A_1. This way the spline is totally determined within interval $[a+h, a+2h]$. Iterating this process, we can construct the matrix spline approximation taking $[a+(k-1)h, a+kh]$ as the last subinterval. For the subsequent subinterval $[a+kh, a+(k+1)h]$, we define the

corresponding matrix spline as

$$S_{\big|_{[a+kh,a+(k+1)h]}}(x) = \sum_{i=0}^{3} \frac{S^{(i)}_{\big|_{[a+(k-1)h,a+kh]}}(a+kh)}{i!}(x-(a+kh))^i + \sum_{j=4}^{m-1} \frac{\overline{Y^{(j)}(a+kh)}}{j!}(x-(a+kh))^j + \frac{A_k}{m!}(x-(a+kh))^m,$$

where

$$\begin{aligned}
\overline{Y^{(4)}(a+kh)} &= A\,(a+kh)\,S_{\big|_{[a+(k-1)h,a+kh]}}(a+kh) + B\,(a+kh)\,,\\
\overline{Y^{(5)}(a+kh)} &= g_1\left(a+kh, S_{\big|_{[a+(k-1)h,a+kh]}}(a+kh), S'_{\big|_{[a+(k-1)h,a+kh]}}(a+kh)\right),\\
&\vdots\\
\overline{Y^{(m-1)}(a+kh)} &= g_{m-5}\left(a+kh, S_{\big|_{[a+(k-1)h,a+kh]}}(a+kh),\ldots,S^{(m-5)}_{\big|_{[a+(k-1)h,a+kh]}}(a+kh)\right).
\end{aligned}$$

With this definition, the matrix spline $S(x) \in \mathscr{C}^4([a, a+(k+1)h])$ fulfills the differential equation (2) at point $x = a + kh$. As an additional requirement, we assume that $S_{\big|_{[a+kh,a+(k+1)h]}}(x)$ satisfies (2) at point $x = a + (k+1)h$:

$$S^{(4)}_{\big|_{[a+kh,a+(k+1)h]}}(a+(k+1)h) = A\,(a+(k+1)h)\,S_{\big|_{[a+kh,a+(k+1)h]}}(a+(k+1)h) + B\,(a+(k+1)h)\,,$$

and expanding this expression gives

$$A_k = \frac{(m-4)!}{h^{m-4}}\left[A(a+(k+1)h)\left(\sum_{i=0}^{3}\frac{S^{(i)}_{\big|_{[a+(k-1)h,a+kh]}}(a+kh)}{i!}h^i + \sum_{j=4}^{m-1}\frac{\overline{Y^{(j)}(a+kh)}}{j!}h^j + \frac{A_k}{m!}h^m\right) + B(a+(k+1)h) - \overline{Y^{(4)}(a+kh)} - \cdots - \frac{h^{m-5}}{(m-5)!}\overline{Y^{(m-1)}(a+kh)}\right]. \quad (13)$$

Observe that the final result (13) relates directly to Eqs. (9) and (12), when setting $k = 0$ and $k = 1$. We will demonstrate that these equations have a unique solution.

We rewrite Eq. (13) in the form

$$\left(I_{r\times r}-\frac{h^4A(a+(k+1)h)}{m(m-1)(m-2)(m-3)}\right)A_k$$

$$=\frac{(m-4)!}{h^{m-4}}\left[A(a+(k+1)h)\left(\sum_{i=0}^{3}\frac{S^{(i)}_{\big|_{[a+(k-1)h,a+kh]}}(a+kh)}{i!}h^i+\sum_{j=4}^{m-1}\frac{\overline{Y^{(j)}(a+kh)}}{j!}h^j\right)\right.$$

$$\left.+B\,(a+(k+1)h)-\overline{Y^{(4)}(a+kh)}-\cdots-\frac{h^{m-5}}{(m-5)!}\overline{Y^{(m-1)}(a+kh)}\right]. \tag{14}$$

Note also that solubility of Eq. (14) is guaranteed by showing that the coefficient matrix $\left(I_{r\times r}-\frac{h^4A(a+(k+1)h)}{m(m-1)(m-2)(m-3)}\right)$ is invertible, for $k=0,1,\ldots,n-1$. To see this, let us denote

$$M=\max\{\|A(x)\|\,;\ x\in[a,b]\}. \tag{15}$$

Then, one obtains

$$\left\|I_{r\times r}-\left(I_{r\times r}-\frac{h^4A(a+(k+1)h)}{m(m-1)(m-2)(m-3)}\right)\right\|\leq\frac{h^4M}{m(m-1)(m-2)(m-3)}.$$

Taking

$$0<h<\sqrt[4]{\frac{m(m-1)(m-2)(m-3)}{M}} \tag{16}$$

according to Lemma 2.3.3 in [3, p. 58], it follows that coefficient matrix

$$\left(I_{r\times r}-\frac{h^4A(a+(k+1)h)}{m(m-1)(m-2)(m-3)}\right)$$

is invertible for $0\leq k\leq n-1$. Therefore Eq. (13) has unique solutions A_k for $k=0,1,\ldots,n-1$, and the matrix spline is completely determined. In summary, we have proved the following theorem:

Theorem 1 *For the fourth-order matrix differential equation (2), let M be the constant defined by (15). We also consider the partition (3) with step size h satisfies (16). Then, the matrix spline $S(x)$ of order $m, 4\leq m\leq s$ exists in each subinterval $[a+kh, a+(k+1)h]$, $k=0,1,\ldots,n-1$, as defined in the previous construction and is of class $\mathscr{C}^4([a,b])$.*

Observe that the so constructed splines have a global error of $O(h^{m-1})$, which follows from an analysis similar to Loscalzo and Talbot's work [5].

3 Numerical Example

A suitable benchmark for testing our method is the following scalar problem:

$$y^{(4)}(x) = \left(x^4 - 6x^2 + 3\right) y(x), y(0) = 1, y'(0) = y^{(3)}(0) = 0, y''(0) = -1, 0 \leq x \leq 1 \tag{17}$$

whose known solution is $y(x) = e^{x^2/2}$. Following [2] and (15), we determine constant $M = \max_{x\in[0,1]} \left\{x^4 - 6x^2 + 3\right\} \leq 3$. For splines of the sixth order ($m = 6$), according to theorem 1 we obtain $h < 3.30975$. We choose $n = 10$ partitions and $h = 0.1$, then condition (16) holds. For the symbolic solutions of the algebraic equations which arise from the algorithm, we use software *Mathematica*. The results are summarized in Table 1. In Table 2, we evaluated the difference between the estimates of our numerical approach and the exact solution. The maximum of these errors are indicated for each subinterval.

Table 1 Approximation for problem (17)

Interval	Approximation
$[0, 0.1]$	$1. - 0.5x^2 + 0.125x^4 - 0.0207122x^6$
$[0.1, 0.2]$	$1. - 3.5462 \times 10^{-8}x - 0.499999x^2 - 0.0000166527x^3 + 0.125143x^4 - 0.0006445x^5 - 0.019517x^6$
$[0.2, 0.3]$	$1. - 3.196534 \times 10^{-6}x - 0.499955x^2 - 0.000336338x^3 + 0.126447x^4 - 0.00346416x^5 - 0.0169916x^6$
$[0.3, 0.4]$	$1. - 0.0000427757x - 0.499603x^2 - 0.00200073x^3 + 0.130856x^4 - 0.00967378x^5 - 0.013358x^6$
$[0.4, 0.5]$	$1.00002 - 0.000260775x - 0.498173x^2 - 0.00699231x^3 + 0.140637x^4 - 0.0198766x^5 - 0.00893112x^6$
$[0.5, 0.6]$	$1.00008 - 0.00102047x - 0.494226x^2 - 0.0179163x^3 + 0.157619x^4 - 0.0339389x^5 - 0.00408475x^6$
$[0.6, 0.7]$	$1.00027 - 0.00297411x - 0.48582x^2 - 0.0371872x^3 + 0.182447x^4 - 0.0509828x^5 + 0.000786428x^6$
$[0.7, 0.8]$	$1.00072 - 0.0069686x - 0.471151x^2 - 0.0658931x^3 + 0.214024x^4 - 0.0694961x^5 + 0.00530606x^6$
$[0.8, 0.9]$	$1.00159 - 0.0136859x - 0.449633x^2 - 0.102634x^3 + 0.249291x^4 - 0.0875411x^5 + 0.00915112x^6$
$[0.9, 1.0]$	$1.00296 - 0.0229897x - 0.423195x^2 - 0.142683x^3 + 0.2834x^4 - 0.103027x^5 + 0.0120793x^6$

Table 2 Approximation error for problem (17)

Interval	$[0, 0.1]$	$[0.1, 0.2]$	$[0.2, 0.3]$	$[0.3, 0.4]$	$[0.4, 0.5]$
Max. error	9.5148×10^{-11}	3.0342×10^{-9}	2.3307×10^{-8}	9.9968×10^{-8}	3.0769×10^{-7}
Interval	$[0.5, 0.6]$	$[0.6, 0.7]$	$[0.7, 0.8]$	$[0.8, 0.9]$	$[0.9, 1.0]$
Max. error	7.6477×10^{-7}	1.6374×10^{-6}	3.1402×10^{-6}	5.5327×10^{-6}	9.1126×10^{-6}

Acknowledgements This work has been supported by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF) under grant TIN2014-59294-P.

References

1. Defez, E., Tung, M.M., Ibáñez, J., Sastre, J.: Approximating and computing nonlinear matrix differential models. Math. Comput. Model. **55**(7), 2012–2022 (2012)
2. Famelis, I., Tsitouras, C.: On modifications of Runge–Kutta–Nyström methods for solving $y^{(4)} = f(x, y)$. Appl. Math. Comput. **273**, 726–734 (2016)
3. Golub, G.H., Loan, C.F.V.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore, MD (1996)
4. Hussain, K., Ismail, F., Senu, N.: Two embedded pairs of Runge-Kutta type methods for direct solution of special fourth-order ordinary differential equations. Math. Probl. Eng. **2015** (2015). doi:10.1155/2015/196595
5. Loscalzo, F.R., Talbot, T.D.: Spline function approximations for solutions of ordinary differential equations. SIAM J. Numer. Anal. **4**(3), 433–445 (1967)
6. Olabode, B., et al.: Implicit hybrid block Numerov-type method for the direct solution of fourth-order ordinary differential equations. Am. J. Comput. Appl. Math. **5**(5), 129–139 (2015)
7. Papakostas, S.N., Tsitmidelis, S., Tsitouras, C.: Evolutionary generation of 7^{th} order Runge - Kutta - Nyström type methods for solving $y^{(4)} = f(x, y)$. In: American Institute of Physics Conference Series, vol. 1702 (2015). doi:10.1063/1.4938985

Evaluation of Steel Buildings by Means of Non-destructive Testing Methods

Markus Doktor, Christian Fox, Wolfgang Kurz, and Christina Thein

Abstract Non-destructive testing methods became popular within the last few years. For steel beams incorporated in buildings there are currently only destructive ways for testing the yield limit as well as for determination of the current stress level. Rise of ultrasonic and micro-magnetic tools for (non-destructive) measurements allows the characterization of the inbuild material especially of old steel bridges as economical maintenance of the infrastructure. It is possible to determine the reserve of residuence of bridges or of any other existing steel buildings in order to upgrade them competitively for future usage by the possibility of a simple way of strengthening by welding or using bolds.

1 Construction in Existing Buildings

Higher static loads can easily be supported by steel constructions with simple strengthening measures. But the used steel and its material characteristics, especially his bearing capacity, are not known.

Furthermore, the current state of load is typically unknown. Modifications of buildings in their previous service lifes caused changed load transfers, which are not incorporated in current construction plans. In many cases, construction plans of the building to be investigated do not exist anymore. Thus, it is impossible to make a serious statement about the load bearing capacity still available in the beams.

M. Doktor
Statistics Research Group, University of Kaiserslautern, 67653 Kaiserslautern, Germany
e-mail: doktor@mathematik.uni-kl.de

C. Fox (✉) • W. Kurz • C. Thein
Institute of Steel Structures, University of Kaiserslautern, 67653 Kaiserslautern, Germany
e-mail: christian.fox@bauing.uni-kl.de; wolfgang.kurz@bauing.uni-kl.de;
thein@mathematik.uni-kl.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_90

2 Non-destructive Testing Methods

To determine the residual carrying capacity of a construction without risk of collapse is just possible with non-destructive testing methods. The determination of the yield strength by micromagnetic measurements and the determination of the current load state by ultrasonic measurements are described below.

2.1 Unknown Material Characteristics

The yield strength is the most important material characteristic of the steel to determine the load-bearing capacity. It is defined via regulations to characterize different steel qualities. If the yield limit is determined by measurements, the ultimate limit state design can be proven by real material characteristics without model uncertainties and not by guaranteed minimum values.

For determination of the yield strength of an investigated beam the ferromagnetic properties of steel are used. There is a causal relation between magnetism and the yield limit to be determined, see [6] for details. For calibration means, the magnetic properties of the different steel types are determined and assigned to their yield limit measured in a classical tensile test.

2.2 Determination of the Current Load State

The determination of the current state of stress of a built-in steel beam is possible via ultrasonic measurements, compare [2–4]. The influence of the stress and strain conditions on the speed of the propagation of the ultrasonic waves is used for the ultrasonic stress analysis. Comparing the speed of the ultrasonic waves in a beam without load v_0 and the speed in a beam with load v_l, the current state of stress σ can be determined using the following equation:

$$\sigma \left[\frac{N}{mm^2}\right] = \frac{(v_0 - v_l)}{v_0} \cdot \frac{K}{H} \left[\frac{N}{mm^2}\right] \tag{1}$$

$\frac{K}{H}$ is a linear factor that is determined in lab tests. In these tests the speed v_0 is measured in beams without load. In practice, there is no possibility for a direct measurement as there are no in-built beams without load. Due to the anti-symmetric stress distribution in the cross section of a beam, the speed v_0 can be determined by the mean value of symmetrically distributed measurement points in a loaded beam assuming that there is no axial force (Fig. 1).

The texture due to the rolling process of a steel beam causes the directional property of the ultrasonic measuring results (anisotropism). The difference of the

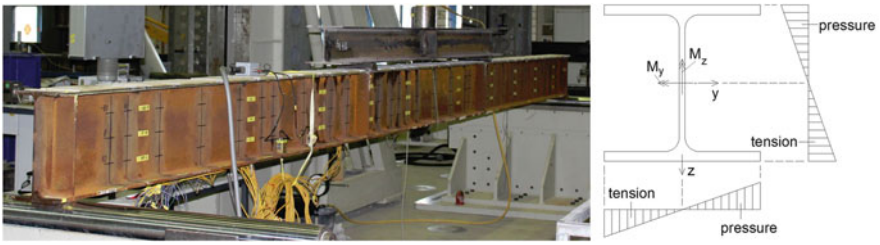


Fig. 1 Beam in lab test and anti-symmetric stress distribution in the cross section

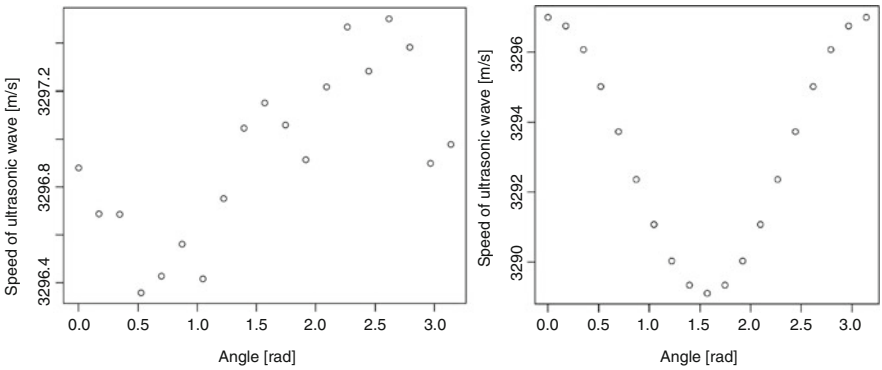


Fig. 2 *Left*: speed variation influenced by stress. *Right*: speed variation due to texture

speed according to the texture is shown in Fig. 2, right and the difference of the speed according to the stress due to loading is shown in Fig. 2, left. Ultrasonic measurement values are the addition of both of them. The influence of texture is up to ten times higher than the influence of stresses. To show a sufficient bearing resistance, the much greater influence of texture has to be eliminated.

3 Mathematical Modeling and Applications

To fit a regression model to the data, we need to develop a stochastic model of the ultrasonic measurements. The time-of-flight of the ultrasonic waves in every point measured should be related to the local stress in the beam. Additionally, the residuals have to be taken into account. Finally, a segmented regression approach based on the statical system and statistical decision criteria, was developed, for details see [2, 5].

3.1 The Regression Model

Due to the statical system we expect the stress curve to show interval-wise different behaviour. The points of change of the (local) regression function, driven by the stress curve, are well known through the statical system. For measured ultrasonic times-of-flight $x_1, \ldots, x_N$, we want to estimate the local stresses $y_1, \ldots, y_N$, i.e. for all $1 \leq j \leq N : y_j = f(x_j) + \epsilon_j$, where ϵ_j is the residual term. The classical minimization problem in linear regression for a third order polynomial looks as follows:

$$\min_{a,b,c,d} \sum_{k=1}^{N} (y_k - a - bx_k - cx_k^2 - dx_k^3)^2$$

Due to practical reasons, we restrict ourselves to (piecewise) degree three polynomials and a fixed natural number K of segments, the permitted number of different regression functions. The equation to be minimized, where the I_js are in increasing order with $\bigcup_j I_j = \{x_1, \ldots, x_k\}$, is

$$\min_{L=1,\ldots,K} \sum_{j=1}^{L} \min_{a_j,b_j,c_j,d_j} \sum_{k \in I_j} (y_k - a_j - b_j x_k - c_j x_k^2 - d_j x_k^3)^2$$

Further, a_j, b_j, c_j, d_j are the coefficients of the jth regression function, i.e. the ones for the interval I_j. Additional knowledge of the statical system reduces the computational effort: only in load introduction points of the system a change of the function is permitted. This is an additional criterion in the minimization as well as the R^2-statistic to be minimized. This leads to a multicriterial optimization problem. The result can be seen in Fig. 5, we refer to [2–4, 8] for further details.

3.2 Statistical Tests for Outlier Detection

The devices used for data-recording are error-prone. This means that a lot of work has to be done to eliminate/minimize the systematic error occurring due to the measuring devices. At a first glance physically non-plausible records were eliminated using a fixed-deviation criterion based on the literature, compare Fig. 3. This reduces on the one hand the amount of errors recorded, on the other hand the number of records available. A statistical test based on the median and asymptotics of order statistics has been developed to solve this problem, for details on multiple testing see [3]. It works as follows: for a dataset $x = (x_1, \ldots, x_N)$ compute the median $\tilde{x}$, the 25% and 75%-quantiles and their distance $\tilde{x}_{25}$, $\tilde{x}_{75}$ and $\tilde{x}_{diff}$, respectively. Those recorded values satisfying $|x_i - \tilde{x}| > b \cdot \tilde{x}_{diff}$ where $b > 0$ is the bandwidth (chosen according to the asymptotics of order statistics to get the desired

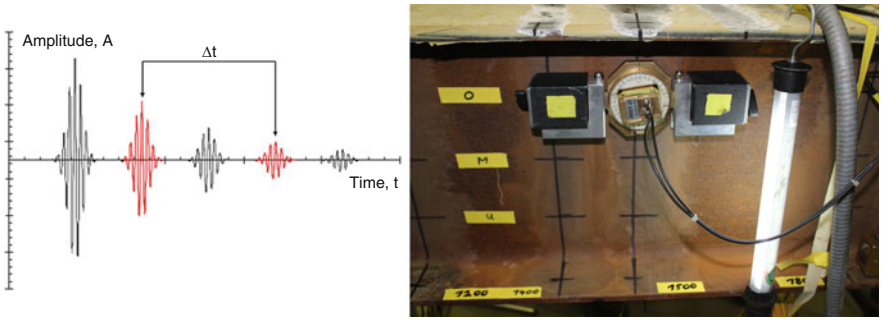


Fig. 3 *Left*: one source of errors is the skip of one of the echos which is physically non-plausible, handled using the plausibility-criterion. *Right*: precise preparation is necessary for feasible measurements

α-level) are classified as outliers. This significantly improves the results. As they have not been applicable to practical problems due to numerical inconsistencies, especially the computation of the regression curve, the error bands have lengths of roughly $600\,\mathrm{N/mm^2}$. This is not practicable as the yield limit of inbuilded steel is often $235\,\mathrm{N/mm^2}$. Furthermore, the distribution of the residuals has been statistically proven, via adapted Kolmogorov-Smirnoff test, to be a mixture of two gaussian distributions. Solutions to both of the problems above mentioned are presented in the following paragraph.

3.3 Improvement with Micromagnetics

The determination of valid confidence bands is a necessity for proper safety concepts in civil engineering. Further, the classification of residuals observed in ultrasonics has to be done properly to avoid economical and ecological disadvantages. Based on [1], additional micromagnetic measurements are used to link techniques and concepts.

3.3.1 Reduction of Complexity

The devices used to record micromagnetic quantities measure 42 different quantities. They use four different implemented sensors which are rather expensive. To make the device handable in practice, the goal is the reduction of the number of quantities necessary without significant loss of quality. For this purpose we applied multiple statistical tests of independence, based on χ^2-tests, for details, we refer to [7]. With a (combined) level of 1%, 36 quantities are identified to be stochastically dependent. The quantities left are driven by two sensors only: the Barkhausen effect and incremental permeability, which halves the amount of sensors required, see [6] for details.

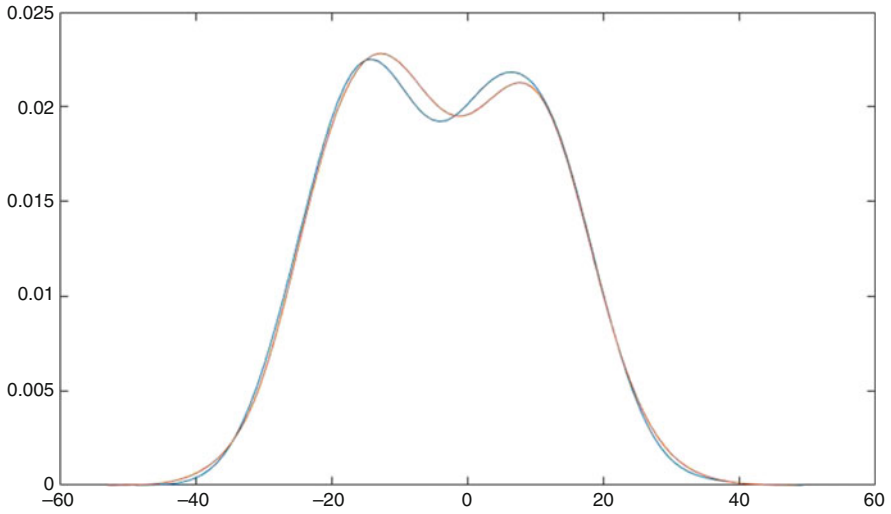


Fig. 4 The density estimated using a gaussian kernel and measurements is given in *blue*, the mixture of the densities estimated using the link-function in *orange*. The difference is minimal due to bias-effects which vanishes asymptotically

3.3.2 Combining Ultrasonic Waves and Micromagnetism

The link function is based on stochastic moments of the kernel density estimate of the micromagnetic measurements. It allows the adaption of the confidence band given the distribution of the residuals specifically for all points (Figs. 4 and 5). Under some restrictions, which are satisfied in practice, the estimation of the parameters of the mixture of gaussian distributions can be done consistent. Moreover, the estimates are asymptotically normally distributed. This allows combined confidence bands, see [8] for details: for the regression function given the best/worst case in parameter estimation. The last construction used to improve the results is a non-causal Moving-Average process of order 4 to model the residuals (compared to finite-element computations), i.e.

$$\epsilon_k = \sum_{j=-2,\, j\neq 0}^{2} \alpha_j \cdot \epsilon_{k+j} + \delta_k$$

The results of Sects. 3.2 and 3.3 look as given in Fig. 5.

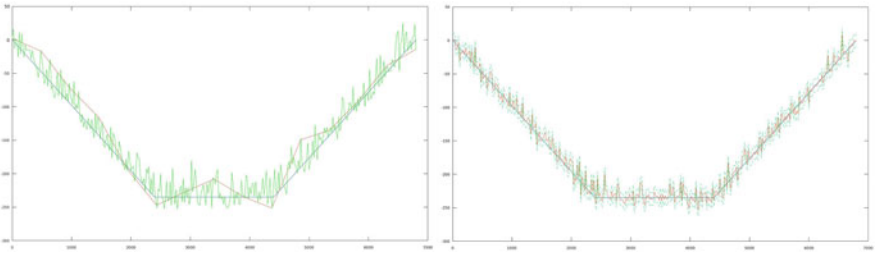


Fig. 5 *Left*: the ideal (*blue*) curve, the recorded data (*green*) and the estimated curve (*orange*) *Right*: the ideal (*blue*) curve, the improved estimation one (*orange*) and the lower and upper bounds for the worst-case confidence band (*green*)

4 Conclusion

Non-destructive testing is always in competition with destructive testing. Using a combination of micromagnetic and ultrasonic measurements it has been shown that the reserve of resistance of existing steel buildings can be determined non-destructively. Although the quality of records is malfunctioning the statistical tools developed provide good results. To get confidence bands for proving the load bearing capacity, the determined stress curve is smoothened via residual analysis and a non-causal time series approach. Further, we are able to give confidence bands based on micromagnetic clustering which shortens the confidence band. This ends up in new and improved security concepts for existing buildings.

References

1. Ackermann, J.: Die Barkhausen-Rauschanalyse zur Ermittlung von Eigenspannungen im Stahlbau. Ph.D. thesis, Department of Civil Engineering, University of Darmstadt (2008)
2. Doktor, M., Fox, C., Kurz, W., et al.: FOSTA Bericht P859 Bestandsbewertung von Stahlbauwerken mithilfe zerstörungsfreier Prüfverfahren (2016)
3. Fox, C., Kurz, W.: Evaluation of steel buildings by means of non-destructive testing methods. Schriftenreihe des Studiengangs Bauingenieurwesen der TU Kaiserslautern, Band 18 (2014)
4. Fox, C., Doktor, M., Kurz, W., Schneider, E.: Beitrag zur Bewertung von Stahlbauwerken durch zerstörungsfreie Prüfung von Spannungszuständen. In: STAHLBAU, vol. 85(1), pp. 1–15 (2016)
5. Lerman, P.M.: Fitting segmented regression models by grid search. J. R. Stat. Soc. Ser. C (Appl. Stat.) **29**(1), 77–84 (1089)
6. Mayer, J.-P.: Aufbau und Kalibrierung eines magnetischen Hall-Sensors zur Detektion und Bewertung von Schädigung an Stahlbauwerken. Seminar Project, Institute of Materials Science and Engineering, University of Kaiserslautern (2016)
7. Roy, S.N., Bargmann, R.E.: Tests of multiple independence and the associated confidence bounds. Ann. Math. Stat. **29**(2), 491–503 (1958)
8. Thein, C.: Sicherheitskonzept eines zerstörungsfrei geführten Tragfähigkeitsnachweises. B.Sc. thesis, Department of Civil Engeneering, University of Kaiserslautern (2016)

A Novel Multi-Scale Strategy for Multi-Parametric Optimization

Rosa Donat, Sergio López-Ureña, and Marc Menec

Abstract The motion of a sailing yacht is the result of an equilibrium between the aerodynamic forces, generated by the sails, and the hydrodynamic forces, generated by the hull(s) and the appendages (such as the keels, the rudders, the foils, etc.), which may be fixed or movable and not only compensate the aero-forces, but are also used to drive the boat. In most of the design, the 3D shape of an appendage is the combination of a plan form (2D side shape) and a planar section(s) perpendicular to it, whose design depends on the function of the appendage. We often need a section which generates a certain quantity of lift to fulfill its function, but the lift comes with a penalty which is the drag. The efficiency, equilibrium and speed of a sailing boat depend on the appendage hence on the planar section. We describe a multi-scale strategy to optimize the shape of a section in a multi-parametric setting by embedding the problem within a discrete multiresolution framework. We show that our strategy can be easily implemented and, combined with appropriate optimization techniques, provides a fast algorithm to obtain an 'optimal' perturbation of the original shape.

1 Introduction

Appendages of a sailing yacht such as keels and rudders are designed from planar sections [1, 4]. The shape of these sections determines the drag and lift generated by the appendage and, hence, the boat's efficiency and performance. The search for an 'optimal' shape, which minimizes the drag generated by the section, is thus an important problem in yacht design.

Sections are closed planar curves, which may be parametrized as $\alpha(t) = (x(t), y(t))$, $t \in [0, 1]$. We shall assume that the first coordinate, x, runs along the

R. Donat • S. López-Ureña (✉)
Universitat de València, 50th Doctor Moliner street, 46100 València, Burjassot, Spain
e-mail: donat@uv.es; sergio.lopez-urena@uv.es

M. Menec
IS3DE ENG., 5th Mare Nostrum avenue, València, Spain
e-mail: marc.menec@is3de.com

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_91

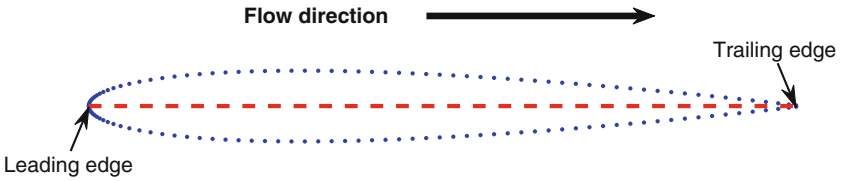


Fig. 1 An example of section: NACA0010 with $N = 129$ points

fluid direction and the second coordinate, y, is perpendicular to the direction of x. We also assume that $\alpha(0) = \alpha(1)$ corresponds to the trailing edge (Fig. 1).

Our goal is to perturb the shape of a given section so that the associated *drag coefficient*, computed by the CFD code `xfoil`,[1] is 'optimally' low. Codes like `xfoil` work in a discrete framework where the closed profile α is simply given by a finite number of points in the plane $\alpha = (\alpha_i)$, $\alpha_i = (x_i, y_i)$, $1 \le i \le N$. If the shape of the section is given by the parametrization $\alpha(t)$, $t \in [0, 1]$, then

$$\alpha_i = \alpha(t_i), \quad t_i < t_{i+1}, \quad t_i \in [0, 1]. \tag{1}$$

Hence, we may assume that we have at our disposal an underlying mesh on $[0, 1]$, given by a finite number of nodes t_i, $1 \le i \le N$, such that α_i is associated t_i.

We seek an 'optimal shape' by considering perturbations of the original shape $\alpha = (\alpha_i)$ of the form $\alpha^\varepsilon := (\alpha_i + \varepsilon_i)_i$, $\varepsilon_i \in \mathbb{R}^2$. Our goal is to find

$$\varepsilon_* \in \mathbb{R}^{2\times N} : \mathscr{D}(\alpha^{\varepsilon_*}) = \min_{\varepsilon\in\mathbb{R}^{2\times N}} \mathscr{D}(\alpha^\varepsilon), \tag{2}$$

where $\mathscr{D}(\beta)$ is the drag coefficient associated to a section $\beta = (\beta_i)$. The structural limitations of the appendage have to be taken into account in the optimization process. For example, for a keel section we would like to preserve the inertia[2] and the chord length of the initial shape. Hence (2) is, in general, a multi-parametric constrained optimization problem.

For a 'typical' section shape ($N = 129$) finding the solution of (2) may be (very) slow due to the large number of free parameters. The purpose of this paper is to describe a *multi-scale optimization* (MSO) strategy, which reduces the computational cost of solving the minimization problem (2).

The paper is organized as follows: In Sect. 2 we briefly recall the main ingredients of Harten's Multiresolution Framework (H-MRF), and in particular its connection with subdivision schemes, which is the main tool used in our MSO strategy. Section 3 provides the description of our technique and in Sect. 4 we carry out several examples of application of our strategy. We demonstrate the efficiency of

[1] XFOIL is an interactive program for the design and analysis of subsonic isolated airfoils.

[2] Inertia is the resistance of any physical object to any change in its state of motion.

our MSO strategy in an academic example. An example involving a more realistic situation is also shown.

2 Harten's Framework for Multiresolution

A multiresolution (MR) decomposition of a discrete data set α^L is an equivalent representation that encodes the discrete information contained in α^L as a *coarse* representation of this data set, α^0, together with a sequence of details at each resolution level (d^j)

$$\alpha^L \equiv M\alpha^L := (\alpha^0, d^0, d^1, \ldots, d^{L-1}), \tag{3}$$

where L represents the finest resolution level and 0 the coarsest one. In H-MRF the levels of resolution are induced from a hierarchy of *nested*[3] meshes $(\mathscr{G}^j)_{j=0}^L$ on an underlying spatial domain. In H-MRF, the representation in (3) is obtained after a repeated application of the operators which connect two consecutive resolution levels.

- *Decimation* (from fine to coarse): $\alpha^{j-1} = D_j^{j-1}\alpha^j$. D_j^{j-1} are linear operators that define a coarser representation α^{j-1} from a finer representation α^j.
- *Prediction* (from coarse to fine): $\tilde{\alpha}^j = P_{j-1}^j \alpha^{j-1}$. P_{j-1}^j are (possibly nonlinear) operators that define new data at a finer level, from discrete data at a coarser resolution level.

The multiresolution transform, and its inverse, are obtained after a recursive application of the 2-level transform

$$\begin{cases} \alpha^{j-1} &= D_j^{j-1}\alpha^j \\ e^j &= (I - P_{j-1}^j D_j^{j-1})\alpha^j \end{cases} \quad \leftrightarrow \quad \begin{cases} \alpha^j &= P_j^{j-1}\alpha^{j-1} + e^j. \end{cases} \tag{4}$$

An essential property in H-MRF is *consistency*: $D_j^{j-1}P_{j-1}^j = I$, that is, decimation after prediction must preserve the original data set. Prediction operators are often obtained by using subdivision schemes [3] which are *consistent* with the decimation. The detail coefficients (d^j) represent the non-redundant information in the prediction error (e^j), and $\alpha^j \equiv (\alpha^{j-1}, d^{j-1})$, $\forall j$, and both sequences have exactly the same number of components.

For further details on Harten's MRF, we refer the interested reader to [5] and references therein.

[3] $\{\mathscr{G}^j\}_j$ is nested if $\mathscr{G}^j \subset \mathscr{G}^{j+1}$. An example of nested grids on $[0, 1]$ is $\mathscr{G}^j = (i2^{-j})_{i=0}^{2^j}$.

3 The Multi-Scale Optimization (MSO) Technique

As mentioned before, using a standard optimization tool to solve problem (2) might be very costly, or even unfeasible, when N is large. This is mainly due to the large number of parameters involved, and the cost of evaluating the functional $\mathscr{D}$.

In this section we describe a feasible mechanism to solve problem (2), under appropriate conditions. Our MSO technique is based on an embedding of the problem within a conveniently chosen MRF, specified by given decimation and prediction operators. Here, we only provide a description of the final algorithm, which can be seen as a multi-scale 'parameter-reduction' approach to the original minimization problem. We refer the interested reader to [2] for specific details on the design of the proposed technique, and only mention here that, although the design of our MSO technique is based on (3), the decimation operator does not play a direct role in the practical application of the technique.

We assume that we are given an initial (discrete) shape, $\alpha = (\alpha_i)_{i=1}^N$ and that the underlying grid $(t_i)_{i=1}^N$ in (1) defines the finest resolution level in our chosen MRF. Then, starting at the coarsest level, we solve a sequence of optimization problems that involve a number of free parameters which increases when climbing up the MR ladder. More specifically, at the coarsest level we compute

$$\varepsilon_*^0 = \operatorname{argmin}_{\varepsilon^0 \in \mathbb{R}^{2\times N_0}} \mathscr{D}(\alpha^L + \Pi_{j=1}^L P_{j-1}^j \varepsilon^0), \quad \text{initial guess } \varepsilon_0^0 = 0. \tag{5}$$

Notice that, if N_0 is sufficiently small, (5) may be solved very fast. Then, for $k = 1, 2, \ldots, L$, we compute

$$\varepsilon_*^k = \operatorname{argmin}_{\varepsilon^k \in \mathbb{R}^{2\times N_k}} \mathscr{D}(\alpha^L + \Pi_{j=k+1}^L P_{j-1}^j \varepsilon^k), \quad \text{initial guess } \varepsilon_0^k = P_{k-1}^k \varepsilon_*^{k-1}. \tag{6}$$

We notice that the k-level optimal set of parameters, ε_*^k serves to construct a (discrete) curve at the highest resolution level, $\alpha^{L,k} := \alpha^L + \Pi_{j=k+1}^L P_{j-1}^j \varepsilon_*^k$, which represent the *best* kth-*level approximation* to the solution of problem (2). For $j = L$, $\alpha^{L,L}$ is the 'optimal' curve which solves problem (2).

It is expected that $\alpha^{L,k}$ gets closer to the 'optimal' curve when k increases. Hence, even though the successive optimization problems in (6) involve more parameters, when k increases, the initial guess is closer to the final solution, which reduces the total cost. The performance of the proposed MSO technique will be illustrated in the following section.

4 Numerical Experiments

The setting in the numerical experiments considered in this section is as follows:

- We seek to minimize a given functional with a specific minimization tool (in this work we use those available in Matlab) and using an initial guess provided by the user.
- As a stopping criteria we consider the max-norm of the difference between two consecutive iterates and the absolute value of the corresponding functional evaluations. We stop the minimization process when both quantities are less than a specified tolerance (*tol*). In addition, we stop the process (and consider it unfeasible) if the number of evaluations of the functional reaches 10^5.
- The prediction operator is always given by a B-spline subdivision scheme of order 5 [3] to ensure smooth perturbations of the initial shapes throughout the process.

4.1 An Academic Example

We start with an academic example that shows the advantages of using our proposed MSO technique, as opposed to using directly the minimization code to solve the problem in a large parameter space. The setting here is one-dimensional, and a bit simpler than that considered in (2): We seek to find the minimum of the functional

$$F(z) := \|z_i - \cos(2\pi t_i)\|_2^2, \qquad (t_i)_{i=0}^{2^L} = (i2^{-L})_{i=0}^{2^L},$$

starting from the initial shape $\bar{z} = (\lambda \cos(2\pi t_i))_{i=1}^{N}$. The parameter λ controls how far the initial guess is from the minimum of the functional, $z_{\min} = (\cos(2\pi t_i))_i$. The MATLAB `fminsearch` minimization tool has been selected to solve

$$\text{Find } \varepsilon_* \in \mathbb{R}^N : \ F(z^{\varepsilon *}) = \min_{\varepsilon \in \mathbb{R}^N} F(z^{\varepsilon}), \ \ z^{\varepsilon} := z + \varepsilon, \ \ N = 2^L + 1, \, L = 7, \qquad (7)$$

using the following minimization strategies ($tol = 10^{-4}$):

- DM: a Direct use of the chosen minimization tool on $\mathbb{R}^N$.
- MSO: using the same minimization code within our MSO strategy.

In Fig. 2, we show the output of the MSO strategy (always on top of the true solution, $z_{\min}$). In Fig. 2b, c the output of the DM strategy after 10^5 iterations is displayed. Notice that the DM strategy may be unfeasible when the initial guess is not sufficiently close to the desired solution.

We consider next the computational cost, measured by the number of functional evaluations (in more realistic cases, the remaining operations have a negligible cost

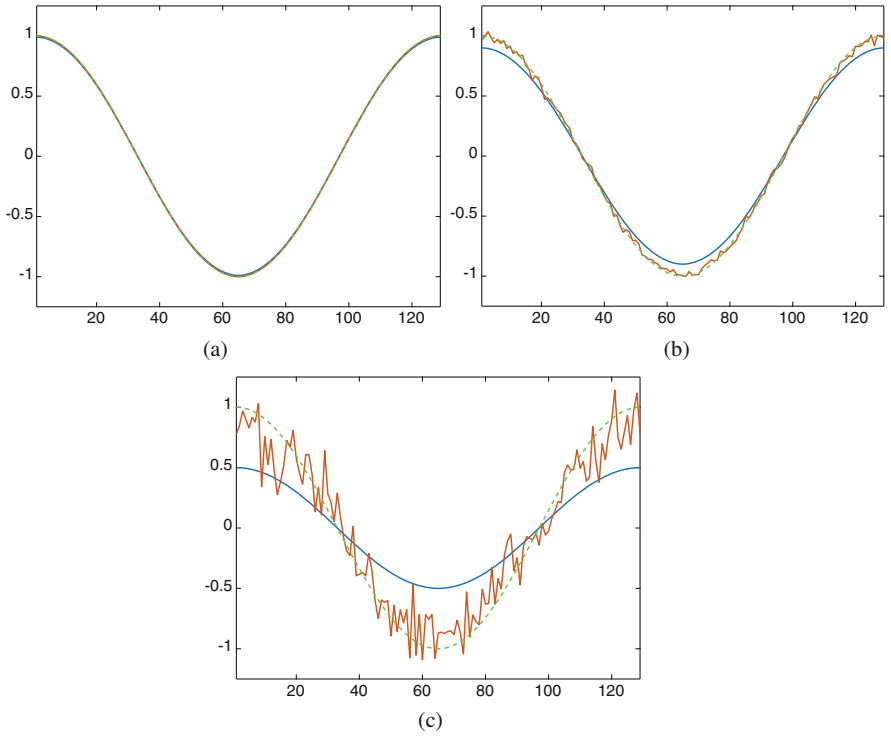


Fig. 2 Initial guess (*blue line*), output of DM algorithm (*red line*) and optimal solution/output of MSO algorithm (*green dashed line*) for (**a**) $\lambda = 0.999$, (**b**) $\lambda = 0.9$, (**c**) $\lambda = 0.5$

Table 1 F-eval= total number of function evaluations. $F_{min} = F(z_{last})$, $z_{last} = z + \epsilon_{last}$, ϵ_{last} last value computed

	MSO		DM	
λ	F-eval.	F_{min}	F-eval.	F_{min}
0.999	1758	9.48e−11	11,957	3.27e−9
0.9	4248	1.81e−10	*	6.81e−4
0.5	11,343	3.85e−11	*	3.58e−2

The symbol * means that the maximum allowed value of 10^5 has been reached

compared with the evaluation of the functional). The results, for both strategies and for various values of λ are compiled in Table 1. According to Table 1, only for $\lambda = 0.999$ the initial guess is close enough to $z_{\min}$ so that the DM strategy is feasible. For the other values of λ, only the MSO leads to a feasible algorithm. In addition, the MSO strategy leads to a very large computational gain for $\lambda = 0.999$. The results in Table 1 also show the value of $F(z_{last})$, which, for this problem, is also a measure (in the 2-norm) of how close this iterate is with respect to $z_{\min}$.

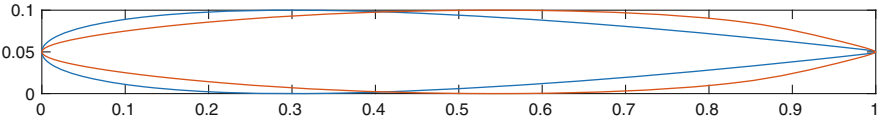


Fig. 3 Initial profile: *blue line*. Output of the MSO of the drag coefficient while preserving chord length and max height: *red line*. The drag coefficient is reduced from $9.21 \cdot 10^{-3}$ to $4.47 \cdot 10^{-3}$

4.2 A Realistic Case

In this section we consider a more realistic experiment, for which a DM strategy is unfeasible. We seek to solve problem (2) where the functional $\mathscr{D}$ provides the drag coefficient of a section, computed with `xfoil`, under a value of the Reynolds number equal to 10^6. In this section $tol = 10^{-5}$, $L = 5$, $N_0 = 2^2 + 1$.

The initial profile, provided by the user is a discrete version of the NACA0010-profile, with $N = 129$ points, shown in Fig. 1. The new sections α^ε generated during the minimization process are required to meet some structural conditions and the selected minimization tool is the MATLAB function `patternsearch`, which allows the specification of linear as well as nonlinear constraints.

In Fig. 3 we display the minimizing shape (in red) when the imposed constraints are that the chord length and maximum height of the initial profile (in blue) should be maintained.

5 Conclusions and Perspectives

We have described a novel multi-scale strategy to solve optimization problems in large parameter spaces. The technique only relies on the definition of an appropriate subdivision scheme, which depends on the type of application under consideration.

By considering an academic toy-problem we have shown that the technique computes the desired minimal shape with a rather small computational effort, compared with a direct strategy that uses the same underlying minimization tool. For realistic scenarios, only the multi-scale strategy is able to provide results.

More analytic work is necessary in order to determine the range of problems for which the technique would provide the desired optimal shape, as well as the influence of the underlying minimization code or the specification of the particular constraints.

Acknowledgements The authors acknowledge support from Project MTM2014-54388 (MINECO, Spain) and the FPU14/02216 grant (MECD, Spain)

References

1. Abbot, I.H., Von Doenhoff, A.E.: Theory of Wing Sections. Dover Publication, New York (1959)
2. Donat, R., López-Ureña, S.: A novel multi-scale strategy for multi-parametric optimization. (in preparation)
3. Dyn, N.: Subdivision Schemes in Computer-Aided Geometric Design. Advances in Numerical Analysis, vol. II. Lancaster (1990), pp. 36–104. Oxford Science Publication, Oxford University Press, New York (1992)
4. Fossati, F.: Aero-hydrodynamics and the performance of sailing Yachts: the science behind sailboats and their design. A&C Black, London (2009)
5. Harten, A.: Multiresolution representation of data: a general framework. SIAM J. Numer. Anal. **33**(3), 1205–1256 (1996)

Optimal Shape Design for Polymer Spin Packs

Robert Feßler, Christian Leithäuser, and René Pinnau

Abstract A shape optimization approach for the design of cavities with a specified wall shear stress profile is presented. Applications are in the design of spin pack geometries with low and homogeneous residence times and without dead spaces to prevent polymer degradation for sensitive materials. Furthermore, a related operator is studied which suggests that the set of attainable wall shear stresses is rather large.

1 Introduction

Polymer spin packs are widely used for the production of synthetic fibers and nonwoven materials. Polymer melt is extruded through a pipe into the spin pack geometry where it is first distributed along the whole cross-sectional area before it passes several layers of filter material and is finally spun into fibers by the spinneret plate which consists of a large number of very small nozzles. The whole spin pack is heated in order to prevent premature solidification of the melt. However, the influence of heat can lead to polymer degradation if the residence times are too long. For sensitive polymers with interesting properties this issue can be the limiting factor which prevents innovations due to the fact that spinning is not possible. This can be resolved by designing special spin packs with low residence time profiles. An important part in the spin pack design is the cavity which distributes polymer from the inlet pipe onto the whole cross-sectional area. This part of the geometry is in particular vulnerable for dead spaces and regions with slow flow velocities where degradation can take place. These problematic flow regions usually occur in close proximity to the walls and the wall shear stress can be used for quantification. In general a low wall shear stress coincides with a slow velocity zone close to the wall. On the other hand being able to design cavities with a sufficiently high level

R. Feßler • C. Leithäuser (✉)
Fraunhofer ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
e-mail: robert.fessler@itwm.fraunhofer.de; christian.leithaeuser@itwm.fraunhofer.de

R. Pinnau
TU Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany
e-mail: pinnau@mathematik.uni-kl.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_92

of stress throughout its wall is an effective tool against dead spaces and thus against large residence times and polymer degradation.

In the following we present a shape optimization approach which can be used to design cavities with specified wall shear stress profiles. We begin in Sect. 2 by introducing the model. Section 3 gives a result on the approximate controllability of the space of attainable wall shear stress profiles for a linearized operator. The numerical approach for solving the shape optimization problem is presented in Sect. 4 and results are shown in Sect. 5.

2 Modeling

Let $\Omega \subset \mathbb{R}^d$ with $d = 2, 3$ be a bounded domain whose boundary Γ decomposes into an inlet part Γ^{in}, a wall part Γ^{wall} and an outlet part Γ^{out} (see Fig. 1). Let $\mathbf{n}$ be the outward pointing normal vector. Since the viscosity η is large compared to the flow velocities Stokes equations are used to model the flow within the spin pack. For velocity $\mathbf{u}$ and pressure p we solve

$$\begin{aligned} -\eta\Delta\mathbf{u} + \nabla p &= 0 && \text{in } \Omega \\ \operatorname{div}\mathbf{u} &= 0 && \text{in } \Omega \\ \mathbf{u} &= \mathbf{n}u_0 && \text{on } \Gamma^{in} \\ \mathbf{u} &= 0 && \text{on } \Gamma^{wall} \\ (\mathbf{n}\cdot\mathbf{u})\frac{a}{A} &= -\frac{k}{\eta}\frac{p-p_0}{L} && \text{on } \Gamma^{out} \\ \mathbf{n}\times\mathbf{u} &= 0 && \text{on } \Gamma^{out}, \end{aligned} \tag{1}$$

where u_0 is a given inflow profile and the parameters a, A, k and L are used to model the spinneret plate as a porous medium through a Darcy law. The wall shear stress can then be computed by

$$\sigma = \eta\,|\nabla\times\mathbf{u}| \qquad \text{on } \Gamma^{wall}.$$

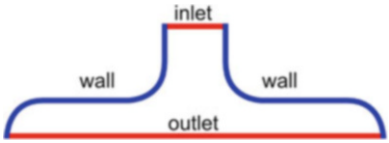


Fig. 1 Sketch of a distributor geometry with boundary partition

3 Approximate Controllability

Before we deal with the numerical approach we want to investigate the operator $\mathbf{S} : \Omega \mapsto \sigma$ which maps any admissible shape to the wall shear stress. We derive results for the linearization $d\mathbf{S}$ of $\mathbf{S}$ and these results suggest that the set of attainable wall shear stresses is rather large which is a good sign for the performance of the shape design algorithm. Our approach is inspired by Chenais and Zuazua [1] where the authors study a similar operator for a potential flow problem. In the following we define the operators $\mathbf{S}$ and $d\mathbf{S}$ and state the result on the approximate controllability of $d\mathbf{S}$ in Theorem 1. Further details including the proof can be found in [3].

Definition 1 (Approximate Controllability) Let $F : X \to Y$ be a linear operator. Then, F is approximately controllable if and only if $\operatorname{im} F$ lies dense in Y.

For this section let $\Omega_0 \subset \mathbb{R}^2$ be a bounded domain of class $C^{6,1}$. We assume that there is no Γ_0^{out} boundary with a Darcy law boundary condition (cf. (1)). The outlet is realized by a velocity boundary condition similar to the inlet, which is admissible for the Stokes problem if the flows are balanced. Define

$$\mathcal{V} = \{\mathbf{V} \in C^{4,1}(\mathbb{R}^2, \mathbb{R}^2); \mathbf{V}|_{\Gamma_0^{in}} = 0; \mathbf{n} \times \mathbf{V}|_{\Gamma_0^{wall}} = 0\}$$
$$\Theta = \{\theta \in \mathcal{V}; \|\theta\|_{C^{4,1}(\mathbb{R}^2,\mathbb{R}^2)} < 0.5\}.$$

Then for every $\theta \in \Theta$ the map $\mathrm{Id} + \theta : \mathbb{R}^2 \to \mathbb{R}^2$ is invertible and it is furthermore a $(4, 1)$-diffeomorphism (see [6]). We can use this to define our space of admissible shapes by

$$\mathcal{D} := \{\Omega_\theta = (\mathrm{Id} + \theta)(\Omega_0); \theta \in \Theta\} \tag{2}$$

as perturbations of the reference shape Ω_0. By definition these perturbations leave Γ_0^{in} fixed and are normal on Γ_0^{w}. We define the shape-to-wall-shear-stress operator by

$$\begin{aligned} \mathbf{S} : \mathcal{D} &\to L^2(\Gamma_0^{wall}) \\ \Omega_\theta &\mapsto (\nabla \times \mathbf{u}(\theta)|_{\Gamma_\theta^{wall}}) \circ (\mathrm{Id} + \theta), \end{aligned} \tag{3}$$

where $\mathbf{u}(\theta)$ and $p(\theta)$ are the solutions of the Stokes problem on Ω_θ

$$\begin{aligned} -\Delta \mathbf{u}(\theta) + \nabla p(\theta) &= 0 && \text{in } \Omega_\theta \\ \operatorname{div} \mathbf{u}(\theta) &= 0 && \text{in } \Omega_\theta \\ \mathbf{u}(\theta) &= \mathbf{n} u_0 && \text{on } \Gamma_\theta^{in} \\ \mathbf{u}(\theta) &= 0 && \text{on } \Gamma_\theta^{wall}. \end{aligned} \tag{4}$$

Let

$$d\mathbf{S} : \Theta \to L^2(\Gamma_0^{wall})$$
$$\theta \mapsto \lim_{s \searrow 0} \frac{\mathbf{S}(\Omega_{s\theta}) - \mathbf{S}(\Omega_0)}{s}$$

be the linearization of $\mathbf{S}$. Both operators are well-defined as shown in [3].

Theorem 1 *Let Ω_0 be bounded of class $C^{6,1}$ and assume that $\mathbf{S}(0) \neq 0$ on Γ_0^{wall}. Then, the operator $d\mathbf{S} : \mathcal{V} \to L^2(\Gamma_0^{wall})/\mathcal{Z}$ is approximately controllable. Here $\mathcal{Z}$ is a finite dimensional subspace of $L^2(\Gamma_0^{wall})$.*

The proof relies on an adjoint argument (cf. [4, 5]) and details are given in [3]. This result shows that the linearized operator $d\mathbf{S}$ is approximate controllable except for some finite dimensional subspace $\mathcal{Z}$. This suggests that the set of attainable wall shear stresses for the original operator $\mathbf{S}$ is also rather large and that we can hope to derive geometries with a wall shear stress close to the desired target wall shear stress with the numerical approach presented in the following.

4 Numerical Shape Optimization

The present shape optimization algorithm for the design of cavities with a specified wall shear stress profile is presented. The goal is to solve the following problem:

$$\min_{\substack{\Omega \\ \text{subject to (1)}}} \quad J(\Omega) = \int_{\Gamma^{wall}} (\eta\,|\nabla \times \mathbf{u}| - \sigma_d)^2\, ds$$

where $\sigma_d \in \mathbb{R}^+$ is a given target wall shear stress. For simplicity σ_d is assumed to be constant but a generalization is straightforward. Let $\Omega_0 \subset \mathbb{R}^3$ be a given initial domain and let Ω_k be the domain after iteration $k \in \mathbb{N}$ of the algorithm. For sufficiently regular shape deformations $\theta : \mathbb{R}^2 \to \mathbb{R}^2$ with $\theta|_{\Gamma_k^{in} \cup \Gamma_k^{out}} = 0$ and $\mathbf{n} \times \theta|_{\Gamma_k^{wall}} = 0$ we want to differentiate $J((\mathrm{Id} + \theta) \circ \Omega_k)$ with respect to θ. Using shape derivatives as introduced in [6, 7] leads to the adjoint problem

$$\begin{aligned}
-\eta \Delta \mathbf{v} + \nabla q &= 0 && \text{in } \Omega_k \\
\operatorname{div} \mathbf{v} &= 0 && \text{in } \Omega_k \\
\mathbf{v} &= 0 && \text{on } \Gamma_k^{in} \\
\mathbf{n} \cdot \mathbf{v} &= 0 && \text{on } \Gamma_k^{wall} \\
\mathbf{v} \times \mathbf{n} &= 2\eta \frac{\sigma - \sigma_d}{\sigma} (\nabla \times \mathbf{u}(0)) && \text{on } \Gamma_k^{wall} \\
q - \eta_{out}(\mathbf{n} \cdot \mathbf{v}) &= 0 && \text{on } \Gamma_k^{out} \\
\mathbf{v} \times \mathbf{n} &= 0 && \text{on } \Gamma_k^{out}.
\end{aligned} \tag{5}$$

The gradient θ_{grad} is then given on Γ_k^{wall} by

$$\begin{aligned}\theta_{grad} = &-\eta\,(\partial_{\mathbf{n}}\mathbf{u}\times\mathbf{n})\cdot\nabla\times\mathbf{v} + 2\eta^2\frac{\sigma-\sigma_d}{\sigma}\nabla\times\mathbf{u}\cdot\partial_{\mathbf{n}}(\nabla\times\mathbf{u})\\ &- 2(\sigma-\sigma_d)\partial_{\mathbf{n}}\sigma_d + \kappa(\sigma-\sigma_d)^2.\end{aligned} \tag{6}$$

Finally a line search is used to find a step size $s \in \mathbb{R}^+$ such that

$$\Omega_{k+1} = (\mathrm{Id} + s\theta_{grad}) \circ \Omega_k$$

minimizes $J(\Omega_{k+1})$. The optimization problem is solved by iterating over k. In practice some effort is necessary to preserve the regularity of Ω_k. Details on this are given in [2].

5 Numerical Results

With the presented shape optimization approach it becomes possible to design cavities with a specific wall shear stress profile. Figures 2 and 3 show two cavities with a constant target wall shear stress of $\sigma_d = 0.1$ and $\sigma_d = 0.2$, respectively. The wall shear stress is indicated by color and we see that it is possible to get a very close agreement between target and actual wall shear stress throughout wide parts of the geometries. This agrees with Theorem 1 which suggests that the space of attainable wall shear stress profiles is rather large. The wall shear stress has been used as an auxiliary quantity to improve the residence times. Figure 4 compares the residence times between the two optimized cavities with a simple flat reference cavity.

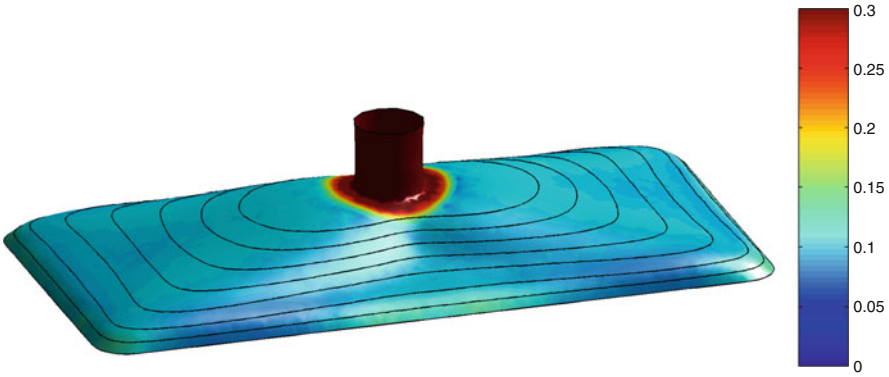


Fig. 2 Cavity A: optimal distributor geometry for a target wall shear stress of $\sigma_d = 0.1$

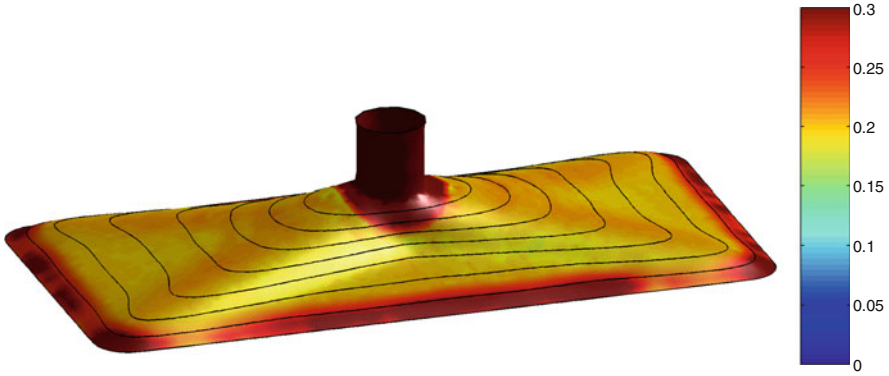


Fig. 3 Cavity B: optimal distributor geometry for a target wall shear stress of $\sigma_d = 0.2$

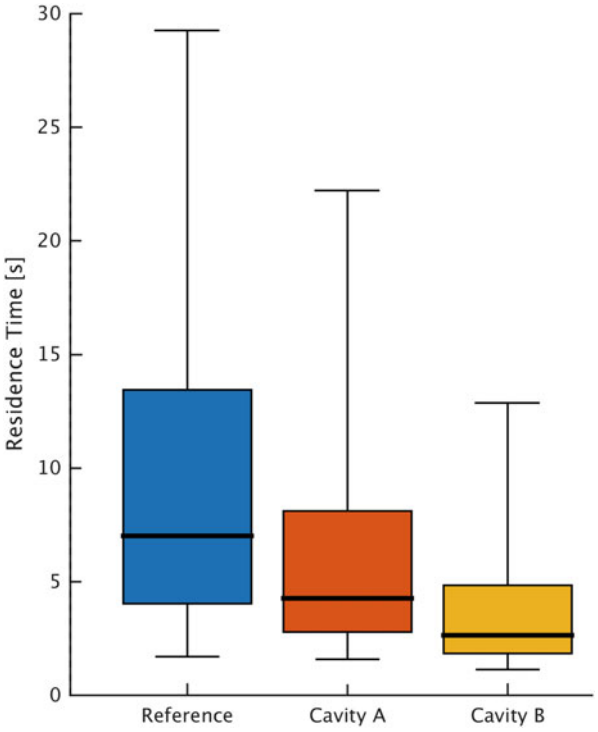


Fig. 4 Residence time in Cavity A and Cavity B compared to a flat reference cavity

6 Conclusion

The presented shape optimization approach is able to generate cavities with specific wall shear stress profiles. This has proven to be an effective tool in the optimization of polymer spin packs and filter devices for sensitive applications. It is now possible to design geometries without dead spaces and with short and more homogeneous residence times. Our optimization approach uses Stokes equation and no complex rheology. However, when dealing with industrial applications we always validate the final geometry with a full Navier-Stokes simulation and a complex rheology model. The validation confirms that using the simplified model in the optimization step is sufficient for our field of applications.

Acknowledgements This work was supported by the German Federal Ministry of Education and Research (BMBF) grant no. 03MS606F and by the German Federal Ministry for Economic Affairs and Energy (BMWI) grant no. IGF 17629 N.

References

1. Chenais, D., Zuazua, E.: Controllability of an elliptic equation and its finite difference approximation by the shape of the domain. Numer. Math. **95**(1), 63–99 (2003)
2. Leithäuser, C.: Controllability of shape-dependent operators and constrained shape optimization for polymer distributors. PhD Thesis, TU Kaiserslautern (2013)
3. Leithäuser, C., Pinnau, R., Feßler, R.: Approximate controllability of linearized shape-dependent operators for flow problems. ESAIM: COCV **23**(3), 751–771 (2017)
4. Osses, A., Puel, J.: Boundary controllability of a stationary stokes system with linear convection observed on an interior curve. J. Optim. Theory Appl. **99**(1), 201–234 (1998)
5. Osses, A., Puel, J.: On the controllability of the Laplace equation observed on an interior curve. Rev. Mat. Complut. **11**(2), 403–441 (1998)
6. Simon, J.: Differentiation with respect to the domain in boundary value problems. Numer. Funct. Anal. Optim. **2**(7), 649–687 (1980)
7. Sokolowski, J., Zolesio, J.: Introduction to Shape Optimization: Shape Sensitivity Analysis, vol. 16. Springer, Berlin (1992)

A Fast Ray Tracing Method in Phase Space

Carmela Filosa, Jan ten Thije Boonkkamp, and Wilbert IJzerman

Abstract Ray tracing is a widely used method in geometric optics to simulate the propagation of light through a non-imaging optical system. We present a new ray tracing approach based on the source and the target phase space (PS) representation of two-dimensional optical systems. The new PS method allows tracing far fewer rays through the system compared to the classical ray tracing approach. The efficiency of the PS method is demonstrated by numerical simulations that compare the new approach with Monte Carlo (MC) ray tracing. The results show that the PS method is very accurate and much faster than MC.

1 Introduction

Ray tracing is a general technique to describe the propagation of light in geometric optics. The light emitted by the source consists of rays and, to compute the photometric variables at the target of the system, the interactions of every ray with the optical surfaces are considered. The method can be implemented both in two and three dimensions. In non-imaging optics Monte Carlo (MC) ray tracing is very often used. This method is quite robust but it is also a very slow procedure [4].

In this paper a new ray tracing method which employs the phase space (PS) representation of the source and the target of the system is proposed for two-dimensional optical systems. In 2D the PS is a two-dimensional space where every ray is described by a unique point, the coordinates of which describe the position and the direction of the ray, see [5]. The parts of the target PS reached by the rays that leave the source correspond to the regions with positive luminance, for the other parts the luminance is equal to zero. The idea of tracing only rays close to the discontinuities of the luminance leads to a significant reduction of the number of rays. The output intensity in PS is provided for a two-dimensional Compound

C. Filosa (✉) • J. ten Thije Boonkkamp
CASA, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: c.filosa@tue.nl; j.h.m.tenthijeboonkkamp@tue.nl

W. IJzerman
Philips Lighting, High Tech Campus 7, 5656 AE Eindhoven, The Netherlands
e-mail: wilbert.ijzerman@philips.com

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_93

Parabolic Concentrator (CPC) and the result is compared with a reference intensity. Numerical results show that the method converges proportionally to the inverse of the number of rays traced versus a speed of convergence proportional to the inverse of the square root of the number of rays traced for MC ray tracing.

The paper is organized as follows. In Sect. 2 the geometry of the CPC is described. In the same section MC ray tracing and PS ray tracing are explained; in Sect. 3 numerical results are provided; in the last section conclusions and future perspectives are given.

2 Ray Tracing for the CPC

In this section both MC ray tracing and the new PS method are described for a two-dimensional CPC the profile of which is depicted in Fig. 1.

2.1 The Classical Ray Tracing Approach: MC Ray Tracing

For a Cartesian coordinate system (x, z) a two-dimensional and symmetric optical system is defined. Given the optical system and the distribution of the rays at the source, ray tracing computes the distribution of the rays at the target. The CPC in Fig. 1 consists of the source $\mathscr{S}$ (line 1), the target $\mathscr{T}$ (line 4) and two reflectors (lines

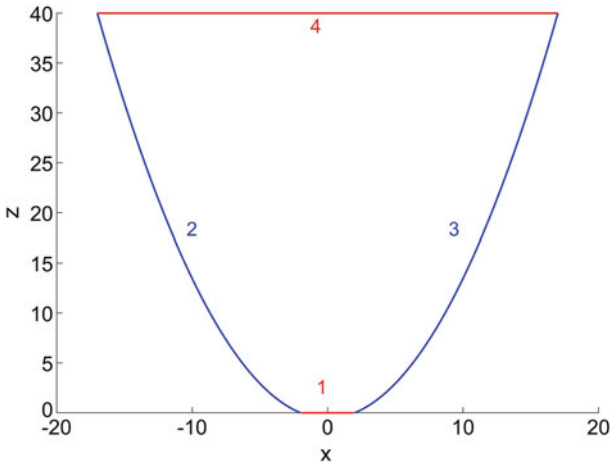


Fig. 1 Shape of the CPC. Each line of the system is labeled with a number. The source $\mathscr{S} = [-1.9, 1.9]$ (*line 1*) is located at $z = 0$. The target $\mathscr{T} = [-17, 17]$ (*line 4*) is parallel to the source and it is located at a height $z = 40$. The left and the right reflectors (*lines 2 and 3*) are arcs of parabolas. All the lines are located in air

2 and 3). The source $\mathscr{S}$ and the target $\mathscr{T}$ are horizontal lines perpendicular to the optical axis. Moreover the source $\mathscr{S} = [-a, a]$ with $a = 1.9$ is located at $z = 0$ and is Lambertian. The target $\mathscr{T} = [-b, b]$ with $b = 17$ is located at a height $z = 40$. The left and the right reflectors are arcs of parabolas. All the lines are located in a medium with index of refraction $n = 1$ (air). The position coordinates of the rays on the line i ($i \in \{1, \ldots, 4\}$) are indicated with (x_i, z_i) and the direction vector of the rays that leave line i is denoted with $\mathbf{s}_i = (-\sin t_i, \cos t_i)$, where t_i is the angle that the ray forms with the optical axis, measured counterclockwise. Therefore, a ray segment between (x_i, z_i) and (x_j, z_j) is parameterized by:

$$\mathbf{r}(s) = \begin{pmatrix} x_i - s\sin(t_i) \\ z_i + s\cos(t_i) \end{pmatrix} \qquad s \geq 0\,, \tag{1}$$

where s denotes the arc-length. Ray tracing relates the position coordinates (x_1, z_1) and the direction vector $\mathbf{s}_1$ of every ray at the source $\mathscr{S}$ with their corresponding position (x_4, z_4) and direction $\mathbf{s}_4$ at the target $\mathscr{T}$. To do this, when a ray propagates from a line i to a line j with $j \neq i$ the position coordinates (x_j, z_j) of the ray on line j and the new direction $\mathbf{s}_j$ need to be calculated. For the CPC in Fig. 1 the expressions for all the lines $i \in \{1, \ldots, 4\}$ are known. Therefore, computing the intersection between the ray's parametrization in Eq. (1) and the equation that describes the line j, the coordinates (x_j, z_j) are found. Once a ray hits the line j, its direction changes according to the reflection law. The new position and direction of the rays are calculated until the ray reaches the target $\mathscr{T}$. The procedure explained above is applied for every ray traced inside the optical system.

Once we know the output position and direction of every ray, the photometric variables at the target are computed. Considering a Lambertian source, the intensity emitted by $\mathscr{S}$ in the direction t_1 is $I(t_1) = 2aL\cos(t_1)$ where L is the luminance. In non-imaging optics a frequently used ray tracing approach is Monte Carlo (MC) ray tracing. In this procedure the rays are traced randomly from $\mathscr{S}$ to $\mathscr{T}$ and the intensity computation is based on a probabilistic interpretation of the ray distribution at the source. Therefore, MC ray tracing is a very slow technique.

2.2 A New Approach: Ray Tracing in Phase Space

In two-dimensions the phase space (PS) of a line i is a 2D space where every ray is described by its position coordinate x on that line and its direction $\tau_i = n\sin(t_i)$, where n is the index of refraction in which the line is located and t_i is the angle that the ray forms with the optical axis, see [5]. We indicate with $\mathscr{P}_\mathrm{s} = [-a, a] \times [-0.99\, n_\mathrm{s}, 0.99\, n_\mathrm{s}]$ the source PS and with (x_1, τ_1) the coordinates of the rays on $\mathscr{P}_\mathrm{s}$. Similarly, $\mathscr{P}_\mathrm{t} = [-b, b] \times [-0.99\, n_\mathrm{t}, 0.99\, n_\mathrm{t}]$ is the target PS and (x_4, τ_4) the ray's coordinates on $\mathscr{P}_\mathrm{t}$, where $\tau_1 = n_\mathrm{s}\sin(t_1)$, $\eta = n_\mathrm{t}\sin(t_4)$ with n_s, n_t the index of refraction of the medium in which the source and the target are located, and t_1 and

t_4 the angles that the rays form with the optical axis at the source and the target, respectively. From now on, we assume that $n_s = n_t = 1$ as both $\mathscr{S}$ and $\mathscr{T}$ of the CPC in Fig. 1 are located in air, furthermore we indicate with $(q, \eta) = (x_4, \tau_4)$ the target coordinates of the rays on $\mathscr{P}_t$. Defining the ray's path as the sequence of the lines that the ray hits during its propagation, we note that the phase spaces can be partitioned into regions according to the path followed by the rays. We indicate with p the maximum number of possible paths, with $j = 1, \ldots, p$ the index of the paths and, with $\mathscr{R}_{s,j}$ and $\mathscr{R}_{t,j}$ the regions formed by the rays that follow path j in $\mathscr{P}_s$ and $\mathscr{P}_t$, respectively. Note that multiple reflections along the lines 2 and 3 of the CPC are possible. Therefore, many different regions are presented in target phase space.

Ray tracing in PS is based on the edge ray principle which states that there exists an optical map $\mathscr{M} : \mathscr{P}_s \mapsto \mathscr{P}_t$ such that $\mathscr{M}(x_1, \tau_1) = (q, \eta)$ for every $(x_1, \tau_1) \in \mathscr{P}_s$, [3]. Therefore, the regions $\mathscr{R}_{s,j}$ in $\mathscr{P}_s$ are mapped into the regions $\mathscr{R}_{t,j}$ in $\mathscr{P}_t$ by the map $\mathscr{M}$. Furthermore, the edge ray principle guarantees that also the boundaries $\partial\mathscr{R}_{s,j}$ are mapped into the boundaries $\partial\mathscr{R}_{t,j}$. Our method employs the map $\mathscr{M}$ to calculate the output photometric variables. In PS the luminance $L_t(q, \eta)$ at the target $\mathscr{T}$ is given by

$$L_t(q,\eta) > 0 \quad \text{for}(q,\eta) \in (\mathscr{R}_{t,j})_{j=1,\ldots,p}, \qquad L_t(q,\eta) = 0 \qquad \text{otherwise.} \tag{2}$$

Furthermore, the luminance is conserved along rays, hence, for a Lambertian source the luminance is constant inside the regions $\mathscr{R}_{t,j}$. The target intensity along a given direction η is computed through an integration of the target luminance L_t defined in Eq. (2). Denoting with $q_{j,1}$ and $q_{j,2}$ the position coordinates in $\mathscr{P}_t$ of the rays located on the boundaries $\partial\mathscr{R}_{t,j}$ where $q_{j,2} > q_{j,1}$, the target intensity along the direction η is given by:

$$I_{PS}(\eta) = \sum_{j=1}^{p} \int_{q_{j,1}(\eta)}^{q_{j,2}(\eta)} L_t(q,\eta)\mathrm{d}q = \sum_{j=1}^{p} \left(q_{j,2}(\eta) - q_{j,1}(\eta)\right), \tag{3}$$

where the second equality holds assuming a Lambertian source with $L_s = L_t = 1$. Note that relation (3), where the intensity is a function of the direction η instead of t, is valid when only two points $q_{j,1}(\eta)$ and $q_{j,2}(\eta)$ on the boundaries $\mathscr{R}_{t,j}$ are found (which is the case of the CPC in Fig. 1). In case $\partial\mathscr{R}_{t,j}$ has more than two intersection points with $\eta = const$, relation (3) should be modified accordingly.

For the computation of the output intensity the boundaries $\partial\mathscr{R}_{t,j}$ need to be determined for every $j = 1, \ldots, p$. To this purpose a triangulation is constructed on $\mathscr{P}_s$. The procedure starts dividing $\mathscr{P}_s$ into two equal triangles and tracing the rays located at the corner of every triangle. Then, the paths followed by these rays are considered. If there are at least two different paths, then one boundary is expected to cross the triangle. In that case three more rays with coordinates equal to the coordinates of the middle point of every side of the triangle are traced and again the paths of those rays are considered. The refinement procedure continues until the

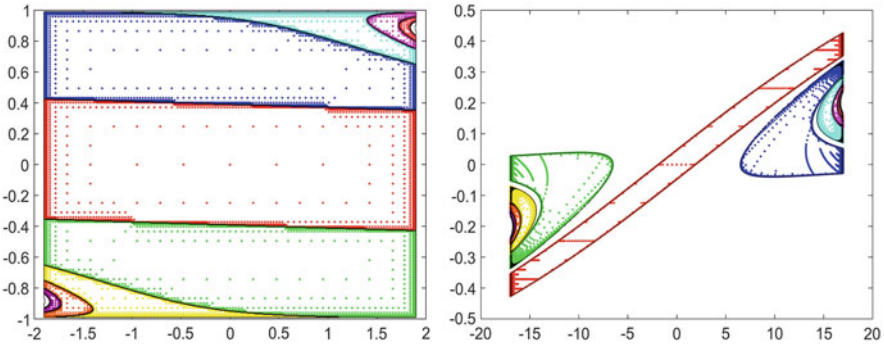


Fig. 2 $\mathscr{P}_s$ and $\mathscr{P}_t$ for the CPC (*left and right picture*). Around $2.9 \cdot 10^4$ rays are traced through the system and 13 different paths are found (multiple reflections with the reflectors can occur). All rays that follow the same path are depicted with the same color. The boundaries are depicted with *black lines* and they are determined connecting all the rays that follow the same path

paths of the rays located at the vertices of the triangle are equal. In case the rays located at the corners of the triangle have the same path, there is no need to refine again unless the size of the triangle is bigger than a fixed parameter. Therefore, two different parameters need to be determined to establish the size of the smallest and the largest triangle. To this purpose we employ the energy conservation from the source to the target, see [2] for details.

$\mathscr{P}_s$ and $\mathscr{P}_t$ of the CPC are shown in the left and the right pictures in Fig. 2, respectively. The rays that follow the same path are depicted with the same color. The triangulation refinement allows tracing more rays close to the boundaries of the regions $\mathscr{R}_{t,j}$. In Fig. 2 around $2.9{\cdot}10^4$ rays are traced and 13 different paths are found. Note that the rays that leave the extremes of the source with an angle close to $\pi/2$ or $-\pi/2$ can have multiple reflections along the reflectors. These rays are located in small regions $\mathscr{P}_s$ and $\mathscr{P}_t$. The regions are smaller when more reflections are considered. The finer the triangulation refinement is the more paths are detected and a better approximation of the boundaries is obtained. When the number of rays traced goes to infinity, more paths are found, however, the size of the regions formed by the rays that follow those paths goes to zero and then the intensity given by these rays can be neglected.

Every boundary $\mathscr{R}_{t,j}$ in Fig. 2 is approximated by a broken line where the vertices are the coordinates of the rays of the triangulation that follow the path j and, the intersection points between the line $\eta = \eta_k$ and $\partial\mathscr{R}_{t,j}$ are computed for every $j = 1, \ldots, p$. To calculate the intensity for all possible directions, a uniform partitioning $P : -0.99 \leq \eta_0 < \eta_1 < \cdots < \eta_{Nb} \leq 0.99$ of the interval $J = [-0.99, 0.99]$ is considered. Eventually, the intensity for each η_h, with $h = 0, 1, \ldots, Nb$, is obtained using relation (3). In order to compare ray tracing in PS with the already existing MC ray tracing, the intensity for MC needs to be determined. The same partitioning P of J is considered and, the number of rays that

fall into each bin $([\eta_h, \eta_{h+1}])_{h=0,\dots,Nb-1}$ is calculated for all $h \in \{0, \dots, Nb-1\}$. The intensity in the direction $\eta_k \in [\eta_h, \eta_{h+1}]$ is then given by:

$$\hat{I}_{MC}(\eta_k) = \frac{Nr([\eta_h, \eta_{h+1}])}{Nr([-0.99, 0.99])}, \tag{4}$$

for every $\left(\eta_k = \frac{1}{2}(\eta_{h+1} + \eta_h)\right)_{k=1,2,\dots,Nb}$, where we have indicated the number of rays that fall into the bin $[\eta_h, \eta_{h+1}]$ with $Nr([\eta_h, \eta_{h+1}])$ and the total number of rays with $Nr([-0.99, 0.99])$. To compare ray tracing on PS with MC ray tracing, the intensity I_{PS} is calculated over every bin $([\eta_h, \eta_{h+1}])_{h=0,\dots,Nb-1}$ and, the normalized intensity $\hat{I}_{PS}$ is considering dividing I_{PS} by the throughput at the target which in PS is defined as the area covered by regions with positive luminance (see [1]).

3 Numerical Results: Intensity at the Target of the CPC

In this section we provide numerical results that show a comparison between the PS and MC methods. The intensities found with the two methods are compared with a reference intensity $\hat{I}_{\text{ref}}$ obtained running MC ray tracing for 10^8 rays.

In Fig. 3 the intensity distribution at the target as a function of the angular parameter $\eta = \sin(t_4)$ is provided both for PS and MC ray tracing with a dotted red line and a blue line, respectively. The two intensities overlap each other, therefore the PS method computes the intensity correctly.

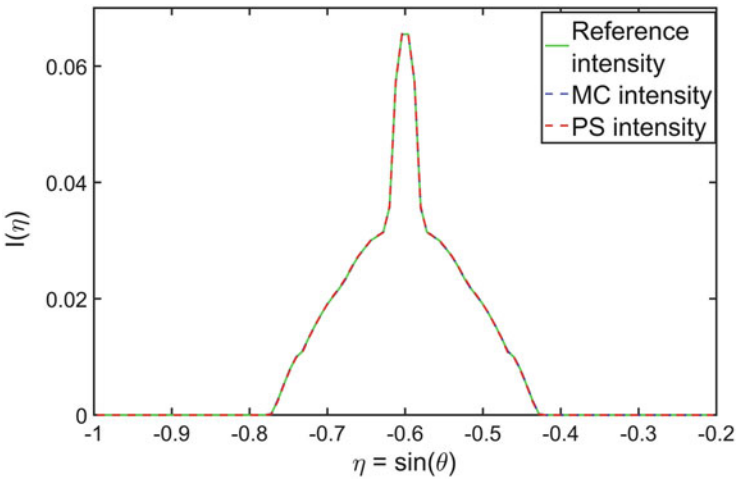


Fig. 3 Intensity distribution at the target of the CPC. The *dotted red line* represents the PS intensity ($5.84 \cdot 10^4$ rays are traced), the *dotted blue line* represents the MC intensity (around 10^7 rays are traced). The *green line* is the reference intensity obtained using MC ray tracing with $5 \cdot 10^8$ rays

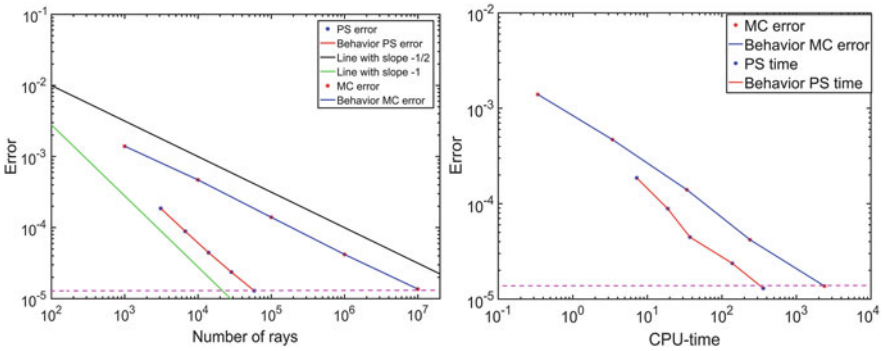


Fig. 4 Errors between the approximated intensities and the reference intensity both for MC ray tracing (*blue line*) and ray tracing in PS (*red line*). *Left*: error as a function of the number of rays traced. *Right*: error as a function of the CPU-time. Ray tracing on PS is four times faster than MC ray tracing. The numerical simulations are implemented using Matlab 2016a

The error between the approximate intensity $\hat{I}_{\mathrm{A}}$ ($A = PS, MC$) and the reference intensity is:

$$\text{error} = \frac{\sum_{h=1}^{Nb} |\hat{I}_{\mathrm{A}}(\eta_h) - \hat{I}_{\mathrm{ref}}(\eta_h)|}{Nb}. \tag{5}$$

In order to show how the error decreases increasing the number of rays traced, the approximated intensities are computed for different sets of rays both for the MC and PS method. The corresponding error is calculated using Eq. (5). The results are shown in Fig. 4 where we depict with a blue line the error for MC ray tracing and with a red line the error for ray tracing in PS. In the left figure the errors are depicted as a function of the number of rays traced in a logarithmic scale. A line with slope $-1/2$ (the black line) is drawn to demonstrate that MC ray tracing converges proportionally to $1/\sqrt{Nr}$, with Nr the number of rays traced. Then, drawing a line with slope -1 (the green line) we show that ray tracing in PS converges proportionally to $1/Nr$. Note that to obtain an accuracy given by an error equal to, say $1.3 \cdot 10^{-5}$, around 10^7 ray need to be traced for MC and only $5.8 \cdot 10^4$ rays for the PS method obtaining a reduction of more than 100. In the right picture in Fig. 4 the error as a function of the CPU-time is shown both for MC (blue line) and PS (red line) method. For our method, to avoid tracing the rays randomly, we need to take into account the path followed by any ray and compute the boundaries of all the regions formed by the rays that follow the same path. Despite this, we observe advantages in terms of the CPU-time for our method compared to the classical ray tracing approach.

4 Conclusions

A new ray tracing method which employs the phase space representation of the source and the target of the optical system was presented. The method was tested for a two-dimensional CPC. Numerical results show that the PS method outperforms MC ray tracing. A computation of the error between a reference intensity and the approximated intensities is provided both for ray tracing in PS and MC ray tracing. For the PS method the error as a function of the number of rays traced decreases proportionally to the inverse of the number of rays traced versus a speed of convergence proportional to the inverse of the square root of the number of rays traced obtained with MC ray tracing. A computation of the error as function of the CPU-time shows that ray tracing in PS is significantly faster than MC ray tracing. For the optical system considered in this work, only the reflection law was involved. The method is valid also for systems where refraction occurs (see [2]). Furthermore, ray tracing in PS can be generalized to systems with a non-constant luminance. Future perspectives are to take Fresnel reflection and scattering phenomena into account and to extend the method to optical systems in 3D.

References

1. Chaves, J.: Introduction to Nonimaging Optics. CRC Press, Boca Raton (2008)
2. Filosa, C., ten Thije Boonkkamp, J.H.M., IJzerman, W.L.: Ray tracing method in phase space for two-dimensional optical systems. Appl. Opt. **13**, 3599–3606 (2016)
3. Ries, H., Rabl, A.: Edge-ray principle of nonimaging optics. J. Opt. Soc. Am. A **11**, 2627–2632 (1994)
4. Ting, D.Z., McGill Jr., T.G.: Monte Carlo simulation of lighting emitting diode. Opt. Eng. **34**, 3545–3553 (1995)
5. Wolf, K.B.: Geometric Optics on Phase Space. Springer, Berlin (2004)

Homogenization of a Hyperbolic-Parabolic Problem with Three Spatial Scales

Liselott Flodén, Anders Holmbom, Pernilla Jonasson, Marianne Olsson Lindberg, Tatiana Lobkova, and Jens Persson

Abstract We study the homogenization of a certain linear hyperbolic-parabolic problem exhibiting two rapid spatial scales $\{\varepsilon, \varepsilon^2\}$. The homogenization is performed by means of evolution multiscale convergence, a generalization of the concept of two-scale convergence to include any number of scales in both space and time. In particular we apply a compactness result for gradients. The outcome of the homogenization procedure is that we obtain a homogenized problem of hyperbolic-parabolic type together with two elliptic local problems, one for each rapid scale.

1 Introduction

We will study the homogenization of the following hyperbolic-parabolic problem:

$$\begin{cases} \partial_{tt}^2 u_\varepsilon(x,t) + \partial_t u_\varepsilon(x,t) - \nabla \cdot \left(a\left(\frac{x}{\varepsilon}, \frac{x}{\varepsilon^2}\right) \nabla u_\varepsilon(x,t)\right) = f(x,t) & \text{in } \Omega_T, \\ u_\varepsilon(x,t) = 0 & \text{on } \partial\Omega_T, \\ u_\varepsilon(x,0) = g(x) & \text{in } \Omega, \\ \partial_t u_\varepsilon(x,0) = h(x) & \text{in } \Omega \end{cases} \tag{1}$$

where $f \in L^2(\Omega_T)$, $g \in H_0^1(\Omega)$ and $h \in L^2(\Omega)$. Moreover, we assume that a fulfils the following structure conditions:

i) $a \in C_\sharp\left(Y^2\right)^{N\times N}$

ii) $a\left(y^2\right)\xi \cdot \xi \geq \alpha |\xi|^2$ for all $y^2 \in \mathbb{R}^{2N}$, all $\xi \in \mathbb{R}^N$ and some $\alpha > 0$.

Homogenization means studying the limit process as $\varepsilon \to 0$ and finding the limit problem which admits the limit u, of the sequence u_ε, as its solution. The coefficient

L. Flodén • A. Holmbom • P. Jonasson • M.O. Lindberg (✉) • T. Lobkova • J. Persson
Department of Quality Technology and Management, Mechanical Engineering and Mathematics, Mid Sweden University, S-83125 Östersund, Sweden
e-mail: Lotta.Floden@miun.se; Anders.Holmbom@miun.se; Pernilla.Jonasson@miun.se; Marianne.OlssonLindberg@miun.se; Tatiana.Lobkova@miun.se; Jens.Persson@miun.se

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_94

b in this limit problem is characterized by so-called local problems. In our problem we have two rapid spatial scales and hence we will obtain two local problems.

Notations We denote $\Omega_T = \Omega \times (0,T)$ and $\partial\Omega_T = \partial\Omega \times (0,T)$. Moreover, we let $y^n = y_1, y_2, \ldots, y_n$, $dy^n = dy_1 dy_2 \cdots dy_n$, $s^m = s_1, s_2, \ldots, s_m$, $ds^m = ds_1 ds_2 \cdots ds_m$, $\mathscr{Y}_{n,m} = Y^n \times S^m$ with $Y^n = Y_1 \times Y_2 \times \cdots \times Y_n$ and $S^m = S_1 \times S_2 \times \cdots \times S_m$ where $S_1 = S_2 = \cdots = S_m = (0,1)$ and $Y_1 = Y_2 = \cdots = Y_n = (0,1)^N$.

2 Preliminaries

Our approach for the homogenization procedure in Sect. 3 is based on evolution multiscale convergence. We give the following definition.

Definition 1 (Evolution Multiscale Convergence) A sequence $\{u_\varepsilon\}$ in $L^2(\Omega_T)$ is said to $(n+1, m+1)$-scale converge to $u_0 \in L^2(\Omega_T \times \mathscr{Y}_{n,m})$ if

$$\int_{\Omega_T} u_\varepsilon(x,t)\, v\left(x, t, \frac{x}{\varepsilon_1}, \ldots, \frac{x}{\varepsilon_n}, \frac{t}{\varepsilon_1'}, \ldots, \frac{t}{\varepsilon_m'}\right) dxdt$$
$$\to \int_{\Omega_T} \int_{\mathscr{Y}_{n,m}} u_0(x,t,y^n,s^m)\, v(x,t,y^n,s^m)\, dy^n ds^m dxdt$$

for any $v \in L^2\left(\Omega_T; C_\sharp(\mathscr{Y}_{n,m})\right)$. We write

$$u_\varepsilon(x,t) \overset{n+1,m+1}{\rightharpoonup} u_0(x,t,y^n,s^m).$$

To proceed we need to distinguish some types of relationships between the scales. The scales in the lists $\{\varepsilon_1, \ldots, \varepsilon_n\}$ and $\{\varepsilon_1', \ldots, \varepsilon_m'\}$ are assumed to be microscopic, i.e. they tend to zero as ε does. We also assume that the lists are lists of well-separated scales. (For a definition of well-separatedness see [1].) Furthermore, the lists are also jointly well-separated according to Definition 2.8 in [4] and Definition 2.58 in [5].

The following theorem provides a characterization of evolution multiscale limits for gradients. Here the space $W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$ is defined as

$$W_2^1(0,T;H_0^1(\Omega),L^2(\Omega)) = \left\{v \in L^2\left(0,T;H_0^1(\Omega)\right) \mid \partial_t v \in L^2\left(0,T;H^{-1}(\Omega)\right)\right\}.$$

Theorem 1 *For $\{u_\varepsilon\}$ bounded in $W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$ and under the assumption of jointly well-separatedness on the lists of scales it holds that, up to a subsequence,*

$$u_\varepsilon(x,t) \to u(x,t) \qquad \text{in } L^2(\Omega_T),$$
$$u_\varepsilon(x,t) \rightharpoonup u(x,t) \qquad \text{in } L^2(0,T;H_0^1(\Omega))$$

and

$$\nabla u_\varepsilon(x,t) \overset{n+1,m+1}{\rightharpoonup} \nabla u(x,t) + \sum_{k=1}^{n} \nabla_{y_k} u_k\left(x,t,y^k,s^m\right),$$

where $u \in W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$, $u_1 \in L^2(\Omega_T \times S^m; H_\sharp^1(Y_1)/\mathbb{R})$ *and* $u_k \in L^2(\Omega_T \times \mathscr{Y}_{k-1,m}; H_\sharp^1(Y_k)/\mathbb{R})$ *for* $k = 2,\ldots,n$.

Proof See Theorem 4 in [2] or Theorem 2.74 in [5].

To handle certain problems with rapid scales in both space and time (see e.g. [2] and the last paragraph of this paper) another type of convergence is sometimes needed, which is called very weak evolution multiscale convergence. Here we give the definition for an arbitrary number of scales in both space and time.

Definition 2 (Very Weak Evolution Multiscale Convergence) A sequence $\{w_\varepsilon\}$ in $L^1(\Omega_T)$ is said to $(n+1,m+1)$-scale converge very weakly to $w_0 \in L^1(\Omega_T \times \mathscr{Y}_{n,m})$ if

$$\int_{\Omega_T} w_\varepsilon(x,t)\, v\left(x,\frac{x}{\varepsilon_1},\ldots,\frac{x}{\varepsilon_{n-1}}\right) c\left(t,\frac{t}{\varepsilon_1'},\ldots,\frac{t}{\varepsilon_m'}\right) \varphi\left(\frac{x}{\varepsilon_n}\right) dxdt$$
$$\to \int_{\Omega_T}\int_{\mathscr{Y}_{n,m}} w_0\left(x,t,y^n,s^m\right) v\left(x,y^{n-1}\right) c\left(t,s^m\right) \varphi\left(y_n\right) dy^n ds^m dxdt$$

for any $v \in D(\Omega; C_\sharp^\infty(Y^{n-1}))$, $\varphi \in C_\sharp^\infty(Y_n)/\mathbb{R}$ and $c \in D(0,T; C_\sharp^\infty(S^m))$ where

$$\int_{Y_n} w_0\left(x,t,y^n,s^m\right) dy_n = 0.$$

We write

$$w_\varepsilon(x,t) \overset{n+1,m+1}{\underset{vw}{\rightharpoonup}} w_0\left(x,t,y^n,s^m\right).$$

Here follows a compactness result for very weak evolution multiscale convergence.

Theorem 2 *For* $\{u_\varepsilon\}$ *bounded in* $W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$ *and under the assumption of jointly well-separatedness on the lists of scales it holds that, up to a subsequence,*

$$\frac{u_\varepsilon(x,t)}{\varepsilon_n} \overset{n+1,m+1}{\underset{vw}{\rightharpoonup}} u_n\left(x,t,y^n,s^m\right),$$

where $u_1 \in L^2(\Omega_T \times S^m; H_\sharp^1(Y_1)/\mathbb{R})$ *and* $u_n \in L^2(\Omega_T \times \mathscr{Y}_{n-1,m}; H_\sharp^1(Y_n)/\mathbb{R})$, $n = 2,3,\ldots$ *are the same as in Theorem 1.*

Proof See Theorem 7 in [2] or Theorem 2.78 in [5].

3 Homogenization

We are now prepared to homogenize (1). The weak formulation of (1) means that we search for u_ε such that

$$\begin{aligned}\int_{\Omega_T} & u_\varepsilon(x,t)\, v(x)\, \partial_{tt}^2 c(t) - u_\varepsilon(x,t)\, v(x)\, \partial_t c(t) \\ & + a\left(\frac{x}{\varepsilon}, \frac{x}{\varepsilon^2}\right) \nabla u_\varepsilon(x,t) \cdot \nabla v(x)\, c(t)\, dxdt \\ & = \int_{\Omega_T} f(x,t)\, v(x)\, c(t)\, dxdt\end{aligned} \tag{2}$$

for all $v \in H_0^1(\Omega)$ and $c \in D(0,T)$. It turns out that (see [3]) there exists for each $\varepsilon > 0$ a unique solution u_ε such that the sequence $\{u_\varepsilon\}$ is bounded uniformly with respect to the $L^\infty\left(0,T,H_0^1(\Omega)\right)$-norm and $\{\partial_t u_\varepsilon\}$ with respect to the $L^\infty\left(0,T,L^2(\Omega)\right)$-norm. This implies that the sequence $\{u_\varepsilon\}$ is bounded in $W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$ and hence Theorems 1 and 2 can be applied.

In the proof of the following theorem we will choose different test functions in the weak formulation (2) to obtain the homogenized problem and the two local problems.

Theorem 3 *Let $\{u_\varepsilon\}$ be a sequence of solutions in $W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$ to (1). Then it holds, up to a subsequence, that*

$$\begin{aligned} u_\varepsilon(x,t) &\to u(x,t) \qquad \text{in } L^2(\Omega_T), \\ u_\varepsilon(x,t) &\rightharpoonup u(x,t) \qquad \text{in } L^2(0,T;H_0^1(\Omega)) \end{aligned}$$

and

$$\nabla u_\varepsilon(x,t) \overset{3,1}{\rightharpoonup} \nabla u(x,t) + \sum_{k=1}^{2} \nabla_{y_k} u_k\left(x,t,y^k\right),$$

where $u = u(x,t) \in W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$ is the unique solution to

$$\begin{cases} \partial_{tt}^2 u(x,t) + \partial_t u(x,t) - \nabla \cdot (b \nabla u(x,t)) = f(x,t) & \text{in } \Omega_T, \\ u(x,t) = 0 & \text{on } \partial\Omega_T, \\ u(x,0) = g(x) & \text{in } \Omega, \\ \partial_t u(x,0) = h(x) & \text{in } \Omega. \end{cases}$$

Here the coefficient b is identified by the local problems

$$-\nabla_{y_2}\cdot\left(a\left(y^2\right)\left(\nabla u\left(x,t\right)+\sum_{k=1}^{2}\nabla_{y_k}u_k\left(x,t,y^k\right)\right)\right)=0,$$

and

$$-\nabla_{y_1}\cdot\int_{Y_2}a\left(y^2\right)\left(\nabla u\left(x,t\right)+\sum_{k=1}^{2}\nabla_{y_k}u_k\left(x,t,y^k\right)\right)dy_2=0,$$

through

$$b\nabla u\left(x,t\right)=\int_{Y^2}a(y^2)\left(\nabla u\left(x,t\right)+\sum_{k=1}^{2}\nabla_{y_k}u_k\left(x,t,y^k\right)\right)dy^2,$$

where $u_1\in L^2(\Omega_T;H^1_\sharp(Y_1)/\mathbb{R})$ *and* $u_2\in L^2(\Omega_T\times Y_1;H^1_\sharp(Y_2)/\mathbb{R})$.

Proof From (4.12)–(4.14) in the proof of Theorem 4.1 in [3], the structure conditions on the coefficient a, stated in the list i)–ii) above, ensures that $\{u_\varepsilon\}$ is bounded in $W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$. This, and the fact that the scales are well-separated, implies, according to Theorem 1, that

$$\begin{aligned}u_\varepsilon\left(x,t\right)&\to u\left(x,t\right)\qquad&&\text{in } L^2(\Omega_T),\\ u_\varepsilon\left(x,t\right)&\rightharpoonup u\left(x,t\right)\qquad&&\text{in } L^2(0,T;H_0^1(\Omega))\end{aligned}$$

and

$$\nabla u_\varepsilon\left(x,t\right)\overset{3,1}{\rightharpoonup}\nabla u\left(x,t\right)+\sum_{k=1}^{2}\nabla_{y_k}u_k\left(x,t,y^k\right)$$

up to a subsequence, where $u\in W_2^1(0,T;H_0^1(\Omega),L^2(\Omega))$, $u_1\in L^2(\Omega_T;H^1_\sharp(Y_1)/\mathbb{R})$ and $u_2\in L^2(\Omega_T\times Y_1;H^1_\sharp(Y_2)/\mathbb{R})$.

For the homogenized problem, we let $\varepsilon\to 0$ in (2) and use Theorem 1 to obtain

$$\begin{aligned}&\int_{\Omega_T}u\left(x,t\right)v\left(x\right)\partial_{tt}^2c\left(t\right)-u\left(x,t\right)v\left(x\right)\partial_tc\left(t\right)\\ &+\int_{Y^2}a\left(y^2\right)\left(\nabla u\left(x,t\right)+\sum_{k=1}^{2}\nabla_{y_k}u_k\left(x,t,y^k\right)\right)dy^2\cdot\nabla v\left(x\right)c\left(t\right)dxdt\\ &=\int_{\Omega_T}f\left(x,t\right)v\left(x\right)c\left(t\right)dxdt.\end{aligned}$$

To find the local problems we choose the test functions in the weak form (2) as follows:

$$v(x) = \varepsilon^p v_1(x)\, v_2\left(\frac{x}{\varepsilon}\right) v_3\left(\frac{x}{\varepsilon^2}\right),$$

where $v_1 \in D(\Omega)$, $v_2 \in C^\infty_\sharp(Y_1)$, $v_3 \in C^\infty_\sharp(Y_2)/\mathbb{R}$, for different values of p. For the first local problem we choose $p = 2$. After differentiating we let $\varepsilon \to 0$ and obtain, from Theorem 1,

$$\int_{\Omega_T}\int_{Y^2} a\left(y^2\right)\left(\nabla u(x,t) + \sum_{k=1}^{2} \nabla_{y_k} u_k\left(x,t,y^k\right)\right) v_1(x)\, v_2(y_1)$$

$$\cdot \nabla_{y_2} v_3(y_2)\, c_1(t)\, dy^2 dx dt = 0.$$

By applying the Variational Lemma (Proposition 18.2 in [6]) we arrive at

$$\int_{Y_2} a\left(y^2\right)\left(\nabla u(x,t) + \sum_{k=1}^{2} \nabla_{y_k} u_k\left(x,t,y^k\right)\right) \cdot \nabla_{y_2} v_3(y_2)\, dy_2 = 0$$

a.e. in $\Omega_T \times Y_1$ for all $v_3 \in C^\infty_\sharp(Y_2)/\mathbb{R}$.

For the second local problem we choose $p = 1$, $v_2 \in C^\infty_\sharp(Y_1)/\mathbb{R}$, and $v_3 \equiv 1$. Again, differentiating and letting $\varepsilon \to 0$, according to Theorem 1, we get

$$\int_{\Omega_T}\int_{Y^2} a\left(y^2\right)\left(\nabla u(x,t) + \sum_{k=1}^{2} \nabla_{y_k} u_k\left(x,t,y^k\right)\right) \cdot v_1(x)\, \nabla_{y_1} v_2(y_1)\, c_1(t)\, dy^2 dx dt = 0$$

and, according to the Variational Lemma,

$$\int_{Y^2} a\left(y^2\right)\left(\nabla u(x,t) + \sum_{k=1}^{2} \nabla_{y_k} u_k\left(x,t,y^k\right)\right) \cdot \nabla_{y_1} v_2(y_1)\, dy^2 = 0$$

a.e. in Ω_T for all $v_2 \in C^\infty_\sharp(Y_1)/\mathbb{R}$.

3.1 A Note on a Case with Rapid Oscillations in Time

The results can be extended to cases with rapid temporal oscillations such as the one we obtain if we replace the coefficient in (1) with e.g. $a\left(\frac{x}{\varepsilon}, \frac{x}{\varepsilon^2}, \frac{t}{\varepsilon}, \frac{t}{\varepsilon^3}\right)$ if we postulate that $\{u_\varepsilon\}$ is still bounded in $W_2^1(0,T; H_0^1(\Omega), L^2(\Omega))$. The essential difference appears in the local problems while the homogenized problem is subject

only to the most obvious changes. To find the local problems we use test functions of the type

$$v(x) = \varepsilon^p v_1(x)\, v_2\left(\frac{x}{\varepsilon}\right) v_3\left(\frac{x}{\varepsilon^2}\right) \text{ and } c(t) = c_1(t)\, c_2\left(\frac{t}{\varepsilon}\right) c_3\left(\frac{t}{\varepsilon^3}\right),$$

where $v_1 \in D(\Omega)$, $v_2 \in C^\infty_\sharp(Y_1)$, $v_3 \in C^\infty_\sharp(Y_2)/\mathbb{R}$, $c_1 \in D(0,T)$, $c_2 \in C^\infty_\sharp(S_1)$ and $c_3 \in C^\infty_\sharp(S_2)$, for different values of p in the weak form of the problem. We first choose $p = 4$ and apply very weak $(3,3)$-scale convergence to remove the dependence of s_2 in u_2. Then we choose $p = 5$, let $v_3 \equiv 1$, $v_2 \in C^\infty_\sharp(Y_1)/\mathbb{R}$ and apply very weak $(2,3)$-scale convergence to remove the dependence of s_2 in u_1. Then we identify the two local problems. We first choose $p = 4$, $c_3 \equiv 1$ and apply $(3,2)$-scale convergence and very weak $(3,2)$-scale convergence to obtain the first local problem

$$-\nabla_{y_2} \cdot \left(\left(\int_{S_2} a\left(y^2, s^2\right) ds_2 \right) \left(\nabla u(x,t) + \sum_{k=1}^{2} \nabla_{y_k} u_k\left(x, t, y^k, s_1\right) \right) \right) = 0.$$

Next we let $p = 1$, $c_3 = v_3 \equiv 1, v_2 \in C^\infty_\sharp(Y_1)/\mathbb{R}$ and utilize $(2,2)$-scale convergence and very weak $(2,2)$-scale convergence. We arrive at the second local problem

$$\partial^2_{s_1 s_1} u_1 - \nabla_{y_1} \cdot \int_{Y_2} \left(\int_{S_2} a\left(y^2, s^2\right) ds_2 \right) \left(\nabla u(x,t) + \sum_{k=1}^{2} \nabla_{y_k} u_k\left(x, t, y^k, s_1\right) \right) dy_2 = 0.$$

References

1. Allaire, G., Briane, M.: Multiscale convergence and reiterated homogenisation. Proc. R. Soc. Edinb. Sect. A **126**, 297–342 (1996)
2. Flodén, L., Holmbom, A., Olsson Lindberg, M., Persson, J.: Homogenization of parabolic equations with an arbitrary number of scales in both space and time. J. Appl. Math. **2014**, 16 pp. (2014)
3. Migórski, S.: Homogenization of hyperbolic-parabolic equations in perforated domains. Univ. Iagel. Acta Math. **1996**, 59–72 (1996)
4. Persson, J.: Homogenisation of Monotone parabolic problems with several temporal scales. Appl. Math. **57**, 191–214 (2012)
5. Persson, J.: Selected topics in homogenization [Ph.D. thesis]. Mid Sweden University, Sundsvall (2012)
6. Zeidler, E.: Nonlinear Functional Analysis and Its Applications IIA. Springer, Berlin (1990)

A Heuristic Method to Optimize High-Dimensional Expensive Problems: Application to the Dynamic Optimization of a Waste Water Treatment Plant

Alberto Garre, Pablo S. Fernandez, Julio R. Banga, and Jose A. Egea

Abstract The mathematical description of industrial processes usually requires the use of models consisting of large systems of differential and algebraic equations. The numerical simulations of such models may lead to high computation times, therefore, making optimization unaffordable using classical optimization methods.

This contribution describes an evolutionary approach for the optimization of computationally expensive, highly dimensional problems. The performance of the algorithm has been compared against well known surrogate based optimization methods using classical benchmark functions. The results show that our method outperforms the reference methods, specially for the high dimensional case.

The proposed algorithm has been applied to the optimization of the operation parameters of a waste water treatment plant, using dynamic profiles. The algorithm has been able to produce better solutions than those obtained previously using static profiles.

1 Introduction

The models describing many industrial processes often consist in large systems of differential and algebraic equations (DAEs). The numerical simulation of such models in standard workstations can be very time consuming. Performing optimization

A. Garre • P.S. Fernandez
Departamento de Ingeniería de Alimentos y del Equipamiento Agrícola, Instituto de Biotecnología Vegetal, Universidad Politécnica de Cartagena (ETSIA), Paseo Alfonso XIII, 48, 30203 Cartagena, Spain
e-mail: alberto.garre@upct.es; pablo.fernandez@upct.es

J.R. Banga
BioProcess Engineering Group, IIM-CSIC, C/ Eduardo Cabello 6, 36208 Vigo, Spain
e-mail: julio@iim.csic.es

J.A. Egea (✉)
Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Antiguo Hospital de Marina (ETSII), Av. Dr. Fleming S/N, 30202 Cartagena, Spain
e-mail: josea.egea@upct.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_95

tasks with these models results in unaffordable computation times (from minutes to hours), thus, there is a need of optimization methods that reduce the number of simulations to locate optimal solutions [3]. The use of surrogate model-based algorithms, and in particular kriging and radial basis functions, has been subject of study by several authors in different disciplines [5]. They can be defined as statistical interpolation/prediction methods which make use of a surrogate model to estimate the location of the optimal solution in the real model. One of the main drawbacks of these methods is that the overhead associated to building the surrogate models increases with the problem dimension and with the number of simulations performed. Since many problems in engineering optimization and design are high dimensional, the applicability of these methods is limited. Even if recent advances have been made in the field of large-scale expensive optimization [6], it is still a subject under research.

Equation (1) presents a general formulation for an optimization problem including DAEs. It consists in the minimization of a cost function $C(y, x)$, where y and x are, respectively, the state and decision variables of the system. These variables are related through the differential-algebraic system $f(\dot{y}, y, x)$. The initial values at time t_0 of the state variables are given by y_0. The decision variables are subjected to the equality conditions $h(y, x)$ and the inequality conditions $g(y, x)$. Moreover, they must remain between the bounds defined by x_L and x_U belonging to $\mathbb{R}^d$, where d is the problem dimension.

$$\begin{aligned} \min C(y, x) \\ f(\dot{y}, y, x) = 0 \\ y(t_0) = y_0 \\ h(y, x) = 0 \\ g(y, x) \leq 0 \\ x_L \leq x \leq x_U \in \mathbb{R}^d \end{aligned} \tag{1}$$

In this contribution, an evolutionary approach designed for computationally expensive, high dimensional problems is presented. Its performance has been compared against other commonly used algorithms in a set of benchmark functions. The algorithm has been applied to the optimization of the operation parameters of a waste water treatment plant (WWTP).

2 Description of the Algorithm

The optimization algorithm presented here is intended for highly dimensional, computationally expensive optimization problems. For this purpose, three heuristic strategies have been implemented: (1) a solution combination method for finding

promising points, (2) a performance index (PI) acting as pseudo-surrogate model and (3) an evolutionary population update scheme.

Points 1 and 3 are typical steps in evolutionary algorithms for optimization. The computation begins with an initial population of n members (x^1, x^2, ..., x^n) of dimension d each, that can be generated using common experimental design schemes. In this paper we have used a Latin square sampling. The cost function is, then, evaluated at every solution of the population. Next, the population is sorted according to the solutions' fitness function, so that, in a minimization problem, x^1 has the lowest fitness function and x^n has the highest. In a next step, the combination method is called for the generation of new candidates to join the population. Finally, the population is updated following a heuristic scheme. If the stopping condition is not fulfilled, the new solutions are sorted and new candidates are, again, generated using the combination method.

The solution combination method was described in [2]. The offspring solutions are generated within a hypercube defined from a pair of "parent" solutions (x^i, x^j). The position of the vertexes of the hypercube (c_1, c_2) are defined in Eq. (2). It is centred on x^i and its size depends on the distance between the "parent" solutions (through parameter p), as well as on their relative quality (parameters α and β). The parameters are defined as follows: $p = (x^i - x^j)/2$. $\beta = (|j-1|-1)(n-2)$ evaluates the relative quality of both points. Finally, α equals 1 if the cost function is lower at x^i than at x^j and -1 otherwise.

$$\begin{aligned} c_1 &= x^i - p\,(1 + \alpha\,\beta) \\ c_2 &= x^i + p\,(1 + \alpha\,\beta) \end{aligned} \tag{2}$$

The evaluation of the cost function at each offspring solution can be computationally expensive. To avoid low quality costly simulations, a performance index (PI) for each offspring solution (x_i) is used as surrogate model for each new created solution. This index is constructed as presented in Eq. (3), where $x_{c,i}$ represents the point of the population closest to point x_i according to the norm $\mathcal{N}$. This norm is defined as the minimum difference between the components of the two vectors. It has been selected due to its low computational cost and its good scaling with the number of dimensions. Note that the PI has been defined for positive cost functions (like the ones studied in this contribution).

$$PI\,(x_i) = \frac{C\,(x_{c,i})}{\mathcal{N}\,(x_i, x_{c,i})^k} \tag{3}$$

This Performance Index can be seen as a ratio between quality of the proposed point (numerator) and the knowledge of the area (denominator). Reducing the value of the coefficient k assigns more weight the quality of the proposed point. In the calculations presented here, it has been set to 1.

In every iteration of the algorithm, the best offspring solution (according to the PI) can replace its parent solution x^i if the former has a lower PI. Once this has been

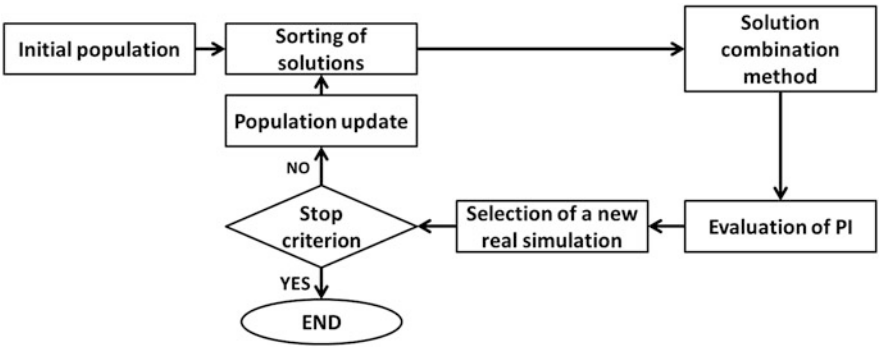


Fig. 1 Overview of the optimization algorithm proposed

performed over all the offspring solution, the real fitness function is evaluated at the offspring with the best PI. If the norm to the closest point in the population is greater than a given tolerance, it is added to the population (i.e. its size is increased in one unit, $n = n + 1$). This scheme controls the population size, allowing only it to increase if really high quality points are included.

The key point of this procedure is to avoid the computation of a surrogate model (e.g. a kriging model) that can also be expensive to build when the problem dimension and/or the number of solutions used to build it is high. The performance index computed here is much cheaper than the mentioned method and other typically used in surrogate-based optimization. The whole procedure is illustrated in Fig. 1.

3 Benchmark Against Common Optimization Functions

The performance of the optimization algorithm proposed has been compared to the one of other commonly used optimization algorithms, namely the Kriging and EGO algorithms from the SURROGATES toolbox for Matlab [8].

The Griewank, Levy and Ackley functions, commonly used for benchmarking optimization algorithms [4], have been selected for the comparison. The three functions are defined for $\mathbb{R}^d$ and have a global minimum $f^\star = 0$. The following values have been assigned for the parameters of the Ackley function: $a = 20$, $b = 0.2$ and $c = 2\pi$. The three functions have been evaluated in the hypercube recommended in [4].

The optimization of the benchmark functions has been performed for the case with 2 and 10 dimensions using the Kriging and EGO methods implemented in SURROGATES and our proposed method. Figure 2 presents the convergence plot obtained on each case. For the Griewank function, every method converges to the global optimum. However, the proposed method clearly outperforms the other two,

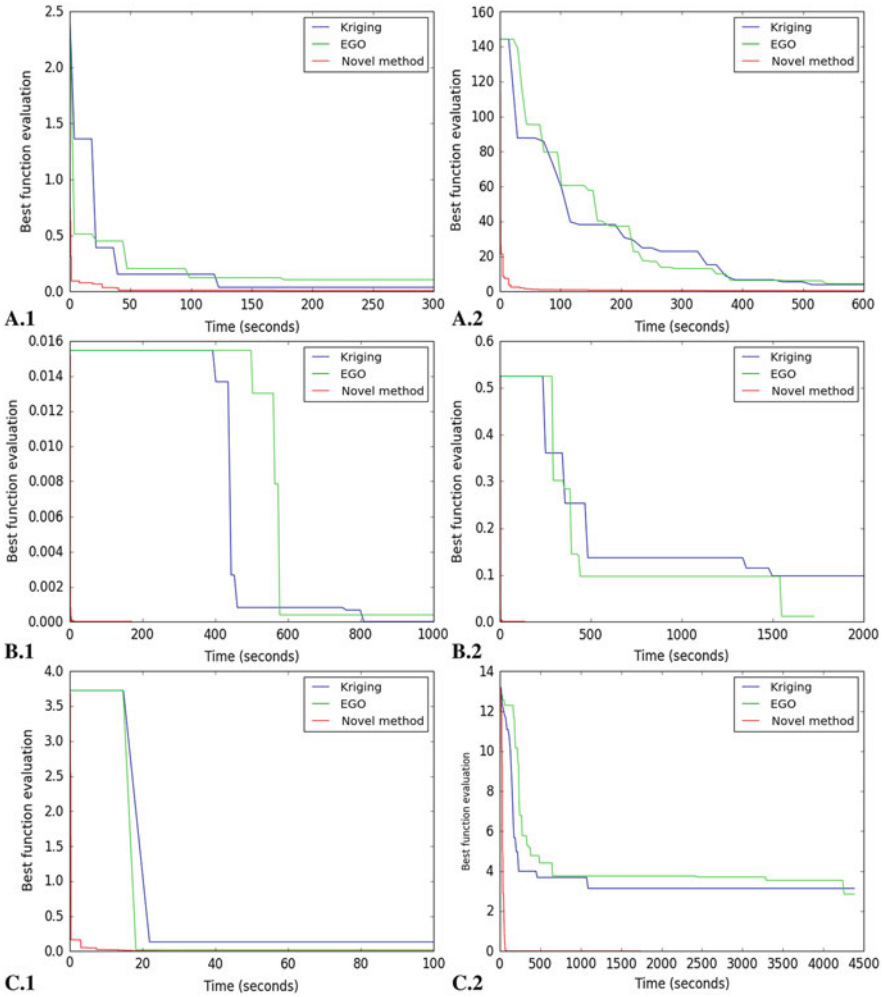


Fig. 2 Convergence plot of the optimization of the benchmark functions: Griewank (**a**), Levy (**b**) and Ackley (**c**) for two variables (1) and ten variables (2) cases

reaching the optimum faster. Better performance is achieved when a large number of variables is considered (Fig. 2(a2, b2 and c2)). The better performance of the proposed method is even more evident for Levy and Ackley functions. The Kriging and EGO methods did not converge to the global optimum for both functions in the iterations performed for the 10 dimensional case, while our method was able to reach the global optimum on every case considered.

The Kriging and EGO methods produced an out of memory error when 20 or more decision variables were considered. The proposed method was able to calculate a solution for up to 40 variables (results not shown).

4 Case Study: Dynamic Optimization of a Waste Water Treatment Plant

The optimization algorithm presented has been applied to the optimization of the operation parameters of a WWTP. The model developed in [1] represents a chemical engineering process and it is highly nonlinear. It consists of a systems of around 100 DAEs, making its computation costly (over 30 s on a 3.40 GHz i7-3770 PC). The cost function described in [7], which considers both economic and environmental factors, has been used.

The decision variables selected are those defining the optimal aeration ($KLA5$) and recirculation (Q_{intr}) profiles during 14 days. Stepwise profiles have been used, divided in 10 different periods whose length is also an optimization variable. Therefore, 40 optimization variables (length of 10 time periods and the magnitude of both variables). The problem is formulated as a dynamic optimization problem. Bounds for the state variables have been defined to keep physically feasible results.

Due to the large number of decision variables in the case study (40), the EGO and Kriging methods could not be applied. Hence, only the proposed model was used. The convergence was assessed by launching eight optimizations using different initial populations and different values of the parameters of the optimization algorithm. The convergence plot shown in Fig. 3 shows that the evolution of the solution was consistent through every case.

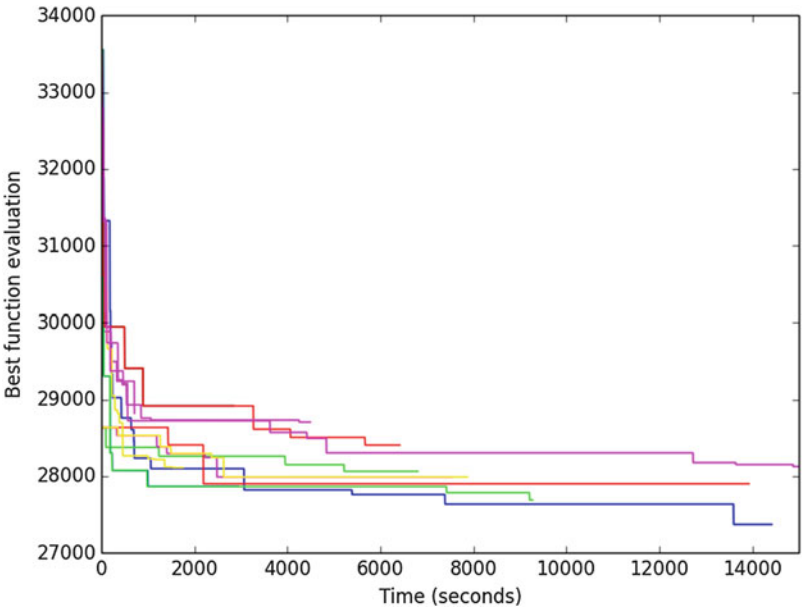


Fig. 3 Convergence plot for the optimization of the WWTP using different parameters of the optimization algorithm and initial populations

5 Conclusions

An algorithm for the optimization of computationally expensive, highly dimensional models has been presented in this contribution. The algorithm implements three heuristic strategies mainly based on evolutionary computation: a solution combination method, a performance index acting as pseudo-surrogate and population update scheme.

The proposed algorithm has been compared with two commonly used optimization methods (Kriging and EGO) using as benchmark Griewank, Levy and Ackley functions. The proposed method outperforms tow popular algorithms implemented in the SURROGATES toolbox, even for the two dimensional case. Those algorithms were unable to produce a solution for 20 or more variables, whereas the proposed method could calculate a solution for the 40 dimensional case.

The method has successfully been applied to the optimization of the operation parameters of a WWTP. The problem has been designed using dynamic optimization profiles, with 40 decision variables in total, consistently obtaining excellent solutions using a low number of function evaluations.

Acknowledgements The financial support of this research work was provided by the Ministry of Economy and Competitiveness (MINECO) of the Spanish Government and European Regional Development Fund (ERDF) through projects AGL2013-48993-C2-1-R and DPI2011-28112-C04-(03, 04). Alberto Garre (BES-2014-070946) is grateful to the MINECO for awarding him a predoctoral grant.

References

1. Copp, J.: The COST Simulation Benchmark Description and Simulator Manual. COST (European Cooperation in the field of Scientific and Technical Research). Brussels, Belgium (2001)
2. Egea, J.A., Marti, R., Banga, J.R.: An evolutionary method for complex-process optimization. Comput. Oper. Res. **37**, 315–324 (2010)
3. Forrester, A.I.J., Keane, A.J.: Recent advances in surrogate-based optimization. Prog. Aerosp. Sci. **45**, 50–79 (2009)
4. HeuristicLab: a paradigm-independent and extensible environment for heuristic optimization. Available via DIALOG. http://dev.heuristiclab.com/trac.fcgi/wiki/Documentation/Reference/Test%20Functions. Cited 22 Oct 2016
5. Jones, D.R.: A taxonomy of global optimization methods based on response surfaces. J. Global Optim. **21**, 345–383 (2001)
6. Shan, S., Wang, G.G.: Survey of modeling and optimization strategies for high-dimensional design problems. Struct. Multidiscip. Optim. **41**, 219–241 (2010)
7. Vanrolleghem, P.A., Jeppsson, U., Carstensen, J., Carlsson, B., Olsson, G.: Integration of wastewater treatment plant design and operation – a systematic approach using cost functions. Water Sci. Technol. **34**(3–4), 159–171 (1996)
8. Viana, FAC: SURROGATES Toolbox User's Guide, Gainesville, FL, USA, version 3.0 (2011) Available via DIALOG. https://sites.google.com/site/srgtstoolbox. Cited 21 Oct 2016

Reduced Basis Method Applied to a Convective Instability Problem

Henar Herrero, Yvon Maday, and Francisco Pla

Abstract Numerical reduced basis methods are instrumental to solve parameter dependent partial differential equations problems in case of many queries. Bifurcation and instability problems have these characteristics as different solutions emerge by varying a bifurcation parameter. Rayleigh-Bénard convection is an instability problem with multiple steady solutions and bifurcations by varying the Rayleigh number. In this paper the eigenvalue problem of the corresponding linear stability analysis has been solved with this method. The resulting matrices are small, the eigenvalues are easily calculated and the bifurcation points are correctly captured. Nine branches of stable and unstable solutions are obtained with this method in an interval of values of the Rayleigh number. Different basis sets are considered in each branch. The reduced basis method permits one to obtain the bifurcation diagrams with much lower computational cost.

1 Introduction

Bifurcations and instabilities in differential equations are features that allow the explanation of many phenomena in nature and in particular in fluid dynamics [6]. An example is the Rayleigh-Bénard convection problem [3, 16]. The model equations in this case are the incompressible Navier-Stokes equations coupled with a heat equation under the Boussinesq approximation. Here the conductive solution becomes unstable for a critical vertical temperature gradient beyond a certain threshold and therefore a convective motion sets in, and, depending on boundary

H. Herrero • F. Pla (✉)
Dpto. Matemáticas, U. Castilla-La Mancha, 13071 Ciudad Real, Spain
e-mail: Henar.Herrero@uclm.es; Francisco.Pla@uclm.es

Y. Maday
Laboratoire Jacques-Louis Lions, Sorbone Universités, UPMC Univ Paris 06, UMR 7598, F-75005 Paris, France

Institut Universitaire de France, Paris, France

Division of Applied Maths, Brown University, Providence, RI, USA
e-mail: maday@ann.jussieu.fr

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_96

conditions and other external physical parameters, new convective patterns occur [6]. In [12, 13] a Rayleigh-Bénard problem is studied under the perspective of looking for the bifurcation diagrams. The different solutions and successive bifurcations when the temperature gradients increase are obtained based on a branch continuation technique.

The reduced basis method is a meaningful numerical technique to solve problems of partial differential equation for a large amount of values of the bifurcation parameter with a reduced cost [1, 2, 8–11, 14, 15]. This method consists of the construction of a basis of solutions for different values of this parameter. In [7] the reduced basis method has been applied to this problem to obtain a branch of stable solutions, allowing one to prove the efficiency of the given method to obtain solutions for different values of the parameters. The aim of the present paper is to complete the bifurcations study with the calculation of the whole bifurcation diagram with all the solutions including the unstable ones, their linear stability, bifurcations among them and the capturing of the bifurcation points. All the steps solved with reduced basis.

2 Formulation of the Problem

The domain is a rectangle $\Omega = [0, \Gamma] \times [0, 1]$ containing a fluid that is heated from below at temperature T_0 and on the upper plate the temperature is $T_1 = T_0 - \Delta T = T_0 - \beta d$, where β is the vertical temperature gradient. The equations governing the system are the incompressible Navier-Stokes equations with the Boussinesq approximation coupled with a heat equation. The variables present in the problem are u_x and u_z, the components of the velocity vector field $\mathbf{u}$, θ the temperature, P the pressure, x and z the spatial coordinates and t the time. The magnitudes are expressed in dimensionless form as explained in [7]:

$$0 = \nabla \cdot \mathbf{u}, \qquad \text{in } \Omega, \tag{1}$$

$$\frac{1}{Pr}\left(\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}\right) = R\,\theta \mathbf{e}_z - \nabla P + \Delta \mathbf{u}, \qquad \text{in } \Omega, \tag{2}$$

$$\partial_t \theta + \mathbf{u} \cdot \nabla \theta - u_z = \Delta \theta, \qquad \text{in } \Omega. \tag{3}$$

Here $\mathbf{e}_z$ is the unitary vector in the vertical direction, R is the Rayleigh number that measures the effect of buoyancy and Pr is the Prandtl number. Motivated by mantle convection modelling, the Prandtl number Pr is considered infinite as in [12, 13], then the left hand side term in Eq. (2) can be made equal to zero. The bottom plate is rigid and the upper surface is non deformable and free slip. These conditions along with the thermal boundaries on the top and bottom plates are expressed as,

$$\mathbf{u} = \mathbf{0},\ \theta = 0 \text{ on } z = 0; \quad \theta = \partial_z u_x = u_z = 0 \text{ on } z = 1, \tag{4}$$

$$\partial_x \theta = \partial_x u_z = u_x = 0, \text{ on } x = 0, \text{ and on } x = \Gamma. \tag{5}$$

Equations in problem (1)–(5) are invariant under the symmetry, $\gamma : (x, z, u_x, u_z, \theta, P) \rightarrow (-x, z, -u_x, u_z, \theta, P)$. Then, if Φ is solution of the problem, $\gamma\Phi$ is a solution as well.

The stationary version of Eqs. (1)–(5) is discretized using the Legendre spectral method [4, 5]. In this work we have considered expansions of order $n = 35$ in the x-direction and $m = 13$ in the z-direction. The stability of the stationary solutions of (1)–(5) is determined by linear stability analysis as in [12]. In this analysis the stationary solutions are perturbed as follows:

$$\mathbf{u}(x, z, t) = \mathbf{u}^b(x, z) + \bar{\mathbf{u}}(x, z) \exp(\sigma t), \tag{6}$$

and similarity for the other fields, here superscript b indicates the corresponding solutions of the stationary problem and the bar refers to the perturbation. Expressions for the perturbed fields are introduced in Eqs. (1)–(3) and boundary conditions and the resulting equations are linearized. Therefore, we get the following eigenvalue problem (with bars now omitted):

$$\nabla \cdot \mathbf{u} = 0, \qquad \text{in } \Omega, \tag{7}$$

$$R\,\theta \mathbf{e}_z - \nabla P + \Delta \mathbf{u} = 0, \qquad \text{in } \Omega, \tag{8}$$

$$\Delta\theta - \mathbf{u^b} \cdot \nabla\theta - \mathbf{u} \cdot \nabla\theta^b + u_z = \sigma\theta, \qquad \text{in } \Omega. \tag{9}$$

together with boundary conditions (4)–(5). The resulting problem is an eigenvalue problem in σ. The sign of the real part of σ determines the stability of the stationary solution. If $Re(\sigma) < 0$ for any eigenvalue of the stationary state, then it is stable, while if there exists a value of σ such that $Re(\sigma) > 0$ then the stationary state becomes unstable. The condition $Re(\sigma) = 0$ defines the critical threshold of the bifurcation at which $Im(\sigma)$ may be either zero or not.

The aspect ratio considered is $\Gamma = 3.495$ as in [7, 12] and let the Rayleigh number R vary in the interval: [1000, 3000].

3 Numerical Method

The numerical method is based on the approximation of all the solutions by the linear combination of some appropriate preliminary computed solutions for some values of R, the same ones for every variable $\mathbf{u}$, θ, P. The reduced basis $\{\Psi_1, \Psi_2, \ldots, \Psi_N\}$ in the corresponding discrete space $X_N \equiv X_N^{\mathbf{u}} \times X_N^{\theta} \times X_N^{P}$ are obtained in a greedy fashion as it is explained in [7].

Once the reduced basis has been calculated, the Galerkin method with expansions in this basis is implemented to solve the eigenvalue problem (7)–(9) described in

Sect. 2. The eigenvalue problem is presented in its corresponding variational form as follows:

$$\begin{aligned} R\int_\Omega \theta_N v_z - \int_\Omega \nabla \mathbf{u}_N \cdot \nabla \mathbf{v} = 0, \quad \forall \mathbf{v} \in X_N^{\mathbf{u}}, \\ \int_\Omega (\mathbf{u}_N \cdot \nabla \theta_N^b + \mathbf{u}_N^b \cdot \nabla \theta_N - u_{N,z}) \cdot \phi + \int_\Omega \nabla \theta_N \cdot \nabla \phi \\ = -\sigma \int_\Omega \theta_N \cdot \phi, \quad \forall \phi \in X_N^\theta, \end{aligned} \tag{10}$$

where $\mathbf{u}_N = \sum_{i=1}^N \alpha_i \Psi_i^{\mathbf{u}}$ and $\theta_N = \sum_{i=1}^N \beta_i \Psi_i^\theta$. The eigenvalue problem (10) is then transformed into its discrete form as:

$$\begin{pmatrix} A_1 & B_1 \\ A_2 & B_2 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \sigma \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{b} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \tag{11}$$

where $\mathbf{b} = -\int_\Omega \theta_N \cdot \phi$ and A_1, B_1, A_2, B_2, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are defined as follow:

$$A_i, B_i \in \mathscr{M}_{N\times N}, i = 1, 2; \ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N); \ \boldsymbol{\beta} = (\beta_1, \ldots, \beta_N).$$

We are interested in the presence of eigenvalues that have a positive real part. If such eigenvalue exists the solution is unstable otherwise the solution is stable.

A rectification post-processing strategy has been incorporated. This post-processing is presented in [7], which consists of a change of basis from the reduced basis to Galerkin solutions on the values of R of the basis obtained.

4 Numerical Results and Conclusions

Figure 1 shows the bifurcation diagram, where the value of the combination of two coefficients of the Legendre expansion of the field u_x as a function of the Rayleigh number is presented. Solutions Φ are calculated by reconstructing all the solutions in each branch by the Galerkin approach to the corresponding post-processed reduced basis. The bifurcation points are exactly captured with the reduced basis method as can be seen in Fig. 1: namely a first pitchfork bifurcation at $R_{c1} = 1101$, a second pitchfork bifurcation at $R_{c2} = 1252$ and finally a secondary inverse pitchfork bifurcation at $R_{c3} = 1538$.

Figure 2 shows the norm of the difference between the stationary solution obtained with Legendre collocation and that obtained with a post-processed reduced basis method based on the Legendre collocation method for the branch Φ_2. The errors are $\mathscr{O}(10^{-7})$ in the stable part of the branch as can be seen in Fig. 2a, $\mathscr{O}(10^{-4})$ in the unstable part of the branch, and near the bifurcation point the error increases to $\mathscr{O}(10^{-2})$, as can be seen in Fig. 2b.

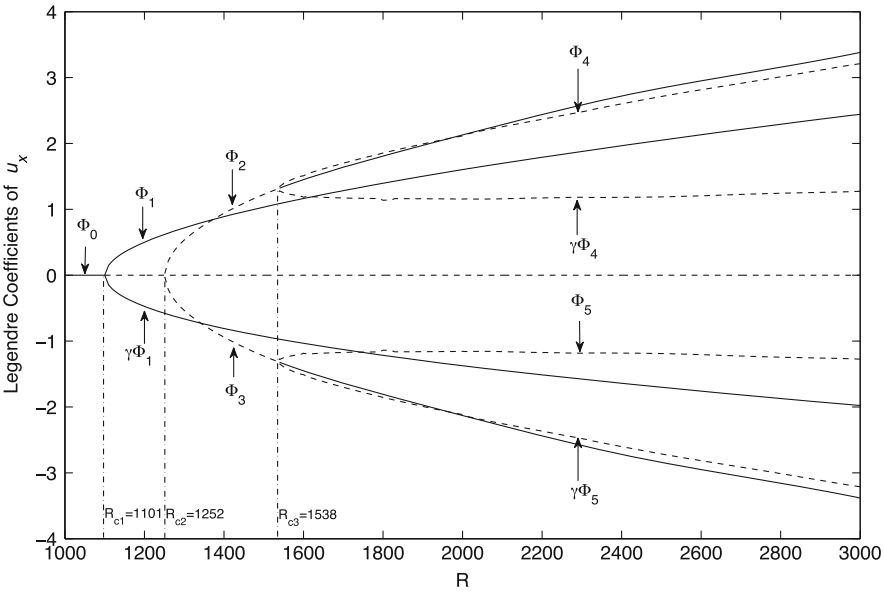


Fig. 1 Bifurcation diagram where the value of a combination of coefficients of the Legendre expansion of the field u_x as a function of the Rayleigh number is presented

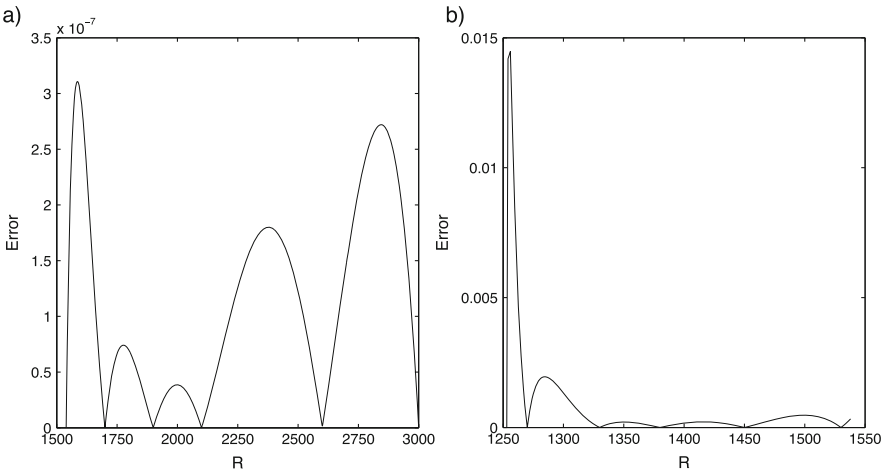


Fig. 2 (**a**) Norm of the difference between the stationary solution obtained with Legendre collocation and with a post-processed reduced basis method based on Legendre collocation for the stable part of branch Φ_2 in the interval of R [1539, 3000]. (**b**) Norm of the difference between the stationary solution obtained with Legendre collocation and with a post-processed reduced basis method based on Legendre collocation for the unstable part of branch Φ_2 in the interval of R [1253, 1538]

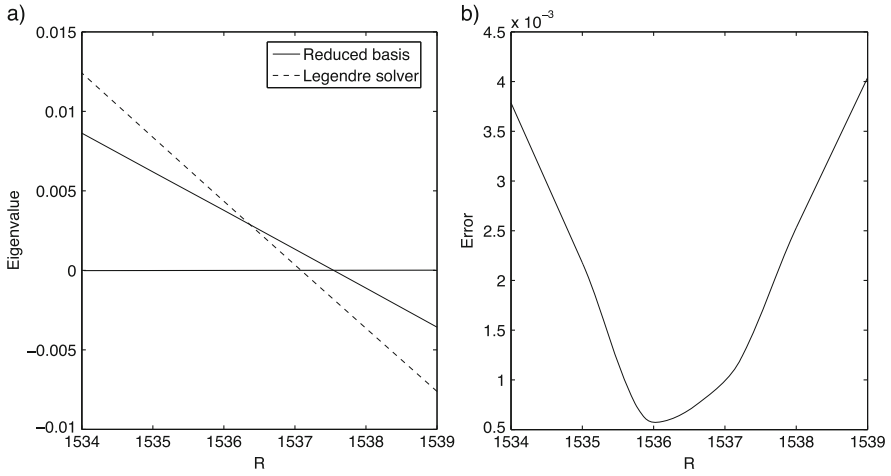


Fig. 3 (**a**) Eigenvalue with maximum real part near the bifurcation point. *Dashed line* for the Legendre solver and *solid line* for the post-processed reduced basis. (**b**) Norm of the difference between the real part of the eigenvalue with largest real part obtained with Legendre collocation and with a post-processed reduced basis method based on Legendre collocation

Figure 3a presents the eigenvalues as a function of R for the branch Φ_2. For both calculations, i.e., Legendre collocation solver and post-processed reduced basis method, eigenvalues cross the axis for the same value, $R_{c3} = 1538$. The same behaviour is observed for the rest of the bifurcation points at the values of $R_{c1} = 1101$ and $R_{c2} = 1252$. Therefore, we may conclude that the bifurcation points are correctly captured. Figure 3b shows the norm of the difference between the real part of the eigenvalue with largest real part obtained via Legendre collocation and by a post-processed reduced basis method based on Legendre collocation for the branch Φ_2 in the interval of R [1534, 1539] near the bifurcation point at $R_{c3} = 1538$. This difference is $\mathcal{O}(10^{-4})$ (where both approximations cross near the bifurcation point), increasing as we move away from the bifurcation point.

A significant advantage of solving the eigenvalue problem is provided by the use of a reduced basis, which is due to a large reduction in computational cost. This reduction arises due to the size of the matrices in the eigenvalue problem; in Legendre collocation the size of the matrix is 2016, while using a reduced basis with 8 elements it is only 16. This is reflected in the temporal computational cost, which is 122 s when using Legendre collocation and only 6 s using a reduced basis. Therefore the reduction is of a factor of 20 in time.

Acknowledgements This work was partially supported by the Research Grants MINECO (Spanish Government) MTM2015-68818-R, which include RDEF funds.

References

1. Almroth, B.O., Stern, P., Brogan, F.A.: Automatic choice of global shape functions in structural analysis. AIAA J. **16**, 525–528 (1978)
2. Barrett, A., Reddien, G.: On the reduced basis method. Z. Angew. Math. Mech. **75**(7), 543–549 (1995)
3. Bénard, H.: Les tourbillons cèllulaires dans une nappe liquide. Rev. Gen. Sci. Pures Appl. Bull. Assoc. **11**, 1261–1271 (1900)
4. Bernardi, C., Maday, Y.: Approximations spectrales de problèmes aux limites elliptiques. Springer, Berlin (1991)
5. Canuto, C., Hussaine, M.Y., Quarteroni, A., Zang, T.A.: Spectral Methods in Fluid Dynamics. Springer, Berlin (1988)
6. Chandrasekhar, S.: Hydrodynamic and Hydromagnetic Stability. Dover, New York (1982)
7. Herrero, H., Maday, Y., Pla, F.: RB (Reduced Basis) applied to RB (Rayleigh-Bénard). Comput. Methods Appl. Mech. Eng. **261–262**, 132–141 (2013)
8. Noor, A.K., Peters, J.M.: Bifurcation and post-buckling analysis of laminated composite plates via reduced basis techniques. Comput. Methods Appl. Mech. Eng. **29**, 271–295 (1981)
9. Noor, A.K., Peters, J.M.: Multiple-parameter reduced basis technique for bifurcation and post-buckling analysis of composite plates. Int. J. Numer. Methods Eng. **19**, 1783–1803 (1983)
10. Noor, A.K., Balch, C.D., Shibut, M.A.: Reduction methods for non-linear steadystate thermal analysis. Int. J. Numer. Methods Eng. **20**, 1323–1348 (1984)
11. Rozza, G., Huynh, D.B.P., Patera, A.T.: Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. Arch. Comput. Methods Eng. **15**(3), 229–275 (2008)
12. Pla, F., Mancho, A.M., Herrero, H.: Bifurcation phenomena in a convection problem with temperature dependent viscosity at low aspect ratio. Physica D **238**(5), 572–580 (2009)
13. Pla, F., Herrero, H., Lafitte, O.: Theoretical and numerical study of a thermal convection problem with temperature-dependent viscosity in an infinite layer. Physica D **239**(13), 1108–1119 (2010)
14. Porsching, T.A., Lin Lee, M.Y.: The reduced-basis method for initial value problems. SIAM J. Numer. Anal. **24**, 1277–1287 (1987)
15. Prud'homme, C., Rovas, D., Veroy, K., Maday, Y., Patera, A.T., Turinici, G.: Reliable real-time solution of parametrized partial differential equations: reduced-basis output bounds methods. J. Fluids Eng. **124**(1), 70–80 (2002)
16. Rayleigh, L.: On convection currents in a horizontal layer of fluid when the higher temperature is on the under side. Philos. Mag. **32**, 529–546 (1916)

Independent Loops Search in Flow Networks Aiming for Well-Conditioned System of Equations

Jukka-Pekka Humaloja, Simo Ali-Löytty, Timo Hämäläinen, and Seppo Pohjolainen

Abstract We approach the problem of choosing linearly independent loops in a pipeflow network as choosing the best-conditioned submatrix of a given larger matrix. We present some existing results of graph theory and submatrix selection problems, based on which we construct three heuristic algorithms for choosing the loops. The heuristics are tested on two pipeflow networks that differ significantly on the distribution of pipes and nodes in the network.

1 Introduction

In boiler design it is essential that we can estimate the flow rates in each of the pipes related to the boiler. By estimating the flow rates the pipes can be made sufficiently durable so that they withstand the flow, while the pipes should not be made needlessly durable so that the material costs can be held at a reasonable level. Thus, it is possible to deliver durable boilers at a competitive prize to the potential customers.

A very simple example of a boiler modeled as a pipeflow network is given in Fig. 1. There the network consists of $m = 12$ pipes and $n = 6$ nodes, and the heat surface is modeled by the seven adjacent pipes between the nodes 3 and 4. The behavior of the flow in such a network can be depicted by two simple principles—the first one being the continuity of flow on each node, i.e., flow in equals flow out, and the second one being that the pressure loss over any closed circuit has to be zero.

We obtain $n - 1$ linearly independent linear equations from the continuity of flow, and therefore we need $m - n + 1$ equations based on the pressure losses (see [4] for details on obtaining the equations). Even though the pressure loss equations are nonlinear with respect to the flow rates, we want to have as many equations

J.-P. Humaloja (✉) • S. Ali-Löytty • 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; simo.ali-loytty@tut.fi; timo.t.hamalainen@tut.fi; seppo.pohjolainen@tut.fi

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_97

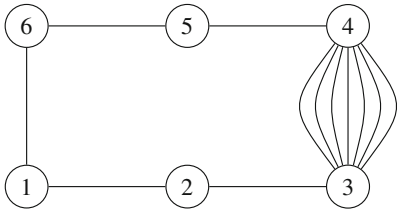


Fig. 1 A very simple example of a boiler modeled as a pipeflow network. In this case, the model consists of 12 pipes corresponding the 12 edges and 6 nodes corresponding the 6 vertices. The heating surface is modeled by the seven adjacent pipes between nodes 3 and 4

as we have unknown variables so that we can solve the linearized version of the equations. As a whole the flow rates are solved iteratively by subsequently solving the linearized version of the equations by following the procedure depicted in [4]. In order to improve the convergence and numerical stability of the iterative procedure, we will present heuristics for improving the condition number of the set of linearized equations.

The structure of this paper is as follows. In Sect. 2 we consider the problem from a theoretical standpoint and present some existing results as a basis for our heuristics. The heuristics for choosing the pressure loss loops are presented in Sect. 3 where we also consider the effect of scaling on the condition number. In Sect. 4 we test our heuristics in two different pipeflow networks, and in Sect. 5 we conclude our work.

Throughout this paper we denote the number of pipes in a network by m and the number of nodes by n. The singular values of a matrix are denoted by σ_i and the condition number of the matrix is determined by the quotient of its largest and smallest singular values. We denote the inner product of two vectors x and y by $\langle x, y\rangle$, and the Euclidean (2-)norm $||\cdot||$ is induced by the inner product, i.e., $||x||^2 = \langle x, x\rangle$. The maximum ($\infty$-)norm is denoted by $||\cdot||_\infty$.

2 Theoretical Background

In this section we will consider some required background to understand our approach to the problem of choosing linearly independent pressure loss loops. The section is divided into two subsections considering graph theory and the problem of choosing the best-conditioned submatrix. As the reference of graph theory we use the recent book by Santamu [2], and in the submatrix selection problem our reference is the article by Šrámek et al. [3].

2.1 Concepts of Graph Theory

In this subsection we will review some required background to graph theory in order to present the concept of fundamental cycles. First of all, we call the minimum set of edges that connects all the vertices a *spanning tree*. There are $n - 1$ edges in a spanning tree, and the remaining $m - n + 1$ edges form the *cospanning tree*. When an edge from the cospanning tree is added to the spanning tree, a cycle is formed. By separately adding the vertices of the cospanning tree to the spanning tree we obtain a cycle for each of the vertices, and together these cycles form the set of *fundamental cycles*. As each fundamental cycle contains at least one unique edge, the set of fundamental cycles is linearly independent under the ring sum operation. Similarly the set of vectors of $\mathbb{R}^m$ corresponding to the fundamental cycles are linearly independent in the usual sense. Note also that as the spanning tree is not unique, the set of fundamental cycles is not unique either.

Based on the above we can choose the pressure loss loops in such a way that they form a set of fundamental cycles, which will make them linearly independent. However, we do not know which set of fundamental cycles is the most linearly independent. In our approach we aim to choose the most linearly independent cycles from the set of an excessive number of cycles by utilizing the ideas presented in the next subsection.

2.2 Choosing the Best-Conditioned Submatrix

Since we aim to choose the most linearly independent set of cycles from a larger set, our problem is similar to choosing the best-conditioned submatrix of a larger matrix. Assume that we have an $M \times m$ ($M > m$) matrix of which we aim to find the best-conditioned $m \times m$ submatrix. The problem is in fact shown to be NP-hard in [3], but the authors have also presented a heuristic to find a well-conditioned submatrix in (Mm^2) time. The heuristic is based on finding such a submatrix A whose rows are as orthogonal as possible. Thus, the heuristic tends to maximize $\text{vol}(A) = \Pi_{i=1}^{m} \sigma_i$ whereas the condition number of A is defined as $\text{cond}(A) = \sigma_{\max}/\sigma_{\min}$, and therefore the heuristic usually provides suboptimal solutions.

Consider the problem of finding a well-conditioned $m \times m$ submatrix B of an $M \times m$ matrix A. The heuristic presented in [3] repeats the following steps m times:

1. On the i:th step, find the row a_j of A that has the largest norm. Choose the row as the i:th row of B, i.e., $b_i = a_j$.
2. For each row of A, subtract its projection to b_i from itself, i.e., update the remaining rows a_k by $a_k = a_k - \langle a_k, b_i \rangle ||b_i||^{-2} b_i$.

Thus, on each step, the heuristic finds the row of A that has the largest orthogonal component with respect to the previously chosen rows. Therefore the rows of the submatrix B should eventually be rather orthogonal, and B should be fairly

well-conditioned. We will be using this algorithm as a basis for our own heuristics for choosing the pressure loss loops.

3 Heuristics for Improving the Condition Number

In this section we consider three heuristics for choosing the pressure loss loops based on the previously presented algorithm from [3]. Before presenting the heuristics in Sect. 3.2 we will consider improving the condition number by scaling the equations.

3.1 Scaling the Equations

Consider the linearized set of equations for the flow rates in the form $Ax = b$. The $m - n + 1$ rows of A corresponding to the pressure loss equations contain elements of the scale $10^3 - 10^5$, whereas the $n - 1$ continuity of flow equations consist of elements $-1, 0$ and 1. Thus, the matrix A may be badly scaled, and therefore we consider simple row scaling to improve its condition number.

A heuristic given in [1] suggests that a scaling matrix S should be chosen in such a way that every row of SA has approximately the same ∞-norm. Thus, by choosing $S = \text{diag}(||a_1||_\infty^{-1}, ||a_2||_\infty^{-1}, \ldots, ||a_m||_\infty^{-1})$ the ∞-norm of each row of SA is 1. As a comparison we will also consider normalizing the rows of A with respect to the Euclidean (2-)norm. The results of scaling in two different cases are presented in Table 1. The two cases are considered in more detail in Sect. 4.

The table shows that normalization with respect to both ∞- and 2-norms improves the condition number by a factor of 100. Furthermore, the table indicates that 2-norm actually provides better condition numbers than ∞-norm, which is somewhat unexpected based on the above heuristic.

Table 1 The effect of scaling on the condition number

	Case 1 $m = 618, n = 235$	Case 2 $m = 953, n = 78$
I	$(9.4 \pm 1.8) \cdot 10^5$	$(13.4 \pm 2.2) \cdot 10^5$
S_∞	$(15.0 \pm 3.0) \cdot 10^3$	$(15.6 \pm 4.2) \cdot 10^3$
S_2	$(12.2 \pm 2.7) \cdot 10^3$	$(7.7 \pm 2.4) \cdot 10^3$

The numbers represent the mean condition number and its standard deviation for 100 matrices generated from the same network. The second and third rows display the effect of normalization with respect to ∞ and 2-norms. For reference, the first row shows the condition number without scaling

3.2 *Choosing the Pressure Loss Loops*

In this subsection we present three heuristics for choosing the pressure loss loops. All the heuristics utilize the submatrix selection algorithm presented in [3]. In each of the heuristics we assume that we have generated an excessive number of loops and have a matrix A the rows of which correspond to these loops. Our task is to find the $m - n + 1$ rows of A that together with the $n - 1$ rows obtained from the continuity of flow equations result in the best-conditioned $m \times m$ matrix.

Heuristic 1: The first heuristic directly utilizes the algorithm presented in [3]. Denote the $n - 1 \times m$ matrix corresponding to the continuity of flow equations by C. We begin by subtracting the projection of A to C from A, i.e., we update A by $A = A - (C^T(CC^T)^{-1}CA^T)^T$, which makes the rows of A orthogonal to the rows of C. Now we simply use the algorithm of [3] to choose the $m - n + 1 \times m$ submatrix of A, which yields the choice of the pressure loss loops.

Heuristic 2: In the second heuristic we assume that we have an initial $m \times m$ matrix A_0 that can be created by choosing a set of fundamental circuits as the pressure loss loops. After the initialization, the heuristic is described as follows.

1. Find the row j of A_0 that satisfies $j = \max_{n \le j \le m} \sum_{k=1,k\neq j}^{m} |\langle a_{0j}||a_{0j}||^{-1}, a_{0k}||a_{0k}||^{-1}\rangle|$ and remove the row a_{0j} from A_0. Denote $\tilde{A}_0 = A_0 \setminus a_{0j}$.
2. Compute the norm of the orthogonal component of a_{0j} to the rows of $\tilde{A}_0$, i.e., $r = ||a_{0j}^T - \tilde{A}_0^T(\tilde{A}_0\tilde{A}_0^T)^{-1}\tilde{A}_0 a_{0j}^T||$.
3. Find the row a_i of A that has norm-wise the largest orthogonal component to the rows of $\tilde{A}_0$.
4. If $||a_i^T - \tilde{A}_0^T(\tilde{A}_0\tilde{A}_0^T)^{-1}\tilde{A}_0 a_i^T|| > r$, swap a_{0j} and a_i and go to step 2, otherwise stop.

That is, as long as we can find a row of A that is more orthogonal to A_0 than the least orthogonal row of A_0, we swap the rows.

Heuristic 3: The third heuristic is essentially a combination of the first two heuristics. First, we generate an initial matrix A_0 as in the previous heuristic. Then compute

$$\sum_{k=1,k\neq j}^{m} |\langle a_{0j}||a_{0j}||^{-1}, a_{0k}||a_{0k}||^{-1}\rangle| \text{ for } n \le j \le m. \tag{1}$$

Next, find N largest of these values, where N is predetermined, and move the corresponding rows from A_0 to A. Denote $\tilde{A}_0 = A_0 \setminus A_{0,moved}$, and update A by $A = A - (\tilde{A}_0^T(\tilde{A}_0\tilde{A}_0^T)^{-1}\tilde{A}_0 A^T)^T$. Finally, use the algorithm presented in [3] to find a well-conditioned $N \times m$ submatrix of A that yields the choice of the N loops that replace the ones that were removed earlier.

4 Test Cases and Results

We will test the heuristics on the two cases we inspected the effect of scaling on. The cases are chosen due to their different distributions of pipes and nodes. We will perform the choosing of loops 100 times for both test cases, and initially we will generate approximately $10(m - n)$ loops. The parameter N for Heuristic 3 is chosen based on the mean value of the quantities computed in Eq. (1). Additionally, it should be noted that since row normalization with respect to 2-norm was found effective on Sect. 3.1, we will do that in the test cases as well.

4.1 Case 1: 618 Pipes, 235 Nodes

In Case 1 the pipeflow network consists of 618 pipes and 235 nodes, i.e., the number of nodes is relatively high. The results on the test case are shown in Table 2 where we see that while Heuristic 1 provides the best-conditioned matrix, it also requires the most iterations and computational time. Heuristic 2 performs also rather well as it yields almost as good results as Heuristic 1 while it requires the least time and iterations. Heuristic 3 does clearly the worst as it is worse than Heuristic 2 on all the aspects. Regardless, it should be noted that all the heuristics yield clearly better results than choosing a matrix randomly as seen by comparing the two bottom rows of the table. Note also that the cond(A_0) for Heuristic 1 is left blank as the heuristic does not use any initial matrix, and the standard deviation on the number of iterations is not displayed as the heuristic always requires $m - n + 1$ iterations.

4.2 Case 2: 953 Pipes, 78 Nodes

In Case 2 the pipeflow network consists of 953 pipes and 78 nodes, i.e., the number of nodes is relatively low. In this case Heuristic 1 provides clearly the best-conditioned matrices but it also requires clearly the most computational time and

Table 2 Results on Case 1

	Heuristic 1	Heuristic 2	Heuristic 3
Iterations	384	51.0 ± 31.8	216.7 ± 36.7
Time (s)	30.6 ± 1.2	8.0 ± 4.8	19.0 ± 3.5
cond(A)	6905 ± 4228	7119 ± 3910	9332 ± 4437
cond(A_0)	–	12312 ± 3083	12312 ± 3083

The numbers represent the mean values and standard deviations of the number of iterations, computational time, condition number of the $m \times m$ submatrix A and condition number of the initial $m \times m$ matrix A_0

Table 3 Results on Case 2

	Heuristic 1	Heuristic 2	Heuristic 3
Iterations	876	2.33 ± 2.62	82.57 ± 9.98
Time (s)	258.9 ± 7.9	2.1 ± 1.6	27.1 ± 3.2
cond(A)	1976 ± 462	7813 ± 2161	8454 ± 3470
cond(A_0)	–	7852 ± 2152	7852 ± 2152

The numbers represent the mean values and standard deviations of the number of iterations, computational time, condition number of the $m \times m$ submatrix A and condition number of the initial $m \times m$ matrix A_0

iterations (Table 3). Heuristic 2 performs only a few iterations, which explains the short computational time and only the slight improvement from the initial cond(A_0). Heuristic 3 similarly does relatively few iterations, but oddly the resulting matrix has on average a worse condition number than the initial matrix. However, it should be noted that the heuristic aims to increase the orthogonality of the matrix, which may be achieved even if the condition number is increased.

5 Conclusions

We presented three heuristics for choosing linearly independent loops in a pipeflow network. The heuristics were tested on two cases with differently distributed pipes and nodes. The first heuristic provided the best results but also required the most time, whereas the second heuristic was the fastest one but could not always provide as good results as Heuristic 1. The third heuristic was clearly the worst of the three.

Acknowledgements This work was carried out in the research program Flexible Energy Systems (FLEXe) in collaboration with Valmet, and it was supported by Tekes—the Finnish Funding Agency for Innovation.

References

1. Golub, G.H., Van Loan, C.F.: Matrix Computations, 4th edn. The Johns Hopkins University Press, Baltimore (2013)
2. Santamu, S.R.: Graph Theory with Algorithms and Its Applications: In Applied Science and Technology. Springer, New Delhi (2013)
3. Šrámek, R., Fisher, B., Vicari, E., Widmayer, P.: Optimal transitions for targeted protein quantification: best conditioned submatrix selection. In: Nqo, H.Q. (Ed.) COCOON 2009, LNCS 5609, pp. 287–296. Springer, Berlin/Heidelberg (2009)
4. Stephenson, D.: Pipeflow Analysis, 1st ed. Elsevier Science, New York (1984)

Modeling and Optimization Applied to the Design of Fast Hydrodynamic Focusing Microfluidic Mixer for Protein Folding

Benjamin Ivorra, María Crespo, Juana L. Redondo, Ángel M. Ramos, Pilar M. Ortigosa, and Juan G. Santiago

Abstract In this paper, we are interested in the design of a microfluidic mixer based on hydrodynamic focusing which is used to initiate the folding process (i.e., changes of the molecular structure) of a protein by diluting a protein solution to decrease its denaturant concentration to a given value in a short time interval we refer to as mixing time. Our objective is to optimize this mixer by choosing suitable shape and flow conditions in order to minimize its mixing time. To this end, we first introduce a numerical model that enables computation of the mixing time of a considered mixer. Then, we define a mixer optimization problem and solve it using a hybrid global optimization algorithm.

1 Introduction

Proteins are composed of chains of amino acids which can assume complex three-dimensional (3D) structures. Protein folding refers to the processes by which inactive proteins (unfolded chains of amino acids) acquire the 3D shapes (called folded) enabling them to perform a wide range of biological functions [1]. The applications of protein folding in research and industry are numerous, including: drug discovery, DNA sequencing and amplification, molecular diagnostics and food

B. Ivorra (✉) • M. Crespo • Á.M. Ramos
Instituto de Matemática Interdisciplinar (IMI) & Departamento de Matemática Aplicada, Universidad Complutense de Madrid, Plaza de Ciencias, 3, 28040 Madrid, Spain
e-mail: ivorra@mat.ucm.es; mcresp01@ucm.es; angel@mat.ucm.es

J.L. Redondo • P.M. Ortigosa
Dpto. de Arquitectura de Computadores y Electrónica, Universidad de Almeria, ceiA3, Ctra. Sacramento, La Cañada de San Urbano, 04120 Almería, Spain

J.G. Santiago
Mechanical Engineering Department, Stanford University, 440 Escondido Mall, Stanford, CA 94305-3030, USA

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_98

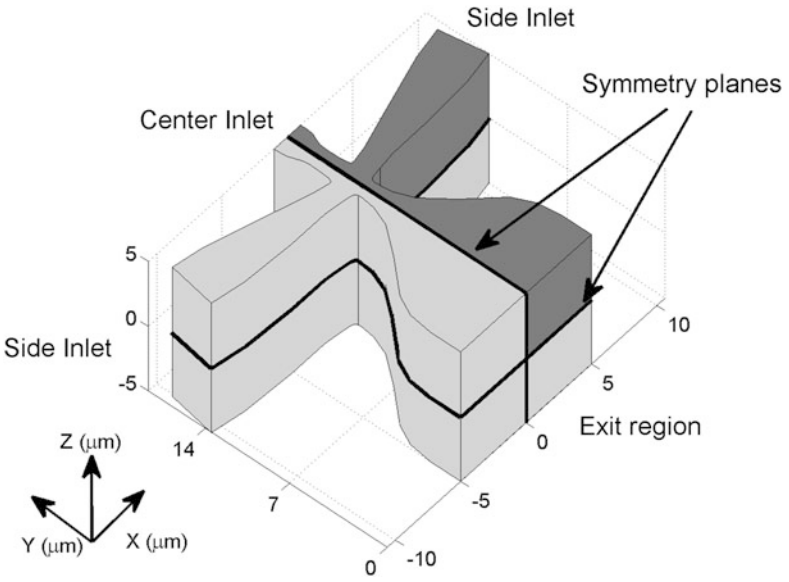


Fig. 1 Typical domain representation of the microfluidic mixer geometry considering the 3D model: in *dark gray* we represent the domain $\Omega_{3D,s}$ used for numerical simulations. The geometry's symmetry planes are highlighted and labeled

engineering. Protein folding can be initiated, for instance, by using changes in chemical potential (such as concentration of a chemical specie).

The original concept of a micromixer based on diffusion from (or to) a hydrodynamically focused stream was first proposed by Brody et al. in [2]. As shown in Fig. 1, this kind of mixer is composed of three inlet channels and a common outlet channel. It is symmetric with respect to its center channel. In the center inlet channel a mixture of unfolded proteins and a chemical denaturant is injected, whereas in the two side inlet channels a background buffer is introduced. The objective is to rapidly decrease the denaturant concentration in order to initiate protein folding in the outlet channel. Since the publication of Brody et al., there have been significant advances in this field. For example, while the original mixer of Brody et al. [2] showed mixing times greater than 10 μs (given the mixing measures used here), Hertzog et al. [4] obtained mixing times of 1.2 μs.

In this article, we present a modeling for the optimization of the shape and flow conditions of a particular hydrodynamic focused microfluidic mixer. Our objective is to improve a specified mixing time of this device taking into account that, currently, the best mixer designs exhibit mixing times of approximately 1.0 μs [4]. To do so, we first introduce a mathematical model which computes mixing time for a given mixer geometry and injection velocities. Then, we define the considered optimization problem based on the model. This problem is solved by considering

a hybrid global optimization method which is itself an improvement of a technique previously used for designing microfluidic mixers [5].

2 Mathematical Modeling

We consider the microfluidic hydrodynamic focusing mixer introduced in Sect. 1.

Let Ω_{3D} be the domain defined by the mixer shape in 3D. A typical representation of Ω_{3D} is depicted in Fig. 1. The mixer geometry has two symmetry planes that can be used to reduce the simulation domain. Therefore, it is only necessary to study a quarter of the mixer, denoted by $\Omega_{3D,s}$ and represented in dark gray in Fig. 1. Furthermore, $\Omega_{3D,s}$ can be approximated considering a 2D projection, as suggested in other works [4]. A representation of this projection, denoted by $\Omega_{2D,s}$, is shown in Fig. 2.

In order to simplify the notations, we introduce $\Omega = \Omega_{2D,s}$. In the boundary of Ω, denoted by Γ, we define: Γ_c the boundary representing the center inlet; Γ_s the boundary representing the side inlet; Γ_e the boundary representing the outlet; Γ_{w1} the boundary representing the wall defining the lower corner; Γ_{w2} the boundary representing the wall defining the upper corner; Γ_a the boundary representing the Y-axis symmetry. A geometrical representation of these boundaries is given in Fig. 2.

We assume the mixer liquid flow is incompressible [3]. Thus, the concentration distribution of the denaturant is described by using the incompressible Navier-Stokes equations coupled with the convective diffusion equation. Since we do not need the behavior of the device during its transient set up, only steady configurations are considered. More precisely, we approximate the flow velocity and the denaturant

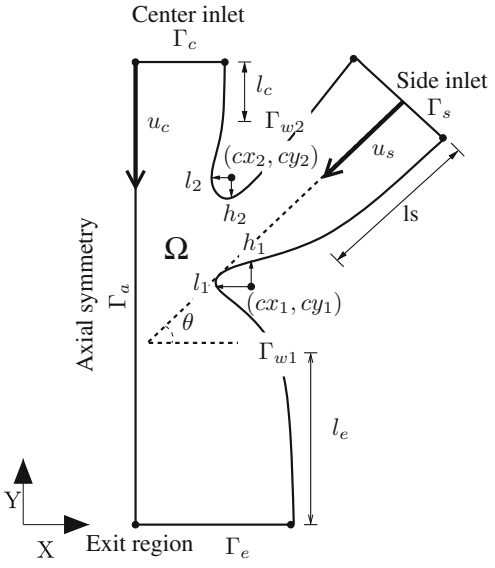


Fig. 2 Typical representation of the domain $\Omega_{2D,s}$ and parameterization of the microfluidic mixer considered for the optimization process

concentration distribution by considering the solution of the following system of equations [3, 4]:

$$\begin{cases} -\nabla \cdot (\eta(\nabla \mathbf{u} + (\nabla \mathbf{u})^\top) - p) + \rho(\mathbf{u} \cdot \nabla)\mathbf{u} = 0 & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega, \\ \nabla \cdot (-D\nabla c) + \mathbf{u} \cdot \nabla c = 0 & \text{in } \Omega, \end{cases} \tag{1}$$

where c is the denaturant normalized concentration distribution, $\mathbf{u}$ is the flow velocity vector ($\mathrm{m\,s^{-1}}$), p is the pressure field (Pa), D is the diffusion coefficient of the denaturant in the background buffer ($\mathrm{m^2\,s^{-1}}$), η is the denaturant dynamic viscosity ($\mathrm{kg\,m^{-1}\,s^{-1}}$) and ρ is the denaturant density ($\mathrm{kg\,m^{-3}}$).

System (1) is completed by the following boundary conditions:

For the flow velocity $\mathbf{u}$:

$$\begin{cases} \mathbf{u} = 0 & \text{on } \Gamma_{w1} \cup \Gamma_{w2}, \\ \mathbf{u} = -u_s \mathrm{para}_1 \mathbf{n} & \text{on } \Gamma_s, \\ \mathbf{u} = -u_c \mathrm{para}_2 \mathbf{n} & \text{on } \Gamma_c, \\ p = 0 \text{ and } (\eta(\nabla \mathbf{u} + (\nabla \mathbf{u})^\top))\mathbf{n} = 0 & \text{on } \Gamma_e, \\ \mathbf{n} \cdot \mathbf{u} = 0 \text{ and } \mathbf{t} \cdot (\eta(\nabla \mathbf{u} + (\nabla \mathbf{u})^\top) - p)\mathbf{n} = 0 & \text{on } \Gamma_a, \end{cases} \tag{2}$$

where u_s and u_c are the maximum side and center channel injection velocities ($\mathrm{m\,s^{-1}}$), respectively; para_1 and para_2 are the laminar flow profiles (parabolas) equal to 0 in the inlet border and unity in the inlet center; and $(\mathbf{t}, \mathbf{n})$ is the local orthonormal reference frame along the boundary.

For the concentration c:

$$\begin{cases} \mathbf{n} \cdot (-D\nabla c + c\mathbf{u}) = -c_0 \mathbf{u} & \text{on } \Gamma_c, \\ c = 0 & \text{on } \Gamma_s, \\ \mathbf{n} \cdot (-D\nabla c) = 0 & \text{on } \Gamma_e, \\ \mathbf{n} \cdot (-D\nabla c + c\mathbf{u}) = 0 & \text{on } \Gamma_{w1} \cup \Gamma_{w2} \cup \Gamma_a, \end{cases} \tag{3}$$

where $c_0 = 1$ is the initial denaturant normalized concentration in the center inlet. We note that the first equality in (3) corresponds to the inward denaturant flux in the center inlet channel and the third equality to the convective flux leaving the outlet channel.

3 Optimization Problem

In this work, the objective is to design a microfluidic mixer described by parameters $\phi \in \Phi$, where $\Phi \subset \mathbb{R}^N$ and $N \in \mathbb{N}$, that minimizes the mixing time function J, all defined below. Thus, the associated optimization problem can be written as:

$$\min_{\phi \in \Phi} J(\phi). \tag{4}$$

The mixer shape is described by rational Bézier curves and two ellipsoids. The latter are denoted as ellipsoids 1 and 2, where part of the ellipsoid 1 joins, in Γ_{w1}, the outlet and side channels, and part of the ellipsoid 2 joins, in Γ_{w2}, the center and side channels. These curves are determined by the following parameters (see Fig. 2 for their geometrical representation), suitably bounded to avoid non-admissible shapes (i.e., shape with intersected curves): the angle $\theta \in [0, \pi/3]$ between Γ_c and the direction normal to Γ_s; the length of the center inlet channel $l_c \in [2.5\,\mu\text{m}, 5\,\mu\text{m}]$; the length of the side inlet channel $l_s \in [1\,\mu\text{m}, 9\,\mu\text{m}]$; the length of the outlet channel $l_e \in [0.1\,\mu\text{m}, 20\,\mu\text{m}]$; the coordinates of the center of the ellipsoid i, with $i = 1, 2$, (cx_i, cy_i), where $cx_1 \in [0.8\,\mu\text{m}, 3\,\mu\text{m}]$, $cy_1 \in [l_e\,\mu\text{m}, l_e + 2\,\mu\text{m}]$, $cx_2 \in [0.8\,\mu\text{m}, 0.9\,\mu\text{m}]$ and $cy_2 \in [cy_1 + 1\,\mu\text{m}, (cy_1 + 3)\,\mu\text{m}]$; the radius l_i in the X-axis of the ellipsoid i, with $i = 1, 2$, satisfies $l_i \in [0\ \mu\text{m}, (cx_i - 0.5)\ \mu\text{m}]$; the radius h_i in the Y-axis of the ellipsoid i, with $i = 1, 2$, h_i, satisfies $h_1 \in [0\,\mu\text{m}, (cy_2 - cy_1 - 1)\,\mu\text{m}]$ and $h_2 \in [0\,\mu\text{m}, (cy_2 - cy_1 - 1 - h_1)\,\mu\text{m}]$. In addition to those parameters, we also consider the maximum injection velocities u_s and u_c as design variables. Furthermore, in order to maintain laminar flow and to avoid secondary flows in the outlet channel, such as Dean vortices, we constrained the typical flow Reynolds Re to less than 15. We define $Re = \rho u_s L/\eta$, where $L = 3\,\mu\text{m}$ is the side channel nozzle width. This implies that $u_s \leq \eta Re/\rho L\ \text{m s}^{-1}$. Moreover, in practice, u_c should be at least 10 times lower than u_s to ensure a good mixing between fluids [4]. Therefore, we impose that $u_s \in [0, \eta Re/\rho L]\ \text{m s}^{-1}$ and $u_c = p \times u_s$, where $p \in [0.001, 0.1]$. Thus, the set of parameters defining a particular mixer design is denoted by

$$\phi = \{u_s, p, \theta, l_c, l_s, l_e, cx_1, cy_1, l_1, h_1, cx_2, cy_2, l_2, h_2\} \in \Phi,$$

where $\Phi = \Pi_{i=1}^{14}[\underline{\Phi}(i), \overline{\Phi}(i)] \subset \mathbb{R}^{14}$ is the admissible space; and $\underline{\Phi}(i) \in \mathbb{R}$ and $\overline{\Phi}(i) \in \mathbb{R}$ are the upper and lower constraint values of the i-th parameter in ϕ described previously, respectively.

Here, the mixing time is defined as the time required to change the denaturant normalized concentration of a typical Lagrangian stream fluid particle situated in the symmetry streamline at depth $z = 0\ \mu\text{m}$ (halfway between the top and the bottom walls) from $\alpha \in [0, 1]$ to $\omega \in [0, 1]$ [3–5]. Thus, the mixing time of a particular mixer described by the parameters $\phi \in \Phi$, and denoted by J, is computed by:

$$J(\phi) = \int_{c_\omega^\phi}^{c_\alpha^\phi} \frac{\text{d}y}{\mathbf{u}^\phi(y)}, \tag{5}$$

where $\mathbf{u}^\phi$ and c^ϕ denote the solution of System (1)–(3), in its iD version, when considering the mixer defined by ϕ; and c_α^ϕ and c_ω^ϕ denote, the points situated along the symmetry streamline where the denaturant normalized concentration is α and ω.

4 Numerical Experiments

The numerical versions of the model presented above is implemented by coupling Matlab scripts with COMSOL Multiphysics 5.0 models (www.comsol.com).

During this work, we have considered guanidine hydrochloride (GdCl) as the denaturant. Indeed, GdCl is a Chaotropic agent which is frequently used for protein folding. The thermophysical variables of GdCl can be approximated as those of water. More precisely, its density is $\rho = 1010\,\mathrm{kg\,m^{-3}}$ and its dynamic viscosity is $\eta = 9.8 \times 10^{-4}\,\mathrm{kg\,m^{-1}\,s^{-1}}$. Furthermore, the diffusion coefficient of GdCl in the background buffer (assumed to be similar to water) is $D = 2 \times 10^{-9}\,\mathrm{m^2\,s^{-1}}$. According to those coefficients and the restriction $Re = \rho v L/\eta \leq 15$ introduced previously, the maximum side injection velocity is $u_s \leq 7\,\mathrm{m\,s^{-1}}$. Finally, the values of α and ω in Eq. (5) adapted to GdCl are considered here as 0.9 and 0.3, respectively [4, 5].

The optimization problem (4) has been solved by considering the Multi-layer secant algorithm (MSA) method presented in [7]. The number of evaluations of J used by MSA was about 6000 and the optimization process spent around 60 h. We denote by ϕ_o the result obtained at then end of the optimization process.

The values of ϕ_o are reported on Table 1. The shape of the optimized microfluidic mixer and its concentration distribution are depicted in Fig. 3. The maximum side injection velocity is $u_s = 5.2\,\mathrm{m\,s^{-1}}$, whereas the maximum central injection velocity is $u_c = 0.038\,\mathrm{m\,s^{-1}}$. The Reynolds number Re is around 9. The mixing time associated to this mixer is about 0.10 µs. This value is 10 times lower than the mixing times achieved by previous mixer designs with a similar model [3–5].

Table 1 Values of the optimized microfluidic mixer parameters

Parameter	u_s	p	θ	l_c	l_s	l_e	cx_1	cy_1	l_1	h_1	cx_2	cy_2	l_2	h_2
Value	5.2	7.3×10^{-3}	0.6	2.5	9.1	16.3	1.1	16.6	0.5	0.3	0.9	18.9	0.1	1.1

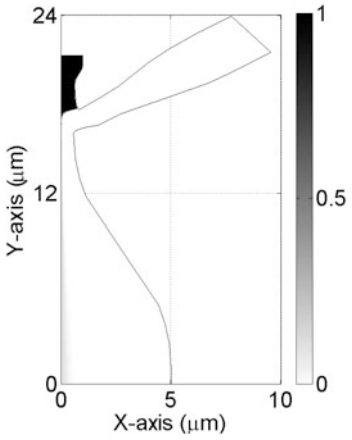


Fig. 3 Optimized mixer: shape of the optimized mixer with a superposed color plot of the denaturant concentration distribution

In those works, the mixing times were each greater than $1\,\mu$s. We attribute this improvement mainly to three factors: (i) the angle θ of the inlet side channels, whose value is about $\pi/5$ radians (this angle was fixed to 0 in [3–5]); (ii) the width of the mixing region (i.e., the area, defined by $(x, y) \in [0, 2] \times [14, 19]\,\mu$m where both fluids are mainly mixed) which reaches a minimum value of about $1.1\,\mu$m; and (iii) the choice of adequate injection velocities.

A detailed version of this work can be found in [6, 8].

Acknowledgements This work was carried out thanks to the financial support of the "Spanish Ministry of Economy and Competitiveness" under projects MTM2011-22658 and MTM2015-64865-P; the "Junta de Andalucía" and the European Regional Development Fund through project P12-TIC301; and the research group MOMAT (Ref.910480) supported by "Banco Santander" and "Universidad Complutense de Madrid".

References

1. Berg, J., Tymoczko, J., Stryer, L.: Biochemistry, 5th edn. W.H. Freeman, New York (2002)
2. Brody, J., Yager, B., Goldstein, R., Austin, R.: Biotechnology at low Reynolds numbers. Biophys. J. **71**(6), 3430–3441 (1996)
3. Hertzog, D., Michalet, X., Jäger, M., Kong, X., Santiago, J., Weiss, S., Bakajin, O.: Femtomole mixer for microsecond kinetic studies of protein folding. Anal. Chem. **76**(24), 7169–7178 (2004)
4. Hertzog, D., Ivorra, B., Mohammadi, B., Bakajin, O., Santiago, J.: Optimization of a microfluidic mixer for studying protein folding kinetics. Anal. Chem. **78**(13), 4299–4306 (2006)
5. Ivorra, B., Mohammadi, B., Santiago, J., Hertzog, D.: Semi-deterministic and genetic algorithms for global optimization of microfluidic protein folding devices. Int. J. Numer. Methods Eng. **66**(2), 319–333 (2006)
6. Ivorra, B., Redondo, J.L., Santiago, J.G., Ortigosa, P.M., Ramos, A.M.: Two- and three-dimensional modeling and optimization applied to the design of a fast hydrodynamic focusing microfluidic mixer for protein folding. Phys. Fluids **25**(3), 032001 (2013)
7. Ivorra, B., Mohammadi, B., Ramos, A.M.: A multi-layer line search method to improve the initialization of optimization algorithms. Eur. J. Oper. Res. **247**(3), 711–720 (2015)
8. Ivorra, B., Redondo, J.L., Ramos, A.M., Santiago, J.G.: Design sensitivity and mixing uniformity of a micro-fluidic mixer. Phys. Fluids **28**(1), 012005 (2016)

A Second Order Fixed Domain Approach to a Shape Optimization Problem

Henry Kasumba, Godwin Kakuba, and John Mango Magero

Abstract A fixed second order domain approach for solving optimal shape design problems is presented. The original optimal shape design problem is converted to an optimal control problem defined on a fixed domain. First and second order optimality conditions are derived. Numerical results are presented which demonstrate the robustness of the second order optimality conditions.

1 Introduction

In this work, we utilize a controllability type approach [4] to find a numerical solution to

$$\min_{E\subset\Omega\subset D} J(y,\Omega) := \frac{1}{2}\int_E |y-y_d|^2\,dx, \tag{1}$$

where y solves

$$-a_1\Delta y_1 + a_0 y_1 = f \quad \text{in} \quad \Omega, \tag{2}$$

$$-a_2\Delta y_2 + a_0 y_2 = f \quad \text{in} \quad D\setminus\Omega, \tag{3}$$

$$a_1\frac{\partial y_1}{\partial n} = a_2\frac{\partial y_2}{\partial n}; \quad y_1 = y_2 \text{ in } \partial\Omega\setminus(\partial\Omega\cap\partial D), \tag{4}$$

$$a_i\frac{\partial y_i}{\partial n} = 0 \quad \text{in} \quad \partial D, \quad i=1,2, \tag{5}$$

without a recourse to computation of shape derivatives [2]. Problem (2)–(5) models a body made of two materials occupying the regions Ω and $D\setminus\Omega$, respectively, having different physical properties expressed by the different coefficients a_1, a_2, associated to each region.

H. Kasumba (✉) • G. Kakuba • J. Mango Magero
Department of Mathematics, Makerere University, P.O. Box 7062, Kampala, Uganda
e-mail: kasumba@cns.mak.ac.ug; kakuba@cns.mak.ac.ug; mango@cns.mak.ac.ug

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_99

In [4], the first order optimality conditions for the problem (1)–(5) were derived and the numerical solution was computed for different study cases. As it is typical of first order gradient type methods, the optimization algorithm in [4] required a lot of optimization steps to converge. In this paper, we compute second order optimality conditions for problem under consideration and construct a Newton-type algorithm for solving this problem. We remark here that the computation of the optimality conditions using the controllability type approach is simpler and cheaper as compared to the computation of the shape derivative and Hessian. The rest of the paper is organized as follows. Section 2 describes the setting of the optimization problem. First order optimality conditions are given in Sect. 3 and in Sect. 4, second order variation are derived. In Sect. 5, a numerical algorithm and results is presented.

2 Setting of the Problem

If χ is the characteristic function of Ω in D, then problem (1)–(5) can be written in terms of χ:

$$\min_{\chi\in\mathbb{E}} J(y,\chi) := \int_E |y(\chi) - y_d|^2\, dx, \tag{6}$$

with $y(\chi)$ the solution of the variational problem

$$\int_D ((a_1\chi + a_2(1-\chi))\nabla y\nabla w + a_0 yw - fw)\, dx = 0 \quad \text{for all } w \in H^1(D), \tag{7}$$

and

$$\mathbb{E} := \{g : D \mapsto \mathbb{R};\ g(x) = 0 \text{ or } g(x) = 1, \text{ for all } x \in D\}.$$

The new minimization problem (6)–(7) can now be thought of as a control in coefficients problem, where the control is the characteristic function χ. In the new setting, the constraint set $\mathbb{E}$ makes the problem difficult to handle [3]. To circumvent this difficulty, a new regularization of characteristic functions was introduced in [4]. The idea of this approach is as follows. Suppose $\mathscr{X}(D)$ is a functional space in D such that the continuous and bounded functions $C(\bar{D}) \supset \mathscr{X}(D)$. If Ω is a subset of the fixed domain D, then it is possible to find some mapping $u : D \mapsto \mathbb{R}$ with $u \in \mathscr{X}(D)$, such that $u > 0$ in Ω, $u = 0$ on $\partial\Omega$ and $u < 0$ in $D \setminus \Omega$. One can say that u is a parametrization of Ω and the admissible domains can be written as

$$\bar{\Omega}_u = \{x \in D \mid u(x) \geq 0\}.$$

The parametrization of Ω can also be achieved with the help of the signed distance function

$$H(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ otherwise,} \end{cases}$$

and admissible domains can equivalently be written as

$$\bar{\Omega}_u = \{x \in D \mid u(x) \geq 0\} = \{x \in D | H(u) \geq 1\}.$$

Note that $H(u) \in L^\infty(D)$ and the weak form (7) can be expressed as

$$\int_D ([a_1 H(u) + a_2(1-H(u))]\nabla y \nabla w + a_0 yw - fw)\, dx = 0 \quad \text{for all } w \in H^1(D). \tag{8}$$

As $H(\cdot)$ is just Lipschitz, a further smoothing is possible. For second order methods, a Lipschitzian C^2 function H_ε defined as

$$H_\varepsilon(x) = \begin{cases} 1, & x \geq 0, \\ 1 + \frac{10}{\varepsilon^3}x^3 + \frac{15}{\varepsilon^4}x^4 + \frac{6}{\varepsilon^5}x^5, & -\varepsilon < x < 0, \\ 0, & x \leq -\varepsilon, \end{cases} \tag{9}$$

suffices. Consequently, we approximate (8) by

$$\int_D (a(u)\nabla y_\varepsilon \nabla \phi + a_0 y_\varepsilon \phi - f\phi)\, dx = 0 \quad \text{for all } \phi \in H^1(D), \tag{10}$$

where $a(u) := a_1 H_\varepsilon(u) + a_2(1 - H_\varepsilon(u))$. The optimization problem now writes

$$\min_{u \in U_{ad}} J(y_\varepsilon, u) := \frac{1}{2}\int_E |y_\varepsilon - y_d|^2\, dx, \tag{11}$$

subject to (10), where $U_{ad} = \{u : \ u \in L^\infty(D), \ |u(x)| \leq 1\}$. For the second order methods, the convexity of the cost can be important. The optimization problem (11)–(10) is generally not valid and we introduce an approximation of the cost via the Tikchonov regularization:

$$\min_{E \subset \Omega \subset D} J(y_\varepsilon, u) := \frac{1}{2}\int_E |y_\varepsilon - y_d|^2\, dx + \frac{\beta}{2}\int_D u^2\, dx, \tag{12}$$

subject to (10). Here, $\beta > 0$ is a small regularization parameter.

Note that for simplicity of presentation, from now on, the subscript ε on the state y shall be omitted.

3 First Order Optimality Conditions

The following theorem establishes the existence of the Lagrange multiplier p.

Theorem 1 *Let $X := H^1(D) \times L^\infty(D)$. Suppose that $x^* = (y^*, u^*) \in X$ is a local solution to (12). Then there exists a unique Lagrange multiplier $p^* \in H^1(D)$ such*

that (x^*, p^*) *is a stationary point of the Lagrangian* $\mathscr{L} : X \times H^1(D) \mapsto \mathbb{R}$ *defined as*

$$\mathscr{L}(x,p) = \int_D \Big(\frac{\chi_E}{2}|y - y_d|^2 + \frac{\beta}{2}u^2 + a(u)\nabla y \nabla p + a_0 yp - fp \Big)\, dx. \tag{13}$$

The following lemma guarantee that the Lagrange functional is twice continuously Frechet differentiable.

Lemma 1 *The cost J is twice continuously Fréchet differentiable and the mapping* $x := (y,u) \mapsto J''(x)$ *is Lipschitz continuous on X. The first and second derivatives of J at* $x = (y,u) \in X$ *are given by*

$$J'(x)(x_\delta) = \int_E (y - y_d)y_\delta + \beta u u_\delta \, dx, \quad J''(x)(x_\delta, \tilde{x}_\delta) = \int_E y_\delta \tilde{y}_\delta + \beta u_\delta \tilde{u}_\delta \, dx,$$

for directions $x_\delta = (y_\delta, u_\delta) \in X$ *and* $\tilde{x}_\delta = (\tilde{y}_\delta, \tilde{u}_\delta) \in X$.

Lemma 2 *For every* $x = (y,u) \in X$, *the operator* $e(y,u) : X \mapsto H^1(D)^*$ *defined as*

$$(e(x), \phi) := \int_D (a(u)\nabla y \nabla \phi + a_0 y\phi - f\phi)\, dx, \text{ for any } \phi \in H^1(D),$$

is continuously Fréchet differentiable with Lipschitz continuous derivative. The action of the first derivative of e with respect to x is given by

$$(e_x(x)(w,s), \phi) := \int_D (a'(u)s\nabla y \nabla \phi + a(u)\nabla w \nabla \phi + a_0 w\phi)\, dx,$$

where $(w,s) \in X$, $\phi \in H^1(D)$.
Differentiability of the Lagrange functional guarantees the following first order optimality conditions.

Theorem 2 *Let* $(y^*, u^*) \in X$ *be an optimal solution for (12). Then it satisfies the following first optimality system in variational sense:*

$$\begin{cases} -\nabla \cdot (a(u^*)\nabla y^*) + a_0 y^* = f \quad in \quad D, \\ a_1 \frac{\partial y_1^*}{\partial n} = a_2 \frac{\partial y_2^*}{\partial n}; \quad y_1^* = y_2^* \ in \ \partial\Omega \setminus (\partial\Omega \cap \partial D), \\ a_i \frac{\partial y_i^*}{\partial n} = 0 \quad in \quad \partial D, \ \ i = 1,2, \end{cases} \tag{14}$$

$$\begin{cases} -\nabla \cdot (a(u^*)\nabla p^*) + a_0 p^* = -\chi_E(y^* - y_d) \quad in \quad D, \\ a_1 \frac{\partial p_1^*}{\partial n} = a_2 \frac{\partial p_2^*}{\partial n}; \quad p_1^* = p_2^* \ in \ \partial\Omega \setminus (\partial\Omega \cap \partial D), \\ a_i \frac{\partial p_i^*}{\partial n} = 0 \quad in \quad \partial D, \ \ i = 1,2, \end{cases} \tag{15}$$

$$a'(u^*)[\beta u^* + \nabla y^* \nabla p^*] = 0, \tag{16}$$

where $a'(u^*) = (a_1 - a_2)H_\varepsilon'(u^*)$ *and* χ_E *is the characteristic function for the set* $E \subset D$.

4 Second Order Variations

Here, we search for an update direction $(\hat{y}, \hat{u}, \hat{p}) \in X \times H^1(D)$ from the following Newton step

$$\mathscr{L}''(y_k, u_k, p_k)[(\hat{y}, \hat{u}, \hat{p}), (\tilde{y}, \tilde{u}, \tilde{p})] = -\mathscr{L}'(y_k, u_k, p_k)(\tilde{y}, \tilde{u}, \tilde{p}), \tag{17}$$

for all variations $(\tilde{y}, \tilde{u}, \tilde{p}) \in X \times H^1(D)$, where $\mathscr{L}'$ and $\mathscr{L}''$ denote the first and second variations of the Lagrangian (13).

Theorem 3 *The operator $e(y, u) : X \mapsto H^1(D)^*$ is twice continuously differentiable with Lipschitz continuous second derivative. The action of the second derivative of e is given by*

$$(e_{xx}(x)(w, s)(v, r), \phi) := \int_D (a''(u)sr\nabla y\nabla\phi + a'(u)r\nabla w\nabla\phi + a'(u)s\nabla v\nabla\phi)\, dx,$$

where $(w, s), (v, r) \in X$ and $\phi \in H^1(D)$.

As a consequence of Lemma 1 and Theorem 3, it follows that the Lagrange functional is twice continuously Fréchet differentiable. Moreover, the mapping $(x, p) \mapsto \mathscr{L}''(x, p)$ is Lipschitz continuous on $X \times H^1(D)$ and the Newton step in (17) precisely writes:

Find $(\hat{y}, \hat{u}, \hat{p}) \in X \times H^1(D)$ as a solution to

$$\begin{aligned}
(\chi_E\hat{y}, \tilde{y}) + (a(u_k)'\hat{u}\nabla p_k, \nabla\tilde{y}) + (a(u_k)\nabla\hat{p}, \nabla\tilde{y}) + (a_0\hat{p}, \tilde{y}) &= -D_y\mathscr{L}(y_k, u_k, p_k)\tilde{y},\\
(a(u_k)'\nabla\hat{y}\nabla p_k, \tilde{u}) + (\zeta\hat{u}, \tilde{u}) + (a(u_k)'\nabla y_k\nabla\hat{p}, \tilde{u}) &= -D_u\mathscr{L}(y_k, u_k, p_k)\tilde{u},\\
(a(u_k)\nabla\hat{y}, \nabla\tilde{p}) + (a_0\hat{y}, \tilde{p}) + (a(u_k)'\hat{u}\nabla y_k, \nabla\tilde{p}) &= -D_p\mathscr{L}(y_k, u_k, p_k)\tilde{p},
\end{aligned}$$

for all $(\tilde{y}, \tilde{u}, \tilde{p}) \in X \times H^1(D)$ and $\zeta := \big(a(u_k)''[\beta u_k + \nabla y_k\nabla p_k]\big)$. If we assume that the state and adjoint systems are feasible at the k^{th} step, i.e.,

$$D_p\mathscr{L}(y_k, u_k, p_k)\tilde{p} = 0, \text{ for all } \tilde{p} \in H^1(D), \text{ and } D_y\mathscr{L}(y_k, u_k, p_k)\tilde{y} = 0, \text{ for all } \tilde{y} \in H^1(D),$$

then we solve the following coupled system to obtain the Newton step $(\hat{y}, \hat{u}, \hat{p})$:

$$\begin{aligned}
&(\chi_E\hat{y}, \tilde{y}) + (a(u_k)'\hat{u}\nabla p_k, \nabla\tilde{y}) + (a(u_k)\nabla\hat{p}, \nabla\tilde{y}) + (a_0\hat{p}, \tilde{y}) = 0,\\
&(a(u_k)'\nabla\hat{y}\nabla p_k, \tilde{u}) + (\zeta\hat{u}, \tilde{u}) + (a(u_k)'\nabla y_k\nabla\hat{p}, \tilde{u}) = -D_u\mathscr{L}(y_k, u_k, p_k)\tilde{u},\\
&(a(u_k)\nabla\hat{y}, \nabla\tilde{p}) + (a_0\hat{y}, \tilde{p}) + (a(u_k)'\hat{u}\nabla y_k, \nabla\tilde{p}) = 0.
\end{aligned} \tag{18}$$

5 Numerical Algorithm and Results

The algorithmic steps for the implementation of the second order variations are summarized in Algorithm 1.

Algorithm 1 Newton algorithm

1. Initialization: Choose u_0. Solve

$$D_p\mathscr{L}(y, u_0, p)\tilde{p} = 0, \quad D_y\mathscr{L}(y_0, u_0, p)\tilde{y} = 0, \text{ for } (y_0, p_0),$$

and set $k = 0$.

Iteration: for $k = 0, 1, 2, \ldots$

2. Newton step: Solve for $(\hat{y}, \hat{u}, \hat{p})$ in (18).
3. Set $u_{k+1} = u_k + \hat{u}$.
4. Feasibility step: Solve for (y^{k+1}, p^{k+1}):

$$D_p\mathscr{L}(y, u_{k+1}, p)\tilde{p} = 0, \quad D_y\mathscr{L}(y_{k+1}, u_{k+1}, p)\tilde{y} = 0.$$

5. Stop or set $k = k + 1$ and go to 2.

5.1 Example

We set $D = (-1, 1) \times (-1, 1)$, $a_1 = 5$, $a_2 = a_0 = 1$, $E = \Omega$, and $f = x_1^2 + x_2^2$. The function H_ε is chosen as in (9) with $\varepsilon = 10^{-2}$. The computational domain is discretized using triangular elements generated by the bi-dimensional anisotropic mesh generator [1]. The regularized state and adjoint problems are discretized using the Galerkin finite element method. We use piecewise linear polynomials for the approximation of the state and adjoint variables and piecewise constant polynomials for the approximation of the control variable u. We choose y_d to be the solution of the transmission problem on the geometry depicted in Fig. 1a. The desired state computed on this geometry is depicted in Fig. 1b. Two runs with 3868 and 6,1738 elements were done. In both cases the initial guess $u = 0$ was used. The obtained results are reported in Table 1.

The results in Table 1 indicate faster convergence of the second order method compared to the first order method.

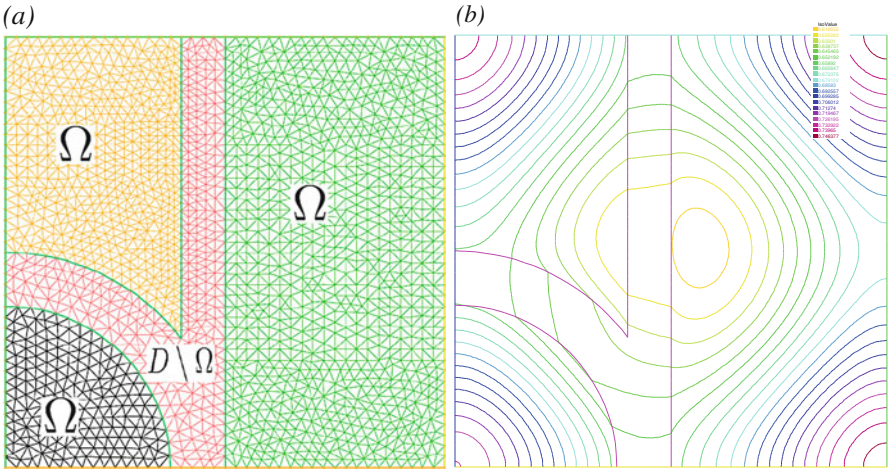


Fig. 1 Test domain and desired solution. (**a**) Geometry with 3868 elements. (**b**) Desired state

Table 1 Comparison of first and second order methods, $\varepsilon = 10^{-2}$, $\beta = 10^{-3}$

Order	No. of elements	Initial cost	Final cost	Iterations
First order	3868	0.00120951	1.128×10^{-14}	21
	61,738	0.00121337	2.358×10^{-11}	28
Second order	3868	0.00120951	3.12×10^{-13}	2
	61,738	0.00121337	1.34×10^{-14}	3

Acknowledgements The authors appreciate the research and financial assistance extended by the Makerere University Sida phase IV Bilateral program project 316-2014, Capacity Building in Mathematics and Its Applications.

References

1. Hecht, F.: Bamg: bidimensional anisotropic mesh generator. http://www.rocq.inria.fr/gamma/cdrom/www/bamg/eng.html.Cited15April2016 (1998)
2. Ito, K., Kunisch, K., Peichl, G.: Variational approach to shape derivatives. ESAIM Control Optim. Calc. Var. **14**, 517–539 (2008)
3. Pironneau, O.: Optimal Shape Design for Elliptic Systems. Springer, Berlin (1984)
4. Tiba, D., Neittaanmäki, P., Mäkinen, R.: A fixed domain approach in an optimal shape design problem. In: Vichnevetsky, R., Miller, J. (eds.) Proceedings of 13th IMACS World Congress on Computation and Applied Mathematics, pp. 1658–1660. Criterion Press, Dublin (1991)

Semi-Discretized Stochastic Fiber Dynamics: Non-Linear Drag Force

Felix Lindner, Holger Stroot, and Raimund Wegener

Abstract We analyze a spatially discretized model for the dynamics of a thin, long, elastic, inextensible fiber in a turbulent flow as occurring, e.g., in the spunbond production process of non-woven textiles. It consists of a high-dimensional stochastic differential equation with a non-linear algebraic constraint and an associated Lagrange multiplier term. We prove existence and uniqueness of a global strong solution for the case of a non-linear underlying drag force model. Our result generalizes previous findings which are based on a simplified linear drag force model.

1 Introduction

In [3] we introduced and investigated a spatially discretized model for the stochastic dynamics of slender, elastic, inextensible fibers in turbulent flows. Such dynamics occur, e.g., in industrial production processes of technical textiles, where polymer fibers are subject to turbulent airflows. Using a simplified underlying drag force model of linear type, we showed existence and uniqueness of a global strong solution. In the present article we extend the results from [3] to a large class of non-linear drag force models.

The semi-discretized model in [3] is formally deduced from a space-time beam model with non-linear algebraic constraint and white noise forcing term as presented in [5]. A theoretical analysis of the full model does not exist; first results concerning the deterministic counterpart have been obtained recently in [1]. The spatially discretized stochastic model describes the fiber curve at time $t \geq 0$ via a finite sequence of associated fiber points $r(t) = (\mathbf{r}_i(t))_{i=1,\dots,N}$ in terms of a polygon line. The dynamics of the $\mathbb{R}^{3N}$-valued stochastic process $(r(t))_{t\geq 0}$ is determined by the

F. Lindner
Fachbereich Mathematik, TU Kaiserslautern, D-67653 Kaiserslautern, Germany
e-mail: lindner@mathematik.uni-kl.de

H. Stroot (✉) • R. Wegener
Fraunhofer ITWM, Fraunhofer Platz 1, D-67663 Kaiserslautern, Germany
e-mail: holger.stroot@itwm.fraunhofer.de; raimund.wegener@itwm.fraunhofer.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_100

stochastic differential algebraic equation

$$\begin{aligned} \mathrm{d}r(t) &= v(t)\,\mathrm{d}t \\ \mathrm{d}v(t) &= a(t, r(t), v(t))\,\mathrm{d}t + B(t, r(t), v(t))\,\mathrm{d}w(t) + \nabla g(r(t))\,\mathrm{d}\mu(t) \\ g(r(t)) &= 0, \end{aligned} \tag{1}$$

with $\mathbb{R}^{3N}$-valued velocity process $(v(t))_{t\geq 0}$ and $\mathbb{R}^{3N}$-valued standard Brownian motion $(w(t))_{t\geq 0}$. The constraint function $g\colon \mathbb{R}^{3N} \to \mathbb{R}^{N-1}$ is defined by $g(x) = \left(\frac{1}{2} - \frac{\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2}{2(\Delta s)^2}\right)_{i=1,\dots,N-1}$, $x = (\mathbf{x}_i)_{i=1,\dots,N} \in \mathbb{R}^{3N}$, so that the pointwise constraint $g(r(t)) = 0$ ensures a fixed distance $\Delta s > 0$ between the fiber nodes. By $\nabla g(x) \in \mathbb{R}^{3N\times(N-1)}$ we denote the transpose of the Jacobian matrix of g at $x \in \mathbb{R}^{3N}$, and $(\mu(t))_{t\geq 0}$ is an $\mathbb{R}^{N-1}$-valued continuous semimartingale (w.r.t. the natural filtration generated by w) which is implicitly given as the Lagrange multiplier to the inextensibility constraint and thus describes the inner traction between the fiber points. The drift function $a\colon \mathbb{R}_0^+ \times \mathbb{R}^{3N} \times \mathbb{R}^{3N} \to \mathbb{R}^{3N}$ involves the deterministic part of the drag force and other external forces as well as the bending stiffness; the diffusion function $B\colon \mathbb{R}_0^+ \times \mathbb{R}^{3N} \times \mathbb{R}^{3N} \to \mathbb{R}^{3N\times 3N}$ involves the stochastic part of the drag force. Related constrained stochastic differential equations are considered in the context of molecular dynamics, see, e.g., [2].

In this article, we consider a class of non-linear drag force models and show that the resulting functions a and B are such that Eq. (1) has a unique global strong solution (r, v, μ) in the sense of [3, Definition 3.1]. While we assumed linear growth and local Lipschitz conditions for a and B in [3], this does not work in the present context where a might have superlinear growth. However, we are able to show that a satisfies a one-sided linear growth condition (also called monotone condition, cf. [4]), and it turns out that the analysis in [3] can be generalized to this case. Let us note that further superlinearly growing terms appear in Eq. (1) if one represents the Lagrange multiplier process μ explicitly in terms of r and v, cf. [3, Sect. 3].

In the sequel, we first present a class of non-linear drag force models for our semi-discretized stochastic fiber dynamics in Sect. 2. Existence and uniqueness of a global strong solution (r, v, μ) to (1) are shown in Sect. 3.

2 A Class of Non-Linear Drag Force Models

Notation We denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the Euclidean norm and the corresponding inner product in finite-dimensional vector spaces which will be clear from the context. $\mathscr{M} = \{x \in \mathbb{R}^{3N} : g(x) = 0\}$ is the configuration manifold of the fiber and $T_x\mathscr{M} = \{y \in \mathbb{R}^{3N} : [\nabla g(x)]^\top y = 0\}$ is the tangent space of $\mathscr{M}$ at x. Position and velocity of the fiber points are denoted by $r = (\mathbf{r}_i)_{i=1,\dots,N}$ and $v = (\mathbf{v}_i)_{i=1,\dots,N} \in \mathbb{R}^{3N}$. The midpoint, the mean velocity and the tangential direction associated to an edge between two fiber nodes are denoted by $\mathbf{r}_{i+1/2} =$

$\frac{1}{2}(\mathbf{r}_{i+1}+\mathbf{r}_i), \mathbf{v}_{i+1/2} = \frac{1}{2}(\mathbf{v}_{i+1}+\mathbf{v}_i)$ and $\boldsymbol{\tau}_{i+1/2} = \frac{\mathbf{r}_{i+1}-\mathbf{r}_i}{\|\mathbf{r}_{i+1}-\mathbf{r}_i\|}$. The relative velocity $\mathbf{v}^{\mathrm{rel}}$ is given by $\mathbf{v}^{\mathrm{rel}} = \mathbf{u}(t,\mathbf{r}) - \mathbf{v}$, where $\mathbf{u}\colon \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}^3$ is the mean velocity of the turbulent flow surrounding the fiber. The tangential and normal part of the relative velocity are given by $\mathbf{v}_\tau^{\mathrm{rel}} = \langle \boldsymbol{\tau}, \mathbf{v}^{\mathrm{rel}} \rangle \boldsymbol{\tau}$ and $\mathbf{v}_n^{\mathrm{rel}} = \mathbf{v}^{\mathrm{rel}} - \mathbf{v}_\tau^{\mathrm{rel}}$.

We now introduce a class of drag force models for the spatially discrete system (1), including in particular the model used in [5], see Example 1 and 2. For simplicity we consider stress free boundary conditions at both ends of the fiber, see Remark 1 for details.

Model 1 (Deterministic forces) *Since we are mainly interested in the drag forces due to the turbulent flow, it is convenient to split the drift function* $a\colon \mathbb{R}_0^+ \times \mathbb{R}^{3N} \times \mathbb{R}^{3N} \to \mathbb{R}^{3N}$ *in* (1) *into a term* a^{drag} *arising from the drag and another term* a^{rem} *arising from the remaining forces, i.e.,*

$$a(t,r,v) = a^{\mathrm{drag}}(t,r,v) + a^{\mathrm{rem}}(t,r,v). \tag{2}$$

The drag part $a^{\mathrm{drag}} = (a_1^{\mathrm{drag}}, \dots, a_N^{\mathrm{drag}})$ *is modeled by*

$$a_i^{\mathrm{drag}}(t,r,v) = \frac{1}{2}\left(f(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2}) + f(t,\mathbf{r}_{i-1/2},\mathbf{v}_{i-1/2},\boldsymbol{\tau}_{i-1/2})\right),$$

where we use ghost points generated by the boundary conditions for $i = 1$ *and* $i = N$, *cf. Remark 1. The local force* $f\colon \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ *is modeled by applying a resistance coefficient* β *to the directional relative velocities separately, i.e.,* $\beta\colon \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}_0^+ \times \mathbb{R}_0^+$, $\beta = (\beta_\tau, \beta_n)$ *and*

$$f(t,\mathbf{r},\mathbf{v},\boldsymbol{\tau}) = \beta_\tau(t,\mathbf{r},\mathbf{v},\boldsymbol{\tau})\mathbf{v}_\tau^{\mathrm{rel}} + \beta_n(t,\mathbf{r},\mathbf{v},\boldsymbol{\tau})\mathbf{v}_n^{\mathrm{rel}}.$$

The remainder part a^{rem} *may contain, e.g., bending and gravity forces, cf. [3].*

For our main result, Theorem 1 in Sect. 3, we use the following regularity assumptions on Model 1.

Assumption 1 *The mean flow velocity function* $\mathbf{u}$ *is measurable, bounded and locally Lipschitz continuous in the spatial variable, uniformly on every compact time interval.*[1] *The resistance coefficient* β *is measurable, locally Lipschitz continuous in the spatial variables* $(\mathbf{r},\mathbf{v},\boldsymbol{\tau})$, *uniformly on every compact time interval, and fulfills the linear growth condition*

$$\left(\beta_n(t,\mathbf{r},\mathbf{v},\boldsymbol{\tau}) + \beta_\tau(t,\mathbf{r},\mathbf{v},\boldsymbol{\tau})\right) \le C_\beta(1+\|\mathbf{v}\|) \tag{3}$$

[1] i.e., for all $T, K \in (0,\infty)$ there exists a finite constant $C_{T,K}$ such that $\sup_{t\in[0,T]} \|\mathbf{u}(\mathbf{x},t) - \mathbf{u}(\mathbf{y},t)\| \le C_{T,K}\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^3$ with $\max(\|\mathbf{x}\|,\|\mathbf{y}\|) \le K$.

with $C_\beta \in (0,\infty)$ independent of t, $\mathbf{r}$, $\mathbf{v}$ and $\boldsymbol{\tau}$. The function a^{rem} is measurable and fulfills the local Lipschitz condition (4a) and the one-sided linear growth condition (5a), stated at the beginning of Sect. 3 below.
Further assumptions on the resistance coefficient β are made in the context of our model for the stochastic forces, see Assumption 2. The regularity assumptions above are fulfilled in practical applications, as shown in the following example.

Example 1 In [5] one considers a resistance coefficient that only depends on $\|\mathbf{v}_n^{\mathrm{rel}}\|$, i.e., $\beta(t,\mathbf{r},\mathbf{v},\boldsymbol{\tau}) = \tilde{\beta}(\|\mathbf{v}_n^{\mathrm{rel}}\|)$ for some $\tilde{\beta}\colon \mathbb{R}_0^+ \to \mathbb{R}_0^+ \times \mathbb{R}_0^+$. This resistance coefficient is based on physical models, valid for certain ranges of $\|\mathbf{v}_n^{\mathrm{rel}}\|$, and on a reasonable interpolation between these ranges. The function $\tilde{\beta}$ constructed in [5] is continuously differentiable with $\tilde{\beta}'(0) = 0$. In particular, the resulting coefficient β is locally Lipschitz continuous in the spatial variables, uniformly on compact time intervals. For large values of $\|\mathbf{v}_n^{\mathrm{rel}}\|$ the resistance coefficient β is comparable to $(\|\mathbf{v}_n^{\mathrm{rel}}\|^{1/2}, \|\mathbf{v}_n^{\mathrm{rel}}\| + \|\mathbf{v}_n^{\mathrm{rel}}\|^{1/2})$ which grows at most linearly, i.e., Eq. (3) is fulfilled. The remainder part a^{rem} corresponding to the model in [5] coincides with the remainder part of a considered in [3], where we proved the conditions (4a) and (5a) from Sect. 3 below.
Next, we present a model for the stochastic part of the drag force.

Model 2 (Stochastic Forces) *The $\mathbb{R}^{3N\times 3N}$-valued diffusion function B in (1) consists of 3×3 block matrices $B_{i,j}$, $i,j \in \{1,\dots,N\}$ such that $B_{i,j} = 0$ for $i \neq j$ and*

$$B_{i,i}(t,r,v) = \frac{1}{2}\Big(L_f(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2})D(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2}) + L_f(t,\mathbf{r}_{i-1/2},\mathbf{v}_{i-1/2},\boldsymbol{\tau}_{i-1/2})D(t,\mathbf{r}_{i-1/2},\mathbf{v}_{i-1/2},\boldsymbol{\tau}_{i-1/2})\Big)$$

Here, $L_f\colon \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ is the derivative of the force function f from Model 1 in the velocity component, and $D\colon \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$, the so-called turbulent drag amplitude, arises from the underlying turbulence model. As in Model 1 we use ghost points which depend on the boundary conditions to define $B_{1,1}$ and $B_{N,N}$, cf. Remark 1.
The following regularity assumptions on Model 2 will be used below.

Assumption 2 *The resistance coefficient β (introduced in Model 1) is continuously differentiable in the velocity component such that the derivative $D_v\beta$ is bounded and locally Lipschitz continuous in the spatial variables $(\mathbf{r},\mathbf{v},\boldsymbol{\tau})$, uniformly on compact time intervals. The turbulent drag amplitude D is as well measurable, bounded and locally Lipschitz continuous in the spatial variables $(\mathbf{r},\mathbf{v},\boldsymbol{\tau})$, uniformly on compact time intervals.*
In the sequel we continue Example 1 and show that Assumption 2 is fulfilled for the model considered in [5].

Example 2 Recall that the resistance coefficient in [5] is of the form $\beta(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau}) = \tilde{\beta}(\|\mathbf{v}_n^{\mathrm{rel}}\|)$, where $\tilde{\beta}$ comes from an interpolation of physical models. The interpolation coefficients are chosen such that the derivative of β in the velocity component is bounded and locally Lipschitz continuous in the spatial variables $(\mathbf{r}, \mathbf{v}, \boldsymbol{\tau})$, uniformly on compact time intervals. As mentioned above, the turbulent drag amplitude D depends on the underlying turbulence model which is a k-ε-model in [5]. Under reasonable assumptions on the k-ε-model, D is bounded and locally Lipschitz continuous in the spatial variables $(\mathbf{r}, \mathbf{v}, \boldsymbol{\tau})$, uniformly on compact time intervals. For instance, the boundedness follows from $\|D(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau})\|^2 \leq C_{k,\varepsilon}\|\varphi\|_{L_1} \sup_{\zeta \in \mathbb{R}_0^+} \int_0^\infty \frac{E(x,\zeta)}{x} \mathrm{d}x$ given suitable integrability assumptions on the spectral energy E and the time delay function φ; see [5] for details.

Remark 1 (Boundary Conditions) For $i = 1$ and $i = N$, the values of a_i^{drag} and $B_{i,i}$ in Model 1 and 2 are defined with the help of ghost points $\mathbf{r}_0$, $\mathbf{r}_{N+1}$, $\mathbf{v}_0$ and $\mathbf{v}_{N+1}$. The ghost points are determined by the considered boundary conditions, cf. [3, page 7]. In the case of stress free fiber ends we have $\mathbf{r}_0 = 2\mathbf{r}_1 - \mathbf{r}_2$, $\mathbf{r}_{N+1} = 2\mathbf{r}_N - \mathbf{r}_{N-1}$ and set $\mathbf{v}_0 = \mathbf{v}_1$, $\mathbf{v}_{N+1} = \mathbf{v}_N$.

3 Existence and Uniqueness of a Global Solution

The proof of existence and uniqueness of a global strong solution to Eq. (1) is based on verifying the following local Lipschitz and (one-sided) linear growth conditions: For all $T, K > 0$ there exist finite constants $C_{T,K}$ and C_T such that

$$\sup_{t \in [0,T]} \|a(t, x_1, y_1) - a(t, x_2, y_2)\| \leq C_{T,K}(\|x_1 - x_2\| + \|y_1 - y_2\|) \tag{4a}$$

$$\sup_{t \in [0,T]} \|B(t, x_1, y_1) - B(t, x_2, y_2)\| \leq C_{T,K}(\|x_1 - x_2\| + \|y_1 - y_2\|) \tag{4b}$$

for all $x_1, x_2, y_1, y_2 \in B_K(0) := \{x \in \mathbb{R}^{3N} : \|x\| \leq K\}$, and

$$\sup_{t \in [0,T]} \langle (x, y), (y, a(t, x, y)) \rangle \leq C_T(1 + \|x\|^2 + \|y\|^2) \tag{5a}$$

$$\sup_{t \in [0,T]} \|B(t, x, y)\| \leq C_T(1 + \|x\| + \|y\|) \tag{5b}$$

for all $x \in \mathscr{M}$ and $y \in T_x\mathscr{M}$.

Theorem 1 *Consider the drift and diffusion function a and B from Model 1 and 2 and suppose that Assumptions 1 and 2 hold. Then, given $r_0 \in \mathscr{M}$ and $v_0 \in T_{r_0}\mathscr{M}$, there exists a unique global strong solution (r, v, μ) to Eq. (1) with initial condition (r_0, v_0) (in the sense of [3, Definition 3]).*

Proof A careful study of the proof of Theorem 3.1 in [3] shows that the arguments therein remain valid under the conditions (4) and (5) above. As shown in Lemma 1 and 2 below these conditions are fulfilled in the present context.

Lemma 1 *In the setting of Theorem 1, the local Lipschitz condition* (4a) *and the one-sided linear growth condition* (5a) *are fulfilled.*

Proof The local Lipschitz continuity of a in the spatial variables follows from the local Lipschitz continuity of β, $\mathbf{u}$ and a^{rem}, hence a fulfills (4a). The splitting (2) yields

$$\langle (r,v),(v,a(t,r,v))\rangle = \langle r,v\rangle + \langle v,a^{\mathrm{drag}}(t,r,v)\rangle + \langle v,a^{\mathrm{rem}}(t,r,v)\rangle.$$

By assumption we have $\langle r,v\rangle + \langle v,a^{\mathrm{rem}}(t,r,v)\rangle \le C(1+\|r\|^2+\|v\|^2)$. It remains to verify a similar estimate for $\langle v,a^{\mathrm{drag}}(t,r,v)\rangle$. Note that

$$\begin{aligned}\left\langle v,a^{\mathrm{drag}}(t,r,v)\right\rangle &= \sum_{i=1}^{N-1}\left\langle \mathbf{v}_{i+1/2},f(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2})\right\rangle \\ &\quad + \frac{1}{2}\left\langle \mathbf{v}_1,f(t,\mathbf{r}_{1/2},\mathbf{v}_{1/2},\boldsymbol{\tau}_{1/2})\right\rangle + \frac{1}{2}\left\langle \mathbf{v}_N,f(t,\mathbf{r}_{N+1/2},\mathbf{v}_{N+1/2},\boldsymbol{\tau}_{N+1/2})\right\rangle.\end{aligned}$$

We first focus on the sum and treat the boundary part afterward. Recall that f splits in two terms, one in tangential direction and one in normal direction of the edge. We decompose $\langle \mathbf{v}_{i+1/2},f(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2})\rangle = (*)_n + (*)_\tau$ accordingly. The first term can be written as

$$\begin{aligned}(*)_n &= \beta_n(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2})\Big\langle \mathbf{v}_{i+1/2},(\mathbf{u}(t,\mathbf{r}_{i+1/2})-\mathbf{v}_{i+1/2}) \\ &\qquad - \langle \boldsymbol{\tau}_{i+1/2},(\mathbf{u}(t,\mathbf{r}_{i+1/2})-\mathbf{v}_{i+1/2})\rangle \boldsymbol{\tau}_{i+1/2}\Big\rangle \\ &= \beta_n(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2})\Big(\langle \mathbf{v}_{i+1/2},\mathbf{u}(t,\mathbf{r}_{i+1/2})\rangle - \|\mathbf{v}_{i+1/2}\|^2 \\ &\qquad - \langle (\mathbf{u}(t,\mathbf{r}_{i+1/2})-\mathbf{v}_{i+1/2}),\boldsymbol{\tau}_{i+1/2}\rangle\langle \mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2}\rangle\Big).\end{aligned}$$

After rearranging these terms we can apply Cauchy-Schwarz inequality and use the linear growth (3) of β_n and $\|\boldsymbol{\tau}_{i+1/2}\| = 1$ to get

$$\begin{aligned}(*)_n &\le C_\beta(1+\|\mathbf{v}^{\mathrm{rel}}_{n,i+1/2}\|)\|\mathbf{v}_{i+1/2}\|\|\mathbf{u}(t,\mathbf{r}_{i+1/2})\| \\ &\quad + C_\beta(1+\|\mathbf{v}^{\mathrm{rel}}_{n,i+1/2}\|)\|\mathbf{v}_{i+1/2}\|\|\mathbf{u}(t,\mathbf{r}_{i+1/2})\|\|\boldsymbol{\tau}_{i+1/2}\|^2 \\ &\quad + \beta_n(t,\mathbf{r}_{i+1/2},\mathbf{v}_{i+1/2},\boldsymbol{\tau}_{i+1/2})\|\mathbf{v}_{i+1/2}\|^2(\|\boldsymbol{\tau}_{i+1/2}\|^2-1) \\ &\le 4C_\beta \sup_{\mathbf{x}\in\mathbb{R}^3}(1+\|\mathbf{u}(t,\mathbf{x})\|^2)(1+\|\mathbf{v}_{i+1/2}\|^2).\end{aligned}$$

Note that $\beta_n(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau})\|\mathbf{v}\|^2$ may have cubic growth in the velocity component; this is the reason why it is necessary to consider one-sided linear growth instead of linear growth. The second term $(*)_\tau$ can be estimated analogously, using the properties of β_τ and the Cauchy-Schwarz inequality. The boundary parts can be estimated using the concrete values found in Remark 1.

Lemma 2 *In the setting of Theorem 1, the local Lipschitz condition* (4b) *and the linear growth condition* (5b) *are fulfilled.*

Proof By Assumption 2, L_f and D are locally Lipschitz continuous in the spatial variables, uniformly on compact time intervals. Since this property is invariant under multiplications the local Lipschitz condition (4b) is fulfilled. Since D is bounded it remains to show that L_f grows at most linearly. A direct calculation yields

$$\begin{aligned} D_v f(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau}) = \mathbf{v}_\tau^{\mathrm{rel}} \otimes D_v \beta_\tau(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau}) - \beta_\tau(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau}) \boldsymbol{\tau} \otimes \boldsymbol{\tau} \\ + \mathbf{v}_n^{\mathrm{rel}} \otimes D_v \beta_n(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau}) - \beta_n(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau})(\mathrm{Id} - \boldsymbol{\tau} \otimes \boldsymbol{\tau}). \end{aligned}$$

Hence the boundedness of $D_v \beta$ yields linear growth of L_f, i.e.,

$$\|L_f(t, \mathbf{r}, \mathbf{v}, \boldsymbol{\tau})\|^2 \leq C(1 + \|\mathbf{r}\|^2 + \|\mathbf{v}\|^2).$$

Conclusion: We presented a class of non-linear drag force models for the semi-discretized stochastic fiber dynamics (1), including the model in [5] as a particular example. Based on the verification of suitable Lipschitz and growth conditions for the coefficients a and B, we showed existence and uniqueness of a global strong solution. A careful analysis was necessary to handle superlinearly growing terms.

Acknowledgements This work has been supported by German DFG, project 251706852, MA 4526/2-1, WE 2003/4-1.

References

1. Grothaus, M., Marheineke N.: On a nonlinear partial differential algebraic system arising in technical textile industry: analysis and numerics. IMA J. Num. Anal. (2015) doi: 10.1093/imanum/drv056
2. Lelièvre, T., Rousset, M., Stoltz, G.: Langevin dynamics with constraints and computation of free energy differences. Math. Comput. **81**, 2071–2125 (2012)
3. Lindner, F., Marheineke, N., Stroot, H., Vibe, A., Wegener, R.: Stochastic dynamics for inextensible fibers in a spatially semi-discrete setting. Stochastics Dyn. (2016) doi: 10.1142/S0219493717500162
4. Mao, X.: Stochastic Differential Equations and Applications. Woodhead Publishing, Cambridge (2007)
5. Marheineke, N., Wegener, R.: Modeling and application of a stochastic drag for fibers in turbulent flows. Int. J. Multiph. Flow **37**, 136–148 (2011)

Simulating Heat and Mass Transfer with Limited Amount of Sensor Data

Vanessa López

Abstract We consider the problem of dynamically modeling the distribution of temperature and concentration of water vapor inside a building. It is assumed that the building is equipped with a network of sparsely located sensors and a data management system recording measurements of temperature, relative humidity, and air flow. The measurements serve as input data for a time-dependent boundary value problem proposed to simulate heat and mass transfer inside a building.

1 Introduction

Buildings nowadays can be equipped with a sensor network and a data management system which gathers the sensor data [3, 4]. Such data typically includes measurements of temperature and relative humidity, but may also contain measurements of air flow. In addition, a collection of models is included to aid with operational tasks. To this end, we are concerned with physics-based models, coupled with sensor measurements, for simulating heat and mass transfer inside buildings. Motivation for the present study stems from current interest on the impact that micro-climatic conditions may have on works of art housed in indoor environments [4, 5].

Specifically, we consider the problem of coupling sensor measurements of temperature, relative humidity, and air flow with a boundary value problem (BVP) for simulating the distribution of temperature and concentration of water vapor in rooms within a building. The measurements serve as input, in the form of time-dependent boundary data. The sensors are assumed to be few in number and *sparsely located* throughout the (inner) room boundaries [4]. Hence, information available for prescribing boundary conditions is minimal and typical approaches for prescribing them may not be suitable. We thus investigate here the use of BVPs in 2D, coupled with the sensor data, to prescribe the boundary conditions for the 3D model.

V. López (✉)
IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
e-mail: lopezva@us.ibm.com

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_101

2 Mathematical Model

A brief description of the proposed model for simulating the distribution of temperature and concentration of water vapor in (humid) air inside buildings follows. For a discussion on the related theory of heat and mass transfer and the derivation of the equations, the reader may consult [6]. We adopt the commonly used Boussinesq approximation for modeling natural convection [6]. The air velocity equations thus include a term that takes the effects of variations in temperature into account. The equations comprising the model consist of the incompressible Navier-Stokes equations (1)–(2) for air velocity, the Eq. (3) for temperature, and the Eq. (4) for the mass fraction for the concentration of water vapor in (humid) air,

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = \nu\nabla^2\mathbf{v} - \nabla(p'/\hat{\rho}) - \beta T'\mathbf{g} \qquad \text{in } \Omega \tag{1}$$

$$\nabla\cdot\mathbf{v} = 0 \qquad \text{in } \Omega \tag{2}$$

$$\hat{\rho}c_p\left(\frac{\partial T'}{\partial t} + \mathbf{v}\cdot\nabla T'\right) = \nabla\cdot(\kappa\nabla T' - \tilde{\mu}\mathbf{i}) + Q \qquad \text{in } \Omega \tag{3}$$

$$\hat{\rho}\left(\frac{\partial c}{\partial t} + \mathbf{v}\cdot\nabla c\right) = -\nabla\cdot\mathbf{i} \qquad \text{in } \Omega\,, \tag{4}$$

where $\mathbf{i} = -\hat{\rho}\,D\,(\nabla c + (k_T/T)\nabla T')$ is a diffusion flux and the 3D domain Ω is a room (or, generally, a building). We consider the system (1)–(4) coupled with the (time-dependent) boundary conditions

$$\mathbf{n}\cdot\mathbf{v} = v_b \qquad \text{on } \partial\Omega \tag{5}$$

$$p'/\hat{\rho} = \beta g z T' \qquad \text{on } \partial\Omega \tag{6}$$

$$T = T_b \qquad \text{on } \partial\Omega \tag{7}$$

$$c = c_b \qquad \text{on } \partial\Omega\,, \tag{8}$$

where $\partial\Omega$ denotes the boundary of the domain Ω.

In the model, the unknowns are (a) the air velocity $\mathbf{v}$; (b) T', which is a variation from a constant mean temperature $\widehat{T}$, so that $T = \widehat{T} + T'$ is the temperature; (c) a variation in pressure p'; and (d) the mass fraction c for the concentration of one component in a mixture of two fluids, namely water vapor and dry air within the present context. Here we take c to be the mass fraction for the concentration of water vapor in (humid) air, hence that of dry air is $1 - c$.

The density is $\rho = \hat{\rho} + \rho'$, where $\hat{\rho}$ is constant and $\rho' = -\hat{\rho}\beta T'$, with $\beta = -1/\rho(\partial\rho/\partial T)_p$ denoting the thermal expansion coefficient of air. The variations T' in temperature and ρ' in density are assumed to be small. The pressure is $p = p' - \hat{\rho}gz + p_{ref}$, where z is height, g is the gravitational acceleration constant, and p_{ref} is a constant reference pressure. The remaining notation, and required input,

is: (a) the viscosity ν, specific heat c_p, and thermal conductivity κ of air; (b) heat sources or sinks Q; (c) the gravitational acceleration vector $\mathbf{g}$; (d) the diffusivity D of water vapor in air and thermal diffusion coefficient k_T for the diffusion flux $\mathbf{i}$; and (e) $\tilde{\mu} = k_T(\partial\mu/\partial c)_{p,T}$, where μ is the chemical potential of the fluid mixture [6].

The Dirichlet boundary condition (7) for temperature (similarly (8) for concentration) requires the values T_b of temperature at each point on the boundary $\partial\Omega$ of the (discretized) domain. These could come from measurements for a *dense set* of points on the boundary; see, for example, [2]. While meaningful, this may not always be practical (due to aesthetic or economic reasons, for instance) since a large number of sensors would need to be in place. Hence, we consider the solution of 2D BVPs, coupled with data from *sparsely located* sensors, to prescribe the boundary conditions (7) and (8) for the 3D model. The details are provided in Sect. 3.

Finally, for subsets of $\partial\Omega$ acting as sources or sinks of air, measurements for air flow can be used to prescribe the Neumann boundary condition (5) for air velocity [8]. The numerical method used to solve the model requires prescribing boundary conditions for pressure. Here we take the restriction to $\partial\Omega$ of the first order approximation (6) of the equation of state $p'/\hat{\rho} = gz(\beta T' - \rho'/\hat{\rho} + \ldots)$, after neglecting $\rho'/\hat{\rho} << 1$. The value of T' on $\partial\Omega$ is taken from the solution of the aforementioned 2D BVP for temperature at the boundary.

3 Numerical Solution Techniques

We build upon the OpenFOAM [10] solvers for heat transfer and buoyancy-driven flows to solve the model (1)–(8), and use the NETGEN software [9] to generate a mesh for the 3D domain, as well as one for the domain boundary. We propose prescribing the boundary conditions (7)–(8) for the 3D model via the solution of 2D BVPs, coupled with sensor data. The 2D BVPs, namely (9) and (10), describe conduction of heat and diffusion of mass. Thus, it is natural to use them to model temperature and concentration of water vapor on the boundary. However, the implementation is not straightforward, for the boundary of the 3D domains of interest to us can be complicated. This aspect was addressed successfully, as explained next.

Having generated a mesh for the 3D domain Ω, let $\mathscr{E}$ denote the union of all mesh element faces triangulating the boundary $\partial\Omega$ of Ω. Suppose that at each sensor location both temperature and relative humidity measurements are recorded. (Usually, temperature and humidity sensors are placed in the same container box.) For each point (x_i, y_i, z_i) on $\partial\Omega$ associated with the physical location of a sensor (labeled by the index i), identify the element faces $\triangle_j^{(i)} \in \mathscr{E}$ such that their barycenter is contained in a small disc (or ball) of radius $r > 0$, centered at said point. The radius r can be thought of as accounting for the finite size of the sensor. Let $\Gamma^{(i)} = \bigcup_j \triangle_j^{(i)}$ be the union of all element faces satisfying the aforementioned property (see Fig. 1). Denote by $\mathscr{E}_1 = \bigcup_i \Gamma^{(i)}$ the union (over the set of all sensors)

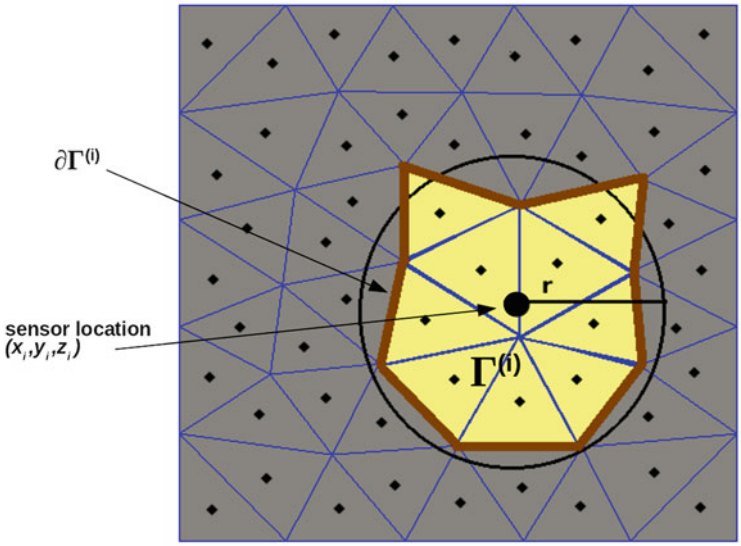


Fig. 1 Set $\Gamma^{(i)}$ of mesh element faces associated with the i-th sensor. An element face is assigned to $\Gamma^{(i)}$ if its barycenter falls within a disc of radius r centered at (x_i, y_i, z_i). (The small diamond inside each element face denotes its barycenter)

of all such triangulating surface pieces. Then let $\Gamma := \mathscr{E} \setminus \mathscr{E}_1$ be the triangulating surface which is the complement to $\mathscr{E}_1$ in $\mathscr{E}$. Note that the (1D) boundary $\partial\Gamma$ of Γ is the union (over the set of all sensors) of the boundaries of the small triangulating surface pieces around each sensor location, taken with the opposite direction. That is, $\partial\Gamma = -\bigcup_i \partial\Gamma^{(i)}$.

To obtain, at each time step, the needed values T_b in (7) for prescribing temperature on $\partial\Omega$, we solve the 2D BVP

$$\nabla \cdot (k \nabla T_b) \;=\; 0 \quad \text{in } \Gamma, \qquad T_b \;=\; T_s \quad \text{on } \partial\Gamma. \tag{9}$$

The values T_s for the Dirichlet boundary condition in (9) are provided directly by temperature sensor measurements. Similarly, to obtain the values c_b in (8) we solve

$$\nabla \cdot (\sigma \nabla c_b) \;=\; 0 \quad \text{in } \Gamma, \qquad c_b \;=\; c_s \quad \text{on } \partial\Gamma. \tag{10}$$

The values c_s for prescribing the Dirichlet boundary condition in (10) are determined from the absolute humidity, that is, the mass of water vapor per unit volume. In other words, and denoting absolute humidity by H_A, one has $c_s = H_A/\rho_H$, where ρ_H denotes the density of humid air. Both H_A and ρ_H can be quantified as functions of measurements of relative humidity and temperature at the same point in space and time; the interested reader is referred to [12] for the details. We compute solutions

of the BVPs (9) and (10) using our own (2D finite elements) solver, called from the OpenFOAM software before solving for the unknowns inside the 3D domain.

Finally, one could consider other types of boundary conditions for temperature (similarly, concentration of water vapor), like the mixed boundary conditions

$$T = T_s \quad \text{on } \mathscr{E}_1, \qquad \mathbf{n} \cdot (\nabla T) = \gamma \quad \text{on } \mathscr{E} \setminus \mathscr{E}_1, \tag{11}$$

for a given value (e.g., zero) of γ. Though more straightforward to implement, the formulation (11) is prone to yielding a non-physical temperature distribution on the boundary, due to the small number of sensors available for prescribing the Dirichlet boundary condition in (11). An illustration is provided in Sect. 4 (see Fig. 3).

4 Numerical Simulations

To assess the proposed methodology, we consider a test problem designed to be representative of the physical spaces of interest to us. The domain Ω is a room of dimension 6.096 m $\times$ 3.048 m $\times$ 3.048 m, assumed to be equipped with a small number, namely 16, of temperature/humidity sensor pairs. We note that the domain size was restricted by the computing resources available to us; neither the model nor the proposed solution techniques set this limit. It may nevertheless be thought of as representing a small room. Depictions and further details are provided in Fig. 2.

To carry out the dynamical, 3D simulations as time advances, we generated a set of synthetic data representing a time series of temperature and relative humidity sensor measurements. The set of synthetic temperature data was designed to mimic trends observed in a series of unpublished charted measurements, while that for relative humidity was defined using a psychrometric chart [11]. As an initial condition for the dynamical simulations, we computed a steady-state solution of the 3D model.

Recall that in the model the temperature variation is $T' = T - \widehat{T}$, for a constant mean temperature $\widehat{T}$. We set $\widehat{T} = 19°\text{C}$. Other parameters were set from tabulated values [7], except as noted next. First, artificial conduction/diffusion was added to maintain stability, since it was not feasible to work with an increasingly refined mesh. Second, an air velocity field that captures the large-scale (versus fine-scale) pattern of air flow can be suitable within the context of the present work [1, 8]. Hence, we solved for non-turbulent air flow to reduce computational effort.

A (partial) depiction of the resulting temperature distribution on the boundary for the initial condition used in the simulations appears in Fig. 3a. Figure 3 also shows a comparison between results obtained using the proposed methodology for prescribing temperature at the boundary versus the alternative formulation (11) of mixed boundary conditions. Results from the simulations appear in Fig. 4, where both air temperature and concentration of water vapor are seen to decrease over time. The air velocity field is also depicted in Fig. 4a, where one can observe (cold)

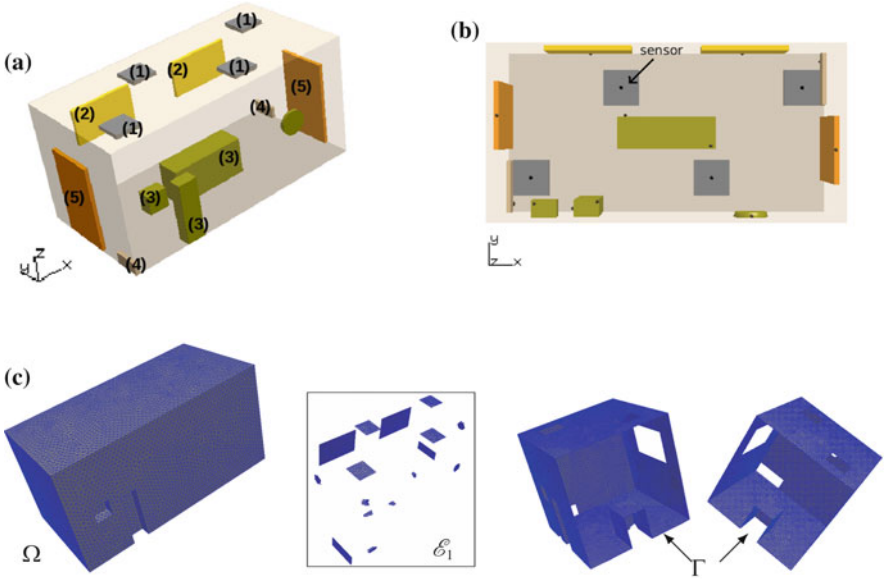


Fig. 2 (**a**) Room layout. Labels: (1) perforated tiles (supply cold air); (2) windows; (3) excluded objects; (4) air intakes; (5) doors. (**b**) Sensor locations: *Top view*; small circles denote sensor locations. (**c**) Computational domain. *Left*: 3D mesh for domain Ω. *Far right*: 2D domain $\Gamma = \mathscr{E} \setminus \mathscr{E}_1$ (cut in two to aid visualization) for solving BVPs (9) and (10)

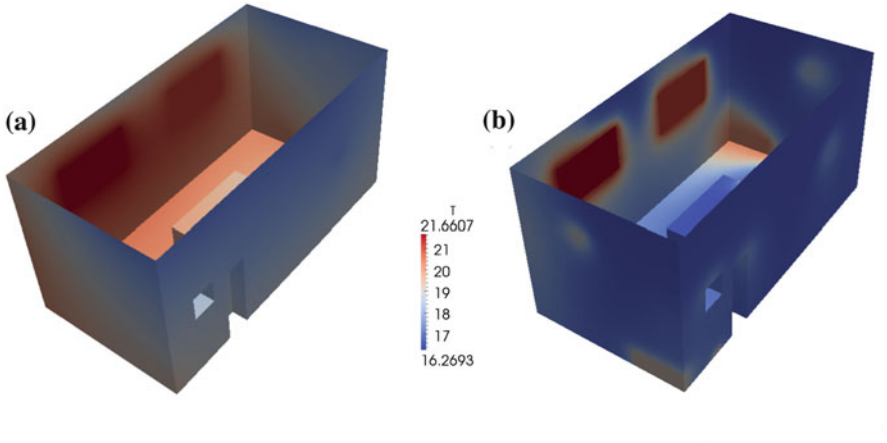


Fig. 3 Results obtained from the use of different boundary conditions. Formulation (11) yields a non-physical distribution of temperature, including sharp differences between temperature at adjacent mesh element faces, as seen in Fig. 3b. Solution of the BVP (9) yields a more smooth and natural temperature distribution, as seen in Fig. 3a. (**a**) Solution of 2D BVP (9). (**b**) Mixed boundary conditions (11)

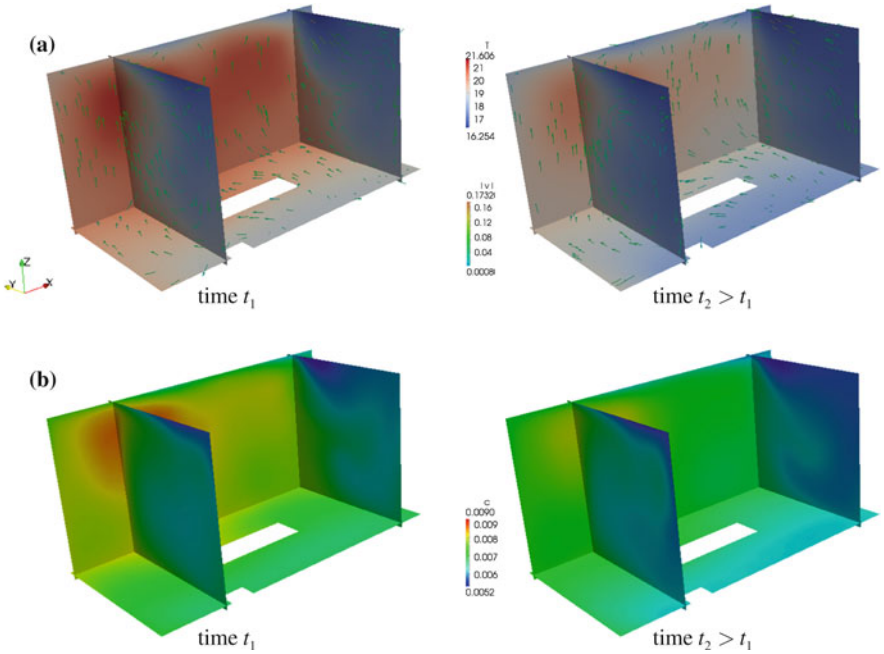


Fig. 4 Simulation results at two different time instances. (**a**) Temperature and air flow: Temperature distribution on slices of the domain, showing the temperature in the room decreasing as time has passed. The plotted arrows indicate the direction of air flow. Cold air flowing from the perforated tiles (*top of the domain*) down into the room is observed, as well as warmer air rising. (**b**) Mass fraction for the concentration of water vapor: Distribution of the mass fraction for the concentration of water vapor, seen to decrease over time

air flowing from the perforated tiles (top of the domain) down into the room. Warmer air rising can also be seen by observing the air velocity field in Fig. 4a.

To further calibrate the model, additional experimentation and validation should be performed using various time series, generated from measurements collected from a sensor network. Such time series could come, for example, from those recorded at a museum [4, 5]. Future directions include fine-tuning the model by incorporating additional terms in the 2D BVPs(9) and (10), such as those for sources/sinks, convection and time derivatives, in order to have a more detailed set of equations describing the distribution of temperature and concentration of water vapor on the boundary. Furthermore, simple turbulence modeling, such as that in the "zero equation" class of methods [1] should be explored.

Acknowledgements The author thanks members of the Physical Analytics group at IBM for discussions that led to the motivation for carrying out this work and Ognyan Stoyanov for helpful comments on a preliminary version of this paper.

References

1. Bonvini, M., Leva, A.: Object-oriented sub-zonal modelling for efficient energy-related building simulation. Math. Comput. Model. Dyn. Syst. **17**(6), 543–559 (2011)
2. Grinzato, E., Volinia, M., Girotto, M., Bison, P., Cadelano, G., Peron, F., Tos, L.: Knowing to prevent: indoor global monitoring by IR thermography. In: Proceedings of ART11 (2011)
3. IBM Research, Physical Analytics Group: http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=6570. Cited 23 Dec 2016
4. IBM Research, Physical Analytics Group: http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=6571. Cited 23 Dec 2016
5. Kargere, L., Leona, M., Tsukada, M., Winslow, A., Hamann, H.F., Klein, L., Robbins, M.A., Vici, P.D., Bermudez-Rodriguez, S.A., Schrott, A.: A technology platform for managing micro-climatic conditions in a museum environment. In: 42nd Annual AIC Meeting (2014)
6. Landau, L.D., Lifshitz, E.M.: Fluid Mechanics. Pergamon Press, Oxford (1959)
7. Lienhard IV, J.H., Lienhard V, J.H.: A Heat Transfer Textbook. Phlogiston Press, Cambridge (2011)
8. López, V., Hamann, H.F.: Heat transfer modeling in data centers. Int. J. Heat Mass Transfer **54**(25–26), 5306–5318 (2011)
9. NETGEN: Open-Source Mesh Generator. http://sourceforge.net/projects/netgen-mesher/. Cited 23 Dec 2016
10. OpenFOAM: Open-Source CFD Toolbox. www.openfoam.org. Cited 23 Dec 2016
11. Parsons, R.: ASHRAE Fundamentals Handbook. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. (ASHRAE), Atlanta (1997)
12. SENSIRION Sensor Company: Introduction to Humidity: Basic Principles on Physics of Water Vapor. www.sensirion.com. Cited 23 Dec 2016

Prototype Model of Autonomous Offshore Drilling Complex

Sergey Lupuleac, Evgeny Toropov, Andrey Shabalin, and Mikhail Kirillov

Abstract The prototype model of autonomous offshore drilling complex consists of several sub models corresponding to different considered phenomena: vibrations of drilling string; circulation of drilling mud; mud filtration; deformation of the liquid filled soil and so on. All sub models are combined into unified prototype model and exchange data during simulations.

The project of prototype model development is launched jointly by St. Petersburg Polytechnic University and Rubin ship design bureau. Specialized in software code is developed and used for simulation by applying methods from different branches of computational mechanics.

1 Introduction

The perspective way of oil and gas field development in the Arctic region is construction of autonomous drilling complexes that are mounted on the seabed and don't have direct connection to the sea surface. The technical specifications of such devices also imply the exploitation below the sheet of ice that makes them universal and independent from weather and ice conditions.

Presently this technology is still in elaborate and has no comparable counterparts. The prototype modelling is extensively used on this stage of development. The main objectives of prototype modelling are the specification of the requirements to drilling complex and design verification by performing of multiple virtual tests of the system and its components.

S. Lupuleac (✉)
Peter the Great St.Petersburg Polytechnic University, 195251, St.Petersburg, Polytechnicheskaya, 29, Russia
e-mail: lupuleac@mail.ru

E. Toropov • A. Shabalin • M. Kirillov
Central Design Bureau for Marine Engineering "Rubin", 191119, St. Petersburg, Marata, 90, Russia
e-mail: evgeny.toropov@ckb-rubin.ru; andrey.shabalin@ckb-rubin.ru; mikhail.kirillov@ckb-rubin.ru

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_102

The prototype model of autonomous offshore drilling complex consists of several sub models corresponding to different considered phenomena: vibrations of drilling string; circulation of drilling mud; mud filtration; deformation of the liquid filled soil and so on. All sub models are combined into unified prototype model and exchange data during simulations.

The project of prototype model development is launched jointly by St. Petersburg Polytechnic University and *Rubin* ship design bureau. Both commercial and in house software are used for simulation by applying methods from different branches of computational mechanics.

The defining feature of considered project is multiphysics. Let's illustrate it by naming the phenomena that need to be accounted during simulation of drill mud flow. The functions of drill mud are the transport of cuttings and compensation of ground pressure inside the well. The scheme of drill mud circulation is given by Fig. 1.

The drill mud is non Newtonian multiphase liquid with complex rheology. It contains solid and gas inclusions. The process of liquid infiltration to/from subsoil layers also needs to be accounted during mud flow simulation because this phenomenon affects both flow rate and physical properties of mud. In turn the simulation of liquid infiltration needs to take into account the formation of filter cakes, anisotropy of subsoil layers and so on. The temperature of subsoil layers dramatically increases during well deepening. So the process is highly nonisothermal and heat exchange with ground also needs to be accounted. The drill mud transports cutting particles of different size, density and shape that also needs

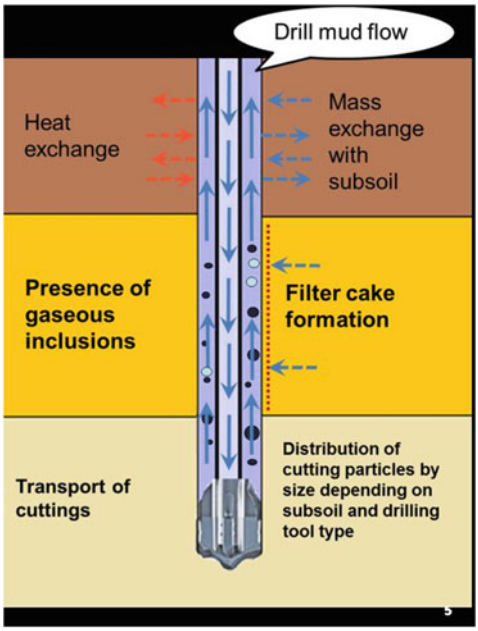


Fig. 1 The scheme of drill mud circulation

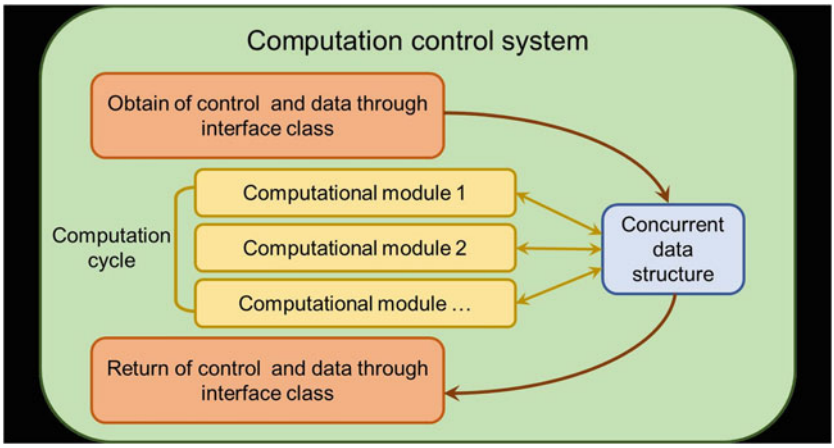


Fig. 2 The scheme of computation control system

to be considered during the simulations. Rate of drilling and distribution of sizes for cutting particles also depend on the loads and moments in the drill string as well as on its rotation speed.

2 Computation Control System

All described above phenomena need to be taken into account during the simulation of drilling process. The only feasible way of realization of such complex model is development of different computation modules combined by special computation control system as it is shown in Fig. 2.

The calculation process appears to be an iterative procedure of execution of computational modules on every time step. The global time step is defined by computation control system. The internal time step is defined individually for every computational module as some fraction of global step. After global step finalizing the computation data are passed to the interface class in order to inform the user about solution process and give him the opportunity to abort solution. Input/output data of all computational modules form unified concurrent data structure. So the computation results of one module are used as input data for other one.

3 Computational Modules

The computational modules included to the model are, as follows:

- Module of mud flow;
- Module of heat exchange;

- Module of mass exchange;
- Module of drill string dynamics;
- Drilling module.

Now let's briefly describe every computational module.

Module of mud flow is the central part of whole computational system. It utilizes the two-phase flow model [7, 8, 14] based on mixture approach [10]. The unknowns in the time-dependent equations are volume fractions of phases, velocities of phases, mixture pressure. All unknowns are functions of time and one space variable. One dimensional model is fully applicable here because of high elongation of flow channel.

Module of heat exchange is based on solving the energy transfer equation inside drill pipes and mining hole annuity. Also heat conductivity equation is solved for pipe walls and surrounding subsoil (see [11, 12]). The unknowns are the temperature of mud inside drill pipes and mining hole annuity, temperatures of pipe walls and surrounding subsoil.

Module of mass exchange is used to simulate the infiltration of flowing liquid into/from borehole. The continuity equations and state equations closed by Darcy or non-linear filtration laws [2, 4] are solved numerically [13] to obtain pressure gradients and rate of drilling mud filtrate invasion. Porous and fracturing rocks are considered.

The beam finite elements are used for simulation of drill string dynamics in the corresponding module [6, 9, 16]. Bending, longitudinal and torsional oscillations are simulated with regard to contact interaction with walls including friction; drag, inertia and damping of mud are also taken into account.

The drilling module simulates the soil crushing by the bore bit [5]. The results of simulation are rate of penetration as well as distribution of size for cutting particles to be transported by the mud. Three types of subsoil layers are considered: rock, sand and mixed. The rock grinding is simulated with Markov's branching process [3, 15]; sand particles are not crushed by the bore bit and grinding of the mixed type soil is simulated by combination previous approaches. The rate of penetration is computed as function of pressure and power of bore bit according the Rehbinder law [1].

4 Conclusion

The developed prototype model has open architecture. It means that the model can be easily supplemented by new modules corresponding to other physical phenomena. It is planned to use it as a basis of new software complex for simulation of different stages of oil production.

References

1. Andrade, E.N., Randall, R., Makin, M.: The Rehbinder effect. Proc. Roy. Soc. B **63**, 990 (1950)
2. Basniev, K.C., Dmitriev, N.M., Rosenberg, G.D.: Nephtegasovaya gidromehanika [Oil and gas hydromechanics] — M.–Igevsk: Inst. komp. issl. (2005)
3. Beznea, L., Deaconu, M., Lupas, O.: Branching processes for the fragmentation equation. Stoch. Process Appl. **125**, 1861–1885 (2015)
4. Faruk, C.: Reservoir Formation Damage. Fundamentals, Modeling, Assessment and Mitigation, 2nd edn. Elsevier Inc., Amsterdam (2007)
5. Filipov, A.F.: O raspredelenii razmerov chastitz pri droblenii (On the distribution of particle sizes by crushing).TVP. **3**(6), 299–318 (1961) (in Russian)
6. Gulyaev, V., Gaidaichuk, V., Solov'ev, I., Gorbunovicha I.: Quasistatic critical states of strings for deep dilling. Strength Mater. **38**(5), 527–534 (2006)
7. Ishii, M., Hibiki, T.: Thermo-Fluid Dynamics of Two-Phase Flow, 2nd edn., 537 pp. Springer, Berlin (2011)
8. Leonov, E.G., Isaev, V.I.: Gidroaeromechinika v burenii[Hydroaeromechanics in drilling]: Ucheb. – M.: Nedra, 304 c. (1987) (in Russian)
9. Li, J., Yan, T., Sun, X., Peng, S.: Finite element analysis on drilling string axial vibration in a crooked hole. Proceedings of ICPTT-. Wuhan, China, pp. 1328–1334 (2012)
10. Manninen, M., Taivassalo, V.: On the Mixture Model for Multiphase Flow, 67 p. Espoo (1996)
11. Marshall, D.W., Bentsen, R.G.: A computer model to determine the temperature distributions in a wellbore. J. Can. Pet. Technol. **21**(1), 63–75 (1982)
12. Mou, Y., Yingfeng, M., Gao, L., et al.: Estimation of wellbore and formation temperatures during the drilling process under lost circulation conditions. Math. Probl. Eng. **2013**, 11 (2013)
13. Partial Differential Equation Toolbox: User's Guide MATLAB R2016a, MathWorks Inc, (2016)
14. Patankar, S.V.: Numerical Heat Transfer and Fluid Flow, 205 pp. CRC Press, Boca Raton (1980)
15. Wagner, W.: Random and deterministic fragmentation models. Monte Carlo Methods Appl. **16**, 399–420 (2010)
16. Wu, A., Hareland, G., Fazaelizadeh, M.: Torque and drag analysis using finite element method. Mod. Appl. Sci. **5**(6), 13–27 (2011)

A Competitive Random Sequential Adsorption Model for Immunoassay Activity

Dana Mackey, Eilis Kelly, and Robert Nooney

Abstract Immunoassays rely on highly specific reactions between antibodies and antigens and are used in biomedical diagnostics applications to detect biomarkers for a variety of diseases. Antibody immobilization to solid interfaces through random adsorption is a widely used technique but has the disadvantage of severely reducing the antigen binding activity and, consequently, the assay performance. This paper proposes a simple mathematical framework, based on the theory known as competitive random sequential adsorption (CRSA), for describing how the activity of immobilized antibodies depends on their orientation and packing density and generalizes a previous model by introducing the antibody aspect ratio as an additional parameter which could influence the assay behaviour.

1 Introduction

Antibodies and antigens are extensively used as biomarkers for a wide variety of immunodiagnostic applications such as pregnancy testing or cancer diagnostics. Many of these technologies rely on the immobilization of antibodies on solid support interfaces, through strategies such as physical adsorption or covalent cross-linking, which result in a random particle distribution and considerably reduced antigen-binding activity. This phenomenon, generally attributed to shielding of antibody active sites and molecule denaturation upon contact with the solid surface, has been widely researched and described in the experimental literature, [8, 11, 12]. Of additional interest is the relationship between antigen-capturing activity and

D. Mackey (✉)
School of Mathematical Sciences, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland
e-mail: dana.mackey@dit.ie

E. Kelly
Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland
e-mail: eilis.kelly@dit.ie

R. Nooney
Biomedical Diagnostics Institute, Dublin City University, Dublin 9, Ireland
e-mail: rob.nooney@dcu.ie

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_103

antibody surface density, with many papers reporting that crowding of antibodies has a detrimental effect on assay signal, [2, 8, 10, 12]. Therefore, one of the primary considerations in immunoassay design is optimizing the concentration of capture antibody in order to achieve maximal antigen binding and, subsequently, improved sensitivity and limit of detection.

Antibodies are large Y-shaped proteins composed of two regions: a fragment crystallizable (Fc) region at the base (responsible for recruiting components of the immune system) and a fragment antigen binding (Fab) region at the top (see Fig. 1). Each arm of the Fab region contains a hypervariable region at its tip, called a paratope, which is capable of binding strongly to one epitope on an antigen. The immobilized antibody orientation is usually described in the immunoassay literature (see, for example, [11, 12]) by one of the following positions: "end on" (the ideal position, with the active Fab region pointing towards the solution), "head on" (or Fab down), "flat on" and "side on" (both of which correspond, in our 1-dimensional model, to a side orientation, as shown in Fig. 1).

The theory known as *random sequential adsorption* (RSA) has been successfully used over the past few decades to describe monolayer particle deposition, with wide applications in many physical and biological settings ([1, 4, 7]). In the standard description, uniform particles are transported to a surface at a constant rate and irreversibly deposited. The adsorption site is chosen randomly but, if the particle overlaps an already deposited one, it is rejected. The process is continued until the *jamming limit* is reached, that is no spaces are left on the substrate. In one dimension, this process is commonly referred to as "the car parking problem" (or interval filling) and the jamming coverage, also known as the Rényi constant, has been calculated in [6] to be $C_R \approx 0.74756$. A generalization known as competitive RSA (see, for example, [3]) deals with mixtures of particles having different sizes and adsorption probabilities.

We have recently presented a new mathematical model for quantifying the activity of antibodies immobilized on a solid surface, based on the RSA framework, [5]. In this model, the antibodies (or any other active molecules) were represented by oriented circles whose diameters cover a 1-dimensional substrate in the same manner as intervals fill a line in the standard car parking problem. A simple procedure was developed for estimating the percentage of antibodies which are

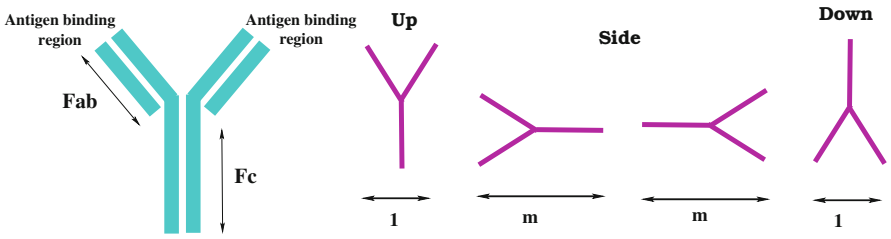


Fig. 1 Antibody structure and immobilization orientations. Our model assumes that an immobilized particle covers a length of either 1 or m on the 1-dimensional substrate

correctly oriented (Fab site accessible to antigens in solution) out of total antibody coverage. Of particular interest is identifying parameter ranges for which the activity (and consequently the assay signal) increases with antibody concentration and then sharply decreases (the so-called "hook effect", see [8, 10]) which indicates the existence of a well-defined optimal surface coverage.

The current paper generalizes our previous model to address the more realistic case where different immobilization configurations occupy intervals of different sizes in the above description. More specifically, the antibody width is assumed larger that its height and so the side-on and head-on orientations can be thought of as two competing adsorption strategies and described within the more general competitive RSA theory ([3]). The antibody aspect ratio is introduced in the analysis as another parameter which might influence the biosensor functionality.

2 A Model for Estimating Surface Antibody Activity

The binary mixture RSA model described in [3] assumes that two particles of sizes 1 and m ($m > 1$) are selected with probabilities q and $p = 1 - q$, respectively, to be deposited randomly on a one dimensional substrate of length l. The function $N(x, t)$ is defined to be the *gap length density function* at time t (so $N(x, t)dx$ represents the mean number of gaps with length between x and $x + dx$) and we let $P(x, t) = N(x, t)/l$. As $l \to \infty$, the time evolution of the gap size density is described by the following kinetic model

$$\frac{\partial P(x,t)}{\partial t} = \begin{cases} -(x-\sigma)P(x,t) + 2q\int_{x+1}^{\infty} P(y,t)\,dy + 2p\int_{x+m}^{\infty} P(y,t)\,dy, & m < x < \infty \\ -q(x-1)P(x,t) + 2q\int_{x+1}^{\infty} P(y,t)\,dy + 2p\int_{x+m}^{\infty} P(y,t)\,dy, & 1 < x \le m \\ 2q\int_{x+1}^{\infty} P(y,t)\,dy + 2p\int_{x+m}^{\infty} P(y,t)\,dy, & 0 < x \le 1, \end{cases} \tag{1}$$

where the terms in all equations represent rates at which gaps of length x can be created or destroyed, while $\sigma = mp + q$ denotes the average size of incoming molecules. (The kinetic equations for $P(x, t)$ in the single particle case were described in [4].) Using the initial conditions $P(x, 0) = 0$ and $\lim_{t\to 0}\int_0^{\infty} xP(x, t)\,dx = 1$ an analytical solution can be found, provided $1 < m \le 2$, as follows:

$$P(x,t) = \begin{cases} t^2 F(t)\, e^{-(x-\sigma)t}, & m < x < \infty \\ 2\, e^{-q(x-1)t}\int_0^t F(s) s\left[qe^{(mp-1)s} + pe^{-qms}\right] e^{-xps}\,ds, & 1 < x \le m \\ 2\int_0^t s\,F(s)\, e^{\sigma s}\left[qe^{-(x+1)s} + pe^{-(x+m)s}\right] ds, & m-1 < x \le 1 \\ 4q\int_0^t e^{qu}\int_0^u sF(s)\frac{qe^{(mp-1)s}+pe^{-qms}}{qu+ps}\times & \\ \quad\times\left[e^{-(qu+ps)(x+1)} - e^{-(qu+ps)m}\right] ds\,du & \\ \quad +2\int_0^t u\,F(u)\left[qe^{(\sigma-m)u} + pe^{(\sigma-x-m)u}\right] du, & 0 < x \le m-1 \end{cases} \tag{2}$$

where

$$F(t) = \exp\left[-2\int_0^t \frac{q(1-e^{-u}) + p(1-e^{-mu})}{u}\,du\right]. \tag{3}$$

The total coverage, $\theta_T(t) = \theta_S(t) + \theta_L(t)$, is comprised of the coverage by small particles, $\theta_S(t)$, and that of the large particles, $\theta_L(t)$, which are calculated in [3] as

$$\begin{aligned}\theta_S(t) &= 2\int_0^t sF(s)e^{-\sigma s}\Big[p + qe^{(m-1)s}\Big] \times \left[\frac{e^{-(m-1)(ps+qt)} - 1}{ps+qt} - \frac{e^{-(m-1)s} - 1}{s}\right] ds \\ &\quad + q\int_0^t F(s)\Big[1 + (m-1)s\Big]e^{-(m-1)qs}\,ds, \qquad (4) \\ \theta_L(t) &= pm\int_0^t F(s)e^{-(m-1)qs}\,ds.\end{aligned}$$

Note that the jamming coverage $\theta_T(\infty)$ in the binary case (which depends on m and q) exceeds the Renyi number for the one particle RSA theory and satisfies

$$0.7475.. = C_R \le \theta_T(\infty) \le C_R + C_R\,(1 - C_R) = 0.937\ldots$$

where the extreme values are obtained for $q = 0$ and $m \to \infty,\ q \to 0$, respectively.

We assume for simplicity that, in our 1-dimensional description, antibodies can only adsorb in one of the four positions shown in Fig. 1 (Fab site pointing up, down, left or right) and can only cover two possible lengths on the substrate (1 or m). We also assume that the probability of an up/down orientation (corresponding to the short side) is q while the probability of a left/right (long side) adsorption is $p = 1-q$. Similar to the model presented in [5], a particle is defined to be active if, either its binding site (Fab region) is pointing up, or else pointing left or right and a gap of length at least δ exists between the binding site and the neighbouring adsorbed molecule, where $\delta \le 1$ depends on the size of the oncoming reactant molecules.

We now write the gap density function as

$$P(x,t) = P_L(x,t) + P_S(x,t),$$

where P_L and P_S denote the density of gaps with a large particle (or small particle, respectively) on the left. We similarly expand P_L as

$$P_L(x,t) = P_{LL}(x,t) + P_{LS}(x,t),$$

where P_{LL}, P_{LS} denote the gaps with a large particle on the left and a large (or small, respectively) particle on the right. The total number of active molecules at any given time, N_{active}, is then calculated by adding all "up" particles, all "left" particles with enough space on their left (which is obtained by multiplying the

percentage of particles pointing left by the total number of gaps $\geq \delta$) and all "right" particles with enough space on their right (obtained in the same way as the left case). The "down" particles are assumed inactive. Moreover, if two adjacent particles are pointing towards the gap between them and this gap is not large enough to fit two antigens then only one of the immobilized particles is considered active. We can then write

$$\frac{1}{l} N_{active} = \frac{1}{2} \theta_S + \int_{\delta}^{\infty} P_L(x,t)\, dx - \frac{1}{4} \int_{\delta}^{2\delta} P_{LL}(x,t)\, dx. \tag{5}$$

The density P_L can be calculated using the same approach as that described in the previous section for $P(x,t)$. For example, when $x > m$, this function satisfies the integro-differential equation

$$\frac{\partial P_L}{\partial t} = -(x-\sigma)P_L(x,t) + q \int_{x+1}^{\infty} P_L(y,t)\, dy + p \int_{x+m}^{\infty} P_L(y,t)\, dy + ptF(t)e^{-(x+m-\sigma)t},$$

with similar equations for other x ranges. This can be solved to give

$$P_L(x,t) = \begin{cases} t\, F_L(t)\, e^{-(x-\sigma)t}, & m < x < \infty, \\ e^{-q(x-1)t} \int_0^t G(m-1,s) e^{-(xp+qm)s}\, ds, & 1 < x \leq m, \\ \int_0^t e^{-(x+m-\sigma)s} G(m-1,s)\, ds, & m-1 < x \leq 1, \\ q \int_0^t e^{-(m-1)qu} \int_0^u \frac{e^{-ms}}{qu+ps} G(m-1,s) & \\ \qquad \times \left[e^{-(x-m+1)(qu+ps)} - 1 \right] ds\, du & \\ \qquad + \int_0^t G(x,u) e^{-(x+m-\sigma)u}\, du, & 0 < x \leq m-1, \end{cases} \tag{6}$$

where

$$F_L(t) = p\, F^{1/2}(t) \int_0^t F^{1/2}(s) e^{-ms}\, ds, \qquad G(z,t) = qF_L(t)e^{zt} + pF_L(t) + ptF(t)$$

and $F(t)$ is defined in (3). This procedure is repeated once more and, finally, we find

$$P_{LL}(x,t) = \begin{cases} 2pe^{-(x-\sigma)t} \int_0^t F_L(s) e^{-ms}\, ds, & m < x < \infty, \\ 2pe^{-q(x-1)t} \int_0^t F_L(s) e^{-s(px+qm)}\, ds & 1 < x \leq m, \\ 2p \int_0^t F_L(s) e^{(\sigma-x-m)s}\, ds, & 0 < x \leq 1. \end{cases} \tag{7}$$

The derivation of N_{active} is greatly simplified if we assume that the gap distribution is independent of the particle size and hence the densities P_L and P_S are proportional to the number ratios of large and small particles, that is

$$P_L(x,t) = \alpha_L(t)\, P(x,t), \qquad P_S(x,t) = \alpha_S(t)\, P(x,t),$$

where

$$\alpha_L(t) = \frac{\theta_L(t)}{\theta_L(t) + m\theta_S(t)}, \qquad \alpha_S(t) = \frac{m\theta_S(t)}{\theta_L(t) + m\theta_S(t)}.$$

With this assumption, the concentration of active particles can be calculated as

$$\frac{1}{l} N_{active} = \frac{1}{2}\,\theta_S + \alpha_L \int_{\delta}^{\infty} P(x,t)\,dx - \frac{1}{4}\,\alpha_L^2 \int_{\delta}^{2\delta} P(x,t)\,dx, \tag{8}$$

which only uses the total gap density function $P(x,t)$ given in (2).

3 Results and Conclusions

The number of active particles is evaluated from (5) or (8), using the previously calculated density functions $P_L(x,t)$, $P_{LL}(x,t)$ or $P(x,t)$. (Note that the two approaches to calculating N_{active} give identical results.) The active particle concentration is then plotted as a function of the total concentration of immobilized antibodies, given by $\theta_S + \frac{\theta_L}{m}$, in Fig. 2. We consider the cases of large antigen particles ($\delta = 1$, so the antigen size is equal to that of the antibody) and small antigen particles ($\delta = 0.1$), each of these relevant to various biosensing applications. As expected, the antibody activity is consistently higher for small antigens, due to improved lateral accessibility (for side-lying particles). Recall that the total antibody coverage (and, hence, concentration), calculated as $\theta_T = \theta_S + \theta_L$ from (4), varies with the parameters m (the aspect ratio) and q (the probability of upright orientation).

The two groups of curves in each figure correspond to $q = 0.1$ (side orientation prevails) and $q = 0.8$ (most antibodies adsorb in the up/down positions). Within each group, the three curves (distinguished by different line styles) correspond to

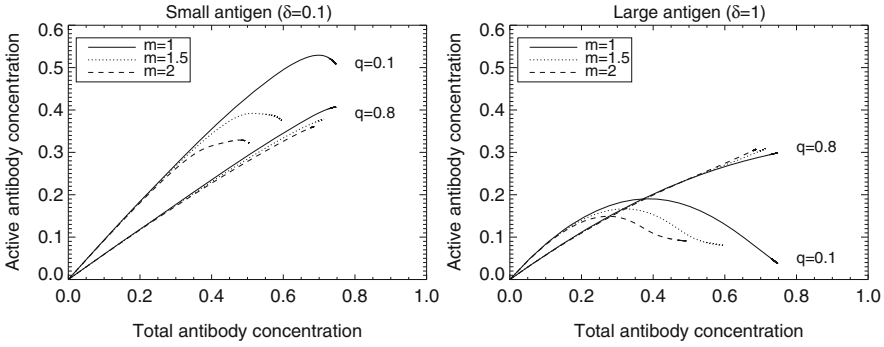


Fig. 2 Active antibody concentration, N_{active}/l, as a function of total antibody concentration, $\theta_S + (\theta_L/m)$, for small ($\delta = 0.1$) and large ($\delta = 1$) antigen particles

$m = 1$, $m = 1.5$ and $m = 2$. (The case $m = 1$ has already been discussed in [5]). The side orientation is generally associated with the hook effect (high activity and antigen-binding rate at low antibody coverage, which then decrease at higher concentrations), which is consistent with results reported in the experimental literature, [8, 12, 13], and explained by increased steric hindrance to antigen access.

The aim of the current paper was to assess the effect of the antibody aspect ratio, m, on the immunoassay behaviour. It can be seen from Fig. 2 that m strongly influences activity when the side orientation predominates but has relatively little importance when antibodies are adsorbed in the up/down configurations. In the latter case, it would be acceptable to approximate the particle aspect ratio by $m = 1$ and use our much simpler model in [5] to simulate the immobilized antibody activity. By contrast, when q is small, it is seen that higher values of m correspond to lower activity and so particles with $m = 1$ seem to be the most efficient.

As discussed in [5], many qualitative features of this modelling framework reproduce immunoassay properties reported in the experimental literature and support the conclusion of [9] that optimal performance is determined by the interplay of factors such as immobilized antibody density, relative size of antigens and method of immobilization. Therefore, a theoretical study of the relationship between these parameters should lead to improved design and functionality of such devices.

Acknowledgements The second author gratefully acknowledges financial support from the Irish Research Council.

References

1. Evans, J.W.: Random and cooperative sequential adsorption. Rev. Mod. Phys. **65**, 1281 (1993)
2. Gubala, V., Crean, C., Nooney, R., Hearty, S., McDonnell, B., Heydon, K., O'Kennedy, R., MacCraith, B.D., Williams, D.E.: Kinetics of immunoassays with particles as labels: effect of antibody coupling using dendrimers as linkers. Analyst **136**, 2533–2541 (2011)
3. Hassan, M.K., Schmidt, J., Blasius, B., Kurths, J.: Jamming coverage in competitive random sequential adsorption of a binary mixture. Phys. Rev. E **65**, 045103 (2002)
4. Krapivsky, P.L.: Kinetics of random sequential parking on a line. J. Stat. Phys. **69**, 135–150 (1992)
5. Mackey, D., Kelly, E., Nooney, R.: Modelling random antibody adsorption and immunoassay activity. Math. Biosci. Eng. **13**, 6, 1159–1168 (2016)
6. Rényi, A.: On a one-dimensional problem concerning random space-filling (In Hungarian). Publ. Math. Inst. Hung. Acad. Sci. **3**, 109–127 (1958)
7. Roach, J.C., Thorsson, V., Siegel, A.F.: Parking strategies for genome sequencing. Genome Res. **10**, 1020–1030 (2000)
8. Saha, B., Evers, T.H., Prins, M.W.J.: How antibody surface coverage on nanoparticles determines the activity and kinetics of antigen capturing for biosensing. Anal. Chem. **86**, 8158–8166 (2014)
9. Schramm, W., Paek, S.: Antibody-antigen complex formation with immobilized immunoglobulins. Anal. Biochem. **205**, 47–56 (1992)

10. Vareiro, M.L.M., Liu, J., Knoll, W., Zak, K., Williams, D., Toby, A., Jenkins, A.: Surface plasmon fluorescence measurements of human chorionic gonadotrophin: role of antibody orientation in obtaining enhanced sensitivity and limit of detection. Anal. Chem. **77**, 2426–2431 (2005)
11. Wiseman, M.E., Frank, C.W.: Antibody adsorption and orientation on hydrophobic surfaces. Langmuir **28**, 1765–1774 (2012)
12. Xu, H., Lu, J.R., Williams, D.E.: Effect of surface packing density of interfacially adsorbed monoclonal antibody on the binding of hormonal antigen human chorionic gonadotrophin. J. Phys. Chem. B **110**, 1907–1914 (2006)
13. Zhao, X., Pan, F., Cowsill, B., Lu, J.R., Garcia-Gancedo, L., Flewitt, A.J., Ashley, G.M., Luo, J.: Interfacial immobilization of monoclonal antibody and detection of human prostate-specific antigen. Langmuir **27**, 7654–7662 (2011)

A Finite Volume Scheme for Darcy-Brinkman's Model of Two-Phase Flows in Porous Media

Houssein Nasser El Dine, Mazen Saad, and Raafat Talhouk

Abstract In this paper, we are interested in the displacement of two incompressible phases in a Darcy-Brinkman flow in a porous media. The equations are obtained by the conservation of the mass and by considering the Brinkman regularization velocity of each phase. This model is treated in its general form with the whole nonlinear terms. This paper deals with construction and convergence analysis of a finite volume scheme together with a phase-by-phase upstream according to the total velocity. Finally, numerical tests illustrate the behavior of the solutions of this proposed scheme.

1 The Darcy-Brinkman Model

Different empiric laws are used to describe the filtration of a fluid through porous media. Darcy law is the most popular one, due to its simplicity. It states that the filtration velocity of the fluid is proportional to the pressure gradient. Darcy law cannot sustain the no-slip condition on an impermeable wall or a transmission condition on the contact with free flow [10, 11]. That motivated H. Brinkmann in 1947 to modify the Darcy law in order to be able to impose the no-slip boundary condition on an obstacle submerged in porous medium. He assumed large

H. Nasser El Dine (✉)
École Centrale de Nantes. UMR 6629 CNRS, Laboratoire de Mathématiques Jean Leray, F-44321 Nantes, France

Doctoral School of Sciences and Technology (EDST), Laboratory of Mathematics, Lebanese University, Beirut, Hadath-Lebanon
e-mail: houssein.nasser-el-dine@ec-nantes.fr

M. Saad
École Centrale de Nantes. UMR 6629 CNRS, Laboratoire de Mathématiques Jean Leray, F-44321 Nantes, France
e-mail: mazen.saad@ec-nantes.fr

R. Talhouk
Doctoral School of Sciences and Technology (EDST), Laboratory of Mathematics, Lebanese University, Beirut, Hadath-Lebanon
e-mail: rtalhouk@ul.edu.lb

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_104

permeability to compare his law with experimental data and assumed that the second viscosity μ equals the physical viscosity of the fluid in the case of monophasic flow. Brinkman law could formally be obtained from the Stokes system describing the microscopic flow, by adding the resistance to the flow [1].

Let $T > 0$, fixed and let Ω be a bounded set of $\mathcal{R}^d$ $(d \geq 1)$. We set $Q_T = (0,T) \times \Omega$, $\Sigma_T = (0,T) \times \partial\Omega$, and η is the outward normal to the boundary $\partial\Omega$. The mass conservation of two incompressible phases with Darcy-Brinkman velocity of each phase can be written in the following system, see [6, 8] for more details,

$$\begin{cases} \phi\partial_t s - \mu\phi\Delta\partial_t s - \operatorname{div}\left(kf(s)\lambda(s)\nabla p\right) - \operatorname{div}\left(k\alpha(s)\nabla s\right) = 0, & \text{in } Q_T \\ -\operatorname{div}\left(k\lambda(s)\nabla p\right) = 0, & \text{in } Q_T \\ \displaystyle\int_\Omega p(t,x)\,\mathrm{d}x = 0, & \text{in } (0,T) \\ k\lambda(s)\nabla p(t,x)\cdot\eta = \pi(t,x), & \text{on } \Sigma_T \\ \mu\phi\partial_\eta\partial_t s + f(s)\pi + k\alpha(s)\nabla s\cdot\eta = h(t,x), & \text{on } \Sigma_T \\ s(0,\mathbf{x}) = s_0,(\mathbf{x}) & \text{in } \Omega. \end{cases} \tag{1}$$

In the above model, the saturation (volume fraction) of one of the phases is represented by $s = s(\mathbf{x},t)$ and the other saturation phase is stated to be $1-s$. Next, $\lambda(s) = k_1(s)/\mu_1 + k_2(s)/\mu_2$ is the total mobility with k_i and μ_i are the relative permeability and the viscosity of the phase $i = 1,2$; and $f(s) = k_1(s)/(\mu_1\lambda(s))$ is the fractional flow. Furthermore, the function $\alpha(s) = k_1(s)k_2(s)p_c'(s)/(\mu_1\mu_2\ \lambda(s))$ is the capillary pressure term.

We give the main assumptions made about the system:

(H1) The functions f, λ and α are continuous bounded and nonnegative.
(H2) The fractional flow satisfies $f(0) = 0$, $f(1) = 1$, and $0 \leq f(u) \leq 1$.
(H3) There exist a positive constant $\lambda_\star$, such that $0 < \lambda_\star \leq \lambda(\cdot)$.
(H4) The function $\pi : \Sigma_T \longrightarrow \mathbb{R}$ is a bounded function in $L^2(\Sigma_T)$ with $\int_{\partial\Omega} \pi\,\mathrm{d}\sigma = 0$.
(H5) The initial function $s_0 \in H^1(\Omega)$, and the function $h \in L^2(\Sigma_T)$.

In the sequel, we use the Lipschitz continuous nondecreasing function $\beta : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ defined by

$$\beta(s) := \int_0^s \alpha(\zeta)\,\mathrm{d}\zeta, \qquad \forall s \in \mathbb{R}^+.$$

Numerical discretization and simulation of multi-phase flow in porous media with Darcy velocity have been the object of several studies during the past decades. There is an extensive literature on the approximation of incompressible two-phase flows with Darcy's velocities for each phase. For instance the finite difference method can be found in the book of Aziz and Settari [2]. The method of finite element methods (mixed or hybrid) is also studied in the past years see Chavent et al. [4] and Chen et al. [5]. For the incompressible immiscible model, let us mention

the work of Eymard, Herbin and Michel [7] where a coupled scheme, consisting in a finite volume method together with a phase-by-phase upstream weighting scheme, is analyzed on orthogonal admissible mesh. In [3], the authors propose to explore the limit of a finite volume scheme of two-phase flow model with discontinuous capillary pressure. For that, they propose to study the limit of the upstream finite volume scheme for the incompressible immiscible model. In [9], the authors suggest a second order accurate finite volume scheme for the Richards equation.

At our knowledge, the numerical analysis of finite volume scheme of the problem (1) is not studied on triangular mesh. Nevertheless, we mention the pioneer work [6] on the convergence analysis of the finite difference scheme of the Brinkman regularization of two incompressible phase flows in porous media. Our aim is to propose a finite volume scheme based on upstream approach of fractional flow with respect to the gradient of the global pressure.

2 The Nonlinear FV Scheme For Darcy-Brinkman System

Now, we want to discretize problem (1). Let $\mathcal{T}$ be a regular and admissible mesh of the domain Ω, constituting of open and convex polygons called control volumes with maximum size (diameter) h. For all $K \in \mathcal{T}$, let x_K be the center of K, $N(K)$ the set of the neighbors of K i.e, the set of cells of $\mathcal{T}$ which have a common interface with K, $N_{int}(K)$ the set of the neighbors of K located in the interior of $\mathcal{T}$, $N_{ext}(K)$ the set of edges of K on the boundary $\partial\Omega$. Furthermore, for all $L \in N_{int}(K)$, let us denote by $d_{K,L}$ the distance between x_K and x_L, by $\sigma_{K,L}$ the interface between K and L, by $\eta_{K,L}$ the unit normal vector to $\sigma_{K,L}$ outward to K. And for all $\sigma \in N_{ext}(K)$, denote by $d_{K,\sigma}$ the distance from x_K to σ.

For all $K \in \mathcal{T}$, we denote by $|K|$ the measure of K. The **admissibility** of $\mathcal{T}$ implies that $\overline{\Omega} = \cup_{K\in\mathcal{T}} \overline{K}$, $K \cap L = \emptyset$ if $K, L \in \mathcal{T}$ and $K \neq L$, and there exist a finite sequence of points $(x_K)_{K\in\mathcal{T}}$ such that **the straight line $\overline{x_K x_L}$ is orthogonal to the edge** $\sigma_{K,L}$. We suppose the classical regularity assumption on the mesh :

$$\min_{K\in\mathcal{T},\ L\in N(K)} \frac{d_{K,L}}{diam(K)} \geq \delta, \text{ for some } \delta \in \mathbb{R}^+.$$

We denote by $H_h(\Omega) \subset L^2(\Omega)$ the space of functions which are piecewise constant on each control volume $K \in \mathcal{T}$. For all $w_h \in H_h(\Omega)$ and for $K \in \mathcal{T}$, we denote by W_K the constant value of w_h in K. Next, we let $K \in \mathcal{T}$ and $L \in N(K)$ with common vertexes $(a_{l,K,L})_{1\leq l\leq I}$ with $I \in \mathbb{N}^*$. We define the diamond $T_{K,L}$ (respectively $T^{ext}_{K,L}$ for $\sigma \in N_{ext}(K)$) to be the open and convex polygon with vertex (x_K, x_L) (respectively x_K) and $(a_{l,K,L})_{1\leq l\leq I}$.

The **discrete gradient** $\nabla_h w_h$ of a function w_h is defined as the constant per diamond $T_{K,L}$:

$$\nabla_h w_h(x) = \begin{cases} d\frac{W_L - W_K}{d_{K,L}}\eta_{K,L} \text{ if } x \in T_{K,L}, \\ d\frac{W_\sigma - W_K}{d_{K,\sigma}}\eta_{K,\sigma} \text{ if } x \in T^{ext}_{K,\sigma}. \end{cases}$$

A finite volume scheme for the discretization of problem (1) is given by the following set of equations with unknowns $P_h = (P_K^{n+1})_{K\in\mathcal{T}}$, $n \in [0, N]$, and $S_h = (S_K^{n+1})_{K\in\mathcal{T}}$, $n \in [0, N]$, for all $K \in \mathcal{T}$ and $n \in [0, N]$

$$|K|\frac{S_K^{n+1} - S_K^n}{\Delta t} - \mu \sum_{L\in N(K)} \frac{|\sigma_{K,L}|}{d_{K,L}}\Big(\frac{S_L^{n+1} - S_L^n}{\Delta t} - \frac{S_K^{n+1} - S_K^n}{\Delta t}\Big)$$

$$- \sum_{L\in N(K)} \frac{|\sigma_{K,L}|}{d_{K,L}}\big(\beta(S_L^{n+1}) - \beta(S_K^{n+1})\big) + \sum_{L\in N(K)} G_s(S_K^{n+1}, S_L^{n+1}, dp_{K,L}^{n+1})$$

$$= |\partial K \cap \partial\Omega| h_K^{n+1}, \tag{2}$$

$$- \sum_{L\in N(K)} G_p(S_K^{n+1}, S_L^{n+1}, dp_{K,L}^{n+1}) = |\partial K \cap \partial\Omega|\pi_K^{n+1}, \tag{3}$$

$$\sum_{k\in\mathcal{T}_h} |K| P_K^{n+1} = 0, \tag{4}$$

with

$$S_K^0 = \frac{1}{|K|}\int_K s_0(x)\,\mathrm{d}x,\ P_K^0 = \frac{1}{|K|}\int_K p_0(x)\,\mathrm{d}x,\ h_K^{n+1} = \frac{1}{|\partial K \cap \partial\Omega|}\int_{\partial K\cap\partial\Omega} h(t,x)\,\mathrm{d}x,$$

$$\pi_K^{n+1} = \frac{1}{|\partial K \cap \partial\Omega|}\int_{\partial K\cap\partial\Omega} \pi(t,x)\,\mathrm{d}x, \quad dp_{K,L}^{n+1} = \frac{|\sigma_{K,L}|}{d_{K,L}}\big(P_L^{n+1} - P_K^{n+1}\big).$$

The numerical fluxes G_p and G_s denote the upstream scheme of fluxes $\lambda(s)$ and $\lambda(s)f(s)$, respectively, according to the gradient of pressure, precisely

$$G_p(a, b, c) = \lambda(b)c^+ - \lambda(a)c^-$$

and

$$G_s(a, b, c) = c^+\big((f\lambda)_\uparrow(a) + (f\lambda)_\downarrow(b)\big) - c^-\big((f\lambda)_\uparrow(b) + (f\lambda)_\downarrow(a)\big),$$

with

$$(f\lambda)_{\uparrow}(z) := \int_0^z ((f\lambda)'(s))^+ \, ds, \ (f\lambda)_{\downarrow}(z) := -\int_0^z ((f\lambda)'(s))^- \, ds,$$

and $s^+ = \max(s, 0), \ s^- = \max(-s, 0)$.

Notice that, we can take a general form of function G_s that satisfies the following main properties: monotony, consistency and conservation.

3 Discrete Estimates, Existence and Convergence of the Scheme

We note that all the upcoming numerical results are discussed and detailed in the forthcoming paper. Now we give some discrete properties on the FV scheme (2)–(4).

Proposition 1 (Estimate on Pressure) *For all $n \geq 0$, there exists a constant $C(\Omega) > 0$ independent of h such that*

$$\sum_{K\in\mathcal{T}_h} \sum_{L\in N(K)} \frac{|\sigma_{K,L}|}{d_{K,L}} \left(P_L^{n+1} - P_K^{n+1}\right)^2 \leq C(\Omega)||\pi(t,\cdot)||^2_{L^2(\partial\Omega)}. \tag{5}$$

Proposition 2 (Estimate on Saturation) *For all $n \geq 0$, there exists a two positive constants C and C_1 independent of h such that*

$$\sum_{K\in\mathcal{T}_h} |K|\left|S_K^n\right|^2 + \sum_{K\in\mathcal{T}_h} \sum_{L\in N(K)} \frac{|\sigma_{K,L}|}{d_{K,L}} \left|S_L^n - S_K^n\right|^2 \leq g(t), \tag{6}$$

$$\sum_{K\in\mathcal{T}_h} |K|\left(\frac{S_K^{n+1} - S_K^n}{\Delta t}\right)^2 + \sum_{K\in\mathcal{T}_h} \sum_{L\in N(K)} \frac{|\sigma_{K,L}|}{d_{K,L}} \left(\frac{S_L^{n+1} - S_L^n}{\Delta t} - \frac{S_K^{n+1} - S_K^n}{\Delta t}\right)^2 \leq C_1 g(t),$$

where

$$g(t) = \Big(\sum_{K\in\mathcal{T}_h} |K|\left|S_K^0\right|^2 + \sum_{K\in\mathcal{T}_h} \sum_{L\in N(K)} \frac{|\sigma_{K,L}|}{d_{K,L}} \left|S_L^0 - S_K^0\right|^2 + \|h(t,\cdot)\|^2_{L^2(\partial\Omega)} + ||\pi(t,\cdot)||^2_{L^2(\partial\Omega)} \Big) \exp\left(\frac{CT}{\mu(1 - C\Delta t_0)}\right).$$

Proposition 3 (Existence of Discrete Solutions) *There exists at least one solution $\left(S_K^{n+1}, P_K^{n+1}\right)_{K\in\mathcal{V}}$ to the scheme (2)–(4).*

Estimates (5) and (6) permit to get weak convergences on the gradient of the pressure and the saturation. To handle with the nonlinear terms, a strong convergence of the saturation should be established. For that, we establish estimations on the time and space translates of saturation.

We are in a position to state the main result on the convergence of the FV scheme (2)–(4) towards the weak solution of the continuous system (1) as the time and space discretization steps go to zero. Firstly, we establish the following lemmas before stating the main theorem. In what follows, $\forall t \in [t_n, t_{n+1}]$, we denote by $\bar{s}_h(t,x) = (s(t_{n+1},x) - s(t_n,x))(t-t_n)/\Delta t + s(t_n,x)$, $s_{h|_{Q_K^n}} = S_K^{n+1}$ and $p_{h|_{Q_K^n}} = P_K^{n+1}$ where $Q_K^{n+1} =]t^n, t^{n+1}] \times K$.

Lemma 1 (Space and Time Translation Estimates) *There exists a positive constant C depending on Ω, T and s_0 such that*

$$\iint_{\Omega' \times [0,T]} \left|\bar{s}_h(t, x+\xi) - \bar{s}_h(t,x)\right|^2 \, \mathrm{d}x \, \mathrm{d}t \leq C|\xi|(|\xi| + 2h),$$

for all $\xi \in \mathbb{R}^d$, with $\Omega' = \{x \in \Omega,\ [x, x+\xi] \subset \Omega\}$, and

$$\int_0^{T-\tau} \int_\Omega |\bar{s}_h(t+\tau, x) - \bar{s}_h(t,x)|^2 \, \mathrm{d}x \, \mathrm{d}t \leq C(\tau + \Delta t), \ \ \forall \tau \in]0, T[.$$

Lemma 2 *There exists a subsequence, still denoted (s_h, p_h), such that, as $h \longrightarrow 0$*

$$\begin{aligned}
&s_h \longrightarrow s \text{ strongly in } L^p(Q_T) \text{ and a.e in } Q_T \text{ for all } p \geq 1,\\
&\nabla_h s_h \rightharpoonup \nabla s \text{ weakly in } (L^2(Q_T))^d,\\
&\nabla_h \beta(s_h) \rightharpoonup \nabla \beta(s) \text{ weakly in } (L^2(Q_T))^d,\\
&p_h \rightharpoonup p \text{ weakly in } L^2(Q_T),\\
&\nabla_h p_h \rightharpoonup \nabla p \text{ weakly in } (L^2(Q_T))^d,\\
&\partial_t \bar{s}_h \rightharpoonup \partial_t s \text{ weakly in } L^2(Q_T),\\
&\nabla \partial_t \bar{s}_h \rightharpoonup \nabla \partial_t s \text{ weakly in } (L^2(Q_T))^d.
\end{aligned}$$

Finally, we arrive to state the main theorem

Theorem 1 *Assume (H1)–(H5) hold. Then the functions p and s defined in lemma 2 constitute a weak solution of the system (1) in the sense*

$$s \in W^{1,2}\big((0,T); H^1(\Omega)\big), \quad p \in L^2\big((0,T); H_M^1(\Omega)\big)$$

satisfying for all $\varphi \in L^2((0,T);H^1(\Omega))$,

$$\int_0^T \int_\Omega \partial_t s\,\varphi \,\mathrm{d}t\,\mathrm{d}x + \mu \int_0^T \int_\Omega \nabla \partial_t s \cdot \nabla \varphi \,\mathrm{d}t\,\mathrm{d}x + \int_0^T \int_\Omega f(s)\lambda(s)\nabla p \cdot \nabla \varphi \,\mathrm{d}t\,\mathrm{d}x$$
$$+ \int_0^T \int_\Omega k\alpha(s)\nabla s \cdot \nabla \varphi \,\mathrm{d}t\,\mathrm{d}x - \int_0^T \int_{\partial\Omega} h\varphi \,\mathrm{d}t\,\mathrm{d}x = 0, \tag{7}$$

$$\int_0^T \int_\Omega \lambda(s)\nabla p \cdot \nabla \varphi \,\mathrm{d}t\,\mathrm{d}x - \int_0^T \int_{\partial\Omega} \pi\varphi \, dt\,\mathrm{d}x = 0. \tag{8}$$

4 Numerical Experiment

In this section, we present a numerical simulation of model (1) in a two dimensional space and show that the finite volume scheme (2)–(4) is effective in computing approximate solutions of the Brinkman regularization and showing the effect of Brinkman regularization on the flow through porous media. To do this, we use the Newton algorithm to approach the solution of the nonlinear system defined by the equations (2)–(3). This algorithm is coupled with a bigradient method to solve linear systems arising from the Newton algorithm process. We simulate the model in a two dimensional normalized domain $\Omega = (0,1) \times (0,1)$ and we consider a nonuniform admissible grid (see Fig. 1). We take the following data for the numerical test : porosity $\phi = 0.7$, permeability $k = 0.15 \times 10^{-8} m^2$, $\mu_1 = 10^{-3}\, Pa \cdot s$ (water viscosity) and $\mu_2 = 9 \times 10^{-5}\, Pa \cdot s$ (gas viscosity). Further,

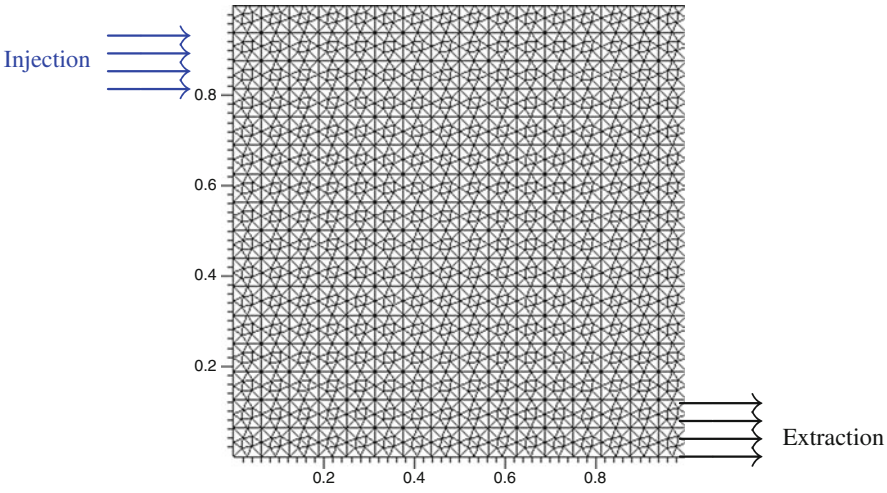


Fig. 1 Admissible mesh with 3584 triangles

we consider a small time step $\Delta t = 0.0001$ and we neglecting capillary pressure, i.e, $p_c = 0$.

In this test, we inject the water from the left up corner in order to extract from the right down corner as we see in Fig. 1. Here we are interested in the evolution of the water saturation versus time and the influence of Brinkman term. For that we take several values of μ and we compare them when $\mu = 0$ (Darcy flow). We observe from Figs. 2 and 3 that the Brinkman term adds diffusing effects by regularizing the

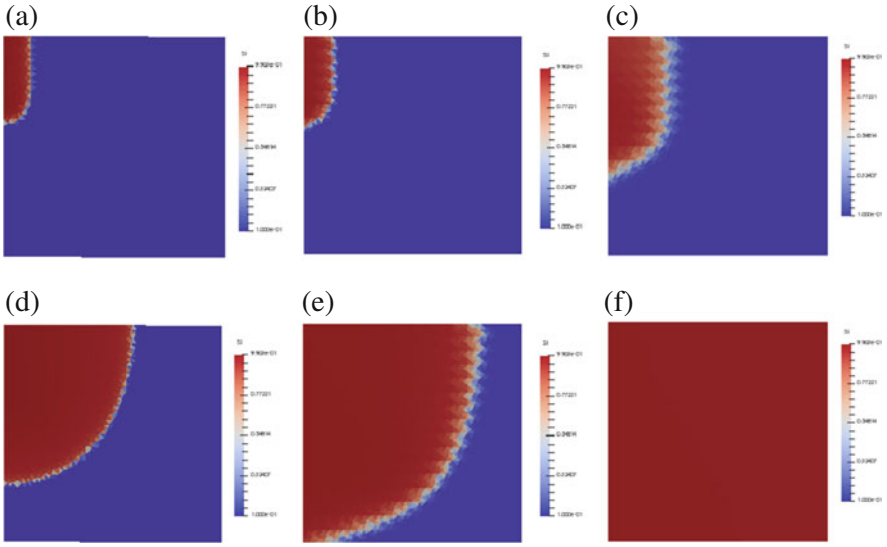


Fig. 2 Evolution of the water saturation versus time in the space domain Ω. We compare the evolution of Darcy flow $\mu = 0$, Brinkmann-Darcy $\mu = 10^{-6}$ and $\mu = 10^{-5}$. (**a**)Darcy at $t = 0.025$s. (**b**)Brinkman at $t = 0.025$s with $\mu = 10^{-6}$. (**c**) Brinkman at $t = 0.025$s with $\mu = 10^{-5}$. (**d**) Darcy at $t = 0.4$s. (**e**) Brinkman at $t = 0.4$s with $\mu = 10^{-6}$. (**f**) Brinkman at $t = 0.4$s with $\mu = 10^{-5}$

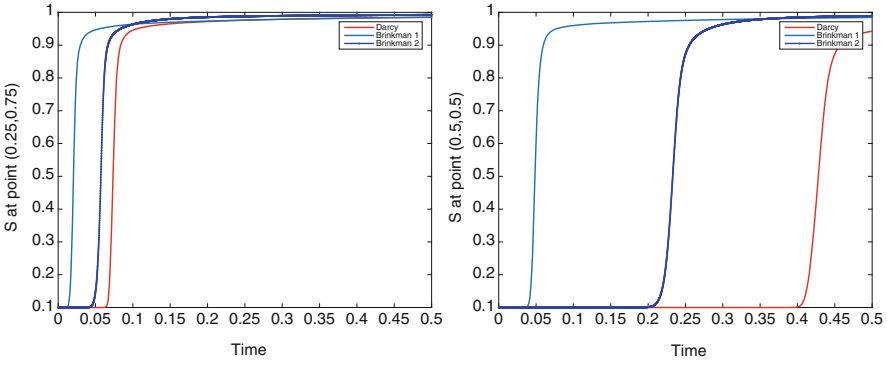


Fig. 3 Evolution of the water saturation versus time at point $(0.25, 0.75)$ (*left*), at $(0.5, 0.5)$ (*right*). We compare the evolution of Darcy flow $\mu = 0$, Brinkman-Darcy $\mu = 10^{-6}$ and $\mu = 10^{-5}$

interface between the two phases and the speed of this front increases with respect to the Brinkman viscosity.

References

1. Auriault, J.L., Geindreau, C., Boutin, C.: Filtration law in porous media with poor separation of scales. Transp. Porous. Med. **60**, 89–108 (2005)
2. Aziz, A., Settari, K.: Treatment of nonlinear terms in the numerical solution of partial differential equations for multiphase flow in porous media. Int. J. Multiphase Flow **1**, 817–844 (1975)
3. Brenner, K., Cancès, C., Hilhorst, D.: Finite volume approximation for an immiscible two-phase flow in porous media with discontinuous capillary pressure. Comput. Geosci. **17**, 573–597 (2013)
4. Chavent, G., Jaffré, J.: Mathematical Models and Finite Elements for Reservoir Simulation: Single Phase, Multiphase and Multicomponent Flows Through Porous Media. Elsevier, Amsterdam (1986)
5. Chen, Z., Ewing, R.E., Espedal, M.: Multiphase flow simulation with various boundary conditions. Comput. Methods Water Resour. 925–932 (1994)
6. Coclite, G.M., Mishra, S., Risebro, N.H.: Analysis and numerical approximation of Brinkman regularization of two-phase flows in porous media. Comput. Geosci. **18**, 637–659 (2014)
7. Eymard, R., Herbin, R., Michel, A.: Mathematical study of a petroleum-engineering scheme. ESAIM-Math. Model. Numer. Anal. **37**, 937–972 (2003)
8. Hannukainen, A., Juntunen, M., Stenberg, R.: Computations with finite element methods for the Brinkman problem. Comput. Geosci. **15**, 155–166 (2011)
9. Misiats, O., Lipnikov, K.: Second-order accurate monotone finite volume scheme for Richards equation. J. Comput. Phys. **239**, 123–137 (2013)
10. Paloka, E., Pazanin, I., Marusic, S.: Comparison between Darcy and Brinkman laws in a fracture. Appl. Math. Comput. **14**, 7538–7545 (2012)
11. Salinger, A.G., Aris, R., Derby, J.: Finite element formulations for large scale, coupled flows in adjacent porous and open fluid domains. Int. J. Numer. Methods Fluids **18**, 1185–1209 (1994)

Optimization and Sensitivity Analysis of Trajectories for Autonomous Small Celestial Body Operations

Anne Schattel, Andreas Cobus, Mitja Echim, and Christof Büskens

Abstract Within this paper, a method for on-board trajectory calculation in the vicinity of a small celestial body is introduced. Therefore, high precision methods of nonlinear optimization and optimal control are used. Additionally, a parametric sensitivity analysis is implemented. This tool allows to approximate a perturbed optimal solution in case of model parameter deviations from nominal values without noticeable computational effort. Parametric sensitivity analysis is a recent research area of great interest. Parameter perturbations that occur in the dynamic of the system as well as in boundary conditions or in state and control constraints can be analyzed. Thus, additional stability information is provided. Furthermore, the fast and reliable approximation of perturbed controls can be used for real-time control in time critical navigation phases.

1 Introduction

The number of space missions designed for the investigation of small celestial bodies is increasing. They provide the opportunity to obtain detailed knowledge about the composition of asteroids and comets and thereby to learn more about the evolution of our Solar System as well as to find rare earth elements. Due to the large spatial distances and signal transmission times, extensive deep space missions require autonomous spacecraft systems [8]. Within this paper, the use of high precision methods of nonlinear optimization and optimal control for realizing on-board trajectory calculations is introduced. Subsequently, an on-board capable parametric sensitivity analysis is implemented for the approximation of the perturbed optimal solution in case of deviations in parameters like the output of the thrusters. For the evaluation of our approach, near-asteroid orbit maneuvers are investigated considering different priorities of energy cost and flight time.

A. Schattel (✉) • A. Cobus • M. Echim • C. Büskens
Center for Industrial Mathematics, University of Bremen, Bibliothekstr. 1, 28359 Bremen, Germany
e-mail: ascha@math.uni-bremen.de; acobus@math.uni-bremen.de; mitja@math.uni-bremen.de; bueskens@math.uni-bremen.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_105

2 Optimization and Optimal Control

First, we need to introduce the formulation of a parametric optimal control problem OCP($\mathbf{p}$) [3]. We define

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & \mathbf{F}(\mathbf{x},\mathbf{u},\mathbf{p}) := \mathbf{g}(\mathbf{x}(t_\mathrm{f})) + \int_{t_0}^{t_\mathrm{f}} \mathbf{f}_0(\mathbf{x}(t),\mathbf{u}(t),\mathbf{p})dt \\ s.t. \quad & \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t),\mathbf{u}(t),\mathbf{p}), \\ & \mathbf{x}(t_0) = \mathbf{x}_0, \\ & \boldsymbol{\Psi}(\mathbf{x}_0,\mathbf{x}(t_\mathrm{f}),\mathbf{p}) = 0, \\ & \mathbf{C}(\mathbf{x}(t),\mathbf{u}(t),\mathbf{p}) \leq 0, \end{aligned} \tag{1}$$

with state $\mathbf{x} : [t_0, t_\mathrm{f}] \to \mathbb{R}^n$, control $\mathbf{u} : [t_0, t_\mathrm{f}] \to \mathbb{R}^m$, dynamic system $\mathbf{f} : \mathbb{R}^{n+m} \times P \to \mathbb{R}^n$, time $t \in [t_0, t_\mathrm{f}]$, data perturbations modeled by a parameter $\mathbf{p} \in P = \mathbb{R}^q$, boundary conditions $\boldsymbol{\Psi} : \mathbb{R}^{n+n} \times P \to \mathbb{R}^r$, and side constraints $\mathbf{C} : \mathbb{R}^{n+m} \times P \to \mathbb{R}^k$. According to [1], an optimal control problem can in general be solved using indirect or direct methods. Within this paper, the latter technique is used, that transcribes the OCP($\mathbf{p}$) into a finite-dimensional parametric nonlinear optimization problem NLP($\mathbf{p}$). For a chosen number of grid points $t_0 = \tau_1 < \ldots < \tau_N = t_\mathrm{f}$, let $\mathbf{x}^i \approx \mathbf{x}(\tau_i)$, $\mathbf{u}^i \approx \mathbf{u}(\tau_i)$ be. Additionally, we define the optimization vector $\mathbf{z} := (\mathbf{x}^1, \ldots, \mathbf{x}^N, \mathbf{u}^1, \ldots, \mathbf{u}^N, t_\mathrm{f})$ and let $\mathbf{G}(\mathbf{z},\mathbf{p}) = (G_1(\mathbf{z},\mathbf{p}), \ldots, G_L(\mathbf{z},\mathbf{p}))$ denote the collection of functions defining the according equality and inequality constraints. The NLP($\mathbf{p}$) then is defined as

$$\begin{aligned} \min_{\mathbf{z}} \quad & \mathbf{F}(\mathbf{z},\mathbf{p}) \\ s.t. \quad & G_i(\mathbf{z},\mathbf{p}) = 0,\ i = 1,\ldots,L_e, \\ & G_i(\mathbf{z},\mathbf{p}) \leq 0,\ i = L_e+1,\ldots,L. \end{aligned} \tag{2}$$

We let the final time t_f free, hence it is used as an additional optimization variable in $\mathbf{z}$. The resulting NLP($\mathbf{p}$) is solved using the NLP solver WORHP [4], which is complemented by the transcriptor TransWORHP [7] and by the sensitivity analysis module WORHP ZEN [9]. Here, we use a multiple shooting method for transcription instead of the common and highly robust full discretization [3], since it provides high accuracy in the calculated solutions [1, 10].

To investigate the behavior of the solution of an optimal control problem with respect to model parameters $\mathbf{p}$, *parametric sensitivity analysis* (PSA) can be used. Parameter perturbations can occur within the dynamics, boundary conditions, or in state and control constraints. For a nominal value $\mathbf{p} = \mathbf{p}_0$, the problem OCP($\mathbf{p}_0$) is called the nominal or unperturbed problem. Fiacco [5] showed that the unperturbed optimal solution $\mathbf{z}(\mathbf{p}_0)$ can be embedded into a family of perturbed optimal C^1-functions $\mathbf{z}(\mathbf{p})$ with respect to $\mathbf{p}_0$. Moreover, he defined how to calculate its

parametric derivatives $\frac{d\mathbf{z}}{d\mathbf{p}}(\mathbf{p}_0)$ by the implicit function theorem without noticeable computational effort. Evaluating the first-order Taylor expansion

$$\mathbf{z}(\mathbf{p}) \approx \mathbf{z}(\mathbf{p}_0) + \frac{d\mathbf{z}}{d\mathbf{p}}(\mathbf{p}_0)\Delta\mathbf{p}, \qquad \Delta\mathbf{p} = \mathbf{p} - \mathbf{p}_0 \tag{3}$$

gives a very fast on-line approximation of the perturbed solution. This way, the sensitivity derivatives provide stability information about the optimal solution. Additionally, they can be used efficiently for real-time control approximations of perturbed solutions. Further details can be found in [3].

3 Modeling

Minor celestial bodies like asteroids have a small gravitational pull. This presents a great challenge for traditional orbiting, but on the other hand gives spacecrafts a substantial scope to implement active control within their vicinity [2]. The objective function $\mathbf{F}$ of (1) is set up to consider the competitive mission aims of short flight time t_f and low energy consumption. The latter depends on the applied thrust

$$\mathbf{T} := (u_1(t), u_2(t), u_3(t))^T, \qquad t \in [t_0, t_\mathrm{f}] \tag{4}$$

in Newton. We then define

$$\mathbf{F} = t_\mathrm{f} w + \int_{t_0}^{t_\mathrm{f}} \left(u_1^2(t) + u_2^2(t) + u_3^2(t)\right) dt(1-w) \tag{5}$$

with a weighting factor $w \in (0, 1]$. The value $w = 0.0$ is excluded, since otherwise the optimal solution would be allowed to take an infinite flight time. Trajectories are calculated at a distance of about 5–20 km above the surface of the asteroid, which is the typical range to perform a detailed characterization [6]. Therefor, we define a respective ellipsoidal security zone around the asteroid, which is included within the side constraints $\mathbf{C}$. The dynamic system $\mathbf{f}$ is defined as

$$\dot{\mathbf{x}} := \begin{pmatrix} \dot{\mathbf{p}}_\mathrm{sc} \\ \ddot{\mathbf{p}}_\mathrm{sc} \\ \dot{m}_\mathrm{sc} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_\mathrm{sc} \\ \mathbf{g}_\mathrm{a} + \mathbf{T} m_\mathrm{sc}^{-1} + \boldsymbol{\varphi} \\ -\|\mathbf{T}\|(g_0 I_\mathrm{sp})^{-1} \end{pmatrix}.$$

Herein, $\mathbf{p}_\mathrm{sc} \in \mathbb{R}^3$ is the asteroid-centered position vector of the spacecraft, $\mathbf{v}_\mathrm{sc} \in \mathbb{R}^3$ its velocity vector, and $m_\mathrm{sc} \in \mathbb{R}$ is its mass. The gravitational influence of the asteroid is given as a point mass $\mathbf{g}_\mathrm{a} = \mu_\mathrm{ast}\mathbf{p}_\mathrm{sc}\|\mathbf{p}_\mathrm{sc}\|^{-3} \in \mathbb{R}^3$ with the standard gravitational parameter $\mu_\mathrm{ast} \in \mathbb{R}$ of the asteroid, $\mathbf{r} := -\mathbf{p}_\mathrm{sc}\|\mathbf{p}_\mathrm{sc}\|^{-1} \in \mathbb{R}^3$ is the spacecraft-asteroid direction, $\boldsymbol{\varphi} \in \mathbb{R}^3$ the approximated acceleration due to radiation pressure, $g_0 = 9.80665\, m\, s^{-2}$ the gravitational acceleration, and $I_\mathrm{sp} \in \mathbb{R}$ the specific impulse (Isp) of the spacecraft in seconds.

4 Numerical Results

A model of the asteroid Itokawa and a spacecraft propulsion system with 12×1 N thrusters and an Isp of 220 s is used. The spacecraft's total mass is 2400 kg, containing 400 kg of fuel. The ellipsoidal security zone is set to 5.7 km×4 km×4 km. We choose a total number of 61 discrete time points and 21 multinodes. This yields a problem with 331 optimization variables, 662 box constraints, and 213 equality constraints. First, we vary the weighting factor w within the objective function in (5) to calculate the Pareto front regarding the flight time t_f and energy demand. The latter is considered in terms of fuel consumption and therefore mass loss Δm_sc, which is given by

$$\Delta m_\mathrm{sc} = \frac{1}{g_0 I_\mathrm{sp}} \int_{t_0}^{t_\mathrm{f}} \sqrt{u_1^2(t) + u_2^2(t) + u_3^2(t)}\, dt. \tag{6}$$

This is followed by a sensitivity analysis regarding the efficiency of the thrusters and concluded by the analysis of the impact of the size of the Isp.

Initially, we investigate solution trajectories for *differently weighted objective priorities*. In Fig. 1, the solutions for $w = 0.001$ and $w = 1.0$ for flying halfway around asteroid Itokawa are shown. Whereas the first solution needs $\Delta m_\mathrm{sc} = 2.7$ kg of fuel and a flight time of $t_\mathrm{f} = 565{:}53$ min:s, the latter yields $\Delta m_\mathrm{sc} = 12.9$ kg and a total flight time of $t_\mathrm{f} = 99{:}42$ min:s. Thus, the solution for $w = 0.001$ needs 10.2 kg less fuel, but 466:11 min:s longer than the solution for $w = 1.0$. The higher Δm_sc for $w = 1.0$ is due to the fact that the inclination of the trajectory changes stronger in order to meet the boundary conditions faster.

An iterative calculation of solutions for an increasing weighting factor $w \in (0, 1]$ can be used to calculate the *Pareto front* according to the problem objective in (5). In Fig. 2, the solutions for 100 values of w, equidistantly distributed over the interval $(0, 1]$, are plotted. The first two plots show the dependence of t_f and Δm_sc on w respectively, while the third plot shows the Pareto front. To speed up calculations, we use sensitivity derivatives introduced in (3) with respect to perturbations in

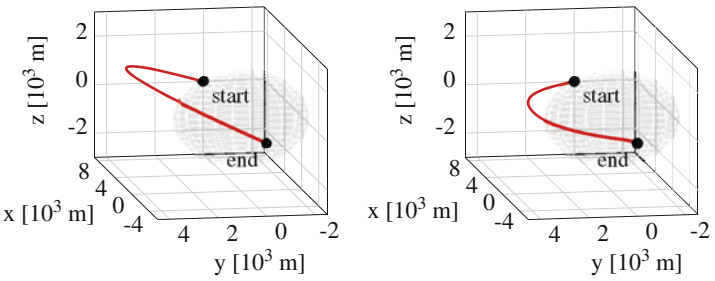


Fig. 1 Optimal trajectories for flying halfway around Itokawa with different weighting factors $w = 0.001$ (*left*) and $w = 1.0$ (*right*)

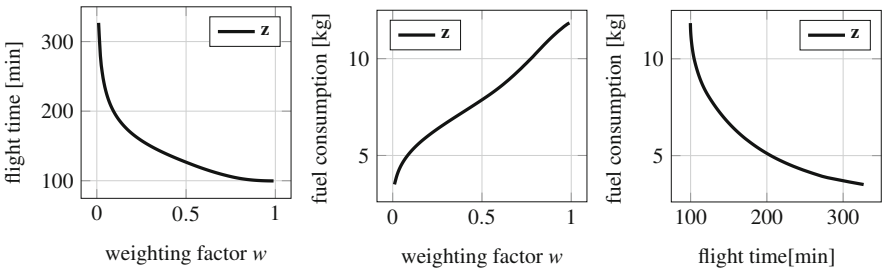


Fig. 2 Dependence of flight time t_f on w, dependence of fuel consumption Δm_{sc} on w, and according Pareto front for 100 values of w, equidistantly distributed over the interval $(0, 1]$

the parameter w to generate an initial guess for the optimization variables and parameters. While thus only 461 s are needed to calculate the whole Pareto front, using the preceding optimization result as initial guess instead would lead to a calculation time of 513 s, and using a trivial initial guess even takes the calculation process 1425 s, while the problem size is still very small. The results were obtained on a 2.9 GHz Intel(R) Xeon(R) CPU E5–4617 0 of a computer with 512 GB RAM using single threading.

In a next step, we perform a *sensitivity analysis of the thruster performance*. An error-proneness lies in the small thrusters used during proximity operations. Thrusters may not work properly or the actuators may not implement the thrust commands exactly. In Fig. 3, a 10% thruster perturbation in the velocity-direction of the spacecraft is examined. On the left, the deviations in position, velocity, and thrust between the nominal and newly calculated perturbed optimal solution are plotted. On the right, the respective deviations between the perturbed optimal solution and the via PSA approximated solution can be seen. While there is a relatively strong deviation in the former solutions, Fig. 3 shows that by using PSA the perturbed solution is approximated very accurately. This enables a stability analysis of the nominal solution as well as an efficient use of the approximated perturbed controls for real-time optimal control, especially in time critical navigation phases.

Furthermore, we can use *parametric sensitivity analysis for spacecraft design*. During trajectory optimization, PSA can be used to analyze the influence of certain model parameters like the Isp of the spacecraft. This parameter is a measure of how efficiently the propellant is used for creating forward thrust. As can be seen in Fig. 4, when decreasing the nominal Isp of 220 s to 200 s and further to 170 s, Δm_{sc} increases linearly, first with 9.08% and further with 22.7%, while t_f stays approximately the same. When increasing the Isp to 240 s and further to 270 s, Δm_{sc} decreases by 9.08% and 22.7% respectively. This way, decisions over the spacecraft design can be finalized and technological progress can be investigated on-the-fly.

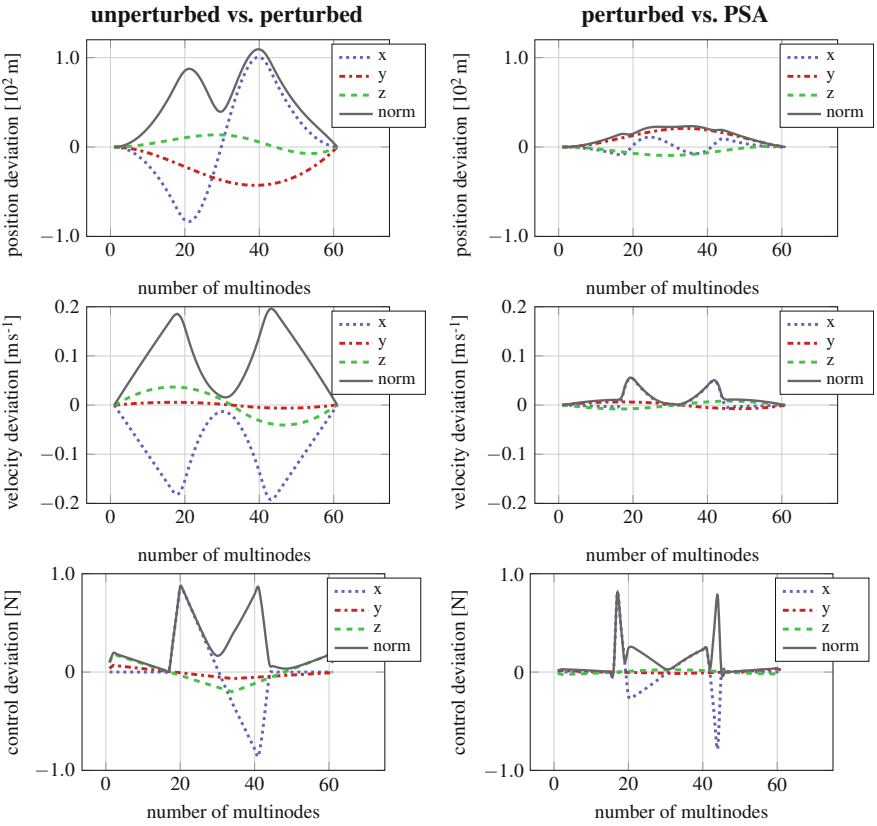


Fig. 3 Impact of 10% thruster perturbation in velocity-direction. On the *left*, *top to bottom*, deviations between unperturbed and perturbed solution in position, velocity, and thrust, on the *right*, deviations between the perturbed solution and the approximation using PSA accordingly

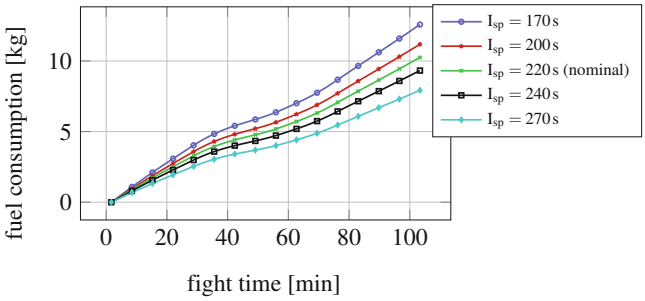


Fig. 4 A linear increase in I_{sp} leads to a linear decrease in Δm_{sc}, and vice versa

5 Conclusion

In this paper, we presented trajectory optimization methods for the vicinity of a small celestial body. Further, we showed how parametric sensitivity analysis can be used for an on-line approximation of perturbed states and controls, where the latter is especially interesting for real-time optimal control in time critical navigation phases. All these results enable autonomous decision making and Fault-Detection, Fault-Isolation, and Recovery Techniques (FDIR) regarding orbit maneuvers, as a range of diverse trajectories for different objectives including stability information of nominal solutions is provided. By changing the dynamics and model properties, the developed algorithms can easily be adapted for applications in autonomous driving and deep sea navigation. A next step will be the performance of actual real-time control on the basis of PSA during the landing phase of a spacecraft on an asteroid.

Acknowledgements This work was supported by the German Aerospace Center (DLR) with financial means of the German Federal Ministry for Economic Affairs and Energy (BMWi), project "KaNaRiA" (grant No. 50 NA 1318).

References

1. Betts, J.T.: Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, 2nd edn. SIAM, Philadelphia (2010)
2. Broschart, S., Scheeres, D.: Control of hovering spacecraft near small bodies: application to asteroid 25143 Itokawa. J. Guid. Control. Dyn. **28**(2), 343–354 (2005)
3. Büskens, C., Maurer, H.: SQP-methods for solving optimal control problems with control and state constraints: adjoint variables, sensitivity analysis and real-time control. J. Comput. Appl. Math. **120**(1), 85–108 (2000). doi:10.1016/S0377-0427(00)00305-8
4. Büskens, C., Wassel, D.: The ESA NLP solver WORHP. In: Fasano, G., Pintr, J.D. (eds.) Modeling and Optimization in Space Engineering. Springer Optimization and Its Applications, vol. 73, pp. 85–110. Springer, New York (2013). doi:10.1007/978-1-4614-4469-5_4
5. Fiacco, A.V.: Introduction to sensitivity and stability analysis in nonlinear programming. In: Mathematics in Science and Engineering, vol. 165. Academic Press, New York (1983)
6. Hussmann, H., Oberst, J., Wickhusen, K., Shi, X., Damme, F., Lüdicke, F., Lupovka, V., Bauer, S.: Stability and evolution of orbits around the binary asteroid 175706 (1996 FG3): implications for the MarcoPolo-R mission. Planet. Space Sci. **70**(1), 102–113 (2012)
7. Knauer, M., Büskens, C.: From WORHP to TransWORHP. In: 5th International Conference on Astrodynamics Tools and Techniques (2012)
8. Probst, A., González Peytaví, G., Nakath, D., Schattel, A., Rachuy, C., Lange, P., Clemens, J., Echim, M., Schwarting, V., Srinivas, A., Gadzicki, K., Förstner, R., Eissfeller, B., Schill, K., Büskens, C., Zachmann, G.: KaNaRiA: identifying the challenges for cognitive autonomous navigation and guidance for missions to small planetary bodies. In: 66th International Astronautical Congress (IAC), pp. IAC-15-A3.IP.15-1–IAC-15-C3.IP.15-14. International Astronautical Federation (2015)
9. Schäfer, R.: Parametrische Sensitivitätsanalyse. Bachelor thesis, Center for Industrial Mathematics, University of Bremen, Germany (2012)
10. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. In: Texts in Applied Mathematics, vol. 12, 3rd edn. Springer Science & Business Media, New York (2002)

A Finite Volume Method with Staggered Grid on Time-Dependent Domains for Viscous Fiber Spinning

Stefan Schiessl, Nicole Marheineke, Walter Arne, and Raimund Wegener

Abstract The spinning of slender viscous fibers can be described by the special Cosserat theory with one-dimensional models consisting of partial differential and algebraic equations on time-dependent spatial domains. We propose a first-order finite volume method on a staggered grid with flux approximation suitable for the underlying differential-algebraic characteristics and a proper geometric handling of the space-time domain. The numerical results confirm the theoretical convergence orders. As exemplary application a rotational spinning scenario is studied.

1 Modeling of Viscous Fiber Spinning

Fiber Model A fiber is a long and thin body. Its three-dimensional description can be reduced to a one-dimensional Cosserat rod model by cross-sectional averaging, for details see [1, 4]. The constitutive elements are a curve specifying the position and an orthonormal director triad for the oriented cross-sections. The kinematics of the rod is combined with balance laws for mass, linear and angular momentum as well as physically reasonable geometric and material laws.

In this work we consider an incompressible viscous fiber growing over time. For that purpose we use a Lagrangian parametrization. The space-time domain is given by $\mathscr{Q} = \{(s,t) \in \mathbb{R}^2 | s \in [-\mathscr{L}(t), -\mathscr{R}(t)], t \in [0,T]\}$ where s denotes the material parameter and t the time. We assume that the given functions $\mathscr{L}, \mathscr{R} : [0,T] \to \mathbb{R}$ satisfy $\mathscr{L}'(t) > 0$ (inflow) and $\mathscr{R}'(t) \geq 0$ (outflow or free end) for all $t \in [0,T]$. To handle the numerical challenges coming from the time-dependent domain we propose a special adjusted formulation of the kinematics. The respective

S. Schiessl • N. Marheineke (✉)
FAU Erlangen-Nürnberg, Lehrstuhl Angewandte Mathematik 1, Cauerstr. 11, D-91058 Erlangen, Germany
e-mail: stefan.schiessl@math.fau.de; marheineke@math.fau.de

W. Arne • R. Wegener
Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Fraunhofer Platz 1, D-67663 Kaiserslautern, Germany
e-mail: walter.arne@itwm.fraunhofer.de; raimund.wegener@itwm.fraunhofer.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_106

dimensionless fiber model stated in the director basis is given by

$$\partial_t \bar{\boldsymbol{r}} = \boldsymbol{R}^T(\mathbf{q}) \cdot \boldsymbol{v} + \partial_s \bar{\boldsymbol{\lambda}}_{\boldsymbol{\tau}}, \qquad \partial_t \mathbf{q} = \mathbf{A}(\boldsymbol{\omega} + \partial_s \boldsymbol{\lambda}_{\boldsymbol{\kappa}}) \cdot \mathbf{q} + \lambda_t \mathbf{q}, \tag{1a}$$

$$\partial_s \bar{\boldsymbol{r}} = \boldsymbol{R}^T(\mathbf{q}) \cdot \boldsymbol{\tau}, \qquad \partial_s \mathbf{q} = \mathbf{A}(\boldsymbol{\kappa}) \cdot \mathbf{q} + \lambda_s \mathbf{q}, \tag{1b}$$

$$\partial_t \boldsymbol{\tau} = \partial_s \boldsymbol{v} + \boldsymbol{\kappa} \times \boldsymbol{v} + \boldsymbol{\tau} \times \boldsymbol{\omega} + C\boldsymbol{R}(\mathbf{q}) \cdot \bar{\boldsymbol{\lambda}}_{\boldsymbol{\tau}}, \tag{1c}$$

$$\partial_t \boldsymbol{\kappa} = \partial_s \boldsymbol{\omega} + \boldsymbol{\kappa} \times \boldsymbol{\omega} + C\boldsymbol{\lambda}_{\boldsymbol{\kappa}}, \tag{1d}$$

$$\mathbf{q} \cdot \mathbf{q} = 1, \tag{1e}$$

$$\partial_t \boldsymbol{p} = \boldsymbol{p} \times \boldsymbol{\omega} + \partial_s \boldsymbol{n} + \boldsymbol{\kappa} \times \boldsymbol{n} + \boldsymbol{f}, \tag{1f}$$

$$\varepsilon^2 \partial_t \boldsymbol{h} = \varepsilon^2 \boldsymbol{h} \times \boldsymbol{\omega} + \partial_s \boldsymbol{m} + \boldsymbol{\kappa} \times \boldsymbol{m} + \boldsymbol{\tau} \times \boldsymbol{n} + \varepsilon^2 \boldsymbol{l}, \tag{1g}$$

$$\partial_t \tau_3 = \frac{\mathrm{Re}}{3\mu} \frac{\tau_3^2}{\sigma_V} n_3, \qquad \tau_1 = 0, \quad \tau_2 = 0, \tag{1h}$$

$$\varepsilon^2 \partial_t \boldsymbol{\kappa} = \frac{\mathrm{Re}}{3\mu} \frac{\tau_3^3}{\sigma_V^2} \boldsymbol{L}_{2/3}^{-1} \cdot \boldsymbol{m}. \tag{1i}$$

For certain model variables a representation in an outer basis is more convenient, these component tuples are indicated here by $\bar{\cdot}$, see, e.g., the fiber curve $\bar{\boldsymbol{r}} : \mathscr{Q} \to \mathbb{R}^3$. The transformation between the representations in director and outer bases is described by the rotation matrix $\boldsymbol{R} \in SO(3)$ that we parameterize with unit quaternions $\mathbf{q} : \mathscr{Q} \to \mathbb{R}^4$, i.e., $z = \boldsymbol{R}(\mathbf{q}) \cdot \bar{z}$, cf. [1].

In a rod model [1] the temporal derivatives of the fiber curve and the director triad (quaternions) induce the linear and angular velocities $\boldsymbol{v}$ and $\boldsymbol{\omega}$ with skew-symmetric matrix $\mathbf{A}$, the spatial derivatives the tangent $\boldsymbol{\tau}$ and the generalized curvature $\boldsymbol{\kappa}$. The last ones can be interpreted as distortion measures. In (1) we use the kinematic relations for $\bar{\boldsymbol{r}}$ and $\mathbf{q}$ stated in (1a)–(1b) with their respectively implied compatibility conditions (1c)–(1d). The unit quaternion constraint (1e) is necessary when dealing with non-isometric numerical realizations. To avoid an overdetermined system we incorporate projection corrections by introducing respective Lagrange multipliers $\lambda_s, \lambda_t, \bar{\boldsymbol{\lambda}}_{\boldsymbol{\tau}}, \boldsymbol{\lambda}_{\boldsymbol{\kappa}}$ and a constant $C \ll 1$ (see Remark 1).

The dynamics of the rod is driven by external loads $\boldsymbol{f}$ and $\boldsymbol{l}$ and inner forces $\boldsymbol{n}$ and couples $\boldsymbol{m}$. In the balance laws (1f)–(1g) the linear and angular momentum line densities are given by $\boldsymbol{p} = \sigma_M \boldsymbol{v}$ and $\boldsymbol{h} = \boldsymbol{J}_M \cdot \boldsymbol{\omega}$. Considering an incompressible viscous fiber with homogeneous circular cross-sections we particular have $\boldsymbol{J}_M = (\sigma_M \sigma_V / \tau_3)\boldsymbol{L}_2$, $\boldsymbol{L}_x = \mathrm{diag}(1, 1, x)/4\pi$ with constant mass line density σ_M, volume line density σ_V and viscosity μ, all three values are chosen here to be one. The fiber behavior is particularly characterized by the dimensionless Reynolds number Re and the slenderness ratio ε. Note that the stated formulation allows the replacement of the material laws (1h)–(1i) without affecting the remaining equations.

Remark 1 (Motivation for New Kinematic Formulation) The partial differential-algebraic system (1) becomes a system of differential-algebraic equations (DAEs)

after semi-discretization (in time or space). When thinking about an appropriate numerical scheme we have to take into account the (differentiation) index of the DAEs. The original formulation of the kinematics, (1a)–(1b) with $\bar{\boldsymbol{\lambda}}_\tau = \boldsymbol{\lambda}_\kappa \equiv \mathbf{0}$, yields index three in time and one in space. The proposed formulation (1a)–(1d) reduces the index in time to two, while avoiding drift-off effects. It is inspired by index-reduction and stabilization techniques for DAEs, [2]. Note that a index-reduction would already be achieved by using (1a)–(1c) with $\boldsymbol{\lambda}_\kappa \equiv \mathbf{0}$, but the presence of both compatibility conditions turns out to be beneficial for our numerics.

Set-Up and Scenario In the following we focus exemplarily on a set-up where a fiber is extruded through a nozzle at a fixed position for $s = -\mathscr{L}(t)$ and has a tension-free end that moves freely in the surrounding medium for $s = -\mathscr{R}(t)$, see Fig. 1. We particularly consider the rotational spinning scenario of [1] where the nozzle is located on the curved face of a cylindrical drum that rotates around its symmetry axis. Introducing an outer basis that rotates with the drum results in a time-independent nozzle position and linear and angular velocities adapted to the rotational frequency. Thus, the external loads cover Coriolis and centrifugal effects—apart from gravity. They are characterized by the dimensionless Rossby number Rb and the Froude number Fr and given by

$$\begin{aligned}
\boldsymbol{f} &= \boldsymbol{f}_g + \boldsymbol{f}_r, \quad \boldsymbol{f}_g = -\mathrm{Fr}^{-2}\sigma_M \boldsymbol{R}\cdot\boldsymbol{e}_1,\\
&\qquad \boldsymbol{f}_r = -2\mathrm{Rb}^{-1}\sigma_M(\boldsymbol{R}\cdot\boldsymbol{e}_1)\times\boldsymbol{v} - \mathrm{Rb}^{-2}\sigma_M\boldsymbol{R}\cdot(\boldsymbol{e}_1\times(\boldsymbol{e}_1\times\bar{\boldsymbol{r}})),\\
\boldsymbol{l} &= \mathrm{Rb}^{-1}\Big[\Big(\boldsymbol{J}_M\cdot(\boldsymbol{R}\cdot\boldsymbol{e}_1)\Big)\times\Big(\boldsymbol{\omega}+\mathrm{Rb}^{-1}(\boldsymbol{R}\cdot\boldsymbol{e}_1)\Big) + \Big(\boldsymbol{J}_M\cdot\boldsymbol{\omega}\Big)\times(\boldsymbol{R}\cdot\boldsymbol{e}_1)\\
&\qquad + \boldsymbol{J}_M\cdot\Big(\boldsymbol{\omega}\times(\boldsymbol{R}\cdot\boldsymbol{e}_1)\Big) - \partial_t\boldsymbol{J}_M\cdot(\boldsymbol{R}\cdot\boldsymbol{e}_1)\Big],
\end{aligned}$$

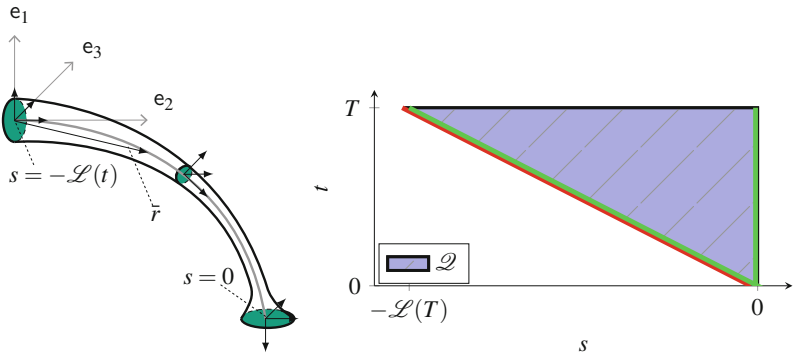


Fig. 1 *Left*: Sketch of a fiber for some time $t > 0$. *Right*: Space-time domain $\mathscr{Q}$ with initial (*red line*) and boundary conditions (*green lines*) at $s = -\mathscr{L}(t) = -t$ and $s = -\mathscr{R}(t) = 0$

with $\boldsymbol{R} = \boldsymbol{R}(\mathbf{q})$ and the canonical basis triples denoted by $\boldsymbol{e}_i$, $i = 1, 2, 3$. Starting with $\mathscr{L}(0) = \mathscr{R}(0) = 0$ the mixed initial-boundary conditions are set to be

$$\bar{\boldsymbol{r}}(-\mathscr{L}(t), t) = \boldsymbol{e}_2, \qquad \mathbf{q}(-\mathscr{L}(t), t) = (1, 1, 0, 0)/\sqrt{2}, \quad \boldsymbol{v}(-\mathscr{L}(t), t) = \mathscr{L}'(t)\boldsymbol{e}_3,$$
$$\boldsymbol{\tau}(-\mathscr{L}(t), t) = \boldsymbol{e}_3, \qquad (\boldsymbol{\kappa}, \boldsymbol{\omega})(-\mathscr{L}(t), t) = \mathbf{0}, \qquad (\boldsymbol{m}, \boldsymbol{n}, \bar{\boldsymbol{\lambda}}_\tau, \boldsymbol{\lambda}_\kappa)(-\mathscr{R}(t), t) = \mathbf{0}.$$

2 Numerical Scheme

The fiber model can be stated in the general form

$$\partial_t \boldsymbol{a}(\mathbf{y}(s,t)) + \partial_s \boldsymbol{f}(\mathbf{y}(s,t)) + \boldsymbol{g}(\mathbf{y}(s,t)) = \mathbf{0} \tag{2}$$

with the vector of unknowns $\mathbf{y} : \mathscr{Q} \to \mathbb{R}^M$ and the vector-valued functions $\boldsymbol{a}, \boldsymbol{f}, \boldsymbol{g} : \mathbb{R}^M \to \mathbb{R}^M$ for the time evolution, spatial fluxes and source terms, respectively.

Employing a finite volume scheme on a staggered spatial grid for the fiber model we consider a partition of the domain $\mathscr{Q}$ with finite space-time cells. For this purpose we introduce the equidistant spatial grid $s_{i+1/2}$, $i \in \mathbb{Z}$ with $s_{i+1/2} < s_{i+3/2}$ and $s_{1/2} = 0$, the spacing is Δs. Additionally, we set $s_i = (s_{i+1/2} + s_{i-1/2})/2$. The temporal grid is also equidistantly given by t_n, $n \in \mathbb{N}$ with $t_n < t_{n+1}$ and $t_0 = 0$, the spacing is Δt (cf. Fig. 2). All discrete space-time tuples enclosed by $\mathscr{Q}$ are collected in the set $\mathscr{Q}^h = \{(s_k, t_n) \in \mathscr{Q} \mid -\mathscr{L}(t_n) \le s_k \le -\mathscr{R}(t_n), t_n \le T, k \in \mathscr{I}, n \in \mathbb{N}\}$ with $\mathscr{I} = \{\ldots, -3/2, -1, -1/2, 0, 1/2, \ldots\}$. We call those tuples with $k \in \mathbb{Z}$ nodes and those with $k \notin \mathbb{Z}$ edges and formalize this choice by

$$Z(k) = \begin{cases} 1, & \text{if } k \in \mathbb{Z}, \quad \text{(node)} \\ 0, & \text{if } k \notin \mathbb{Z}. \quad \text{(edge)} \end{cases}$$

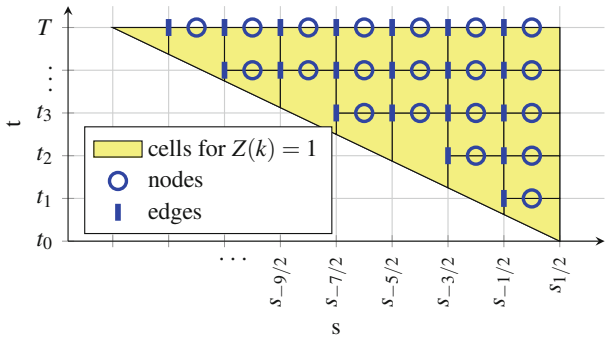


Fig. 2 Illustration of the domain partitioning with cells built around all active nodes (*blue circles*) used for node type equations (analogously for edge type), cf. Fig. 1

Aiming for natural narrow stencils of the spatial fluxes, we decide for every unknown whether it should be of node type or of edge type, implying the decision vector $\boldsymbol{D} = (D_1, \ldots, D_M) \in \{1, 0\}^M$. The associated model equation receives the same type.

To obtain a discrete version of (2), we integrate over the finite volumes that are constructed with help of the nodes and edges. We apply a central spatial flux approximation suitable for DAEs with index 1 and a fully implicit temporal flux approximation suitable for DAEs with index 2 that allows for iterative solving. This yields component-wise for $j = 1, .., M$ and every $(s_k, t_n) \in \mathscr{Q}^{h,A}$ with $D_j = Z(k)$

$$\begin{aligned} \Delta s\big[a_j(\mathbf{Y}_k(t_n)) + a_j(\mathbf{Y}_k(t_{n-1}^k))\big] + \Delta t_n^k \left[f_j(\mathbf{Y}_{k+1/2}(t_n)) - f_j(\mathbf{Y}_{k-1/2}(t_n))\right] \\ + \Delta s \Delta t_n^k g_j(\mathbf{Y}_k(t_n)) = \mathbf{0} \end{aligned} \tag{3}$$

where $\mathbf{Y}_k = (\mathsf{Y}_{k,1}, \ldots, \mathsf{Y}_{k,M}) : [0, T] \to \mathbb{R}^M$ are approximated by

$$\mathsf{Y}_{k,j}(t_n) = \begin{cases} \mathsf{Y}_{k,j}^n, & \text{if } D_j = Z(k), \\ (\mathsf{Y}_{k-1/2,j}^n + \mathsf{Y}_{k+1/2,j}^n)/2, & \text{otherwise} \end{cases}$$

with $\mathsf{Y}_{k,j}^n = \int_{s_{k-1/2}}^{s_{k+1/2}} \mathsf{y}_j(s, t_n)\, ds/\Delta s$. The time step in (3) is given by $\Delta t_n^k = t_n - t_{n-1}^k$ with $t_{n-1}^k = t_{n-1}$ if $(s_k, t_{n-1}) \in \mathscr{Q}^{h,A}$ and $t_{n-1}^k = -\mathscr{L}^{-1}(s_k)$ otherwise. We refer to

$$\begin{aligned} \mathscr{Q}^{h,A} = \big\{(s_k, t_n) \in \mathscr{Q}^h \big| \big[\big((s_{k-1/2}, t_n) \in \mathscr{Q}^h \wedge Z(k-1/2) = 1\big) \vee (s_{k-1}, t_n) \in \mathscr{Q}^h\big] \\ \wedge \big[\big((s_{k+1/2}, t_n) \in \mathscr{Q}^h \wedge Z(k+1/2) = 0\big) \vee (s_{k+1}, t_n) \in \mathscr{Q}^h\big], k \in \mathscr{I}, n \in \mathbb{N}\big\} \end{aligned}$$

as active domain (i.e., its elements are active nodes and edges, exemplarily indicated in blue in Fig. 2). For all other $(s_k, t) \in \mathscr{Q} \backslash \mathscr{Q}^{h,A}$ (non-active) that are required in the discrete formula the initial and boundary conditions are used. Initial values can be set directly due to the adjusted time step, whereas boundary conditions need to be interpolated to reflect the actual distances. The resulting non-linear system for the unknowns $\mathsf{Y}_{k,j}^n$ is solved for every time level t_n using Newton's method.

The proposed scheme is second order for the interior spatial fluxes and first order for the temporal fluxes, source terms and boundary cells due to an applied linear interpolation.

Remark 2 (Choice of Active Domain and Type Assignment) Note that $\mathscr{Q}^{h,A}$, or more precisely, Z for the boundaries has to be chosen appropriately with respect to the model problem. In the numerical approach we need to ensure that all boundary conditions can be coupled naturally through interpolation within the discrete fluxes of (3). For the fiber model at hand we apply the unknowns to equations and type assignment as follows listed in the style [unknowns; equations; type]: $[\bar{\boldsymbol{r}}, \mathbf{q}$; (1a); 1], $[\bar{\boldsymbol{\lambda}}_{\boldsymbol{\tau}}, \boldsymbol{\lambda}_{\boldsymbol{\kappa}}, \lambda_s$; (1b); 0], $[\boldsymbol{\tau}$; (1c); 0], $[\boldsymbol{\kappa}$; (1d); 0], $[\lambda_t$; (1e); 1], $[\boldsymbol{v}$; (1f); 1], $[\boldsymbol{\omega}$; (1g); 1], $[\boldsymbol{n}$; (1h); 0] and $[\boldsymbol{m}$; (1i); 0]. This requires knowledge of the distortion measures at the boundary $s = -\mathscr{R}(t)$. We particularly use here linear extrapolation.

3 Numerical Results

Convergence Properties In [1] a finite volume scheme was developed for the fiber model with index-reduced, but not stabilized kinematic formulation and applied to the rotational spinning scenario. It showed a worrisome temporal order reduction effect that rendered approximations for higher convergence orders useless. Regarding the underlying DAE characteristics and handling the dependencies of initial and boundary conditions appropriately, our new scheme overcomes the hitherto existing restrictions. The numerical investigations show spatial, temporal and combined convergence orders that confirm the theoretically expected results, see Fig. 3. For the spatial convergence we even obtain better results due to the dominating spatial fluxes. This is attributed to the improved approximation quality of the scheme—in particular of the space-time domain—and the stabilized formulation of the kinematics. Our investigations suggest that an overall higher order accuracy could be achieved with suitable approximations of the fluxes and source terms.

Fiber Spinning Simulations In spinning processes predictions of the fiber behavior close to the nozzle and the achievable thickness are of special interest. Since respective simulation studies leave freedom in choosing the boundary conditions for the tension-free end, we explore here two cases: A—without mass outflow ($\mathscr{R}(t) = 0$) and B—with mass outflow ($\mathscr{R}(t) = 0.3t$). In Case A a natural boundary layer at the tension-free end arises because the distortion measures keep their initial values due to (1h)–(1i) (cf. Fig. 4, top right, dashed line). This ultimately causes the simulation to fail—in our example at time $t = 1.52$. In Case B the fiber behaves comparably over its length, but the imminent boundary layer is suppressed due to mass outflow. Thus, this set-up is preferable for parameter studies. However, at some point the elongation and the normal velocities explode when the fiber becomes thinner and cannot be properly resolved with a static grid—in our example at time

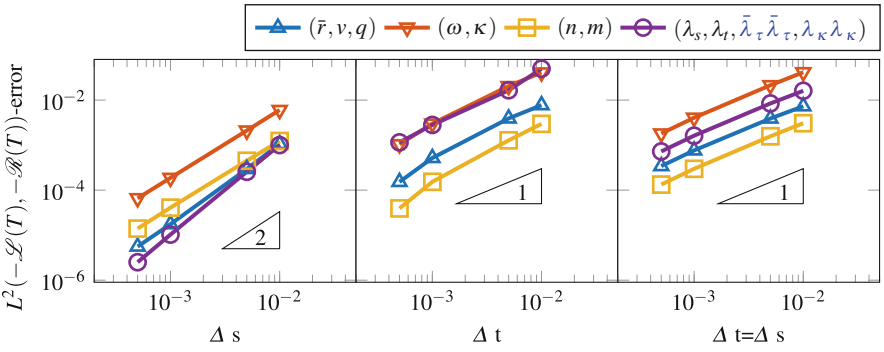


Fig. 3 *From left to right*: Spatial, temporal and combined convergence behavior for rotational spinning scenario with $(\mathscr{L}(t), \mathscr{R}(t)) = (t, 0)$, $(\mathrm{Re}, \mathrm{Rb}, \mathrm{Fr}, \epsilon, T) = (1, 1, 1, 0.05\sqrt{\pi}, 0.5)$ and default $\Delta t = \Delta s = 10^{-2}$. Numerically found orders for unknowns grouped like in the legend: Δs: (1.73, 1.50, 1.50, 2.01), Δt: (1.38, 1.27, 1.48, 1.15), $\Delta t = \Delta s$: (1.05, 1.07, 1.07, 1.06)

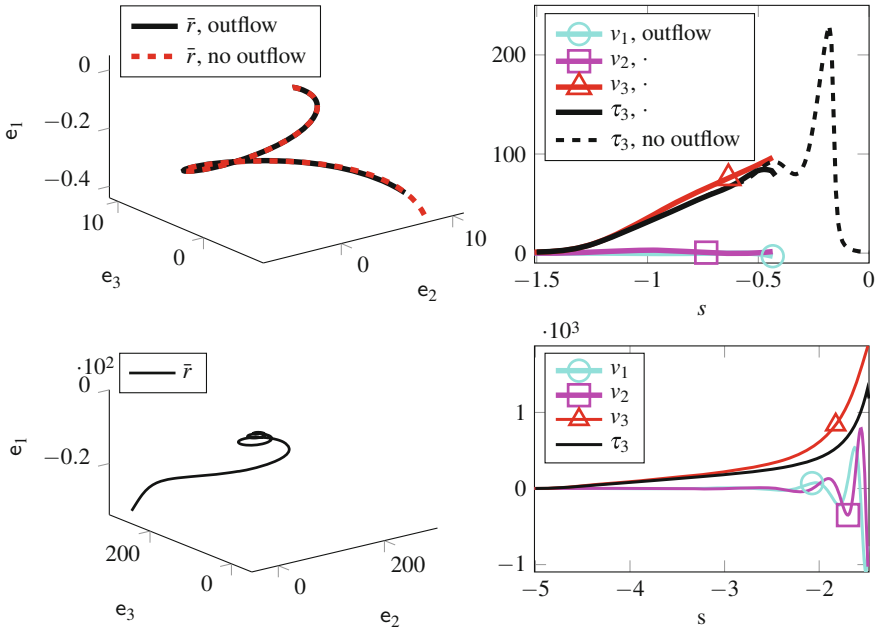


Fig. 4 Rotational spinning scenario with $\mathscr{L}(t) = t$, $(\mathrm{Re}, \mathrm{Rb}, \mathrm{Fr}, \epsilon, T) = (1, 0.1, 1, 0.05\sqrt{\pi}, \infty)$ and $\Delta t = \Delta s = 10^{-2}$. *Top*: Influence of boundary condition with and without outflow on fiber curve (*left*) and on velocity and elongation (*right*) at time $t = 1.51$. *Bottom*: Long-term behavior with mass outflow at time $t = 5.01$

$t = 5.02$ (Fig. 4, bottom). We expect simulations of longer run time when using a refined outflow function (e.g., tracking the thinnest point) and/or a possible adaptive mesh refinement suitable for large gradients. Note that our proposed model and discrete scheme allows the incorporation of moving mesh approaches, [3].

4 Conclusion

The proposed numerical scheme for the fiber model in the index-reduced stabilized kinematic formulation takes into account the underlying differential-algebraic characteristics. It avoids DAE drift-off effects and yields space-time convergence results that promise higher order accuracy when using respective approximations of the fluxes and source terms. It provides a basis for further investigations of industrial spinning processes and parameter ranges. Moreover, the underlying Cosserat rod formulation and the scheme have been developed with the thought of versatility to facilitate the exchange of material laws, geometry models and external loads enhancing the reach of possible applications.

Acknowledgements This work has been supported by German DFG, project 251706852, MA 4526/2-1, WE 2003/4-1.

References

1. Arne, W., Marheineke, N., Meister, A., Schiessl, S., Wegener, R.: Finite volume approach for the instationary Cosserat rod model describing the spinning of viscous jets. J. Comput. Phys. **294**, 20–37 (2015). doi:10.1016/j.jcp.2015.03.042
2. Gear, C.W.: Differential-algebraic equation index transformations. SIAM J. Sci. Stat. Comput. **9**, 39–47 (1988). doi:10.1137/0909004
3. Schiessl, S., Marheineke, N., Wegener, R.: A moving mesh framework based on three parametrization layers for 1d PDEs. In: Progress in Industrial Mathematics at ECMI 2014. Springer, Berlin (2016)
4. Wegener, R., Marheineke, N., Hietel, D.: Virtual production of filaments and fleeces. In: Currents in Industrial Mathematics, pp. 103–162. Springer, Berlin/Heidelberg (2015)

A Variational Approach to the Homogenization of Double Phase $p_h(x)$-Curl Systems in Magnetism

Hélia Serrano

Abstract We introduce a variational approach to study the homogenization of a class of $p_h(x)$-curl systems arising in Magnetism based on the study of the Γ-convergence of the sequence of associated energies. The explicit characterization of the effective coefficients is obtained by means of a three dimensional minimization problem when $p_h(x)$ is a double phase exponent.

1 Introduction

In materials science is crucial to give a macroscopic description of composites with a periodic microstructure made off materials with different properties, see [2, 8]. The mathematical theory of homogenization focuses on the asymptotic analysis of boundary value problems which describe microscopically the behaviour of such heterogeneous media. See the monographs [1, 3, 7].

Let Ω be a simply connected bounded domain in $\mathbb{R}^3$ with a Lipschitz continuous boundary $\partial\Omega$. Assume that Ω is a sample of a composite with a periodic microstructure with relative size $1/h$ formed by two alternate layers, one of material 1 and other of material 2, with relative thickness α/h and $(1-\alpha)/h$, respectively, normal to a given unit vector $\mathbf{e}$. Furthermore, assume that materials 1 and 2 are physically characterized by different positive parameters μ_1 and μ_2, respectively, which may represent for instance the magnetic permeability, so that the composite is characterized by

$$\mu_h(x) = \mu_1 \chi_{(0,\alpha)}(hx \cdot \mathbf{e}) + \mu_2(1 - \chi_{(0,\alpha)}(hx \cdot \mathbf{e}))$$

for a.e. $x \in \Omega$, where $\chi_{(0,\alpha)}(y)$ stands for the characteristic function of the interval $(0,\alpha) \subset (0,1)$.

H. Serrano (✉)
Departamento de Matemáticas, Universidad de Castilla-La Mancha, Av. Camilo José Cela 10, 13071 Ciudad Real, Spain
e-mail: heliac.pereira@uclm.es

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_107

Our aim is to study the asymptotic behaviour, as the parameter h goes to ∞, of the sequence of solutions $\{u_h\}$ of boundary value problems of type

$$\begin{cases} \operatorname{curl}\left(\mu_h(x)|\operatorname{curl} u(x)|^{p_h(x)-2}\operatorname{curl} u(x)\right) = \operatorname{curl} j(x) & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u\cdot\mathbf{n} = 0 & \text{on } \partial\Omega \\ \mu_h|\operatorname{curl} u|^{p_h-2}\operatorname{curl} u\times\mathbf{n} = j\times\mathbf{n} & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where the variable exponent $p_h : \Omega \to [2,+\infty)$ depends on the double phase microstructure of Ω so that

$$p_h(x) = p\chi_{(0,\alpha)}(hx\cdot\mathbf{e}) + q(1-\chi_{(0,\alpha)}(hx\cdot\mathbf{e}))$$

for a.e. $x\in\Omega$, with $2\le p\le q<\infty$, and j is a given vector potential in Ω. We present a variational approach to study the homogenization of $p_h(x)$-curl systems of type (1) which is based on the study of the Γ-convergence (see [4, 5]) of the sequence of associated energies. Namely, the boundary value problem (1) turns out to be the Euler-Lagrange equation associated with the energy E_h defined by

$$E_h(u) = \int_\Omega \left(\frac{\mu_h(x)}{p_h(x)}|\operatorname{curl} u(x)|^{p_h(x)} - j(x)\cdot\operatorname{curl} u(x)\right)dx \tag{2}$$

whenever the field u belongs to the variable exponent space

$$\mathbf{X}^{p_h(\cdot)}(\Omega) = \left\{u\in\mathbf{L}^{p_h(\cdot)}(\Omega) : \operatorname{curl} u\in\mathbf{L}^{p_h(\cdot)}(\Omega),\ \operatorname{div} u{=}0 \text{ in } \Omega,\ u\cdot\mathbf{n}{=}0 \text{ on } \partial\Omega\right\},$$

where the variable exponent Lebesgue space $\mathbf{L}^{p_h(\cdot)}(\Omega)$ is given by

$$\mathbf{L}^{p_h(\cdot)}(\Omega) = \left\{u:\Omega\to\mathbb{R}^3 \text{ measurable} : \int_\Omega |\eta u(x)|^{p_h(x)}dx<\infty \text{ for some } \eta>0\right\}.$$

In this way, we propose to study the behaviour, as h goes to ∞, of the sequence of minimizers u_h of the family of energies E_h through Γ-convergence in order to conclude on the behaviour of the sequence of solutions of the boundary value problem (1).

2 Γ-Convergence of Energies

Our main result is the following one.

Theorem 1 *The sequence of energies $E_h : \mathbf{L}^{p_h(\cdot)}(\Omega)\to\mathbb{R}\cup\{\infty\}$ defined by*

$$E_h(u) = \begin{cases} \displaystyle\int_\Omega \left(\frac{\mu_h(x)}{p_h(x)}|\operatorname{curl} u(x)|^{p_h(x)} - j(x)\cdot\operatorname{curl} u(x)\right)dx & \text{if } u\in\mathbf{X}^{p_h(\cdot)}(\Omega) \\ +\infty & \text{otherwise} \end{cases}$$

Γ-converges, with respect to the strong convergence in $\mathbf{L}^p(\Omega)$, to the energy E defined by

$$E(u) \;=\; \begin{cases} \displaystyle\int_\Omega \left(\psi(\operatorname{curl} u(x)) - j(x)\cdot \operatorname{curl} u(x)\right) dx & \text{if } \operatorname{curl} u \in \mathbf{L}_{hom}(\Omega) \\ +\infty & \text{otherwise,} \end{cases}$$

where $\psi : \mathbb{R}^3 \to [0, +\infty)$ is defined by

$$\psi(\rho) = \min_{\Lambda_1,\Lambda_2\in\mathbb{R}^3} \left\{ \alpha\frac{\mu_1}{p}|\Lambda_1|^p + (1-\alpha)\frac{\mu_2}{q}|\Lambda_2|^q \;:\; \begin{array}{l} \rho = \alpha\Lambda_1 + (1-\alpha)\Lambda_2 \\ (\Lambda_1 - \Lambda_2)\cdot \mathbf{e} = 0 \end{array} \right\}. \tag{3}$$

Notice that the Γ-limit functional E is well defined for all divergence-free fields whose curl belongs to the intermediate space denoted by $\mathbf{L}_{hom}(\Omega)$ such that $\mathbf{L}^q(\Omega) \subset \mathbf{L}_{hom}(\Omega) \subset \mathbf{L}^p(\Omega) + \mathbf{L}^q(\Omega)$.

Thus, the Euler-Lagrange equation associated with the Γ-limit functional E is the following boundary value problem

$$\begin{cases} \operatorname{curl}\left(\nabla\psi\left(\operatorname{curl} u(x)\right)\right) = \operatorname{curl} j(x) & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u\cdot\mathbf{n} = 0 & \text{on } \partial\Omega \\ \nabla\psi\left(\operatorname{curl} u\right)\times\mathbf{n} = j\times\mathbf{n} & \text{on } \partial\Omega, \end{cases}$$

which turns out to be the homogenized problem associated with the family of boundary value problems in (1).

Let us proceed to prove the previous result.

Proof We say that the sequence of energies $\{E_h\}$ is Γ-convergent, with respect to the strong topology in $\mathbf{L}^p(\Omega)$, to the energy E if, for any $u \in \mathbf{L}_{hom}(\Omega)$, it holds

1. for every sequence $\{u_h\} \subset \mathbf{L}^{p_h(\cdot)}(\Omega)$ such that $u_h \to u$ strongly in $\mathbf{L}^p(\Omega)$, then

$$E(u) \;\le\; \liminf_{h\to\infty} E_h(u_h); \tag{4}$$

2. there exists a sequence $\{u_h\} \subset \mathbf{L}^{p_h(\cdot)}(\Omega)$ such that $u_h \to u$ strongly in $\mathbf{L}^p(\Omega)$, and

$$E(u) \;=\; \lim_{h\to\infty} E_h(u_h). \tag{5}$$

Let us start by proving the first statement, i.e. inequality (4). Here, we apply similar techniques and tools used previously in [10] to prove the Γ-convergence of sequences of functionals satisfying standard growth conditions and defined in non-variable exponent spaces $\mathbf{X}^p(\Omega)$.

For any u such that $E(u) < \infty$, consider a sequence $\{u_h\} \subset \mathbf{L}^{p_h}(\Omega)$ strongly convergent to u in $\mathbf{L}^p(\Omega)$ such that $E_h(u_h) < +\infty$, otherwise there is nothing to prove. Let $\{\nu_x\}_{x\in\Omega}$ be the Young measure associated with the sequence $\{(\mu_h, p_h, \operatorname{curl} u_h)\}$ which may be decomposed, for a.e. $x \in \Omega$, as

$$\nu_x = \alpha\delta_{\mu_1} \otimes \delta_p \otimes \theta_{p,x} + (1-\alpha)\delta_{\mu_2} \otimes \delta_q \otimes \theta_{q,x}$$

for some probability measures $\theta_{p,x}$ and $\theta_{q,x}$ supported on $\mathbb{R}^3$. Therefore, applying [9, Theorem 6.11] it holds true

$$\liminf_{h\to\infty} \int_\Omega \frac{\mu_h(x)}{p_h(x)} |\operatorname{curl} u_h(x)|^{p_h(x)}\, dx \geq$$
$$\int_\Omega \left(\alpha \frac{\mu_1}{p} \int_{\mathbb{R}^3} |\lambda|^p d\theta_{p,x}(\lambda) + (1-\alpha)\frac{\mu_2}{q} \int_{\mathbb{R}^3} |\lambda|^q d\theta_{q,x}(\lambda) \right) dx.$$

It follows from the Jensen inequality that

$$\int_\Omega \left(\alpha \frac{\mu_1}{p} \int_{\mathbb{R}^3} |\lambda|^p d\theta_{p,x}(\lambda) + (1-\alpha)\frac{\mu_2}{q} \int_{\mathbb{R}^3} |\lambda|^q d\theta_{q,x}(\lambda) \right) dx \geq$$
$$\int_\Omega \left(\alpha \frac{\mu_1}{p} |\varphi(p,x)|^p + (1-\alpha)\frac{\mu_2}{q} |\varphi(q,x)|^q \right) dx,$$

where we have defined the field $\varphi : \{p,q\} \times \Omega \to \mathbb{R}^3$ by putting $\varphi(y,x) = \displaystyle\int_{\mathbb{R}^3} \lambda d\theta_{y,x}(\lambda)$. In this way, we conclude

$$\liminf_{h\to\infty} \int_\Omega \frac{\mu_h(x)}{p_h(x)} |\operatorname{curl} u_h(x)|^{p_h(x)}\, dx \geq$$
$$\int_\Omega \left(\alpha \frac{\mu_1}{p} |\varphi(p,x)|^p + (1-\alpha)\frac{\mu_2}{q} |\varphi(q,x)|^q \right) dx.$$

Notice that the Young measure $\{\xi_x\}_{x\in\Omega}$ associated with the sequence $\{\operatorname{curl} u_h\}$ may be characterized as $\xi_x = \alpha\theta_{p,x} + (1-\alpha)\theta_{q,x}$ for a.e. $x \in \Omega$, so that the weak limit has to satisfy the equality $\operatorname{curl} u(x) = \alpha\varphi(p,x) + (1-\alpha)\varphi(q,x)$ for a.e. $x \in \Omega$. From the arbitrariness of the field φ, we may take the minimum and thus it follows inequality (4).

Now, let us prove equality (5). Without loss of generality, let u be such that

$$\int_\Omega \psi(\operatorname{curl} u(x))\, dx < \infty$$

where the density ψ is defined in (3). Since Ω is an open bounded domain, we may decompose Ω such that there exists, for each $i \in \mathbb{N}$, a family of points $\{x_i^k\} \subset \Omega \setminus N_i$,

with $|N_i| = 0$, and radius $\{r_i^k\}$ for which $\Omega = \cup_k (x_i^k + r_i^k \Omega) \cup N_i$, and

$$\int_\Omega \psi(\operatorname{curl} u(x))\, dx = \lim_{i\to\infty} \sum_k \psi(\operatorname{curl} u(x_i^k)) \left|x_i^k + r_i^k \Omega\right|,$$

see [9, Lemma 7.9]. For each x_i^k there exists an optimal pair $(\Lambda_1(x_i^k), \Lambda_2(x_i^k)) \in \mathbb{R}^3 \times \mathbb{R}^3$ such that

$$\psi(\operatorname{curl} u(x_i^k)) = \alpha \frac{\mu_1}{p} |\Lambda_1(x_i^k)|^p + (1-\alpha)\frac{\mu_2}{q} |\Lambda_2(x_i^k)|^q,$$

with $\operatorname{curl} u(x_i^k) = \alpha \Lambda_1(x_i^k) + (1-\alpha)\Lambda_2(x_i^k)$ and $(\Lambda_1(x_i^k) - \Lambda_2(x_i^k)) \cdot \mathbf{e} = 0$. Thus, we have

$$\int_\Omega \psi(\operatorname{curl} u(x))\, dx = \lim_{i\to\infty} \sum_k \left(\alpha \frac{\mu_1}{p} |\Lambda_1(x_i^k)|^p + (1-\alpha)\frac{\mu_2}{q} |\Lambda_2(x_i^k)|^q \right) \left|x_i^k + r_i^k \Omega\right|.$$

For each $i, k \in \mathbb{N}$, if we define the sequence $\{(\mu_h, p_h, U_h^{i,k})\}$ in Ω by putting

$$(\mu_h(y), p_h(y), U_h^{i,k}(y)) = (\mu_1, p, \Lambda_1(x_i^k))\chi_{(0,\alpha)}(hy \cdot \mathbf{e}) + (\mu_2, q, \Lambda_2(x_i^k))(1 - \chi_{(0,\alpha)}(hy \cdot \mathbf{e})),$$

then the sequence $\{U_h^{i,k}\}$ converges weakly*, as h goes to ∞, to $\operatorname{curl} u(x_i^k)$ in $\mathbf{L}^\infty(\Omega)$, $\operatorname{div} U_h^{i,k} = 0$ in Ω, provided $(\Lambda_1(x_i^k) - \Lambda_2(x_i^k)) \cdot \mathbf{e} = 0$, and

$$\lim_{h\to\infty} \frac{1}{|\Omega|} \int_\Omega \frac{\mu_h(y)}{p_h(y)} |U_h^{i,k}(y)|^{p_h(y)} dy = \alpha \frac{\mu_1}{p} |\Lambda_1(x_i^k)|^p + (1-\alpha)\frac{\mu_2}{q} |\Lambda_2(x_i^k)|^q.$$

Therefore,

$$\int_\Omega \psi(\operatorname{curl} u(x))\, dx = \lim_{i\to\infty} \lim_{h\to\infty} \sum_k \frac{\left|x_i^k + r_i^k \Omega\right|}{|\Omega|} \int_\Omega \frac{\mu_h(y)}{p_h(y)} |U_h^{i,k}(y)|^{p_h(y)} dy.$$

Now, we may build a sequence $\{V_h\}$ defined in Ω such that it converges weakly* to $\operatorname{curl} u$ in $\mathbf{L}^\infty(\Omega)$, generates locally the same Young measure as $\{U_h^{i,k}\}$, and the sequence $\{\operatorname{div} V_h\}$ converges weakly to 0 in $\mathbf{L}^p(\Omega)$. Thus, it follows from [6, Lemma 2.17] there exists a sequence $\{u_h\} \subset \mathbf{L}^\infty(\Omega)$ so that $\operatorname{curl} u_h = V_h$ in Ω, $\operatorname{div} u_h = 0$ in Ω, and $u_h \cdot \mathbf{n} = 0$ on $\partial\Omega$ such that

$$\limsup_{h\to\infty} \int_\Omega \frac{\mu_h(x)}{p_h(x)} |\operatorname{curl} u_h(x)|^{p_h(x)} dx \le \int_\Omega \psi(\operatorname{curl} u(x))\, dx.$$

□

3 Examples

In this section we apply the main theorem to two different examples. In the first one we consider the same exponent $p = 2$ and recover the well know characterization of the effective coefficient, while in the second one we take $p = 2$ and $q = 3$.

Example 1 In the simplest case, when $p = 2 = q$, $\alpha = \frac{1}{2}$, $\mu_1 = 1$, $\mu_2 = 2$, and $\mathbf{e} = (1, 0, 0)$, the boundary value problem in (1) reads as

$$\begin{cases} \operatorname{curl}(\mu_h(x)\operatorname{curl} u(x)) = \operatorname{curl} j(x) & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \\ \mu_h \operatorname{curl} u \times \mathbf{n} = j \times \mathbf{n} & \text{on } \partial\Omega, \end{cases}$$

and its associated energy

$$E_h(u) = \int_\Omega \left(\frac{\mu_h(x)}{2} |\operatorname{curl} u(x)|^2 - j(x) \cdot \operatorname{curl} u(x) \right) dx$$

if $u \in \mathbf{X}^2(\Omega)$. Therefore, the sequence $\{E_h\}$ is Γ-convergent to the functional E whose density $\psi : \mathbb{R}^3 \to \mathbb{R}$ is defined by $\psi(\rho) = \frac{1}{2} M\rho \cdot \rho$, where the effective matrix M is given by

$$M = \begin{pmatrix} \frac{3}{2} & 0 & 0 \\ 0 & \frac{4}{3} & 0 \\ 0 & 0 & \frac{4}{3} \end{pmatrix},$$

and the homogenized problem is

$$\begin{cases} \operatorname{curl}(M\operatorname{curl} u(x)) = \operatorname{curl} j(x) & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \\ M\operatorname{curl} u \times \mathbf{n} = j \times \mathbf{n} & \text{on } \partial\Omega. \end{cases}$$

Example 2 In the case that $p = 2$, $q = 3$, $\alpha = \frac{1}{3}$, $\mu_1 = 2$, $\mu_2 = 3$, $\mathbf{e} = (1, 0, 0)$, the sequence of energies $\{E_h\}$ is Γ-convergent to the functional E whose density $\psi : \mathbb{R}^3 \to \mathbb{R}$ is defined by

$$\psi(\rho_1, \rho_2, \rho_3) = \min_{x,y \in \mathbb{R}} \left\{ \frac{1}{3}\left(\rho_1^2 + x^2 + y^2\right) + \frac{2}{3}\left(\rho_1^2 + \left(\frac{3\rho_2 - x}{2} \right)^2 + \left(\frac{3\rho_3 - y}{2} \right)^2 \right)^{\frac{3}{2}} \right\},$$

which is equivalent to

$$\psi(\rho_1,\rho_2,\rho_3)=\frac{1}{3}\left(\rho_1^2+9\rho_2^2+9\rho_3^2\right)+\frac{1}{3}r^2+\frac{2}{3}\left(\rho_1^2+\frac{r^2}{4}\right)^{\frac{3}{2}}-2r\sqrt{\rho_2^2+\rho_3^2},$$

where r satisfies the equation $r\left(\rho_1^2+\frac{r^2}{4}\right)^{\frac{1}{2}}+\frac{4}{3}r-4\sqrt{\rho_2^2+\rho_3^2}=0$. In this way, the Γ-limit E is defined in the intermediate space between $\mathbf{L}^2(\Omega)$ and $\mathbf{L}^3(\Omega)$. Namely, we have that $\psi(\rho_1,0,0)=\frac{1}{3}\rho_1^3+\frac{2}{3}\rho_1^2$, for every $\rho_1\in\mathbb{R}$, while $\psi(0,\rho_2,0)$ and $\psi(0,0,\rho_3)$ are quadratic polynomials, for every $\rho_2,\rho_3\in\mathbb{R}$, provided

$$\psi(0,\rho_2,0)=3\rho_2^2-\frac{8}{9}|\rho_2|\sqrt{4+18|\rho_2|}+\frac{8}{3}|\rho_2|-\frac{16}{81}\sqrt{4+18|\rho_2|}+\frac{32}{81}$$

and

$$\psi(0,0,\rho_3)=3\rho_3^2-\frac{8}{9}|\rho_3|\sqrt{4+18|\rho_3|}+\frac{8}{3}|\rho_3|-\frac{16}{81}\sqrt{4+18|\rho_3|}+\frac{32}{81}.$$

Acknowledgements This work was supported by projects MTM2013-47053-P from *Ministerio de Economía y Competitividad*, and PEII-2014-010-P from *Junta de Comunidades de Castilla-La Mancha* (Spain).

References

1. Bensoussan, A., Lions, J.L., Papanicolaou, G.: Asymptotic Analysis for Periodic Structures. North-Holland P.C., Amsterdam (1978)
2. Cessenat, M.: Mathematical Methods in Electromagnetism. World Scientific Publishing, Singapore (1996)
3. Cioranescu, D., Donato, P.: An Introduction to Homogenization. Oxford University Press, Oxford (1999)
4. Dal Maso, G.: An Introduction to Γ-Convergence. Birkhäuser, Basel (1993)
5. Focardi, M.: Γ-convergence: a tool to investigate physical phenomena across scales. Math. Methods Appl. Sci. **35**, 1613–1658 (2012)
6. Fonseca, I., Müller, S.: A-quasiconvexity, lower semicontinuity, and Young measures. SIAM J. Math. Anal. **30**, 1355–1390 (1999)
7. Jikov, V.V., Kozlov, S.M., Oleinik, O.A.: Homogenization of Differential Operators and Integral Functionals. Springer, Berlin/Heidelberg (1994)
8. Milton, G.: The Theory of Composites. Cambridge University Press, Cambridge (2004)
9. Pedregal, P.: Parametrized Measures and Variational Principles. Birkhäuser, Basel (1997)
10. Serrano, H.: A variational approach to the homogenization of laminate metamaterials. Nonlinear Anal. Real World Appl. **18**, 75–85 (2014)

Modelling of Combustion and Diverse Blow-Up Regimes in a Spherical Shell

Yuri N. Skiba and Denis M. Filatov

Abstract Physical phenomena with critical blow-up regimes simulated by the 3D nonlinear diffusion equation in a spherical shell are studied. For solving the model numerically, the original differential operator is split along the radial coordinate, as well as an original technique of using two coordinate maps for solving the 2D subproblem on the sphere is involved. This results in 1D finite difference subproblems with simple periodic boundary conditions in the latitudinal and longitudinal directions that lead to unconditionally stable implicit second-order finite difference schemes. A band structure of the resulting matrices allows applying fast direct (non-iterative) linear solvers using the Sherman-Morrison formula and Thomas algorithm. The developed method is tested in several numerical experiments. Our tests demonstrate that the model allows simulating different regimes of blow-up in a 3D complex domain. In particular, heat localisation is shown to lead to the breakup of the medium into individual fragments followed by the formation and development of self-organising patterns, which may have promising applications in thermonuclear fusion, nonlinear inelastic deformation and fracture of loaded solids and media and other areas.

1 Introduction

The phenomenon of diffusion has many manifestations in nature. One of the most obvious is from gas dynamics and atmospheric environment. The air, as it is known, is a mixture of different gases, such as carbon dioxide, oxygen, hydrogen, nitrogen and particles of dust. However, due to diffusion the atmospheric composition at any given altitude is fairly homogeneous. The driving force for diffusion is the

Yu.N. Skiba
Centre for Atmospheric Sciences, National Autonomous University of Mexico, Av. Universidad 3000, Mexico City, C.P. 04510, MEXICO
e-mail: skiba@unam.mx

D.M. Filatov (✉)
Sceptica Scientific Ltd, Carpenter Court, 1 Maple Road, Bramhall, Stockport, Cheshire SK7 2DH, UK
e-mail: denis.filatov@sceptica.co.uk

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_108

difference between the thermodynamic potentials. By means of the redistribution of the substance a system tends to equalise local differences of the potentials, and, consequently, is approximated to the thermodynamic equilibrium, and this alignment is carried out by diffusion.

For certain diffusion (heat transfer) models, the charge and momentum, taken in blow-up regimes, are the quantities able to model the formation and evolution of tornadoes and lightnings. Specifically, in [7] it is shown that in the life cycle of tornadoes experiencing in blow-up modes the tornado elephant trunk descends from the atmosphere to the ground. For atmospheric flows of regional and global scales simulation in spherical geometry is appropriate, which leads to 2D (on the sphere) or 3D (in a spherical shell) nonlinear diffusion models.

More information on nonlinear diffusion phenomena and applications can be found, e.g., in [1–6, 9–12, 17–19].

The development of numerical methods for solving the diffusion equation on a sphere (2D) or in a spherical shell (3D) has its own features. The point is that the sphere is not a doubly periodic domain, since it is periodic in the longitude, but is not in the latitude due to the presence of two poles. Therefore, a numerical procedure designed to solve the original 2D problem will be computationally cumbersome, because the matrix of the resulting linear system will be of a general type, hence not permitting to apply some or other fast linear solvers. As for making some or other modifications to the model prior to computing, these normally bring to the necessity of constructing mathematically and physically correct boundary conditions or involving special numerical procedures near the poles (e.g., matrix bordering). Both are always a problem, as the poles represent an artificial boundary appearing exclusively due to the latitudinal-longitudinal coordinate system, and the construction of proper artificial boundary conditions is a serious independent question.

Some of the three-dimensional nonlinear diffusion equations considered in the spherical geometry are convenient to solve by separating the diffusion operator into the spherical and radial components. The main idea consists in splitting the original differential operator by coordinates and subsequent constructing finite difference schemes for the split 1D subproblems using two different coordinate maps for the same sphere when computing in the longitudinal and latitudinal directions. The key advantage of this technique is that each split 1D equation can be equipped with a periodic boundary condition, despite the sphere being not a doubly periodic domain. Therefore, unlike the existing methods, this one does not require applying special numerical procedures for careful computing the solution near the poles, which is always a serious challenge. The developed algorithm is cheap-to-implement from the computational point of view [12–15]. The theoretical results are confirmed numerically by simulating various nonlinear diffusion processes.

2 Nonlinear Diffusion Model

We study the 3D nonlinear diffusion equation in the spherical shell written in the coordinates r (the radius), λ (the longitude), φ (the latitude)

$$\frac{\partial T}{\partial t} = (A_r + A_\lambda + A_\varphi)T + f, \tag{1}$$

where

$$A_r T = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 D \frac{\partial T}{\partial r}\right), \tag{2}$$

$$A_\lambda T = \frac{1}{r\cos\varphi}\frac{\partial}{\partial \lambda}\left(\frac{D}{r\cos\varphi}\frac{\partial T}{\partial \lambda}\right), \tag{3}$$

$$A_\varphi T = \frac{1}{r\cos\varphi}\frac{\partial}{\partial \varphi}\left(\frac{D\cos\varphi}{r}\frac{\partial T}{\partial \varphi}\right). \tag{4}$$

Here $D \sim (T)^\alpha$ is the diffusion coefficient, while f is the external forcing. Using the standard method of operator splitting, we define the grid spacing in time as $\tau = t_{n+1} - t_n$ and in every double time interval (t_{n-1}, t_{n+1}) linearise (1) and then split the linearised equation by coordinates; upon this we assume $D \sim (T^{n-1})^\alpha$, $T^{n-1} = T(\lambda, \varphi, t_{n-1})$. The split equations are considered to be solved in time successively: the solution to the radial subproblem (i.e. solved in r) is the initial condition for the longitudinal subproblem; that, in turn, after being solved in λ, will provide the initial condition for the latitudinal subproblem; the latter, solved in φ, will supply the initial condition for the forcing equation. Next, with the aim to increase the approximation order in time, the order of splitting changes to the inverse one, which yields the bicyclic splitting schemes [8] (see (7)–(13) below).

Because the metric term $r\cos\varphi$ vanishes at $\varphi = \pm\pi/2$, expressions (3)–(4) are meaningless at the poles. Therefore, for covering the entire sphere we assume $\Delta\lambda = \lambda_{k+1} - \lambda_k$ and $\Delta\varphi = \varphi_{l+1} - \varphi_l$ and define on S a grid, making a half step shift in φ, i.e.

$$S^{(1)}_{\Delta\lambda,\Delta\varphi} = \left\{(\lambda_k, \varphi_l) : \lambda_k \in \left[\tfrac{\Delta\lambda}{2}, 2\pi + \tfrac{\Delta\lambda}{2}\right), \varphi_l \in \left[-\tfrac{\pi}{2} + \tfrac{\Delta\varphi}{2}, \tfrac{\pi}{2} - \tfrac{\Delta\varphi}{2}\right]\right\}. \tag{5}$$

Thereby we exclude the pole singularities, so that the subsequent finite difference equations will have sense everywhere on $S^{(1)}_{\Delta\lambda,\Delta\varphi}$. For dealing with (3) one has, evidently, to use the periodic boundary condition in λ, as the sphere S is a periodic domain in the longitude.

As for dealing with (4), the point is that the sphere is not a periodic domain in φ. Several approaches can be used to solve the latitudinal subproblem, e.g., one could involve a matrix bordering procedure or try to paste the solution from the opposite meridians at the poles. A disadvantage of these techniques is that in the case of an implicit temporal approximation the resulting matrix will be of a general type, and so no fast algorithms of linear algebra can be used for

computing the solution. Alternatively, an explicit temporal discretisation can be used, but this will impose serious restrictions on the timestep, especially when α is large. However, this is exactly the method of splitting that allows to avoid these undesired procedures which, besides, if performed inaccurately, may easily result in introducing nonphysical modes into the solution. Due to the splitting, for computing the solution in φ we change the coordinate map from (5) to

$$S^{(2)}_{\Delta\lambda,\Delta\varphi} = \left\{(\lambda_k, \varphi_l) : \lambda_k \in \left[\tfrac{\Delta\lambda}{2}, \pi - \tfrac{\Delta\lambda}{2}\right], \varphi_l \in \left[-\tfrac{\pi}{2} + \tfrac{\Delta\varphi}{2}, \tfrac{3\pi}{2} + \tfrac{\Delta\varphi}{2}\right)\right\}. \quad (6)$$

Obviously, grid (6) contains the same nodes as (5). The use of the two coordinate maps allows employing the same numerical algorithm with the periodic boundary conditions in both directions, λ and φ [13–15]. The only change we have to make in (4) if using (6) is to replace $\cos\varphi$ with $|\cos\varphi|$, as well.

For the second-order finite difference schemes we take the approximations

$$\frac{\partial}{\partial r}\left(r^2 D\frac{\partial T}{\partial r}\right) \approx \frac{1}{\Delta r}\left((r^2D)_{p+\frac{1}{2}}\frac{T_{p+1}-T_p}{\Delta r} - (r^2D)_{p-\frac{1}{2}}\frac{T_p - T_{p-1}}{\Delta r}\right),$$

$$\frac{\partial}{\partial\lambda}\left(D\frac{\partial T}{\partial\lambda}\right) \approx \frac{1}{\Delta\lambda}\left(D_{k+\frac{1}{2}}\frac{T_{k+1}-T_k}{\Delta\lambda} - D_{k-\frac{1}{2}}\frac{T_k - T_{k-1}}{\Delta\lambda}\right),$$

$$\frac{\partial}{\partial\varphi}\left(D|\cos\varphi|\frac{\partial T}{\partial\varphi}\right) \approx \frac{1}{\Delta\varphi}\left((D|\cos\varphi|)_{l+\frac{1}{2}}\frac{T_{l+1}-T_l}{\Delta\varphi} - (D|\cos\varphi|)_{l-\frac{1}{2}}\frac{T_l - T_{l-1}}{\Delta\varphi}\right).$$

Supplementing them with the bicyclic splitting (a sort of the Strang splitting [16])

$$\frac{T^{n-\frac{3}{4}}_{pkl} - T^{n-1}_{pkl}}{\tau} = A_r\frac{T^{n-\frac{3}{4}}_{pkl} + T^{n-1}_{pkl}}{2}, \quad (7)$$

$$\frac{T^{n-\frac{2}{4}}_{pkl} - T^{n-\frac{3}{4}}_{pkl}}{\tau} = A_\lambda\frac{T^{n-\frac{2}{4}}_{pkl} + T^{n-\frac{3}{4}}_{pkl}}{2}, \quad (8)$$

$$\frac{T^{n-\frac{1}{4}}_{pkl} - T^{n-\frac{2}{4}}_{pkl}}{\tau} = A_\varphi\frac{T^{n-\frac{1}{4}}_{pkl} + T^{n-\frac{2}{4}}_{pkl}}{2}, \quad (9)$$

$$\frac{T^{n+\frac{1}{4}}_{pkl} - T^{n-\frac{1}{4}}_{pkl}}{2\tau} = f^n_{pkl}, \quad (10)$$

$$\frac{T^{n+\frac{2}{4}}_{pkl} - T^{n+\frac{1}{4}}_{pkl}}{\tau} = A_\varphi\frac{T^{n+\frac{2}{4}}_{pkl} + T^{n+\frac{1}{4}}_{pkl}}{2}, \quad (11)$$

$$\frac{T^{n+\frac{3}{4}}_{pkl} - T^{n+\frac{2}{4}}_{pkl}}{\tau} = A_\lambda\frac{T^{n+\frac{3}{4}}_{pkl} + T^{n+\frac{2}{4}}_{pkl}}{2}, \quad (12)$$

$$\frac{T^{n+1}_{pkl} - T^{n+\frac{3}{4}}_{pkl}}{\tau} = A_r\frac{T^{n+1}_{pkl} + T^{n+\frac{3}{4}}_{pkl}}{2}, \quad (13)$$

we achieve the second order accuracy in time in the interval (t_{n-1}, t_{n+1}) [8, 16], provided $D \sim (T^{n-1})^{\alpha}$.

The solution T^{n+1} to (7)–(13) converges to the solution to the unsplit 3D linearised differential problem in the interval (t_{n-1}, t_{n+1}) under $\tau \to 0$.

3 Numerical Tests

In Fig. 1 we provide the initial and final solutions of equation (1) for the phenomenon of combustion [11]. There are three different mode of combustion depending on the parameter β that introduces nonlinearity with respect to the sources $f \sim T^{\beta}$: the case $\beta < \alpha + 1$ corresponds to the HS-regime—the area of combustion is getting expanded while burning; the case $\beta > \alpha + 1$ corresponds to the LS-regime—the area of combustion is getting narrow; the case $\beta = \alpha + 1$ corresponds to the constant area of burning. In all the cases the given source leads to an infinite increase of the temperature T while the time grows, that is a blow-up occurs [11].

In Fig. 2 another nonlinear diffusion phenomenon is simulated, namely temperature waves. Unlike the phenomenon of combustion, here $f \sim T - T^3$, which results

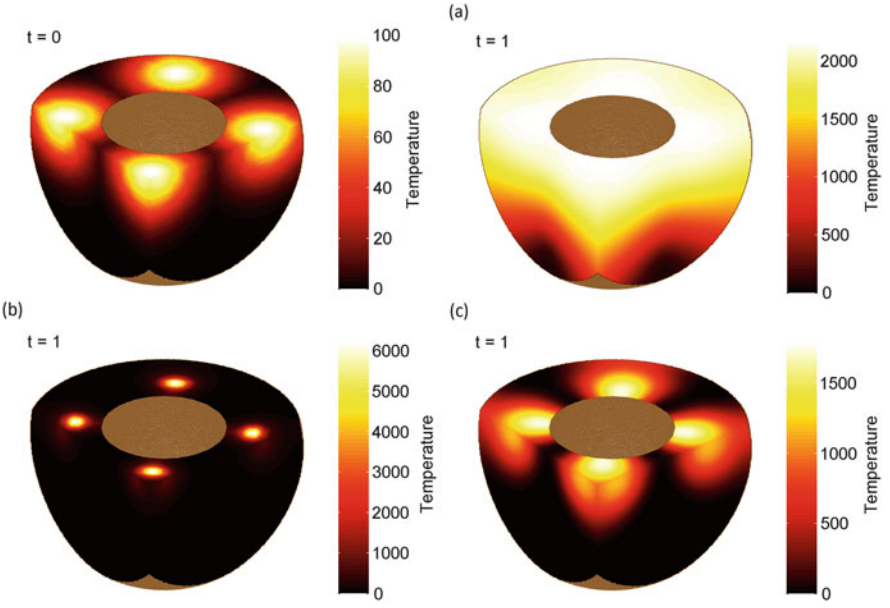


Fig. 1 The initial condition ($t = 0$) and the final solution ($t = 1$) of equation (1) with $f \sim T^{\beta}$, where $\beta < \alpha + 1$ (the area of combustion is expanding, (**a**)), $\beta > \alpha + 1$ (the area of combustion is reducing, (**b**)) and $\beta = \alpha + 1$ (the area of combustion is constant, (**c**))

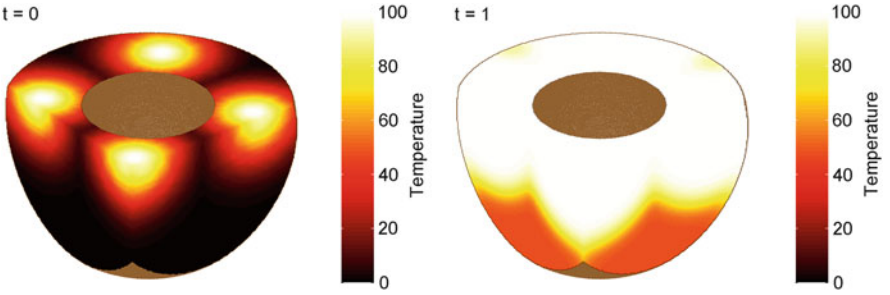


Fig. 2 The initial condition ($t = 0$) and the final solution ($t = 1$) of equation (1) with $f \sim T - T^3$. This simulates the phenomenon of temperature waves with a constant front

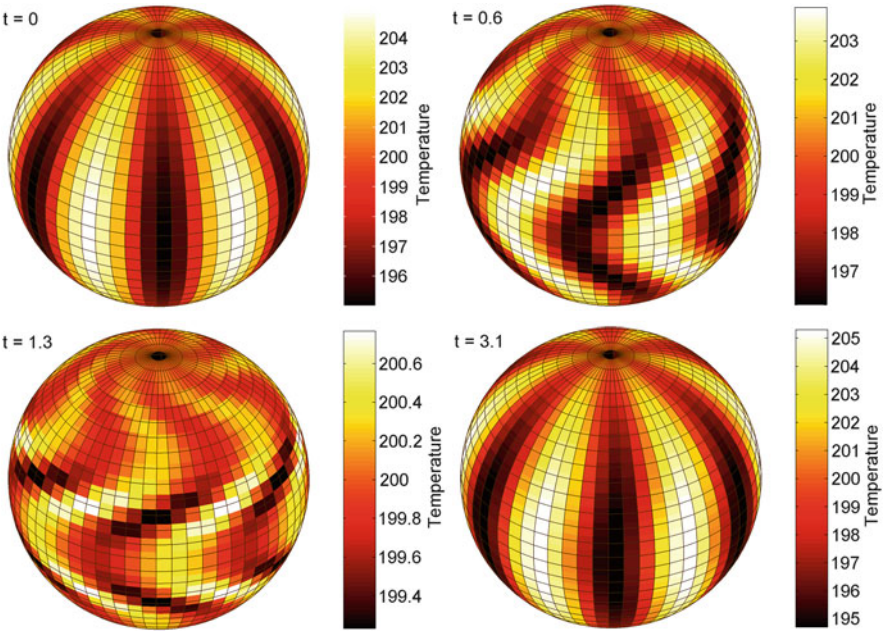


Fig. 3 The initial condition ($t = 0$) and several solution of equation (1) for the spherical travelling waves phenomenon. This is an example of nonlinear interaction between diverse mechanisms of the model—external forcing and dissipation, which results in solutions with periodic breakups and renewals of complex structures via self-organisation

in another sort of solution: while the area of the temperature is expanding (similarly to the HS-mode of combustion), the maximum value of T does not change, i.e. the wave front has a constant value of temperature.

In Fig. 3 we show a yet another phenomenon, spherical travelling waves. These are modelled by the diffusion equation with a specially chosen right-hand side function f. The sinusoid begins travelling clockwise at the low latitudes and near the poles, while at the intermediate latitudes it goes anticlockwise. Then the

processes alters, and rotation in the opposite direction takes place (not shown). This phenomenon evinces the interaction between three different mechanisms—nonlinearity, external forcing and dissipation, generating solutions with periodic breakups and renewals of complex structures via self-organisation.

Acknowledgements Yuri Skiba is grateful to the System of National Researchers of Mexico (SNI-CONACYT) for partial financial support.

References

1. Agoshkov, V.I., Saleri, F.: Recent developments in the numerical simulation of shallow water equations. Part III: boundary conditions and finite element approximations in the river flow calculations. Math. Model. **8**, 3–24 (1996)
2. Bear, J.: Dynamics of Fluids in Porous Media. Courier Dover Publications, New York (1988)
3. Catté, F., et al.: Image selective smoothing and edge detection by nonlinear diffusion. SIAM J. Numer. Anal. **29**, 182–193 (1992)
4. Glicksman, M.E.: Diffusion in Solids: Field Theory, Solid-State Principles and Applications. John Wiley & Sons, Hoboken (2000)
5. King, J.R.: 'Instantaneous source' solutions to a singular nonlinear diffusion equation. J. Eng. Math. **27**, 31–72 (1993)
6. Lacey, A.A., Ockerdon, J.R., Tayler, A.B.: 'Waiting-time' solutions of a nonlinear diffusion equation. SIAM J. Appl. Math. **42**, 1252–1264 (1982)
7. Lykov, I.A.: Blow-up modes in atmospheric motions: particularities of mathematical and numerical modelling by means of nonlinear dynamics. PhD Thesis, Ekaterinburg, Russia (2013) (in Russian)
8. Marchuk, G.I.: Methods of Computational Mathematics. Springer, Berlin (1982)
9. Rudykh, G.A., Semenov, E.I.: Non-self-similar solutions of multidimensional nonlinear diffusion equations. Math. Notes **67**, 200–206 (2000)
10. Samarskii, A.A.: Nonlinear effects of blow-up and localization processes in burning problems. In: Brauner, C.-M., Schmidt-Lainé, C. (eds.) Mathematical Modeling in Combustion and Related Topics, pp. 217–231. Martinus Nijhoff Publishers, Leiden (1988)
11. Samarskii, A.A., et al.: Blow-up in Quasilinear Parabolic Equations. Walter de Gruyter, Berlin (1995)
12. Skiba, Y.N.: Dual oil concentration estimates in ecologically sensitive zones. Environ. Monit. Assess. **43**, 139–151 (1996)
13. Skiba, Y.N., Filatov, D.M.: On splitting-based mass and total energy conserving arbitrary order shallow-water schemes. Numer. Methods Partial Differ. Equ. **23**, 534–552 (2007)
14. Skiba, Y.N., Filatov, D.M.: Conservative arbitrary order finite difference schemes for shallow-water flows. J. Comput. Appl. Math. **218**, 579–591 (2008)
15. Skiba, Y.N., Filatov, D.M.: Simulation of nonlinear diffusion on a sphere. In: Proceedings of the 1st International Conference on Simulation and Modelling Methodologies, Technologies and Applications (SIMULTECH 2011), Noordwijkerhout, the Netherlands, pp. 24–30 (2011)
16. Strang, G.: On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5**, 506–517 (1968)
17. Vorob'yov, A.K.: Diffusion Problems in Chemical Kinetics. Moscow University Press, Moscow (2003) (in Russian)
18. Wu, Z., et al.: Nonlinear Diffusion Equations. World Scientific Publishing, Singapore (2001)
19. Zheng, X., Palffy-Muhoray, P.: Nonlinear diffusion on a sphere. Bull. Am. Phys. Soc. **57**, B50.7 (2012). http://meetings.aps.org/Meeting/MAR12/Session/B50.7, http://meetings.aps.org/Meeting/MAR12/PersonIndex/2299

Parameterized Model Order Reduction by Superposition of Locally Reduced Bases

Tino Soll and Roland Pulch

Abstract We present an approach for model order reduction of parameterized linear dynamical systems which combines local reduced bases using singular value decompositions. The local reduced bases are computed for a fixed set of sample points in the parameter domain by moment matching techniques. Covering the parameter domain with sample points may yield a very large set of samples. We investigate a superposition approach that takes only a few of the local models into account to keep the resulting reduced dimension small. Our approach will be compared to one existing approach which directly interpolates local bases in some illustrating numerical experiments.

1 Introduction

In many applications dynamical systems are derived for modelling physical or chemical processes, for example. Automatic model generation or semi-discretizations of partial differential equations typically yield high-dimensional systems. Consequently, simulations for many different parameter values require high computational costs. The aim of parameterized model order reduction (pMOR) is to find dynamical systems of much lower dimension while preserving the dependency on physical or geometrical parameters. Thus more efficient simulations are feasible over a broad range of parameter values.

In pMOR of linear dynamical systems, often local bases are determined for a finite set of parameter samples. These local bases may be used either to construct a global basis for all possible parameter values or only for a subset of the parameter domain, see e.g. [4, 6]. In this paper we utilize the second approach by superposition of the parameter domain with the closest sample points to the considered parameter value.

T. Soll (✉) • R. Pulch
Institut für Mathematik und Informatik, Ernst-Moritz-Arndt-Universität Greifswald,
Walther-Rathenau-Str. 47, 17489 Greifswald, Germany
e-mail: tino.soll@uni-greifswald.de; roland.pulch@uni-greifswald.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_109

We introduce the problem setup and specify the aim of pMOR in the following section. There are several approaches which make use of techniques of ordinary model order reduction to compute local models for fixed parameter values. We will briefly introduce an existing approach from [1, 4] in Sect. 3.1 where these local models are combined by some interpolation technique. Alternatively, our approach combines local models by applying singular value decompositions (SVDs) in Sect. 3.2. In Sect. 4 both approaches are applied to a test example for some numerical simulations and to compare the performances.

2 Problem Setup

Let a dynamical system with P parameters $\mu = (\mu_1, \ldots, \mu_P) \in \Pi \subset \mathbb{R}^P$ be given of the form

$$\begin{aligned} E(\mu)\,\dot{x}(t,\mu) &= A(\mu)\,x(t,\mu) + B(\mu)\,u(t) \\ y(t,\mu) &= C(\mu)\,x(t,\mu), \end{aligned} \tag{1}$$

which is linear in state. The state-vector is denoted by $x(t,\mu) \in \mathbb{R}^N$, whereas $u(t) \in \mathbb{R}^M$ and $y(t,\mu) \in \mathbb{R}^K$ represent the M inputs and the K outputs (quantities of interest). We call the size N of the state-vector the dimension of the parametric model (1). The system matrices $E(\mu), A(\mu) \in \mathbb{R}^{N\times N}$, $B(\mu) \in \mathbb{R}^{N\times M}$ and $C(\mu) \in \mathbb{R}^{K\times N}$ are parameter-dependent. We assume that the matrices $E(\mu)$ are regular and that the system (1) is asymptotically stable for each $\mu \in \Pi$, meaning that all eigenvalues of $E(\mu)^{-1}A(\mu)$ have negative real parts (see [2, 4]). The parameter domain Π is usually bounded, for example, a compact cuboid. The dimension of the dynamical system is often very large, and thus simulations for many different parameter values or varying input functions may become too expensive.

To study the input-output-behavior of a linear dynamical system we consider the *transfer function* of (1) in the frequency domain, cf. [2] for the non-parametric case,

$$H(s,\mu) := C(\mu)\,\left(s\,E(\mu) - A(\mu)\right)^{-1}\,B(\mu) \in \mathbb{C}^{K\times M},$$

where $s \in S(\mu) \subseteq \mathbb{C}$ with $S(\mu) := \{s \in \mathbb{C} \,:\, \det(s\,E(\mu) - A(\mu)) \neq 0\}$.

The aim of parametric model reduction is to find a dynamical system

$$\begin{aligned} E_{\mathrm{red}}(\mu)\,\dot{z}(t,\mu) &= A_{\mathrm{red}}(\mu)\,z(t,\mu) + B_{\mathrm{red}}(\mu)\,u(t) \\ y_{\mathrm{red}}(t,\mu) &= C_{\mathrm{red}}(\mu)\,z(t,\mu) \end{aligned} \tag{2}$$

of dimension $n \ll N$ with $y_{\mathrm{red}}(t,\mu) \approx y(t,\mu)$ for all $t \geq 0$, all $\mu \in \Pi$ and a broad class of inputs u. This means finding a dynamical system of much lower dimension which has a similar input-output-behavior as the original system and preserves the parameter dependency, i.e., $H_{\mathrm{red}}(s,\mu) \approx H(s,\mu)$ for all parameter

values $\mu \in \Pi$ and a broad range of frequency points $s \in \mathbb{C}$, where H_{red} denotes the transfer function of the reduced system (2). We may achieve this approximation by using two projection matrices $V(\mu), W(\mu) \in \mathbb{C}^{N\times n}$, which also depend on the parameter μ, and the projected matrices

$$\begin{aligned} E_{\text{red}}(\mu) &= W(\mu)^{\text{H}} E(\mu) V(\mu), \; A_{\text{red}}(\mu) = W(\mu)^{\text{H}} A(\mu) V(\mu), \\ B_{\text{red}}(\mu) &= W(\mu)^{\text{H}} B(\mu), \qquad C_{\text{red}}(\mu) = C(\mu) V(\mu). \end{aligned} \tag{3}$$

We discuss techniques for the computation of the projection matrices $V(\mu)$ and $W(\mu)$ for the reductions (3) in the following section.

3 Basis Computation

For a fixed tuple $\mu \in \Pi$, a non-parametric linear dynamical system (1) occurs for which projection matrices $V(\mu)$ and $W(\mu)$ can be computed by techniques of ordinary morel order reduction (MOR). One class is given by *moment maching techniques* or methods based on *Krylov subspaces*. This approach is to find a reduced system of dimension n, whose transfer function is a local approximation of the original transfer function in a neighborhood of a fixed frequency point s_0, see [5].

Other approaches to build local projection matrices are, for example, *balanced truncation*, where Lyapunov equations have to be solved, or *proper orthogonal decomposition*, where an SVD is applied to a matrix of so-called snapshots in the time domain to obtain the projection matrices, see [2, 4].

In this paper we consider approaches where the projection matrices $V(\mu), W(\mu)$ are computed efficiently for each parameter value $\mu \in \Pi$. The starting point of the following two approaches is the computation of local projection matrices for some well chosen parameter values $\mu_1, \mu_2 \ldots, \mu_Q \in \Pi$. As we concentrate on Krylov-subspace methods for MOR, we additionally need several expansion points in the frequency domain to get a more uniform approximation.

For each sample point $\mu_j, j = 1, \ldots, Q$, we apply a method of MOR to the system (1) with $\mu = \mu_j$ and some expansion points $\{s_1, s_2, \ldots\}$ to obtain projection matrices $V_j \in \mathbb{C}^{N\times n_j}$ and $W_j \in \mathbb{C}^{N\times n_j}$, which can be used to construct a local reduced model of dimension n_j. These local models are combined subject to some given parameter value $\hat{\mu} \neq \mu_j, j = 1, \ldots, Q$.

In the first subsection one existing approach which directly interpolates these local bases to compute the projection matrices $V(\hat{\mu})$ and $W(\hat{\mu})$ is given which can be found in [1, 4]. In the second subsection we present our approach to compute these projection matrices using SVDs. Without loss of generality, we assume that $\mu_1, \ldots, \mu_R$ are the R closest sample points to $\hat{\mu}$ with non-decreasing distances, i.e., $\|\hat{\mu} - \mu_1\| \leq \|\hat{\mu} - \mu_2\| \leq \cdots \leq \|\hat{\mu} - \mu_R\|$.

3.1 Interpolation of Local Bases

The interpolation approach requires the local projection matrices to be of the same dimension $n < N$. Given the local projection matrices $V_1, \ldots, V_R \in \mathbb{C}^{N \times n}$ and the parameter value $\hat{\mu}$ for which the projection matrix $V(\hat{\mu})$ shall be computed, the main idea is that $V(\hat{\mu})$ interpolates $V_1, \ldots, V_R$ on a *tangent subspace* to the Grassmann manifold $\mathscr{G}_{N,r}$. We define $\mathscr{R}_k$ to be the subspace which is spanned by the columns of V_k and choose $\mathscr{R}_1$ to be the reference point for the further computations. Let $\mathscr{T}_{\mathscr{R}_1}$ be the tangent space of $\mathscr{G}_{N,r}$ at $\mathscr{R}_1$. All subspaces $\mathscr{R}_2, \ldots, \mathscr{R}_R$ are mapped onto the tangent space defined by $\mathscr{R}_1$ which describes the mapping $\mathscr{X}_k = \mathrm{Log}_{\mathscr{R}_1}(\mathscr{R}_k) \in \mathscr{T}_{\mathscr{R}_1}$ for $k = 1, \ldots, R$.

To compute the matrix T_k for $k = 1, \ldots, R$, which represents the point $\mathscr{X}_k \in \mathscr{T}_{\mathscr{R}_1}$, first a thin SVD $(I - V_1 V_1^{\mathrm{H}}) V_k (V_1^{\mathrm{H}} V_k)^{-1} = U_k \Sigma_k Z_k^{\mathrm{H}}$ is determined. Then let $T_k = U_k \arctan(\Sigma_k) Z_k^{\mathrm{H}}$. For a new parameter point $\hat{\mu} \in \Pi$ the method interpolates the local bases in their representations $T_1, \ldots, T_R$. The interpolation can be done using the parameter points $\mu_1, \ldots, \mu_R$, for example,

$$T(\hat{\mu}) = \sum_{k=1}^{R} L_k(\hat{\mu})\, T_k, \tag{4}$$

where $L_k(\hat{\mu})$ are the Lagrange basis functions. The resulting local basis $V(\hat{\mu})$ is achieved by computing a thin SVD

$$T(\hat{\mu}) = \hat{U}\, \hat{\Sigma}\, \hat{Z}^{\mathrm{H}}, \tag{5}$$

followed by

$$V(\hat{\mu}) = V_1 \hat{Z} \cos(\hat{\Sigma}) + \hat{U} \sin(\hat{\Sigma}). \tag{6}$$

The trigonometric functions in (6) are applied component-wise to the singular values on the diagonal of $\hat{\Sigma}$ only. Using the Galerkin approach the matrix $W(\hat{\mu})$ is obtained by setting $W(\hat{\mu}) := V(\hat{\mu})$.

3.2 Concatenation of Local Bases by Superposition of Local Models

Let $V_1, \ldots, V_R$ and $W_1, \ldots, W_R$ be the local bases associated to the R closest sample points $\mu_1, \ldots, \mu_R$ to $\hat{\mu}$ where $V_j, W_j \in \mathbb{C}^{N \times n_j}$ may have different dimensions $n_j \in \mathbb{N}$ for $j = 1, \ldots, R$. By concatenating these matrices we obtain

$$\widetilde{V} := [V_1\ V_2\ \cdots\ V_R] \qquad \text{and} \qquad \widetilde{W} := [W_1\ W_2\ \cdots\ W_R]. \tag{7}$$

We apply a thin SVDs

$$\widetilde{V} = \widehat{V}\,\Sigma_V\,Z_V^{\mathrm{H}} \quad \text{and} \quad \widetilde{W} = \widehat{W}\,\Sigma_W\,Z_W^{\mathrm{H}} \tag{8}$$

to remove linear dependencies and approximative redundancies which may occur by stringing together the local matrices. Then the columns of $\widehat{V}$ and $\widehat{W}$ are orthonormal bases spanning the same subspaces as $\widetilde{V}$, $\widetilde{W}$ and are called left singular vectors. In the matrices Σ_V, Σ_W we assume that the singular values are decreasing monotonically with respect to the column ordering, i.e., $\sigma_1^V \geq \sigma_2^V \geq \ldots \geq 0$ and $\sigma_1^W \geq \sigma_2^W \geq \ldots \geq 0$. Hence the first n columns belong to the n largest singular values. Typically the sequence of singular values in Σ_V and Σ_W decreases quite fast. Consequently it is reasonable to take only the left singular vectors of those singular values that are larger than a predetermined tolerance $\varepsilon > 0$.

We define

$$n := \max\{n_V, n_W\}, \tag{9}$$

$$n_V := \max\{j \,:\, \sigma_j^V > \varepsilon\}, \tag{10}$$

$$n_W := \max\{j \,:\, \sigma_j^W > \varepsilon\} \tag{11}$$

and finally set

$$V(\hat{\mu}) := [\hat{v}_1, \ldots, \hat{v}_n] \qquad \text{and} \qquad W(\hat{\mu}) := [\hat{w}_1, \ldots, \hat{w}_n] \tag{12}$$

where $\hat{v}_j, \hat{w}_j$ denote the columns of $\widehat{V}$ and $\widehat{W}$, respectively.

The resulting reduced dimension n strongly depends on the number of chosen sample parameters R. Compared to the interpolation approach in Sect. 3.1 the reduced dimension will be larger for increasing R.

One benefit of the superposition approach is that if two different parameter values are related to the same closest points, the same subspaces are spanned by the columns of the concatenated bases in (7). Therefore only the SVDs (8) are required for determination the projection matrices in (12). Whereas in the interpolation approach the Lagrange basis functions in (4) must be determined for each value separately so the SVDs in (5) cannot be reused.

4 Numerical Simulations

We apply the two approaches above to a benchmark example called anemometer, see [8], which can also be found in [7]. An anemometer is a device which is used for measuring flow velocities. It consists of one heating component and two temperature sensors. The convection diffusion equation is solved by a semi-discretization, which yields a linear dynamical system of the form (1). The system is a *single input single output* system, i.e., $K = M = 1$ and it has dimension $N = 29008$. We consider

the case of three parameters $\mu = (c, c\,v, \kappa) \in [0,1] \times [0,2] \times [1,2]$. Then only the matrices $A(\mu)$ and $E(\mu)$ are parameter dependent, while the matrices B, C are constant.

For the computations of the local bases the domain of each parameter was discretized by equidistant grids with 3 points for c, 3 points for $c\,v$ and 18 points for κ. The resulting set of samples Π_S contains every combination of these discrete values and has cardinality $Q := |\Pi_S| = 126$, because $c = 0$ indicates $c\,v = 0$ so there are redundant combinations which can be omitted. For each sample point $\mu_j \in \Pi_S$ local projection matrices V_j, W_j were computed using the tangential interpolation method in [3] with 80 imaginary frequency points $s = \mathrm{i}\omega$, $\omega \in [10^0, 10^5]$ which are logarithmically distributed on the imaginary axis. For the determination of the projection matrices $V(\hat{\mu})$ and $W(\hat{\mu})$ for some given parameter value $\hat{\mu} \neq \mu_1, \ldots, \mu_Q$ we take the $R = 4$ matrices which are associated to the closest samples for both the interpolation approach and the superposition approach.

Since the interpolation approach can only be applied if the local bases have the same dimension the local bases were truncated to the minimum of the reduced dimensions. Similar to the procedure in Sect. 3.2, see Eqs. (9)–(11), we applied thin SVDs and take the first n left singular vectors which are associated to the singular values larger than $\varepsilon = 10^{-12}$. The SVDs were already done during the computation of the local models. Thus the only procedure to be done is comparing the dimensions n_j of V_j and truncate

$$V_j := \left[v_1^{(j)}, \ldots, v_n^{(j)} \right] \qquad \text{for } j = 1, \ldots, 4$$

where $v_i^{(j)}, i = 1, \ldots, n$ denote the columns of V_j, $j = 1, \ldots, 4$, and $n := \min_{1 \leq j \leq 4} n_j$.

We tested both the interpolation approach and the superposition approach by evaluating the resulting transfer functions for 26,130 test parameters $\hat{\mu}^{(k)} \in \Pi$ where the domain for each component of μ was discretized by 30 equidistant grid points. These values are compared to the values of the transfer function of the full-order model by determining the relative error, respectively, for 100 frequency points $s = \mathrm{i}\omega$ for $\omega \in [10^{-2}, 10^5]$ which form a logarithmic grid in the given domain.

Figure 1 shows the histograms of the occurring maximum relative errors for each tested point with respect to all frequency points. To illustrate the resulting order of the maximum relative errors the ordinate is logarithmically scaled meaning that the jth bin counts the occurrences maximum relative errors in $\left[10^{j-1}, 10^j\right]$ for $j = -9, \ldots, 2$. On the right-hand side of Fig. 1 we see that in most of the test points the maximum relative errors are lower than 10^{-2}. More precisely, for about 16,000 tested points the maximum relative error was smaller then 10^{-4} and for about 2000 even smaller than 10^{-6}. Whereas on the left-hand side for less than 2000 of the tested parameter values the relative error is smaller than 10^{-1}. The mean values of the maximum relative errors are 1.57 at a mean reduced dimension of 42.9 for the interpolation approach and $8.63 \cdot 10^{-4}$ at a mean reduced dimension of 174.2 for the superposition approach.

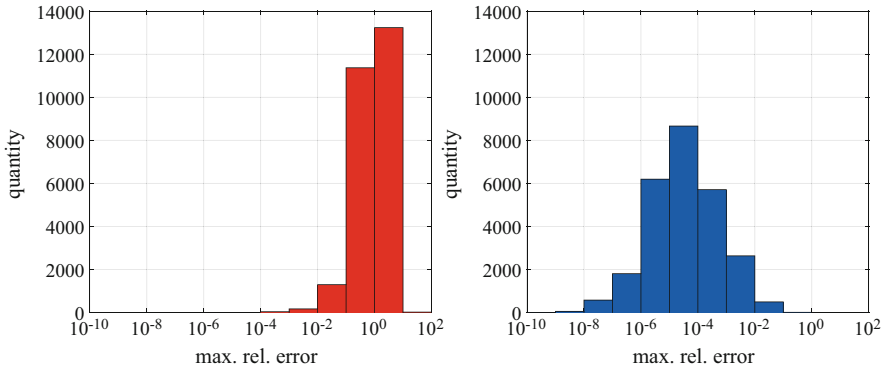


Fig. 1 Histograms of maximum relative errors using the interpolation approach (*left*) and the superposition approach (*right*) for 26,130 test parameter samples

One reason for the better results in the second method might be that the reduced systems using the superposition approach have about four times as high as the dimension of the reduced systems of the interpolation approach. Another reason might be that in the superposition approach we used two-sided methods to compute the local models whereas the interpolation approach requires the Galerkin method.

5 Conclusions

The presented superposition approach yield lower errors for the most tested sample points. The final dimension of the reduced systems are higher than in the first approach which means higher computational costs for the evaluation of the reduced systems. However, the higher costs yield significantly better approximations so there is potential for the second approach.

References

1. Ansalem, D., Farhat, C.: Interpolation method for adapting reduced-order models and application to aeroelasticity. AIAA J. **46**(7), 1803–1813 (2008)
2. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM Publications, Philadelphia (2005)
3. Baur, U., Beattie, D., Benner, P., Gugercin, S.: Interpolatory projection methods for parameterized model reduction, SIAM J. Sci. Comput. **33**, 2489–2518 (2011)
4. Benner, P., Gugercin, S., Willcox, K.: A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Rev. **57**(4), 483–531 (2015)
5. Freund, R.: Model reduction methods based on Krylov subspaces. Acta Numer. **12**, 267–319 (2003)

6. Lohmann, B., Eid, R.: Efficient order reduction of parametric and nonlinear models by superposition of locally reduced models. In: Roppenecker, G., Lohmann, B. (eds.) Methoden und Anwendungen der Regelungstechnik. Erlangen-Münchener Workshops 2007 und 2008. Shaker, Aachen (2009)
7. Moosmann, C.: ParaMOR - Model Order Reduction for parameterized MEMS applications. PhD thesis, University of Freiburg, Germany (2007)
8. http://morwiki.mpi-magdeburg.mpg.de/morwiki/index.php/Anemometer. Cited 20 Dec 2016

A Preliminary Statistical Evaluation of GPS Static Relative Positioning

M. Filomena Teodoro and Fernando M. Gonçalves

Abstract The objective of this work is to evaluate GPS static relative positioning (Hofmann-Wellenhof et al, GNSS-Global Navigation Satellite Systems GPS, GLONASS, Galileo, and more. Springer Verlag-Wien, New York, 2008; Kaplan and Hegarty, Understanding GPS: Principles and Applications. Artech House, Norwood, 2006; Leick, GPS Satellite Surveying. John Wiley & Sons, New Jersey, 2004), regarding accuracy, as the equivalent of a Real Time Kinematic (RTK) network and to address the practicality of using either a continuously operating reference stations (CORS) or a passive control point for providing accurate positioning control. The precision of an observed 3D relative position between two global navigation satellite systems (GNSS) antennas, and how it depends on the distance between these antennas and on the duration of the observing session, was studied. We analyze the performance of the software for each of the six chosen ranges of length in each of the four scenarios created, considering different intervals of observation time. The relation between observing time and baseline length is established. In this work are applied different statistical techniques, such as data analysis and elementary/intermediate inference level techniques (Tamhane and Dunlop, Statistics and Data Analysis: From Elementary to Intermediate. Prentice Hall, New Jersey, 2000) or multivariate analysis (Turkman and Silva, Modelos Lineares Generalizados da teoria a prática. Sociedade Portuguesa de Estatística, Lisboa, 2000; Anderson, An Introduction to Multivariate Analysis. Jonh Wiley & Sons, New York, 2003).

M.F. Teodoro (✉)
CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, Almada, Portugal

CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Lisbon, Portugal
e-mail: maria.alves.teodoro@marinha.pt

F.M. Gonçalves
NGI, Nottigham Geospatial Institute, University of Nottingham, Nottingham, UK
e-mail: geofactor@hotmail.com

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_110

1 Introduction

RTK networks are common in Europe but this is not the case in emerging economies where huge construction projects are running requiring geodetic support. In such cases, the easiest way to ensure that kind of support still is the static relative positioning using a single reference station. This technique provides surveyors the ability to determine the 3D coordinates of a new point with centimeter-level accuracy relative to a control point located several hundred kilometers away, which in turn can be associated with another GNSS receiver of a CORS operated by some institution.

Today the global navigation satellite systems play a fundamental role in the way that surveyors measure positional coordinates. It is now possible to determine the 3D coordinates of a new point with centimeter-level accuracy relative to a control point located several hundred kilometers away, which in turn can be associated with another GNSS receiver of a CORS operated by some institution. Examples of such networks are the ordnance survey (OS) Network across the UK [7] or, globally, the International GNSS Service Network [4].

In this research the coordinates of the OS active stations were used as 'true' values to address the practicality of using either a CORS or a passive control point for providing accurate positioning control and, implicitly, the performance of the software used. The precision of an observed 3D relative position between two GNSS antennas, and how it depends on the distance between these antennas and on the duration of the observing session, was studied. These results were attained through using commercial software LGO to process 105 single baselines, ranging from 61 to 898 km, according to observing sessions of varying lengths. ABEP was used as a reference station, with fixed coordinates, and the values obtained for the rover stations compared with those provided by OS. Also, to address the differences between using broadcast or precise ephemerides and computing the tropospheric effects or for simply applying a tropospheric model, the data processing was repeated for all different strategies.

Generally results show, whatever the strategy followed, that the length of the baseline matters, regarding the rate of successful baselines processed for a priori given values of 1D (ellipsoidal height accuracy) and 2D (compound of longitude and latitude accuracy). While distance matters, under the conditions of this experiment, the results also indicate that the duration of the observing session does not present the same pattern for 1D and 2D. In addition to the length of the baseline and the duration of the observing session, positioning precision depends on several other factors, including the methodology and the software used for processing GPS data, in this case the LGO. Biases associated with meteorological effects (ionosphere and troposphere) also play an important role in the total error budget of positioning precision.

This work investigates the performance of commercial software LGO when baselines are processed in static mode [10]. The parameter to be tested is the time of observation needed to achieve a given accuracy (1D and 2D) for a set of baseline length ranges. Four different scenarios were created, as follow:

- Broadcast ephemerides and Hopfield model (BH);
- Broadcast ephemerides and Computing the troposphere (BC);
- Precise ephemerides and Hopfield model (PH);
- Precise ephemerides and Computing the troposphere (PC).

Summarizing, the present work is comprised of introduction and conclusion sections, a section with background information, another describing the methodology adopted and two sections containing specific tests and results.

2 GNSS Overview

In this section, we provide an introduction of GPS, the navigation system used in this research. As there are a number of relevant references available, e.g. [3, 5, 6, 10], only a very brief discussion on the basics of the system will be given, with particular emphasis on the parts which are relevant to observation modeling of systematic biases and errors affecting GPS measurements. The various types of GPS observables of interest on baseline determination in static relative positioning are also described, as are some of their possible combinations. The possible usefulness of Precise Ephemerides, in terms of the increased accuracy in long baselines, is also evaluated.

There are numerous sources of measurement errors that influence GPS performance. Both observables types, code and phase, are affected by many systematic biases and errors, different in their source and suitable method of treatment. The most important of these biases and errors are briefly reviewed here. The orbital errors and tropospheric effects will be discussed later with more detail.

The sum of all systematic biases and errors contributing to the measurement error is referred to as a range bias. In [2] Bingley argues that this bias is caused by a physical phenomenon, as is the case, for example, in ionospheric or tropospheric delays, and error is the quantity remaining after the bias has been mitigated to some extent, which is the case, for example, for errors in broadcast ephemerides. According to the same author, the systematic biases and errors affecting GPS measurements can be grouped into three main categories: satellite related, atmospheric related and station related.

3 The Data

OS active stations were used to investigate the relation between time of observation and length of the baseline. A total of 105 baselines were processed using LGO, separated into six range groups ($R_i,\ i = 1, \ldots, 6,$) according with their lengths in kilometers:

- $R_1 = [000 - 100] \rightarrow$(5 baselines)
- $R_2 = [100 - 200] \rightarrow$(14 baselines)
- $R_3 = [200 - 300] \rightarrow$(27 baselines)
- $R_4 = [300 - 400] \rightarrow$(29 baselines)
- $R_5 = [400 - 500] \rightarrow$(14 baselines)
- $R_6 = [500 - 900] \rightarrow$ (16 baselines)

All the stations are permanent stations of clear sky visibility and with low multipath conditions. The quality of the data is therefore expectedly high. Day 13/06/2013 of receiver independent exchange (RINEX) data of GPS week 1744 was downloaded from the data archive of the active GPS network of Ordnance Survey (OS Net) for each of the 106 stations (http://www.ordnancesurvey.co.uk/gps/os-net-rinex-data/). These RINEX data include phase measurement of the carrier waves L_1 and L_2, P_1, P_2 and C/A pseudo-range code at a 30 s interval.

For this experiment, 24 h of dual-frequency GPS carrier phase observations for each of 105 baselines formed by ABEP, chosen as reference station, and all the other active stations, designated as rover, from OS Network were used. These 105 baselines range in length from 61 *km* to 898 *Km* and correspond to all active stations considered 'healthy' on the 13^{th} June 2013. The data for each baseline comprised the same 24 h session that was further subdivided into periods of time of 1, 2, 3, 4, 6, 8, 12 and 24 h as follow, where the two first digits represent the beginning of the observation period and the last two the end:

- 1 h periods: [0001], [0607], [1213], [1819];
- 2 h periods: [0002], [0608], [1214], [1820];
- 3 h periods: [0003], [0609], [1215], [1821];
- 4 h periods: [0004], [0408], [0812], [1216], [1620], [2024];
- 6 h periods: [0006], [0612], [1218], [1824];
- 8 h periods: [0008], [0816], [1624];
- 12 h periods: [0012], [1224];
- 24 h period: [0024].

The division of time in this way was done in order to evaluate the performance of the software for different lengths of observation time and for similar lengths but at different times of the day experiencing diverse atmospheric conditions.

A preliminary experiment shows that to obtain high accurate relative positioning 3D coordinates for long baselines in static mode with LGO at least 4 h of observation are recommended. Therefore, special focus was given to periods of this magnitude and over. These cover the whole day in non overlapping periods, whereas for the 1, 2 and 3 h intervals only representative samples were chosen.

The criteria followed to select the reference station were primarily based on location. Thus ABEP, on the west coast of England, was chosen, because of its high altitude and location, providing a well distributed range of radial vectors to all the other active stations, either in latitude and longitude. Its 3D positional coordinates were fixed to the official values adopted by OS.

In order to evaluate at what range of baseline lengths the use of precise ephemerides become worthwhile, both results using broadcast and precise ephemerides are presented as well. The corresponding SP3 files were downloaded from the data archive of IGS (http://igscb.jpl.nasa.gov/components/prods_cb.html). These include precise ephemerides at a sampling interval of 15 min and the high-rate precise satellite clocks with a sampling of 30 s.

Hence, the four different scenarios can be compared as follow: direct comparison of the results obtained using the broadcast ephemerides and the precise ephemerides (BH versus PH and BC versus PC); direct comparison of the results obtained using Hopfield model and computing the troposphere (BH versus BC and PH versus PC).

At starting points 1D, 2D and 3D accuracy criteria were established for each baseline, as only successful processed baselines are of interest for this research. The chosen values were set to 1D and 2D accuracies to be better than 3 cm and 3D better than 4.5 cm. These are realistic values, as the OS active stations have 1D accuracy of about 2 cm in magnitude and close to 1 cm in 2D. Therefore, assuming the 3 cm as 1D and 2D threshold seems to be reasonable due the fact that this tolerance allows for the 'absorption' of errors inherent to the coordinates of the stations. Despite how perfectly the baseline was calculated an error of up to 4 cm in height and 2 cm in plan could arise due to the uncertainty associated with the coordinates.

The published coordinates of each of these stations (in Cartesian format on the header of the corresponding RINEX file) are assumed as 'true' and used to compute the errors (1D, 2D and 3D) in the solutions processed by LGO.

In Fig. 1 we present the percentage of successful baselines in 1D (dashed lines) and 2D (full lines). There is a clear trend for fewer successful baselines as the length increases, regardless of the strategy adopted, either in 1D or 2D.

In a preliminary approach, it was found that the different ranges led to significantly different results. Were used parametric tests to compare proportions (t-test). With some small samples in certain ranges, were also applied some nonparametric tests that allow us to compare location measures, or a chi-square test and a Kruskal-Wallis to evaluate if the proportions of success are the same in the different ranges; a chi-square independence test were also used to evaluate the relation between the proportion of success and range. In Table 1 are described the $p - values$ obtained when it is tested the difference of the proportion of success for different ranges.

In general, different strategies conduce to similar results: almost all comparisons have the same conclusion—the proportions of success in different ranges are not equal except when the ranges are sequential of each other. Also were performed similar tests comparing different strategies considering the same range. Generally, the proportions of success for the same range, but with different strategies conducted to significant tests, meaning that there is statistical evidence of different proportions of success per different strategies for same range.

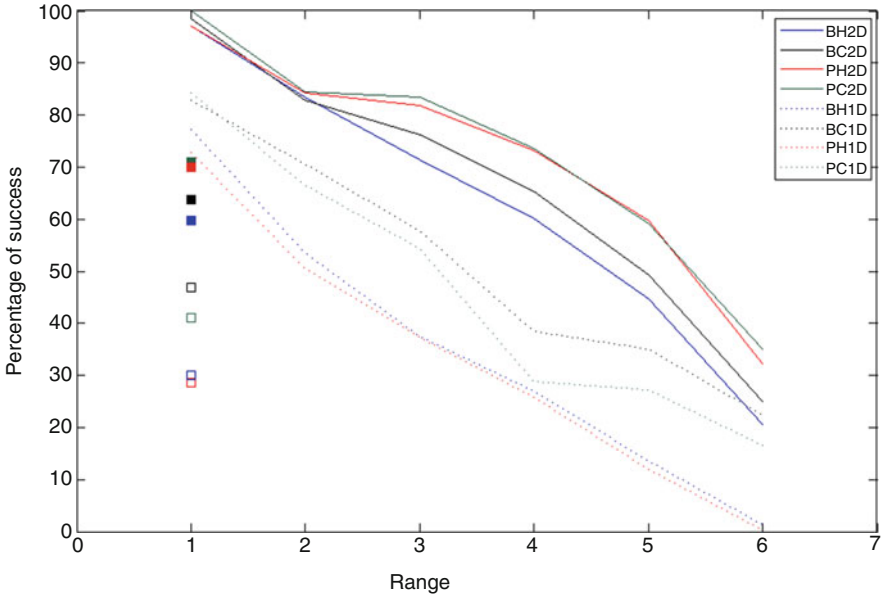


Fig. 1 Percentage of successful baselines in 1D (*dashed lines*) and 2D (*full lines*)

In Fig. 2 the yellow triangles represent successful baselines in 1D and 2D simultaneously; the red triangles represent successful baselines only in 1D considering the 24 h session. Figure 2 (left) shows that, considering the 24 h session, Hopfield is only acceptable up to a certain distance (Wales and Southern England). On the other hand, the computed option is acceptable over greater distances (Wales, Southern England, North England and Southern Scotland), as can be seen in Fig. 2 (right).

4 Conclusions

This work studied the relation for single baselines between lengths ranges an between the different ranges and the observation time required to obtain high-accurate positioning, using commercial software LGO. The analysis for different observation time is not reproduced in this paper. The results are valid for this specific software and under the conditions of the experiments. Four different strategies were established and evaluated through the processing of a total of 11,760 baselines. The data processing and testing used several options concerning the best thresholds for accuracy. The LGO results were compared with the published coordinates by Ordnance Survey and the baselines passing the accuracy criteria were isolated. Other test and techniques were developed, inquiring about the significance of the hour of the day, the amplitude of time interval of exposure, considering the four strategies.

Table 1 Tests of proportions of success difference for distinct ranges. Strategies[a] BH, BC, PH and PC

Strategies	Ranges	R_2	R_3	R_4	R_5	R_6
BH	R_1	0.3204	0.0679	0.0204	0.0073	0.0008
	R_2		0.3271	0.0988	0.0200	0.0007
	R_3			0.4097	0.0745	0.0006
	R_4				0.2767	0.0052
	R_5					0.2140
BC	R_1	0.5652	0.2053	0.0267	0.0367	0.0065
	R_2		0.4100	0.0395	0.0530	0.0054
	R_3			0.1452	0.1585	0.0158
	R_4				0.8250	0.2491
	R_5					0.4492
PH	R_1	0.3640	0.1159	0.0357	0.0122	0.0017
	R_2		0.4222	0.1208	0.0229	0.0008
	R_3			0.3533	0.0538	0.0003
	R_4				0.2541	0.0033
	R_5					0.1900
PC	R_1	0.4018	0.1237	0.0048	0.0116	0.0018
	R_2		0.4451	0.0168	0.0320	0.0034
	R_3			0.0502	0.0844	0.0071
	R_4				0.9174	0.3316
	R_5					0.4812

[a]Four strategies considering 1D

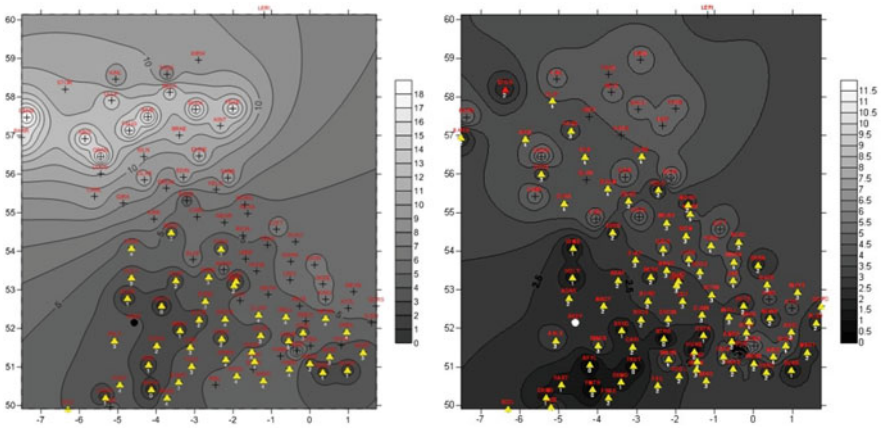


Fig. 2 Successful baselines simultaneously in 1D and 2D (*yellow*) or just in 1D (*red*). Strategy: BH (*left*), BC (*right*)

MANOVA was applied with several factors such as range, strategies, amplitude of interval time exposure [8]. Also was developed a GLM model [1, 9]. Such statistical approach details will be found in a future extended version of this manuscript.

Acknowledgements This work was supported by Portuguese funds through the *Center for Computational and Stochastic Mathematics* (CEMAT), *The Portuguese Foundation for Science and Technology* (FCT), University of Lisbon, Portugal, project UID/Multi/04621/2013, and *Center of Naval Research* (CINAV), Naval Academy, Portuguese Navy, Portugal.

References

1. Anderson, T.W.: An Introduction to Multivariate Analysis. Jonh Wiley & Sons, New York (2003)
2. Bingley, R.M.: GNSS Principles and Observables: Systematic Biases and Errors. Short Course, University of Nottingham, Nottingham Geospatial Institute, Nottingham (2013)
3. Hofmann-Wellenhof, B., Lichtenegger, H., Wasle, H.: GNSS-Global Navigation Satellite Systems GPS, GLONASS, Galileo, and more. Springer Verlag-Wien, New York (2008)
4. IGS: International GNSS Service. http://igscb.jpl.nasa.gov/ (2016). Cited 19/08/2016
5. Kaplan, E.D., Hegarty, C.J.: Understanding GPS: Principles and Applications. Artech House, Norwood (2006)
6. Leick, A.: GPS Satellite Surveying. John Wiley & Sons, New Jersey (2004)
7. OS Net Business and Government: Ordnance Survey. http://www.ordnancesurvey.co.uk/oswebsite/products/os-net/index.html (2016). Cited 19/08/2016
8. Tamhane, A.C., Dunlop, D.D.: Statistics and Data Analysis: From Elementary to Intermediate. Prentice Hall, New Jersey (2000)
9. Turkman, M.A., Silva, G.: Modelos Lineares Generalizados da teoria a prática. Sociedade Portuguesa de Estatística, Lisboa (2000)
10. Wells, D.E., et al.: Guide to GPS Positioning. Canadian GPS Associates, Fredericton. http://plan.geomatics.ucalgary.ca/papers/guide_to_gps_positioning_book.pdf (1986). Cited 1/10/2016

Numerical Simulation of Flow Induced Vocal Folds Vibration by Stabilized Finite Element Method

Jan Valášek, Petr Sváček, and Jaromír Horáček

Abstract This paper is interested in the numerical simulation of human vocal folds vibration induced by fluid flow. A two-dimensional problem of fluid-structure interaction is mathematically formulated. An attention is paid to the description of a robust method based on finite element approximation. In order to capture the typical flow velocities involved in the physical model resulting in high Reynolds numbers, the modified Streamline-Upwind/Petrov-Galerkin stabilization is applied. The mathematical description of the considered problem is presented, where the arbitrary Lagrangian-Euler method is used to describe the fluid flow in time dependent domain. The viscous incompressible fluid flow and linear elasticity models are considered. Further, the numerical scheme is described for the fluid flow and the elastic body motion. The implemented coupling procedure is explained. The numerical results are shown.

1 Introduction

This contribution deals with the numerical simulation of flow induced vocal folds vibration. It is one specific example of wide class of fluid-structure interaction problem (FSI), which can be plentifully found in nature as well as in technical applications, see e.g. [2].There are many possibilities how to solve this problem, see e.g. [4]. Along many others the partitioned approach with finite element method (FEM) as basic dicretization procedure has gained bigger popularity, see [4, 7]. It consists of decoupling of the FSI problem into elastic and fluid parts, which are then solved in sequential order coupled by boundary conditions until

J. Valášek (✉) • P. Sváček
Faculty of Mechanical Engineering, CTU in Prague, Karlovo nam. 13, 121 35 Praha 2, Czech Republic
e-mail: valasek.jan@volny.cz; petr.svacek@fs.cvut.cz

J. Horáček
Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Dolejškova 5, 182 00 Praha 8, Czech Republic
e-mail: jaromirh@it.cas.cz

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*, Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_111

a convergence criteria in each time step is reached. Although subiterations are practically necessary, the convergence can be often fast, see [4].

The essential part of FSI is to incorporate time varying fluid domain. The most popular method is the arbitrary Lagrangian Eulerian (ALE) method, which is easy to implement, see [8]. The main attention of this contribution is paid to the stabilization of the FEM in order to overcome the possible numerical instabilities due to high Reynolds number flows. The streamline upwind/Petrov-Galerkin (SUPG), pressure-stabilization/Petrov-Galerkin (PSPG) together with 'div-div' stabilization are applied to enable simulation with higher velocities and to obtain more accurate results, see [5]. The dominant frequencies of obtained numerical results are analyzed with Fourier transformation.

2 Mathematical Model

We consider 2D model problem of a fluid flow interacting with an elastic body shown in Fig. 1. The domain Ω^s_{ref} represents the undeformed structure (vocal fold) with the fixed bottom part Γ^s_{Dir} and the common interface Γ_{W} between the structure and the fluid. The fluid occupies domain Ω^f_{ref} at the beginning but due to the structure deformation the domain changes to Ω^f_t at time t. The interface is deformed to $\Gamma_{\mathrm{W_t}}$ at time t, has also t-index. The other fixed parts of boundary $\partial\Omega^f_t$ are the channel wall Γ^f_{Dir}, the inlet part Γ^f_{In} and the outlet part Γ^f_{Out}.

ALE Method The ALE method describes the time development of the fluid domain with the aid of a smooth mapping A_t. This diffeomorphism of the reference domain Ω^f_{ref} onto the deformed one Ω^f_t at arbitrary time instant t is moreover assumed to satisfy the following conditions $\frac{\partial A_t}{\partial t} \in C(\overline{\Omega^f_{ref}}), \quad A_t(\partial\Omega_{ref}) = \partial\Omega^f_t,\ t \in [0, \mathrm{T}]$. The ALE domain velocity $\mathbf{w}_D$, which express the velocity of the point with a

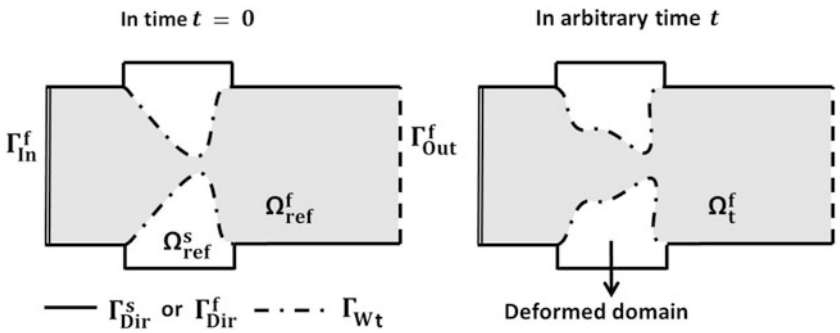


Fig. 1 The scheme of vocal fold model with marked boundaries. On the *left* the reference fluid and structure configurations are shown. On the *right* the deformed structural and fluid domains are depicted

given reference $X \in \Omega_{ref}$, is then defined by

$$\mathbf{w}_D(x,t) = \hat{\mathbf{w}}_D(A_t^{-1}(x),t); \quad \hat{\mathbf{w}}_D(X,t) = \frac{\partial}{\partial t}A_t(X), \quad t \in (0,\mathrm{T}),\ x \in \Omega_t^f. \tag{1}$$

Similarly, the ALE derivative of a function $f(x,t)$ is derivative with respect to a reference point $X \in \Omega_{ref}^f$ and it satisfies the relation

$$\frac{D^A f}{Dt}(x,t) = \frac{\partial f}{\partial t}(x,t) + \mathbf{w}_D(x,t)\cdot\nabla f(x,t). \tag{2}$$

For more details or practical construction of mapping A_t respectively, see [8] or [4, 9] resp.

Fluid Flow The motion of viscous incompressible fluid in time dependent domain is described by Navier-Stokes equations (here in the ALE form)

$$\frac{D^A\mathbf{v}}{Dt} + ((\mathbf{v}-\mathbf{w}_D)\cdot\nabla)\mathbf{v} - \nu^f\Delta\mathbf{v} + \nabla p = \mathbf{0}, \qquad \operatorname{div}\mathbf{v} = 0 \quad \text{in } \Omega_t^f, t \in (0,\mathrm{T}), \tag{3}$$

where $\mathbf{v}(x,t)$ denotes the fluid velocity, $p(x,t)$ is the kinematic pressure and ν^f is the kinematic fluid viscosity.

Equations (3) are enclosed with initial and boundary conditions

$$\text{(a)} \qquad \mathbf{v}(x,t) = \mathbf{v}_{\mathrm{Dir}}(x,t) \qquad \text{for } x \in \Gamma_{\mathrm{In}}^f \cup \Gamma_{\mathrm{Dir}}^f,\ t \in (0,\mathrm{T}), \tag{4}$$

$$\text{(b)} \quad p(x,t)\mathbf{n}^f - \nu^f\frac{\partial\mathbf{v}}{\partial\mathbf{n}^f}(x,t) = -\frac{1}{2}\mathbf{v}(\mathbf{v}\cdot\mathbf{n}^f)^- \qquad \text{for } x \in \Gamma_{\mathrm{Out}}^f,\ t \in (0,\mathrm{T}),$$

where $\mathbf{n}^f = (n_1^f, n_2^f)$ denotes outer unit normal to boundary $\partial\Omega^f$. The condition (4 b) is the modified do-nothing boundary condition, see [1]. This boundary condition increases the stability of the computation in the case of occurrence of backward inlet through the outlet part of the boundary. The boundary condition prescribed on $\Gamma_{\mathrm{W_t}}$ is given in next paragraphs.

Structure Motion The vocal fold is modelled as an elastic body whose motion is governed by the partial differential equations

$$\rho^s\frac{\partial^2 u_i}{\partial t^2} - \frac{\partial\tau_{ij}^s}{\partial X_j} = f_i^s \quad \text{in } \Omega^s \times (0,\mathrm{T}), \tag{5}$$

where ρ^s is the structure density, $\mathbf{u}(X,t) = (u_1,u_2)$ is the sought displacement, $\mathbf{f^s} = (f_1^s, f_2^s)$ denotes the acting volume force and $\boldsymbol{\tau}^s = (\tau_{ij}^s)$ is the Cauchy stress tensor. It can be expressed by the generalized Hooke's law for isotropic material as

$$\tau_{ij}^s = \lambda^s(\operatorname{div}\mathbf{u})\,\delta_{ij} + 2\mu^s e_{ij}^s(\mathbf{u}), \tag{6}$$

where λ^s, μ^s are Lame constants, e^s_{ij} are components of the strain tensor and δ_{ij} is Kronecker delta. Under the assumption of small displacement we set $e^s_{jk}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_j}{\partial X_k} + \frac{\partial u_k}{\partial X_j}\right)$. Equations (5) are completed by suitable initial conditions, see [9]. The prescribed boundary conditions are following

$$\text{(a)} \qquad \mathbf{u}(X,t) = \mathbf{0} \quad \text{for } X \in \Gamma^s_{\text{Dir}},\ t \in (0,\text{T}), \tag{7}$$

$$\text{(b)} \qquad \tau^s_{ij}(X,t)\, n^s_j(X) = q^s_i(X,t), \qquad \text{for } X \in \Gamma_{\text{W}},\ t \in (0,\text{T}),$$

where $\mathbf{n}^s = (n^s_1, n^s_2)$ is the unit outer normal to the boundary $\partial\Omega^s$ and the definition of q^s_i is explained in next paragraph.

Interface Conditions The FSI problem is coupled via boundary conditions on the common interface $\Gamma_{\text{W}_\text{t}}$. First, the location of the interface depends on the establishing force equilibrium between the aerodynamic and elastic forces and it is implicitly given by the deformation $\mathbf{u}$, see [9].

Further, the impact of the aerodynamic forces on the structure is expressed in the form of Neumann boundary condition (7) with the q^s_i given by

$$q^s_i(X,t) = \sum_{j=1}^{2} \rho^f \left(p\,\delta_{ij} - \nu^f \left(\frac{\partial \mathbf{v}_i}{\partial x_j} + \frac{\partial \mathbf{v}_j}{\partial x_i} \right) \right) n^f_j. \tag{8}$$

The correspondence between fluid and elastic body motion on boundary $\Gamma_{\text{W}_\text{t}}$ (where $\mathbf{u} = \mathbf{w}_D$) enforces the Dirichlet boundary condition prescribed for fluid

$$\mathbf{v}(x,t) = \mathbf{w}_D(x,t) = \frac{\partial \mathbf{u}}{\partial t}(X,t) \qquad \text{for } x \in \Gamma_{\text{W}_\text{t}},\ t \in (0,\text{T}). \tag{9}$$

3 Numerical Model

The numerical approximation of both subproblems (3) and (5) is treated with the aid of FEM. The time discretization is based on the equidistant time stepping with time step Δt.The approximation of the solution $\mathbf{u}, \mathbf{v}, p$ at time $t_n = n\Delta t$ is denoted by $\mathbf{u}^n, \mathbf{v}^n, p^n$.

Elastic Body Equation (5) is spatially reformulated in a weak sense in standard way, see [9]. Then the solution is sought in the finite element space, i.e. the numerical solution is expressed as a linear combination of basis functions $\mathbf{u}_h(x,t) = \sum_{j=1}^{N_h} \alpha_j(t)\boldsymbol{\varphi}_j(x)$. This results in the system of ordinary differential equations of second order for the unknown coefficients $\alpha_j(t)$

$$\mathbb{M}^T\ddot{\alpha} + \mathbb{K}^T\alpha = \mathbf{b}(t), \tag{10}$$

where the vector $\mathbf{b}(t)$ has components $b_i(t) = (\mathbf{f}^s, \boldsymbol{\varphi}_i)_{\Omega^s} + (\mathbf{q}^s, \boldsymbol{\varphi}_i)_{\Gamma^s_{\mathrm{Neu}}}$ and the elements of matrices $\mathbb{M} = (m_{ij}), \mathbb{K} = (k_{ij})$ are given by

$$m_{ij} = (\rho^s \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)_{\Omega^s}, \quad k_{ij} = (\lambda^s (\mathrm{div}\, \boldsymbol{\varphi}_j)\, \delta_{rl} + 2\mu^s e^s_{rl}(\boldsymbol{\varphi}_j), e^s_{rl}(\boldsymbol{\varphi}_i))_{\Omega^s}. \tag{11}$$

The notation $(\cdot,\cdot)_{\mathscr{D}}$ means scalar product in the space $L^2(\mathscr{D})$ or $\mathbf{L}^2(\mathscr{D})$. System (11) is numerically solved by the Newmark method, see [3].

Flow Motion For the sake of clarity let us restrict here to the case of discretization at (arbitrary) fixed time step $t = t_{n+1}, \Omega^f = \Omega^f_{t_{n+1}}$. Further, in Eq. (3) we replace the ALE derivative by the second order backward differentiation formula and the convective term is linearized by $(\mathbf{v}\cdot\nabla)\mathbf{v}|_{t_{n+1}} \approx (\mathbf{v}^n \cdot \nabla)\mathbf{v}^{n+1}$. Then the weak formulation of problem (3) according to [1, 9] is introduced: find solution $V = (\mathbf{v}, p) = (\mathbf{v}^{n+1}, p^{n+1})$ such that $\mathbf{v}$ satisfies boundary conditions (4 a, b), (9) and

$$a(V, \Phi) = f(\Phi) \qquad \text{for all } \Phi = (\boldsymbol{\varphi}, q) \in \mathbf{X} \times L^2(\Omega^f). \tag{12}$$

Here, the space $\mathbf{X} = \{\boldsymbol{\psi} \in \mathbf{W}^{1,2}(\Omega^f_t) |\ \boldsymbol{\psi} = \mathbf{0} \text{ on } \Gamma^f_{\mathrm{Dir}} \cup \Gamma^f_{\mathrm{In}} \cup \Gamma^f_{\mathrm{W_t}}\}$ and the form f is given by $f(\Phi) = \left(\frac{4\overline{\mathbf{v}}^n - \overline{\mathbf{v}}^{n-1}}{2\Delta t}, \boldsymbol{\varphi}\right)_{\Omega^f}$ with the used notation $\overline{\mathbf{v}}^i(x) = \mathbf{v}^i(A_{t_i}(A^{-1}_{t_{n+1}}(x)))$. The bilinear form $a(\cdot,\cdot)$ reads

$$a(V,\Phi) = \left(\frac{3\mathbf{v}}{2\Delta t}, \boldsymbol{\varphi}\right)_{\Omega^f} + \frac{1}{2}((\overline{\mathbf{v}}^n - 2\mathbf{w}_D)\cdot\nabla)\mathbf{v}, \boldsymbol{\varphi})_{\Omega^f} - \frac{1}{2}((\overline{\mathbf{v}}^n\cdot\nabla)\boldsymbol{\varphi}, \mathbf{v})_{\Omega^f} +$$
$$+\frac{1}{2}\left((\overline{\mathbf{v}}^n \cdot \mathbf{n}^f)^+ \mathbf{v}, \boldsymbol{\varphi}\right)_{\Gamma^f_{\mathrm{Out}}} + \nu^f(\nabla\mathbf{v}, \nabla\boldsymbol{\varphi})_{\Omega^f} - (p, \mathrm{div}\,\boldsymbol{\varphi})_{\Omega^f} + (q, \mathrm{div}\,\mathbf{v})_{\Omega^f}. \tag{13}$$

Stabilization The high Reynolds numbers flows is difficult to numerically approximate partially due to dominating convection. Also the finite element solution can be numerically unstable in that case. This is principally caused by unresolved velocity gradients due to too coarse grid. The other source of instability is the necessity of satisfying the Babuška-Brezzi (BB) condition, see [6]. In order to overcome the possible numerical instability the residual based stabilization is applied, namely SUPG, PSPG and 'div-div' stabilization. These methods keep method stable, consistent and still accurate (not too diffusive), see e.g. [4, 5]. The stabilization is realized with the aid of additional terms included to weak formulation (12). The stabilized problem reads: find $V = (\mathbf{v}, p)$ such that

$$a(V,\Phi) + \sum_{K\in\mathscr{T}_h} \delta_K \left(\frac{3\mathbf{v}}{2\Delta t} + ((\overline{\mathbf{v}}^n - \mathbf{w}_D)\cdot\nabla)\mathbf{v} + \nabla p - \nu\Delta\mathbf{v},\ \boldsymbol{\zeta}(\Phi)\right)_K + \tag{14}$$
$$+ \sum_{K\in\mathscr{T}_h} \tau_K (\nabla\cdot\mathbf{v}, \nabla\cdot\boldsymbol{\varphi})_K = f(\Phi) + \sum_{K\in\mathscr{T}_h} \delta_K \left(\frac{4\overline{\mathbf{v}}^n - \overline{\mathbf{v}}^{n-1}}{2\Delta t},\ \boldsymbol{\zeta}(\Phi)\right)_K,$$

for all $\Phi = (\boldsymbol{\varphi}, q) \in \mathbf{X} \times L^2(\Omega^f)$, where short notation $\boldsymbol{\zeta}(\Phi) := ((\overline{\mathbf{v}}^n - \mathbf{w}_D) \cdot \nabla)\boldsymbol{\varphi} + \nabla q$ and $\mathscr{T}_h$ for regular triangulation was used. The parameters τ_K and δ_K are locally defined using the local element length h_K, see [4]. The BB stable P1-bubble/P1 element was implemented.

4 Results

Numerical simulation was performed on computational domain with vocal fold model M5 of the height 6 mm and the width of glottis 1 mm. The vocal folds were divided into two parts—epithelium (thin layer around interface) with $E^s = 100\,\text{kPa}$, $\sigma = 0.4$ and muscle (the rest of vocal fold) with $E^s = 12\,\text{kPa}$, $\sigma = 0.4$. The further parameters were set as follows: $\Delta t = 5 \cdot 10^{-5}\,\text{s}$, $\nu^f = 1.47 \cdot 10^{-5}\,\text{m/s}^2$, $\rho^f = 1.7\,\text{kg/m}^3$, $\rho^s = 1050\,\text{kg/m}^3$.

The FSI problem was solved with the prescribed inlet parabolic profile of three different maximal magnitudes: 1.5 and 2.4 and 3.0 m/s. Figure 2 illustrates typical distribution of the flow velocity magnitude during closing and opening motions. The jet and associated vortex structures arises as flow passes through glottis.

Figure 3 shows comparison of x-displacements of the point chosen from the top of the bottom vocal fold. In the same figure on the right the Fourier transformation

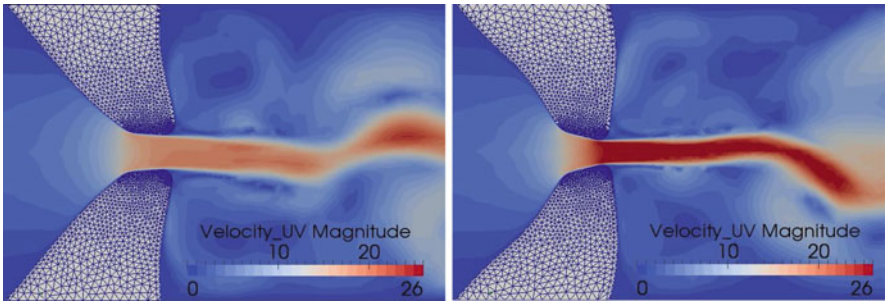


Fig. 2 The typical distribution of the flow velocity magnitude and the deformation of vocal fold for case $v = 3.0\,\text{m/s}$ in opened (*left*) and closed (*right*) positions

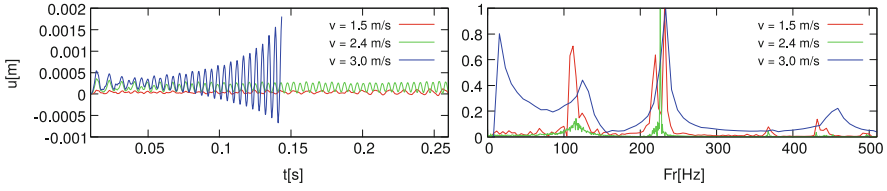


Fig. 3 On the *left*—time evolution of displacement of chosen point from the top of bottom vocal fold in x-direction. On the *right*—(normalized) Fourier transformation of time signal from the left

of the time signal from the left is plotted. In the case of $v = 3.0\,\mathrm{m/s}$ the fluid flow induced exponential growing vibration amplitude insofar, that computation shuttered due to too distorted mesh. In other two cases the vocal fold vibration was stable, only magnitudes were different. The frequency spectra reveal that in all cases the dominant frequency corresponding to the second eigenfrequency was excited. The different shape of curve in third case ($v = 3.0\,\mathrm{m/s}$) is caused by short time signal.

5 Conclusion

This contribution presents the mathematical description and FEM based numerical scheme for solution of FSI. One special section is dedicated to the stabilization of FEM by SUPG, PSPG and 'div-div' method. These methods introduce additional terms in weak formulation of the problem and help to obtain more accurate results for different inlet velocities. The simulation of induced vibration of vocal folds by fluid flow computed by in-house developed program is shown.

Acknowledgements This work was supported by grant No. GA 16-01246S of the Czech Science Foundation and by grant No. SGS16/206/OHK2/3T/12 of CTU in Prague.

References

1. Braack, M., Mucha, P.B.: Directional do-nothing condition for the Navier-Stokes equations. J. Comput. Math. **32**, 507–521 (2014)
2. Clark, R., Dowell, E.H.: A Modern Course in Aeroelasticity. Springer, Netherlands (2004)
3. Curnier, A.: Computational Methods in Solid Mechanics. Springer, Netherlands (1994)
4. Feistauer, M., Sváček, P., Horáček, J.: Numerical simulation of fluid-structure interaction problems with applications to flow in vocal folds. In: Bodnár, T., Galdi, G.P., Nečasová, S. (eds.) Fluid-Structure Interaction and Biomedical Applications, pp. 312–393. Birkhauser, Basel (2014)
5. Gelhard, T., Lube, G., Olshanskii, M.A., Starcke, J.H.: Stabilized finite element schemes with LBB-stable elements for incompressible flows. J. Comput. Appl. Math. **177**(2), 243–267 (2005). doi:10.1016/j.cam.2004.09.017
6. Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations. Springer, Berlin/Heidelberg (1986)
7. Sváček, P., Horáček, J.: Stabilized finite element approximations of flow over a self-oscillating airfoil. In: Kreiss, G., Ljastedt, P., Malqvist, A., Neytcheva, M. (eds.) Numerical Mathematics and Advanced Applications 2009, pp. 847–854. Springer, Berlin/Heidelberg (2010)
8. Takashi, N., Hughes, T.J.R.: An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. Comput. Methods Appl. Mech. Eng. **95**, 115–138 (1992)
9. Valášek, J., Sváček, P., Horáček, J.: Numerical solution of fluid-structure interaction represented by human vocal folds in airflow. In: Veselý M., Dančová P. (eds.) EFM15 Experimental Fluid Mechanics 2015, vol. 114. EPJ Web of Conferences (2016). doi:10.1051/epjconf/201611402130

Wiener Chaos Expansion for an Inextensible Kirchhoff Beam Driven by Stochastic Forces

Alexander Vibe and Nicole Marheineke

Abstract In this work we study the feasibility of the Wiener Chaos Expansion for the simulation of an inextensible elastic slender fiber driven by stochastic forces. The fiber is described as Kirchhoff beam with a 1d-parameterized, time-dependent curve, whose dynamics is given by a constrained stochastic partial differential equation. The stochastic forces due to a surrounding turbulent flow field are modeled by a space-time white noise with flow-dependent amplitude. Using the techniques of polynomial chaos, we derive a deterministic system which approximates the original stochastic equation. We explore the numerical performance of the approximation and compare the results with those obtained by Monte-Carlo simulations.

1 Introduction

Large systems of fibers subjected to stochastic forces occur in various applications. In non-woven manufacturing, for example, fibers are exposed to turbulent air flows which have different purposes depending on the actual production process, such as cooling, entangling, mixing, stretching and/or transporting the fibers. The material properties of the resulting non-woven fabric depend strongly on the stochastic nature of the air flow. Strategies for predicting these material properties usually involve numerical simulations of the fiber systems. Thus, it is of great interest to study how the respective computational costs might be reduced. The so-called Wiener Chaos Expansion (WCE) for stochastic differential equations yields an effective splitting between deterministic and stochastic effects, allowing an efficient computation of the mean behavior of the system under consideration. It is based on a Fourier expansion of the underlying white noise with respect to a deterministic basis, giving stochastic coefficients. Expanding the unknowns in polynomials of these white noise coefficients leads then to a deterministic system of differential equations which inherits the complete information of the stochastic system [2, 3, 6]. In this work, we explore the feasibility of a WCE-based numerical method to a Kirchhoff

A. Vibe (✉) • N. Marheineke
Department Mathematik, FAU Erlangen-Nürnberg, Cauerstr. 11, D-91058 Erlangen, Germany
e-mail: vibe@math.fau.de; marheineke@math.fau.de

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_112

(constrained) beam model under the influence of a planar white noise for stochastic inextensible fibers and compare its performance with Monte-Carlo simulations.

2 Fiber Model

An inextensible elastic fiber can be modeled as Kirchhoff beam by a one-dimensionally parametrized, time-dependent curve $\mathbf{r} : [0, L] \times \mathbb{R}_0^+ \to \mathbb{R}^d$, $d \in \{2, 3\}$ whose dynamics in a turbulent air flow is governed by a constrained partial differential equation with noise [7, 8], i.e.,

$$\partial_{tt}\mathbf{r} = \partial_s(\lambda \partial_s \mathbf{r}) - \alpha^{-2}\partial_{ssss}\mathbf{r} + \mathbf{f} + \mathbf{B} \cdot \boldsymbol{\xi}, \qquad \|\partial_s \mathbf{r}\|^2 = 1.$$

The unknown scalar-valued inner traction λ acts hereby as Lagrangian multiplier to the point-wise constraint of inextensibility. In the stated dimensionless form the bending number α is the ratio of inertial and bending forces. The external deterministic forces $\mathbf{f}$ due to gravity and aerodynamics are in general described by a space and time-dependent function of fiber position, velocity and orientation $\mathbf{f} = \mathbf{f}(\mathbf{r}, \partial_t \mathbf{r}, \partial_s \mathbf{r}, s, t)$. The turbulent effects are additionally incorporated by a space-time Gaussian white noise $\boldsymbol{\xi}$ with flow-dependent amplitude $\mathbf{B} = \mathbf{B}(\mathbf{r}, \partial_t \mathbf{r}, \partial_s \mathbf{r}, s, t)$. The matrix-vector product is denoted by $\cdot$. Since the feasibility of a WCE approach for the constrained stochastic system stands in the focus of this paper, we restrict here to a linear drag model and a linear flow field for simplicity. We particularly consider $\mathbf{f} = \mathrm{Fr}^{-2}\mathbf{e}_g + \mathrm{Dr}^{-2}(\mathbf{A} \cdot \mathbf{r} - \partial_t \mathbf{r})$ and $\mathbf{B} = \beta \mathrm{Dr}^{-2}\mathbf{I}$ with direction of gravity $\mathbf{e}_g$, $\|\mathbf{e}_g\| = 1$, a constant flow matrix $\mathbf{A}$ and the identity matrix $\mathbf{I}$. The dimensionless Froude Fr and drag Dr numbers give the ratios of inertial and gravitational forces as well as inertial and mean drag forces, β is the turbulent fluctuation number [5].

Definition 1 (Gaussian White Noise, Planar Wiener Process) [5, 10]. Consider $\mathscr{L}^2(\Omega, \mathscr{A}, \mu; \mathbb{R}^d)$ to be the space of $\mathbb{R}^d$-valued random variables with finite second moments on a sufficiently rich probability space $(\Omega, \mathscr{A}, \mu)$, and let $\mathscr{B}_0([0, L] \times \mathbb{R}_0^+)$ be the family of Borel sets with finite Lebesgue measure on $[0, L] \times \mathbb{R}_0^+$. Then a mapping $\boldsymbol{\xi} : \mathscr{B}_0([0, L] \times \mathbb{R}_0^+) \to \mathscr{L}^2(\Omega, \mathscr{A}, \mu; \mathbb{R}^d)$ is called Gaussian white noise, if it fulfills:

1. For any set $B \in \mathscr{B}_0$, $\boldsymbol{\xi}(B)$ is a $\mathscr{N}(\mathbf{0}, |B|\mathbf{I})$ distributed random variable, where $|B|$ denotes the Lebesgue measure of B.
2. For any disjoint sets $B, C \in \mathscr{B}_0$, $\boldsymbol{\xi}(B)$ and $\boldsymbol{\xi}(C)$ are independent and $\boldsymbol{\xi}(B \cup C) = \boldsymbol{\xi}(B) + \boldsymbol{\xi}(C)$ holds.

Furthermore, $\mathbf{w} : [0, L] \times \mathbb{R}_0^+ \to \mathscr{L}^2(\Omega, \mathscr{A}, \mu; \mathbb{R}^d)$ given by $\mathbf{w}(s, t) = \boldsymbol{\xi}((0, s] \times (0, t])$ is called a planar Wiener process. The associated filtration is denoted by $\mathscr{F}_{s,t}$ for $0 \le s \le L$ and $t \ge 0$.

Formalizing the stochastic Kirchhoff model with respect to Definition 1—for a one-sided clamped ($s = 0$) fiber with stress-free end ($s = L$), we consider $\mathscr{F}_{s,t}$-measurable random fields $\mathbf{r} : [0, L] \times \mathbb{R}_0^+ \times \Omega \to \mathbb{R}^d$ and $\lambda : [0, L] \times \mathbb{R}_0^+ \times \Omega \to \mathbb{R}$ which almost surely solve

$$\partial_{tt}\mathbf{r} = \partial_s(\lambda \partial_s \mathbf{r}) - \alpha^{-2}\partial_{ssss}\mathbf{r} + \mathrm{Fr}^{-2}\mathbf{e}_g + \mathrm{Dr}^{-2}(\mathbf{A} \cdot \mathbf{r} - \partial_t \mathbf{r}) + \beta \mathrm{Dr}^{-2}\boldsymbol{\xi}, \tag{1a}$$

$$\|\partial_s \mathbf{r}\|^2 = 1, \tag{1b}$$

subjected to the boundary and initial conditions

$$\mathbf{r}(0, .) = \hat{\mathbf{r}}\ ,\ \partial_s \mathbf{r}(0, .) = \boldsymbol{\tau}\ ,\ \partial_{ss}\mathbf{r}(L, .) = \mathbf{0}\ ,\ \partial_{sss}\mathbf{r}(L, .) = \mathbf{0}\ ,\ \lambda(L, .) = 0, \tag{1c}$$

$$\mathbf{r}(., 0) = \mathbf{r}_0\ ,\quad \partial_t \mathbf{r}(., 0) = \mathbf{0}\ ,\quad \lambda(., 0) = 0. \tag{1d}$$

3 Wiener Chaos Expansion for a Planar White Noise

The Wiener Chaos Expansion is based on an orthogonal decomposition of the underlying space $\mathscr{L}^2(\Omega, \mathscr{A}, \mu; \mathbb{R})$. Following [2] we denote by $\mathscr{H}$ a Gaussian Hilbert space, i.e., a complete linear subspace of $\mathscr{L}^2(\Omega, \mathscr{A}, \mu; \mathbb{R})$ containing centered Gaussian random variables. Then the set $P_n(\mathscr{H})$ of M-variate polynomials $p(\kappa_1, \ldots, \kappa_M)$ for $M \in \mathbb{N}$ of degree $\deg(p) \leq n$ with $\kappa_i \in \mathscr{H}$ is a linear subspace of $\mathscr{L}^2$. This holds also for its closure $\overline{P}_n(\mathscr{H})$. Setting $H_0 = \overline{P}_0(\mathscr{H})$ and $H_n = \overline{P}_n(\mathscr{H}) \cap P_{n-1}^{\perp}$ for $n \geq 1$, with orthogonal complement $^{\perp}$, we arrive at a pairwise orthogonal decomposition of $\mathscr{L}^2(\Omega, \sigma(\mathscr{H}), \mu; \mathbb{R})$ with σ-algebra $\sigma(\mathscr{H})$ generated by $\mathscr{H}$. For $\sigma(\mathscr{H}) = \mathscr{A}$, the Cameron-Martin theorem [1] holds, stating $\mathscr{L}^2(\Omega, \mathscr{A}, \mu; \mathbb{R}) = \cup_n H_n$. This statement can be extended to $\mathscr{L}^2(\Omega, \mathscr{A}, \mu; \mathbb{R}^d)$, since every component of a d-dimensional random variable has a finite second moment on the corresponding one-dimensional space. To derive an appropriate set $\mathscr{H}$, we follow [3] and introduce the Fourier expansion of the planar white noise.

Definition 2 (Fourier Expansion of Planar White Noise) Let $\{b_{ij}\}_{i,j\in\mathbb{N}}$ be an orthonormal basis of $\mathscr{L}^2([0, L] \times [0, T]; \mathbb{R})$ with $b_{ij}(s, t) = \phi_i(s)\psi_j(t)$, where $\{\phi_i\}_{i\in\mathbb{N}}$ and $\{\psi_i\}_{i\in\mathbb{N}}$ form orthonormal bases of $\mathscr{L}^2([0, L]; \mathbb{R})$ and $\mathscr{L}^2([0, T]; \mathbb{R})$ respectively. The Fourier expansion of the white noise components $(\mathbf{w})_i = w_i\ ,\ 1 \leq i \leq d$ is given in terms of stochastic coefficients $\kappa_{ik\ell}$ that are determined by an Itô integral (for the construction of a multi-parameter Itô integral see, e.g., [4, 10]),

$$w_i(s, t, \omega) = \sum_{k\ell} \kappa_{ik\ell}(\omega, L, T) \int_0^s \int_0^t \phi_k(\sigma)\psi_\ell(\tau)\, \mathrm{d}\sigma \mathrm{d}\tau\ , \kappa_{ik\ell} = \int_0^L \int_0^T \phi_k \psi_\ell\, \mathrm{d}\xi_i\ .$$

Remark 1 (Properties of the Fourier Expansion)

1. The stochastic coefficients of the white noise are independent identically distributed $\mathcal{N}(0,1)$ random variables as can be shown using the Itô-isometry [10].
2. For the trigonometric basis $\phi_1(s) \equiv L^{-1/2}$, $\phi_i(s) = (2/L)^{1/2}\cos((i-1)\pi s/L)$, $i \geq 2$, analogously for ψ_i, the mean square error of the truncated expansion w_i^a of the white noise behaves as $\mathbb{E}[w_i - w_i^a]^2 = \mathcal{O}(LT(N_L^{-1} + N_T^{-1}))$, where N_L and N_T denote the number of basis functions ϕ_i and ψ_i, respectively.

For the construction of the orthogonal polynomial basis, we choose the Gaussian Hilbert space as $\mathscr{H} = \overline{\text{span}\{\kappa_{ik\ell}\}}$ and introduce a bijective mapping $\eta : \{1,\dots,d\} \times \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ to enumerate each component. In this case $\mathscr{F}_{L,T} = \sigma(\mathscr{H})$ holds [4] and the system of orthonormal polynomials on $\mathscr{L}^2(\Omega, \mathscr{F}_{L,T}, \mu; \mathbb{R}^d)$ is given by products of Hermite polynomials [2]. We denote by $h_n : \mathbb{R} \to \mathbb{R}$ the nth order Hermite polynomial which is normalized with respect to the Gaussian measure [9], i.e., $\{h_n\}$ fulfill

$$\int_{\mathbb{R}} h_n(x)h_m(x)(2\pi)^{-1/2}e^{-x^2/2}\,\mathrm{d}x = \delta_{nm},$$

with Kronecker delta δ_{nm}. Furthermore, we consider a set of multiindices $J \subset \mathbb{N}_0^{\mathbb{N}}$ by $J = \{\iota = (\iota_i)_{i\geq 1} | \iota_i \in \mathbb{N}_0, |\iota| < \infty\}$ with $0_J = (0,0,\dots) \in J$, $|\iota| = \sum_i \iota_i$ and introduce the so-called Wick polynomial $p_\iota(\kappa) = \Pi_{i=1}^{\infty} h_{\iota_i}(\kappa_{\eta^{-1}(i)})$ indicated by $\iota \in J$ for a multivariable $\kappa \in \mathbb{R}^d \times \mathbb{R}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}}$. This way we can conclude the following statement from the Cameron-Martin-Theorem.

Lemma 1 *For $(s,t) \in [0,L] \times [0,T]$ let $r_i(s,t), \lambda(s,t) \in \mathscr{L}^2(\Omega, \mathscr{F}_{L,T}, \mu; \mathbb{R})$, $i = 1,\dots,d$ be $\mathscr{F}_{s,t}$ adapted solutions of (1). Then $\mathbf{r}$ and λ are given as*

$$\mathbf{r}(s,t,\omega) = \sum_{\iota\in J} \mathbf{r}_\iota(s,t)p_\iota(\kappa(\omega)), \quad \lambda(s,t,\omega) = \sum_{\iota\in J} \lambda_\iota(s,t)p_\iota(\kappa(\omega)), \tag{2}$$

with the deterministic coefficients $\mathbf{r}_\iota = \mathbb{E}[\mathbf{r}p_\iota]$, $\lambda_\iota = \mathbb{E}[\lambda p_\iota]$ and κ as in Definition 2.

We aim a system of equations that describes the behavior of the deterministic coefficients of the expansion (2). Therefore, we formally rewrite (1a) into a system of first order partial differential equations in space or time, integrate over suitable time and space intervals accounting for the initial and boundary conditions, multiply each equation with the polynomial p_ι and take the expectation. The algebraic constraint (1b) is handled by inserting the Fourier-Hermite expansion, multiplying with p_ι and taking the expectation. To obtain the final system of equations, we apply the following properties of Wick polynomials [6].

Lemma 2 *The product of two random variables a, b which are expressed in terms of their Fourier-Hermite expansion fulfills*

$$ab = \sum_{\iota} \left(\sum_{\mu} \sum_{0_J \le \nu \le \iota} C(\iota, \mu, \nu) a_{\iota-\nu+\mu} b_{\nu+\mu} \right) p_\iota \,,$$

$$C(\iota, \mu, \nu) = \left[\binom{\iota}{\nu} \binom{\mu+\nu}{\mu} \binom{\iota+\mu-\nu}{\mu} \right]^{1/2},$$

where the operators $+, -, \le$ *are defined component-wise and* $\iota! = \iota_1! \iota_2! \cdots$. *The expectations of the product of two Wick polynomials and of the product of the Wiener process and a Wick polynomial satisfy*

$$\mathbb{E}[p_\iota(\kappa) p_\mu(\kappa)] = \delta_{\iota,\mu}, \qquad \text{with } \delta_{\iota,\mu} = \Pi_i \delta_{\iota_i,\mu_i},$$

$$\mathbb{E}[p_\iota w_i(s,t)] = \sum_{k\ell} \delta_{\iota,\nu} \int_0^s \int_0^t \phi_k \psi_\ell \, \mathrm{d}\sigma \mathrm{d}\tau, \qquad \text{with } \nu = (\delta_{\eta(i,k,\ell),j})_{j \in \mathbb{N}}.$$

Consequently, the deterministic coefficients associated to the Kirchhoff beam model are described by

$$\partial_{tt} \mathbf{r}_\iota = \sum_{\mu} \sum_{0_J \le \nu \le \iota} C(\iota, \mu, \nu) \partial_s (\lambda_{\iota-\nu+\mu} \partial_s \mathbf{r}_{\nu+\mu}) - \alpha^{-2} \partial_{ssss} \mathbf{r}_\iota + \delta_{\iota,0_J} \mathrm{Fr}^{-2} \mathbf{e}_g \tag{3a}$$

$$+ \mathrm{Dr}^{-2} (\mathbf{A} \cdot \mathbf{r}_\iota - \partial_t \mathbf{r}_\iota) + \beta \mathrm{Dr}^{-2} \sum_{k\ell} \sum_{i=1}^{d} \mathbf{e}_i \delta_{\iota,(\delta_{\eta(i,k,\ell),j})_{j\in\mathbb{N}}} \phi_k \psi_\ell,$$

$$\sum_{\mu} \sum_{0_J \le \nu \le \iota} C(\iota, \mu, \nu) \partial_s \mathbf{r}_{\iota-\nu+\mu} \cdot \partial_s \mathbf{r}_{\nu+\mu} = 0, \tag{3b}$$

where $\mathbf{e}_i$ denotes the Euclidean unit vector. System (3) is completed by the boundary and initial conditions (1c)–(1d) for $\mathbf{r}_{0_J}$ and λ_{0_J} and vanishing boundary and initial conditions for $\iota \neq 0_J$.

4 Numerical Results and Discussion

Concerning the numerical simulations, we use the same spatial and temporal discretization for both, the WCE system (3) and the original constrained stochastic system (1). In space we apply a finite volume approach in combination with finite differences. To achieve small discretization stencils, we particularly use a staggered grid, assigning the cell averages of $\mathbf{r}, \partial_t \mathbf{r}$ to the cell nodes and the corresponding values of λ and $\partial_s \mathbf{r}$ to the cell edges. In time we take an implicit Euler scheme.

See [5] for details. We choose a truncation of the expansion (2). For that purpose, we fix $N_s, N_t \in \mathbb{N}$ and consider a subset of the basis on $\mathscr{L}^2([0,L]\times[0,T])$ given by $\{\phi_i\psi_j\}$ for $i \le N_s, j \le N_t$. In addition, we restrict the degree of the Wick polynomials by setting $|\iota| \le n \in \mathbb{N}$, which results in the truncated multiindex set $J_{N_sN_t,n}$.

Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e_3}\}$ be the canonical Euclidean unit basis. As benchmark scenario, we consider a two-dimensional setting with a rotational flow $\mathbf{A} = \mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_2 \otimes \mathbf{e}_1$ and the gravitational direction $\mathbf{e}_g = -\mathbf{e}_2$. The fiber is assumed to be clamped at $\hat{\mathbf{r}} = \mathbf{0}$ with $\boldsymbol{\tau} = -\mathbf{e}_2$, initially hanging straight in direction of gravity $\mathbf{r}_0(s) = -s\mathbf{e}_2$, $s \in [0, L]$. The dimensionless characteristic parameters governing the model are chosen as $(\alpha, \beta, \mathrm{Dr}, \mathrm{Fr}) = (4, 0.1, 0.1, 3)$. The fiber length and final time are $L = 1, T = 0.1$. In the simulations we use an equidistant discretization with spatial and temporal grid size $\Delta s = 10^{-2}$ and $\Delta t = 10^{-3}$, respectively. For the Fourier expansion of the Wiener process we choose the trigonometric basis, see Remark 1.

Solving the WCE system (3) once allows the creation of an arbitrary number of stochastic realizations of (1) by generating the corresponding Wiener processes, calculating their Fourier stochastic coefficients and applying Lemma 1. Additionally, the expressions (2) and Lemma 2 yield algebraic formulas for the calculation of the first and second moments of $\mathbf{r}$ and λ in terms of polynomials of their deterministic coefficients. Thus the WCE approach is in principle efficiently superior to classical Monte-Carlo simulations (MCS). To study the feasibility of WCE for our constrained stochastic beam model in industrial applications we investigate the performance of the approach (accuracy, computational costs) in comparison to MCS. Thereby, the results from 1000 MCS turn out to be suitable as reference. In Fig. 1 we see that WCE with first order Wick polynomials yields already good results for the fiber curve $\mathbf{r}$, this is confirmed by the behavior of

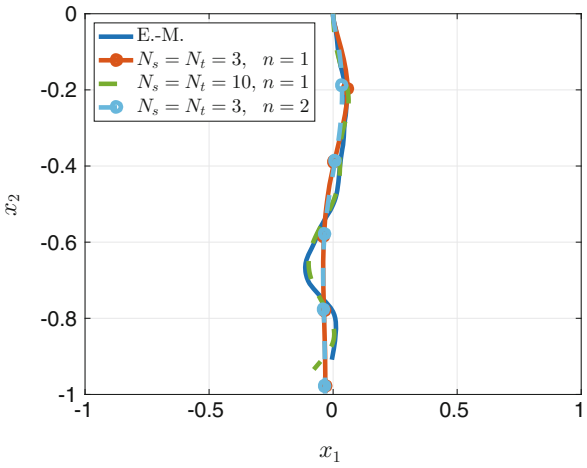


Fig. 1 Fiber curve $\mathbf{r}(., T, \omega)$. Solution of original system (1) (Euler-Maruyama method, E.-M.) and WCE system (3) for different N_s, N_t and n

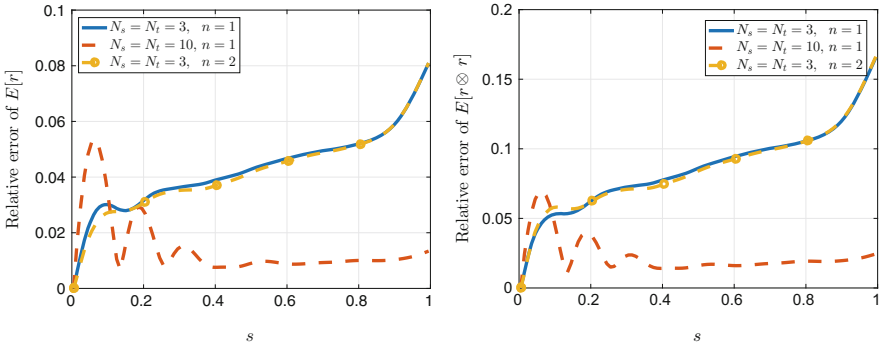


Fig. 2 Relative error for expectation $\mathbb{E}[\mathbf{r}(.,T)]$ (*left*) and second moment $\mathbb{E}[(\mathbf{r} \otimes \mathbf{r})(.,T)]$ (*right*) between WCE-approximations and (reference) results from 1000 MCS

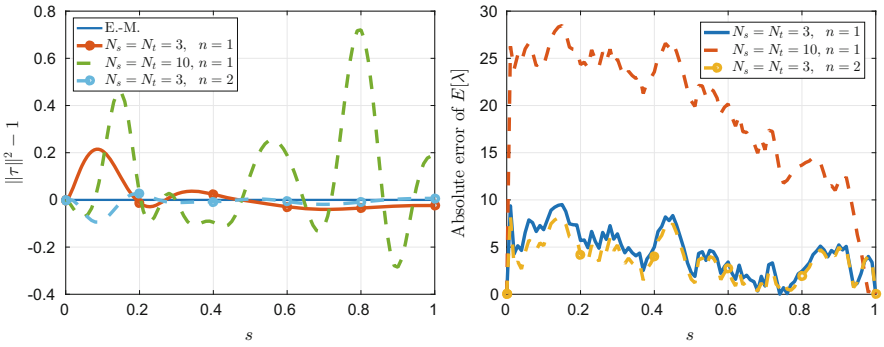


Fig. 3 *Left*: Error in the algebraic constraint of the solution to (1) (E.-M.) and the WCE-approximations at $t = T$. *Right*: Absolute error for expectation $\mathbb{E}[\lambda(.,T)]$ between WCE-approximations and (reference) result from 1000 MCS

the respective first and second moments (Fig. 2). In particular, better accuracy is achieved when increasing N_s and N_t. However, first order Wick polynomials with larger N_s and N_t are not able to improve the approximation quality of the constraint, instead we observe a deterioration. This is also seen in the behavior of the Lagrangian multiplier λ, Fig. 3.

The results indicate that the WCE approach (3) yields only reasonable approximations of the constrained system (1), given higher order polynomials (and N_s, N_t sufficiently large) are used. This poses a severe numerical challenge, since the number of coefficient systems grows as $(n+z)!/(n!z!)$ with $z = dN_sN_t$. Thus, in cases where the number of degrees of freedom is large due to the required spatial resolution, the WCE approach becomes impracticable in contrast to MCS. Another restriction on the applicability of WCE in industrial problems is given by the evaluation of the mixed moment of the force term and a Wick polynomial $\mathbb{E}[\mathbf{f}p_t]$ (and the treatment of the stochastic amplitude). The drag forces depend in general on fluid

flow simulations in complicated geometries. Hence, an approximation of such forces is required to formulate a closed WCE system analogously to (3). Consequently, WCE is not suited as a direct numerical technique for accurate stochastic fiber simulations but instead it could be applied for moment approximation or for variance reduction in a hybrid WCE-MCS approach as proposed in [6].

Acknowledgements This work has been supported by German DFG, project 251706852, MA 4526/2-1 and by the German BMBF, project OPAL 05M13.

References

1. Cameron, R.H., Martin, W.T.: The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals. Ann. Math. **48**, 385–392 (1947)
2. Ernst, O.G., Mugler, A., Starkloff, H.J., Ullmann, E.: On the convergence of generalized polynomial chaos expansions. ESAIM Math. Model. Numer. Anal. **46**, 317–339 (2012)
3. Hou, T.Y., Luo, W., Rozovskii, B., Zhou, H.M.: Wiener Chaos expansions and numerical solutions of randomly forced equations of fluid mechanics. J. Comput. Phys. **216**, 687–706 (2006)
4. Janson S.: Gaussian Hilbert Spaces. Cambridge University Press, Cambridge (1997)
5. Lindner, F., Marheineke, N., Stroot, H., Vibe, A., Wegener, R.: Stochastic dynamics for inextensible fibers in a spatially semi-discrete setting. Stochastics Dyn. (2016). doi:10.1142/S0219493717500162
6. Luo, W.: Wiener Chaos Expansion and Numerical Solutions of Stochastic Partial Differential Equations. California Institute of Technology, Pasadena (2006)
7. Marheineke, N., Wegener, R.: Fiber dynamics in turbulent flows: general modeling framework. SIAM J. Appl. Math. **66**, 1703–1726 (2006)
8. Marheineke, N., Wegener, R.: Modeling and application of a stochastic drag for fibers in turbulent flows. Int. J. Multiphase Flow **37**, 136–148 (2011)
9. Szegö, G.: Orthogonal Polynomials, 4th edn. American Mathematical Society, Providence, RI (1985)
10. Wong, E., Zakai, M.: Martingales and stochastic integrals for processes with a multi-dimensional parameter. Z. Wahr. Verw. Gebiete **29**, 109–122 (1974)

A Methodology for Fasteners Placement to Reduce Gap Between the Parts of a Wing

Nadezhda Zaitseva and Sergey Berezin

Abstract We concerned with the methodology of automatic fastener placement that reduces the gap between the parts of the wing. The gap is assumed to be of stochastic nature, therefore, it is modeled as a random field. The major issue we ran into is lack of sufficient amount of data, thus, a special small sample statistical estimator for the random field parameters is used. As a final result, the procedure is proposed to successively install the fasteners in such a way that their placement suits multiple wings at once, reducing the gap to a certain level.

1 Introduction

The assembly of an aircraft wing is a very expensive and time-consuming process, important part of which is attachment of the outer wing panel to its frame. The latter involves installation of several fastening elements, and it is necessary that the resulting gap between the attached parts be small, otherwise critical stresses may appear causing damage to the wing.

In industrial production of an aircraft, engineers often use the set of common instructions in order to reduce assembly time consumption. Particularly, there is the template for fastening element placement, which serves for all the wings within a specific series. That being said, notice that because of various uncontrollable errors in manufacturing, in positioning of the wing panel with respect to its frame, and another types of errors, the gap that needs to be eliminated will be varying from one wing to another even if these wings are taken from the same series. This raises the question of how to design a methodology of fastener placement, so that it works well for all the similar wings.

In this paper we address such a question, and propose a methodology of automatic fastener placement, that allows us to reduce the gap to a certain small level, taking into account variations within the series. In order to design this methodology it is crucial to develop a mathematical model of the gap and only then,

N. Zaitseva (✉) • S. Berezin
Peter the Great St.Petersburg Polytechnic University, Saint-Petersburg, Russia
e-mail: zaitseva.n.i@mail.ru; servberezin@yandex.ru

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3_113

the placement algorithm itself. It turns out that the most natural way to model the gap for all the wings in the series at once is to use so-called random functions [6]. More precisely, the gap is thought of as being a random field, or a random surface, that is a random function of two real variables. Further, such a random field is going to be called the *gap field*.

The main idea of our placement procedure is this. First, we put a fastener to the most probable position of the gap global maximum. Further, we estimate probability that the maximum of the field after this fastener installation is less than a certain given value. If this probability is sufficiently close to one, then we stop the procedure. Otherwise, we repeat all the actions.

When it comes to actual use of the mathematical model, one needs to estimate its parameters. This involves statistical analysis of some experimental data, measurements of the gap for real wings. The latter is very hard to get. Indeed, these measurements can only be taken manually, that involves interruption of the production at an intermediate step and leads to significant financial losses for the producer. This means that only limited amount of experimental data will be available. Therefore, in order to estimate the model parameters, special methods of the small sample statistics should be used.

2 Mathematical Model

In this section, we describe our stochastic approach of the gap modeling and the placement algorithm.

For simplicity, within this paper a very simple and concrete model of the gap is going to be used. Let us considered homogeneous isotropic Gaussian random field $\xi(x, y)$, $(x, y) \in D \subset R^2$, and its correlation function is $\rho(r, \alpha) = e^{-\alpha^2 r^2/2}$. Such a field is completely defined with three scalar parameters: mean μ, variance σ^2 and parameter α. Let us also introduce the probability $p(a, b)$ of that the ordinate of the random field is in a range from a to b. The value of these parameters can be estimated from given experimental data (see description in Sect. 3)

The easiest way to find the distribution of the gap global maximum is to work with the big sample of the gap $G(x, y)$. Though, as an initial data we have only a few realizations for $G(x, y)$ before fastening. However, after we model this gap as a random field $\xi(x, y)$, we can generate as many realizations of the gap, as we need: $G_i(x, y) = \xi_i(x, y), i = 1, 2, \ldots$

From the engineering point of view, this stochastic gap modeling is not typical and, therefore, difficult in implementation. Hence, we introduce one more gap parameter γ—the 99th percentile of the gap value (i.e. the gap is less than γ with 0.99 probability). It allows an engineer to understand the amplitude of the gap and then select level ε to which the gap must be reduced.

The following describes steps of the procedure of the automatic fasteners placement that reduces the gap to a certain level ε.

Step 1. Calculate the probability p that the global maximum value of the gap is less than ε. If this probability is sufficient large ($p \approx 1$), then stop the procedure.
Step 2. Estimate the most probable position of the global maximum for the current gap and install a fastener into this position.
Step 3. Use a software package ASRP [2] to obtain the gap value after the fastener installation. Go to Step 1.

This procedure stops when the gap is reduced to ε with probability p. The result is a list of fasteners that were installed during this process.

In practice, we start from generating a sample $\xi_1(x, y), \ldots, \xi_L(x, y)$ of Gaussian random fields with specified parameters. There are several statistical methods that allow us to do so. We choose the parametric modeling method [8] as the most convenient way, since we know the exact form of the correlation function. We consider this sample as a current gap sample: $G_1(x, y) = \xi_1(x, y), \ldots, G_L(x, y) = \xi_L(x, y)$.

For the random gap $G(x, y)$, the global maximum value M and global maximum position (X, Y) are random variables. We sample these variables:

$$M_i = \max_{(x,y)\in D} G_i(x, y), \quad (X_i, Y_i) = \arg\max_{(x,y)\in D} G_i(x, y), \quad i = 1, \ldots L, \tag{1}$$

and estimate the corresponding cumulative distributions function through kernel estimation:

$$f_M(x) = \frac{1}{Lh_1} \sum_{i=1}^{L} K\left(\frac{x - M_i}{h_1}\right), \quad f_{(X,Y)}(x, y) = \frac{1}{Lh_2^2} \sum_{i=1}^{L} K\left(\frac{x - X_i}{h_2}\right) K\left(\frac{y - Y_i}{h_2}\right). \tag{2}$$

In this case, we use an estimator from [4]: the Epanechnikov kernel function and corresponding smoothing parameters $h_1 \sim L^{-1/5}$, $h_2 \sim L^{-1/6}$, which are optimal in case of a big sample.

Thereafter, we calculate required values: desired probability $p = \int_{-\infty}^{\varepsilon} f_M(x)dx$ and the most probable position of the global maximum $(\bar{x}, \bar{y}) = \arg\max f_{(X,Y)}(x, y)$.

After each fastener installation, the gap between the parts changes its value. We use a special software package called ASRP to solve this contact problem: for one gap realization $G_i(x, y)$ we receive the value of changed gap $\bar{G}_i(x, y)$. Thus, we get a new sample with gap realizations and can start another loop of the procedure with this sample.

Further, we offer several improvements of this procedure.

Before we start fastener placement, we need to set the level ε. In the industrial wing production, the engineers who control the manufacturing process should make this decision. We offer them to do this choice based on γ-value, as it shows the range of the gap before fastening. In case, γ can be calculated from the equation $p(-\infty, \gamma) = 0.99$.

The other improvement is the acceleration of the procedure. Before the beginning of the placement procedure, we use parameter $p(a, b)$ to calculate the probability that the gap is less than ε: $p(-\infty, \varepsilon)$. This value does not provide an accurate estimate of the probability that the gap maximum does not exceed ε, but it can be used as an upper bound. If this probability is large, we do not need to install fasteners at all.

3 Model Parameters Estimation

As it has been said before, due to lack of experimental data, we bound to use methods of the small sample statistics [3, 5].

Let $\xi_1(x, y), \ldots, \xi_N(x, y)$ be the realizations of the random field $\xi(x, y)$, measured at the points of the discrete set $\{(x_i, y_i)\}_{i=1}^n$. These realizations can be grouped into the sample of Gaussian random vectors $\{\boldsymbol{\xi}^{(k)}\}_{k=1}^N$, where

$$\boldsymbol{\xi}^{(k)} = [\xi^{(k)}(x_1, y_1), \ldots, \xi^{(k)}(x_n, y_n)]^T. \tag{3}$$

From homogeneity and isotropy of the field $\xi(x, y)$ it follows that all the components of the random vector $\boldsymbol{\xi}$ have the same Gaussian distribution $N(\mu, \sigma^2)$. Besides, these components are dependent random variables. Their correlation matrix has the form $\kappa(\alpha) = \{\kappa_{i,j}(\alpha)\}_{i,j=1}^n$, where $\kappa_{i,j}(\alpha) = e^{-\alpha^2 ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/2}$. The main problem of this section is to obtain estimators for the parameters μ, σ^2, α, and $p(a, b)$.

Let us fix i-th component $\xi_i^{(k)}$ of each vector $\boldsymbol{\xi}^{(k)}$, then we will get one-dimensional sample $\xi_i^{(1)}, \ldots, \xi_i^{(N)}$. In order to estimate the mean μ and variance σ^2 we are going to use the following estimators:

$$\hat{\mu}_i = \frac{1}{N} \sum_{k=1}^{N} \xi_i^{(k)}, \quad \hat{\sigma}_i^2 = \frac{1}{N+1} \sum_{k=1}^{N} (\xi_i^{(k)} - \mu_i)^2. \tag{4}$$

Notice that the first one is quite standard, but the other one is not. This modification was proposed in [3], and it was shown that (4) yields more accurate results in terms of the relative mean square error (RMSE) for a small sample case, comparing to the standard estimator with the leading coefficient either $1/N$ or $1/(N-1)$.

Now we can use the least squares method to get to the final estimators

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}_i^2. \tag{5}$$

In order to estimate the parameter α, the idea is to select the pair of random variables $(\xi_i^{(k)}, \xi_j^{(k)})$ for each vector $\xi^{(k)}$. Each of the pairs is two dimensional Gaussian random vector, and, in order to get the maximum likelihood estimator

for $\alpha_{i,j}$, we write the logarithmic likelihood function [10]

$$L_{i,j}(\mu,\sigma^2,\kappa_{i,j}(\alpha)) = C - N\ln\sigma^2 + \frac{N}{2}\ln(1-\kappa_{i,j}^2(\alpha)) - \frac{1}{2\sigma^2(1-\kappa_{i,j}^2(\alpha))}\times$$
$$\times\sum_{k=1}^{N}\left((\xi_i^{(k)}-\mu)^2 + (\xi_j^{(k)}-\mu)^2 - 2\kappa_{i,j}(\alpha)(\xi_i^{(k)}-\mu)(\xi_j^{(k)}-\mu)\right). \tag{6}$$

Thereafter, the estimates for $\alpha_{i,j}$ can be calculated numerically by maximization of (6) with respect to $\alpha_{i,j}$. Again, the estimator for α can be found by the least squares method:

$$\hat{\alpha} = \frac{1}{n(n-1)}\sum_{l,p=1,l\neq p}^{n}\hat{\alpha}_{l,p}. \tag{7}$$

In order to estimate probability $p(a,b)$, the parametric estimator

$$\hat{p}(a,b) = \int_a^b \tilde{f}(x)\,dx, \quad \tilde{f}(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}}e^{-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}} \tag{8}$$

is usually used. It was shown in [3, 5] that for a small sample problems the quality of this estimator can be improved if one uses a kernel density estimator instead of $\tilde{f}$.

Further we are going to need the way to compare different estimators of the density function. The natural quantity to do so is the relative mean integral square error (RMISE) [7]

$$\delta^2(\hat{f},f) = \frac{E\|\hat{f}-f\|^2}{\|f\|^2}, \tag{9}$$

where $\|\cdot\|$ denotes the $L_2(\mathbb{R})$-norm.

A kernel estimator is uniquely defined in terms of the kernel function $K(x)$ and smoothing parameter h (see [4, 9]). In this paper, we are considering two problems. The first one is to choose the optimal parameter when the kernel function is specified, and the other one is to choose both the kernel function and the parameter. In either of the cases, RMISE (9) is going to be minimized.

Within the first approach we use the kernel estimator written in the form

$$\hat{f}_{K,h}(x) = \frac{1}{nNh}\sum_{k=1}^{N}\sum_{i=1}^{n}K\left(\frac{x-\xi_i^{(k)}}{h}\right), \tag{10}$$

and the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. Then, the optimization problem can be formulated as

$$\inf_h \delta^2(\hat{f}_{K,h},f), \tag{11}$$

that is, we need to find the optimal value of h in order to minimize RMISE δ^2. It can be shown that $\delta^2(\hat{f}_{K,h},f) = \delta^2(\eta,\kappa)$, where $\eta = h/\sigma$, can be rewritten in the form [10]

$$\delta^2(\hat{f}_{K,h},f) = 1+\frac{(N-1)}{N\sqrt{1+\eta^2}}-\frac{2\sqrt{2}}{\sqrt{2+\eta^2}}+\frac{1}{n^2N}\sum_{\substack{l_1=1\\l_2=1}}^{n}(1-\kappa_{l_1,l_2}(\alpha)+\eta^2)^{-1/2}. \quad (12)$$

Let us do some simplification and consider a *minimax* problem similar to the one employed in [1]: $\inf\limits_{h}\sup\limits_{k\in Q}\delta^2(\eta,k)$, where Q is the set of symmetric positive-definite matrices with absolute values of their elements not exceeding one and with the unit elements in the diagonal. From (12) we are going to have

$$\delta^2(\eta) = \sup_{k\in Q}\delta^2(\eta,k) = 1+\frac{N-1}{N}\frac{1}{\sqrt{1+\eta^2}}-\frac{2\sqrt{2}}{\sqrt{2+\eta^2}}+\frac{1}{N\eta}. \quad (13)$$

The optimal value for h then can be calculated as $h_{opt} = \eta_{opt}\sigma$, where $\eta_{opt} = \arg\min_{\eta>0}\delta^2(\eta)$. The last problem can be solved numerically.

Within the second approach, we are looking for an optimal kernel function, therefore, it make sense to use the kernel estimator of the following form:

$$\hat{f}_K(x) = \frac{1}{nN}\sum_{k=1}^{N}\sum_{i=1}^{n}K\left(x-\xi_i^{(k)}\right). \quad (14)$$

The parameter h has already been added into the kernel function, so there is no need to consider it separately.

In this case the problem of the RIMSE $\delta^2(\hat{f}_K,f)$ minimization

$$\inf_{K\in L_2(\mathbb{R})}\delta^2(\hat{f}_K,f) \quad (15)$$

can be solved analytically [10]:

$$K_{opt}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\frac{1}{n^2N}\sum_{l_1=1,l_2=1}^{n}e^{\sigma^2\kappa_{l_1,l_2}z^2}+1-\frac{1}{N}\right)^{-1}e^{-izx}dz. \quad (16)$$

Now, we have two estimators for the density function, and, hence, two estimates for the probability $p(a,b)$, besides the parametric one.

All these estimators were tested using numerical experiments. In Fig. 1 one can see the relative mean square error comparison for the standard unbiased estimator and for estimator (5) of the variance. The comparison of different estimators of the parameter $p(a,b)$ is presented in Fig. 2. Apparently, for the small sample case

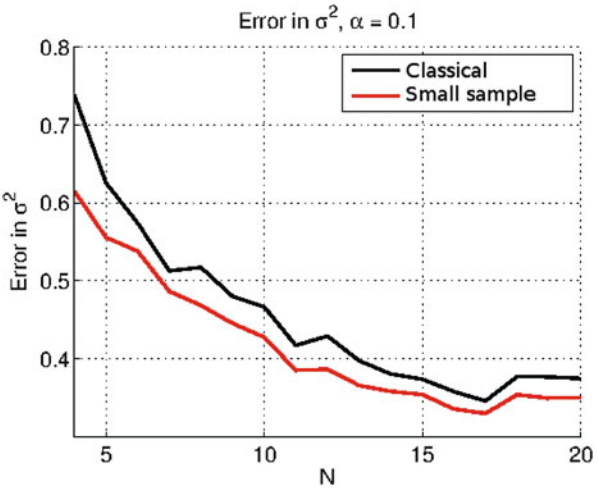


Fig. 1 Relative error for small sample estimator and standard unbiased estimator for σ^2

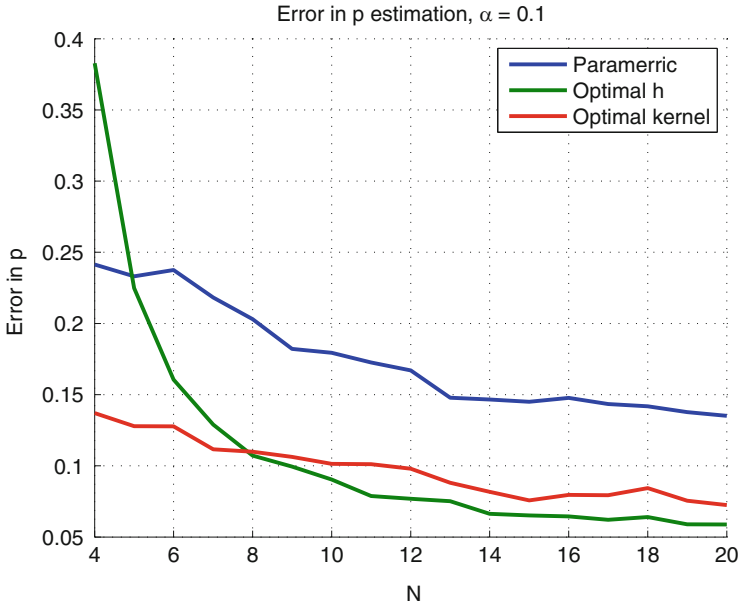


Fig. 2 RIMSE for estimators through h_{opt}, K_{opt} and the standard parametric one for $p(-\infty, 2)$

obtained estimators give better accuracy than the parametric one. Notice that the estimator of $p(a, b)$ based on use of the optimal kernel function (16) has more stable behavior with respect to the number N of given realizations of the field $\xi(x, y)$.

4 Conclusions

In this work, we have developed a methodology for fasteners placement that reduces the gap between two parts of the wing. Notice, that more generally any appropriate two parts of the plane can be considered. This methodology is based on the mathematical model of the gap and offers an explicit procedure for the automatic fastener placement. In order to estimate the parameters of the mathematical model an accurate method based on the small sample statistics was proposed.

The methodology that was created can be used to improve technology of the wing parts assembly, and, therefore, can reduce the overall production cost.

References

1. Bentkus, R., Kazbaras, A.: Optimal statistical estimates of the density funtion under a priori information. Lith. Math. J. **22**(3), 29–39 (1981)
2. Bretagnol, B., Lupuleac, S., Petukhova, M., Shinder, Y., Smirnov, A., Yakunin, S.: Numerical approach for airframe assembly simulation. J. Math. Ind. **4**, 8 (2014)
3. Cheremenskij, V.G., Zayats, O.I.: On the application of the small sample methods to the problems of reliability berthing structures (in Russian). Nauchno-tekhnicheskie problemy, Moscow, Transport, pp. 66–75 (1988)
4. Epanechnikov, V.A.: Nonparametric estimates of a multivariate probability density. Theory Probab. Appl. **14**, 153–158 (1969)
5. Gaskarov, D.V., Shapovalov, V.I.: Small Sample (in Russian). Statistic, Moscow (1978)
6. Husu, A.P., Palmov, V.A., Wittenberg, Y.R.: The Surface Roughness (Theoretical-Probabilistic Approach) (in Russian). Nauka, Moscow (1975)
7. Nadaraya, E.A.: On the integrated mean square error of some nonparametric estimates for the density function. Theory Probab. Appl. **21**, 133–141 (1974)
8. Palagin, Y.I., Shalygin, A.S.: Applied Methods of Statistical Modeling (in Russian). Mashinostroenie, Leningrad (1986)
9. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. **21**, 832–837 (1956)
10. Zaitseva, N.I.: Estimation of Gaussian field parameters with a small sample of discrete measurements (in Russian). Master dissertation, Peter the Great Saint-Petersburg Polytechnic University, Saint-Petersburg (2016)

Index

P. Quintela et al. (eds.), *Progress in Industrial Mathematics at ECMI 2016*,
Mathematics in Industry 26, DOI 10.1007/978-3-319-63082-3