Mathematics in Industry 32
The European Consortium for Mathematics in Industry

Giuseppe Nicosia
Vittorio Romano
Editors

Scientific Computing in Electrical Engineering

SCEE 2018, Taormina, Italy, September 2018

ECMI
EUROPEAN CONSORTIUM FOR
MATHEMATICS IN INDUSTRY

Springer

Mathematics in Industry

The European Consortium for Mathematics in Industry

Volume 32

The *ECMI* subseries of the *Mathematics in Industry* series is a project of *The European Consortium for Mathematics in Industry*. *Mathematics in Industry* focuses on the research and educational aspects of mathematics used in industry and other business enterprises. Books for *Mathematics in Industry* are in the following categories: research monographs, problem-oriented multi-author collections, textbooks with a problem-oriented approach, conference proceedings. Relevance to the actual practical use of mathematics in industry is the distinguishing feature of the books in the *Mathematics in Industry* series.

More information about this subseries at http://www.springer.com/series/4651

Giuseppe Nicosia • Vittorio Romano
Editors

Scientific Computing in Electrical Engineering

SCEE 2018, Taormina, Italy, September 2018

Springer

ECMI
EUROPEAN CONSORTIUM FOR
MATHEMATICS IN INDUSTRY

Editors

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

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

ISSN 1612-3956 ISSN 2198-3283 (electronic)
Mathematics in Industry
The European Consortium for Mathematics in Industry
ISBN 978-3-030-44100-5 ISBN 978-3-030-44101-2 (eBook)
https://doi.org/10.1007/978-3-030-44101-2

Mathematics Subject Classification (2020): 65-06, 65Lxx, 65Mxx, 65Nxx, 65L06, 65L12, 65L15, 65L60, 65L80, 65M06, 65M60, 78-06

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

Preface

The 12th International Conference on *Scientific Computing in Electrical Engineering* (SCEE) was organized by the Group of Applied Mathematics at the University of Catania, Italy, jointly with the Interdepartmental Center of Mathematics for Technology "A. M. Anile" (CIMAT). It was held in the period 23–27 September 2019 in Taormina, Italy, a charming town on a cliff by the sea on the east coast of the island of Sicily, with a mythical atmosphere spread all around which has enchanted visitors from all over the world for years and years.

The SCEE series of conferences has an interdisciplinary focus and provides a platform for sharing the results of the latest scientific research about modeling and numerical simulation related to electrical engineering in a broad sense. That in Taormina was the 12th edition after the conference series started as a national meeting in 1997 in Darmstadt, Germany. The other previous editions were held in

- Wolfgang, Austria (2016),
- Wuppertal, Germany (2014),
- Zurich, Switzerland (2012),
- Toulouse, France (2010),
- Espoo, Finland (2008),
- Sinaia, Romania (2006),
- Capo D'Orlando, Italy (2004),
- Eindhoven, Netherlands (2002),
- Warnemünde, Germany (2000),
- Berlin, Germany (1998).

The main topics of the conferences concern:

- computational electromagnetics,
- circuit and device modeling and simulation,
- coupled problems and multi-scale approaches in space and time,
- mathematical and computational methods including uncertainty quantification,
- model order reduction, and
- industrial applications.

The scientific program included invited and contributed talks, poster sessions, and an industrial day. More than 60 scientists attended SCEE 2018. Nine invited speakers covered all the subjects of the conference with inspiring and visionary presentations (in alphabetical order):

– Bilen Emek Abali, Technische Universität Berlin, *Modeling Mechanochemistry in Li-Ion Batteries*;
– Matthias Auf der Maur, University of Rome Tor Vergata, *Current Developments in Device Simulation: Degeneracy, Arbitrary Density of States and Multi-Particle Drift-Diffusion*;
– Tonio Biondi, Maxim Integrated, Catania, *Data Center Power*;
– Steffen Börm, University of Kiel, *GCA-H^2Matrix Compression for Electrostatic Simulations*;
– Peter Gangl, Technische Universität Graz, *Topology and Shape Optimization of Electrical Machines*;
– Jay Gopalakrishnan, Portland State University, *Techniques for Modeling Fiber Laser Amplifiers*;
– Sarah Grundel, Max Planck Institute Magdeburg, *Simulation and Model Order Reduction of Power Systems*;
– Tudor Ionescu, University Politehnica of Bucharest, *Model Reduction for Non-linear Systems—A time-Domain Moment Matching Perspective*;
– Omar Morandi, University of Florence, *Description of the Trajectories of Quantum Particles by a Quantum Lagrangian Approach*.

In addition, there have been 26 oral contributed talks, selected by the scientific committee, and 22 posters. More details can be found in the webpage of the conference.[1] The scientific program was completed with a social tour through the streets of Taormina and a visit to the magnificent ancient Greek-Roman theater.

This book contains a selection of the contributions to SCEE 2018, after a regular peer-review process. The contributions are divided into five parts according to the specific investigated subject:

I Computational Electromagnetics
II Device Modeling and Simulation
III Circuit Simulation
IV Mathematical and Computational Methods
V Model Order Reduction

We would like to thank all the attendees, whose contributions have made the event a scientifically successful one, the scientific committee, for their valuable work during the selection of the scientific contributions, and the organizing committee, for the help which has allowed a smooth carrying on of the conference.

We acknowledge the financial contribution from the Department of Mathematics and Computer Science, University of Catania, and from the Gruppo Nazionale di

[1] https://scee2018.icas.xyz.

Fisica Matematica (GNFM). A special thanking is addressed to the association *Associazione Angelo Marcello Anile* (ASSOAMA) which very generously supported the conference.

We are grateful to Dr. Giovanni Nastasi whose work, in collecting all the contributions and unifying their styles to merge them into a unique book, has been precious for the redaction of the present proceedings.

Catania, Italy
Catania, Italy
November 2019

Giuseppe Nicosia
Vittorio Romano

Contents

Part III Circuit Simulation

Part IV Mathematical and Computational Methods

Part V Model Order Reduction

Contributors

Karthik V. Aadithya Sandia National Laboratories, Albuquerque, NM, USA

Bilen Emek Abali Technische Universität Berlin, Institute of Mechanics, MS 2, Berlin, Germany

Pasquale Claudio Africa MOX Modelling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

Andreas Bartel Bergische Universität Wuppertal, Wuppertal, Germany

Tamara Bechtold Jade University of Applied Sciences, Department of Engineering, Wilhelmshaven, Germany

University of Rostock, Institute for Electronic Appliances and Circuits, Rostock, Germany

Theo Beelen Eindhoven University of Technology, Eindhoven, The Netherlands

Brigitte Bidégaray-Fesquet Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, Grenoble, France

Institute of Engineering Univ. Grenoble Alpes, Grenoble, France

Kai Bittner University of Applied Sciences of Upper Austria, Wels, Austria

Andreas Blaszczyk ABB Corporate Research, Baden-Dättwil, Switzerland

P. Bolcato Mentor Graphics, Grenoble, France

Maxime Bonnet Laboratory on Plasma and Conversion of Energy, Toulouse Cedex, France

Steffen Börm Department of Mathematics, University of Kiel, Kiel, Germany

Hans Georg Brachtendorf University of Applied Sciences of Upper Austria, Wels, Austria

M. S. Bruzón Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

Mario Caironi Center for Nano Science and Technology @PoliMi, Istituto Italiano di Tecnologia, Milano, Italy

Thomas Christen ABB Corporate Research, Baden-Dättwil, Switzerland

Sven Christophersen Department of Mathematics, University of Kiel, Kiel, Germany

Marco Coco Università degli Studi di Firenze, Dipartimento di Matematica e Informatica “Ulisse Dini”, Firenze, Italy

N. T. K. Dang Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

Carlo de Falco MOX Modelling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

R. de la Rosa Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

Herbert Egger Dept. of Mathematics, TU Darmstadt, Darmstadt, Germany

Armin Fohler Linz Center of Mechatronics, Linz, Austria

M. L. Gandarias Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

Peter Gangl Institute of Applied Mathematics, TU Graz, Graz, Austria

Jay Gopalakrishnan Portland State University, Portland, OR, USA

Tathagata Goswami Portland State University, Portland, OR, USA

Jacob Grosek Directed Energy Directorate, Air Force Research Laboratory, Kirtland Air Force Base, Albuquerque, NM, USA

Michael Günther Bergische Universität Wuppertal, Wuppertal, Germany

Martin S. Hilario Air Force Research Laboratory, Kirtland AFB, Albuquerque, NM, USA

Carole Hénaux Laboratory on Plasma and Conversion of Energy, Toulouse Cedex, France

Brad W. Hoff Air Force Research Laboratory, Kirtland AFB, Albuquerque, NM, USA

Siyang Hu Jade University of Applied Sciences, Department of Engineering, Wilhelmshaven, Germany

University of Rostock, Institute for Electronic Appliances and Circuits, Rostock, Germany

Guillermo Indalecio CITIUS, University of Santiago de Compostela, Santiago de Compostela, Spain

Xavier Jonsson Mentor, A Siemens Business, Grenoble, France

Clément Jourdana Univ. Grenoble Alpes, CNRS, Grenoble INP[†], LJK, Grenoble, France

Institute of Engineering Univ. Grenoble Alpes, Grenoble, France

Kole Keita Univ. Jean Lorougnon Guédé (UJLoG), Daloa, Ivory Coast

Eric R. Keiter Sandia National Laboratories, Albuquerque, NM, USA

Hans Kosina Institute for Microelectronics, TU Wien, Vienna, Austria

Jan Kühn Bergische Universität Wuppertal, Wuppertal, Germany

Petra Kumi Worcester Polytechnic Institute, Worcester, MA, USA

Ulrich Langer Institute of Computational Mathematics, JKU Linz, Linz, Austria

Liliana Luca Department of Mathematics and Computer Science, Università degli Studi di Catania, Catania, Italy

Angelos Mantzaflaris Institute of Applied Geometry, JKU Linz, Linz, Austria

Université Côte d'Azur, Inria Sophia Antipolis - Méditerranée, Valbonne, France

Giovanni Mascali Università della Calabria, Dipartimento di Matematica e Informatica, Rende, Italy

INFN-Gruppo c. Cosenza, Cosenza, Italy

J. M. L. Maubach Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

Ting Mei Sandia National Laboratories, Albuquerque, NM, USA

Frédéric Messine Laboratory on Plasma and Conversion of Energy, Toulouse Cedex, France

Hans Kristian Meyer Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Omar Morandi University of Florence, Florence, Italy

Orazio Muscato Università degli Studi di Catania, Dipartimento di Matematica e Informatica, Catania, Italy

Giovanni Nastasi Department of Mathematics and Computer Science, Università degli Studi di Catania, Catania, Italy

Dario A. Natali Dipartimento di Elettronica, Informazione, Bioingegneria, Politecnico di Milano, Milano, Italy

Center for Nano Science and Technology @PoliMi, Istituto Italiano di Tecnologia, Milano, Italy

Roland Pulch Universität Greifswald, Institute of Mathematics and Computer Science, Greifswald, Germany

Piotr Putek Bergische Universität Wuppertal, Wuppertal, Germany

Bogdan Radu Graduate School for Computational Engineering, TU Darmstadt, Darmstadt, Germany

Ian M. Rittersdorf Naval Research Laboratory, Washington, DC, USA

Vittorio Romano Università degli Studi di Catania, Dipartimento di Matematica e Informatica, Catania, Italy

J. Rommes Mentor Graphics, Grenoble, France

W. H. A. Schilders Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

Rainer Schneckenleitner RICAM, Austrian Academy of Sciences, Linz, Austria

Wim Schoenmaker MAGWEL N.V., Leuven, Belgium

Michael Schueller University of Applied Sciences Rapperswil, Rapperswil, Switzerland

E. Jan W. ter Maten Bergische Universität Wuppertal, Wuppertal, Germany

R. Tracinà Dipartimento di Matematica e Informatica, Università di Catania, Catania, Italy

Anuj Kumar Tyagi Eindhoven University of Technology, Eindhoven, The Netherlands

Jonathan S. Venne Worcester Polytechnic Institute, Worcester, MA, USA

Vadim V. Yakovlev Worcester Polytechnic Institute, Worcester, MA, USA

Rtimi Youness Laboratory on Plasma and Conversion of Energy, Toulouse Cedex, France

Chengdong Yuan Jade University of Applied Sciences, Department of Engineering, Wilhelmshaven, Germany

University of Rostock, Institute for Electronic Appliances and Circuits, Rostock, Germany

Walter Zulehner Johannes Kepler University Linz, Institute of Computational Mathematics, Linz, Austria

Part I
Computational Electromagnetics

Seven papers form this part dedicated to the contributions falling into the field of the Computational Electromagnetics (CE).

In the contribution *Surface Charging Formulations for Engineering Applications. Validation by Experiments and Transient Models* by A. Blaszczyk et al. electrostatic BEM (Boundary Element Method) formulations are presented for the calculation of dielectric surface charging, including saturation and restrike phenomena. The simulation results turn out to be in agreement with surface potential measurements in a simple rod-barrier-plane configuration, where lightning impulses initiate streamers and charge accumulation on the barrier. The usefulness of the given BEM-formulation is additionally supported by transient charging simulations in the framework of an electric carrier drift model.

In *A Mass-Lumped Mixed Finite Element Method for Maxwell's Equations* by H. Egger and B. Radu a novel mass-lumping strategy for a mixed finite element approximation of Maxwell's equations is proposed. On structured orthogonal grids it coincides with the spatial discretization of the Yee scheme but generalizes naturally to unstructured grids and anisotropic materials, thus yielding a natural variational extension of the Yee scheme for these situations.

In the contribution *Adaptive Mesh Refinement for Rotating Electrical Machines Taking into Account Boundary Approximation Errors* by A. Fohler and W. Zulehner an error estimator for a class of second order quasilinear elliptic problems in 2D is presented. The computational domain consists of two parts—called rotor and stator in the framework of electrical motors—separated by a curvilinear interface. For the coupling of the rotor and the stator on the interface a Nitsche technique is employed. The residual error estimator is constructed with adaptations due to the coupling strategy. The error estimator takes into account the polygonal approximation of the stator and the rotor using ideas from hierarchical error estimates.

Future e-mobility calls for efficient electrical machines. For different areas of operation, these machines have to satisfy certain desired properties that often depend on their design. In the contribution *Isogeometric Simulation and Shape Optimization with Applications to Electrical Machines* by P. Gangl et al. the authors investigate the use of multipatch Isogeometric Analysis (IgA) for the simulation and shape optimization of the electrical machines. In order to get fast simulation and optimization results, non-overlapping domain decomposition (DD) methods are used to solve the large systems of algebraic equations arising from the IgA discretization of underlying partial differential equations. The DD is naturally related to the multipatch representation of the computational domain and provides the framework for the parallelization of the DD solvers.

Since a full-featured simulation using the Maxwell system on a realistic fiber is beyond reach, simplified models using Coupled Mode Theory (CMT) form the state of the art. Numerical techniques for simulation of electromagnetic wave propagation within fiber amplifiers are discussed in the contribution *Techniques for Modeling Fiber Laser Amplifiers* by J. Gopalakrishnan et al. who present a novel concept of an equivalent short fiber, namely an artificial fiber which imitates a longer fiber in the essential characteristics. A CMT simulation on an equivalent short fiber requires only a fraction of the computational resources needed to simulate the full length fiber.

There exist different models which approximate the phenomenon of magnetic hysteresis. Only some of them inherently consider the thermal dependency of hysteresis or have been extended with that respect. In *A Thermal Extension of Tellinen's Scalar Hysteresis Model*, authors J. Kühn et al., Tellinen's scalar magnetic hysteresis model is reviewed and illustrated. A temperature dependent scalar magnetic hysteresis model is deduced and investigated.

In the contribution *Computational Characterization of a Composite Ceramic Block for a Millimeter Wave Heat Exchanger* by P. Kumi et al. electromagnetic and electromagnetic-thermal coupled problems are solved by the finite-difference time-domain technique for an AlN:Mo composite ceramic block backed by a thin metal plate and irradiated by a high-power W-band plane wave. Computation is based on experimental data on temperature-dependent complex permittivity, specific heat, and thermal conductivity. Non-uniformity of patterns of dissipated power and temperature is quantified via standard-deviation-based metrics.

Surface Charging Formulations for Engineering Applications: Validation by Experiments and Transient Models

Andreas Blaszczyk, Thomas Christen, Hans Kristian Meyer, and Michael Schueller

Abstract Electrostatic BEM (Boundary Element Method) formulations are presented for the calculation of dielectric surface charging, including saturation and restrike phenomena. The simulation results turn out to be in agreement with surface potential measurements in a simple rod-barrier-plane configuration, where lightning impulses initiate streamers and charge accumulation on the barrier. The usefulness of the given BEM-formulation is additionally supported by transient charging simulations in the framework of an electric carrier drift model.

1 Introduction

Surface charges (SC) on solid insulator surfaces can significantly influence the dielectric performance of medium and high voltage power devices. They can mitigate discharge inception effects during a lightning impulse test, as well as enhance them for applied voltages with reversed polarity. Unfortunately, the simulation of the intrinsically transient charging, which may occur via a zoo of different gas discharge processes like streamers, leaders, ion motion and combinations thereof, is a complex task and thus requires simplified approaches for application to real devices.

Recently a simplified engineering approach based on the saturation-charge boundary-condition has been proposed [1]. It works because saturation is a rather robust extremal stage of SC accumulation that allows assessment of possible changes in field distribution without performing the full analysis of the charging

A. Blaszczyk (✉) · T. Christen
ABB Corporate Research, Baden-Dättwil, Switzerland
e-mail: Andreas.Blaszczyk@ch.abb.com; Thomas.Christen@ch.abb.com

H. K. Meyer
Norwegian University of Science and Technology (NTNU), Trondheim, Norway
e-mail: hans.meyer@ntnu.no

M. Schueller
University of Applied Sciences Rapperswil, Rapperswil, Switzerland
e-mail: michael.schueller@hsr.ch

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_1

process. By neglecting the influence of the space charge, a simple electrostatic computation based on integral approach is possible (without meshing the gas volume).

In this paper we present a new formulation of the saturated SC for the 3D boundary element method (BEM), which can be efficiently applied in an industrial design environment. In addition to the saturation stage, computational models for modification of the accumulated charge due to restrikes (back discharges after changes of electrode potentials) are considered. The new formulation is validated with experiments and transient models.

2 BEM Formulations

The formulations presented below are explained and validated for a simple, air-insulated arrangement of rod-barrier-plane shown in Fig. 1 where the standardized lightning impulse (LI) 1.2/50 µs is applied to the active rod electrode. During the transient load and the dynamically developing discharge we distinguish a few representative stages shown in Fig. 1a–e. They occur within a time frame starting from 1 µs, when the applied voltage reaches the peak, up to 2–3 min lasting until the final measurement. Each of these stages can be represented by a steady state (electrostatic) formulation that is described in the following subsections. Such steady state formulations are applicable for arbitrary 3D geometries and can be efficiently used in engineering simulations [2].

The presented equations are limited to boundaries between gas and the solid dielectric where the surface charge distribution has to be obtained or specified. Other parts of the traditional BEM formulation including the Fredholm integral equation of the first order for electrode boundaries as well as formulation for floating potentials are described in [2–5].

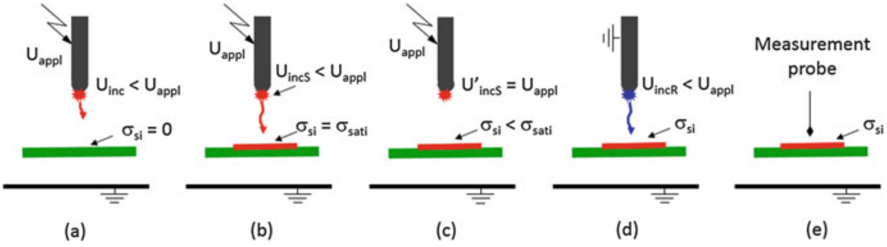


Fig. 1 Rod-barrier-plane configuration for different stages of discharge development. (**a**) Background field. (**b**) Saturation. (**c**) Subsaturation. (**d**) Restrikes. (**e**) Surface potentials

2.1 Background Field

For a point i on a dielectric boundary the Gauss law must be fulfilled according to the following equation:

$$\varepsilon_{r,Ins}\varepsilon_0 E_{ni,Ins} = \varepsilon_{r,Gas}\varepsilon_0 E_{ni,Gas} + \sigma_{si} \tag{1}$$

where $\varepsilon_{r,Ins}$ and $\varepsilon_{r,Gas}$ are the relative permittivities of the solid insulation and gas, $E_{ni,Ins}$ and $E_{ni,Gas}$ are normal components of the electric field in the corresponding medium, and σ_{si} is the density of the accumulated surface charge, which is equal to zero in case of the initial background field, see Fig. 1a.

The indirect BEM formulation introduces a concept of virtual charge "in vacuo" that specifies the density σ_i for all points i on boundaries of the model as a solution of the equation system including (1) . Once this charge is obtained the field can be computed as a superposition of all virtual charge contributions. An essential feature is handling of the singularity when the field is computed exactly at the point i. This is solved by introducing a jump term resulting from the Gauss law applied to the virtual charge located on a small, flat surface area around the point i [4]. Consequently, the normal field components in (1) can be computed as a sum of a jump term and the normal electric field component E^-_{ni} obtained by integrating all virtual charges σ_j except of the charge located within the small, flat surface area around the point i:

$$E_{ni,Ins} = E^-_{ni} - \frac{\sigma_i}{2\varepsilon_0} \quad \text{and} \quad E_{ni,Gas} = E^-_{ni} + \frac{\sigma_i}{2\varepsilon_0} \tag{2}$$

with

$$E^-_{ni} = \frac{1}{4\pi\varepsilon_o}\sum_j \int_{S_j} \frac{\mathbf{n_i}\cdot\mathbf{r_{ij}}}{r^3_{ij}}\sigma_j dS \tag{3}$$

where $\mathbf{n_i}$ is the normal vector at collocation point i pointing into the gas and r_{ij} is the distance between collocation point i and the surface element represented by the integration point j.[1] After applying (2) to (1) and assuming $\sigma_{si} = 0$ one can obtain the Fredholm integral equation of the second order as follows:

$$E^-_{ni} - \frac{\varepsilon_{r,Ins} + \varepsilon_{r,Gas}}{\varepsilon_{r,Ins} - \varepsilon_{r,Gas}}\frac{\sigma_i}{2\varepsilon_0} = 0. \tag{4}$$

[1] Rigorous mathematical formulations denote the integral included in (3) as the adjoint double layer potential operator. Since our focus is on physical and engineering models the mathematical technique of computing this integral is beyond of scope of this paper. For more details we refer to literature [4, 5].

As shown in Fig. 1a, a streamer discharge will start to propagate if the applied voltage U_{appl} (peak value) is larger than the inception voltage U_{inc} at the rod tip (estimated according to [1]). This will deliver the charge to be accumulated along the barrier surface.

2.2 Saturation

We assume that the saturation stage at the dielectric boundary is achieved when the amount of accumulated charge is so large that the normal component of the electric field in the gas $E_{ni,Gas}$ is zero. Physically, it means that the accumulated charge changes the background field in such a way that the streamer discharge instead of hitting the barrier will follow the field lines going parallel to its surface, which will prevent further accumulation. Consequently, Eq. (1) can be split into the following two equations where the prescribed surface charge density σ_{si} is replaced by the unknown saturation charge density σ_{sati}:

$$\varepsilon_{r,Gas}\varepsilon_0 E_{ni,Gas} = 0 \quad \text{and} \quad \varepsilon_{r,Ins}\varepsilon_0 E_{ni,Ins} = \sigma_{sati} \tag{5}$$

After applying (2) to (5) we get a system of integral equations where the unknowns are both charge densities, the virtual σ_i related to BEM and the physical σ_{sati} representing the accumulated charge:

$$E_{ni}^{-} + \frac{\sigma_i}{2\varepsilon_0} = 0 \quad \text{and} \quad E_{ni}^{-} - \frac{\sigma_i}{2\varepsilon_0} - \frac{\sigma_{sati}}{\varepsilon_0 \varepsilon_{r,Ins}} = 0 \tag{6}$$

An example of the computed saturation charge distribution σ_{sati} has been shown in Fig. 2b (bell-shaped curve). The saturation charge will mitigate the field strength at the rod tip and increase the inception voltage from U_{inc} to U_{incS}. The saturation charge can be considered as a good approximation of the accumulated charge if $U_{incS} < U_{appl}$, see Fig 1b. Otherwise, the formulation presented in the next subsection should be followed.

2.3 Subsaturation

In case of $U_{incS} > U_{appl}$ the streamers delivering charge to the barrier may be extinguished. Consequently, the saturation may not be achieved and the extremal value of σ_{sati} (calculated from (6)) has to be reduced as follows:

$$\sigma_{si} = k_{si}\sigma_{sati} \cong k_{sConst}\sigma_{sati} \tag{7}$$

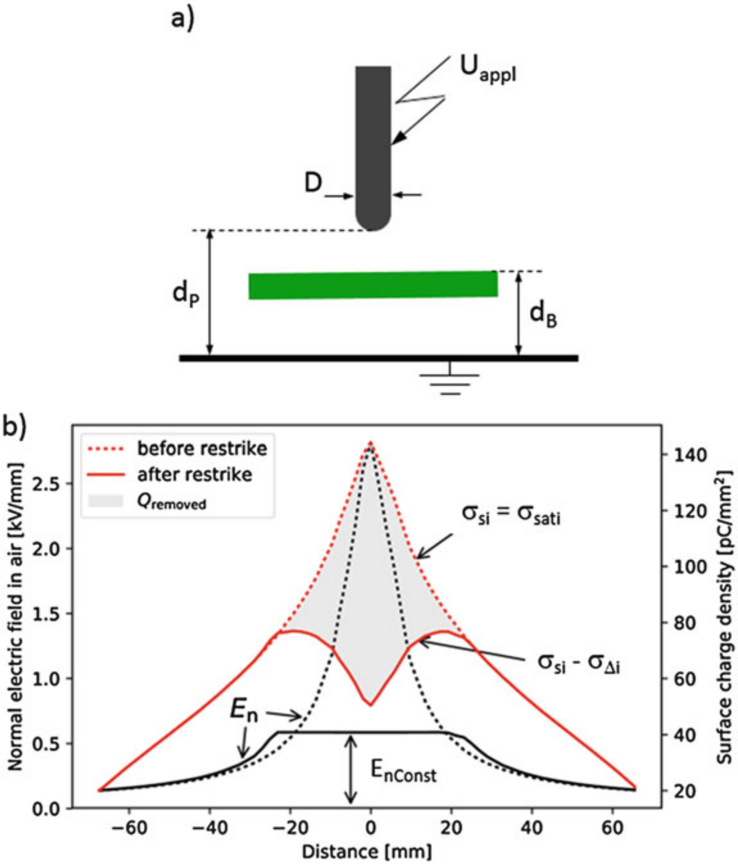


Fig. 2 (**a**) Rod-barrier-plane configuration. (**b**) distributions of charge density and normal field strength before and after the restrike calculated for: U_{appl}=35 kV, D=4 mm, d_P=10 mm, d_B=5 mm

where the reduction factor k_{si} is approximated, for simplification, by a constant $k_{sConst} \leq 1$ (to be independent of location i). The value of k_{sConst} needs to be estimated iteratively so that the original inception voltage in saturation stage U_{incS} will decrease to a new value $U^{'}_{incS}$ where $U^{'}_{incS} = U_{appl}$. $U^{'}_{incS}$ is the inception voltage calculated with the presence of the reduced surface charge (7), see Fig 1c. If in the saturated stage $U_{incS} < U_{appl}$ then $k_{sConst} = 1$ and no iterations are required.

The reduced charge (7) applied together with (2) to the continuity equation (1) leads to the following BEM formulation:

$$E^{-}_{ni} - \frac{\varepsilon_{r,Ins} + \varepsilon_{r,Gas}}{\varepsilon_{r,Ins} - \varepsilon_{r,Gas}} \frac{\sigma_i}{2\varepsilon_0} = \frac{\sigma_{si}}{\varepsilon_0(\varepsilon_{r,Ins} - \varepsilon_{r,Gas})}. \tag{8}$$

2.4 Restrikes

Restrikes, called also back discharges, may occur due to changes of the applied voltage. For example, the maximum voltage applied to the rod during the standard lightning impulse test 1.2/50 µs lasts approximately a few microseconds. The rod is grounded after a few hundreds microseconds, see Fig 1d. Due to the charge accumulated on the barrier a new inception may be initiated at the grounded rod tip. The new discharge will bring the charge of the opposite polarity to the dielectric, which will recombine with the previously accumulated charge reducing its total amount by a value of $Q_{removed}$. We assume that in the new equilibrium the normal field strength component at the dielectric will converge to a constant value E_{nConst} within a surface region affected by the charge removal. For a collocation point i within this region the charge density σ_{si} calculated in saturation or subsaturation stage will be reduced by an unknown value $\sigma_{\Delta i}$. With these assumptions the continuity equation can be split in two separate equations like in (5), but the additional terms related to E_{nConst} and $\sigma_{\Delta i}$ must be included as follows:

$$\varepsilon_{r,Gas}\varepsilon_0 E_{ni,Gas} = \varepsilon_{r,Gas}\varepsilon_0 E_{nConst} \tag{9}$$

$$\varepsilon_{r,Ins}\varepsilon_0 E_{ni,Ins} = \varepsilon_{r,Gas}\varepsilon_0 E_{nConst} + \sigma_{si} - \sigma_{\Delta i} \tag{10}$$

After introducing (2) and moving all unknowns to the left hand side the following BEM formulation can be obtained:

$$E_{ni}^{-} + \frac{\sigma_i}{2\varepsilon_0} - E_{nConst} = 0 \tag{11}$$

$$E_{ni}^{-} - \frac{\sigma_i}{2\varepsilon_0} + \frac{\sigma_{\Delta i}}{\varepsilon_0\varepsilon_{r,Ins}} - \frac{\varepsilon_{r,Gas}}{\varepsilon_{r,Ins}} E_{nConst} = \frac{\sigma_{si}}{\varepsilon_0\varepsilon_{r,Ins}} \tag{12}$$

The unknown value of E_{nConst} requires an additional equation specifying the amount of removed charge as a fraction of the total accumulated charge:

$$\sum_i \sigma_{\Delta i} S_i = Q_{removed} = (1 - k_{rConst}) Q_{total} \tag{13}$$

where $\sigma_{\Delta i}$ is the surface charge density removed in a point i, S_i is the surface area assigned to point i and Q_{total} is the total amount of surface charge $Q_{total} = \sum_i \sigma_{si} S_i$. The factor k_{rConst}, representing the fraction of the remaining charge, has a value in the range between 0 and 1, which has to be estimated iteratively using the similar criterion like in Sect. 2.3: the inception voltage initiating the re-strike U_{incR} should be equal to U_{appl}. The whole restrike computation can be skipped if initially $U_{incR} > U_{appl}$. Examples of the computed charge density (volcano-shaped curve) and normal field distributions are shown in Fig. 2b.

2.5 Surface Potentials

The last step in evaluation of surface charging is computation of the measured surface potentials. Typically a measurement has to be performed in a different geometrical configuration, which may significantly differ from the initial one used for background field, saturation and re-strikes. An example is shown in Fig. 1e where the active rod electrode has been replaced by a measurement probe (neglected in simulations). This requires re-computation of the whole model while preserving the already computed surface charge. (The charge remains unchanged because in this stage all discharge activities are finished and no additional charge is delivered.) For all charged points i Eqs. (7) and (8) can be used with the factor $k_{sConst} = 1$ or smaller if decaying effects should be considered.

3 Iterative Procedure

When using a static approach only snapshots of the final or intermediate charging stages can be evaluated. For complex geometrical configurations such an analysis is not straightforward and may require several computational steps in order to properly reflect the process of surface charge accumulation and the related discharge development. We propose an iterative procedure consisting of the following steps:

1. Compute electrostatic background field without any surface charge (4).
2. Find a location of saturation boundary condition and compute the corresponding saturation charge density according to (6):

   (a) Evaluate the critical spots and identify points with the lowest inception voltage
   (b) Select a discharge path starting from the most critical point and ending at a dielectric
   (c) Find and verify the surface patch for saturation boundary condition. This patch must fulfill the following criteria:

   - it must include the point where the discharge arrived (“seed point”)
   - the initial patch includes all neighboring points with the same orientation of the normal field component as in the “seed point”
   - the polarity of the resulting charge density must be the same as the polarity of the discharge; points with the opposite polarity of surface charge density must be rejected
   - surface patches detached from the “seed point” must be rejected
   - all points within the patch must fulfill the stability field criterion [1]: distance from the discharge start point is not larger than $U_{appl}/E_{stability}$.

> **Note:** For complex geometries the above procedure may require several steps (typically 2–4) including re-computation of saturation charge for the corrected patch. For simple examples like in Fig. 2 the surface patch represented by a circle of approximately 70 mm radius ($= 35$ kV/0.5 kV mm^{-1}) could be correctly defined within the first iteration.

3. Compute sub-saturation according to (7)–(8) if required.
4. Repeat steps 2 and 3 above if new inception points and possible discharges appeared due to computed surface charge. For example, the charge accumulated on the top of the barrier in Fig. 2a can trigger a new inception below the barrier, which will bring the charge of opposite polarity to the barrier bottom.
5. Compute re-strikes according to (11)–(13) if required.
6. Compute surface potentials for comparison with measurements.

4 Experimental Validation

The experimental test arrangement includes a HV rod with diameter D=7 mm (or 4 mm), a dielectric barrier 600 × 600 × 5 mm with $\epsilon_{r,Ins}$=3, and a grounded plate electrode. The rod-barrier distance, d_B, and rod-plate distance, d_p, vary between 0 and 100 mm. A standard lightning voltage impulse (LI) with 1.2/50 µs and a peak value in the range between 20 and 100 kV is applied to the rod. The positive streamer discharge initiated at the spherical rod tip r=3.5 mm (or 2 mm) deposes SC at the barrier surface. After the impulse and a possible restrike the barrier together with the grounded plane are moved to another location where the surface potential due to accumulated charge is scanned by a robot-driven measurement probe. Before applying the next impulse the barrier is cleaned with alcohol in order to remove the SC.

For comparison between computations and experiments we selected three geometrical configurations representing different combinations of physical effects that have to be considered in the iterative procedure of Sect. 3: (a) subsaturation with $k_{sConst} = 0.975$, Steps: 1, 2, 3, 6; (b) re-strike with $k_{rConst} = 0.95$, Steps: 1, 2, 5, 6; (c) charge accumulated on both barrier sides due to inception triggered by a small protrusion placed at grounded plate under the rod, Steps: 1, 2, 4, 6. The corresponding comparisons presented in Fig. 3 show reasonable agreement. Multiple measurement curves illustrate the statistical behavior obtained when repeating the experiments. More experimental results are included in [6].

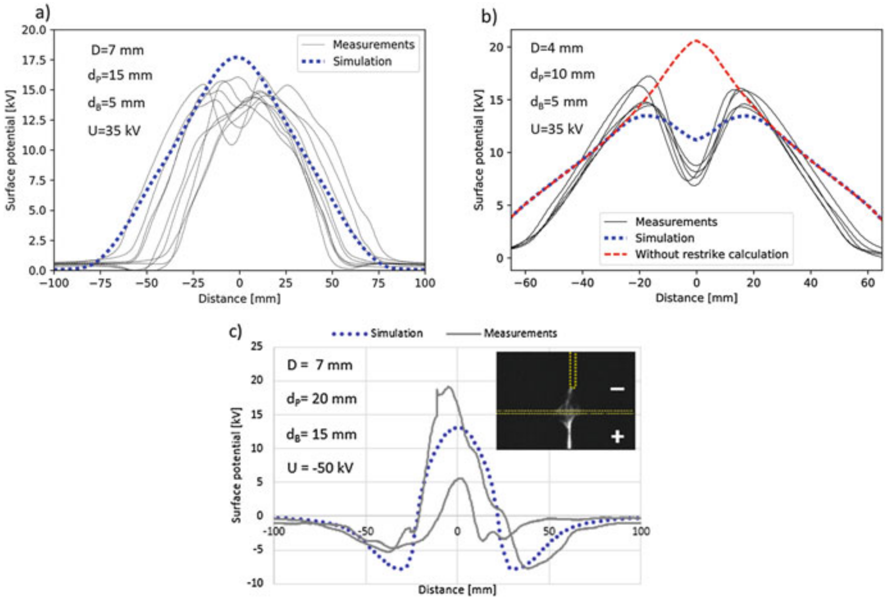


Fig. 3 Surface potential distributions for configurations with: (**a**) subsaturation, (**b**) re-strike, (**c**) charge accumulation on both barrier sides. **Note:** in case (**c**) the measurement and computation have been performed for the positively charged bottom side after removing the rod and turning the barrier around

5 Validation with a Transient Drift-Diffusion Model

Surface charging is, in general, a dynamic process, and should thus be simulated with a transient simulation. Note that the iterative procedure discussed in Sect. 3 mimics a kind of transient charging. Of course, there are different types of charging processes, e.g., by streamers, Corona, or DC ion drift, etc. with different underlying physics and which may thus lead to different details of the final charge distributions. Here we show for a specific illustrative example that the previous approach, i.e., the nullification of the normal field component at the dielectric surface, reproduces well the result, which is obtained from a drift-diffusion model for space charge in a transient simulation. The details of the drift-diffusion model are described in Refs. [7, 8] and will not be re-iterated here. It consists of the drift-diffusion equation for charge carriers with a mobility μ, which are injected from the contact. In principle, one can take into account in this model [7, 8] the effect of space charge in the Poisson equation, the effect of suppression of the inception in the electrode boundary condition model for charge injection, and the stability field in the carrier drift model. But we will include here only the effect of the surface charge density, σ, in the Poisson equation, disregard all other effects, and compare the result with the assumption of normal field nullification used in Sect. 3. Surface charging is modeled by a local surface-charge source term on the solid dielectric surface, $d\sigma/dt = j$,

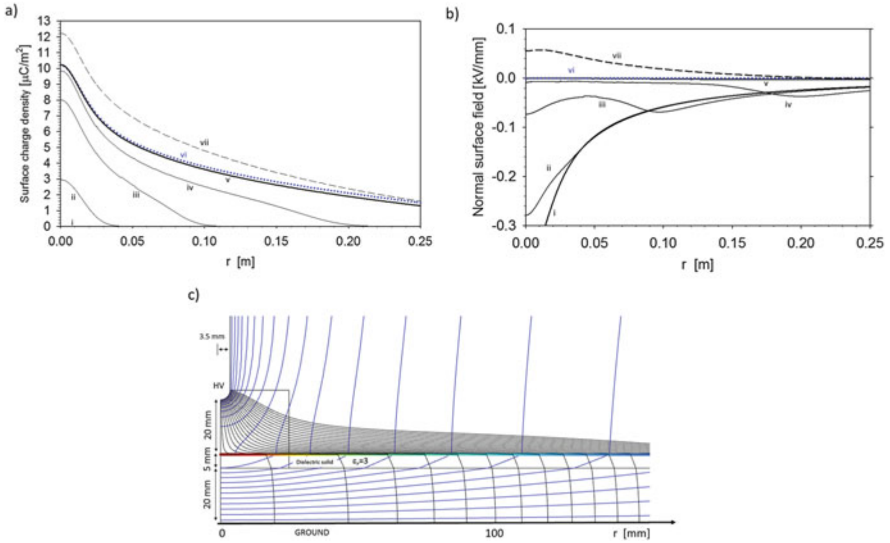


Fig. 4 (**a**) Surface charge and (**b**) normal surface field at different times during the transient charging up at 30 kV. i: capacitive state, ii: 0.2 µs, iii: 1 µs, iv: 10 µs, v: 100 µs, vi (blue dots): exact normal field nullification (Sect. 3), vii (dashed): with space charge (see text). (**c**) Simulated final state: equipotential curves (blue), field lines (black), and surface charge (color) (the box contains a refined mesh)

where j is the normal component of the current density onto the dielectric surface. The surface charge density is thus just the time integral of the current density.

The cylindrically symmetric geometry allows to perform the simulations in 2d cylindrical coordinates (r, z); the geometrical details ($D = 7$ mm, $d_p = 45$ mm, $d_B = 25$ mm) are sketched in Fig. 4. Furthermore, although we will not discuss details of the charging dynamics, we mention that there are two quantities which affect the duration of the charging process: the speed of the charge propagation, and the injection current density. The speed is generally very high for streamers as compared to, e.g., ion drift velocities. Although it is rather artificial to model streamers by a charge density cloud, we will assume a carrier mobility of $\mu \approx 1$ m^2/Vs, which leads in fields of the order of a few kilo-volts per millimeter to velocities which are comparable to typical streamer velocities. Nevertheless, due to the artificiality of the model, the mobility value should not be taken too serious but rather as a mean to control the characteristic time scale.

The simulation results are shown in Fig. 4. Parts (a) and (b) provide the space charge and field distributions, respectively, at different times during charging up. The final saturated state (curve v) is in good accordance with the normal field nullification approach obtained from a separate simulation, shown as curve vi. Of course, if one includes further phenomena, like space charge effects, the normal field component does not necessarily vanish on charged surfaces. As an example, the dashed curves in Fig. 4a and b show the result for a case study, where the transient is simulated with taking space charge into account. The presence of space charge increases the accumulated saturation charge density by approximately 20% (curve vii in Fig. 4a). After charging (steady state) the space charge is removed, such that the final field distribution is only due to the applied voltage and the surface charge. The normal field, which nullifies in presence of space charge, leads to a nonzero reversed field when the positive space charge is removed (curve vii in Fig. 4b). However, the inclusion of space charge can lead to a strongly nonlinear behavior (e.g., the formation of space charge limited currents [7, 8]), and requires additional justification and validation which is not the purpose here.

6 Conclusion

A comparison with experiments and transient modelling indicates that the numerically efficient steady-state surface charging model based on the discussed saturation concept can be used for a reasonable prediction of field characteristics during high voltage tests.

References

1. Blaszczyk, A., Ekeberg, J., Pancheshnyi, S., Saxegaard, M.: Virtual high voltage lab. In: SCEE 2016. Mathematics in Industry. Springer, Heidelberg (2018)
2. De Kock, N., Mendik, M., Andjelic, Z., Blaszczyk, A.: Application of 3D boundary element method in the design of EHV GIS components. IEEE Mag. Electr. Insul. **14**(3), 17–22 (1998)
3. Blaszczyk, A.: Region-oriented BEM formulation for numerical computations of electric fields. In: SCEE 2008. Mathematics in Industry. Springer, Heidelberg (2010)
4. Tozoni, O.B., Mayergoyz, I.D.: Calculation of Three Dimensional Electromagnetic Fields. Technika, Kiev (1974) (in Russian language)
5. Andjelic, Z., Krstajic, B., Milojkovic, S., Blaszczyk, A., Steinbigler, H., Wohlmuth, M.: Integral Methods for the Calculation of Electric Fields. Scientific Series of the International Bureau. Research Center Jülich, Jülich (1992). ISBN 3-89336-084-0
6. Meyer, H.K., Blaszczyk, A., Schueller, M., Mauseth, F., Pedersen, A.: Surface charging of dielectric barriers in short rod-plane air gaps - experiments and simulations. In: IEEE Conf. on High Voltage Engineering and Application, ICHVE, Athens, September 2018
7. Christen, T.: FEM simulation of space charge, interface and surface charge formation in insulating media. In: XVth International Symposium on High Voltage Engineering, Ljubliana, 7, T8-54 (2007)
8. Christen, T.: Nonstandard high-voltage electric insulation models. In: Comsol Conference, Milano (2012)

A Mass-Lumped Mixed Finite Element Method for Maxwell's Equations

Herbert Egger and Bogdan Radu

Abstract A novel mass-lumping strategy for a mixed finite element approximation of Maxwell's equations is proposed which on structured orthogonal grids coincides with the spatial discretization of the Yee scheme. The proposed method, however, generalizes naturally to unstructured grids and anisotropic materials and thus yields a natural variational extension of the Yee scheme for these situations.

1 Introduction

We consider the propagation of electromagnetic radiation through a linear non-dispersive and non-conducting medium described by Maxwell's equations

$$\epsilon \partial_t \mathbf{E} = \operatorname{curl} \mathbf{H}, \tag{1}$$

$$\mu \partial_t \mathbf{H} = -\operatorname{curl} \mathbf{E}. \tag{2}$$

Here $\mathbf{E}$, $\mathbf{H}$ denote the electric and magnetic field intensities and ϵ, μ are the symmetric and positive definite permittivity and permeability tensors. For ease of notation, we assume that $\mathbf{E} \times \mathbf{n} = 0$ on the boundary. The space discretization of (1)–(2) usually leads to finite dimensional differential equations of the form

$$\mathrm{M}_\epsilon \partial_t \mathbf{e} = \mathrm{C}' \mathbf{h}, \tag{3}$$

$$\mathrm{M}_\mu \partial_t \mathbf{h} = -\mathrm{C}\mathbf{e}. \tag{4}$$

H. Egger (✉)
Department of Mathematics, TU Darmstadt, Darmstadt, Germany
e-mail: egger@mathematik.tu-darmstadt.de

B. Radu
Graduate School for Computational Engineering, TU Darmstadt, Darmstadt, Germany
e-mail: radu@gsc.tu-darmstadt.de

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_2

Due to the particular structure of the system, the stability of such discretization schemes can easily be ensured by the simple algebraic conditions

(i) $\mathrm{C}' = \mathrm{C}^\top$,
(ii) $\mathrm{M}_\epsilon, \mathrm{M}_\mu$ symmetric and positive definite.

In order to enable an efficient solution of (3)–(4) by explicit time-stepping methods, one additionally has to assume that

(iii) $\mathrm{M}_\epsilon^{-1}, \mathrm{M}_\mu^{-1}$ can be applied efficiently.

The finite difference approximation of (1)–(2) on staggered orthogonal grids yields approximations of the form (3)–(4) satisfying the conditions (i)–(iii) with diagonal matrices M_ϵ, M_μ [13]. Moreover, the entries $\mathbf{e}_i$, $\mathbf{h}_j$ in the solution vectors yield second order approximations for the line integrals of $\mathbf{E}$, $\mathbf{H}$ along edges of the primal and dual grids [3, 12]. An extension to unstructured grids and anisotropic coefficients is in principle possible, but these approaches rely on the use of two sets of unstructured grids [2, 11] which makes a rigorous convergence analysis rather difficult.

The finite element approximation of (1)–(2) on the other hand yields systems of the form (3)–(4) satisfying conditions (i)–(ii) automatically and a rigorous convergence analysis is possible in rather general situations [7–9]. Although the matrices M_ϵ and M_μ are usually sparse, condition (iii) is here in general not valid. The resulting lack of efficiency can however be overcome by appropriate *mass-lumping* [4, 6], which aims at approximating M_ϵ and M_μ by diagonal or block-diagonal matrices. These approaches are usually based on an enrichment of the approximation spaces and appropriate quadrature; see [3] for details and further references.

In this paper, we present a novel mass-lumping strategy for a mixed finite element approximation of (1)–(2) that yields properties (i)–(iii) without such an increase of the system dimension. We further show that in special cases, i.e., for orthogonal grids and scalar coefficients, the resulting scheme reduces to the staggered-grid finite difference approximation of the Yee scheme.

2 A Mass-Lumped Mixed Finite Element Method

As a preliminary step, we consider a mass-lumped mixed finite element approximation based on enriched approximation spaces and numerical quadrature. We seek for approximations $\widetilde{\mathbf{E}}_h(t) \in \widetilde{V}_h$, $\widetilde{\mathbf{H}}_h(t) \in \widetilde{Q}_h$ satisfying

$$(\epsilon \partial_t \widetilde{\mathbf{E}}_h(t), \widetilde{\mathbf{v}}_h)_h = (\widetilde{\mathbf{H}}_h(t), \operatorname{curl} \widetilde{\mathbf{v}}_h) \qquad \forall \widetilde{\mathbf{v}}_h \in \widetilde{V}_h, \tag{5}$$

$$(\mu \partial_t \widetilde{\mathbf{H}}_h(t), \widetilde{\mathbf{q}}_h)_{h,*} = -(\operatorname{curl} \widetilde{\mathbf{E}}_h(t), \widetilde{\mathbf{q}}_h) \qquad \forall \widetilde{\mathbf{q}}_h \in \widetilde{Q}_h, \tag{6}$$

for all $t > 0$. Here, $\widetilde{V}_h \subset H_0(\mathrm{curl}; \Omega)$ and $\widetilde{Q}_h \subset L^2(\Omega)$ are appropriate finite dimensional subspaces and $(\mathbf{a}, \mathbf{b})_h$, $(\mathbf{a}, \mathbf{b})_{h,*}$ are approximations for usual the scalar product $(\mathbf{a}, \mathbf{b}) = \int_\Omega \mathbf{a}(x) \cdot \mathbf{b}(x)\, dx$ to be defined below.

In the sequel, we restrict our discussion to problems where $\mathbf{E} = (E_x, E_y, 0)$ and $\mathbf{H} = (0, 0, H_z)$ with E_x, E_y, H_z independent of z, which allows to represent the fields in two dimensions. The extension to three dimensions will be discussed in Sect. 5.

Let $\mathcal{T}_h = \{T\}$ be a conforming mesh of Ω consisting of triangles and parallelograms. Any element $T \in \mathcal{T}_h$ is the image $F_T(\widehat{T})$ of a reference triangle or reference square under an affine mapping $F_T(\widehat{x}) = a_T + B_T\widehat{x}$ with $a_T \in \mathbb{R}^2$ and $B_T \in \mathbb{R}^{2\times 2}$. We denote by h the maximal element diameter and assume uniform shape regularity.

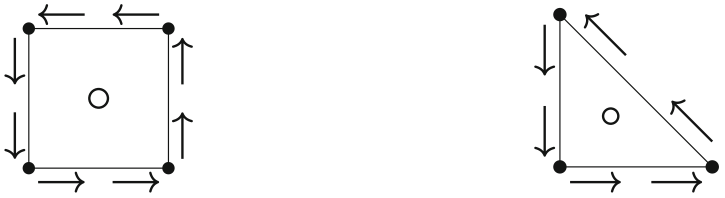
To every element T_j, $j = 1, \ldots, n_T$ of the mesh, we associate one basis function $\widetilde{\psi}_j$ of the space $\widetilde{Q}_h$ with $\widetilde{\psi}_j|_{T_k} = \delta_{jk}$. For every interior edge $e_i = T_l \cap T_r$, $i = 1, \ldots, n_e$ of the mesh, we further define two basis functions $\widetilde{\phi}_i, \widetilde{\phi}_{i+n_e}$ which are defined by

$$\widetilde{\phi}_{i+\ell\cdot n_e}|_T = B_T^{-\top}\widehat{\phi}_{\alpha,\gamma}, \qquad \ell = 0, 1, \tag{7}$$

on $T \in \{T_l, T_r\}$ and vanish identically on all other elements. Here $\alpha \in \{1, \ldots, \widehat{n}_e\}$ refers to the number of the edge e_i on the reference element $\widehat{T}$ and $\gamma \in \{0, 1\}$ depends on ℓ and the orientation of the edge e_i. The functions $\widehat{\phi}_{\alpha,\gamma}$ are defined in Fig. 1. Similar approximation spaces in three dimensions were utilized in [8, 9]. We further set $(\mathbf{a}, \mathbf{b})_{h,*} = (\mathbf{a}, \mathbf{b})$ and define $(\mathbf{a}, \mathbf{b})_h = \sum_T (\mathbf{a}, \mathbf{b})_{h,T}$ with

$$(\mathbf{a}, \mathbf{b})_{h,T} = |T| \sum_{l=1}^{\widehat{n}_p} \mathbf{a}(F_T(\widehat{x}_l)) \cdot \mathbf{b}(F_T(\widehat{x}_l))\, w_l, \tag{8}$$

where $w_l = 1/\widehat{n}_p$ denote the quadrature weights and $\widehat{x}_l$, $l = 1, \ldots, \widehat{n}_p$ the quadrature points on the reference element, depicted by dots in Fig. 1.



$$\widehat{\phi}_{1,0} = \tfrac{1}{2}\binom{-y^2+y}{-2xy+2x}, \quad \widehat{\phi}_{1,1} = \tfrac{1}{2}\binom{y^2-y}{2xy},$$
$$\widehat{\phi}_{2,0} = \tfrac{1}{2}\binom{-2xy}{-x^2+x}, \quad \widehat{\phi}_{2,1} = \tfrac{1}{2}\binom{-2y+2xy}{x^2-x},$$
$$\widehat{\phi}_{3,0} = \tfrac{1}{2}\binom{y^2-y}{2xy-2y}, \quad \widehat{\phi}_{3,1} = \tfrac{1}{2}\binom{-y^2+y}{2x+2y-2xy-2},$$
$$\widehat{\phi}_{4,0} = \tfrac{1}{2}\binom{-2x-2y+2xy+2}{x^2-x}, \quad \widehat{\phi}_{4,1} = \tfrac{1}{2}\binom{-2xy+2x}{-x^2+x}.$$

$$\widehat{\phi}_{1,0} = \tfrac{1}{2}\binom{0}{x}, \quad \widehat{\phi}_{1,1} = \tfrac{1}{2}\binom{-y}{0},$$
$$\widehat{\phi}_{2,0} = \tfrac{1}{2}\binom{-y}{-y}, \quad \widehat{\phi}_{2,1} = \tfrac{1}{2}\binom{0}{x+y-1},$$
$$\widehat{\phi}_{3,0} = \tfrac{1}{2}\binom{1-x-y}{0}, \quad \widehat{\phi}_{3,1} = \tfrac{1}{2}\binom{x}{x}.$$

Fig. 1 Degrees of freedom and basis functions for the unit triangle and unit square. The black dots at the vertices represent the quadrature points for the quadrature formula introduced below

Using the bases defined above, all functions in $\widetilde{V}_h$ and $\widetilde{Q}_h$ can be represented as

$$\widetilde{\mathbf{E}}_h = \sum\nolimits_i \widetilde{\mathbf{e}}_i \widetilde{\phi}_i + \widetilde{\mathbf{e}}_{i+n_e} \widetilde{\phi}_{i+n_e} \qquad \text{and} \qquad \widetilde{\mathbf{H}}_h = \sum\nolimits_j \widetilde{\mathbf{h}}_j \widetilde{\psi}_j. \tag{9}$$

This allows to rewrite the variational problem (5)–(6) in algebraic form as

$$\widetilde{\mathrm{M}}_\epsilon \partial_t \widetilde{\mathbf{e}} = \widetilde{\mathrm{C}}^\top \widetilde{\mathbf{h}}, \tag{10}$$

$$\widetilde{\mathrm{M}}_\mu \partial_t \widetilde{\mathbf{h}} = -\widetilde{\mathrm{C}}\, \widetilde{\mathbf{e}}, \tag{11}$$

with matrices $(\widetilde{\mathrm{M}}_\epsilon)_{ij} = (\epsilon \widetilde{\phi}_j, \widetilde{\phi}_i)_h$, $(\widetilde{\mathrm{M}}_\mu)_{ij} = (\mu \widetilde{\psi}_j, \widetilde{\psi}_i)$, and $(\widetilde{\mathrm{C}})_{ij} = (\operatorname{curl} \widetilde{\phi}_j, \widetilde{\psi}_i)$. As a direct consequence of the particular choice of the basis functions, we obtain

Lemma 1 *Let $\widetilde{\mathrm{M}}_\epsilon$, $\widetilde{\mathrm{M}}_\mu$, and $\widetilde{\mathrm{C}}$ be defined as above. Then conditions (i)–(iii) hold.*

Proof The properties (i)–(ii) follow directly from the definition of the matrices and the symmetric positive definiteness of the material tensors. Since the basis functions for $\widetilde{Q}_h$ are supported only on single elements, one can see that $\widetilde{\mathrm{M}}_\mu$ is diagonal. To see the block-diagonal structure of $\widetilde{\mathrm{M}}_\epsilon$, let us refer to Fig. 2. In the left plot, the degrees of freedom for $\widetilde{V}_h$ are depicted by the red arrows and the quadrature points by blue circles. By definition, the corresponding basis functions are zero in all vertices, except the one which the arrow representing the corresponding degree of freedom originates from. Together with the nodal quadrature formula, this reveals that only groups of basis functions associated with same vertex yield non-zero contributions to the mass matrix $\widetilde{\mathrm{M}}_\epsilon$. In the right plot of Fig. 2, we depict the structure of the inverse of $\widetilde{\mathrm{M}}_\epsilon^{-1}$. Each block here corresponds to the degrees of freedom associated to one of the vertices in the mesh and the size of the block is determined by the number of edges incident to the corresponding vertex. Note

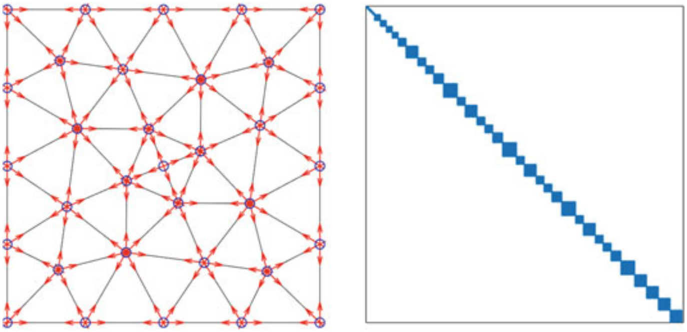


Fig. 2 Location of degrees of freedom for the basis function of the space $\widetilde{V}_h$ (left), and structure of the matrix $\widetilde{\mathrm{M}}_\epsilon^{-1}$ after appropriate numbering of degrees of freedom (right)

that an appropriate numbering of the degrees of freedom is required to see the block diagonal structure so clearly. □

Let us mention that the quadrature rule satisfies $(\mathbf{a}, \mathbf{b})_{h,T} = \int_T \mathbf{a}(x) \cdot \mathbf{b}(x)\, dx$ when $\mathbf{a}(x) \cdot \mathbf{b}(x)$ is affine linear. This ensures that the method (5)–(6) also has good approximation properties. By a slight adoption of the results given in [5], we obtain

Lemma 2 *Let* $\mathbf{E}$, $\mathbf{H}$ *be a smooth solution of (1)–(2) and let* $\widetilde{\mathbf{E}}_h(0)$ *and* $\widetilde{\mathbf{H}}_h(0)$ *be chosen appropriately. Then*

$$\|\widetilde{\mathbf{E}}_h(t) - \mathbf{E}(t)\|_{L^2(\Omega)} + \|\widetilde{\mathbf{H}}_h(t) - \mathbf{H}(t)\|_{L^2(\Omega)} \le Ch,$$

for all $0 \le t \le T$ *with* $C = C(\mathbf{E}, \mathbf{H}, T)$. *Moreover,* $\|\widetilde{\mathbf{H}}_h(t) - \pi_h^0 \mathbf{H}(t)\|_{L^2(\Omega)} \le Ch^2$ *where* $\pi_h^0 \mathbf{H}$ *denotes the piecewise constant approximation of* $\mathbf{H}$ *on the mesh* $\mathcal{T}_h$.

Remark 1 For structured meshes and isotropic coefficients, one can observe second order convergence also for line integrals of the electric field along edges of the mesh. Second convergence for the electric field can also be obtained for unstructured meshes by a non-local post-processing strategy; see [5] for details.

3 A Variational Extension of the Yee Scheme

The method of the previous section already yields a stable and efficient approximation. We now show that one degree of freedom per edge can be saved without sacrificing the accuracy or efficiency of the method. To this end, we construct approximations $\mathbf{E}_h(t) \in V_h$, $\mathbf{H}_h(t) \in Q_h$ in spaces $V_h \subset \widetilde{V}_h$ and $Q_h = \widetilde{Q}_h$.

We again define one basis function ψ_j of Q_h for every element T_k by $\psi_j|_{T_k} = \delta_{jk}$. To any edge $e_i = T_l \cap T_r$, we now associate one single basis function ϕ_i defined by

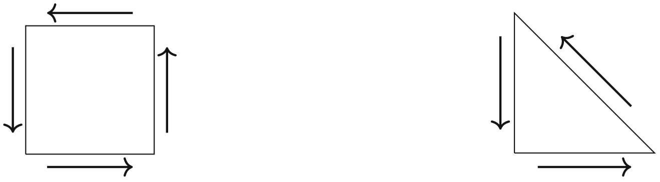
$$\phi_i = \widetilde{\phi}_i + \widetilde{\phi}_{i+n_e}. \tag{12}$$

Using the construction of $\widetilde{\phi}_i$, one can give an equivalent definition of ϕ_i via

$$\phi_i|_T = B_T^{-\top} \widehat{\phi}_\alpha, \qquad T \cap e_i \neq \emptyset, \tag{13}$$

with basis functions $\widehat{\phi}_\alpha = \widehat{\phi}_{\alpha,0} + \widehat{\phi}_{\alpha,1}$ defined on the reference element in Fig. 3. Note that the space V_h coincides with the Nedelec space of lowest order [1, 10]. Any function $\mathbf{E}_h \in V_h$ and $\mathbf{H}_h \in Q_h$ can now be expanded as

$$\mathbf{E}_h = \sum\nolimits_i \mathbf{e}_i \phi_i \qquad \text{and} \qquad \mathbf{H}_h = \sum\nolimits_j \mathbf{h}_j \psi_j. \tag{14}$$



$\widehat{\phi}_1 = \binom{0}{x}, \quad \widehat{\phi}_2 = \binom{-y}{0},$
$\widehat{\phi}_3 = \binom{0}{x-1}, \quad \widehat{\phi}_4 = \binom{1-y}{0}.$

$\widehat{\phi}_1 = \frac{1}{2}\binom{-y}{x}, \quad \widehat{\phi}_2 = \frac{1}{2}\binom{-y}{x-1}$
$\widehat{\phi}_3 = \frac{1}{2}\binom{1-y}{x}.$

Fig. 3 Degrees of freedom and basis functions on the unit triangle and unit square

As a consequence of (12), any $\mathbf{E}_h \in V_h$ can be interpreted as function $\widetilde{\mathbf{E}}_h \in \widetilde{V}_h$ by

$$\mathbf{E}_h = \sum\nolimits_i \mathrm{e}_i \phi_i = \sum\nolimits_i \mathrm{e}_i (\widetilde{\phi}_i + \widetilde{\phi}_{i+n_e}) = \sum\nolimits_i \mathrm{e}_i \widetilde{\phi}_i + \mathrm{e}_i \widetilde{\phi}_{i+n_e} = \widetilde{\mathbf{E}}_h. \tag{15}$$

The coordinates of $\widetilde{\mathbf{E}}_h$ and $\mathbf{E}_h$ are thus simply connected by $\widetilde{\mathrm{e}}_i = \widetilde{\mathrm{e}}_{i+n_e} = \mathrm{e}_i$. Vice versa, we can associate to any function $\widetilde{\mathbf{E}}_h \in \widetilde{V}_h$ a function $\mathbf{E}_h = \Pi_h \widetilde{\mathbf{E}}_h \in V_h$ by defining its coordinates as $\mathrm{e}_i = \frac{1}{2}(\widetilde{\mathrm{e}}_i + \widetilde{\mathrm{e}}_{i+n_e})$. In linear algebra notation, this reads

$$\mathrm{e} = \mathrm{P}\widetilde{\mathrm{e}}, \tag{16}$$

with projection matrix P defined by $\mathrm{P}_{ij} = \frac{1}{2}$ if $j = i$ or $j = i + n_e$, and $\mathrm{P}_{ij} = 0$ else.

We now define system matrices for the system (3)–(4) by $(\mathrm{M}_\mu)_{ij} = (\mu\psi_j, \psi_i)$, $\mathrm{C}_{ij} = \mathrm{C}'_{ji} = (\operatorname{curl}\phi_j, \psi_i)$, and $\mathrm{M}_\epsilon^{-1} = \mathrm{P}\widetilde{\mathrm{M}}_\epsilon^{-1}\mathrm{P}^\top$, where $\widetilde{\mathrm{M}}_\epsilon$ is defined as in the previous sections. This construction has the following properties.

Lemma 3 *Let* M_μ*,* C*,* C′*, and* M_ϵ^{-1} *be defined as above, and set* $\mathrm{M}_\epsilon = (\mathrm{M}_\epsilon^{-1})^{-1}$*. Then the conditions (i)–(iii) are satisfied.*

Proof Condition (i) follows by construction. The matrix M_μ is diagonal and positive definite and therefore M_μ^{-1} has the same properties. This verifies (ii) and (iii) for the matrix M_μ. Since P is sparse and has fully rank and $\widetilde{\mathrm{M}}_\epsilon^{-1}$ is block diagonal, symmetric, and positive definite, one can see that also M_ϵ^{-1} is sparse, symmetric, and positive-definite. This verifies conditions (ii) and (iii) for M_ϵ. □

In the following, we investigate more closely the relation of the system (3)–(4) with matrices as defined above and the system (10)–(11) discussed in the previous section. We start with an auxiliary result.

Lemma 4 *Let* C*,* P*, and* $\widetilde{\mathrm{C}}$ *be defined as above. Then one has* $\widetilde{\mathrm{C}} = \mathrm{CP}$.

Proof The result follows directly from the construction. □

As a direct consequence, we can reveal the following close connection between the methods (3)–(4) and (10)–(11) discussed in the preceding sections.

Lemma 5 *Let $\widetilde{\mathbf{e}}(t)$, $\widetilde{\mathbf{h}}(t)$ be a solution of (10)–(11). Then $\mathbf{e}(t) = \mathrm{P}\widetilde{\mathbf{e}}(t)$, $\mathbf{h}(t) = \widetilde{\mathbf{h}}(t)$ solves (3)–(4) with matrices M_ϵ, M_μ, and C as defined above.*

Proof From Eq. (10), the definition of $\mathbf{e}$, $\mathbf{h}$, and Lemma 4, we deduce that

$$\partial_t \mathbf{e} = \mathrm{P}\partial_t \widetilde{\mathbf{e}} = \mathrm{P}\widetilde{\mathrm{M}}_\epsilon^{-1}\widetilde{\mathrm{C}}^\top\widetilde{\mathbf{h}} = \mathrm{P}\widetilde{\mathrm{M}}_\epsilon^{-1}\mathrm{P}^\top\mathrm{C}^\top\widetilde{\mathbf{h}} = \mathrm{M}_\epsilon^{-1}\mathrm{C}^\top\mathbf{h}.$$

This verifies the validity of Eq. (3). Using Eq. (11), we obtain

$$\mathrm{M}_\mu\partial_t\mathbf{h} = \widetilde{\mathrm{M}}_\mu\partial_t\widetilde{\mathbf{h}} = -\widetilde{\mathrm{C}}\,\widetilde{\mathbf{e}} = -\mathrm{CP}\,\widetilde{\mathbf{e}} = -\mathrm{C}\,\mathbf{e},$$

which verifies the validity of Eq. (4). Finally, using the discrete stability of the projection completes the proof. □

Remark 2 The vectors $\mathbf{e}(t)$, $\mathbf{h}(t)$ computed via (3)–(4) with the above choice of matrices correspond to finite element approximations $\mathbf{E}_h(t) \in V_h$, $\mathbf{H}_h(t) \in Q_h$. Therefore, the procedure described above can be interpreted as a mixed finite element method with mass-lumping based on the approximation spaces V_h and Q_h.

As an immediate consequence of Lemma 5 and the approximation results of Lemma 2, we now obtain the following assertions.

Lemma 6 *Let $\mathbf{e}(t)$, $\mathbf{h}(t)$ denote the solutions of (3)–(4) with appropriate initial conditions and set $\mathbf{E}_h(t) = \sum_i \mathbf{e}_i(t)\phi_i$, $\mathbf{H}_h(t) = \sum_j \mathbf{h}_j(t)\psi_j$. Then*

$$\|\mathbf{E}_h(t) - \mathbf{E}(t)\|_{L^2(\Omega)} + \|\mathbf{H}_h(t) - \mathbf{H}(t)\|_{L^2(\Omega)} \le Ch,$$

for all $0 < t \le T$. In addition, $\|\pi_h^0\mathbf{H}(t) - \mathbf{H}_h(t)\|_{L^2(\Omega)} \le Ch^2$ where $\pi_h^0\mathbf{H}$ denotes the piecewise constant approximation of $\mathbf{H}$ on the mesh $\mathcal{T}_h$.

By some elementary computations, one can verify the following observation.

Lemma 7 *Let $\mathcal{T}_h$ be a uniform mesh consisting of orthogonal quadrilaterals T of the same size. Furthermore, let ϵ and μ be positive constants. Then the matrices M_ϵ, M_μ, and C, defined above coincide with those obtained by the finite difference approximation on staggered grids; see [3] for the two dimensional version.*

The method proposed in this section therefore can be understood as a variational extension of the Yee scheme in the sense of [3]. In the two dimensional setting, one degree of freedom $\mathbf{e}_i$ is required for every edge, and one value $\mathbf{h}_j$ for every element.

4 Numerical Validation

Consider the domain $\Omega = (-1,1)^2 \setminus \{(x,y) : (x-0.6)^2 + y^2 \leq 0.25^2\}$, which is split by an interior boundary into $\Omega = \Omega_1 \cup \Omega_2$; see Fig. 4 for a sketch. We set $\epsilon = 1$ on Ω_1, $\epsilon = 3$ on Ω_2 and $\mu = 1$ on Ω, and consider a plane wave that enters the domain from the left boundary. The wave gets slowed down and refracted, when entering the domain Ω_2, and reflected at the circle $\partial\Omega_0$, where we enforce a perfect electric boundary conditions. For the spatial discretization, we choose the method presented in Sect. 3, while for the time discretization, we choose the leapfrog scheme. A very small time step is chosen to suppress the additional errors due to time discretization. Convergence rates for the numerical solution are depicted in table of Fig. 5 and a few snapshots of the solution are depicted Fig. 6. The error is measured in the norm $|||e||| := \max_{0\leq t^n \leq T} \|e(t^n)\|_{L^2(\Omega)}$.

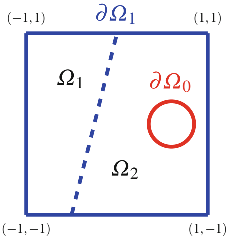


Fig. 4 Geometry

h	DOF	$\lvert\lvert\lvert \mathbf{E}_h - \pi_h \mathbf{E}_{h^*} \rvert\rvert\rvert$	eoc	$\lvert\lvert\lvert \pi_h^0(\mathbf{H}_h - \pi_h \mathbf{H}_{h^*}) \rvert\rvert\rvert$	eoc
2^{-3}	2246	0.158291	—	0.242490	—
2^{-4}	8884	0.057465	1.46	0.069676	1.80
2^{-5}	35368	0.025145	1.19	0.017157	2.02
2^{-6}	141136	0.011835	1.08	0.004064	2.07

Fig. 5 Errors and estimated order of convergence (eoc) with respect to a fine solution $(\mathbf{E}_{h^*}, \mathbf{H}_{h^*})$ for $h^* = 2^{-8}$. The total number of degrees of freedom (DOF) is also given

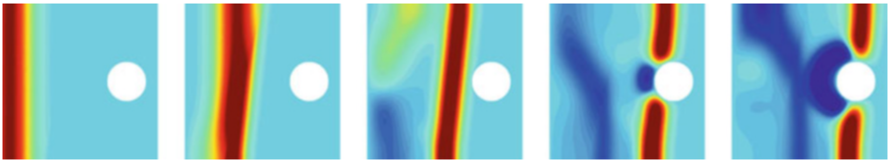


Fig. 6 Snapshots of the magnetic field intensity $\mathbf{H}_h$ for time $t = 0.8, 1.2, 1.6, 2.4, 2.8$

5 Discussion

Before we conclude, let us briefly discuss an alternative formulation and the extension to three dimensions and higher order approximations.

Remark 3 Eliminating $\mathbf{h}$ from (3)–(4) leads to a second order equation

$$\mathrm{M}_\epsilon \partial_{tt}\mathbf{e} = \mathrm{K}_{\mu^{-1}}\mathbf{e} \tag{17}$$

for the electric field vector $\mathbf{e}$ alone, with $\mathrm{K}_{\mu^{-1}} = C'\mathrm{M}_\mu^{-1}\mathrm{C}$. A sufficient condition for the stability of the scheme (17) is

(iv) M_ϵ and $\mathrm{K}_{\mu^{-1}}$ are symmetric and positive definite, respectively, semi-definite,

and for an efficient numerical integration of (17), one now requires that

(v) M_ϵ^{-1} and $\mathrm{K}_{\mu^{-1}}$ can be applied efficiently.

The conditions (iv) and (v) can be seen to be a direct consequence of the conditions (i)–(iii), and the special form $\mathrm{K}_{\mu^{-1}} = \mathrm{C}'\mathrm{M}_\mu^{-1}\mathrm{C}$ of the matrix $\mathrm{K}_{\mu^{-1}}$.

Remark 4 Using the definition of the matrices M_μ, C, and $\mathrm{C}' = \mathrm{C}^\top$ given in the previous section, one can verify that $\mathrm{K}_{\mu^{-1}}$ is given by $(\mathrm{K}_{\mu^{-1}})_{ij} = (\mu^{-1}\operatorname{curl}\phi_j, \operatorname{curl}\phi_i)$. Thus $\mathrm{K}_{\mu^{-1}}$ can be assembled without constructing C or M_μ explicitly. Moreover, the conditions (iv) and (v) for $\mathrm{K}_{\mu^{-1}}$ are satisfied automatically. The essential ingredient for a mass-lumped mixed finite element approximation of (1)–(2) thus is the construction of a positive definite and sparse matrix M_ϵ^{-1}.

Remark 5 The construction of the approximation M_ϵ discussed in Sect. 3 immediately generalizes to three space dimensions. Like in the two dimensional case, two basis functions $\widetilde{\phi}_i$, $\widetilde{\phi}_{i+n_e}$ of the space $\widetilde{V}_h$ are defined for every edge e_i of the mesh [9, 10] and the approximation $(\cdot,\cdot)_h$ is defined via numerical quadrature by the vertex rule. The lumped mass matrix given by $(\widetilde{\mathrm{M}}_\epsilon)_{ij} = (\epsilon\widetilde{\phi}_j, \widetilde{\phi}_i)_h$ then is again block-diagonal. As before, the basis functions for the space V_h are then defined by $\phi_i = \widetilde{\phi}_i + \widetilde{\phi}_{i+n_e}$ and the inverse mass matrix for the reduced space is again given by $\mathrm{M}_\epsilon^{-1} = \mathrm{P}\widetilde{\mathrm{M}}_\epsilon^{-1}\mathrm{P}^\top$ with projection matrix P of the same form as in two dimensions.

Acknowledgments The authors are grateful for support by the German Research Foundation (DFG) via grants TRR 146, TRR 154, and Eg-331/1-1 and through grant GSC 233 of the "Excellence Initiative" of the German Federal and State Governments.

References

1. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics, vol. 44. Springer, Heidelberg (2013)
2. Codecasa, L., Politi, M.: Explicit, consistent, and conditionally stable extension of FD-TD to tetrahedral grids by FIT. IEEE Trans. Magn. **44**, 1258–1261 (2008)

3. Cohen, G.: Higher-Order Numerical Methods for Transient Wave Equations. Springer, Heidelberg (2002)
4. Cohen, G., Monk, P.: Gauss point mass lumping schemes for Maxwell's equations. Numer. Methods Partial Diff. Equat. **14**, 63–88 (1998)
5. Egger, H., Radu, B.: A mass-lumped mixed finite element method for acoustic wave propagation (2018). arXive:1803.04238
6. Elmkies, A., Joly, P.: éléments finis d'arête et condensation de masse pour les équations de Maxwell: le cas de dimension 3. C. R. Acad. Sci. Paris Sér. I Math. **325**, 1217–1222 (1997)
7. Joly, P.: Variational methods for time-dependent wave propagation problems. In: Topics in Computational Wave Propagation, LNCSE, vol. 31, pp. 201–264. Springer, Berlin (2003)
8. Monk, P.: Analysis of a finite element methods for Maxwell's equations. SIAM J. Numer. Anal. **29**, 714–729 (1992)
9. Mur, G., de Hoop, A.T.: A finite-element method for computing three-dimensional electromagnetic fields in inhomogeneous media. IEEE Trans. Magn. **21**(6), 2188–2191 (1985)
10. Nedelec, J.C. : Mixed finite elements in $\mathbb{R}^3$. Numer. Math. **35**(6), 315–341 (1980)
11. Schuhmann, R., Weiland, T.: A stable interpolation technique for FDTD on non-orthogonal grids. Int. J. Numer. Model. **11**, 299–306 (1998)
12. Weiland, T.: Time domain electromagnetic field computation with finite difference methods. Int. J. Numer. Model. **9**, 295–319 (1996)
13. Yee, K.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **AP-16**, 302–307 (1966)

Adaptive Mesh Refinement for Rotating Electrical Machines Taking into Account Boundary Approximation Errors

Armin Fohler and Walter Zulehner

Abstract In this work we present an error estimator for a class of second order quasilinear elliptic problems in 2D. The computational domain consists of two parts—called rotor and stator in the framework of electrical motors—separated by a curvilinear interface. For the coupling of the rotor and the stator on the interface we use a Nitsche technique as described in Hollaus et al. (Nitsche-type mortaring for Maxwell's equations, In: Progress in electromagnetics research symposium proceedings, Cambridge, pp 397–402, 5–8 July 2010). The residual error estimator is constructed similarly to the approach used in Houston et al. (IMA J Numer Anal 28:245–273, 2008) with adaptations due to the coupling strategy. The error estimator takes into account the polygonal approximation of the stator and the rotor using ideas from hierarchical error estimates.

1 Motivation and Introduction

During an optimization process of a rotating electrical machine with respect to specific requirements on the performance many different designs have to be simulated. In each simulation a complete rotation cycle of the motor has to be performed. The vast amount of simulations calls for efficient concepts to reduce the computational costs.

Our focus is on the introduction of a domain decomposition approach that allows an independent meshing of the two motor parts. This decomposition gives us the flexibility to simulate the whole rotation cycle without remeshing. Furthermore, to reduce the complexity of the simulation while attaining the same accuracy of the calculation we are interested in a mesh that is as coarse as possible and as fine as

A. Fohler (✉)
Linz Center of Mechatronics, Linz, Austria
e-mail: armin.fohler@lcm.at

W. Zulehner
Johannes Kepler University Linz, Institute of Computational Mathematics, Linz, Austria
e-mail: zulehner@numa.uni-linz.ac.at

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_3

necessary. To achieve an optimal distribution of the degrees of freedom in our finite element mesh we use an adaptive mesh refinement procedure.

2 Model Problem

To model the behaviour of the electrical machine we use the magnetostatic case of Maxwell's equations in 2D. Since we are interested in modelling a rotating electrical machine our computational domain Ω decomposes—as already mentioned—into two different parts. One part is fixed—we will call it stator Ω_S—and one part is moving—the rotor Ω_R. Both parts consist of various material domains Ω_i, $i \in I$. In our considerations we added to the physical stator and rotor an outer and inner air domain as well as parts of the motor-air-gap to each of them respectively. Additionally we introduce the interface

$$\Gamma = \partial \overline{\Omega_S} \cap \partial \overline{\Omega_R},$$

where $\partial \Omega_k$ represents the boundary of the respective domain.

Remark 1 In our application we consider only interfaces Γ that are circular (Fig. 1).

The variational formulation for the continuous problem then has the following form: Find $u \in H_0^1(\Omega)$ such that

$$a(u, w) = \langle f, w \rangle \qquad\qquad \forall w \in H_0^1(\Omega) \tag{1}$$

with the semi-linear form

$$a(u, w) = \sum_{i \in I} \left(\int_{\Omega_i} \nu_i(|\nabla u_i|) \nabla u_i \cdot \nabla w_i \right)$$

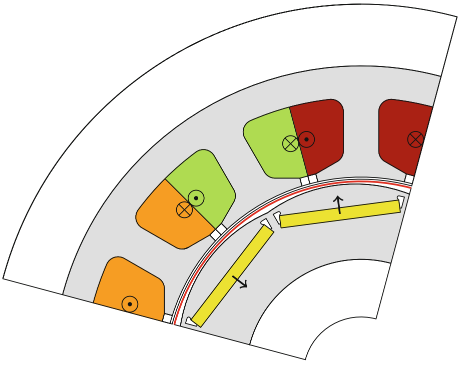


Fig. 1 Example of the cross-section of a motor sector. Iron parts are colored in gray, magnets are depicted in yellow. Current domains and their connections are indicated in orange, green and red

and the linear form

$$\langle f, w\rangle = \sum_{i\in I}\left(\int_{\Omega_i} J_{3,i} w_i + \int_{\Omega_i} \nu_M \mathbf{M}_{i,\perp}\nabla w_i\right).$$

Here u_i is the restriction of the solution u to the domain Ω_i. The right-hand side includes contributions from the current $J_{3,i}$ in the coils of the electrical machine and from the magnetization of domains representing permanent magnets, denoted by $\mathbf{M}_i = (M_{1,i}, M_{2,i})$. The material properties in Ω_i is described by the function ν_i, which is non-linear in iron. In air we use a constant value ν_0. $H^s(\Omega)$ and $H_0^s(\Omega)$ denote the usual Sobolev spaces.

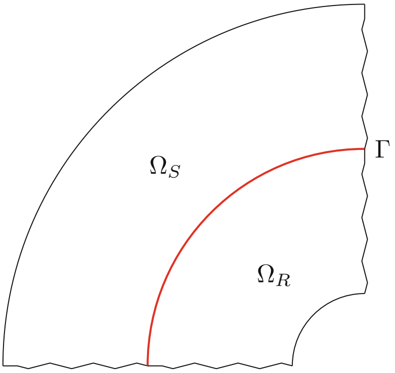
2.1 Nitsche Coupling

The Nitsche coupling is a domain decomposition technique see, e.g., [2, 3]. The idea is to penalize the jump of the solution on the two adjacent domains Ω_S and Ω_R across the interface Γ. For this the trace of u on Γ is introduced as a new variable λ (Fig. 2).

2.1.1 Discrete Problem

We approximate the domains Ω_S and Ω_R by admissible triangular meshes $\mathcal{T}_{S,h}$ and $\mathcal{T}_{R,h}$. We require that all nodes of an element T lie inside one material domain Ω_i. With this assumption we write ν_T instead of ν_i for this triangle. We assume that the meshes are shape-regular. The interface Γ is subdivided by elements $\mathcal{E}_\Gamma = \{e\}$ with $\Gamma = \bigcup_{e\in\mathcal{E}_\Gamma} \overline{e}$. Note that the elements of $\mathcal{E}_\Gamma$ are curvilinear and represent Γ accurately. So the overall mesh is described by the triple $\mathcal{M}_h = (\mathcal{T}_{S,h}, \mathcal{T}_{R,h}, \mathcal{E}_\Gamma)$ (Fig. 3).

Fig. 2 Schema of the two different motor parts: the rotor and the stator with the circular interface Γ in between



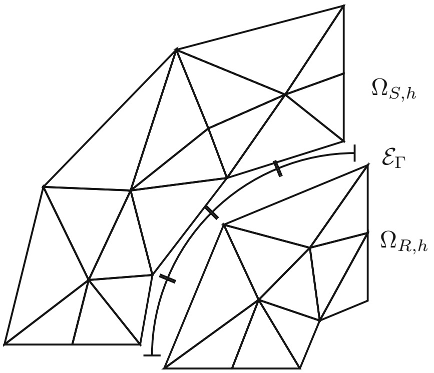


Fig. 3 Example of a polygonal approximation $\mathcal{T}_h$ for the individual parts. The meshes are separated for better presentation, in reality all interface vertices of rotor and stator lie on Γ

For the discretization we use continuous piece-wise linear elements, in details: Find $\mathbf{u}_h = (u_{S,h}, u_{R,h}, \lambda_h) \in X_h(\mathcal{M}_h) = V_h\left(\mathcal{T}_{S,h}\right) \times V_h\left(\mathcal{T}_{R,h}\right) \times W_h\left(\mathcal{E}_\Gamma\right)$, such that

$$a_h\left(\mathbf{u}_h, \mathbf{w}_h\right) = \langle f, \mathbf{w}_h\rangle \qquad \forall \mathbf{w}_h = (w_{S,h}, w_{R,h}, \psi_h) \in X_h(\mathcal{M}_h) \tag{2}$$

with

$$V_h\left(\mathcal{T}_{k,h}\right) = \left\{w_h \in H^1_{0,\partial\Omega\cap\partial\Omega_{k,h}}(\Omega_{k,h}) : w_h\big|_T \in \mathcal{P}^1(T) \text{ for } T \in \mathcal{T}_{k,h}\right\},$$

where $\mathcal{P}^1(T)$ is the space of polynomials of degree less than 1 restricted to the element T and

$$W_h\left(\mathcal{E}_\Gamma\right) = \left\{w_h \in L^2(\Gamma) \cap C^0(\Gamma) : w_h\big|_e \circ F_e \in \mathcal{P}^1(\widehat{e}) \text{ for } e \in \mathcal{E}_\Gamma\right\},$$

where $\widehat{e} = [0, 1]$ denotes the reference element and F_e maps $\widehat{e}$ onto e. The semilinear form and the linear form are given by

$$\begin{aligned} a_h\left(\mathbf{u}_h, \mathbf{w}_h\right) = \sum_{k\in\{R,S\}} \Bigg(& \sum_{T\in\mathcal{T}_{k,h}} \int_T \nu_T(\left|\nabla u_{k,h}\right|)\nabla u_{k,h} \cdot \nabla w_{k,h} \\ & - \sum_{e\in\mathcal{E}_\Gamma} \int_e \nu_0 \nabla \overline{u}_{k,h} \cdot n_k \left(\overline{w}_{k,h} - \psi_h\right) + \beta \sum_{e\in\mathcal{E}_\Gamma} \int_e \nu_0 \left(\overline{u}_{k,h} - \lambda_h\right) \nabla \overline{w}_{k,h} \cdot n_k \\ & + \frac{\alpha}{h} \sum_{e\in\mathcal{E}_\Gamma} \int_e \nu_0 \left(\overline{u}_{k,h} - \lambda_h\right)\left(\overline{w}_{k,h} - \psi_h\right)\Bigg) \end{aligned} \tag{3}$$

and

$$\langle f_h, \mathbf{w}_h \rangle = \sum_{k\in\{R,S\}} \sum_{T\in\mathfrak{T}_{k,h}} \left(\int_T J_3 w_{k,h} + \int_T \nu_M \mathbf{M}_\perp \nabla w_{k,h} \right).$$

Here n_k stands for the normal vector on $\partial\Omega_k$ that points in outward direction. $\overline{u}_{k,h}$ and $\overline{w}_{k,h}$ denotes the extension of the function $u_{k,h}$ and $w_{k,h}$ respectively as defined by

Definition 1 (Extensions of u_h) The extension of $u_{k,h}$ from $\Omega_{k,h}$ to $\Omega_k \cup \Omega_{k,h}$ is given in the following way:

$$\overline{u}_{S,h} = \begin{cases} u_{S,h} & \text{on } \Omega_{S,h} \\ 0 & \text{on } \Omega_S \setminus \Omega_{S,h} \end{cases} \qquad \overline{u}_{R,h} = \begin{cases} u_{R,h} & \text{on } \Omega_{R,h} \\ u_{R,h}^{ext} & \text{on } \Omega_R \setminus \Omega_{R,h}, \end{cases}$$

where u^{ext} denotes the linear extension of the function on the triangle to the curved domain.

The second term on the right-hand side in (3) is called *consistency term*, the third term is the *symmetry term*, since in the linear case we get a symmetric bilinear form by choosing $\beta = -1$. The last term in (3) is referred to as *penalization term*.

We use $\beta = -1$ and we choose the parameter $\alpha = 20$. This choice results in a symmetric, positive definite stiffness matrix. In the notation of discontinuous Galerkin methods this corresponds to a *symmetric interior penalty Galerkin* (*SIPG*) method.

3 A Posteriori Error Estimator

The error is estimated in the following energy norm (see [2, p.6])

$$\|(w,\psi)\|_{E(\Omega,\Gamma)} = \left(\sum_{k\in\{R,S\}} \sum_{i\in I_k} \nu_{ref,i}^2 \, \|\nabla w_i\|_{L^2(\Omega_i)}^2 + \nu_0^2 \, |w-\psi|_{\frac{1}{2},\Gamma}^2 \right)^{\frac{1}{2}}, \tag{4}$$

where $\nu_{ref,i}$ is a reference value for each material domain, and

$$|w|_{\frac{1}{2},\Gamma}^2 = \sum_{k\in\{S,R\}} \frac{1}{h} \, |w_k|_{L^2(\partial\Omega_k\cap\Gamma)}^2 .$$

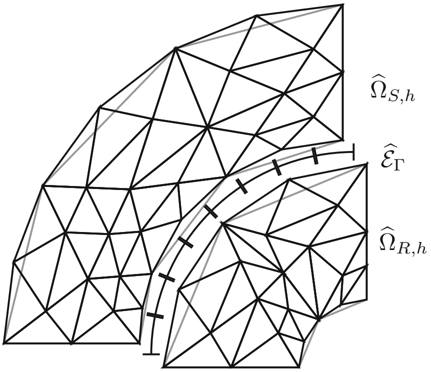


Fig. 4 $\widehat{\mathcal{M}}_h$ with one boundary refinement. The two meshes that are touching the interface are depicted separately

Additionally we introduce the jump notation for vector-valued functions q:

$$[\![q]\!]_e = \begin{cases} q_i \cdot n_i + q_j \cdot n_j & \text{for } e \in \overline{\Omega}_i \cap \overline{\Omega}_j, \\ q_i \cdot n_i & \text{for } e \in \partial\Omega \cap \partial\Omega_i. \end{cases}$$

For the error estimator we introduce another mesh $\widehat{\mathcal{M}}_h$ which is finer near the curvilinear boundaries than $\mathcal{M}_h$ (see Fig. 4). In particular we start in each part Ω_k from the mesh $\mathcal{T}_{k,h}$ and refine all elements with at least two nodes on a curvilinear boundary uniformly. Additionally the neighbouring elements are refined in such a way that we end up with admissible meshes. The interface elements are refined uniformly. This refinement strategy is repeated until the boundary is approximated sufficiently well. By $\widehat{\mathbf{u}}_h$ we will denote an approximate solution of the discrete problem on this finer mesh (calculated by using a few steps of a symmetric Gauss-Seidel-Iteration), extended to $\Omega \cup \widehat{\Omega}_h$ as described in Definition 1.

Theorem 1 (A Posteriori Error Estimation) *Let $u \in H_0^1(\Omega)$ be the solution of problem (1), and let $\mathbf{u}_h \in X_h(\mathcal{M}_h)$ denote the solution of (2) and $\widehat{\mathbf{u}}_h \in X_h(\widehat{\mathcal{M}}_h)$ denotes the solution on the finer mesh. Then there exists a constant C such that for $\overline{\mathbf{e}}_h = \mathbf{u} - \overline{\mathbf{u}}_h$ the following a posteriori error bound holds:*

$$\|\overline{\mathbf{e}}_h\|_{E(\Omega,\Gamma)} \leq \left(\eta^2 + \sum_{k\in\{R,S\}} \mathcal{O}_\eta(f_k, \widehat{u}_h) + \mathcal{O}(h^3)\right)^{\frac{1}{2}} \tag{5}$$

with

$$\eta^2 = \sum_{k\in\{R,S\}} \left(\sum_{T\in\mathcal{T}_{k,h}} \eta_T^2 + \sum_{e\in\overline{\mathcal{E}}_\Gamma} \eta_{e,k}^2 \right),$$

where

$$\eta_T^2 = Ch^2 \left\| f \right\|_{L^2(T)}^2 + \sum_{\widehat{T} \in \mathcal{C}(T)} \left(\sum_{e \in \partial \widehat{T}} \frac{C}{2} h \left\| [\![\nu_{\widehat{T}}(|\nabla \widehat{u}_h|) \nabla \widehat{u}_h]\!] \right\|_{L^2(e)}^2 \right.$$

$$\left. + \left\| \nu_{ref} \nabla \widehat{u}_h - \nu_{ref} \nabla \overline{u}_h \right\|_{L^2(\widehat{T})} + \frac{1}{2} h \sum_{e \in \partial \widehat{T}} \left\| [\![\nu_{\widehat{T}}(|\nabla \overline{u}_h|) \nabla \overline{u}_h - \nu_{\widehat{T}}(|\nabla \widehat{u}_h|) \nabla \widehat{u}_h]\!] \right\|_{L^2(e)}^2 \right)$$

$$\eta_{e,k}^2 = 2Ch^{-1} \left\| \left(\widehat{u}_{k,h} - \lambda_h \right) \right\|_{L^2(e)}^2 + \nu_0 h^{-1} \left\| \left(\widehat{u}_{k,h} - \widehat{\lambda}_h - \overline{u}_{k,h} + \overline{\lambda}_h \right) \right\|_{L^2(e)}^2 .$$

The set $\mathcal{C}(T)$ consists of all child-triangles of T in $\widehat{\mathfrak{T}}_{h,k}$. For the oscillation terms $\mathcal{O}_\eta(f_k, \widehat{u}_h)$ see [4, p.6].

Remark 2 In [1] an adaptive strategy including a posteriori controlled boundary approximation was presented for the Poisson problem with pure Dirichlet conditions. This technique cannot be adopted to the case of mixed boundary and interface conditions.

3.1 Marking and Refinement Strategy

The constant C in Theorem 1 plays the role of a weighting factor. For the numerical experiments C was replaced by a computable heuristic approximation based on information on the coarsest mesh leading to corresponding computable approximations which—for simplicity—are denoted by the same symbols η_T and $\eta_{e,k}$. With these quantities we assign local error indicators $\widetilde{\eta}_T$, $\widetilde{\eta}_e$ for each triangle T and each edge e from $\mathcal{M}_h$ in the following way:

1. for all triangles T with at most one vertex on Γ we set $\widetilde{\eta}_T = \eta_T$,
2. for every $e \in \mathcal{E}_\Gamma$ we set $\widetilde{\eta}_e = \eta_{e,S} + \eta_{e,R}$,
3. for the remaining triangles we set $\widetilde{\eta}_T$ equal to η_T enlarged by contributions $\widetilde{\eta}_e$ from edges $e \in \mathcal{E}_\Gamma$ weighted proportional to the length of the edge e between the two point of T on Γ.

As marking strategy we use Dörfler marking. For refinement we used classical red- and green-refinement.

4 Numerical Experiments

The error estimator was tested for the motor geometry shown in Fig. 1 for different rotor-to-stator positions. The initial mesh was generated with Netgen [5] and consists of 562 degrees of freedom. The mesh generator takes material interfaces into account leading to a non-uniform distribution of elements mainly concentrated around the air gap as depicted in Fig. 5.

For the first rotor-to-stator position the final mesh after 7 refinement steps is depicted in Fig. 6 and the reduction of the error in the energy-norm compared to the solution on a uniform refined mesh shown in Fig. 7.

Fig. 5 Initial Mesh with 562 degrees of freedom

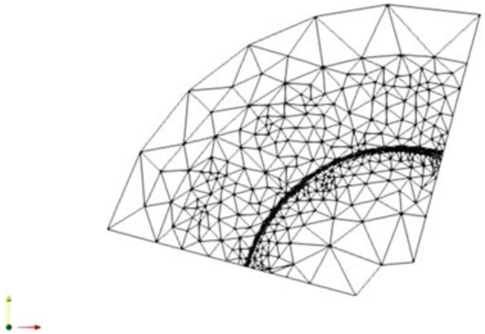


Fig. 6 Final Mesh for $\varphi = 0$, $\alpha = 20$

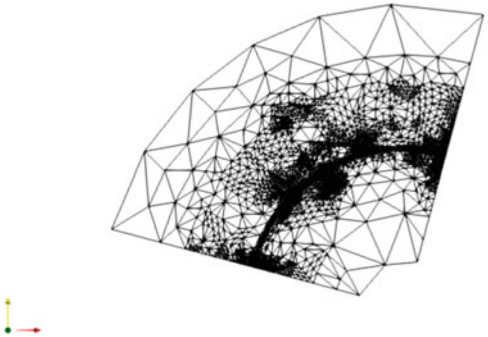
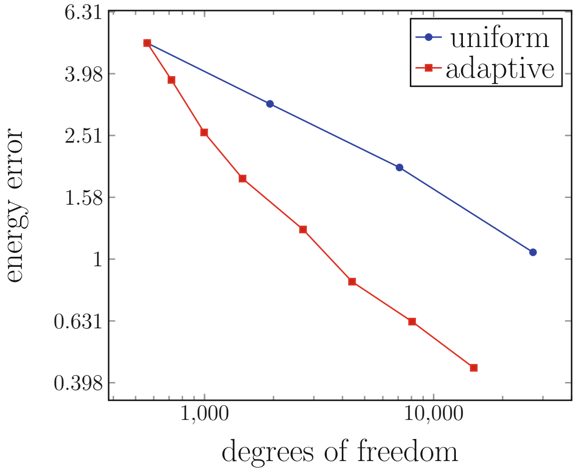


Fig. 7 Convergence for $\varphi = 0, \alpha = 20$



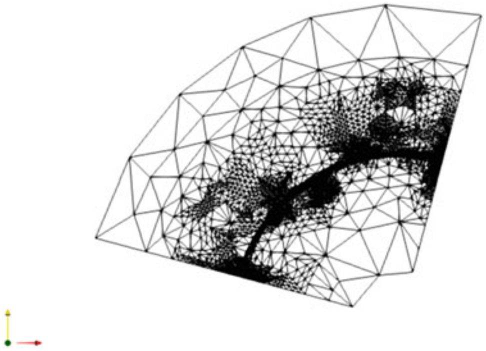


Fig. 8 Final mesh for $\varphi = 60$, $\alpha = 20$

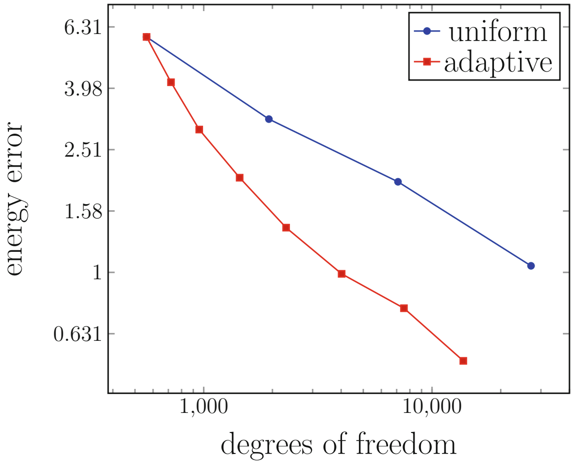


Fig. 9 Convergence for $\varphi = 60$, $\alpha = 20$

Starting from the same initial mesh, with the rotor part rotated by $\varphi = 60°$ we arrive after 7 refinement steps at a mesh depicted in Fig. 8 with an error reduction rate given in Fig. 9.

5 Conclusions

The adaptive refinement strategy show promising results. As expected due to the varying current in the coils for different φ as well as the natural change in the overall geometry the refined mesh looks quite different for different φ. Since it would be preferable to have one refined mesh for all rotation angles we are working on strategies to refine the mesh in an optimal way for the whole rotation cycle.

Acknowledgments This work has been supported by the "LCM - K2 Center for Symbiotic Mechatronics" within the framework of the Austrian COMET-K2 program.

References

1. Dörfler, W., Rumpf, M.: An adaptive strategy for elliptic problems including a posteriori controlled boundary approximation. Math. Comput. **67**, 1361–1382 (1998)
2. Egger, H.: A class of hybrid mortar finite element methods for interface problems with non-matching meshes. Preprint AICES-2009-2, Jan 2009
3. Hollaus, K., Feldengut, D., Schöberl, J., Wabro, M., Omeragic, D.: Nitsche-type mortaring for Maxwell's equations. In: Progress in Electromagnetics Research Symposium Proceedings, Cambridge, pp. 397–402, 5–8 July 2010
4. Houston, P., Süli, E., Wihler, T. P.: A posteriori error analysis of hp-version discontinuous Galerkin finite-element methods for second-order quasi-linear elliptic PDEs. IMA J. Numer. Anal. **28**, 245–273 (2008)
5. Netgen: Online document (2018). https://ngsolve.org/. Visited 30 Jan 2019

Isogeometric Simulation and Shape Optimization with Applications to Electrical Machines

Peter Gangl, Ulrich Langer, Angelos Mantzaflaris, and Rainer Schneckenleitner

Abstract Future e-mobility calls for efficient electrical machines. For different areas of operation, these machines have to satisfy certain desired properties that often depend on their design. Here we investigate the use of multipatch Isogeometric Analysis (IgA) for the simulation and shape optimization of the electrical machines. In order to get fast simulation and optimization results, we use non-overlapping domain decomposition (DD) methods to solve the large systems of algebraic equations arising from the IgA discretization of underlying partial differential equations. The DD is naturally related to the multipatch representation of the computational domain, and provides the framework for the parallelization of the DD solvers.

1 Introduction

Isogeometric Analysis (IgA) is a relatively new approach for discretizing partial differential equations (PDEs). IgA was introduced in [2]. It can be seen as an alternative to the more classical Finite Element Method (FEM). The idea in IgA

P. Gangl
Institute of Applied Mathematics, TU Graz, Graz, Austria
e-mail: gangl@math.tugraz.at

U. Langer
Institute of Computational Mathematics, JKU Linz, Linz, Austria

RICAM, Austrian Academy of Sciences, Linz, Austria
e-mail: ulanger@numa.uni-linz.ac.at

A. Mantzaflaris
Institute of Applied Geometry, JKU Linz, Linz, Austria

Université Côte d'Azur, Inria Sophia Antipolis - Méditerranée, Valbonne, France
e-mail: angelos.mantzaflaris@jku.at

R. Schneckenleitner (✉)
RICAM, Austrian Academy of Sciences, Linz, Austria
e-mail: rainer.schneckenleitner@ricam.oeaw.ac.at

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_4

is to use the same basis functions for both representing the geometry of the computational domain and solving the PDEs. This aspect makes IgA especially interesting for design optimization procedures. In practice, it is often the case that one performs design optimization and geometric modeling simultaneously. State-of-the-art computer aided design (CAD) software uses B-splines or Non-Uniform Rational B-splines (NURBS) for the modeling process whereas the design optimization requires an analysis suitable representation of the model. So far the design optimization is mainly done using FEM as discretization method. Hence, the B-spline or NURBS representation of the geometric model has to be converted into a suitable mesh for the Finite Element Analysis. This conversion is in general very computationally demanding. The new IgA paradigm circumvents these problems. Therefore, IgA is very beneficial for the simulation and optimization when the representation of the computational domain comes from CAD software; see [1, 10] for applications to electrical machines.

Since practical optimization problems tend to be very large, the numerical solution of the underlying PDEs becomes computationally very expensive. Moreover, in PDE-constrained shape optimization processes, there are more than one PDE to solve. In particular, line search requires to solve the magnetostatic PDE constraint several times. In order to get fast optimization results, we use Dual-Primal IsogEometric Tearing and Interconnecting (IETI-DP) methods for the solution of the linear algebraic systems arising from the IgA discretization. The IETI-DP solvers are non-overlapping domain decomposition methods; see [5, 6]. IETI-DP methods are closely related to the FEM-based FETI-DP methods; see, e.g., [9] and the references therein. We show that IETI-DP methods are superior to sparse direct solvers with respect to computational time and memory requirement. Moreover, IETI-DP provides a natural framework for parallelization. Indeed, our numerical experiments on a distributed memory computer show an excellent scaling behavior of this method.

The remainder of the paper is organized as follows. In Sect. 2, we describe our model problem and the shape optimization method that is based on the shape derivative. Section 3 is devoted to the IETI-DP solver and its performance on parallel computers. Finally, in Sect. 4, we use IETI-DP within the interior point optimizer Ipopt [11] yielding an efficient shape optimization procedure.

2 Shape Optimization via Gradient Descent

2.1 Problem Description

We investigate the simulation and shape optimization of an interior permanent magnet (IPM) electric motor by means of IgA. The IgA approach seems to be very attractive for such practical problems. The most beneficial aspect of IgA in the context of optimization is the fact that the same basis functions which are used

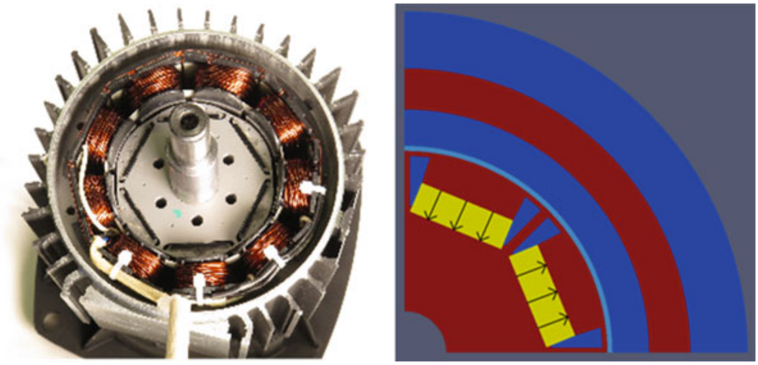


Fig. 1 Real world IPM electric motor (see Acknowledgement section) on the left, and a quarter of the cross section of a similar electric motor with 8 magnetic poles that is used in our numerical tests on the right

to represent the geometry of the IPM electric motor are also exploited to solve the underlying PDEs. In the optimization procedure, we want to optimize the shape of the motor in order to maximize the runout performance, i.e. to maximize the smoothness of the rotation of the motor. An example of an IPM electric motor is given in Fig. 1 (left). One possible way to optimize the runout performance of an IPM electric motor is to minimize the squared L^2-distance between the radial component of the magnetic flux B in the air gap and a desired smooth reference function B_{d}. The resulting optimization problem is subject to the 2d magnetostatic PDE as constraint.

Mathematically, the arising optimization problem can be expressed as follows:

$$\min_{D} J(u) := \int_{\Gamma} |B(u) \cdot \mathrm{n}_{\Gamma} - B_{\mathrm{d}}|^2 \mathrm{d}s = \int_{\Gamma} |\nabla u \cdot \tau_{\Gamma} - B_{\mathrm{d}}|^2 \mathrm{d}s \tag{1}$$

$$\text{s.t. } u \in H_0^1(\Omega) : \int_{\Omega} \nu_{\mathrm{D}}(x) \nabla u \cdot \nabla \eta \,\mathrm{d}x = \langle F, \eta \rangle \quad \forall \eta \in H_0^1(\Omega), \tag{2}$$

where J denotes the objective function, Γ is the midline of the air gap, Ω denotes the whole computational domain, and D is the domain of interest also called design domain. The variational problem (2) is nothing but the 2d linear magnetostatic problem with the piecewise constant magnetic reluctivity $\nu_D(x) = \chi_{\Omega_f(D)}(x)\nu_1 + \chi_{\Omega_{\mathrm{mag}}}(x)\nu_{\mathrm{mag}} + \chi_{\Omega_{\mathrm{air}}(D)}(x)\nu_0$. Here, Ω_f, Ω_{mag} and Ω_{air} denote the ferromagnetic, permanent magnet and air subdomains, respectively, and ν_1, ν_{mag} and ν_0 denote the corresponding reluctivity values. Note that the shape D enters the optimization problem via the function ν_D and influences the objective function via the solution u. The right hand side $F \in H^{-1}(\Omega)$ in (2) is defined by the linear functional

$$\langle F, \eta \rangle := \int_{\Omega} (J_3 \eta + \nu_{\mathrm{mag}} M^{\perp} \cdot \nabla \eta) \,\mathrm{d}x \tag{3}$$

for all $\eta \in H_0^1(\Omega)$. Here, $M^\perp$ denotes the perpendicular of the magnetization M, which is indicated in Fig. 1 and vanishes outside the permanent magnets, and J_3 is the third component of the impressed current density in the coils. Note that the solution u is the third component of the magnetic vector potential, i.e. $B(u) = \text{curl}((0, 0, u)^T)$. Moreover, $n_\Gamma = (n_1, n_2, 0)^T$ and $\tau_\Gamma = (\tau_1, \tau_2)^T$ denote the outward unit normal and unit tangential vectors along the air gap, respectively.

We are interested in the radial component of the magnetic flux density along the air gap due to the permanent magnetization. For that reason, we set $J_3 = 0$ and consider the coil regions as air. Figure 1 (right) shows a quarter of a cross section of a simplified IPM electric motor that is provided by CAD software. Hence, this geometry representation is suitable for IgA simulation. The red-brown areas represent ferromagnetic material (Ω_f), the blue areas consist of air (Ω_{air}), the yellow areas are the permanent magnets (Ω_{mag}). The air gap of the motor is highlighted in light blue. In this initial model for the optimization, the design domain D is the ferromagnetic area right above the permanent magnets. In order to get a smoother rotation we are looking for a better shape of this part D.

2.2 The Shape Derivative

For the optimization of the IPM electric motor, we use gradient based optimization techniques. Hence, we need the derivative of the objective J with respect to a change of the current shape. The shape derivative in tensor form [4, 7, 10] of our optimization problem is given by

$$dJ(D)(\phi) = \int_\Omega \mathcal{S}(D, u, p) : \partial\phi \mathrm{d}x, \quad \forall \phi \in H_0^1(\Omega, \mathbb{R}^2) \tag{4}$$

with $\mathcal{S}(D, u, p) = (\nu_{\text{D}}(x)\nabla u \cdot \nabla p - \nu_{\text{mag}} \nabla p \cdot M^\perp)\mathcal{I} + \nu_{\text{mag}} \nabla p \otimes M^\perp - \nu_{\text{D}}(x)\nabla p \otimes \nabla u - \nu_{\text{D}}(x)\nabla u \otimes \nabla p$, where $\mathcal{I}$ denotes the identity, the state u solves the constraint (2), and p solves the adjoint problem

$$\int_\Omega \nu_{\text{D}}(x)\nabla p \cdot \nabla\eta \,\mathrm{d}x = -2\int_\Gamma (B(u) \cdot n_\Gamma - B_{\text{d}})(B(\eta) \cdot n_\Gamma)\,\mathrm{d}s \quad \forall \eta \in H_0^1(\Omega). \tag{5}$$

In (4), $\mathcal{S}(D, u, p) : \partial\phi$ means Frobenius' scalar product of the 2×2 matrices $\mathcal{S}(D, u, p)$ and $\partial\phi = (\frac{\partial\phi_i}{\partial x_j})_{i,j=1}^2$, defined by $A : B := \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ij}$ for general $n \times n$ matrices $A = (a_{ij})_{i,j=1}^n$ and $B = (b_{ij})_{i,j=1}^n$, whereas $a \otimes b := (a_i b_j)_{i,j=1}^n$ for vectors $a = (a_i)_{i=1}^n$ and $b = (b_j)_{j=1}^n$ from $\mathbb{R}^n$.

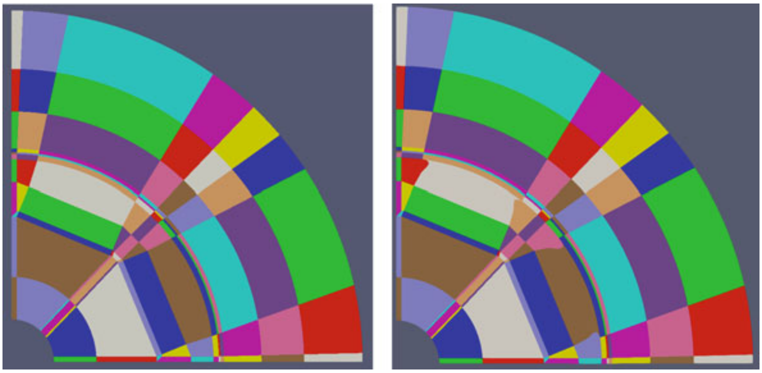


Fig. 2 Initial and final design of an IPM motor

2.3 Numerical Shape Optimization

We used a continuous Galerkin (cG) IgA discretization for both the simulation and optimization problems. The implementation is done in **G+Smo**.[1] Figure 2 (left) shows a possible computational domain suitable for cG. The shown multipatch domain consists of 93 patches. For each of these patches, we used a B-spline mapping from a reference patch with splines of degree 3. For the optimization, we need the shape gradient $\nabla J \in V := H_0^1(\Omega, \mathbb{R}^2)$ which can be computed by solving the auxiliary problem: find $\nabla J \in V$ such that

$$b(\nabla J, \psi) = -\mathrm{d}J(D)(\psi) \quad \forall \psi \in V. \tag{6}$$

The expression on the right hand side of (6) is the negative shape derivative whereas the expression $b(\cdot, \cdot)$ on the left hand side is some V-elliptic, V-bounded bilinear form which must be chosen appropriately. For our studies, we used

$$b(\phi, \psi) = \int_\Omega \phi \cdot \psi \,\mathrm{d}x + \int_\Omega \alpha(\partial\phi : \partial\psi)\,\mathrm{d}x \tag{7}$$

with a patchwise constant function $\alpha \in L^\infty(\Omega)$.

In the right picture of Fig. 2, we can see the optimized shape with respect to the runout performance compared to the initial domain on the left. We were able to reduce the objective from $4.236 \cdot 10^{-4}$ down to $2.781 \cdot 10^{-4}$.

[1] Mantzaflaris, A. et al.: G+Smo (geometry plus simulation modules) v0.8.1., http://gs.jku.at/gismo, 2017 Jun 19 2018.

3 Fast Numerical Solutions by IETI-DP

Up to now, we have solved the arising PDEs by means of a sparse direct solver. One drawback of a direct solution method is that it is rather slow for large-scale systems. In particular, in shape optimization, we have to solve the state equation (2), the adjoint equation (5), and the auxiliary problem (6) for the shape gradient, which decouples into two scalar problems, in every iteration of the optimization algorithm. Moreover, during a line search procedure, it might be the case that the state equation has to be solved several times. To overcome the issue of a slow performance, we were looking for a fast and suitable solver for our simulation and optimization processes. We chose the IETI-DP technique for solving the PDEs [5]. IETI-DP is a non-overlapping domain decomposition technique which introduces local subspaces which are then again coupled using additional constraints. A comparison between the sparse direct solver SuperLU [3] and IETI-DP for solving the state equation (2) on a full cross section of an IPM electric motor clearly shows that the recently developed IETI-DP method [5] performs much better as can be seen in Table 1. The numerical results displayed in Tables 1 and 2 were obtained on RADON1 (https://www.ricam.oeaw.ac.at/hpc/overview/) a high performance computing cluster with 1168 computing cores and 10.7 TB of memory. Table 1 also shows that, with an increasing number of degrees of freedom, the proposed IETI-DP technique solves the problem much faster than the sparse direct solver. Moreover, it can be seen that, with too many degrees of freedom, the sparse direct solver ran out of memory whereas IETI-DP could provide the solution to the problem. The solution to the state equation is shown in Fig. 3 (right).

Moreover, IETI-DP provides a natural framework for parallelization. Because of the multipatch structure of the computational domains in IgA, each patch can be seen as a subdomain in the IETI-DP approach. Then one can create suitable subdomains consisting of a certain number of patches for each processor, e.g., one possible choice is to group the patches to subdomains according to their number of degrees of freedom which means that the degrees of freedom are almost evenly distributed over the number of processors. Table 2 shows the strong scaling behavior of the IETI-DP solver. In this experiment, we solved the constraint equation (2) on

Table 1 SuperLU vs. IETI-DP on a single core

# dofs	SuperLU	IETI-DP	speedup
72,572	36.0 s	17.0 s	2.12
250,844	193.0 s	69.8 s	2.77
928,796	1943.0 s	463.0 s	4.20
3,570,332	–	1179.0 s	–

Table 2 Strong scaling with IETI-DP and 3,570,332 dofs

# cores	1	2	4	8	16	32	64	128
Time [s]	1179	577	325	164	89	43	22	14
Rate	–	2.04	1.78	1.98	1.84	2.07	1.95	1.57

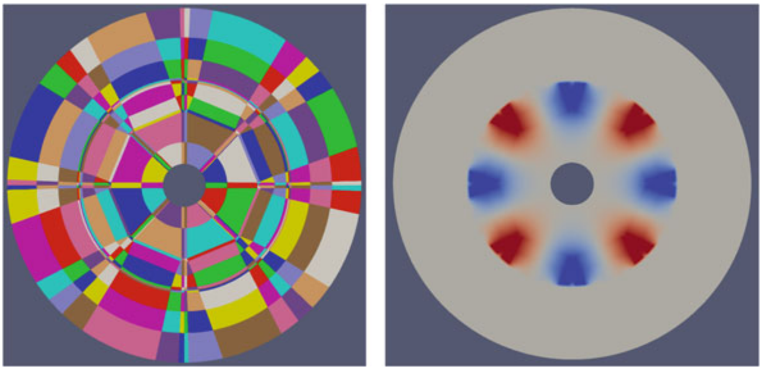


Fig. 3 Whole initial cross section as well as the solution

the full cross section of an IPM electric motor with 3,570,332 degrees of freedom. From Table 2, we can see the expected performance, i.e., if we double the number of processors the computation time reduces nearly by a factor of two.

4 Shape Optimization Based on Ipopt and IETI-DP

In this section, we point out the usage of the interior point optimizer Ipopt [11], for the shape optimization using IETI-DP as underlying PDE solver. If we perform shape optimization without any additional considerations, then we might run into troubles. More precisely, it can happen that we get self-intersections in the final shape even if the objective decreases.

To prevent such self-intersections, we consider the Jacobian determinant of the geometry transformation in the design domain and its neighboring air regions. The Jacobian determinant of these patches must have the same sign in each iteration. If the sign changes from one iteration to the next, then we reduce the step size until the Jacobian determinant of the new design has the same sign as in the initial configuration. In this way, we are able to ensure that the shape is technically feasible.

In the first naive approach, all control coefficients of the multipatch domain are considered as design variables, and the vector field computed by (6) is applied globally. The computational effort for the optimization can be reduced by applying the computed vector field only on the important interfaces between the design domain and the neighboring air regions. This reduces the number of design variables from approximately 28,000 to 128 in the coarsest setting. The inner control coefficients of the design area and the bordering air regions are rearranged via a spring patch model [8].

In a first test setting, Ipopt stops at an optimal solution after 95 iterations using a BFGS method. We set the NLP error tolerance to 10^{-6}, the relative error in the objective change to the same value, and we decided to exit the optimization loop

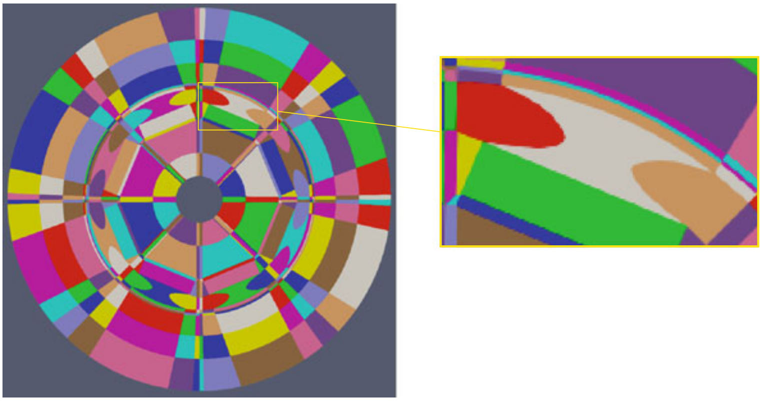


Fig. 4 Optimal shape after 130 iterations with relaxed bounds (left), zoom into one of the design regions (right)

after three iterations within these error bounds. The objective value dropped from $4.266 \cdot 10^{-4}$ down to $2.587 \cdot 10^{-4}$

Furthermore, we tried an additional experiment were we relaxed the bounds on the constraints a bit. In particular, we set the `bound_relax_factor` in Ipopt to 1. The result of this experiment can be seen in Fig. 4. We may observe from Fig. 4 that we get a very smooth final shape with even a smaller objective value of $2.436 \cdot 10^{-4}$. We point out that, if we adjust the different optimization parameters, we may get different optimal shapes and different objective values in the end.

Acknowledgments This work was supported by the Austrian Science Fund (FWF) via the grants NFN S117-03 and the DK W1214-04. We also acknowledge the permission to use the Photo in Fig. 1 (left) taken by the Linz Center of Mechatronics (LCM). The motor was produced by Hanning Elektro-Werke GmbH & Co KG.

References

1. Bontinck, Z., Corno, J., Schöps, S., De Gersem, H.: Isogeometric analysis and harmonic stator–rotor coupling for simulating electric machines. Comput. Methods Appl. Mech. Eng. **334**, 40–55 (2018)
2. Cottrell, J.A., Hughes, T.J.R., Bazilevs, Y.: Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput. Methods Appl. Mech. Eng. **194**, 4135–4195 (2005)
3. Demmel, J.W., Eisenstat, S. C., Gilbert, J.R., Li, X.S., Liu, J. W. H.: A supernodal approach to sparse partial pivoting. SIAM J. Matrix Anal. Appl. **20**(3), 720–755 (1999)
4. Gangl, P.: Sensitivity-based topology and shape optimization with application to electrical machines. Ph.D. thesis, Johannes Kepler University Linz (2016)
5. Hofer, C., Langer, U.: Dual-primal isogeometric tearing and interconnecting solvers for multipatch dG-IgA equations. Comput. Methods Appl. Mech. Eng. **316**, 2–21 (2017)

6. Kleiss, S., Pechstein, C., Jüttler B., Tomar S.L: IETI–isogeometric tearing and interconnecting. Comput. Methods Appl. Mech. Eng. **247**, 201–215 (2012)
7. Laurain, A., Sturm, K.: Distributed shape derivative via averaged adjoint method and applications. Esaim Math. Model. Numer. Anal. **50**(4), 1241–1267 (2016)
8. Nguyen, D.M., Gravesen, J., Evgrafov, A.: Isogeometric analysis and shape optimization in electromagnetism. Ph.D. thesis, Technical University of Denmark (2012)
9. Pechstein, C.: Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems. Springer, Berlin (2013)
10. Schneckenleitner, R.: Isogeometrical analysis based shape optimization. Master thesis, Johannes Kepler University Linz (2017)
11. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line search algorithm for large scale nonlinear programming. Math. Program. **106**(1), 25–57 (2006)

Techniques for Modeling Fiber Laser Amplifiers

Jay Gopalakrishnan, Tathagata Goswami, and Jacob Grosek

Abstract Numerical techniques for simulation of electromagnetic wave propagation within fiber amplifiers are discussed. Since a full-featured simulation using the Maxwell system on a realistic fiber is beyond reach, simplified models using Coupled Mode Theory (CMT) form the state of the art. This work presents a novel concept of an *equivalent short fiber,* namely an artificial fiber which imitates a longer fiber in essential characteristics. A CMT simulation on an equivalent short fiber requires only a fraction of the computational resources needed to simulate the full length fiber.

1 Fiber Amplifiers

The ability of solid-state fiber laser amplifiers to deliver high output power has been exploited and studied over the last few decades [4]. Currently, the main roadblock to power scaling these amplifiers is the transverse mode instability (TMI), a sudden breakdown in beam quality at high power operation, first observed experimentally [2]. These observations have led to intensive speculations on the cause of TMI, the prevailing theory being that the cause is a temperature-induced grating. Reliable numerical simulation of TMI and other nonlinear optical effects within fibers can provide important insights for validating or rejecting various physical hypotheses put forth to explain these effects. The simulation techniques must however be able to numerically solve the field propagation within a long fiber a vast number of times.

J. Gopalakrishnan (✉) · T. Goswami
Portland State University, Portland, OR, USA
e-mail: gjay@pdx.edu; tgoswami@pdx.edu

J. Grosek
Directed Energy Directorate, Air Force Research Laboratory, Kirtland Air Force Base, Albuquerque, NM, USA
e-mail: jacob.grosek.1@us.af.mil

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_5

While numerical modeling of fiber amplifiers has been effectively used by many [5, 6, 9], accurate simulation of full length fibers remains cumbersome due to its high demands on computational resources and long simulation times. A full Maxwell simulation of Raman gain in a fiber was attempted in [7]: more than five million degrees of freedom was needed to simulate a fiber 80 wavelengths long (less than 0.0001 m). Clearly, a full Maxwell model of a realistically long ($\sim$10 m) fiber is beyond the reach of today's simulation capabilities. The need for simplified models is evident. Indeed the state of the art in fiber amplifier simulation consists of beam propagation methods and simplified CMT-based models (see Sect. 2). Yet, even these simplified models are computationally too demanding. This paper contributes to the search for a faster numerical technique by developing a new concept of equivalent short fiber (see Sect. 3) that can speed up these computations a 1000-fold.

The highest power beam combinable amplifiers are large mode area (LMA) circularly symmetric step-index fibers. They have a cylindrical core (usually doped—see Example 2 below) of radius r_{core} and a cladding region enveloping the core extending to radius r_{clad}. We set up our axes so that the longitudinal direction of the fiber is the z-axis. The transverse coordinates will be denoted x, y while using Cartesian coordinates. The refractive index n of the fiber is a piecewise constant function that equals n_{core} in the core and n_{clad} in the cladding. There is also a polymer coating surrounding the inner cladding; however, this will be ignored in our model, which focuses on the guided light in the core region.

At its inlet ($z = 0$ cross section), the fiber core region is seeded by a continuous wave input of highly coherent laser light, which is typically denoted as the "signal". We are interested in how the signal light is amplified by energy transfer from "pump" light while it propagates through the fiber, through the process called "gain". The pump light is also injected at the beginning of the fiber in a "co-pumped" configuration. The signal light is injected into the core, while the pump light goes into both the core and the cladding.

Let $\mathcal{E}_s, \mathcal{H}_s$ and $\mathcal{E}_p, \mathcal{H}_p$ denote the electric and magnetic fields of the signal and pump light, respectively. They are time harmonic of frequencies ω_s and ω_p, respectively, i.e., $\mathcal{E}_\ell(x, y, z, t) = \text{Re}\big[\mathbf{E}_\ell(x, y, z)e^{-\mathrm{i}\omega_\ell t}\big], \mathcal{H}_\ell(x, y, z, t) = \text{Re}\big[\mathbf{H}_\ell(x, y, z)e^{-\mathrm{i}\omega_\ell t}\big]$, for $\ell \in \{s, p\}$, so we may focus on their spatial dependence. We assume that the signal field $\mathbf{E}_s, \mathbf{H}_s$ and the pump field $\mathbf{E}_p, \mathbf{H}_p$ each satisfy the time-harmonic Maxwell system and that they are coupled only through a polarization term $\mathbf{P}_\ell \equiv \mathbf{P}_\ell(\mathbf{E}_s, \mathbf{E}_p)$:

$$\operatorname{curl} \mathbf{E}_\ell - \mathrm{i}\omega_\ell \mu_0 \mathbf{H}_\ell = 0, \qquad \operatorname{curl} \mathbf{H}_\ell + \mathrm{i}\omega_\ell \epsilon_0 \mathbf{E}_\ell = -\mathrm{i}\omega_\ell \mathbf{P}_\ell, \qquad \ell \in \{s, p\},$$

where ϵ_0 is the electric permittivity and μ_0 is the magnetic permeability (in vacuum).

Since the fiber is a dielectric medium, the standard linear background polarization term must be taken into account: $\mathbf{P}_\ell^{\text{bg}} = \epsilon_0(n^2 - 1)\mathbf{E}_\ell$. We set the total polarization $\mathbf{P}_\ell$ to $\mathbf{P}_\ell = \mathbf{P}_\ell^{\text{bg}} - \mathrm{i}\epsilon_0 \mathbf{E}_\ell n g_\ell c/\omega_\ell$ where the gain term g_ℓ depends nonlinearly on $\mathbf{E}_s$, $\mathbf{E}_p$ and $c = 1/\sqrt{\epsilon_0 \mu_0}$ is the speed of light. Examples of g_ℓ are given below.

Eliminating $\mathbf{H}_\ell$, we obtain the second order equation curl curl $\mathbf{E}_\ell - \omega_\ell^2 \epsilon_0 \mu_0 \mathbf{E}_\ell = \omega_\ell^2 \mu_0 \mathbf{P}_\ell$ for the electric field alone, which by virtue of the expression for $\mathbf{P}_\ell$ simplifies to curl curl $\mathbf{E}_\ell - k_\ell^2 n^2 \mathbf{E}_\ell + \mathrm{i} k_\ell n g_\ell \mathbf{E}_\ell = 0$ with $k_\ell = \omega_\ell / c$. A further assumption, frequently used in the theory of fiber optics, is that $\mathbf{E}_\ell$ is linearly polarized, i.e., using $\hat{e}_x$ to denote the unit vector in the x-direction, $\mathbf{E}_\ell(x, y, z) = U_\ell(x, y, z)\hat{e}_x$. Its also standard to neglect grad div$(U_\ell \hat{e}_x)$ since the dominant variations in $\mathbf{E}_\ell$ are in the z-direction. These assumptions yield the scalar Helmholtz equation for U_ℓ,

$$-\Delta U_\ell - k_\ell^2 n^2 U_\ell + \mathrm{i} k_\ell n g_\ell U_\ell = 0. \tag{1}$$

Examples (below) of g_ℓ we have in mind are expressed in terms of the irradiance $I_\ell = n|U_\ell|^2/\mu_0 c$. Light of high irradiance can perturb the refractive index causing many interesting nonlinear effects in optical fibers, such as in Example 1 below. However, of primary interest to us in the simulation of fiber amplifiers is *active gain*, occurring in fibers whose core is doped with lanthanide rare-earth metallic elements, such as Thulium (Tm) or Ytterbium (Yb)—see Example 2. They result in much larger gain due to the pump light driving dopant ions to excited radiative states followed by stimulated emission into signal photons.

Example 1 (Raman Gain) As described in [7, 10], the nonlinear Raman gain can be modeled using a measurable "bulk Raman gain coefficient" g_R by

$$g_\ell = \Upsilon_\ell g_R I_{\ell^c}, \qquad \ell \in \{s, p\} \tag{2}$$

where $\Upsilon_p = -\omega_p/\omega_s$, $\Upsilon_s = 1$, and $\ell^c \in \{s, p\} \setminus \{\ell\}$, the complementary index of ℓ.

Example 2 (Ytterbium-Doped Fiber) Yb-doped fiber amplifiers are usually pumped at λ_p =976 nm to move ions from a ground state (manifold ${}^2F_{7/2}$) to an excited state (manifold ${}^2F_{5/2}$) [3, 8]. After undergoing a rapid non-radiative transition to a lower energy state, the amplifier can lase around λ_s =1064 nm very efficiently. Denoting the constant, uniformly distributed ion population as $N_{\text{total}} = N_{\text{excited}} + N_{\text{ground}}$, the active gain can be modeled by

$$g_\ell = \sigma_\ell^{\text{ems}} N_{\text{excited}} - \sigma_\ell^{\text{abs}} N_{\text{ground}} = N_{\text{total}} \big[\sigma_\ell^{\text{ems}} \varepsilon - \sigma_\ell^{\text{abs}} (1 - \varepsilon)\big] \tag{3}$$

where the excited ion fraction $\varepsilon = N_{\text{excited}}/N_{\text{total}}$ is calculated in terms of a radiative lifetime (τ), and absorption and emission cross sections ($\sigma_\ell^{\text{abs}}, \sigma_\ell^{\text{ems}}$) as $\varepsilon = c^{\text{abs}}/(c^{\text{abs}} + c^{\text{ems}} + \tau^{-1})$, where $c^{\text{e/a}} = \sigma_p^{\text{e/a}} I_p/\hbar\omega_p + \sigma_s^{\text{e/a}} I_s/\hbar\omega_s$, for e/a $\in$ {ems, abs} and $\omega_\ell = 2\pi c/\lambda_\ell$. A commercial Yb-doped fiber, branded Nufern™ (nufern.com), offers realistic parameters for our numerical simulations: $n_{\text{core}} = 1.45097$, $n_{\text{clad}} = 1.44973$, $r_{\text{core}} = 12.5$ µm, $r_{\text{clad}} = 16 r_{\text{core}}$. (Other parameters: τ = 8e-4 s; σ_p^{abs}, σ_s^{abs}, σ_p^{ems}, σ_s^{ems} = 1.429e-24, 6e-27, 1.776e-24, 3.58e-25 m^2/ion; N_{total} = 6.25e25 ions/m^3.)

2 CMT Model

Coupled Mode Theory (CMT) uses the *transverse core modes* of the fiber to construct an electric field ansatz. These fiber modes $\varphi_l(x, y)$ are non-trivial functions that, together with their accompanying (positive) *propagation constants* β_l, solve the eigenproblem $(\Delta_{xy} + k_s^2 n^2)\varphi_l = \beta_l^2 \varphi_l$, where $\Delta_{xy} = \partial_{xx} + \partial_{yy}$ denote the transverse Laplacian. Since the modes we expect to see decay exponentially in the cladding region, the eigenproblem may be supplemented with zero Dirichlet boundary conditions. There can only be finitely many such modes, which we index using $l = 1, 2, \ldots, M$. For step-index fibers, these modes are called the linearly polarized (LP) transverse guided core modes [1]. The field ansatz is

$$U_s(x, y, z) = \sum_{m=1}^{M} A_m(z)\varphi_m(x, y)e^{\mathrm{i}\beta_m z}. \tag{4}$$

Furthermore, we assume that each A_m is so slowly varying in z that we may neglect the second derivative $\mathrm{d}^2 A_m/\mathrm{d}z^2$ for all $m = 1, \ldots, M$. Since we may precompute the modes φ_l, the ansatz (4) reduces the field computation to the numerical computation of $A_l(z)$. Substituting (4) into (1) and simplifying using the L^2 orthogonality of the modes, we find that A_l satisfies the system of ordinary differential equations (ODE)

$$\frac{\mathrm{d}A_l}{\mathrm{d}z} = \sum_{m=1}^{M} e^{\mathrm{i}(\beta_m - \beta_l)z}\, K_{lm}(A, I_p)\, A_m, \qquad 0 < z < L, \tag{5}$$

for $l = 1, \ldots, M$, where the mode coupling coefficient K_{lm} is given by

$$K_{lm}(A, I_p) = \frac{k_s}{2\beta_l} \int_{\Omega_z} g_s(I_s(x, y, A), I_p)\, n(x, y)\varphi_m(x, y)\overline{\varphi_l(x, y)}\, dx\, dy. \tag{6}$$

Here Ω_z represents the fiber cross section having the constant longitudinal coordinate value of z. Note that I_s depends on x, y and also on z through $A \equiv \{A_l\}$, i.e., $I_s \equiv I_s(x, y, A) = \frac{n}{\mu_0 c}|\sum_{m=1}^{M} A_m(z)e^{\mathrm{i}\beta_m z}\varphi_m(x, y)|^2$. Note that the “mode beating” term on the right hand side of (5), namely $e^{\mathrm{i}(\beta_m - \beta_l)z}$, oscillates at a wavelength not smaller than the *mode beat length* $2\pi / \max_{l,m} |\beta_m - \beta_l|$. An ODE solver must take enough steps per mode beat length to safeguard accuracy.

As in previous works [6, 9], we use a drastically simplified model of pump light: the effect of pump is modeled only through its irradiance $I_p(z)$ after assuming it to be independent of x and y, leading to the ODE

$$\frac{\mathrm{d}I_p}{\mathrm{d}z} = \langle g_p \rangle I_p \tag{7}$$

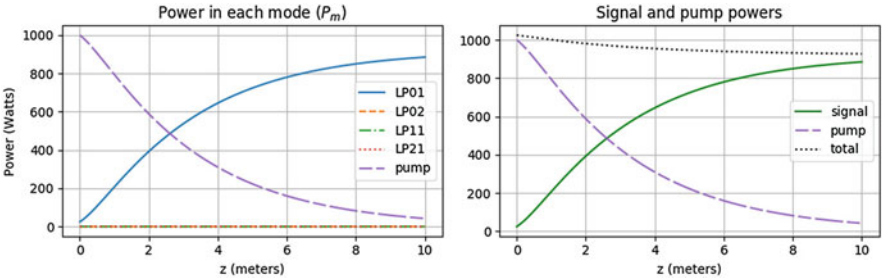


Fig. 1 Results from simulation of the full length ($L = 10$ m) Nufern fiber

where $\langle g_p \rangle(z)$ denotes the mean value of g_p over Ω_z.

Equations (5)–(7) were solved numerically for a 10 m long Nufern fiber of Example 2. This fiber has 4 modes, LP01, LP02, LP11 and LP21, enumerated as $\varphi_1, \ldots, \varphi_4$, respectively. We set initial values $A_m(0)$ such that 25 W of power is injected into the LP01 mode, while the remaining modes receive no power at inlet. Pump light is injected at 1000 W at $z = 0$. Lagrange finite elements of degree 5 were used to approximate φ_l and the mode overlap integral. All our simulations used 50 steps of the 4th order Runge-Kutta scheme per mode beat length. Over 400,000 ODE steps were needed to traverse 10 m. Results are shown in Fig. 1. Clearly, the signal power amplifies as z increases, while the power in pump light depletes.

3 A Scale Model: Equivalent Short Fiber

Physical or numerical scale models of an object preserve some of the important properties of the object while not preserving the original dimensions of the object. In the context of fiber amplifiers, a miniature scale model that reduces fiber length (while preserving the remaining dimensions) would be highly valuable in numerical computations. By reducing the number of steps within the ODE solver, an equivalent shorter fiber can bring about drastic reductions in computational cost.

At the outset, consider a quick dimensional analysis of (5). Its left hand side has dimension V/m (Volts per meter), so K_{lm} must have units of m^{-1}. Therefore, by writing out a non-dimensional formulation, we suspect that a shorter fiber of $\tilde{L} \ll L$ might, in some ways, behave similarly to the original fiber of length L, provided its coupling coefficient is magnified by $L/\tilde{L}$.

We need to understand better in what way the behaviour is similar and what properties need not be preserved. Let $\zeta(\tilde{z}) = \tilde{z}L/\tilde{L}$. A fiber of length L, under the variable change $\tilde{z} = \zeta^{-1}(z) = z\tilde{L}/L$ becomes one of length $\tilde{L}$. The original system (5)–(7) under the variable change, becomes $\mathrm{d}I_p(\tilde{z}L/\tilde{L})/\mathrm{d}\tilde{z} =$

$(L/\tilde{L})\langle g_p \rangle I_p(\tilde{z}L/\tilde{L})$ and

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{z}} A_l(\tilde{z}L/\tilde{L}) = \sum_{m=1}^{M} \frac{L}{\tilde{L}} K_{lm}(A(\tilde{z}L/\tilde{L}), I_p(\tilde{z}L/\tilde{L}))\, e^{\mathrm{i}(\beta_m-\beta_l)(L/\tilde{L})\tilde{z}}\, A_m(\tilde{z}L/\tilde{L}), \tag{8}$$

for all $0 < \tilde{z} < \tilde{L}$. Letting $\hat{A}_l = A_l \circ \zeta$ and $\hat{I}_p = I_p \circ \zeta$, we may rewrite these as

$$\frac{\mathrm{d}\hat{A}_l}{\mathrm{d}\tilde{z}} = \sum_{m=1}^{M} e^{\mathrm{i}(\beta_m-\beta_l)L\tilde{z}/\tilde{L}}\, \frac{L}{\tilde{L}}\, K_{lm}(\hat{A}, \hat{I}_p)\, \hat{A}_m, \quad \frac{\mathrm{d}\hat{I}_p}{\mathrm{d}\tilde{z}} = \frac{L}{\tilde{L}}\langle g_p \rangle \hat{I}_p, \quad 0 < \tilde{z} < \tilde{L}. \tag{9}$$

Thus, (9) on the shorter domain $0 < \tilde{z} < \tilde{L}$ is completely equivalent to (5)–(7). (Same initial data at $z = \tilde{z} = 0$ is assumed throughout.) Indeed, its solution $\hat{A}_l$, after changing variables is the same as the original solution A_l of (5). Unfortunately, (9) is not an improvement over (5) for numerical simulation. This is because the mode beat length is now reduced by $\tilde{L}/L$ in (9). Therefore an ODE solver, keeping the same number of steps per mode beat length, must now perform $L/\tilde{L}$ times the number of original steps, thus destroying the advantage of shortening the fiber to length $\tilde{L}$.

Hence we formulate another mode coupling system on the shorter fiber, with the same mode beat length as the original system (5)

$$\frac{\mathrm{d}\tilde{A}_l}{\mathrm{d}\tilde{z}} = \sum_{m=1}^{M} e^{\mathrm{i}(\beta_m-\beta_l)\tilde{z}} \frac{L}{\tilde{L}} K_{lm}(\tilde{A}, \tilde{I}_p)\tilde{A}_m, \quad \frac{\mathrm{d}\tilde{I}_p}{\mathrm{d}\tilde{z}} = \frac{L}{\tilde{L}}\langle g_p \rangle \tilde{I}_p, \quad 0 < \tilde{z} < \tilde{L}. \tag{10}$$

Since the phase factors in (9) and (10) are different, we cannot expect $\tilde{A}_l(\tilde{z})$ to be the same as the pullback $A_l \circ \zeta$ of the original solution A_l. Thus (10) is not completely equivalent to the original system (5): it does not preserve the solution. Yet the phase information lost in this new formulation is not of significant importance experimentally. Hence, we proceed to argue that (10) is a practically useful scale model of (5) by showing that it preserves some features of the solution under certain conditions.

Let $a_l(z) = A_l(z)e^{\mathrm{i}\beta_l z}$. Elementary calculations show that (5) implies

$$\frac{\mathrm{d}|a_l|^2}{\mathrm{d}z} = 2\sum_{m=1}^{M} \mathrm{Re}\big[K_{lm}(A, I_p)\, \overline{a}_l a_m\big]. \tag{11}$$

Let P be the vector function whose l^{th} component, $P_l(z)$, is the power contained in the l^{th} mode, namely $P_l(z) = \int_{\Omega_z} \frac{n}{\mu_0 c}|A_l(z)\varphi_l(x, y)|^2\, dx\, dy = |a_l|^2 \Phi_l$, where

$\Phi_l = \int_{\Omega_z} \frac{n}{\mu_0 c} |\varphi_l|^2 \, dx\, dy$. Equation (11) can be expressed using P_l as

$$\frac{1}{2}\frac{\mathrm{d}P_l}{\mathrm{d}z} = K_{ll}(A, I_p)P_l + \Phi_l \sum_{m \neq l} \mathrm{Re}\big[K_{lm}(A, I_p)\,\overline{a}_l a_m\big]. \tag{12}$$

Recall that $K_{lm}(A, I_p)$ is defined using $g_s(I_s(x, y, A), I_p)$—see (6). In some circumstances (see below), $I_s(x, y, A) = \frac{n}{\mu_0 c} |\sum_{m=1}^{M} a_m \varphi_m|^2$ can be approximated by

$$\mathfrak{I}_s(P) = \sum_{m=1}^{M} \frac{n}{\mu_0 c} |a_m \varphi_m|^2 = \sum_{m=1}^{M} \frac{n}{\mu_0 c \Phi_m} P_m \, |\varphi_m|^2 \, .$$

Let $\gamma_\ell(P, I_p) = g_\ell(\mathfrak{I}_s(P), I_p)$ for $\ell \in \{s, p\}$ and let κ_{lm} be defined exactly as K_{lm} but with g_s replaced by γ_s. Then (12) may be rewritten as

$$\frac{1}{2}\frac{\mathrm{d}P_l}{\mathrm{d}z} = \kappa_{ll}(P)P_l + \eta_l, \qquad l = 0, 1, \ldots, M, \qquad \text{where} \tag{13}$$

$$\eta_l = \Bigg[K_{ll}(A, I_p) - \kappa_{ll}(P)\Bigg]P_l + \Phi_l \sum_{m \neq l} \mathrm{Re}\Bigg[K_{lm}(A, I_p)\,\overline{a}_l a_m\Bigg] \tag{14}$$

for $l = 1, \ldots, M$. For the index $l = 0$, we set $P_0(z) = \int_{\Omega_z} I_p(z)\, dxdy$, the pump power, thereby absorbing (7) into (13) after setting $\eta_0 = \frac{1}{2}\big[\langle g_p \rangle - \langle \gamma_p \rangle\big]P_0$. We are interested in the case of small η_l. Then (13) is a perturbation of an autonomous system.

Repeating the same procedure starting from (10) using $\tilde{a}_l(z) = \tilde{A}_l(z)e^{\mathrm{i}\beta_l z}$, we find that the corresponding powers $\tilde{P}_l = |\tilde{a}_l|^2 \Phi_l$ and $\tilde{P}_0 = \int_{\Omega_z} \tilde{I}_p \, dx\, dy$ satisfy

$$\frac{1}{2}\frac{\mathrm{d}\tilde{P}_l}{\mathrm{d}\tilde{z}} = \frac{L}{\tilde{L}}\kappa_{ll}(\tilde{P})\tilde{P}_l + \tilde{\eta}_l, \qquad l = 0, 1, \ldots, M, \qquad \text{where} \tag{15}$$

$$\tilde{\eta}_l = \left[\frac{L}{\tilde{L}}K_{ll}(\tilde{A}, \tilde{I}_p) - \frac{L}{\tilde{L}}\kappa_{ll}(\tilde{P})\right]\tilde{P}_l + \Phi_l \sum_{m \neq l} \mathrm{Re}\left[\frac{L}{\tilde{L}}K_{lm}(\tilde{A}, \tilde{I}_p)\,\overline{\tilde{a}}_l \tilde{a}_m\right].$$

To compare (15) with (13), we apply the change of variable ζ to (13) to find that $\frac{1}{2}\mathrm{d}(P_l \circ \zeta)/\mathrm{d}\tilde{z} = (L/\tilde{L})\kappa_{ll}(P_l \circ \zeta)P_l \circ \zeta + (L/\tilde{L})\eta_l \circ \zeta$. This means that when η and $\tilde{\eta}$ are small, $P_l \circ \zeta$ and $\tilde{P}_l$ solve approximately the same equation, so $P_l \circ \zeta \approx \tilde{P}_l$.

For this reason we shall call (10) an *equivalent short fiber model,* even if the electric fields generated are generally not the same. Note that, when considering real fiber amplifiers, power is the quantity of interest (measurable experimentally), not the electric field amplitude and phase. To summarize, *in the equivalent short fiber, the power P_l contained in the l^{th} mode is preserved from the original fiber model* (5) through a change of variable, under the above assumptions. Moreover, by estimating $\tilde{\eta}$ in equivalent fiber computations, we can gauge the reliability of the equivalent short fiber model a posteriori.

To simulate the equivalent short fiber model (10), we only need to multiply K_{lm} and $\langle g_p \rangle$ by $L/\tilde{L}$. In Example 1, this is accomplished by scaling the bulk Raman gain coefficient, i.e., replace the physical g_R by $\tilde{g}_R = g_R L/\tilde{L}$ in (2). Whereas for an active gain amplifier (like that of Example 2) this may be accomplished by scaling the total number of dopant ions, i.e., by replacing the physical N_{total} by $\tilde{N}_{\text{total}} = N_{\text{total}} L/\tilde{L}$.

One scenario where the assumption that η is small is justified is when most of the power is carried in one mode. Indeed, typical experimental setups of LMA fiber amplifiers do operate them as near-single mode fibers by filtering out the higher-order modes through differential bend losses induced by fiber coiling. When all except one a_i is small, cross terms involving $\bar{a}_l a_m$ are small for all $l \neq m$, so the last term in (14) is small. Moreover, the approximation of I_s by $\mathcal{I}_s$ where similar cross terms are neglected, is also accurate, so all terms defining (14) are small.

Figure 2 shows simulation results from the equivalent short fiber model of length $\tilde{L} = 0.01$ m mimicking the physical Nufern fiber of length $L = 10$ m we simulated at the end of Sect. 2. We see that the power distribution plots (bottom row) are identical to that of the physical fiber in Fig. 1. The cost of computation has however been reduced by a factor of $\tilde{L}/L = 1/1000$ (keeping the same number of ODE steps per mode beat length, see Sect. 2). When $\tilde{L}$ is changed to 0.005—see Fig. 3—we obtain similar power distribution plots again, although the solution components (A_m) have notably changed, in agreement with our analysis above. In further experiments (unreported here for brevity), we observed good performance

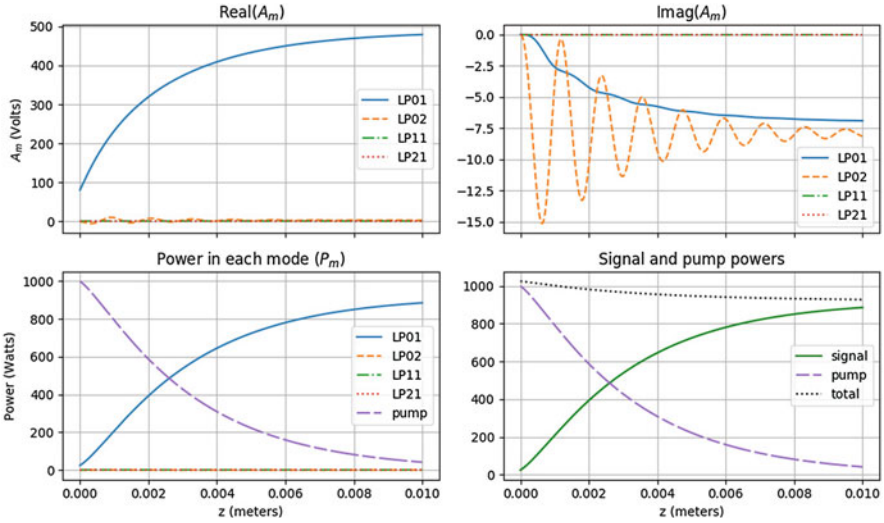


Fig. 2 Results from a short fiber of length $\tilde{L} = 0.01$ equivalent to an $L = 10$ m Nufern fiber

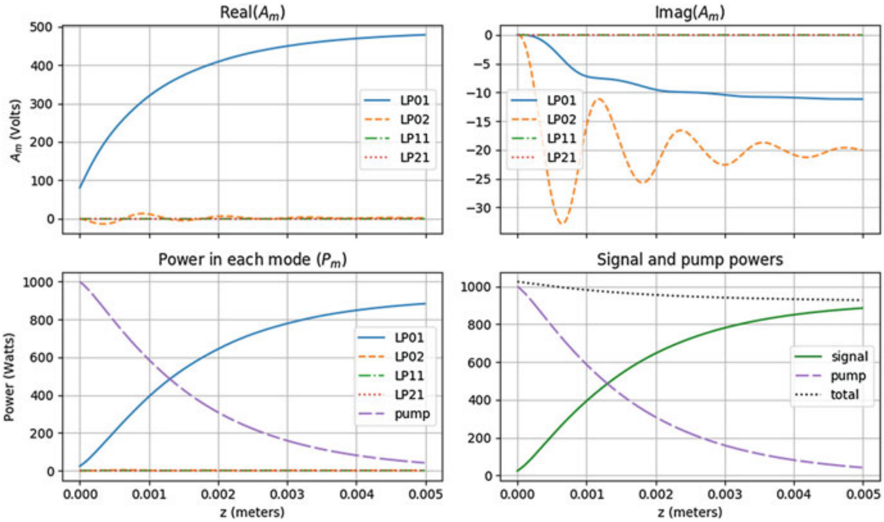


Fig. 3 Results from a short fiber of length $\tilde{L} = 0.005$ equivalent to an $L = 10$ m Nufern fiber

of equivalent fiber of length 0.1 m even when power was distributed equally among the modes. Note that all our simulations considered the worst-case scenario of no differential mode bend loss, i.e., any tendency of a bent fiber toward single-mode operation is left unmodeled.

Acknowledgments This work was supported in part by AFOSR grant FA9550-17-1-0090.

References

1. Agrawal, G.P.: Nonlinear Fiber Optics, 5th edn. Elsevier, New York (2013)
2. Eidam, T., Wirth, C., Jauregui, C., Stutzki, F., Jansen, F., Otto, H., Schmidt, O., Schreiber, T., Limpert, J., Tünnermann, A.: Experimental observations of the threshold-like onset of mode instabilities in high power fiber amplifiers. Opt. Express **19**(14), 13218–13224 (2011)
3. Grosek, J., Naderi, S., Oliker, B., Lane, R., Dajani, I., Madden, T.: Laser simulation at the Air Force Research Laboratory. In: Proc. SPIE, vol. 10254, p. 102540N-1 (2017)
4. Jauregui, C., Limpert, J., Tünnermann, A.: High-power fibre lasers. Nat. Photonics **7**(11), 861–867 (2013)
5. Jauregui, C., Otto, H.-J., Stutzki, F., Limpert, J., Tünnermann, A.: Simplified modelling the mode instability threshold of high power fiber amplifiers in the presence of photodarkening. Opt. Express **23**(16), 20203–20218 (2015)
6. Naderi, S., Dajani, I., Madden, T., Robin, C.: Investigations of modal instabilities in fiber amplifiers through detailed numerical simulations. Opt. Express **21**(13), 16111–16129 (2013)
7. Nagaraj, S., Grosek, J., Petrides, S., Demkowicz, L., Mora, J.: A 3D DPG Maxwell approach to nonlinear Raman gain in fiber laser amplifiers. Preprint, arXiv:1805.12240 (2018)

8. Pask, H., Carman, R., Hanna, D., Tropper, A., Mackechnie, C., Barber, P., Dawes, J.: Ytterbium-doped silica fiber lasers: versatile sources for the 1-1.2 μm region. IEEE J. Sel. Top. Quantum Electron. **1**(1), 2–13 (1995)
9. Smith, A.V., Smith, J.J.: Mode instability in high power fiber amplifiers. Opt. Express **19**(11), 10180–10192 (2011)
10. Verdeyen, J.T.: Laser Electronics, 3rd edn. Prentice Hall, Englewood Cliffs (1995)

A Thermal Extension of Tellinen's Scalar Hysteresis Model

Jan Kühn, Andreas Bartel, and Piotr Putek

Abstract There exist different models which approximate the phenomenon of magnetic hysteresis. Only some of them inherently consider the thermal dependency of hysteresis or have been extended with that respect. In this paper, Tellinen's scalar magnetic hysteresis model is reviewed and illustrated. Thereby, special focus is laid upon the physical motivation. Afterwards, the underlying concept is adapted and extended w.r.t. thermal behavior. In the end, a temperature dependent scalar magnetic hysteresis model is deduced and investigated.

1 Introduction and Motivation

In many applications, ferromagnetic materials are exposed to a broad range of temperatures. In extreme cases the temperature approaches or exceeds the Curie point. Then, the material loses its ferromagnetic properties and transits to a paramagnetic state. But even for smaller temperature changes, the magnetic properties are affected. This has been and is still being researched [2, 3]. By now, several magnetic hysteresis models exist, e.g., [1, 4, 5, 9]. However, only some of them originally incorporated thermal influence or have been extended in that sense (e.g. [6, 7]). To our knowledge a temperature dependent Tellinen's model [9] has not yet been proposed. We have chosen Tellinen's model because it is easy to understand, simple to implement and overall fast to compute. Despite of that, it is still competitive in comparison to more complicated models [8]. Tellinen's model is motivated by phenomenological observations and can be more or less described graphically. It is very adaptable w.r.t. the input parameters. We will derive an extended model, such that it includes temperature dependence, while retaining the above mentioned

J. Kühn (✉) · A. Bartel
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: kuehn@math.uni-wuppertal.de; bartel@math.uni-wuppertal.de

P. Putek
Universität Rostock, Rostock, Germany
e-mail: piotr.putek@uni-rostock.de

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_6

beneficial properties. The following sections are organized as follows: First, the original model is introduced and the approaches used are presented. The thermal extension is then modeled on the basis of a similar idea. Afterwards, some aspects for embedding the model in a magnetic field simulation are discussed.

2 Tellinen's Hysteresis Model

Originally, Tellinen's hysteresis model was defined in [9]. In order to prepare the thermal extension, we summarize some important model properties.

Input Data We start from a magnetically fully negatively saturated material, let say $h, b \ll 0$ with a scalar magnetic flux density b and field strength h. Then, the transition to the saturated state of opposite polarity, $h, b \gg 0$, is described by a material specific function $b = B_{\text{sat}}^{+}(h)$ (for a monotone transition from $h \ll 0$ to $h \gg 0$). It is called the saturation or limiting curve. For B_{sat}^{+} to be a valid input for Tellinen's model, a few constraints must be fulfilled. Firstly, since both magnetic polarization directions are interchangeable, there must exist an analogous function for the transition from $h \gg 0$ to $h \ll 0$, say $B_{\text{sat}}^{-}(h)$ which can be defined by symmetry, i.e.,

$$B_{\text{sat}}^{-}(h) := -B_{\text{sat}}^{+}(-h)\,. \tag{1}$$

Physically, B_{sat}^{-} must lie above B_{sat}^{+}, such that eventually a loop is formed, i.e.,

$$B_{\text{sat}}^{+}(h) < B_{\text{sat}}^{-}(h), \qquad \lim_{|h|\to\infty} \big(B_{\text{sat}}^{-}(h) - B_{\text{sat}}^{+}(h)\big) = 0, \tag{2}$$

see Fig. 1. Secondly, B_{sat}^{+} must be differentiable (C^1) with bounds from below on the derivative by μ_0 (permeability of vacuum):

$$\frac{d}{dh} B_{\text{sat}}^{+}(h) \geq \mu_0 > 0\,, \qquad \lim_{|h|\to\infty} \frac{d}{dh} B_{\text{sat}}^{+}(h) = \mu_0\,. \tag{3}$$

Due to (3), Tellinen's model is valid for ferromagnetic and paramagnetic materials. In the further course, let the given $B_{\text{sat}}^{\pm}$ satisfy the constraints above.

Realization and Physical Motivation Tellinen's model is based upon two major principles. Firstly, the derivative $\frac{db}{dh}$ for the current working point (h, b) is calculated depending on whether h is de- or increased. This reflects the fact that hysteresis is not reversible. Secondly, the values for the derivative $\frac{db}{dh}$ are fixed on both boundaries B_{sat}^{+} and B_{sat}^{-} and intermediate values are determined by lin. interpolation. The choice of the boundary values is physically motivated and also derived from input data.

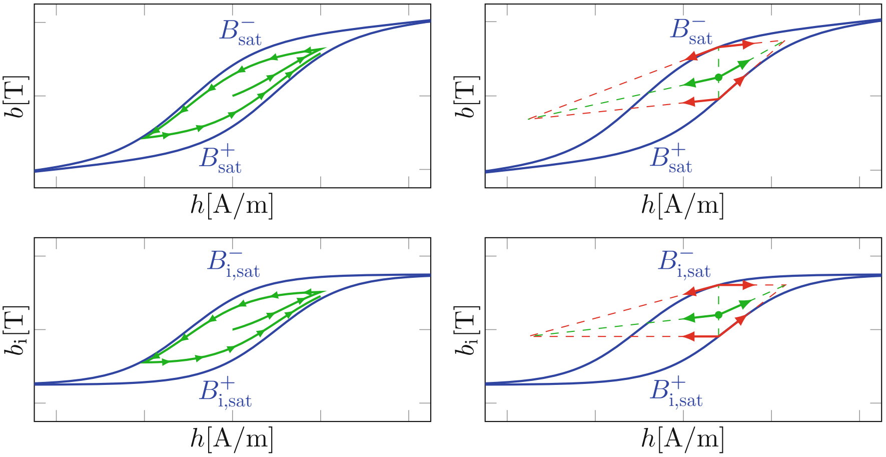


Fig. 1 The first row represents b/h-curves, while the second row gives b_{i}/h-curves for intrinsic induction. The left column depicts a sample hysteresis loop starting from the origin. The right column shows the idea of Tellinen's model. For a given working point (h, b) the resulting slope μ_{diff} is depicted (inner arrows), depending on increasing or decreasing field strength

To define Tellinen's model, we introduce the set of all valid states I within the boundaries B^+_{sat} and B^-_{sat} by

$$I := \left\{(h, b) \in \mathbb{R}^2 \;\middle|\; B^+_{\text{sat}}(h) \leq b \leq B^-_{\text{sat}}(h)\right\}. \tag{4}$$

For any $(h, b) \in I$, the relative (vertical) position between the boundaries B^+_{sat} and B^-_{sat} is calculated by

$$\lambda = \frac{B^-_{\text{sat}}(h) - b}{B^-_{\text{sat}}(h) - B^+_{\text{sat}}(h)} \in [0, 1]\,. \tag{5}$$

Now, combining the values of $\frac{db}{dh}$ on the boundary, i.e., for $\lambda = 0$ and $\lambda = 1$ and linear interpolation, we obtain for the distinguished $dh < 0$ and $dh > 0$ cases [9]

$$\frac{db}{dh} = \begin{cases} \mu^+_{\text{diff}} = \lambda \dfrac{dB^+_{\text{sat}}(h)}{dh} + (1-\lambda)\mu_0 & \text{if } dh > 0\,, \\[2ex] \mu^-_{\text{diff}} = \lambda\mu_0 + (1-\lambda)\dfrac{dB^-_{\text{sat}}(h)}{dh} & \text{if } dh < 0\,. \end{cases} \tag{6}$$

Tellinen's model is fully defined by ((5)–(6)). By construction, an analytical solution starting at $(h_0, b_0) \in I$, progressed by (6), always stays in I. To physically motivate (6), we delve further into the derivation [9]. To this end, we examine the intrinsic induction $b_{\text{i}} := b(h) - \mu_0 h$. Here, the pure vacuum term $\mu_0 h$ is subtracted. On the boundary (saturation), we have

$$B^{\pm}_{\text{i,sat}}(h) = B^{\pm}_{\text{sat}}(h) - \mu_0 h\,, \qquad\qquad \frac{dB^{\pm}_{\text{i,sat}}}{dh} = \frac{dB^{\pm}_{\text{sat}}}{dh} - \mu_0\,. \tag{7}$$

All properties of $B^{\pm}_{\text{sat}}$ and $\frac{dB^{\pm}_{\text{sat}}}{dh}$ can easily be transferred to $B^{\pm}_{\text{i,sat}}$ and $\frac{dB^{\pm}_{\text{i,sat}}}{dh}$, see Fig. 1. In [9], it is argued that a saturated material, i.e., $b_{\text{i}} = B^{+}_{\text{i,sat}}(h)$ or $b_{\text{i}} = B^{-}_{\text{i,sat}}(h)$, resists a magnetization of opposing polarization. Thus, if the material is in the state (h, b_{i}) with $b_{\text{i}} = B^{+}_{\text{i,sat}}(h)$, we model $\frac{db_{\text{i}}}{dh} = \frac{dB^{+}_{\text{i,sat}}}{dh}$ for $dh > 0$, because this has been measured by $B^{+}_{\text{i,sat}}$. But for $dh < 0$, we explicitly set $\frac{db_{\text{i}}}{dh} = 0$, because the material opposes the magnetization, and therefore, does not contribute to b_{i}. The analogous approach is used for $b_{\text{i}} = B^{-}_{\text{i,sat}}(h)$. Together with (7) this results in (6).

Implementation Details and Drawbacks The implementation of Tellinen's model is easy, but the integration into a simulation tool, e.g., based on finite elements, has to be done carefully. For a critical discussion, we mention two details:

(a) A priori in a time step, it is unknown if h is locally increasing or decreasing, and thus, if μ^{+}_{diff} or μ^{-}_{diff} has to be used. An iterative solver remains applicable, but the discontinuity may hinder the solver and may lead to a not converging sequence. To fix this, one can create a smooth function μ_{diff}, e.g., by employing a sigmoid-like transition between μ^{+}_{diff} and μ^{-}_{diff}. Unfortunately, this increases the sensitivity of the problem, and thus can lead to a significantly reduced time step size. Moreover, if the time step sizes are chosen too small, μ_{diff} might use the artificial values of the transition zone for a longer period yielding inappropriate overall simulation results. For these reasons, we have not pursued this concept any further.

In our current implementation, μ^{+}_{diff} or μ^{-}_{diff} is used if the h value was in- or decreased in the last time step, respectively. Hence, this approach can only react to a change of $\text{sgn}(\frac{dh}{dt})$ with a delay of one time step. In a simulation, μ_{diff} needs to be computed on a discrete set of points (for each element) and we need to store the data from the previous time step. We globally initialize $\mu_{\text{diff}} = \frac{1}{2}(\mu^{+}_{\text{diff}} + \mu^{-}_{\text{diff}})$, and then, update μ_{diff} element-wise after each calculation by

$$\mu_{\text{diff}} = \begin{cases} \mu^{+}_{\text{diff}} & \text{if } h_{\text{new}} > h_{\text{old}}, \\ \mu^{-}_{\text{diff}} & \text{if } h_{\text{new}} < h_{\text{old}}. \end{cases} \tag{8}$$

The advantage of this procedure is that μ_{diff} is fixed locally and per time step. Therefore, it does not hinder the solver in any way and it allows the use of non-iterative solvers.

(b) Tellinen's model does not calculate μ, b or h. For a given state (h, b) and direction $\operatorname{sgn}(\frac{dh}{dt})$, it outputs the change of b, i.e., $\mu_{\text{diff}} = \frac{db}{dh}$. To obtain the new field values, we suggest a linear approach

$$b_{\text{new}} = b_{\text{old}} + \mu_{\text{diff}}(h_{\text{new}} - h_{\text{old}}) \quad \text{or} \quad h_{\text{new}} = h_{\text{old}} + \frac{1}{\mu_{\text{diff}}}(b_{\text{new}} - b_{\text{old}}). \tag{9}$$

For example, for the magnetoquasistatic approximation of Maxwell's equations in terms of the magnetic vector potential $\mathbf{A}$, one obtains the curl-curl equation:

$$\sigma \frac{d\mathbf{A}}{dt} + \nabla \times \left(\frac{1}{\mu} (\nabla \times \mathbf{A}) \right) = \mathbf{J} \tag{10}$$

with conductivity σ, applied current density $\mathbf{J}$ and $\mathbf{B} = \nabla \times \mathbf{A}$. Using (9), this yields

$$\sigma \frac{d\mathbf{A}}{dt} + \nabla \times \left(\frac{1}{\mu_{\text{diff}}} (\nabla \times \mathbf{A}) \right) = \mathbf{J} + \nabla \times \left(\frac{1}{\mu_{\text{diff}}} \mathbf{B}_{\text{old}} - \mathbf{H}_{\text{old}} \right). \tag{11}$$

Here, the incorporation of Tellinen's model into (10) is simple. Since the left-hand side remains the same, the solvability should not be worsened.

Possible Further Improvements Tellinen's model needs very few inputs, namely B_{sat}^{+} and its derivative. Thus a combination with other models as preprocessor is possible if the inputs are computable. For example in [8], the Jiles-Atherton model [4] is used to generate B_{sat}^{+}. Thus, the Tellinen model can be easily compared and combined with other models as long as B_{sat}^{+} is shared.

Moreover, it is possible to add the data of intermediate first-order reversal curves to the measured saturation curves B_{sat}^{+} and B_{sat}^{-}, see Fig. 2. Then, the linear interpolation for a given operation point (b, h) can be restricted to the two adjacent curves. This approach is necessary if the material exhibits a pronounced non-linear behavior along h =const and $B_{\text{sat}}^{+}(h) \leq b \leq B_{\text{sat}}^{-}(h)$. The complexity of this approach is low and the additional computation cost at runtime is small.

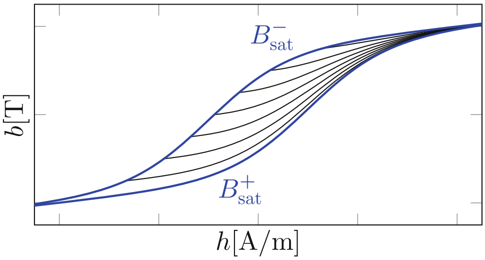


Fig. 2 An example of saturation curves B_{sat}^{+}, B_{sat}^{-} and intermediate curves, so-called first-order reversal curves

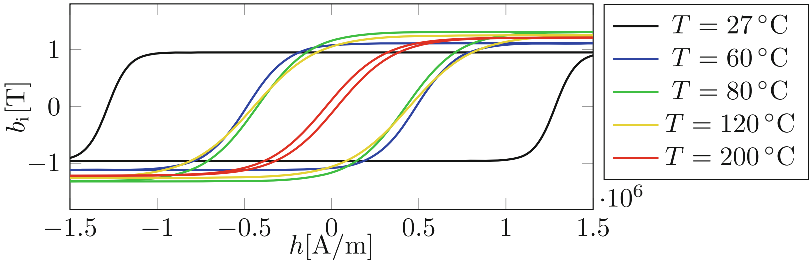


Fig. 3 Sample curves depicting the intrinsic induction b_i versus the magnetic field strength h for various temperatures. These curves are based upon measurements of NdFeB magnets [2], but simplified for demonstration

These improvements can as well be applied to the thermal extension developed next.

3 Thermal Extension of Hysteresis

Using the same concepts as Tellinen, we develop a thermal extension by defining the partial derivatives $\frac{\partial b}{\partial T}$ dependent on the $sgn(\partial T)$. Again, a physical motivation is given based on measured data for the saturation boundary and based on linear interpolation for the intermediate points. Ultimately, the extended model shall approximate the material behavior in terms of b if a current state (h, b, T) is changed w.r.t. to h and T (Fig. 3).

Input Data Let $B^+_\mathrm{sat} = B^+_\mathrm{sat}(h, T) \in C^1$ be a function that defines the saturation value of b for a given field strength h and temperature T in the case where h is monotonously increased from $h \ll 0$ to $h \gg 0$. Analogously to (3), the derivative w.r.t. h has to be bounded from below:

$$\frac{\partial}{\partial h} B^+_\mathrm{sat}(h, T) \geq \mu_0\,, \qquad \lim_{|h|\to\infty} \tfrac{\partial}{\partial h} B^+_\mathrm{sat}(h, T) = \mu_0\,.$$

As in (2), $B^-_\mathrm{sat}(h, T)$ can be defined by symmetry and shall satisfy

$$B^+_\mathrm{sat}(h, T) < B^-_\mathrm{sat}(h, T) := -B^+_\mathrm{sat}(-h, T), \quad \lim_{|h|\to\infty} \left(B^-_\mathrm{sat}(h, T) - B^+_\mathrm{sat}(h, T)\right) = 0.$$

Here, the valid states (4) become

$$I_\mathrm{T} = \left\{(h, b, T) \in \mathbb{R}^3 \;\middle|\; B^+_\mathrm{sat}(h, T) \leq b \leq B^-_\mathrm{sat}(h, T)\right\}.$$

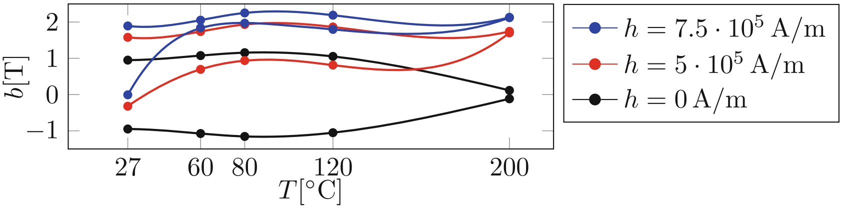


Fig. 4 Shown are $B^+_{\text{sat}}(h, T)$ and $B^-_{\text{sat}}(h, T)$ for fixed values of h. The same dataset as in Fig. 3 is used. The interpolation is based on a spline approach

One possibility to define such $B^+_{\text{sat}}(h, T)$ is to interpolate several b/h-curves of different temperatures. For fixed h, $B^{\pm}_{\text{sat}}(h, \cdot)$ might lack monotonicity, see Fig. 4.

Physical Motivation and Computation For a fixed temperature T and varying h, the non-thermal Tellinen model is used by just replacing $B^{\pm}_{\text{sat}}(h)$ by $B^{\pm}_{\text{sat}}(h, T)$. We will now discuss the situation of fixed h and varying T. The combination of both defines the temperature dependent Tellinen model. We remark that no solution starting from a valid point $(h, b, T) \in I_{\text{T}}$ may leave I_{T}. Now, for given surfaces $B^{\pm}_{\text{sat}}(h, T)$ and for any $(h, b, T) \in I_{\text{T}}$, λ can be calculated analogously to (5). Firstly, for an increasing temperature $\partial T > 0$, it can be argued that the Weiss domains are changing because the domain walls are more likely moving from their pinned state. Therefore, this model tries to follow the measured values on the boundaries. Hence, we define

$$\frac{\partial b}{\partial T} = \lambda \frac{\partial}{\partial T} B^+_{\text{sat}}(h, T) + (1 - \lambda) \frac{\partial}{\partial T} B^-_{\text{sat}}(h, T) \quad \text{for } \partial T > 0 . \tag{12}$$

Secondly, reducing the temperature, the ability of the Weiss domains to change is decreased. They tend to freeze and try to retain their current state. Therefore, for $\partial T < 0$, we like to assign $\frac{\partial b}{\partial T} = 0$. However, there are cases where this will lead to values of b outside of the set I_{T}. To handle this, we assigns the value zero whenever possible. Otherwise, the measured value on the boundary is used, such that I_{T} cannot be left. This reads: (see Fig. 5)

$$\frac{\partial b}{\partial T} = \lambda \min\Big(\tfrac{\partial}{\partial T} B^+_{\text{sat}}(h, T), 0\Big) + (1-\lambda) \max\Big(\tfrac{\partial}{\partial T} B^-_{\text{sat}}(h, T), 0\Big) \quad \text{for } \partial T < 0 . \tag{13}$$

Properties The definition of the partial derivatives ((12)–(13)) ensures that an analytical solution starting in I_{T} always stays within I_{T}. Keeping h constant and increasing T, (12) will change state from (T_0, b_0) to (T_1, b_1). But reducing T back to T_0 via (13), one generally ends up in (T_0, b_2) with $b_0 \neq b_2$, see Fig. 5. Thus, this model reflects a non-reversible behavior. Nevertheless, also stable loops exist. It can be shown that for a given interval $[T_0, T_1]$, there is exactly one stable loop such that

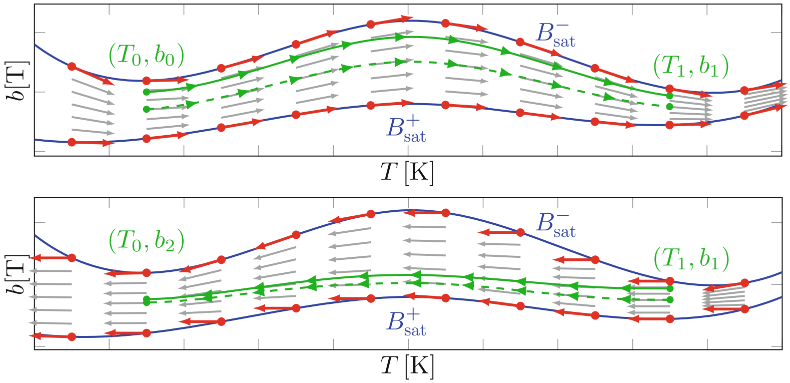


Fig. 5 For constant h, exemplary bounding saturation curves $B^{\pm}_{\text{sat}}(h, T)$ are shown. The arrows represent the assigned values of $\frac{\partial b}{\partial T}$ for an increasing (upper plot) and decreasing (lower plot) temperature T. The inner solid path represent the solution of b starting at (T_0, b_0), and increasing the temperature to T_1 (upper) and going back to T_0 (lower). In general, $b_0 \neq b_2$. The dashed curve shows a stable loop, where the start and end points coincide

$b_0 = b_2$ holds, or all loops are stable. If a loop is unstable, it converges against a stable loop if T is periodically alternated between T_0 and T_1. The situation that all loops for which $B^+_{\text{sat}} \leq b_0 \leq B^-_{\text{sat}}$ are stable, can only occur if $\frac{\partial}{\partial T} B^+_{\text{sat}}(h, T) \leq 0$ and $\frac{\partial}{\partial T} B^-_{\text{sat}}(h, T) \geq 0$ for all $T \in [T_0, T_1]$. In this special case, (12) and (13) are equal and a reversible process is given. For $h = 0$ and uniqueness of the stable curve, this stable curve is defined by $b = 0$ for all T. Thus, any valid b_0 will lead to a gradual depolarization of the material.

Simulation In time domain, the temperature is often slower evolving than the magnetic field. Therefore, we allow that the temperature values are updated only in every k-th time step; elsewhere, they are assumed to be constant. Now, to implement the thermal extension, we perform an additional adjustment of the b field in every k-th time step. This can be realized by the following linear approach

$$b_{\text{new}} = b_{\text{old}} + \frac{\partial b}{\partial T}(T_{\text{new}} - T_{\text{old}}) \,, \tag{14}$$

where $\frac{\partial b}{\partial T}$ is given by (12) or (13). Overall, it is ensured that an analytical solution stays in $\in I_{\text{T}}$. This approach enables multirate simulation (for $k > 1$).

4 Conclusion and Outlook

A thermal extension of Tellinen's hysteresis model was proposed, based on the same ideas as given in the original version. The thermal model describes b in terms of h and T. It reflects fundamental physical phenomena, e.g., the irreversibility and depolarization. Currently, we investigate the computation of magnetic losses based on the extended thermal model. A comparison to other thermal models is also a pending task.

References

1. Bobbio, S., Milano, G., Serpico, C., Visone, C.: Models of magnetic hysteresis based on play and stop hysterons. IEEE Trans. Magn. **33**, 4417–4426 (1997)
2. Ghezelbash, M., Darbani, S., Majd, A., Ghasemi, A.: Temperature dependence of magnetic hysteresis loop of NdFeB with uniaxial anisotropy by LIBS technique. J. Supercond. Nov. Magn. **30**, 1893–1898 (2017)
3. Güntzel, U., Westerholt, K., Methfessel, S.: Temperature dependence of the magnetic hysteresis properties of some metglasses between 4.2 and 300K. J. Magn. Magn. Mater. **38**, 294–300 (1983)
4. Jiles, D.C., Atherton, D.L.: Theory of ferromagnetic hysteresis. J. Appl. Phys. **55**, 2115–2120 (1984)
5. Preisach, F.: Über Die Magnetische Nachwirkung. Z. Phys. **94**, 277–302 (1935)
6. Raghunathan, A., Melikhov, Y., Snyder, J., Jiles, D.: Theoretical model of temperature dependence of hysteresis based on mean field theory. IEEE Trans. Magn. **46**, 1507–1510 (2010)
7. Sixdenier, F., Messal, O., Hilal, A., Martin, C., Raulet, M., Scorretti, R.: Temperature-dependent extension of a static hysteresis model. IEEE Trans. Magn. **52**, 1–4 (2016)
8. Steentjes, S., Hameyer, K., Dolinar, D., Petrun, M.: Iron-loss and magnetic hysteresis under arbitrary waveforms in NO electrical steel: a comparative study of hysteresis models. IEEE Trans. Ind. Electron. **64**, 2511–2521 (2017)
9. Tellinen, J.: A simple scalar model for magnetic hysteresis. IEEE Trans. Magn. **34**, 2200–2206 (1998)

Computational Characterization of a Composite Ceramic Block for a Millimeter Wave Heat Exchanger

Petra Kumi, Jonathan S. Venne, Vadim V. Yakovlev, Martin S. Hilario, Brad W. Hoff, and Ian M. Rittersdorf

Abstract Electromagnetic and electromagnetic-thermal coupled problems are solved by the finite-difference time-domain technique for an AlN:Mo composite ceramic block backed by a thin metal plate and irradiated by a high-power W-band plane wave. Computation is based on experimental data on temperature-dependent complex permittivity, specific heat, and thermal conductivity. Non-uniformity of patterns of dissipated power and temperature is quantified via standard-deviation-based metrics. A 10×10×10±2 mm block of the composite with concentration of Mo from 0.25 to 4% is considered as irradiated by the plane wave with power density on the block's front surface from 0.3 to 1.0 W/mm^2 at 95 GHz. It is shown that the block can be heated up to 1000 ^{o}C highly uniformly for 50–110 s. The composite producing maximum total dissipated power is found to have Mo content about 3%.

1 Introduction

Electromagnetic (EM) heating is employed in applications in food engineering, chemistry, materials science [1, 2]. This phenomenon is also in the core of recently introduced EM heat exchangers (HX) used in solar energy collectors [3], power beaming applications [4], and microwave thermal thrusters [5]. Millimeter-wave (MMW) powered HX are currently under development for the use in power beaming experiments for wireless energy transfer [6–8]. Interactions of the MMW field

P. Kumi · J. S. Venne · V. V. Yakovlev (✉)
Worcester Polytechnic Institute, Worcester, MA, USA
e-mail: pkumi@wpi.edu; jsvenne@wpi.edu; vadim@wpi.edu

M. S. Hilario · B. W. Hoff
Air Force Research Laboratory, Kirtland AFB, Albuquerque, NM, USA
e-mail: martin.hilario@us.af.mil; brad.hoff@us.af.mil

I. M. Rittersdorf
Naval Research Laboratory, Washington, DC, USA
e-mail: ian.rittersdorf@nrl.navy.mil

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_7

with ceramic slabs should be well understood for effectively utilizing EM energy, controllably heating the material, and transferring heat to fluid flowing through the neighboring channels.

In this paper, by the means of EM and coupled EM-thermal modeling, we study the process of heating of a rectangular block of an aluminum nitride-molybdenum (AlN:Mo) composite by a normally incident plane MMW. The back surface of the block is in contact with a thin metal plate. Characterizing this system in terms of homogeneity of heating and a level of absorbed power, we identify the material which appears to be most efficient for using in a physical prototype of a MMW HX.

2 Computational Scenario & Input Data

A considered MMW-powered HX contains an assembly of ceramic tiles and a metal baseplate that is attached to their back surfaces and contains channels with fluid flow (Fig. 1). Due to dielectric losses, EM power is dissipated inside the ceramic blocks and induces thermal field that heats the baseplate and the flowing fluid.

In order to help design the physical prototype, we build a computer model capable of simulating EM and thermal processes in a single ceramic block irradiated by a plane wave and attached to a thin metal plate (Fig. 2). An ultimate goal of the computational experiments is to find a composition of the AlN:Mo composite that leads to maximum absorption of EM energy and, at the same time, to as uniform temperature distribution on the back surface of the block as possible.

The considered factors affecting the system performance are limited to its EM and thermal material parameters: dielectric constant ϵ', the loss factor ϵ'', density ρ, specific heat c, and thermal conductivity k. We examine AlN samples with different levels of Mo doping; by varying percentage of Mo in the AlN:Mo composite, both EM and thermal parameters can be significantly changed.

Temperature characteristics of ϵ' and ϵ'' were obtained with the use of a dedicated apparatus for a free space dielectric measurement [7, 9] at 95 GHz for temperatures T up to nearly 550 °C. The measurements were carried out for six samples of AlN with Mo content from 0.25 to 4% by volume. The characteristics turned out to be almost linearly dependent on T (Fig. 3). Corresponding linear polynomials (also shown in Fig. 3) approximating the experimental data points were used to generate the values of ϵ' and ϵ'' at 20, 100, 200, ..., 1000 °C as input data for the EM model.

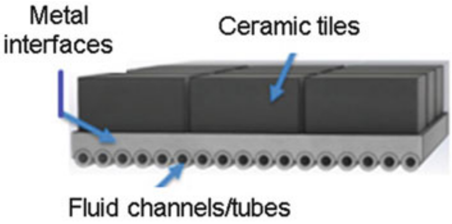


Fig. 1 Conceptual drawing of a section of a MMW-powered HX

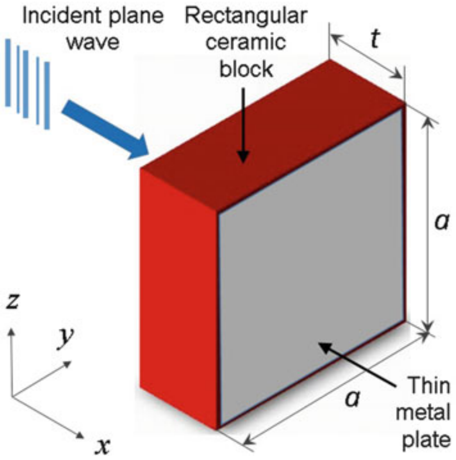


Fig. 2 3D view of the considered computational scenario

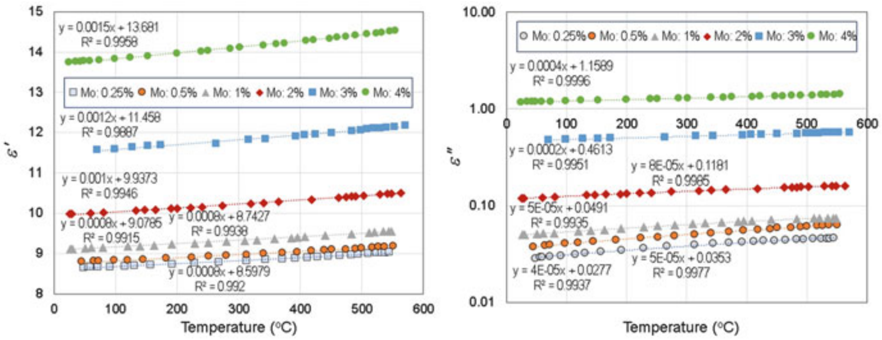


Fig. 3 Measured points and linearly approximated temperature characteristics of dielectric constant ϵ' and the loss factor ϵ'' of six AlN:Mo samples with different concentrations of Mo

Values of specific heat and thermal conductivity were measured for 4% of Mo at seven temperatures between 20 and 1001 °C (large points in Fig. 4), and they were assumed to be similar to those of lower concentrations of Mo. The values of c and k at the intermediate points (i.e., at 50, 150, 200, …, 950 °C) were determined from cubic spline interpolations of the experimental points; those values are also shown in Fig. 4 (small points).

Density ρ was measured for 4% of Mo at room temperature and found to be 3.66 g/cm^3. It was used in all computations under the assumption that it does not depend on concentration of Mo and temperature.

Computations were carried out for the AlN:Mo block with a = 10 mm and t = 8, 10 and 12 mm. Thickness of the thin metal plate was 1 mm. The frequency of the incoming field was 95 GHz. The incident power of the plane wave was set as power density at the front surface of the composite block P_{in} being from 0.3 to 1.0 W/mm^2;

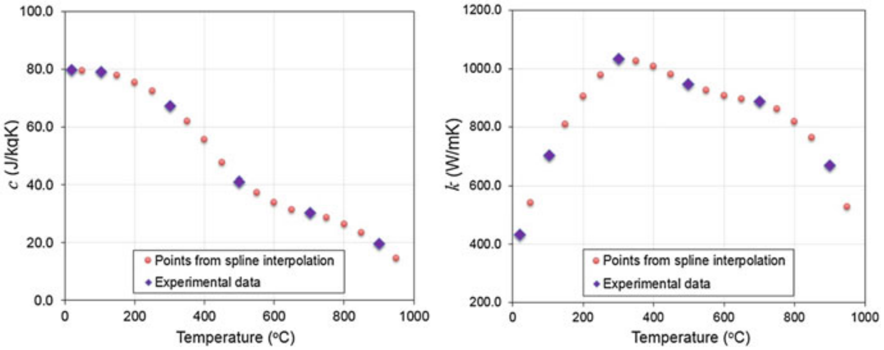


Fig. 4 Measured and spline-interpolated points of temperature characteristics of specific heat c and thermal conductivity k of the AlN:Mo sample with 4% concentrations of Mo

these values were expected to be consistent with the power and radiating antennas typical for MMW HX operations.

3 Computational Techniques

3.1 Electromagnetic and Thermal Problems

The process of MMW heating of an AlN:Mo block is studied with EM and EM-thermal (coupled) models. Both underlying (EM and thermal) problems are numerically solved by a 3D finite-difference time-domain (FDTD) technique. That is, similarly to the fundamental FDTD step with Maxwell's equations, both spatial and temporal partial derivatives in the heat transfer equation are approximated by the central-difference formula, and the differential equation is converted into an algebraic one. Both EM and thermal problems are discretized using the same mesh.

In order to satisfy the Courant condition and minimize computational resources, the scenario is discretized by the mesh with maximum cell sizes of 0.295 and 0.095 mm in air and in the material, respectively. The $t \times a \times a$ mm AlN:Mo block is surrounded by the $(t+8)\times(a+3)\times(a+3)$ mm plane wave box with one $(a+3)\times(a+3)$ mm face responsible for sinusoidal excitation of the incident field. The ceramic block is situated inside the $(t+14)\times(a+9)\times(a+9)$ mm box imitating the Mur with superabsorption boundary condition.

Depending on the size of the block, the model consists of 2.9 to 3.5 mil cells. It is capable of computing patterns of dissipated power density (P_d) in the volume of the composite due to dielectric losses. These patterns are identical to temperature distributions at the initial moment of heat dissipation due to thermal conductivity.

The heat transfer problem is solved under Neumann (adiabatic) boundary conditions on all ceramic-air and ceramic-metal interfaces.

The EM and thermal solvers operate as part of an iterative procedure in which a steady state solution of the EM problem becomes an input for the thermal problem, and, in the repeated runs of the EM solver, material parameters are upgraded in every cell in accordance with the temperature field outputted from the thermal solver. The latter determines temperature distribution induced after each heating time step τ; N steps are needed to reach maximum temperature of the process (1000 °C).

The above computational scheme mimics time evolution of the MMW-induced temperature field under the given temperature-dependent EM and thermal material parameters. The procedure employed for coupling the two solvers is similar to the ones outlined in [10, 11]. The models were built in the environment of the FDTD solver *QuickWave* that includes its thermal units *QW-BHM* and *QW-HFM* [12].

3.2 Metric of Uniformity and Total Dissipated Power

Sufficient homogeneity of heating of the ceramic block is one of the required features of a MMW HX. A formal metric of field (non-) uniformity is necessary for quantitative characterization of the patterns and formulation of corresponding optimization problems. Following discussions in [13], it is introduced for P_d fields as:

$$\lambda_P = \frac{\sigma}{\mu}, \tag{1}$$

where σ is the standard deviation of the P_d values in all FDTD cells in the block from the mean value μ. When dealing with temperature field evolving in time, metric (1) is applied to all n patterns (corresponding to n different temperatures, $n \leq N$, where N is number of heating time steps τ) representing the heating process. The latter is therefore characterized by the metric

$$\lambda_T = \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_i}{\mu_i}, \tag{2}$$

where σ_i is the standard deviation of the T values in all FDTD cells in the block from the mean value in the ith pattern μ_i. Evaluation of uniformity of 2D patterns is based on the values of P_d or T in a particular layer of FDTD cells, and for 3D patterns, $\lambda_{P,T}$ are found from the values in the cells in the block's volume.

Denoting the total power dissipated in the ceramic material as $\bar{P}_d$ and the power brought by the incident wave as P, we introduce parameter η as a measure of energy efficiency of the MMW heating process as

$$\eta = \frac{\bar{P}_d}{P}. \tag{3}$$

Since ϵ' and ϵ'' vary with temperature, the value of $\bar{P}_d$ is different for every T. To characterize the heating process in terms of total dissipated power, we calculate

$$\hat{P}_d = \frac{1}{n}\sum_{i=1}^{n} \bar{P}_{di}, \; n \leq N, \tag{4}$$

where $\bar{P}_{di}$ is the total dissipated power after the ith heating time step, and find η in (3) by using $\hat{P}_d$ instead of $\bar{P}_d$.

4 Results and Discussion

Typical distributions of P_d obtained from the EM model are shown in Fig. 5. For T = 400 and 800 °C, the patterns were computed under the assumption that the material was heated uniformly and therefore the values of complex permittivity were the same throughout the block. Of particular interest is distribution of heat on the back surface, so patterns in Fig. 5 show distributions of dissipated power density in the last layer of the FDTD cells in the ceramic block that is perpendicular to the x-axis.

It appears that P_d patterns, regardless the level of doping, are characterized by vast amount of maxima/minima. When the metal plate is attached, the patterns essentially change. A number of "hot spots" spread over the area of 10×10 mm

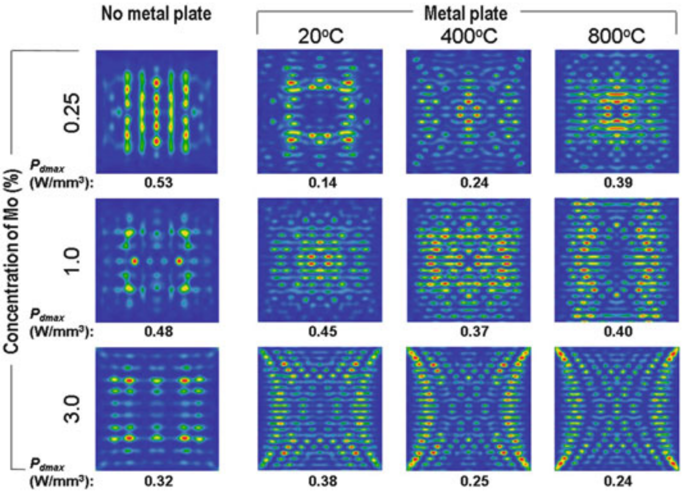


Fig. 5 Patterns of dissipated power density in the yz-plane on the back surface of the AlN:Mo block with 0.25, 1, and 3% of Mo at T = 20, 400, and 800 °C and maximum values of dissipated power density (P_{dmax}); P_{in} = 1 W/mm^2; t = 10 mm

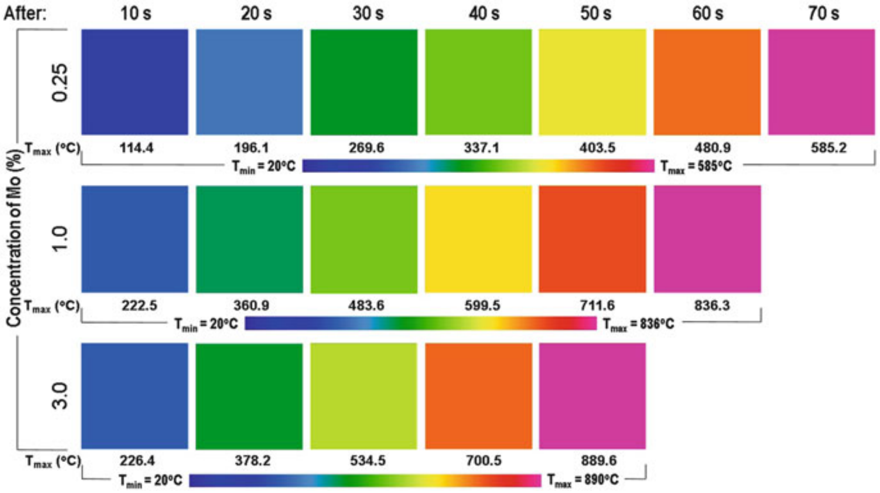


Fig. 6 Normalized temperature distributions in the yz-plane on the back surface of the AiN:Mo block with 0.25, 1, and 3% of Mo and maximum values of temperature (T_{max}); heating time steps 5 s (0.25% Mo), 2.5 s (1.0% Mo), 2 s (3.0% Mo); P_{in} = 1 W/mm^2; t = 10 mm

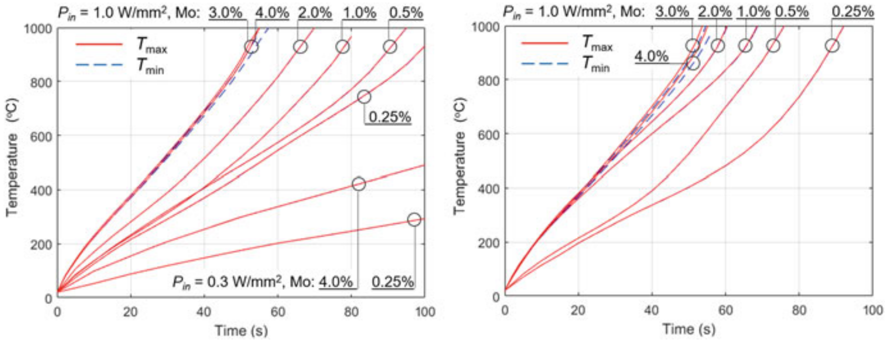


Fig. 7 Time evolution of maximum and minimum temperatures (T_{max} and T_{min}) of the MMW-heated AlN:Mo blocks with different concentrations of Mo (0.25 to 4%) and power densities of the incident wave P_{in} (0.3 and 1 W/mm^2); a thin metal plate is absent (left) and present (right); t = 10 mm

suggests that, due to high value of thermal conductivity of the AlN:Mo composites, the material may be heated sufficiently uniformly.

Temperature distributions obtained from the coupled model and shown in Fig. 6 depict the heating processes of three AlN:Mo blocks. Exceeding the expectation suggested by Fig. 5, the patterns are exceptionally uniform for all compositions of the material. (This, in turn, justifies the assumption about uniformity of heating that was used in computations of dissipated power in the EM model.)

The very high level of uniformity of temperature patterns in Fig. 6 (with λ_T in (2) being of the order of 10^{-4}) is also illustrated by the graphs in Fig. 7. It is seen that

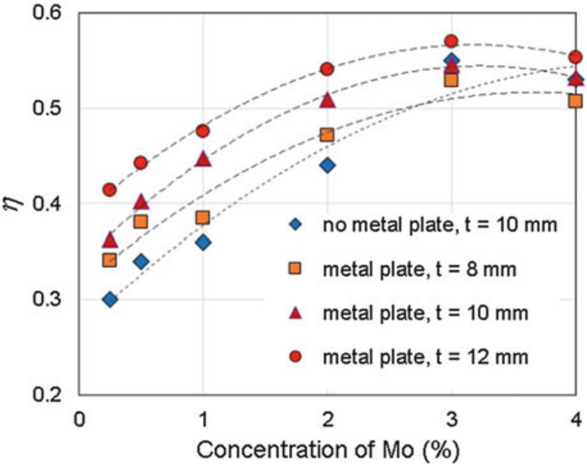


Fig. 8 Energy efficiency of the system with the thin metal plate absent and present

the time increases of minimum and maximum temperatures are almost identical; one can see small (5 to 50 °C) differences between T_{max} and T_{min} only in the composites with higher concentrations of Mo that are heated most rapidly.

When the power density of the incident field is 1 W/mm^2, temperature in the considered block reaches 1000 °C for the time between 50 and 110 s; the higher concentration of Mo, the faster the heating. This is due to the impact of composite's dielectric losses which notably increase with the rise of the contents of Mo.

When a back surface of the block is in contact with a metal plate, the composites with lower concentrations of Mo reach 1000 °C faster. The pace of heating for P_{in} = 1 W/mm^2 is found to be about three times higher than for 0.30 W/mm^2.

Parameter η in (3) computed for six Mo concentrations and the cases of absent and present metal plate reveals a maximum energy efficiency of the composites with Mo near 3% (Fig. 8). Variation in thickness (from 8 to 12 mm) does not affect the position of the maximum. The existence of the extreme value may be caused by a significant decay of both the incident and reflected fields in the blocks with high contents of Mo (3% and higher). The model may be built into a suitable optimization procedure searching for maximum energy efficiency of the system.

5 Conclusions

The performed computational studies allow for the following observations. When the plane wave carries the power of a practical level and that power is represented by the power density on the front surface of the composite block (so that P_{in} is from 0.3 to 1.0 W/mm^2), the AlN:Mo 10×10×10±2 mm samples are heated up to

1000 °C for 50–110 s, depending on the incoming power and concentration of Mo, and with a very high level of uniformity ($\lambda_T \approx 10^{-4}$). When the back surface of the ceramic block is in contact with a metal plate, absorption of MMW power is higher than in its absence.

While the blocks made of all the considered composites are heated highly uniformly, the materials with 3% concentration of molybdenum appear to be more energy efficient: for all thicknesses of the block, the parameter η takes on the largest values. On the other hand, preliminary data in production of the AlN:Mo composites suggests that the materials with Mo $\geq 3\%$ are mechanically more stable and robust. The composites of those concentrations of Mo appear to be good candidates for the use in the physical prototype of a MMW-powered heat exchanger.

The developed EM-thermal coupled model is open for further development, and it is expected to be stimulated by the upcoming experimental studies. That may include the oblique incidence of the plane wave, multiple composite blocks on the metal plate, Gaussian beams, and alternative thermal boundary conditions imitating some particular operational regimes.

Acknowledgments J.S. Venne and V.V, Yakovlev are grateful for the support from the Air Force Research Laboratory (via Leidos, Inc.), Grant No P010200226. The work of P. Kumi was supported by the 2018 WPI Summer Undergraduate Research Fellowship (SURF). B.W. Hoff and M.S. Hilario were funded by the Air Force Office of Scientific Research under Contract FA9550-17RDCOR449 and by the Operational Energy Capability Improvement Fund (OECIF). I.M. Rittersdorf was supported by OECIF.

References

1. Willert-Porada, M.: Advances in Microwave and Radio Frequency Processing. Springer, Berlin (2006)
2. Leadbeater, N.: Microwave Heating as a Tool for Sustainable Chemistry. CRC Press, Boca Raton (2010)
3. Jamar, A., Majid, Z.A.A., Azmi, W.H., Norhafana, M., Razak, A.A.: A review of water heating system for solar energy applications. Int. Commun. Heat Mass Transfer **76**, 178–187 (2016)
4. Jawdat, B., Hoff, B., Hilario, M., Baros, A., Pelletier, P., Sabo, T., Dynys, F.: Composite ceramics for power beaming. In: Proc. 2017 IEEE Wireless Power Transfer Conference, 978-1-5090-4595-3/17
5. Parkin, K.L., DiDomenico, L.D., Culick, F.E.: The microwave thermal thruster concept. In: 2nd Intern. Symp. on Beamed Energy Propulsion, pp. 418–429, 0-7354-0175-6/04 (2004)
6. Hoff, B.W., Hilario, M.S., Jawdat, B., Baros, A.E., Dynys, F.W., Mackey, J.A., Yakovlev, V.V., Andraka, C.E., Armijo, K.M., Savrun, E., Rittersdort, I.M.: Millimeter wave interactions with high temperature materials and their applications to power beaming. In: Proc. 52nd IMPI Microwave Power Symp., Long Beach, CA, pp. 82–83, June 2018
7. Mohekar, A.A., Gaone, J.M., Tilley, B.S., Yakovlev, V.V.: A 2D coupled electromagnetic, thermal and fluid flow model: application to layered microwave heat exchangers. In: IEEE MTT-S Intern. Microwave Symp. Dig., Philadelphia, PA, pp. 1389–1392, June 2018
8. Mohekar, A.A., Gaone, J.M., Tilley, B.S., Yakovlev, V.V.: Multiphysics simulation of temperature profiles in a triple-layer model of a microwave heat exchanger. In: Proc. 52nd IMPI's Microwave Power Symp., Long Beach, CA, pp. 33–35, June 2018

9. Hilario, M.S., Hoff, B.W., Young, M.P., Lanagan, M.T.: W-band free-space dielectric material property measurement techniques for beamed energy applications. In: Proc. 53rd AIAA Aerospace Sciences Meeting, Kissimmee, FL, Jan 2015
10. Celuch, M., Kopyt, P.: Modeling of microwave heating of foods. In: Lorence, M.W., Pesheck, P.S. (eds.) Development of Packaging and Products for Use in Microwave Ovens, pp. 305–348. Woodhead Publishing, Cambridge (2009)
11. Yakovlev, V.V., Allan, S.M., Fall, M.L., Shulman, H.S.: Computational study of thermal runaway in microwave processing of Zirconia. In: Tao J. (ed.) Microwave and RF Power Applications, pp. 303–306 Cépaduès Éditions, Toulouse (2011)
12. QuickWave: QWED Sp. z o. o., http://www.qwed.eu/
13. Wang, S.J., Luechapattanaporn, K., Tang, J.: Experimental methods for evaluating heating uniformity in radio frequency systems. Biosyst. Eng. J. **100**, 58–65 (2008)

Part II
Device Modeling and Simulation

Nine papers form this part dedicated to the contributions falling into the field of Device Modeling and Simulation (DMS).

Mechanochemistry in Li-ion batteries involve interaction of ions migrating in a cell with mechanical stresses as well as electromagnetic fields. B. E. Abali in his contribution *Modeling Mechanochemistry in Li-ion Batteries* models a battery cell by involving thermomechanics with the corresponding balance equations of mass, momentum, energy and electromagnetism with the aid of the Maxwell equations. Before applying specific assumptions relevant for Li-ion batteries, developing a complete theory by using continuum mechanics and thermodynamics is addressed.

Lately, organic semiconductor materials are going to play a relevant role in electron device. In the contribution *Automatic Extraction of Transport Model Parameters of an Organic Semiconductor Material* by P. C. Africa et al. a step-by-step procedure is presented that enables to determine the density of states width, the carrier mobility and the injection barrier height of an organic FET by fitting data from simple measurements to suitable numerical simulations. At each step of the procedure only one parameter value is determined, thus highly simplifying the fitting procedure and enhancing its robustness. The procedure is applied to p-type organic polymers and a very satisfactory fitting of experimental measurements is obtained.

On a Bloch-Type Model with Electron-Phonon Interactions: Modeling and Numerical Simulations the authors B. Bidégaray-Fesquet et al. discuss how to take into account electron-phonon interactions in a Bloch type model for the description of quantum dots. The model consists in coupling an equation on the density matrix with a set of equations on quantities called phonon-assisted densities, one for each phonon mode. After a description of the model, the authors discuss also how to discretize efficiently this non-linear coupling in view of numerical simulations.

Thermal effects are playing an increasing crucial role for the design of electron nanoscale devices. In *Charge and Phonon Transport in Suspended Monolayer Graphene* by M. Coco et al. charge and phonon transport under an applied external electric field is investigated in a suspended monolayer of graphene. A major question is represented by the phonon-phonon collision operator involving in general a three particles scattering mechanism. To model the phonon-phonon interactions a relaxation time approximation is employed. This requires the introduction of a local equilibrium phonon temperature whose definition is still a matter of debate for a general non equilibrium situation. Two different approaches are presented and discussed.

In the contribution *Monte Carlo Simulation of Electron-electron Interactions in Bulk Silicon* by G. Indalecio and H. Kosina a novel Monte Carlo (MC) algorithm to study carrier transport in semiconductors in the presence of electron-electron scattering (EES) is developed. It is well known that the Boltzmann scattering operator for EES is nonlinear in the single-particle distribution function but in terms of the pair distribution function the scattering operator is linear. A kinetic equation for the pair distribution function is formulated. Assuming a spatially homogeneous system a two-particle MC algorithm for the stationary problem and an ensemble MC algorithm for the transient problem are formulated. Both algorithms were implemented and tested for bulk silicon.

Since quantum transport is daunting to numerically face, often macroscopic models are proposed. In *Semi-classical and Quantum Hydrodynamic Modeling of Electron Transport in Graphene* by L. Luca and V. Romano A semi-classical hydrodynamic model has been developed starting from the moment system associated with the Boltzmann equation and obtaining the closure relations with the Maximum Entropy Principle. In a second step the model previously developed has been extended to include quantum corrections.

A different approach is presented by O. Morandi in his contribution *Quantum Model for the Transport of Nearly Localized Particles* where a quantum model based on the Gaussian-Hermite expansion of the wave function of a system of n particles is proposed. The dynamics is described by trajectories in a configuration space. The method is designed to provide some corrections to the classical motion of nearly localized particles. An application to the motion of a nearly localized particle in a 2D confining structure is presented.

Instead a full quantum approach is investigated in *Wigner Monte Carlo Simulation of a Double Potential Barrier*, author O. Muscato, where the Wigner transport equation is solved stochastically by Monte Carlo simulations, based on the generation and annihilation of particles. This creation mechanism has been recently understood in terms of the Markov jump process, producing new stochastic algorithms. One of this is used to investigate the quantum transport through a double potential barrier.

Field effects transistors, where the active region is constituted by a single layer of graphene, is one of the present day challenging problem in the design of electron devices. In the contribution *Simulation of Graphene Field Effect Transistors* by G. Nastasi and V. Romano field effects transistors, where the active area is made of a single layer of graphene, are simulated and the characteristic curves are shown. The current-voltage curves present a behaviour different from that of devices made of classical semiconductors, like Si or GaAs, because of the zero gap in monolayer graphene. The current is no longer a monotone function of the gate voltage but there exists an inversion gate voltage corresponding to which the type of majority carriers changes. A full two-dimensional simulation is presented. The model is based on a system of drift-diffusion equations for electrons and holes. A special treatment of the Poisson equation is adopted for taking into account the charge in the graphene sheet.

Modeling Mechanochemistry in Li-ion Batteries

Bilen Emek Abali

Abstract Mechanochemistry in Li-ion batteries involve interaction of ions migrating in a cell with mechanical stresses as well as electromagnetic fields. We aim at modeling this multiphysics in a battery cell by involving thermomechanics with the corresponding balance equations of mass, momentum, energy and electromagnetism with the aid of the MAXWELL equations. Before applying specific assumptions relevant for Li-ion batteries, we address developing a complete theory by using continuum mechanics and thermodynamics.

1 Introduction

Two differently charged electrodes cause an electric potential leading to flow of electric charge in a circuit—called an electric current. This basic idea is used in batteries, where a positively charged electrode and a negatively charged electrode are separated by a thin microperforated plastic sheet, which allows Li-ions to pass between electrodes. By connecting the battery to an electric circuit, we get power by discharging the battery. Electrons move along the circuit and at the same time, within the battery, ions move through the separator. Electrodes and separator are submerged in an organic solvent, acting as the electrolyte such that the ion transport is possible. One typical example is graphite C_6 and lithium cobalt oxide $LiCoO_2$, where during charging Li^+ ions (and also electrons) leave cathode

$$\mathrm{LiCoO_2} \rightleftarrows \mathrm{CoO_2} + \mathrm{Li}^+ + e^- \tag{1}$$

and move to the anode

$$\mathrm{C_6} + \mathrm{Li}^+ + e^- \rightleftarrows \mathrm{LiC_6} \tag{2}$$

B. E. Abali (✉)
Institute of Mechanics, MS 2, Technische Universität Berlin, Berlin, Germany
e-mail: bilenemek@abali.org
http://bilenemek.abali.org

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_8

by diffusing through the electrolyte (electrons moving through the circuit). During discharging (standard use of the battery) the process is such that the electrons (hence, ions) are transported from the anode to the cathode. Therefore, anode and cathode are also called negative and positive electrodes, respectively. This basic interpretation is intuitive; however, there are several assumptions taken for granted. First, the presented process is based on the equilibrium thermodynamics under the assumption that the amount of lithium insertion (intercalation) and extraction (deintercalation) are equal for a time period, which is actually too long for the underlying process. Intercalation is happening conceivably in a shorter time period such that we claim that forcing an intercalation due to an electric voltage difference is a non-equilibrium process. Second, charging and discharging are seen as a reversible process, which is wrong since we know that the battery degrades over time because of the charging and discharging cycle. Third, at least in classical theories in chemical kinetics, the production rate of constituents is assumed to be a constant without coupling to the underlying process (independent of the stress, temperature, phase transformation).

For the long and polemical history of thermodynamics, we refer to [11]. The modern theory of thermodynamics is started in [14–16] as known under the name of the *thermodynamics of irreversible processes*. One of the general approaches for thermodynamics of mixtures is given in [33]—also see [23, Chap. 7]—leading to mass, momentum, energy balances for each constituent. In simple words, electrolyte with ions is a mixture of ions, each modeled as a separate material, and electrolyte as another material. All materials have different amount evolving in each location of the battery cell. Depending on the process, the system can be simplified by assuming that the ion particles are small in number such that momentum and/or energy balance can be solved for the bulk instead of every single constituent leading to models called Class I, II, III. Such theories are actually not new; already in [40] a similar approach was proposed. However, they are still not popular among chemists. Instead, phase-field models are often used. 150 years ago, introduced by Fick in [19] for dilute solutions, the partial mass balance can be seen as a diffusion equation with a flux term given as a constitutive equation. The diffusivity or mobility is assumed to be a constant material parameter dictating the diffusion flux. There are various theories in mesoscale, so-called phase field models, for a review see the works cited in [4] as well as [37]. It is challenging to obtain the correct form for Fickean fluxes, i.e., the dependencies of the diffusion flux on the thermodynamic state is phenomenological and often arduous to justify. Even if the system is closed, the phase-field models suffer under numerical problems, sometimes called "mosaic instabilities," mostly because of non-symmetric diffusion coefficients [39]. In other words, the interaction of constituents are not equal; concretely, constituent α diffuses into another constituent β in a different rate than β into α. This fact is partly because of the system being in non-equilibrium and also partly because of the irreversible chemical reactions. Moreover, if one constituent is vanishing, the system fails to be well-defined anymore, leading to severe numerical convergence problems. Although the phase-field models are powerful, we will use a continuum mechanics based theory for motivating the governing equations and a thermodynamical theory in order to obtain the full coupling between different types of phenomena.

We consider in this work a start for a modeling in the continuum scale by describing the ion transport and chemical production by using the non-equilibrium thermodynamics by following [10]. In this manner, the general formulation is obtained and different simplifications are discussed by combining chemists' and continuum mechanicians' jargons. We emphasize that the same physico-chemical phenomena are studied with different notation by distinct groups such that the possible understanding is lagging in modeling the Li-ion batteries. Therefore, in this paper, we aim at building bridges between communities by presenting different notations and conventional names for the process in a single Li-ion battery cell. We follow [8] for names in chemistry and [41] for names in continuum mechanics. For modeling the mechanochemical processes alike in batteries, there is a growing attention in use of thermodynamics as in [6, 7, 9, 12, 13, 17, 20, 30, 32, 36, 42]. For a detailed review, we refer to [43]. The proposed strategy is based on the thermodynamics of irreversible processes as well as on the non-equilibrium thermodynamics. The main objective is to clearly demonstrate how to obtain the governing equations for the mechanochemical process in a Li-ion battery cell. We basically use the classical continuum mechanics notation and augment it by chemists' notation in order to demonstrate the differences as well as elucidate the exchange between scientific communities.

2 Modeling the Process in a Battery Cell

Consider a battery cell composed of electrodes, binder, polymer, and ions diffusing between as well as intercalating into the electrodes during charge and discharge. At any given point all constituents build the bulk. Basically, we have different constituents or species denoted by α and their corresponding mass and momentum equations need to be fulfilled. Technically, it is challenging to measure velocity of each constituent, $\boldsymbol{v}^\alpha$, accurately. Therefore, constitutive relations involving this quantity fail to be feasible. In order to overcome this restriction, we introduce a diffusion flux:

$$N_i^\alpha = \rho^\alpha (v_i^\alpha - v_i) \ , \tag{3}$$

with partial mass, ρ^α, and velocity, v_i^α, of each constituent as well as the bulk velocity, v_i, where all velocities are expressed in Cartesian coordinates. The physical meaning of this diffusion flux seems to be adequate as it involves a relative velocity, $\boldsymbol{v}^\alpha - \boldsymbol{v}$, of the constituent with respect to the bulk velocity. Then we aim at finding a constitutive equation relating the diffusion flux to other measurable quantities like concentration such that the constituents' velocities never occur in the formulation. In other words, we simplify the system by circumventing the latter equation by using a constitutive relation instead. Not always indicated, but the *partial mass* *density*, ρ^α in kg/m^3, is a very confusing quantity. It is not a physical variable; but it is introduced to allow a more convenient simplification in the theory as we will see in the following. The partial mass density, $\rho^\alpha = M^\alpha n^\alpha$, is given by the *molar mass*, M^α in kg/mol, and the *molar density*, n^α in mol/m^3, which is also called *molarity* as well as *number density*. Often, chemists call the molar density (number or amount) *concentration*. Molar mass of an atom is found in chemists' periodic table by the weight per 1 mol. It is important to notice that the partial mass density is a completely different concept than the mass density of the constituent. Mass density is the mass per volume of the constituent. However, the partial mass density is the mass of the constituent per the volume of the bulk. By using this concept, rarely called the skeleton approach, we obtain an extensive quantity such that the partial mass densities can be added. Consider a unit volume, V, of the bulk material of mass, m, composed of constituents denoted by the index α and having corresponding masses in this volume, m^α, we have

$$m = \sum_\alpha m^\alpha \Rightarrow \rho = \frac{m}{V} = \sum_\alpha \frac{m^\alpha}{V} = \sum_\alpha \rho^\alpha \ . \tag{4}$$

Another terminology is the concept of the mass fraction, Y^α, being simply the ratio of each mass of the constituent to the total mass of the bulk within the volume V such as we obtain

$$Y^\alpha = \frac{m^\alpha}{m} = \frac{\rho^\alpha}{\rho} \Rightarrow \sum_\alpha Y^\alpha = 1 \ . \tag{5}$$

The mass fraction is called (mass) *concentration* by mechanicians and also by chemists in polymer science. Using the same wording for different yet related terms is very risky and also omits to exchange between different scientific communities. We suggest a pretty simple remedy and never use a notation for the concentration; molar density is denoted by n^α and mass fraction is indicated by Y^α. By using the total (mol) number, n, within the unit volume V, an analogous quantity, *molar fraction* reads

$$\begin{gathered}
\frac{n^\alpha}{n} = X^\alpha \Rightarrow n X^\alpha M^\alpha = \rho^\alpha = \rho Y^\alpha \ , \quad X^\alpha = \frac{\rho Y^\alpha}{n M^\alpha} \ , \\
\sum_\beta \frac{Y^\beta}{M^\beta} = \sum_\beta \frac{\rho^\beta}{\rho M^\beta} = \sum_\beta \frac{n^\beta}{\rho} = \frac{n}{\rho} \ , \\
X^\alpha = \frac{Y^\alpha / M^\alpha}{\sum_\beta Y^\beta / M^\beta} .
\end{gathered} \tag{6}$$

We aim at calculating constituents' mass fraction, Y^α, as the molar fraction is determined by the latter relation. Starting with the balance of mass in the bulk as a conserved quantity, as usual, we suggest a mass balance for each constituent, where the partial mass density is not a conserved quantity leading to production terms for each constituent modeling a chemical reaction as in Eqs. (1), (2) with a reaction rate, k^α, depending on constituents as reactants (educts) or products,

$$\frac{\partial \rho}{\partial t} + \frac{\partial v_i \rho}{\partial x_i} = 0 \Rightarrow \frac{\partial \rho^\alpha}{\partial t} + \frac{\partial v_i^\alpha \rho^\alpha}{\partial x_i} = k^\alpha \ , \tag{7}$$

where and henceforth we employ EINSTEIN's summation convention over repeated Latin indices. A possible connection is generated between the balance of mass (within the bulk) and the balance of partial mass (for each constituent) by assuming

$$\sum_\alpha \rho^\alpha v_i^\alpha = \rho v_i \ , \ \ \sum_\alpha k^\alpha = 0 \ . \tag{8}$$

The former relation introduces the velocity of the bulk, $\boldsymbol{v}$, as a barycentric (center of mass) velocity. The latter relation furnishes that the sum of reactant masses are equal to the sum of product masses such that the total mass or the bulk mass remains constant throughout the chemical process. We circumvent to solve velocities of each constituent by inserting $\boldsymbol{N}^\alpha$ in kg/(m^2 s) from Eq. (3) and using Eq. (7)$_1$ as follows:

$$\begin{aligned}
\frac{\partial \rho^\alpha}{\partial t} + \frac{\partial}{\partial x_i}\Big(v_i \rho^\alpha + N_i^\alpha \Big) &= k^\alpha \ , \\
\frac{\partial \rho Y^\alpha}{\partial t} + \frac{\partial}{\partial x_i}\Big(v_i \rho Y^\alpha + N_i^\alpha \Big) &= k^\alpha \ , \\
\rho \frac{\partial Y^\alpha}{\partial t} + v_i \rho \frac{\partial Y^\alpha}{\partial x_i} + \frac{\partial N_i^\alpha}{\partial x_i} &= k^\alpha \ , \\
\rho \frac{\mathrm{d} Y^\alpha}{\mathrm{d} t} + \frac{\partial N_i^\alpha}{\partial x_i} &= k^\alpha \ ,
\end{aligned} \tag{9}$$

where the total derivative is applied

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial t} + v_i \frac{\partial}{\partial x_i} \ . \tag{10}$$

After introducing specific volume, $\mathrm{v} = \rho^{-1}$, we rewrite the latter,

$$\begin{aligned}
\rho \frac{\mathrm{d} Y^\alpha}{\mathrm{d} t} = \frac{\mathrm{d} \rho Y^\alpha}{\mathrm{d} t} - Y^\alpha \frac{\mathrm{d} \rho}{\mathrm{d} t} = \frac{\mathrm{d} M^\alpha n^\alpha}{\mathrm{d} t} - Y^\alpha \frac{\mathrm{d} \mathrm{v}^{-1}}{\mathrm{d} t} = \\
= M^\alpha \frac{\mathrm{d} n^\alpha}{\mathrm{d} t} + \frac{Y^\alpha}{\mathrm{v}^2} \frac{\mathrm{d} \mathrm{v}}{\mathrm{d} t} = M^\alpha \frac{\mathrm{d} n^\alpha}{\mathrm{d} t} + \frac{M^\alpha n^\alpha}{\mathrm{v}} \frac{\mathrm{d} \mathrm{v}}{\mathrm{d} t} \ , \\
\frac{\mathrm{d} n^\alpha}{\mathrm{d} t} + \frac{n^\alpha}{\mathrm{v}} \frac{\mathrm{d} \mathrm{v}}{\mathrm{d} t} + \frac{1}{M^\alpha} \frac{\partial N_i^\alpha}{\partial x_i} = \frac{k^\alpha}{M^\alpha} \ .
\end{aligned} \tag{11}$$

In the case of no chemical reaction, $k^\alpha = 0$, for a homogeneous and incompressible fluid with $\rho = \text{const}$ leading to $\mathrm{dv} = 0$; we obtain the so-called diffusion equation:

$$\frac{\mathrm{d}n^\alpha}{\mathrm{d}t} = -\frac{\partial j_i^\alpha}{\partial x_i} \ , \quad j_i^\alpha = \frac{N_i^\alpha}{M^\alpha} \ , \tag{12}$$

which is also named after Fick as phenomenologically proposed in [19]. We emphasize that chemists measure the diffusion flux $\boldsymbol{j}^\alpha$ in mol/(m^2 s).

In this setting, we aim at solving $\{\rho, Y^\alpha, \boldsymbol{v}\}$ with $n^\alpha = \rho Y^\alpha / M^\alpha$ by fulfilling balance of mass for the bulk, balances of mass for the constituents, and balance of momentum for the bulk,

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \frac{\partial v_i \rho}{\partial x_i} &= 0 \ , \\ \frac{\partial \rho Y^\alpha}{\partial t} + \frac{\partial}{\partial x_i}\Big(v_i \rho Y^\alpha + M^\alpha j_i^\alpha \Big) &= k^\alpha \ , \\ \frac{\partial \rho v_i}{\partial t} + \frac{\partial}{\partial x_j}\Big(v_j \rho v_i - \sigma_{ji} \Big) &= \mathcal{F}_i \ , \end{aligned} \tag{13}$$

respectively. Apart from the constitutive equations for CAUCHY's stress of the bulk, $\boldsymbol{\sigma}$, and for the diffusion flux of each constituent, $\boldsymbol{N}^\alpha$, subject to a chemical reaction given by the rate k^α; we need to define the interaction between mechanics and electromagnetism by defining the electromagnetic (ponderomotive) force density, $\mathcal{F}$. Especially the definition of this quantity is very challenging and there exists no consensus between the scientific community, see for example [5, 21, 31, 35]. We follow the method of derivation used in [27, Eq. (15)], [25, Chap. 1], [10, Chap. XIV], [22, Chap. 8], [29, Sect. 3.3] in the following and introduce a very general identity from the tensor calculus,

$$\frac{\partial \mathcal{G}_i}{\partial t} = \frac{\partial m_{ji}}{\partial x_j} - \mathcal{F}_i \ , \tag{14}$$

between the electromagnetic momentum density, $\mathcal{G}$, electromagnetic stress, $\boldsymbol{m}$, and the electromagnetic force density, $\mathcal{F}$. This relation is in analogy with the balance of momentum such that the names stress and momentum are justified by analyzing their units. If the electromagnetic momentum, $\mathcal{G}$, is defined, as a consequence of the latter relation, we can propose the electromagnetic stress and the electromagnetic force density.

The measurable electromagnetic fields are the electric field $\boldsymbol{E}$ and the magnetic flux (area density) $\boldsymbol{B}$ defined by solving two of four MAXWELL equations as follows:

$$E_i = -\frac{\partial \phi}{\partial x_i} - \frac{\partial A_i}{\partial t} \ , \quad B_i = \epsilon_{ijk} \frac{\partial A_k}{\partial x_j} \ , \tag{15}$$

where the electromagnetic potentials ϕ, $\boldsymbol{A}$ are sought after, ϵ_{ijk} is the LEVI-CIVITA symbol. In order to solve the scalar (electric) potential, ϕ, we use the balance of electric charge combined by one MAXWELL equation,

$$\frac{\partial \rho z}{\partial t} + \frac{\partial J_i}{\partial x_i} = 0 \ , \quad \rho z = \frac{\partial D_i}{\partial x_i} \ , \tag{16}$$

respectively, where the specific charge, z, is related to the total charge potential, $\boldsymbol{D}$, and the electric current, $\boldsymbol{J}$, of the total charge reads

$$J_i = J_i^{\text{fr.}} + \frac{\partial P_i}{\partial t} + \epsilon_{ijk} \frac{\partial \mathcal{M}_k}{\partial x_j} \ , \tag{17}$$

with the electric current of free charges, $\boldsymbol{J}^{\text{fr.}}$, electric polarization, $\boldsymbol{P}$, and magnetic polarization, $\mathcal{M}$, all to be given by constitutive equations.

In order to solve the vector (magnetic) potential, $\boldsymbol{A}$, we use the final MAXWELL equation augmented by the LORENZ gauge [28],

$$-\frac{\partial D_i}{\partial t} + \epsilon_{ijk} \frac{\partial H_k}{\partial x_j} = J_i \ , \quad \frac{\partial \phi}{\partial t} + \frac{1}{\varepsilon_0 \mu_0} \frac{\partial A_i}{\partial x_i} = 0 \ , \tag{18}$$

respectively, where the LORENZ gauge[1] is an appropriate choice for numerical solutions, see [2] for implementation and applications. The universal constants:

$$\begin{aligned} \epsilon_0 &= 8.85 \cdot 10^{-12}\, \text{A s/(V m)} \ , \\ \mu_0 &= 12.6 \cdot 10^{-7}\, \text{V s/(A m)} \ , \end{aligned} \tag{19}$$

and the MAXWELL–LORENTZ aether relations:

$$D_i = \varepsilon_0 E_i \ , \quad H_i = \frac{1}{\mu_0} B_i \ , \tag{20}$$

are used to combine the electromagnetic fields with the (total) charge potential, $\boldsymbol{D}$, and the (total) current potential, $\boldsymbol{H}$.

We remark herein the use of different notations, this fact is also a challenging hurdle in the electromagnetism community. Conventionally, the total charge (volume) density is denoted by $q = \rho z$ such that the specific charge is z in C/kg. We emphasize the adjective "total" and declare the total charge as free and bound charges together. Free charges move in macroscopic distances like valence electrons in a metal lattice or hopping ions in an electrolyte. Bound charges displace only in atomic distances causing a polarization. As free and bound charges move in different

[1] This gauge condition has been introduced by Lorenz but mostly used by Lorentz, we refer to [34] for historical remarks.

length scales, it is appropriate to separate related effects because of their motion. Total charge potential (it was called dielectric displacement), $\boldsymbol{D}$, incorporates both effects; analogously, both are consolidated in total current potential (also known as magnetic field strength), $\boldsymbol{H}$. In the case of a specific case, where the material is not polarized and bound charges vanish, charge potential, $\boldsymbol{D}$, current potential, $\boldsymbol{H}$, electric current, $\boldsymbol{J}$, consider only free charges and a popular notational choice is to use the symbols, $\boldsymbol{D}$, $\boldsymbol{H}$, $\boldsymbol{J}$, for free charge related phenomena instead of total charge, see for example [24]. Technically, this choice is allowed since for total charge and current potentials, there exist universal relations; but we suggest to use additional notation for total, free, and bound charge related quantities for a clear separation of concerns. For the application discussed herein, there is no big difference in both notations as we will restrict the attention to the unpolarized case.

We have briefly presented the governing equations modeling mechanochemistry in a battery cell. We aim at solving mass density, ρ, and velocity, $\boldsymbol{v}$, of the bulk as well as mass fractions, Y^α, of all constituents taking part in the chemical reaction called intercalation at the electrode. Additionally, we need to obtain electromagnetic fields, $\boldsymbol{E}$, $\boldsymbol{B}$, by solving electromagnetic potentials, ϕ, $\boldsymbol{A}$. For being able to solve the governing equations, we have to close them by defining $\boldsymbol{J}^\alpha$, k^α, $\boldsymbol{\sigma}$, $\boldsymbol{P}$, $\mathcal{M}$, $\boldsymbol{J}^{\text{fr.}}$ as constitutive equations depending on $\{\rho, Y^\alpha, \boldsymbol{v}, \phi, \boldsymbol{A}\}$.

3 Constitutive Theory for Mechanochemistry

Charge transport in a cell means that the bulk is a conductive material. Herein we consider only the transport within the electrolyte. Hence, we neglect polarization effects by setting $\boldsymbol{P} = 0$ and $\mathcal{M} = 0$. This simplification is easily justified as the bulk is conductive, and it allows us to model the interaction between mechanics and electromagnetism relatively easy. In the case of no polarization, we choose the POYNTING vector:

$$\mathcal{G}_i = (\boldsymbol{D} \times \boldsymbol{B})_i \ , \tag{21}$$

leading to the following electromagnetic stress and force,

$$\begin{aligned} m_{ji} &= -\frac{1}{2}\delta_{ji}(H_k B_k + D_k E_k) + H_i B_j + D_j E_i \ , \\ \mathcal{F}_i &= \rho z E_i + \epsilon_{ijk} J_j B_k \ , \end{aligned} \tag{22}$$

by using KRONECKER delta, δ_{ij}, with these definitions, $\boldsymbol{m}$ is called MAXWELL stress and $\mathcal{F}$ is known as the LORENTZ force (volume density). This choice is accepted in the scientific literature in the case of no polarization, there are differing suggestions how to involve polarization, we refer to [18, 38] for a brief explanation of consequences of these different choices. Specifically, for an interaction between electrolyte and electrodes by involving polarization, we refer to [26]. By following [3] we emphasize that different choices are perfectly appropriate.

We follow [1, Chap. 3] for deriving the constitutive equations by using principle of thermodynamics. Herein we involve chemistry by using the approach as in [10]. The first assumption is the axiomatic start with the relation for the internal energy called the GIBBS equation:

$$\mathrm{d}u = T\,\mathrm{d}\eta - p\,\mathrm{d}\mathrm{v} + \sum_\alpha \mathrm{v}\mu^\alpha\,\mathrm{d}n^\alpha\ , \tag{23}$$

under the assumption that the specific internal energy, u, is associated with the reversible changes in the system as composed of the reversible part of the heat flux, which is given by the specific entropy, η, multiplied by the absolute temperature, T, chemical potential, μ^α, causing a reversible change in constituents' amount, n^α. Pressure, p in Pa$\hat{=}$N/m^2, of the bulk is assumed to be the only part in the reversible stress, which is the usual case in viscous fluids. The use of the specific volume, $\mathrm{v} = \rho^{-1}$ in m^3/kg, is critical such that the chemical potential, μ^α, reads in J/mol as measured by chemists. From the latter relation we observe that the internal energy, $u = u(\eta, \mathrm{v}, n^\alpha)$, depends on entropy, specific volume, and number density. In an experiment, it is easier to control temperature instead of the entropy as well as in mixtures it is only possible to steer pressure in place of the specific volume. Hence, often the GIBBS free energy is used,

$$\begin{gathered} g = u - T\eta + p\mathrm{v} \Rightarrow\ \mathrm{d}g = -\eta\,\mathrm{d}T + \mathrm{v}\,\mathrm{d}p + \sum_\alpha \mathrm{v}\mu^\alpha\,\mathrm{d}n^\alpha\ , \\ \frac{\partial g}{\partial T} = -\eta\ ,\quad \frac{\partial g}{\partial p} = \mathrm{v}\ ,\quad \frac{\partial g}{\partial n^\alpha} = \mathrm{v}\mu^\alpha\ . \end{gathered} \tag{24}$$

By measuring the free energy depending on $\{T, p, n^\alpha\}$, we obtain the constitutive relations for the entropy, specific volume, and the chemical potential.

Now we use the balance of internal energy as derived in [1, Sect. 3.3] for non-polarized material,

$$\rho\frac{\mathrm{d}u}{\mathrm{d}t} + \frac{\partial q_i}{\partial x_i} - \rho r = \sigma_{ji}\frac{\partial v_i}{\partial x_j} + \mathcal{J}_i\mathcal{E}_i\ , \tag{25}$$

with the heat flux, q_i, to be defined, the given supply term (external heating), r, the electric field measured on the deforming body, $\mathcal{E} = \boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}$, and the (total) electric current measured with respect to the moving mass, $\mathcal{J} = \boldsymbol{J} - \rho z \boldsymbol{v}$. By inserting

Eq. (23) into Eq. (25), after a straightforward rewriting by inserting Eq. (11), we obtain

$$\begin{aligned}
&\rho T \frac{\mathrm{d}\eta}{\mathrm{d}t} - \rho p \frac{\mathrm{d}\mathrm{v}}{\mathrm{d}t} + \sum_\alpha \mu^\alpha \frac{\mathrm{d}n^\alpha}{\mathrm{d}t} + \frac{\partial q_i}{\partial x_i} - \rho r = \sigma_{ji} \frac{\partial v_i}{\partial x_j} + \mathcal{J}_i \mathcal{E}_i \ , \\
&\rho T \frac{\mathrm{d}\eta}{\mathrm{d}t} - \Big(\rho p + \sum_\alpha \mu^\alpha \rho n^\alpha \Big) \frac{\mathrm{d}\mathrm{v}}{\mathrm{d}t} + \frac{\partial}{\partial x_i} \Big(q_i - \sum_\alpha \mu^\alpha j_i^\alpha \Big) + \sum_\alpha \frac{\partial \mu^\alpha}{\partial x_i} j_i^\alpha + \\
&+ \sum_\alpha \frac{\mu^\alpha k^\alpha}{M^\alpha} - \rho r = \sigma_{ji} \frac{\partial v_i}{\partial x_j} + \mathcal{J}_i \mathcal{E}_i \ ,
\end{aligned} \tag{26}$$

With the aid of the mass balance in the following form:

$$\begin{aligned}
&\frac{\partial \rho}{\partial t} + \frac{\partial \rho v_i}{\partial x_i} = \frac{\mathrm{d}\rho}{\mathrm{d}t} + \rho \frac{\partial v_i}{\partial x_i} = 0 \ , \quad \frac{\mathrm{d}\mathrm{v}^{-1}}{\mathrm{d}t} = -\mathrm{v}^{-1} \frac{\partial v_i}{\partial x_i} \ , \\
&-\mathrm{v}^{-2} \frac{\mathrm{d}\mathrm{v}}{\mathrm{d}t} = -\mathrm{v}^{-1} \frac{\partial v_i}{\partial x_i} \ , \quad \frac{\mathrm{d}\mathrm{v}}{\mathrm{d}t} = \mathrm{v} \frac{\partial v_i}{\partial x_i} = \frac{1}{\rho} \delta_{ij} \frac{\partial v_i}{\partial x_j} \ ,
\end{aligned} \tag{27}$$

we obtain the balance of entropy:

$$\begin{aligned}
&\rho \frac{\mathrm{d}\eta}{\mathrm{d}t} + \frac{1}{T} \frac{\partial}{\partial x_i} \Big(q_i - \sum_\alpha \mu^\alpha j_i^\alpha \Big) + \sum_\alpha \frac{\mu^\alpha k^\alpha}{T M^\alpha} - \rho \frac{r}{T} = \\
&= \frac{1}{T} \Big(\sigma_{ji} + \big(p + \sum_\alpha \mu^\alpha n^\alpha \big) \delta_{ji} \Big) \frac{\partial v_i}{\partial x_j} + \frac{\mathcal{J}_i \mathcal{E}_i}{T} - \sum_\alpha \frac{\partial \mu^\alpha}{\partial x_i} \frac{j_i^\alpha}{T} \ , \\
&\rho \frac{\mathrm{d}\eta}{\mathrm{d}t} + \frac{\partial \Phi_i}{\partial x_i} - \rho \frac{r}{T} = \Sigma \ ,
\end{aligned} \tag{28}$$

with the entropy flux,

$$\Phi_i = \frac{1}{T} \Big(q_i - \sum_\alpha \mu^\alpha j_i^\alpha \Big) \ , \tag{29}$$

and the entropy production,

$$\Sigma = - \sum_\alpha \frac{\mu^\alpha k^\alpha}{T M^\alpha} - \frac{\Phi_i}{T} \frac{\partial T}{\partial x_i} + \frac{\tau_{ji}}{T} \frac{\partial v_i}{\partial x_j} + \frac{\mathcal{J}_i \mathcal{E}_i}{T} - \sum_\alpha \frac{\partial \mu^\alpha}{\partial x_i} \frac{j_i^\alpha}{T} \tag{30}$$

where the viscous part of the stress is given by $\tau_{ji} = \sigma_{ji} + \big(p + \sum_\alpha \mu^\alpha n^\alpha \big) \delta_{ji}$. The entropy balance is the governing equation for calculating the temperature. Now by using the second law of thermodynamics, $\Sigma \geq 0$, as well as the Curie principle, we can propose that the same rank tensors depend to each other in order to simplify and derive the constitutive relations for k^α, $\boldsymbol{\Phi}$, $\boldsymbol{\tau}$, $\mathcal{J}$, and $\boldsymbol{j}^\alpha$. There are several coupling terms inherently obtained by using laws of thermodynamics.

4 Conclusion

In a battery cell, the governing equations are obtained for thermomechanics in Eqs. (13), (28) and for electromagnetism in Eq. (18). Their necessary constitutive equations are briefly shown how to be derived by using thermodynamical principles. We have mainly considered a notation conventional to continuum mechanicians as well as chemists, since the formulation is acquired by mechanicians and the corresponding measurements for the constitutive equations are realized by chemists.

Acknowledgments B. E. Abali's work is partly funded by a grant from the Daimler and Benz Foundation.

References

1. Abali, B.E.: Computational Reality, Solving Nonlinear and Coupled Problems in Continuum Mechanics. Advanced Structured Materials, vol. 55. Springer Nature, Singapore (2017)
2. Abali, B.E., Reich, F.A.: Thermodynamically consistent derivation and computation of electro-thermo-mechanical systems for solid bodies. Comput. Methods Appl. Mech. Eng. **319**, 567–595 (2017)
3. Barnett, S.M.: Resolution of the Abraham–Minkowski dilemma. Phys. Rev. Lett. **104**(7), 070401 (2010)
4. Bazant, M.Z.: Theory of chemical kinetics and charge transfer based on nonequilibrium thermodynamics. Acc. Chem. Res. **46**(5), 1144–1160 (2013)
5. Bethune-Waddell, M., Chau, K.J.: Simulations of radiation pressure experiments narrow down the energy and momentum of light in matter. Rep. Prog. Phys. **78**(12), 122401 (2015)
6. Bothe, D., Dreyer, W.: Continuum thermodynamics of chemically reacting fluid mixtures. Acta Mech. **226**(6), 1757–1805 (2015)
7. Breitkopf, C., Swider-Lyons, K.: Springer Handbook of Electrochemical Energy. Springer, Berlin (2016)
8. Cohen, E.R., Cvitas, T., Frey, J.G., Holmström, B., Kuchitsu, K., Marquardt, R., Mills, I., Pavese, F., Quack, M., Stohner, J., Strauss, H.L., Takami, M., Thor, A.J.: Quantities, Units and Symbols in Physical Chemistry, IUPAC Green Book, 2nd Printing, 3rd edn. IUPAC & RSC Publishing, Cambridge (2008)
9. Datta, R., Vilekar, S.A.: The continuum mechanical theory of multicomponent diffusion in fluid mixtures. Chem. Eng. Sci. **65**(22), 5976–5989 (2010)
10. de Groot, S.R., Mazur, P.: North-Holland, Amsterdam; Wiley, New York (1962)
11. Dreyer, W., Müller, W.H., Weiss, W.: Tales of thermodynamics and obscure applications of the second law. Contin. Mech. Thermodyn. **12**(3), 151–184 (2000)
12. Dreyer, W., Jamnik, J., Guhlke, C., Huth, R., Movskon, J., Gabervsvcek, M.: The thermodynamic origin of hysteresis in insertion batteries. Nat. Mater. **9**(5), 448 (2010)
13. Dreyer, W., Guhlke, C., Müller, R.: A new perspective on the electron transfer: recovering the Butler–Volmer equation in non-equilibrium thermodynamics. Phys. Chem. Chem. Phys. **18**(36), 24966–24983 (2016)
14. Eckart, C.: The thermodynamics of irreversible processes. I. The simple fluid. Phys. Rev. **58**, 267–269 (1940)
15. Eckart, C.: The thermodynamics of irreversible processes. II. Fluid mixtures. Phys. Rev. **58**(3), 269 (1940)

16. Eckart, C.: The thermodynamics of irreversible processes. III. relativistic theory of the simple fluid. Phys. Rev. **58**(10), 919 (1940)
17. Fang, R., Farah, P., Popp, A., Wall, W.A.: A monolithic, mortar-based interface coupling and solution scheme for finite element simulations of lithium-ion cells. Int. J. Numer. Methods Eng. **114**(13), 1411–1437 (2018)
18. Felix, A.R.: Coupling of continuum mechanics and electrodynamics: an investigation of electromagnetic force models by means of experiments and selected problems. Ph.D. thesis, TU Berlin (2017)
19. Fick, A.: Über diffusion. Ann. Phys. **170**(1), 59–86 (1855)
20. Fuhrmann, J.: Comparison and numerical treatment of generalised Nernst–Planck models. Comput. Phys. Commun. **196**, 166–178 (2015)
21. Griffiths, D.J.: Resource letter em-1: electromagnetic momentum. Am. J. Phys. **80**(1), 7–18 (2012)
22. Griffiths, D.J., College, R.: Introduction to Electrodynamics, vol. 3. Prentice Hall, Upper Saddle River (1999)
23. Hutter, K., Jöhnk, K.: Continuum methods of physical modeling: continuum mechanics, dimensional analysis, turbulence. Springer Science & Business Media, New York (2013)
24. Jackson, J.D.: Classical Electrodynamics. Wiley, New York (1999)
25. Jones, D.S.: The Theory of Electromagnetism. Pergamon, New York (1964)
26. Landstorfer, M., Guhlke, C., Dreyer, W.: Theory and structure of the metal-electrolyte interface incorporating adsorption and solvation effects. Electrochim. Acta **201**, 187–219 (2016)
27. Lorentz, H.A.: Versuch einer Theorie der elektrischen und optischen Erscheinungen in bewegten Körpern. Zittungsverlagen Akad. van Wettenschappen 1, 74, 26 Nov 1892. Proc. Acad. Sci. (Amst.) (Engl. version) **6**, 809 (1904)
28. Lorenz, L.: On the identity of the vibrations of light with electrical currents. Lond. Edinb. Dublin Philos. Mag. J. Sci. **34**(230), 287–301 (1867)
29. Low, F.E.: Classical Field Theory: Electromagnetism and Gravitation. Wiley, New York (2004)
30. Maier, J.: Thermodynamics of electrochemical lithium storage. Angew. Chem. Int. Ed. **52**(19), 4998–5026 (2013)
31. Mansuripur, M.: Resolution of the Abraham–Minkowski controversy. Opt. Commun. **283**(10), 1997–2005 (2010)
32. Morozov, A., Khakalo, S., Balobanov, V., Freidin, A.B., Müller, W.H., Niiranen, J.: Modeling chemical reaction front propagation by using an isogeometric analysis. Tech. Mech. **38**(1), 73–90 (2018)
33. Müller, I.: A thermodynamic theory of mixtures of fluids. Arch. Ration. Mech. Anal. **28**(1), 1–39 (1968)
34. Nevels, R., Shin, C.S.: Lorenz, Lorentz, and the gauge. IEEE Antennas Propag. Mag. **43**(3), 70–71 (2001)
35. Obukhov, Y.N.: Electromagnetic energy and momentum in moving media. Ann. Phys. **17**(9-10), 830–851 (2008)
36. Poluektov, M., Freidin, A.B., Figiel, Ł.: Modelling stress-affected chemical reactions in non-linear viscoelastic solids with application to lithiation reaction in spherical Si particles. Int. J. Eng. Sci. **128**, 44–62 (2018)
37. Ramadesigan, V., Northrop, P.W.C., De, S., Santhanagopalan, S., Braatz, R.D., Subramanian, V.R.: Modeling and simulation of lithium-ion batteries from a systems engineering perspective. J. Electrochem. Soc. **159**(3), R31–R45 (2012)
38. Reich, F.A., Rickert, W., Müller, W.H.: An investigation into electromagnetic force models: differences in global and local effects demonstrated by selected problems. Contin. Mech. Thermodyn. **30**(2), 233–266 (2018)
39. Smith, R.B., Bazant, M.Z.: Multiphase porous electrode theory. J. Electrochem. Soc. **164**(11), E3291–E3310 (2017)
40. Stefan, J.: Über das Gleichgewicht und die Bewegung, insbesondere die Diffusion von Gasgemengen. Sitzber. Akad. Wiss. Wien **63**, 63–124 (1871)

41. Truesdell, C., Toupin, R.A.: The classical field theories. In: Encyclopedia of Physics, volume III/1, Principles of Classical Mechanics and Field Theory, pp. 226–790. Springer, Berlin (1960)
42. Weinberg, K., Werner, M., Anders, D.: A chemo-mechanical model of diffusion in reactive systems. Entropy **20**(2), 140 (2018)
43. Zhao, Y., Stein, P., Bai, Y., Al-Siraj, M., Yang, Y., Xu, B.X.: A review on modeling of electro-chemo-mechanics in lithium-ion batteries. J. Power Sources **413**, 259–283 (2019)

Automatic Extraction of Transport Model Parameters of an Organic Semiconductor Material

Pasquale Claudio Africa, Dario A. Natali, Mario Caironi, and Carlo de Falco

Abstract In Africa et al. (Sci Rep 7(1):3803, 2017) a step-by-step procedure was presented that enables to determine the density of states width, the carrier mobility and the injection barrier height of an OTFT structure by fitting data from simple measurements to suitable numerical simulations. At each step of the procedure only one parameter value is determined, thus highly simplifying the fitting procedure and enhancing its robustness. In this study we apply such procedure to p-type organic polymers. A very satisfactory fitting of experimental measurements is obtained, and physically meaningful values for the aforementioned parameters are extracted thus further confirming the soundness of the parameter extraction method.

1 Introduction

The relatively easy and inexpensive processing techniques for organic semiconductors (e.g., they can be deposited by means of printing techniques adapted from graphical arts (ink-jet, screen printing, spray coating, flexography to cite but a few [1]) make them suitable candidates for the development of large-area, low cost, flexible electronics [2]). Their electronic performance has been constantly improving leading to devices which some times even outperform those based on amorphous silicon [3]. Many fundamental physical questions, though,

P. C. Africa · C. de Falco (✉)
MOX Modelling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
e-mail: pasqualeclaudio.africa@polimi.it; carlo.defalco@polimi.it

D. A. Natali
Dipartimento di Elettronica, Informazione, Bioingegneria, Politecnico di Milano, Milano, Italy

Center for Nano Science and Technology @PoliMi, Istituto Italiano di Tecnologia, Milano, Italy
e-mail: dario.natali@polimi.it

M. Caironi
Center for Nano Science and Technology @PoliMi, Istituto Italiano di Tecnologia, Milano, Italy
e-mail: francesco.maddalena@iit.it; mario.caironi@iit.it

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_9

are still debated and there is a strong need for simple yet reliable approaches for extracting physical parameters from experimental measurements [4]. In [5] it was shown that, by fitting Capacitance-Voltage (CV) measurements of Metal-Insulator-Semiconductor (MIS) capacitors, it is possible to extract the width of the Density of States (DOS) exploiting the sensitivity of CV curves to the degree of disorder. Performing measurements on MIS capacitors at low frequency, quasi-equilibrium is ensured so that simulations can be performed in the static regime and phenomena specifically related to carrier transport are negligible for fitting experimental measurements; as a result the DOS extraction is disentangled from carrier transport properties, which makes the fitting procedure substantially simpler and more robust. If the DOS is assumed to be Gaussian, the carrier mobility can be predicted in the framework of the Extended Gaussian Disorder Model (EGDM)[6], and used to successfully fit the transfer characteristic curves of OTFTs in the linear regime. The DOS width extraction requires the accurate knowledge of the device geometrical dimensions, of the insulator and semiconductor permittivities, of the total density of available states and of the metal/semiconductor injection barrier (Φ_B) between the bottom metal and the semiconductor. This latter parameter is the one that is affected by highest level of uncertainty. Actually, metal/semiconductor interfaces are still a subject of debate in the scientific community [7]; due to the various phenomena which may be involved the prediction of Φ_B is a hard task, and its measurement requires expensive dedicated equipment. The uncertainty in Φ_B results in an uncertainty in the extracted value of the DOS width, as shown in [8]: for each given value of Φ_B a different value for the DOS width results from fitting. As reported in Sect. 5, the uncertainty is not negligible indeed: by varying Φ_B from 2 eV down to 0.25 eV, the DOS width reduces from about 9 k_BT down to about 2 k_BT. In [5] it was demonstrated that this uncertainty can be drastically reduced by cooperatively exploiting MIS CV curves, MIS Capacitance-Frequency (CF) curves and OTFT transfer characteristic curves in the linear regime. To this end, as discussed in [8] the simulation domain needs to cover out-of-equilibrium conditions in the framework of the Drift-Diffusion (DD) model. This allows to simulate the whole CF curve of the MIS capacitor. In addition, in the modeling and fitting of OTFT transfer characteristic curves the contact resistance needs to be accounted for in the context of the current crowding regime [9, 10].

In [8] the simultaneous extraction procedure was successfully applied to Poly{[N,N′-bis(2-octyldodecyl)-naphthalene-1,4,5,8-bis(dicarboximide)-2,6-diyl]-alt-5,5′-(2,2′-bithiophene)} (P(NDI2OD-T2)), a printable, prototypical n-type polymer. Here we adapt the model to p-type materials then we assess its accuracy when applied to Poly(2,5-bis(3-tetradecylthiophen-2-yl)thieno[3,2-b]thiophene) (PBTTT) based devices.

To model charge transport in transient regime, we employ the Drift-Diffusion model, which is described in Sect. 2 [11–13]. Transient simulations are used to compute the voltage and frequency dependence of the small-signal capacitance of the MIS capacitor. The most notable feature of the DD model described in this contribution are: (1) the charge injection boundary condition at the metal/semiconductor interface, and (2) the dependence of the mobility and of the diffusion coefficient on the DOS width.

While useful for computing the capacitance over a wide range of frequencies, the full DD model turns out to be of too high complexity and of insufficient numerical accuracy for efficiently fitting measured low-frequency CV curves. For this reason a modified version of the Non-linear Poisson (NLP) model is derived in Sect. 3, which includes a more accurate description of the contact injection barrier with respect to previous work [5], and is therefore fully consistent with the zero-frequency limit of the complete DD model.

The latter extended NLP model naturally describes the effect of the deviation from Einstein's relation but, as it is derived for the quasi-static regime, it does not require to model the mobility coefficient.

The models used for computing the transfer characteristics of the OTFT device are object of discussion in Sect. 4.

2 The Transient Drift-Diffusion Model

The setting for numerically simulating the MIS, shown in Fig. 1, consists of a one-dimensional schematization of the device along normal direction z to the semiconductor/insulator interface [5]. Denote by Ω_{sc}, Ω_{ins} the semiconductor and insulator regions respectively (such that $\Omega = \Omega_{sc} \cup \Omega_{ins}$ is the whole computational domain) and by T the simulated timespan.

The DD model consists of the following system of partial differential equations

$$-\frac{\partial}{\partial z}\left(\varepsilon \frac{\partial \varphi}{\partial z}\right) - \rho = 0 \quad \text{in } \Omega \times T, \tag{1}$$

$$\frac{\partial p}{\partial t} + \frac{1}{\mathrm{q}}\frac{\partial J_p}{\partial z} = 0, \quad \text{in } \Omega_{sc} \times T, \tag{2}$$

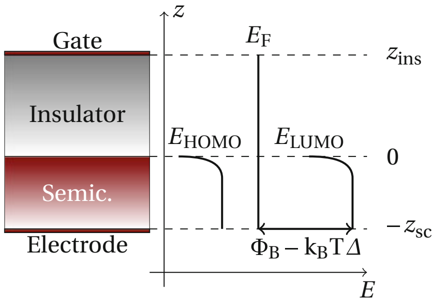


Fig. 1 One-dimensional schematic of the MIS capacitor used for the analysis and related energy levels

for the electrostatic potential φ and the hole density p, where ε is the electrical permittivity, ρ the charge density, q is the quantum of charge and J_p the hole current density. Neglecting trapped charges and dopant ion density, $\rho = +\mathrm{q}p$ in $\varOmega_{\mathrm{sc}}$ while, $\rho = 0$ in insulating regions.

The current density consists of drift and diffusion contributions

$$J_p = -\mathrm{q}\left(D_p \frac{\partial p}{\partial z} + \mu_p p \frac{\partial \varphi}{\partial z}\right), \tag{3}$$

D_p and μ_p denoting the diffusion and mobility coefficients respectively.

In the following we will assume a Gaussian shape for the DOS[5, 14–16] which may be expressed as

$$P(E, E_{\mathrm{HOMO}}) = \frac{N_0}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(E - E_{\mathrm{HOMO}})^2}{2\sigma^2}\right], \tag{4}$$

where N_0 denotes the total density of hopping sites, E_{HOMO} the Highest Occupied Molecular Orbital (HOMO) energy level and σ the DOS width. Having assumed the ansatz (4) for the DOS, it is possible to express the mobility coefficient according to the EGDM [17]

$$\mu_p = \mu_{0,p}\, g_1(p)\, g_2(\mathcal{E}), \tag{5}$$

where $\mu_{0,p}$ is the low-field and low-density mobility and the two enhancement factors g_1 and g_2 account respectively for the dependence on the carrier density p and on the electric field $\mathcal{E}$. The analytical expression for the enhancement factors is given in [5, 8, 17].

Charge injection/extraction at the metal/semiconductor interface is modeled by imposing that carrier density at the contact relaxes with finite velocity v_p to an equilibrium value p_0 which depends on the intensity and direction of the normal electric field at the contact

$$J_p = \mathrm{q}\, v_p \cdot (p - p_0)\,.$$

Following [18] we adopt models for v_n and n_0 that result in a variant of the well known injection model developed by Scott and Malliaras[11, 19]. Further details on the boundary conditions imposed on the DD system are collected in Sect. 2.2.

2.1 The Modified Einstein Relation

The diffusion and mobility coefficients D_p and μ_p in (3) are related via the generalized Einstein relation[17]:

$$D_p = g_3(p(E_{\mathrm{HOMO}}, E_{\mathrm{F}}))\, \frac{\mathrm{k_B T}}{\mathrm{q}}\, \mu_p.$$

Assuming Fermi–Dirac statistics for the occupation probability of electron energy states, the electron density may be expressed as

$$p(E_{\text{HOMO}}, E_{\text{F}}) = \int_{-\infty}^{+\infty} P(E, E_{\text{HOMO}}) \left[1 - \frac{1}{1 + \exp\left(\frac{E - E_{\text{F}}}{\text{k}_{\text{B}}\text{T}}\right)}\right] \text{d}E, \tag{6}$$

where E_{F} denotes the Fermi level, k_{B} the Boltzmann constant and T the temperature, and the dimensionless diffusion enhancement factor g_3 is given by

$$g_3(p) = \left(\text{k}_{\text{B}}\text{T}\,\frac{\partial p}{\partial E_{\text{F}}}\right)^{-1} p.$$

The partial derivative $\partial p / \partial E_{\text{F}}$ can be computed by substituting Eq. (4) into (6) and by applying a suitable change of the integration variable, yielding

$$p(E_{\text{HOMO}}, E_{\text{F}}) = \frac{N_0}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-\eta^2} \left[1 + \exp\left(\frac{\sqrt{2}\sigma\eta - E_{\text{HOMO}} + E_{\text{F}}}{\text{k}_{\text{B}}\text{T}}\right)\right]^{-1} \text{d}\eta.$$

2.2 Boundary Conditions for the Drift-Diffusion Equations

Let z_{sc} and z_{ins} be the thickness of the semiconductor and insulator layer respectively (so that $\Omega_{sc} = \{z : -z_{\text{sc}} \le z \le 0\}$ and $\Omega_{ins} = \{z : 0 < z \le z_{\text{ins}}\}$). The values of the electric potential at the ends of the computational domain are given by

$$\begin{aligned} \varphi|_{z=-z_{\text{sc}}} &= -\Phi_B/\text{q} \\ \varphi|_{z=z_{\text{ins}}} &= V_{\text{g}} + V_{\text{shift}}, \end{aligned}$$

where Φ_B is the zero-field value of the Schottky barrier height and V_{shift} is a model parameter accounting for effects such as permanent dipoles, fixed charge in dielectrics or metal work function mismatch [5]. At the metal/semiconductor interface charge injection/extraction is represented by the following Robin boundary condition

$$J_p\big|_{z=-z_{\text{sc}}} = \text{q}\, v_p \cdot \left(p|_{z=-z_{\text{sc}}} - p_0\right), \tag{7}$$

where p_0 is the equilibrium charge density, expressed as

$$p_0 = \frac{N_0}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-\eta^2} \left[1 + \exp\left(\frac{\sqrt{2}\sigma\eta + \Phi_B}{\text{k}_{\text{B}}\text{T}} - \Delta\right)\right]^{-1} \text{d}\eta.$$

Following the description in [11], the coefficient Δ accounts for Schottky barrier lowering/increase depending on the field at the electrode

$$\Delta = \begin{cases} \sqrt{f}, & \text{if } f \geq 0 \text{ (carrier injection)}, \\ f/4, & \text{if } f < 0 \text{ (carrier extraction)}, \end{cases}$$

where $f = \mathrm{q}\mathcal{E}r_\mathrm{c}/\left(\mathrm{k_B T}\right)$ is the reduced inward electric field and $r_\mathrm{c} = \mathrm{q}^2/\left(4\pi\varepsilon_\mathrm{sc}\mathrm{k_B T}\right)$ the Coulomb radius. The recombination velocity v_p in the injection regime is given by

$$v_p(f) = \frac{4\pi\varepsilon(\mathrm{k_B T})^2\mu_{0,p}}{\mathrm{q}^3}\left(\frac{1}{\psi^2(f)} - f\right), \qquad \psi(f) = f^{-1} + f^{-\frac{1}{2}} - f^{-1}\left(1 + 2f^{\frac{1}{2}}\right)^{\frac{1}{2}},$$

while, in the carrier extraction regime ($f < 0$), $v_p(f) = v_p(0)$. Finally, at the semiconductor/insulator interface

$$J_p\big|_{z=0} = 0. \tag{8}$$

3 The Stationary Non-linear Poisson Model

The DD equations presented above completely describe the devices under study and could in principle be used in any operation regime (stationary, transient, AC, ...). Unfortunately, given the strict tolerances and the large number of simulation runs required by the parameter fitting algorithm described in [5], the DD model may lead to unaffordable computational costs. Fortunately in the simulation of CV curves in the static regime, the model complexity can be significantly reduced. Indeed, in the stationary regime, Eq. (8) implies that $J_n = 0$ everywhere and that the carrier density does not depend on time. As a result the Fermi potential E_F is constant in both space and time (and may be set to 0 without loss of generality). System (1)–(2) reduces to the following NLP equation

$$\begin{aligned}
&-\frac{\partial}{\partial z}\left(\varepsilon\frac{\partial\varphi}{\partial z}\right) - \mathrm{q}\ p(\varphi) = 0, \qquad \text{in } \Omega \\
&p(\varphi) = \frac{N_0}{\sqrt{\pi}}\int_{-\infty}^{+\infty} e^{-\eta^2}\left[1 + \exp\left(\frac{\sqrt{2}\sigma\eta + \mathrm{q}\varphi}{\mathrm{k_B T}} - \Delta\right)\right]^{-1} \mathrm{d}\eta, \\
&\varphi|_{z=-z_\mathrm{sc}} = -\Phi_B/\mathrm{q}, \\
&\varphi|_{z=-z_\mathrm{sc}} = V_\mathrm{g} + V_\mathrm{shift}.
\end{aligned}$$

This model extends the NLP equation presented in [5] in order to account for the Schottky barrier lowering due to charge injection phenomena. We also remark that the boundary condition at the metal/semiconductor interface is consistent with Eq. (7), i.e. $p|_{z=-z_{\mathrm{sc}}} = p_0$. The procedure for accurately deducing the device low-frequency capacitance from charge and potential profiles is described in Ref.[5, 20].

4 OTFT Currents in the Linear Operation Regime

Once the DOS width σ has been extracted by fitting static CV curves, the low-field, low-density mobility $\mu_{0,p}$ can be determined by computing OTFT transfer characteristics in the linear regime. The drain-to-source current is expressed as $I_{\mathrm{DS}}(V_{\mathrm{g}}) = V_{\mathrm{DS}}/R_{\mathrm{tot}}(V_{\mathrm{g}})$, where V_{DS} is the potential drop across the channel and $R_{\mathrm{tot}} = R_{\mathrm{ch}} + R_{\mathrm{C}}$ is the total device resistance, accounting for both the channel and the contact resistance contributions. The channel resistance R_{ch} is given by:

$$R_{\mathrm{ch}} = \left[\frac{W}{L} \int_{-z_{\mathrm{sc}}}^{0} \mathrm{q}\mu_p(z)p(z)\mathrm{d}z \right]^{-1}, \tag{9}$$

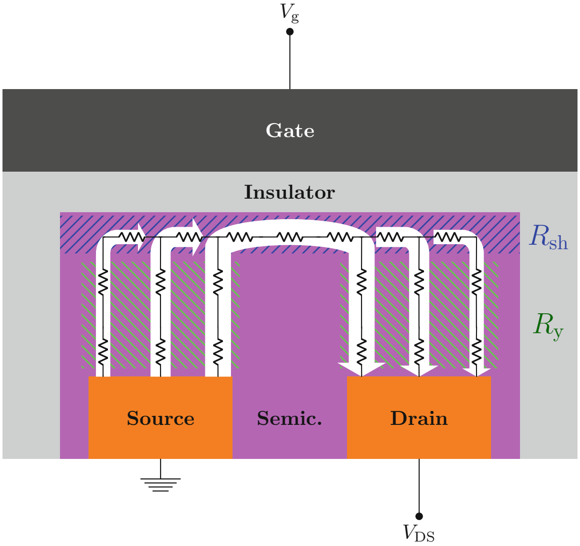
W and L being the channel width and length respectively. As in [8–10], R_{C} accounts for current crowding effects and is obtained considering contributions due to the resistance met by the current flow: (1) across the semiconductor thickness, from the source contact to the accumulated channel (R_{y}); and (2) along the accumulated channel (R_{sh}) in the overlap region between the source contact and the gate contact. It is modeled as $R_{\mathrm{C}} = R_{\mathrm{y}}/\left[WL_0 \tanh(L_{\mathrm{ov}}/L_0)\right]$, where L_{ov} is the overlap length between the gate and the source/drain electrodes, $L_0 = \sqrt{R_{\mathrm{y}}/R_{\mathrm{sh}}}$ and

$$R_{\mathrm{sh}} = \left[\int_{-z_{\mathrm{sc}}}^{0} \mathrm{q}\mu_p(z)p(z)\mathrm{d}z \right]^{-1}, \tag{10}$$

$$R_{\mathrm{y}} = \int_{-z_{\mathrm{sc}}}^{0} \left[\mathrm{q}\mu_p(z)p(z)\right]^{-1} \mathrm{d}z. \tag{11}$$

The integrand functions in Eqs. (9)–(11) are computed by simulating, as highlighted in Fig. 2, two different one-dimensional cross-sections along the z direction corresponding to the middle of the channel (R_{ch}, R_{sh}) and to the source contact (R_{y}) respectively. The mobility coefficient $\mu_p(z)$ is expressed through the EGDM model (5), where $\mathcal{E} = V_{\mathrm{DS}}/L$ is the drain-to-source field. The total current I_{DS} is known up to the multiplicative constant $\mu_{0,p}$, which is then extracted by fitting numerical and experimental Current–Voltage (IV) curves through a least-squares procedure. The fitting residual can finally be exploited to determine the best fitting value of Φ_B, as shown in Fig. 5.

Fig. 2 Sketch of the OTFT under current crowding effects

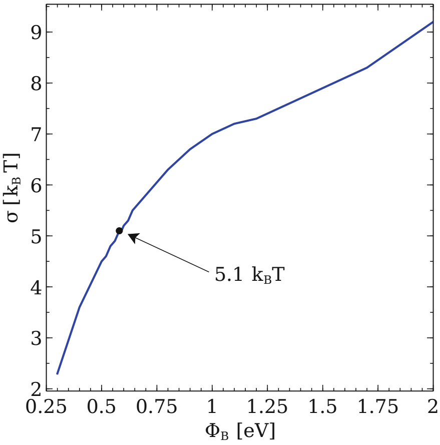


5 Results and Discussion

The parameter estimation procedure described in [8] was applied to the extraction of physical model parameters of the p-type PBTTT semiconductor, starting from experimental data on a MIS capacitor and an OTFT based on this polymer.

Experimental data taken from [5] would lead to estimate a nominal injection barrier $\Phi_B = 0.9$ eV. As described in [8], the uncertainty in determining a physically meaningful value of Φ_B affects other physical properties as obtained by fitting the MIS experimental CV curve, which leads to the dependence of σ and the contact resistance R_C on Φ_B shown in Figs. 3 and 4 respectively.

Fig. 3 Dependence of the fitted Gaussian DOS width σ on the injection barrier Φ_B. The dot on the curve identifies the σ value that simultaneously yields in the best fitting of OTFT transfer characteristic curves and of MIS capacitor CF curves



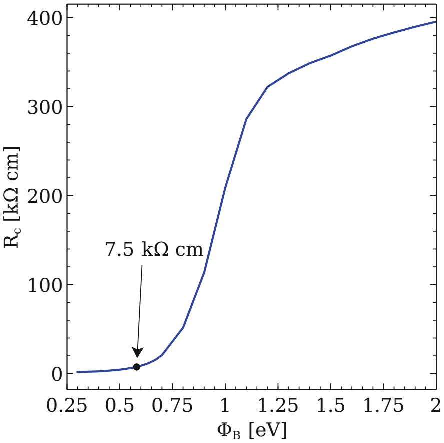


Fig. 4 Contact resistance R_C (at $V_{\text{gate}} = +35$ V) for different values of the injection barrier Φ_B. The dots on the curves identify the Φ_B value that simultaneously yields in the best fitting of OTFT transfer characteristic curves and of MIS capacitor CF curves

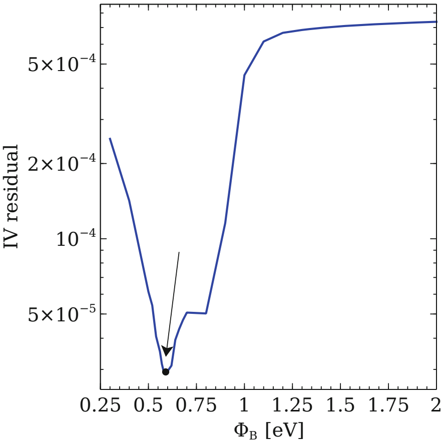


Fig. 5 Residual of the least-squares fit of the OTFT transfer characteristic curves with contact resistance effects taken into account, at different values of the injection barrier Φ_B

However, the uncertainty can be reduced by fitting IV transfer characteristics of the OTFT. The procedure yields to the residual in Fig. 5 that has a unique minimum, corresponding to the optimal barrier value $\Phi_B = 0.58$ eV and, accordingly, $\sigma = 5.1\,k_B T$ and $\mu_{0,p} = 7.84 \cdot 10^{-8}$ cm^2 V^{-1} s^{-1}. The resulting fitted CV and IV curves, compared to experimental data, are shown in Figs. 6 and 7, respectively.

In this case experimental CF curves are limited to a narrow range of frequencies. Therefore, the validation of the values extracted is not as clear as in the case of P(NDI2OD-T2) since both the nominal and the optimal set of parameters seem to provide a good agreement to experimental measurements, as reported in Fig. 8.

Fig. 6 CV curve and derivative, computed for the optimal barrier $\Phi_B = 0.58$ eV. Experimental CV shown for comparison

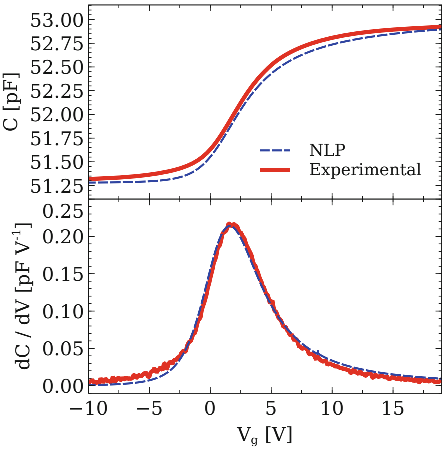
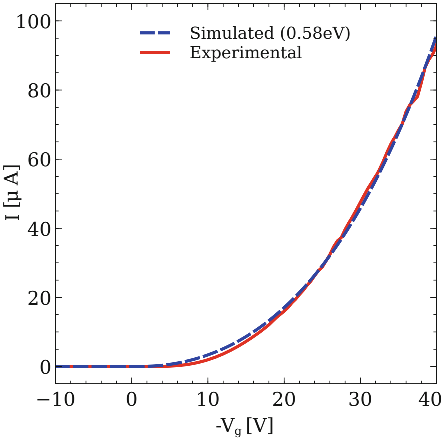


Fig. 7 Comparison between experimental (red) and simulated OTFT transcharacteristics at the optimal barrier value $\Phi_B = 0.58$ eV



Anyhow, the ratio between the capacitance computed in the accumulation and in the depletion regime shows that the nominal barrier leads to a capacitance drop at a low frequency of about 10^3 Hz that is not seen in experimental data, while the optimal barrier results in a plateau extending to a range that perfectly matches the one seen in experiments.

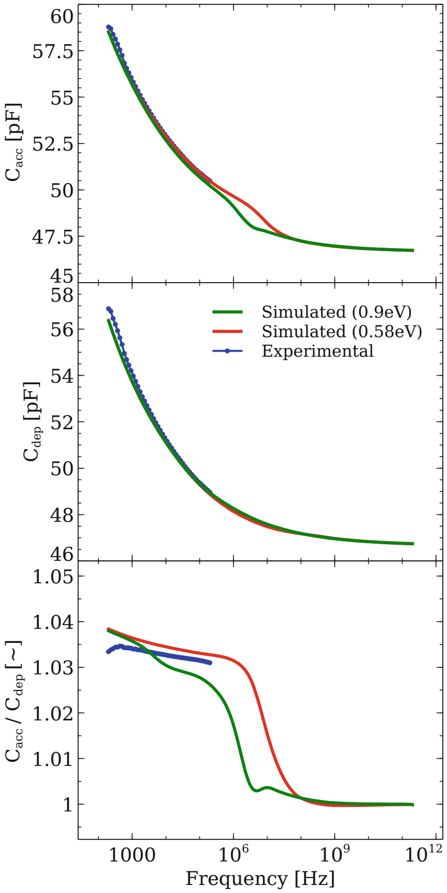


Fig. 8 CF curves in the accumulation regime ($V_{\mathrm{gate}} = +35$ V), the depletion regime ($V_{\mathrm{gate}} = -15$ V) and normalized, computed for the nominal injection barrier $\Phi_B = 0.9$ eV and for the optimal barrier $\Phi_B = 0.58$ eV. Experimental CF characteristics shown for comparison

References

1. Katsuaki, S.: Introduction to Printed Electronics. Springer, New York (2014)
2. Caironi, M., Noh, Y.Y.: Large Area and Flexible Electronics. Wiley–VCH, Weinheim (2015)
3. Paterson, A.F., Singh, S., Fallon, K.J., Hodsden, T., Han, Y., Schroeder, B.C., Bronstein, H., Heeney, M., McCulloch, I., Anthopoulos, T.D.: Recent progress in high-mobility organic transistors: a reality check. Adv. Mater. **30**(36) (2018)
4. de Vries, R.J., Badinski, A., Janssen, R.A.J., Coehoorn, R.: Extraction of the materials parameters that determine the mobility in disordered organic semiconductors from the current-voltage characteristics: accuracy and limitations. J. Appl. Phys. **113**(11), 114505 (2013)
5. Maddalena, F., de Falco, C., Caironi, M., Natali, D.: Assessing the width of Gaussian density of states in organic semiconductors. Org. Electron. **17**, 304–318 (2015)
6. Coehoorn, R., Pasveer, W.F., Bobbert, P.A., Michels, M.A.J.: Charge-carrier concentration dependence of the hopping mobility in organic materials with Gaussian disorder. Phys. Rev. B **72**(15), 155206 (2005)
7. Oehzelt, M., Koch, N., Heimel, G.: Organic semiconductor density of states controls the energy level alignment at electrode interfaces. Nat. Commun. **5**, 4174 (2014)

8. Africa, P.C., de Falco, C., Maddalena, F., Caironi, M., Natali, D.: Simultaneous extraction of density of states width, carrier mobility and injection barriers in organic semiconductors. Sci. Rep. **7**(1), 3803 (2017)
9. Natali, D., Caironi, M.: Charge injection in solution-processed organic field-effect transistors: physics, models and characterization methods. Adv. Mater. **24**(11), 1357–1387 (2012)
10. Jung, K.D., Kim, Y.C., Kim, B.J., Park, B.G., Shin, H., Lee, J.D.: An analytic current-voltage equation for top-contact organic thin film transistors including the effects of variable series resistance. Jpn. J. Appl. Phys. **47**(4S), 3174 (2008)
11. Barker, J.A., Ramsdale, C.M., Greenham, N.C.: Modeling the current-voltage characteristics of bilayer polymer photovoltaic devices. Phys. Rev. B **67**(7), 075205 (2003)
12. de Falco, C., Sacco, R., Verri, M.: Analytical and numerical study of photocurrent transients in organic polymer solar cells. Comput. Methods Appl. Mech. Eng. **199**(25–28), 1722–1732 (2010)
13. de Falco, C., Porro, M., Sacco, R., Verri, M.: Multiscale modeling and simulation of organic solar cells. Comput. Methods Appl. Mech. Eng. **245–246**, 102–116 (2012)
14. Coehoorn, R., Bobbert, P.A.: Effects of Gaussian disorder on charge carrier transport and recombination in organic semiconductors. Phys. Status Solidi A **209**(12), 2354–2377 (2012)
15. Bässler, H., Köhler, A.: Charge transport in organic semiconductors. Top. Curr. Chem. **312**, 1–65 (2012)
16. Baranovskii, S.D.: Theoretical description of charge transport in disordered organic semiconductors. Phys. Status Solidi B **251**(3), 487–525 (2014)
17. van Mensfoort, S., Coehoorn, R.: Effect of Gaussian disorder on the voltage dependence of the current density in sandwich-type devices based on organic semiconductors. Phys. Rev. B **78**(8), 085207 (2008)
18. Santoni, F., Gagliardi, A., der Maur, M.A., Di Carlo, A.: The relevance of correct injection model to simulate electrical properties of organic semiconductors. Org. Electron. **15**(7), 1557–1570 (2014)
19. Scott, J.C., Malliaras, G.G.: Charge injection and recombination at the metal-organic interface. Chem. Phys. Lett. **299**(2), 115–119 (1999)
20. de Falco, C., O'Riordan, E.: Interior layers in a reaction-diffusion equation with a discontinuous diffusion coefficient. Int. J. Numer. Anal. Model. **7**(4), 444–461 (2010)

On a Bloch-Type Model with Electron–Phonon Interactions: Modeling and Numerical Simulations

Brigitte Bidégaray-Fesquet, Clément Jourdana, and Kole Keita

Abstract In this work, we discuss how to take into account electron–phonon interactions in a Bloch type model for the description of quantum dots. The model consists in coupling an equation on the density matrix with a set of equations on quantities called phonon-assisted densities, one for each phonon mode. After a description of the model, we discuss how to discretize efficiently this non-linear coupling in view of numerical simulations.

1 Bloch Model

1.1 Quantum Dot Description

Quantum dots are usually described using electrons and holes. As detailed in [2], we prefer a conduction and valence electron description, where valence electrons can be seen as an absence of holes in a valence band. Due to the 3D confinement, energy levels are quantized for each species of electrons and can be indexed by integers. We denote respectively $(\epsilon_j^c)_{j\in\mathcal{J}^c}$ and $(\epsilon_j^v)_{j\in\mathcal{J}^v}$ the conduction and valence energy levels.

To describe the time evolution of the energy level occupations, we define a global density matrix by

$$\rho = \begin{pmatrix} \rho^c & \rho^{cv} \\ \rho^{vc} & \rho^v \end{pmatrix}. \tag{1}$$

B. Bidégaray-Fesquet · C. Jourdana (✉)
Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, Grenoble, France

Institute of Engineering Univ. Grenoble Alpes, Grenoble, France
e-mail: brigitte.bidegaray@univ-grenoble-alpes.fr; clement.jourdana@univ-grenoble-alpes.fr

K. Keita
Univ. Jean Lorougnon Guédé (UJLoG), Daloa, Ivory Coast

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_10

The matrices ρ^c and ρ^v are respectively the conduction and valence densities. Their diagonal terms, called populations, are the occupation probabilities and their off-diagonal terms, called coherences, describe the intra-band transitions. Finally, ρ^{cv} and $\rho^{vc} = \rho^{cv^*}$ (A^* denoting the Hermitian adjoint of a matrix A) describe the inter-band transitions.

The time-evolution of ρ can be driven by a free electron Hamiltonian associated to electron level energies and the interaction with an electromagnetic wave (see e.g. [2] for details):

$$i\hbar\partial_t\rho = [E_0 + \mathbf{E}\cdot\mathbf{M}, \rho], \tag{2}$$

where $[A, B]$ denotes the commutator $AB - BA$, $E_0 = \operatorname{diag}(\{\epsilon_j^c\}, \{\epsilon_j^v\})$, $\mathbf{M}$ is the dipolar moment matrix (a matrix that can be expressed in terms of the wave functions associated to each energy level) and $\mathbf{E}$ is a time-dependent electric field.

To study the interaction of the quantum dot with an electromagnetic field, Eq. (2) can be coupled with Maxwell equations:

$$\partial_t\mathbf{E} = c^2 \operatorname{curl}\mathbf{B} - \mu_0 c^2 \mathbf{J}, \tag{3}$$

$$\partial_t\mathbf{B} = -\operatorname{curl}\mathbf{E}, \tag{4}$$

$\mathbf{B}$ being the magnetic field, c the speed of light in free space and μ_0 the vacuum permeability. The coupling is expressed via the current density $\mathbf{J}$ which is given by

$$\mathbf{J} = n_a \operatorname{Tr}(\mathbf{M}\partial_t\rho), \tag{5}$$

where n_a is the quantum dot volume density.

Equation (2) is a Liouville equation and it confers a certain number of properties to the solution that have already been extensively studied in the literature. Here, we focus on the addition of electron–phonon (e–ph) interactions in such a model.

1.2 Electron–Phonon Hamiltonian

As in [3] where the addition of Coulomb interactions is discussed, the starting point is to use field quantification to write an e–ph Hamiltonian. We write it in the form $H^{c-ph} + H^{v-ph}$. It reflects that e–ph interactions cannot lead the electron to change species. In this work, only polar coupling to optical phonons is considered since it usually leads to the fastest dynamics in low excitation regime. The corresponding Fröhlich interaction Hamiltonian is given by (see e.g. [4, 6, 7]):

$$H^{c-ph} = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} \sum_{\alpha,\alpha'\in\mathcal{I}^c} G^c_{\mathbf{q},\alpha,\alpha'}\, c^\dagger_\alpha \left(b_\mathbf{q} + b^\dagger_{-\mathbf{q}}\right) c_{\alpha'}\, d\mathbf{q}, \tag{6}$$

$$H^{v-ph}=\frac{1}{|\mathcal{B}|}\int_{\mathcal{B}}\sum_{\alpha,\alpha'\in\mathcal{J}^v} G^v_{\mathbf{q},\alpha,\alpha'}\, v^\dagger_\alpha \left(b_{\mathbf{q}}+b^\dagger_{-\mathbf{q}}\right) v_{\alpha'}\,\mathrm{d}\mathbf{q}. \tag{7}$$

The operators $c^\dagger_j$ and c_j (resp. $v^\dagger_j$ and v_j) are creation and annihilation operators for conduction (resp. valence) electrons and the operators $b^\dagger_{\mathbf{q}}$ and $b_{\mathbf{q}}$ are those for phonons, where the phonon mode $\mathbf{q}$ belongs to the Brillouin zone $\mathcal{B}$ of the underlying crystal. The volume of the Brillouin zone is denoted $|\mathcal{B}|$. For $e \in \{c, v\}$, $G^e_{\mathbf{q}}$ is a matrix whose coefficients are expressed in terms of the wave functions associated to each energy level:

$$G^e_{\mathbf{q},\alpha,\alpha'} = \mathcal{E}_{\mathbf{q}} \int \psi^{e*}_\alpha(\mathbf{r}) \exp(i\mathbf{q}\cdot\mathbf{r})\psi^e_{\alpha'}(\mathbf{r})\,\mathrm{d}\mathbf{r}, \tag{8}$$

$\mathcal{E}_{\mathbf{q}}$ being the Fröhlich constant [6] defined such that $G^{e*}_{\mathbf{q}} = G^e_{-\mathbf{q}}$.

1.3 Phonon-Assisted Densities and Time Evolution

We first recall that the commutation relations between conduction and valence electron operators are the following

$$\forall j,k \in \mathcal{J}^c, \qquad \{c^\dagger_j, c_k\} = \delta_{jk}, \qquad \{c^\dagger_j, c^\dagger_k\} = \{c_j, c_k\} = 0, \tag{9}$$

$$\forall j,k \in \mathcal{J}^v, \qquad \{v_j, v^\dagger_k\} = \delta_{jk}, \qquad \{v_j, v_k\} = \{v^\dagger_j, v^\dagger_k\} = 0, \tag{10}$$

$$\forall j \in \mathcal{J}^c \text{ and } k \in \mathcal{J}^v, \qquad [c^\dagger_j, v^\dagger_k] = [c^\dagger_j, v_k] = [c_j, v^\dagger_k] = [c_j, v_k] = 0, \tag{11}$$

where δ_{jk} is the Kronecker delta function. Contrarily to electrons, phonons are bosons and they obey the following commutation rules

$$\forall\, \mathbf{q},\mathbf{q}' \in \mathcal{B}, \qquad [b_{\mathbf{q}}, b^\dagger_{\mathbf{q}'}] = |\mathcal{B}|\,\delta(\mathbf{q}-\mathbf{q}'), \qquad [b_{\mathbf{q}}, b_{\mathbf{q}'}] = [b^\dagger_{\mathbf{q}}, b^\dagger_{\mathbf{q}'}] = 0, \tag{12}$$

where δ is the Dirac delta function. Indeed, we do not consider here a quantization of phonon modes (even if in order to perform numerical simulations we will discretize in Sect. 2 the Brillouin zone and use a quadrature formula to approximate integrals over $\mathbf{q}$).

As in [5], we introduce phonon-assisted density matrices

$$S_{\mathbf{q}} = \begin{pmatrix} S^{cc}_{\mathbf{q}} & S^{cv}_{\mathbf{q}} \\ S^{vc}_{\mathbf{q}} & S^{vv}_{\mathbf{q}} \end{pmatrix} \tag{13}$$

where $S^{ef}_{\mathbf{q},\alpha,\alpha'} = \langle f^{\dagger}_{\alpha'} b_{\mathbf{q}} e_{\alpha}\rangle$, $e, f \in \{c, v\}$. Then, using intensively the commutativity rules (9)–(12), the time evolution of the density matrix due to e–ph interactions can be cast in a very compact form as

$$i\hbar\partial_t \rho|_{e-ph} = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} [G_{\mathbf{q}}, S_{\mathbf{q}} + S^{*}_{-\mathbf{q}}]\mathrm{d}\mathbf{q} \equiv P(S), \tag{14}$$

where we have introduced the notations

$$G_{\mathbf{q}} = \begin{pmatrix} G^{c}_{\mathbf{q}} & 0 \\ 0 & G^{v}_{\mathbf{q}} \end{pmatrix} \text{ and } S = \{S_{\mathbf{q}},\ \mathbf{q} \in \mathcal{B}\}. \tag{15}$$

We have $P(S)^{*} = -P(S)$ independently of the structure of S. Therefore ρ, which is initially Hermitian, remains Hermitian through time evolution via (14). This equation is also trace preserving since the right-hand side is a combination of trace-free commutators.

To close the system, we now look for the time evolution of phonon-assisted densities. First, using again the commutativity rules (9)–(12), we make explicit the commutators between the e–ph interaction Hamiltonian and the other Hamiltonians involved in the system: the free electron Hamiltonian, the electromagnetic interaction Hamiltonian and the free phonon Hamiltonian. For instance, the free phonon Hamiltonian is given by

$$H^{\mathrm{ph}} = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} E_{\mathbf{q}} b^{\dagger}_{\mathbf{q}} b_{\mathbf{q}}\, \mathrm{d}\mathbf{q}$$

and the involved commutator $[c^{\dagger}_{\alpha}(b_{\mathbf{q}}+b^{\dagger}_{-\mathbf{q}})c_{\alpha'}, H^{\mathrm{ph}}]$ is equal to $E_{\mathbf{q}} c^{\dagger}_{\alpha}(b_{\mathbf{q}}-b^{\dagger}_{-\mathbf{q}})c_{\alpha'}$. Then, we use the Wick theorem [8] to approximate the means involving four operators by sums and products of densities. It closes the system at the cost of rendering it non-linear. After computations, we finally obtain, for each $\mathbf{q} \in \mathcal{B}$, the following equation

$$\begin{aligned} i\hbar\partial_t S_{\mathbf{q}}|_{e-ph} &= E_{\mathbf{q}} S_{\mathbf{q}} + \frac{1}{2}\{G^{*}_{\mathbf{q}}, \rho\} + (\frac{1}{2} + n_{\mathbf{q}})[G^{*}_{\mathbf{q}}, \rho] + C(\rho, G^{*}_{\mathbf{q}}) \\ &\equiv E_{\mathbf{q}} S_{\mathbf{q}} + Q_{\mathbf{q}}(\rho). \end{aligned} \tag{16}$$

The term $\{A, B\}$ denotes the skew-commutator $AB + BA$, $n_{\mathbf{q}} = \langle b^{\dagger}_{\mathbf{q}} b_{\mathbf{q}}\rangle$ is the phonon density expressed in terms of the phonon energy $E_{\mathbf{q}}$ by the Bose–Einstein statistics, and $C(\rho, G^{*}_{\mathbf{q}})$ is a non-linear term expressed as

$$C(\rho, G^{*}_{\mathbf{q}}) = -\widetilde{\rho} G^{*}_{\mathbf{q}} \widetilde{\rho} + \mathrm{Tr}(G^{*}_{\mathbf{q}} \widetilde{\rho})\widetilde{\rho} \tag{17}$$

where $\widetilde{\rho} = \rho \begin{pmatrix} I^c & 0 \\ 0 & -I^v \end{pmatrix}$, I^c and I^v being the identity matrices for the conduction and valence spaces.

To summarize, the e–ph Bloch model consists in coupling an equation on ρ

$$i\hbar\partial_t \rho = [E_0 + \mathbf{E}\cdot\mathbf{M}, \rho] + P(S) \tag{18}$$

with a set of equations on $S_{\mathbf{q}}$ (one for each $\mathbf{q}$)

$$i\hbar\partial_t S_{\mathbf{q}} = E_{\mathbf{q}} S_{\mathbf{q}} + [E_0 + \mathbf{E}\cdot\mathbf{M}, S_{\mathbf{q}}] + Q_{\mathbf{q}}(\rho). \tag{19}$$

2 Numerical Scheme

For simulations, we consider a collection of quantum dots which are scattered in a one dimensional space along the z direction and interact not directly but through the interaction with the electromagnetic field. Therefore, densities depend on time and space and the e–ph Bloch model (18)–(19) is coupled with Maxwell equations (3)–(4).

First, we introduce a uniform discretization of the Brillouin zone $\mathcal{B}$ using N_q points. The integral over $\mathbf{q}$ in (14) is approximated by a simple quadrature formula and consequently the global phonon-assisted density $S = \{S_{\mathbf{q}_l}, l = 1, \cdots, N_q\}$ is computed solving N_q independent equations (19).

We fix a time step $\delta t > 0$ and we discretize uniformly the time using $N_t + 1$ points: $t_n = n\delta t$ for $n \in \{0, \ldots, N_t\}$ with $\delta t = \frac{T}{N_t}$, T being the final time. Analogously, we fix a space step $\delta z > 0$ and we discretize uniformly the space using $N_z + 1$ points: $z_j = j\delta z$ for $j \in \{0, \ldots, N_z\}$ with $\delta z = \frac{L}{N_z}$, L being the length of the quantum dot collection.

A finite difference Yee scheme is used for Maxwell equations (3)–(4): for all $n \in \{0, \cdots, N_t - 1\}$,

$$B^{n+\frac{1}{2}}_{y,j+\frac{1}{2}} = B^{n-\frac{1}{2}}_{y,j+\frac{1}{2}} - \frac{\delta t}{\delta z}(E^n_{x,j+1} - E^n_{x,j}), \qquad \forall j \in \{0, \cdots, N_z - 1\}, \tag{20}$$

$$E^{n+1}_{x,j} = E^n_{x,j} - c^2\frac{\delta t}{\delta z}(B^{n+\frac{1}{2}}_{y,j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{y,j-\frac{1}{2}}) - \mu_0 c^2 \delta t J^{n+\frac{1}{2}}_{x,j}, \quad \forall j \in \{1, \cdots, N_z\}, \tag{21}$$

with $J^{n+\frac{1}{2}}_{x,j} = -\frac{i n_a}{\hbar} Tr\Big(M_x\Big[E_0, \rho^{n+\frac{1}{2}}_j\Big] + M_x P(S^n_j)\Big)$. In (20), we use the convention $B^{-\frac{1}{2}}_{y,j+\frac{1}{2}} = 0$ for all $j \in \{0, \cdots, N_z - 1\}$ and (21) is initialized by $E^0_{x,j}$ and $S^0_j = 0$ for all $j \in \{0, \cdots, N_z\}$. Finally, a time dependent incident wave E_{inc} is injected in the left part of the device by the boundary condition

$$E_{x,0}^{n+1} = E_{x,1}^{n} + E_{inc}^{n+1} - E_{inc}^{n-\frac{\delta z}{c}} + \frac{1 - c\frac{\delta t}{\delta z}}{1 + c\frac{\delta t}{\delta z}}\big(E_{x,0}^{n} - E_{x,1}^{n+1} + E_{inc}^{n+1-\frac{\delta z}{c}} - E_{inc}^{n}\big),$$

for all $n \in \{0, \cdots, N_t - 1\}$. Notice that the discretization steps δt and δz are chosen in order to satisfy the stability condition imposed by this numerical scheme (see e.g. [1]).

We consider a weak coupling between the Maxwell and Bloch equations. It means that E and ρ are not discretized at the same time to avoid a fixed point procedure. Equations (18)–(19) are discretized on a staggered grid in time and each equation is solved using a Strang splitting method:
for all $l \in \{1, \cdots, N_{\mathbf{q}}\}$, for all $n \in \{0, \cdots, N_t - 1\}$, for all $j \in \{0, \cdots, N_z\}$,

$$\begin{aligned} S_{\mathbf{q}_l,j}^{n+1} = {} & \mathcal{A}_3\Big(\frac{\delta t}{2}, E_{\mathbf{q}_l} I\Big)\mathcal{A}_2\Big(\frac{\delta t}{2}, E_0 + \frac{E_{x,j}^{n} + E_{x,j}^{n+1}}{2} M_x\Big) \\ & \mathcal{A}_1\Big(\delta t, Q_{\mathbf{q}_l}(\rho_j^{n+\frac{1}{2}})\Big)\mathcal{A}_2\Big(\frac{\delta t}{2}, E_0 + \frac{E_{x,j}^{n} + E_{x,j}^{n+1}}{2} M_x\Big)\mathcal{A}_3\Big(\frac{\delta t}{2}, E_{\mathbf{q}_l} I\Big) S_{\mathbf{q}_l,j}^{n}, \end{aligned} \tag{22}$$

$$\rho_j^{n+\frac{3}{2}} = \mathcal{A}_2\Big(\frac{\delta t}{2}, E_0 + E_{x,j}^{n+1} M_x\Big)\mathcal{A}_1\Big(\delta t, P(S_j^{n+1})\Big)\mathcal{A}_2\Big(\frac{\delta t}{2}, E_0 + E_{x,j}^{n+1} M_x\Big)\rho_j^{n+\frac{1}{2}}. \tag{23}$$

Equations (22)–(23) are initialized by $S_{\mathbf{q}_l,j}^{0} = 0$ and $\rho_j^{\frac{1}{2}} = \mathcal{A}_2\Big(\frac{\delta t}{2}, E_0 + E_{x,j}^{0} M_x\Big)\rho_j^{0}$. In these expressions, $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ are three semigroups defined by

$$\mathcal{A}_1(t, B)A = A - \frac{it}{\hbar}B, \qquad \mathcal{A}_2(t, B)A = e^{-\frac{itB}{\hbar}} A e^{\frac{itB}{\hbar}} \qquad \text{and} \qquad \mathcal{A}_3(t, B)A = e^{-\frac{itB}{\hbar}} A.$$

They can be computed exactly using matrix exponential formulas. The advantage of this splitting is that it numerically preserves positiveness for each equation.

3 Numerical Simulations

We now perform numerical simulations in order to assess the capability of the Bloch model to describe the interaction of quantum dots with an electromagnetic field. More precisely, we study a Self-Induced Transparency (SIT) case. It is a phenomenon that allows to obtain an exact population inversion with an unchanged electric field, using a light pulse resonant with the electron energy levels.

The propagating field E_{inc} that we inject is a pulse with a specific envelope and a center frequency ω_0. First, we recall some results obtained without e–ph interactions

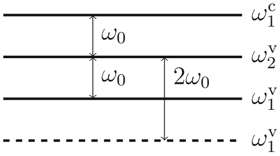


Fig. 1 Energy level description for the two 3-level test cases

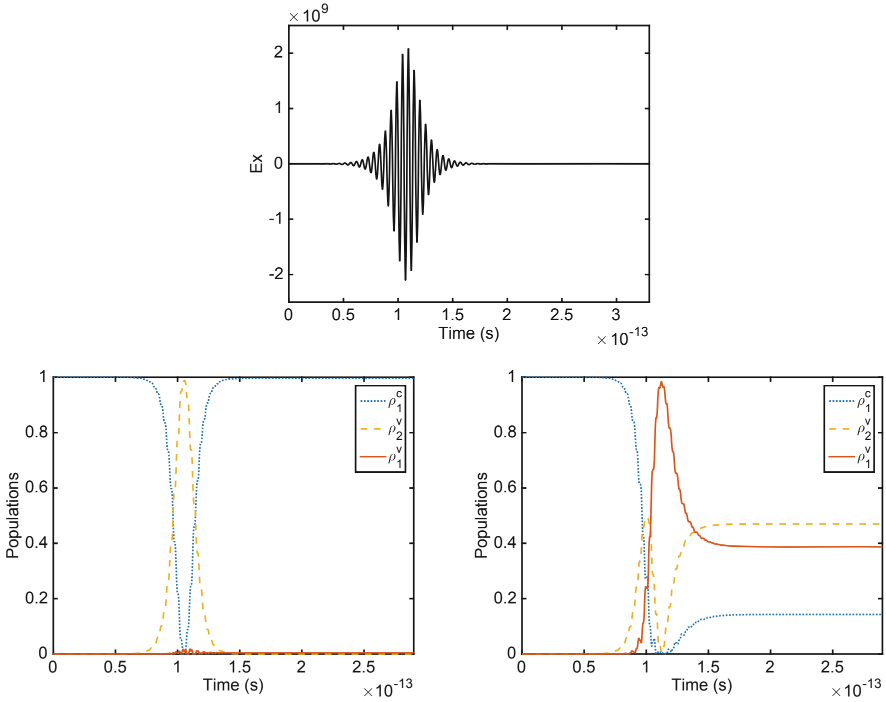


Fig. 2 Time evolution of E_x (at the top) and populations (at the bottom) without e–ph interactions. Left: $\epsilon_2^v - \epsilon_1^v = 2\hbar\omega_0$; Right: $\epsilon_2^v - \epsilon_1^v = \hbar\omega_0$

for two 3-level test cases with a conduction level and two valence levels (see [3] for more details). The energy between the conduction level and the first valence level corresponds to the field frequency ω_0. In the first case (dashed line on the schematic representation Fig. 1), the transition between the two valence levels is $2\hbar\omega_0$. It is instead $\hbar\omega_0$ in the second case (solid line).

In Fig. 2, we represent at the top the time evolution of the electric field E_x for a given quantum dot and at the bottom the corresponding time evolution of populations for the two test cases. In the left picture (corresponding to the case $\epsilon_2^v - \epsilon_1^v = 2\hbar\omega_0$), we observe a complete population inversion due to the light pulse. Instead, when the transition between the two valence levels is resonant with the field (right picture), the SIT phenomenon is destroyed.

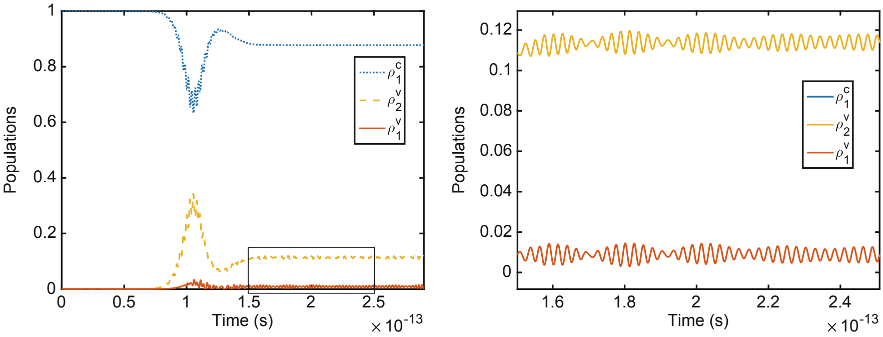


Fig. 3 Time evolution of populations for $\epsilon_2^v - \epsilon_1^v = 2\hbar\omega_0$ and $N_{\mathbf{q}} = 100$ (left) and zoom inside the rectangle (right)

Now, we add the e–ph interactions. For simplicity, we assume that optical phonons are almost dispersionless and take a constant phonon energy $E_{\mathbf{q}}$ for the $N_q = 100$ phonon modes that we consider. The time evolution of populations is presented in Fig. 3 for the test case in which $\epsilon_2^v - \epsilon_1^v = 2\hbar\omega_0$. We observe that e–ph interactions destroy the SIT phenomenon, even for valence levels far apart enough. In addition to a relaxation behavior, fast oscillations are generated for the two valence levels and persist after the electromagnetic pulse (as emphasized in the zoom presented in the right picture of Fig. 3) .

4 Conclusion

To study the interaction of quantum dots with an electromagnetic field taking into account e–ph interactions, we proposed an efficient discretization for the coupling appearing between the equation on the electron density ρ and the set of equations on the phonon-assisted densities $S_{\mathbf{q}}$. For a better modeling, it will be now interesting to investigate how to take into account, via a kinetic equation, the quantum-well wetting layer into which quantum dots are embedded.

References

1. Bidégaray, B.: Time discretizations for Maxwell-Bloch equations. Numer. Methods Partial Differ. Equ. **19**(3), 284–300 (2003)
2. Bidégaray-Fesquet, B.: Positiveness and Pauli exception principle in raw Bloch equations for quantum boxes. Ann. Phys. **325**(10), 2090–2102 (2010)
3. Bidégaray-Fesquet, B., Keita, K.: A nonlinear Bloch model for Coulomb interaction in quantum dots. J. Math. Phys. **55**(2), 021501 (2014)
4. Fröhlich, H.: Electrons in lattice fields. Adv. Phys. **3**(11), 325–361 (1954)

5. Gehrig, E., Hess, O.: Mesoscopic spatiotemporal theory for quantum-dot lasers. Phys. Rev. A **65**(033804), 1–16 (2002)
6. Haug, H., Koch, S.W.: Quantum Theory of the Optical and Electronic Properties of Semiconductors, 5th edn. World Scientific, Singapore (2009)
7. Stauber, T., Zimmermann, R., Castella, H.: Electron-phonon interaction in quantum dots: a solvable model. Phys. Rev. B **62**(11), 7336–7343 (2000)
8. Wick, G.C.: The evaluation of the collision matrix. Phys. Rev. **80**, 268–272 (1950)

Charge and Phonon Transport in Suspended Monolayer Graphene

Marco Coco, Giovanni Mascali, and Vittorio Romano

Abstract Thermal effects are playing a crucial role for the design of electron nanoscale devices. The present contribution deals with charge and phonon transport under an applied external electric field in a suspended monolayer of graphene. A major question is represented by the phonon-phonon collision operator involving in general a three particle scattering mechanism. To model the phonon-phonon interactions a relaxation time approximation is employed. This requires the introduction of a local equilibrium phonon temperature whose definition is still a matter of debate for a general non equilibrium situation. Here, two different approaches are presented and discussed.

1 Introduction

Graphene is one of the most promising materials for future nano-electronics because of its unique electrical and thermal properties. It has been increasingly investigated from different points of view. A correct mathematical description of transport phenomena in graphene is fundamental and there have been a lot of attempts and approaches, starting from the well-established results for other traditional semiconductor materials and devices. In particular, kinetic and macroscopic models have been applied, see for example [1–5]. Also stochastic approaches, as the Direct

M. Coco
Dipartimento di Matematica e Informatica “Ulisse Dini”, Università degli Studi di Firenze, Firenze, Italy
e-mail: marco.coco@unifi.it

G. Mascali
Dipartimento di Matematica e Informatica, Università della Calabria, Rende, Italy

INFN-Gruppo c. Cosenza, Cosenza, Italy
e-mail: giovanni.mascali@unical.it

V. Romano (✉)
Dipartimento di Matematica e Informatica, Università degli Studi di Catania, Catania, Italy
e-mail: romano@dmi.unict.it

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_11

Simulation Monte Carlo (DSMC), are consolidated methods for numerically solving transport equations in graphene. However, many of them do not properly take into account the Pauli exclusion principle [6], that in graphene is no longer negligible for the high values of the electron densities. This problem has been overtaken by a new DSMC procedure which correctly includes the Pauli principle. Cross-validations with deterministic solutions, e.g. the Discontinuous Galerkin-based ones, confirm that this approach is completely satisfactory [7]. The new DSMC approach allows us to describe electron transport in graphene also in cases when graphene is on several types of substrates [8, 9].

Together with the correct inclusion of the Pauli exclusion principle into the charge transport simulations, to have a satisfactory and complete description of transport phenomena in graphene, it is necessary an adequate treatment of the phonon-phonon collision operators. Its complete form involves at least a three particle kernel and this makes the numerics rather complicated. A way to overcome the problem is to replace the original collision operators with a simpler relaxation time approximation. This requires the introduction of a local phonon temperature whose definition is still a matter of debate for a general non equilibrium situation. In this paper two different approaches for defining the local equilbrium phonon temperature are introduced and discussed on the basis of the results reported in [5] and [10, 11].

2 Kinetic Model

Graphene is made of carbon atoms arranged in a honeycomb hexagonal lattice. The most part of electrons are located in the wave vector space around the *Dirac points* K and K', which are the vertices of the hexagonal primitive cell of the reciprocal lattice. At the Dirac points, the valence and conduction band touch each other and, therefore, graphene is a semimetal. In the proximity of the Dirac points the energy bands for electrons can be approximated by a conical band structure and the electrons behave as massless Dirac fermions [12]. We consider Fermi levels high enough to neglect the dynamics of the electrons in the valence band, that we consider fully occupied. This situation is similar to an n-type doping for the traditional semiconductors. Around the equivalent Dirac points the band energy ε_ℓ is approximated by a linear relation

$$\varepsilon_\ell = \hbar\, v_F\, |\mathbf{k} - \mathbf{k}_\ell| \,, \tag{1}$$

and the group velocity is given by

$$\mathbf{v}_\ell = \frac{1}{\hbar}\, \nabla_{\mathbf{k}}\, \varepsilon_\ell \,.$$

Here $\mathbf{k}$ is the electron wave-vector, v_F is the (constant) Fermi velocity, $\hbar$ the Planck constant divided by $2\,\pi$, and $\mathbf{k}_\ell$ is the position of the Dirac point $\ell = K, K'$.

In a semiclassical kinetic setting questa virgola la toglierei, for the electrons in the conduction band, the charge transport in graphene is described by two Boltzmann equations, for the K and K' valleys.

$$\frac{\partial f_\ell(t,\mathbf{x},\mathbf{k})}{\partial t} + \mathbf{v}_\ell \cdot \nabla_{\mathbf{x}} f_\ell(t,\mathbf{x},\mathbf{k}) \; - \frac{e}{\hbar}\mathbf{E}\cdot\nabla_{\mathbf{k}} f_\ell(t,\mathbf{x},\mathbf{k}) = \frac{df_\ell}{dt}(t,\mathbf{x},\mathbf{k})\bigg|_{e-ph} \tag{2}$$

where $f_\ell(t,\mathbf{x},\mathbf{k})$ represents the distribution function of charge carriers in the valley ℓ (K or K'), at position $\mathbf{x}$, time t, and with wave-vector $\mathbf{k}$. We denote by $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{k}}$ the gradients with respect to the position and the wave-vector, respectively. e is the elementary (positive) charge and $\mathbf{E}$ is the externally applied electric field.

All the scattering events between electrons and phonons are described by the collision term, at the right hand side of (2). The scattering of electrons can be with longitudinal or transversal, acoustic and optical phonons, *LA*, *TA*, *LO* and *TO*, respectively. Both the acoustic and optical phonon scatterings are intra-valley and intra-band. Eventually, one has to take into account also the scattering of electrons with K phonons which is inter-valley and, therefore, pushes electrons from a valley to the nearby one. The general form of the collision term can be written as

$$\begin{aligned}\frac{df_\ell}{dt}(t,\mathbf{x},\mathbf{k})\bigg|_{e-ph} = \sum_{\ell'}\bigg[\int_{\mathcal{B}} & S_{\ell',\ell}(\mathbf{k}',\mathbf{k})\, f_{\ell'}(t,\mathbf{x},\mathbf{k}')\,(1 - f_\ell(t,\mathbf{x},\mathbf{k}))\, d\mathbf{k}' \\ & - \int_{\mathcal{B}} S_{\ell,\ell'}(\mathbf{k},\mathbf{k}')\, f_\ell(t,\mathbf{x},\mathbf{k})\,\big(1 - f_{\ell'}(t,\mathbf{x},\mathbf{k}')\big)\, d\mathbf{k}'\bigg],\end{aligned}$$

where the total transition rate $S_{\ell',\ell}(\mathbf{k}',\mathbf{k})$ is given by the sum of the contributions of the several types of scatterings [5]:

$$\begin{aligned}S_{\ell',\ell}(\mathbf{k}',\mathbf{k}) = \sum_\mu \left|G^{(\mu)}_{\ell',\ell}(\mathbf{k}',\mathbf{k})\right|^2 \Big[& \big(g_\mu^- + 1\big)\,\delta\big(\varepsilon_\ell(\mathbf{k}) - \varepsilon_{\ell'}(\mathbf{k}') + \hbar\,\omega_\mu\big) \\ & + g_\mu^+\,\delta\left(\varepsilon_\ell(\mathbf{k}) - \varepsilon_{\ell'}(\mathbf{k}') - \hbar\,\omega_\mu\right)\Big].\end{aligned} \tag{3}$$

The index μ labels the μth phonon species. The $\left|G^{(\mu)}_{\ell',\ell}(\mathbf{k}',\mathbf{k})\right|^2$'s are the electron-phonon coupling matrix elements, which describe the interaction mechanism by which an electron goes, from the state of wave-vector $\mathbf{k}'$ belonging to the valley ℓ' to the state of wave-vector $\mathbf{k}$ belonging to the valley ℓ, through the emission or absorption of a μth phonon. The symbol δ denotes the Dirac distribution, ω_μ is the μth phonon frequency, $g_\mu(\mathbf{q})$ is the phonon distribution for the μ-type phonons with $\mathbf{q}$ the phonon wave-vector belonging to the Brillouin zone $\mathcal{B}$. In (3), $g_\mu^\pm = g_\mu\left(\mathbf{q}^\pm\right)$, where $\mathbf{q}^\pm = \pm\left(\mathbf{k}' - \mathbf{k}\right)$, stemming from the momentum conservation. The K and K' valleys can be treated as equivalent and in the following we consider the population of one single valley.

Similarly, the dynamics of the phonon populations is described by solving the following Boltzmann equations for the phonon distributions $g_\mu(t, \mathbf{x}, \mathbf{q})$

$$\frac{\partial g_{op}}{\partial t} = C_{op}, \quad op = LO, TO, K, \tag{4}$$

$$\frac{\partial g_{ac}}{\partial t} + \mathbf{c}_{ac} \cdot \nabla_{\mathbf{x}} g_{ac} = C_{ac}, \quad ac = LA, TA, \tag{5}$$

$\mathbf{c}_{ac} = \nabla_{\mathbf{q}}\, \omega_{ac}$ is the acoustic phonon group velocity, $\hbar\omega_{ac}$ being the phonon energy and $\mathbf{q}$ the phonon wave vector.

The group velocity of the optical phonons disappears because of the Einstein approximation $\hbar\omega_{op} \approx$ const, which can be used for them, while, regarding the acoustic phonons, the Debye approximation can be used: $\omega_{ac} = c_{ac}|\mathbf{q}|$, with c_{ac} the sound speed of the sth acoustic branch.

In principle also the Z phonons should be included although they do not interact with electrons but they contribute to the total crystal temperature. In the present article the Z phonons will be neglected for the sake of simplicity.

The phonon collision term splits in two parts

$$C_\mu = C_\mu^{p-e} + C_\mu^{p-p}, \quad \mu = LA, TA, LO, TO, K. \tag{6}$$

C_μ^{p-e} represents the phonon-electron collision operator, while C_μ^{p-p} describes the phonon-phonon interactions, that are a very difficult problem to deal with from a numerical point of view. For this reason they are usually treated by means of a Bhatnagar-Gross-Krook (BGK) approximation [13]

$$C_\mu^{p-p} = -\frac{g_\mu - g_\mu^{LE}}{\tau_\mu}.$$

This describes the relaxation of each phonon branch towards an equilibrium condition, that is represented by a local equilibrium distribution g_μ^{LE}, whose temperature, we refer to as the local temperature T_L, is the same for each phonon population.

We assume that the local equilibrium phonon distributions are given by Bose-Einstein distributions

$$g_\mu^{LE} = \left[e^{\hbar\omega_\mu / k_B T_L} - 1 \right]^{-1}. \tag{7}$$

The functions $\tau_\mu = \tau_\mu(T_\mu)$ are the temperature dependent phonon relaxation times. We remark that each relaxation time is supposed to depend only on the temperature T_μ of the same branch.

If we know the phonon distributions g_μ's, we can calculate the average phonon energy densities

$$W_\mu = \frac{1}{(2\pi)^2} \int_{\mathcal{B}} \hbar\omega_\mu\, g_\mu\, d\mathbf{q}, \tag{8}$$

and the temperatures T_μ of each phonon branch are determined from

$$\int_{\mathcal{B}} \hbar\omega_\mu g_\mu(\mathbf{q}) d\mathbf{q} = \int_{\mathcal{B}} \hbar\omega_\mu \left[e^{\hbar\omega_\mu / k_B T_\mu} - 1 \right]^{-1} d\mathbf{q}. \tag{9}$$

From the general properties of the phonon collision operators, the relation

$$\sum_\mu \frac{W_\mu - W_\mu^{LE}}{\tau_\mu} = 0 \tag{10}$$

holds, where W_μ^{LE} is calculated by means of (7). *T_L is obtained by numerically solving the non linear relation arising from (10).*

It is possible to prove that (10) admits a unique solution. For further details we refer to [11] where the previous approach has been adopted to devise a simulation scheme for the electron-phonon transport in graphene.

3 Alternative Form of the Local Temperature

The concept of non equilibrium temperature is a subtle topic and still a matter of debate [14, 15]. In the previous section we have introduced the local temperature by the relation (10) which stems from the properties of the phonon-phonon collision operator. The rational is that the collision operator *pushes* the system, in a characteristic time related to the relaxation times, toward an equilibrium state with a single global temperature. However, in statistical mechanics one of the most reasonable and adopted way to generalise the concept of temperature in a non equilibrium state is relating T_L to the Lagrange multipliers associated to the energy constraint.

For phonon transport in graphene the approach based on the Lagrange multipliers has been followed in [5] (to which the interested reader is referred to for the details) within the application of the Maximum Entropy Principle (MEP) (see [16] for a review of MEP in semiconductors). Let us recall here the main steps.

The temperature of each phonon branch is introduced as in the previous section while the local temperature is defined as follows [13, 17]. *The temperature T_L^* is the common temperature we must assign to each species in order to have*

$$\sum_\mu W_\mu = \sum_\mu W_\mu^{LE}. \tag{11}$$

In other words, T_L^* is the common temperature each phonon species should have if they would be in local thermodynamic equilibrium among them in order to preserve the total energy.

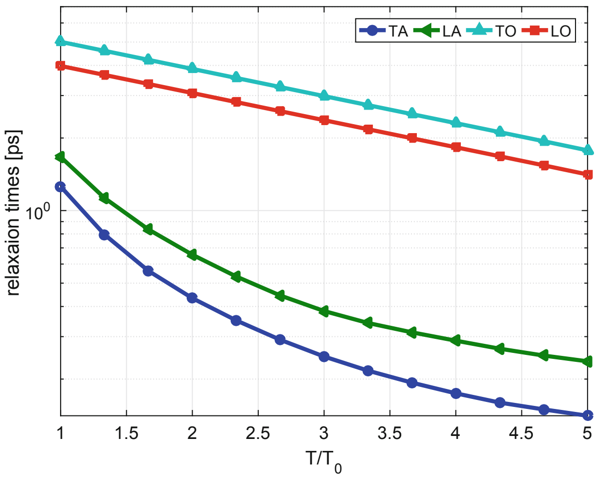


Fig. 1 Relaxation times versus the local temperature normalized with respect to the room one T_0

The new and the old definitions of local temperature are equivalent if all the relaxation times are equal, that is

$$\tau_\mu = \tau, \quad \mu = LO, TO, LA, TA, K, \tag{12}$$

but this assumption is not compatible with experimental data, as clearly indicated by Fig. 1. As a consequence, the two definitions of temperature do not coincide.

T_L^* is related only to the energy of the system and does not take into account any scattering mechanism. If T_L^* is assumed as the correct definition of local temperature, then the relation (10) must be considered as a constraint on the relaxation times.

In the model formulated in [5] for charge and phonon transport in graphene, a certain number of moments of electron and phonon distributions are used as fundamental variables, and the extra fluxes and the production terms, which appear in the corresponding balance equations, are additional unknown quantities and require constitutive relations in terms of the fundamental variables. By resorting to MEP, the electron and phonon occupation numbers can be estimated by the maximum entropy distributions f_{MEP} and $g_{\mu,MEP}$, μ=LO, TO, K, LA, TA, which solve the following maximization problem:

$$\max_{f, g_\mu} S[f, g_\mu],$$

under the constraint that a certain number of moments, the fundamental variables, are known.

$S[f, g_\mu]$ is the total entropy which depends on the electron and phonon distribution functions f, g_μ and whose expression is reported in [5].

If in particular for the phonons we choose the following moments

$$W_{op} = \frac{1}{(2\pi)^2} \int_{\mathcal{B}} \hbar\omega_{op} g_{op} \, d\mathbf{q},$$

$$\mathbf{P}_{op} = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \hbar\mathbf{q} g_{op} \, d\mathbf{q}, \; op = LO, TO, K, \tag{13}$$

$$W_{ac} = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \hbar\omega_{ac} g_{ac} \, d\mathbf{q},$$

$$\mathbf{Q}_{ac} = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \hbar\omega_{ac} \mathbf{c}_{ac} g_{ac} \, d\mathbf{q}, \; ac = LA, TA, \tag{14}$$

and, solving the above constrained maximization problem, we get

$$g_{op,MEP} = \frac{1}{\exp\left(\lambda_{W_{op}} \varepsilon_{op} + \hbar\, \mathbf{q} \cdot \lambda_{\mathbf{P}_{op}}\right) - 1}, \quad op = LO, TO, K,$$

$$g_{ac,MEP} = \frac{1}{\exp\left(\lambda_{W_{ac}} \varepsilon_{ac} + \varepsilon_{ac}\, \mathbf{c}_{ac} \cdot \lambda_{\mathbf{Q}_{ac}}\right) - 1}, \quad ac = LA, TA,$$

where the λ's are the Lagrange multipliers arising from the presence of the constraints.

In order to manage the problem of inverting the constraints, we linearize the occupation numbers around their isotropic part, obtaining

$$g_{op,MEP} \approx \frac{1}{e^{\lambda_{W_{op}} \varepsilon_{op}} - 1} \left[1 - \frac{e^{\lambda_{W_{op}} \varepsilon_{op}}}{e^{\lambda_{W_{op}} \varepsilon_{op}} - 1} \hbar\mathbf{q} \cdot \lambda_{\mathbf{P}_{op}} \right], \; op = LO, TO, K, \tag{15}$$

$$g_{ac,MEP} \approx \frac{1}{e^{\lambda_{W_{ac}} \varepsilon_{ac}} - 1} \left[1 - \frac{e^{\lambda_{W_{ac}} \varepsilon_{ac}}}{e^{\lambda_{W_{ac}} \varepsilon_{ac}} - 1} \varepsilon_{ac}\, \mathbf{c}_{ac} \cdot \lambda_{\mathbf{Q}_{ac}} \right], \quad ac = LA, TA. \tag{16}$$

By substituting (15)–(16) into the constraints (13)–(14) and by solving them with respect to the Lagrange multipliers, one finds

$$\lambda_{W_{op}} = \frac{1}{\varepsilon_{op}} \ln\left(1 + \frac{y\varepsilon_{op}}{W_\eta}\right), \; \eta = LO, TO, K, \tag{17}$$

$$\lambda_{W_{ac}} = \left(\frac{4\pi y \zeta(3)}{\hbar^2 c_{ac}^2}\right)^{\frac{1}{3}} W_{ac}^{-\frac{1}{3}}, \quad ac = LA, TA \tag{18}$$

$$\lambda_{\mathbf{P}_{op}} = -\frac{A^2 \varepsilon_{op}^2}{4\hbar^2 A_1} \frac{y}{W_{op}(W_{op} + Ay\varepsilon_{op})} \mathbf{P}_{op}, \quad op = LO, TO, K, \tag{19}$$

$$\lambda_{\mathbf{Q}_{ac}} = -\frac{2}{3} \left(\frac{4\pi y \zeta(3)}{\hbar^2 c_{ac}^8} W_{ac}^{-4}\right)^{\frac{1}{3}} \mathbf{Q}_{ac}, \quad ac = LA, TA, \tag{20}$$

where $y = \frac{1}{(2\pi)^2}$, $\zeta(\cdot)$ is the zeta function, $A = \frac{8\sqrt{3}}{9}\frac{\pi^2}{a_0^2}$, $A_1 = \frac{20\sqrt{3}}{729}\frac{\pi^4}{a_0^4}$, with $a_0 = 0.142$ nm the nearest neighbor distance between the atoms in graphene.

At equilibrium the phonon temperatures are related to the corresponding Lagrange multipliers by means of

$$T_\mu = \frac{1}{k_B \lambda_{W_\mu}}, \ \mu = LO, TO, K, LA, TA.$$

If we assume that such relations hold even out of equilibrium, the definition of T_L^* can be given in terms of the Lagrangian multipliers as follows.

Definition 3.1 The local temperature of a system of two or more branches of phonons is $T_L^* := \frac{1}{k_B \lambda_{W_L}}$, where λ_{W_L} is the common Lagrange multiplier the occupation numbers of the branches, taken into account, would have if they were in the local thermodynamic equilibrium corresponding to their total energy density, that is

$$W(\lambda_{W_L}) := \sum_\mu W_\mu(\lambda_{W_L}) = \sum_\mu W_\mu(\lambda_{W_\mu}),$$

where the sum is extended to the considered branches and the functions $W_\mu(\lambda_{W_\mu})$ are found from expressions (17)–(18).

In other words, we require that T_L^* is such that by evaluating all the average phonon energy densities with the Lagrange multiplier given by $1/k_B T_L^*$ and by summing up, one gets the value of the total average energy density.

A comparison with experiments is not easy because it is not clear what exactly is measured by the instruments. The comparison between T_L and T_L^* is still under investigation by the authors. The numerical results will be the argument of a forthcoming article.

Acknowledgments The authors acknowledge the support by GNFM (INDAM).

References

1. Lichtenberger, P., Morandi, O., Schürrer, F.: High-field transport and optical phonon scattering in graphene. Phys. Rev. B **84**, 045406 (2011)
2. Barletti, L.: Hydrodynamic equations for electrons in graphene obtained from the maximum entropy principle. J. Math. Phys. **55**, 083303(21) (2014)
3. Morandi, O.: Charge transport and hot-phonon activation in graphene. J. Comput. Theor. Transp. **43**, 162–182 (2014)
4. Coco, M., Majorana, A., Mascali, G., Romano, V.: Comparing kinetic and hydrodynamical models for electron transport in monolayer graphene. In: Schrefler, B., Onate, E., Papadrakakis, M. (eds.) VI International Conference on Computational Methods for Coupled Problems in

Science and Engineering, COUPLED PROBLEMS 2015, Venezia, pp. 1003–1014, 18–20 May 2015

5. Mascali, G., Romano, V.: Charge transport in graphene including thermal effects. SIAM J. Appl. Math. **22**, 593–613 (2014)
6. Lugli, P., Ferry, D.K.: Degeneracy in the ensemble Monte Carlo method for high-field transport in semiconductors. IEEE Trans. Electron Devices **32**(11), 2431–2437 (1985)
7. Romano, V., Majorana, A., Coco, M.: DSMC method consistent with the Pauli exclusion principle and comparison with deterministic solutions for charge transport in graphene. J. Comput. Phys. **302**, 267–284 (2015)
8. Coco, M., Majorana, A., Romano, V.: Cross validation of discontinuous Galerkin method and Monte Carlo simulations of charge transport in graphene on substrate. Ricerche Math. **66**, 201–2020 (2017)
9. Coco, M., Majorana, A., Nastasi, G., Romano, V.: High-field mobility in graphene on substrate with a proper inclusion of the Pauli exclusion principle. Atti della Accademia Peloritana dei Pericolanti, Classe di Scienze Fisiche, Matematiche e Naturali **97**(S1), A6 (2019). ISSN 1825-1242, https://doi.org/10.1478/AAPP.97S1A6
10. Coco, M., Romano, V.: Assessment of the constant phonon relaxation time approximation in electron-phonon coupling in graphene. J. Comput. Theor. Transp. **47**(1–3), 246–266 (2018). https://doi.org/10.1080/23324309.2018.1558253
11. Coco, M., Romano, V.: Simulation of electron-phonon coupling and heating dynamics in suspended monolayer graphene including all the phonon branches. J. Heat Transfer. **140**(9), 092404-092404-10 (2018). https://doi.org/10.1115/1.4040082
12. Castro Neto, A.H., Guinea, F., Peres, N.M.R., Novoselov, K.S., Geim, A.K.: The electronic properties of graphene. Rev. Mod. Phys. **81**, 109–162 (2009)
13. Mascali, G.: A hydrodynamic model for silicon semiconductors including crystal heating. Eur. J. Appl. Math. **26**, 447–496 (2015)
14. Müller, I., Ruggeri, T.: Rational Extended Thermodynamics. Springer, Berlin (1998)
15. Jou, D., Lebon, G., Casas-Vazquez, J.: Extended Irreversible Thermodynamics. Springer, Berlin (2009)
16. Mascali, G., Romano, V.: Exploitation of the maximum entropy principle in mathematical modeling of charge transport in semiconductors. Entropy **19**, 36 (2017). https://doi.org/10.3390/e19010036
17. Coco, M., Mascali, G., Romano, V.: Monte Carlo analysis of thermal effects in monolayer graphene. J. Comput. Theor. Transp. **45**(7), 540–553 (2016)

Monte Carlo Simulation of Electron-Electron Interactions in Bulk Silicon

Guillermo Indalecio and Hans Kosina

Abstract We have developed a novel Monte Carlo (MC) algorithm to study carrier transport in semiconductors in the presence of electron-electron scattering (EES). It is well known that the Boltzmann scattering operator for EES is nonlinear in the single-particle distribution function. Numerical solution methods of the resulting nonlinear Boltzmann equation are usually based on more or less severe approximations. In terms of the pair distribution function, however, the scattering operator is linear. We formulate a kinetic equation for the pair distribution function and related MC algorithms for its numerical solution. Assuming a spatially homogeneous system we derived a two-particle MC algorithm for the stationary problem and an ensemble MC algorithm for the transient problem. Both algorithms were implemented and tested for bulk silicon. As a transient problem we analyzed the mixing of a hot and a cold carrier ensemble. The energy of the hot ensemble relaxes faster with EES switched on. The cold ensemble is temporarily heated by the energy transferred from the hot ensemble. Switching on the electric field rapidly is known to result in an velocity overshoot. We observe that EES enhances the overshoot. The stationary algorithm was used to calculate the energy distribution functions at different field strengths.

1 Introduction

It is commonly accepted that EES alters the high-energy tail of the energy distribution function in a semiconductor device [2]. Since physical models of hot carrier degradation rely on accurate distribution functions as an input it is important

G. Indalecio
CITIUS, University of Santiago de Compostela, Santiago de Compostela, Spain
e-mail: guillermo.indalecio@usc.es

H. Kosina (✉)
Institute for Microelectronics, TU Wien, Vienna, Austria
e-mail: kosina@iue.tuwien.ac.at

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_12

to model EES carefully [5]. In this work we present results of a novel treatment of EES that avoids several of the commonly made approximations.

2 Theory

We study the position-independent case ($\nabla_{\mathbf{r}} f \equiv 0$). Setting $\mathbf{F} = e\mathbf{E}/\hbar$, the Boltzmann equation takes the form

$$\left(\frac{\partial}{\partial t} + \mathbf{F} \cdot \nabla_{\mathbf{k}_1}\right) f(\mathbf{k}_1, t) = Q_{\text{ph}}[f](\mathbf{k}_1, t) + Q_{\text{ee}}[f](\mathbf{k}_1, t) \tag{1}$$

Here, f is the single-particle distribution function, and Q_{ph} denotes the electron-phonon scattering operator. EES is described by the following, nonlinear integral operator.

$$Q_{\text{ee}}[f](\mathbf{k}_1, t) = \int dk_1'\, dk_2'\, dk_2\, S(\mathbf{k}_1, \mathbf{k}_2; \mathbf{k}_1', \mathbf{k}_2') \times \left[f(\mathbf{k}_1', t) f(\mathbf{k}_2', t) - f(\mathbf{k}_1, t) f(\mathbf{k}_2, t)\right] \tag{2}$$

Integration is over all initial states ($\mathbf{k}_2$) and final states ($\mathbf{k}_2'$) of the partner electron and all final states ($\mathbf{k}_1'$) of the sample electron. We restrict our discussion to the non-degenerate case where in the scattering operator (2) Pauli blocking factors of the form $[1 - f(\mathbf{k}, t)]$ are not included.

The two-particle transition rate S_{ee} is derived for a screened Coulomb potential using Fermi's Golden rule [6, 8].

$$S_{\text{ee}}(\mathbf{k}_1, \mathbf{k}_2; \mathbf{k}_1', \mathbf{k}_2') = \frac{e^4 n}{\hbar (2\pi \varepsilon_0 \varepsilon_s)^2} \frac{\delta(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_1' - \mathbf{k}_2')}{(|\mathbf{k}_1 - \mathbf{k}_1'|^2 + \beta_s^2)^2} \times \delta\left[\epsilon(\mathbf{k}_1') + \epsilon(\mathbf{k}_2') - \epsilon(\mathbf{k}_1) - \epsilon(\mathbf{k}_2)\right]$$

The two δ-functions state conservation of momentum and conservation of energy, respectively. In accordance with the principle of detailed balance for energy-conserving transitions, the transition rate is symmetric:

$$S_{\text{ee}}(\mathbf{k}_1, \mathbf{k}_2; \mathbf{k}_1', \mathbf{k}_2') = S_{\text{ee}}(\mathbf{k}_1', \mathbf{k}_2'; \mathbf{k}_1, \mathbf{k}_2)$$

The total scattering rate is obtained by integration over all final states:

$$\Gamma_{\text{ee}}(\mathbf{k}_1, \mathbf{k}_2) = \int dk_1'\, dk_2'\, S_{\text{ee}}(\mathbf{k}_1, \mathbf{k}_2; \mathbf{k}_1', \mathbf{k}_2')$$

One out of the two integrals can be readily evaluated by means of the momentum-conserving δ-function. Assuming a parabolic and isotropic dispersion relation $\epsilon(\mathbf{k})$ which is characterized by the effective mass m^*, the remaining integral can be evaluated analytically.

$$\Gamma_{\mathrm{ee}}(\mathbf{k}_1, \mathbf{k}_2) = \frac{ne^4 m^*}{4\pi\hbar^3(\varepsilon_0\varepsilon_s)^2\beta_s^2} \frac{|\mathbf{k}_2 - \mathbf{k}_1|}{|\mathbf{k}_2 - \mathbf{k}_1|^2 + \beta_s^2}$$

Here, n denotes the electron concentration and β_s the Debye-Hueckel screening parameter. In the presence of EES the Boltzmann equation (1) is nonlinear. Its numerical solution typically requires some iterative method [2, 7].

In this work we chose a different approach. When changing from a single-particle description to a two-particle description the transport equation becomes linear. This step is formally accomplished by replacing the product of two distribution functions by the two-particle distribution function g.

$$f(\mathbf{k}_1, t)f(\mathbf{k}_2, t) \to g(\mathbf{k}_1, \mathbf{k}_2, t)$$

From (1) one can derive a Boltzmann-like kinetic equation for the two-particle distribution function g, which is posed in the six-dimensional momentum space $(\mathbf{k}_1, \mathbf{k}_2)$.

$$\left(\frac{\partial}{\partial t} + \mathbf{F}\cdot\nabla_{\mathbf{k}_1} + \mathbf{F}\cdot\nabla_{\mathbf{k}_2}\right) g(\mathbf{k}_1, \mathbf{k}_2, t) = Q_{\mathrm{ph}}[g](\mathbf{k}_1, \mathbf{k}_2, t) + Q_{\mathrm{ee}}[g](\mathbf{k}_1, \mathbf{k}_2, t) \qquad (3)$$

In the two-particle picture, the nonlinear operator (2) turns into a linear integral operator.

$$Q_{\mathrm{ee}}[g](\mathbf{k}_1, \mathbf{k}_2, t) = 2\int dk_1'\, dk_2'\, S_{\mathrm{ee}}(\mathbf{k}_1, \mathbf{k}_2; \mathbf{k}_1', \mathbf{k}_2')\big[g(\mathbf{k}_1', \mathbf{k}_2', t) - g(\mathbf{k}_1, \mathbf{k}_2, t)\big] \qquad (4)$$

Details about the derivation of (3) and (4) will be presented in a forthcoming publication.

The linear kinetic equation (3) can be transformed into an integral equation of the following form.

$$g(x) = \int g(x')\, K(x', x)\, dx' + g_0(x), \qquad x \equiv (\mathbf{k}_1, \mathbf{k}_2, t) \qquad (5)$$

In this derivation, the very same steps as in the case of the Boltzmann equation are applied [4]. Using the formalism described in [4] we derive a stationary and a transient Monte Carlo algorithm for the solution of the integral equation (5).

In the stationary algorithm, the trajectories of a pair of particles are calculated over a long period of time. Electron-phonon scattering events of the two particles are independent from each other and treated as in the case of the Boltzmann equation. An EES event, however, changes the states of the two particles simultaneously, whereby total momentum and energy are exactly conserved. Averages can be computed using the before-scattering method [3].

In the transient algorithm, an ensemble of trajectory pairs is simulated, starting from a given two-particle initial distribution. Averages are computed as ensemble averages at given points in time.

3 Results and Discussion

In the following simulations an electron concentration of 10^{19} cm^{-3} and a lattice temperature of 300 K are assumed. First we apply the stationary MC algorithm to calculate the momentum distribution functions at different field strengths. In accordance with thermodynamics, in equilibrium a Maxwellian distribution is obtained. EES has no effect on the equilibrium distribution, see Figs. 1 and 2. At 77 K, EES causes a broadening of the non-equilibrium distribution (Fig. 1), whereas as 300 K such a broadening is not observed (Fig. 2). The reason is that at 300 K phonon scattering is much stronger and the relative importance of EES is small.

Figure 3 shows how an ensemble of hot electrons gets cooled down when interacting with the phonons of the crystal lattice and additionally with an electron ensemble at lattice temperature. The mean energy of the hot electrons relaxes faster when EES is present. The mean energy of the cold electrons is temporarily increased by the energy transfer from the hot carriers. Averages are calculated by sampling the two ensembles at equidistant time steps. The number of particle pairs simulated is $2 \cdot 10^4$.

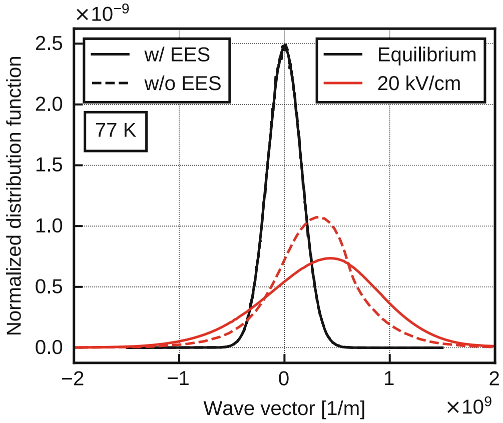


Fig. 1 Momentum distribution functions at equilibrium and at 20 kV/cm, lattice temperature 77 K

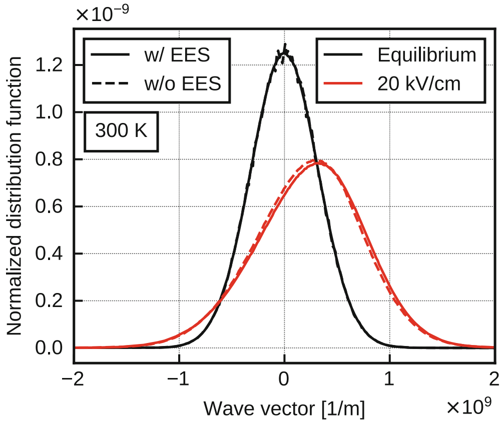


Fig. 2 Momentum distribution functions at equilibrium and at 20 kV/cm, lattice temperature 300 K

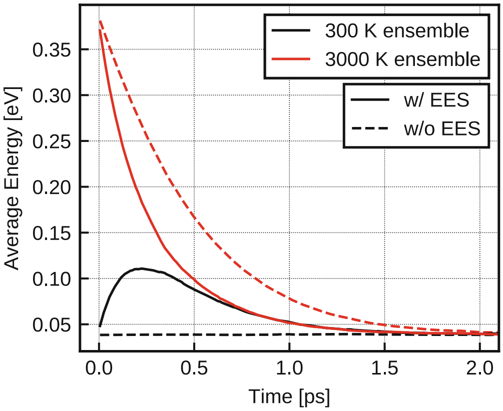
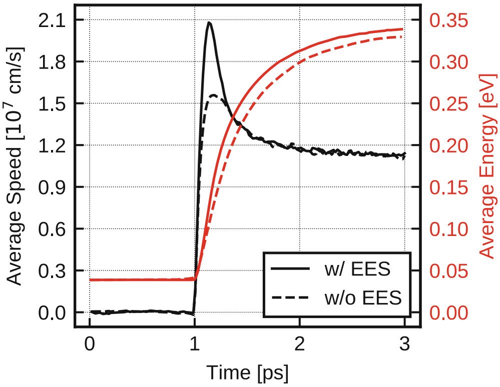


Fig. 3 Relaxation of the mean energy is affected by EES. The initial two-particle distribution function assumed consists of a hot ensemble at 3000 K and a cold ensemble at 300 K

Another application of the transient MC algorithm is the study of the response of the carriers to an abrupt change in the electric field. At 1 ps a field step of 50 kV/cm is applied. During a short period after the field step the carriers experience the high electric field and are accelerated accordingly, whereas the mean energy and thus momentum relaxation due the electron-phonon scattering is still low. In this situation a phenomenon known as the velocity overshoot occurs [1]. Our results indicate that the velocity overshoot even gets enhanced by EES as shown in Fig. 4. We believe that the enhancement in the overshoot can be explained as follows. Two particles entering the high field region experience an EES event where momentum is transferred from one particle to the other. If the momentum transfer is largely oriented along the field direction, one electron gains velocity and the other one is slowed down. The low energetic electron, however, experiences a small electron-phonon scattering rate and has a higher probability to stay in the high field without scattering, so that it will also have a large momentum gain from the field. Therefore, both electrons involved in the EES event eventually reach a higher velocity than they would without the EES event. Figure 4 also shows that EES gives a faster rise of the

Fig. 4 Velocity overshoot (left, black) and energy transient (right, red) after applying an electric field step of 50 kV/cm at 1 ps



mean energy towards the stationary value. Again, in the simulation we sampled an ensemble of $2 \cdot 10^4$ particle pairs at equidistant time steps.

4 Conclusions

We have developed a two-particle Monte Carlo algorithm for the solution of a two-particle kinetic equation that includes electron-electron scattering.

We demonstrate the impact of electron-electron scattering on the transient relaxation of an ensemble of hot carriers, on the velocity overshoot in the presence of a field step, and on the shape the momentum distribution function.

Acknowledgments G.I. acknowledges support by the Programa de Axudas a Etapa Posdoutoral da Xunta de Galicia under grant number 2017/077, and H.K. support by the Austrian Research Promotion Agency (FFG), project MORAFLASH (contract No. 850660).

References

1. Baccarani, G., Wordeman, M.R.: An investigation of steady-state velocity overshoot in silicon. Solid State Electron. **28**(4), 407–416 (1985)
2. Childs, P.A., Leung, C.C.C.: A one-dimensional solution of the Boltzmann transport equation including electron-electron interactions. J. Appl. Phys. **79**, 222 (1996)
3. Jacoboni, C., Lugli, P.: The Monte Carlo Method for Semiconductor Device Simulation. Springer, Wien (1989)
4. Kosina, H., Nedjalkov, M., Selberherr, S.: Theory of the Monte Carlo method for semiconductor device simulation. IEEE Trans. Electron Devices **47**(10), 1898–1908 (2000)
5. Rauch, S.E., La Rosa, G., Guarin, F.J.: Role of e-e scattering in the enhancement of channel hot carrier degradation of deep-submicron NMOSFETs at high V_{GS} conditions. IEEE Trans. Device Mater. Reliab. **1**(2), 113–119 (2001)
6. Ridely, B.K.: Quantum Processes in Semiconductors. Oxford University Press, Oxford (2013)

7. Takenaka, N., Inoue, M., Inuishi, Y.: Influence of inter-carrier scattering on hot electron distribution function in GaAs. J. Phys. Soc. Jpn. **47**(3), 861–868 (1979)
8. Tomizawa, K.: Numerical Simulation of Submicron Semiconductor Devices. Artech House, Boston (1993)

Semi-classical and Quantum Hydrodynamic Modeling of Electron Transport in Graphene

Liliana Luca and Vittorio Romano

Abstract The present work aims at formulating hydrodynamic models for a proper description of charge transport in graphene, which is extremely important for growing technological development in CAD tools. The analysis is carried out in two different steps. Initially a semi-classical hydrodynamic model is developed starting from the moment system associated with Boltzmann equation and obtaining the closure relations with the Maximum Entropy Principle. At this level quantum effects are neglected. In the second step the model previously developed is extended to include quantum effects by incorporating the first order quantum corrections. To asses the validity of this model numerical simulations are under current investigation.

1 Introduction

Graphene, a monolayer of sp^2-bonded carbon atoms with zero band gap, is not only the basis for graphite but also a new material with immense potential in microelectronics for its exceptional electrical transport properties, like high conductivity and high charge mobility. As a result of its promising properties, it seems to be an ideal candidate to take over from silicon for the next generation of faster and smaller electronic devices.

To deal with the basic kinetic transport equations remains too expensive for real life applications. Nevertheless from transport equations it is possible to derive simpler fluid dynamic equations for macroscopic quantities like particle, velocity, or energy densities. They represent a good compromise between physical accuracy and computational cost.

A standard approach to derive macroscopic models, like drift-diffusion, energy transport or hydrodynamic ones, is the moments method. The present work aims at

L. Luca · V. Romano (✉)
Department of Mathematics and Computer Science, Università degli Studi di Catania, Catania, Italy
e-mail: liliana.luca@unict.it; romano@dmi.unict.it

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_13

formulating hydrodynamic models of this type for a proper description of charge transport in graphene, which is extremely important for growing technological development in CAD tools.

The plan of the paper is as follows.

Section 2 focuses on the formulation of a semi-classical hydrodynamic model based on the Maximum Entropy Principle (MEP) without taking into account quantum effects. The model analyzed can be developed by taking into account fully non linear closure relations [11], or its linearized version [2, 12].

To take into account quantum phenomena, in the second section a quantum hydrodynamic model for charge transport in graphene is derived from a moment expansion of the Wigner-Boltzmann equation and the needed closure relations are obtained by adding quantum corrections based on the equilibrium Wigner function to the semiclassical model formulated in [2, 11, 12] by exploiting the Maximum Entropy Principle. The expression of the equilibrium Wigner function which takes into account the energy band of graphene has been obtained by solving the corresponding Bloch equation (see also [1, 17]). In other terms, the strategy adopted for formulating these models combines quantum and semi-classical approaches as shown in Fig. 1.

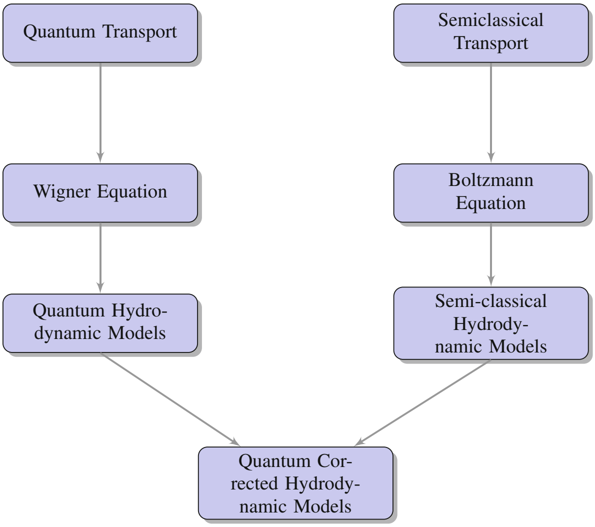

Quantum Transport
Semiclassical Transport
Wigner Equation
Boltzmann Equation
Quantum Hydrodynamic Models
Semi-classical Hydrodynamic Models
Quantum Corrected Hydrodynamic Models

Fig. 1 Schematic representation of the strategy adopted for developing Quantum Corrected Hydrodynamic models

2 A Semi-classical Hydrodynamic Model

The starting point for the derivation of semi-classical hydrodynamic models is the semi-classical Boltzmann equation

$$\frac{\partial f(\mathbf{r},\mathbf{k},t)}{\partial t} + \mathbf{v}\cdot\nabla_{\mathbf{r}} f(\mathbf{r},\mathbf{k},t) + \frac{q}{\hbar}\nabla_{\mathbf{r}}\Phi(\mathbf{r})\cdot\nabla_{\mathbf{k}} f(\mathbf{r},\mathbf{k},t) = \mathcal{C}[f](\mathbf{r},\mathbf{k},t) \quad (1)$$

where $f(\mathbf{r},\mathbf{k},t)$ is the distribution of electrons in the conduction or valence band (the dependence from the Dirac point is omitted), $\nabla_{\mathbf{r}}$ and $\nabla_{\mathbf{k}}$ are the gradients with respect to the space variable $\mathbf{r}$ and wave vector respectively, q is the elementary (positive) charge, $\hbar$ is the reduced Planck's constant, Φ is the electric potential and $\mathbf{v}$ is the microscopic velocity which is related to the energy band by $\mathbf{v}(\mathbf{k}) = \pm\frac{1}{\hbar}\nabla_{\mathbf{k}}\mathcal{E}(\mathbf{k})$. The positive sign refers to the conductions band, the negative sign to the valence one. $\mathcal{C}$ is the collision term representing the interactions of electrons with acoustic (ac) phonons, longitudinal (LO) and transversal (TO) optical phonons and K–phonons (for more details see [2, 11]).

Numerical solutions of Eq. (1) can be obtained, for example, via Direct Monte Carlo Simulation (DSMC)[5, 22] or by finite difference schemes [10] or by discontinuous Galerkin (DG) methods [5]. However, these simulations have been obtained for simple cases such as pristine graphene under the effect of a constant external electric field. With a view of more complex situations, like those represented by a metal-oxide-semiconductor field-effect transistor (MOSFET) with a graphene channel, it is better to benefit from simpler models like drift-diffusion, energy transport or hydrodynamic ones. These directly provide balance equations for macroscopic quantities like electron density, average velocity or current, average energy, etc., and, therefore, are more suited as models for CAD tools.

The macroscopic quantities are related to the distribution function because they represent average values of some functions of the wave vector $\mathbf{k}$. For example, the density $n(\mathbf{r},t)$ is given by

$$n(\mathbf{r},t) = \frac{2}{(2\pi)^2}\int_{\mathbb{R}^2} f(\mathbf{r},\mathbf{k},t)d^2\mathbf{k}.$$

Similarly the average energy $W(\mathbf{r},t)$ is given by the relation

$$n(\mathbf{r},t)W(\mathbf{r},t) = \frac{2}{(2\pi)^2}\int_{\mathbb{R}^2} f(\mathbf{r},\mathbf{k},t)\mathcal{E}(\mathbf{k})d^2\mathbf{k}.$$

Generally speaking, given a weight function $\psi(\mathbf{k})$, the corresponding macroscopic quantity is the expectation value

$$M(\mathbf{r},t) = \frac{2}{(2\pi)^2}\int_{\mathbb{R}^2} \psi(\mathbf{k})\, f(\mathbf{r},\mathbf{k},t)d^2\mathbf{k}.$$

The evolution equation for $M(\mathbf{r}, t)$ is deduced by multiplying Eq. (1) for $\psi(\mathbf{k})$ and by integrating with respect to $\mathbf{k}$

$$\begin{aligned}\frac{\partial M}{\partial t} + \nabla_{\mathbf{r}} \cdot \int_{\mathbb{R}^2} f \, \frac{2\psi(\mathbf{k})}{(2\pi)^2} \, \mathbf{v}(\mathbf{k}) d^2\mathbf{k} - \frac{q}{\hbar} \mathbf{E} \cdot \int_{\mathbb{R}^2} f \nabla_{\mathbf{k}} \, \frac{2\psi(\mathbf{k})}{(2\pi)^2} \, d^2\mathbf{k} \\ = \int_{\mathbb{R}^2} \frac{2\psi(\mathbf{k})}{(2\pi)^2} \, C[f] d^2\mathbf{k}. \end{aligned} \tag{2}$$

Note that the moment equations depend only on the independent variables $\mathbf{r}$, t. This considerably reduces the numerical complexity.

The main issue related to any model based on balance equations deduced as moment equations of type (2) is that there are more unknowns than introduced moments in the evolution equations, and the so-called *closure problem* arises. This comes from expressing the additional unknowns, that is the extra fluxes and production terms

$$\int_{\mathbb{R}^2} f \, \frac{2\psi(\mathbf{k})}{(2\pi)^2} \, \mathbf{v}(\mathbf{k}) d^2\mathbf{k}, \quad \int_{\mathbb{R}^2} f \nabla_{\mathbf{k}} \, \frac{2\psi(\mathbf{k})}{(2\pi)^2} \, d^2\mathbf{k}, \quad \int_{\mathbb{R}^2} \frac{2\psi(\mathbf{k})}{(2\pi)^2} \, C[f] d^2\mathbf{k},$$

as functions of the basic moments.

A systematic way to get the needed closure relations is employing the Maximum Entropy Principle (MEP). It is based on the information theory of Shannon and was devised for application in statistical physics by Jaynes [8] (for a general review of the application of MEP to semiconductors the interested reader is referred to [15]). The central idea of this principle is to predict the distribution of the microstates, which are the particle of the system, on the basis of the knowledge of some macroscopic data. The latter information is specified in the form of some simple moment constraints. Therefore the distribution obtained with MEP is the least biased estimator from the knowledge of a finite number of expectation values.

Let us suppose that a certain number of moments $M_A(\mathbf{r}, t)$, $A = 1, 2, \cdots, N$, relative to the weight functions $\psi_A(\mathbf{k})$, are known. According to MEP, the electron distribution function is estimated with the distribution f_{MEP} obtained by solving the following constrained optimization problem: for fixed $\mathbf{r}$ and t,

$$\max_{f \in \mathcal{F}} S[f] \quad \text{subject to the constraints:}$$

$$0 < f < 1, \tag{3}$$

$$M_A = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} \psi_A(\mathbf{k}) \, f(\mathbf{r}, \mathbf{k}, t) d^2\mathbf{k}, \quad A = 1, 2, \cdots, N, \tag{4}$$

where $S[f]$ is the entropy of the system, which in the semi-classical approximation reads

$$S[f] = -2 \frac{2k_B}{(2\pi)^2} \int_{\mathbb{R}^2} [f \ln f + (1 - f) \ln (1 - f)] \, d^2\,\mathbf{k}.$$

Factor 2 is included to take into account the valley degeneracy. $\mathcal{F}$ is the space of the function $g(\mathbf{k})$ such that $\psi_A(\mathbf{k})g(\mathbf{k}) \in L^1(\mathbb{R}^2)$ for $A = 1, 2, \cdots, N$.

Here with $L^1(\mathbb{R}^2)$ we have denoted the usual Banach space of the summable functions defined over $\mathbb{R}^2$.

To take into account bilateral constraints let us introduce the Lagrange multipliers λ_A, $A = 1, 2, \cdots, N$, and the Legendre transform of S

$$S' = S + \sum_A \lambda_A \left(M_A - \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f \psi_A(\mathbf{k})\, d^2\mathbf{k} \right).$$

Let the variation of S' with respect to f be zero, i.e.,

$$0 = \delta S' = -2 \frac{2k_B}{(2\pi)^2} \int_{\mathbb{R}^2} \left[\ln f - \ln(1-f) + \frac{1}{2\,k_B} \sum_A \psi_A(\mathbf{k}) \lambda_A \right] \delta f \, d^2\,\mathbf{k}.$$

Since δf is arbitrary, the quantity in the square brackets must be zero; we get

$$f_{MEP}(\mathbf{r}, \mathbf{k}, t) = \frac{1}{1 + \exp\left[\sum_A \psi_A(\mathbf{k}) \lambda_A(\mathbf{r}, t)\right]},$$

which also fulfills the unilateral constraints (3).

The multiplicative constant $\frac{1}{2\,k_B}$ has been included into the multipliers for simplicity.

To complete the optimization procedure, it is necessary to invert the relations (4) and express the Lagrangian multipliers as functions of the basic variables. This can generally be achieved only numerically or by some approximation, e.g. expanding around the equilibrium state.

The above problem of inversion apart, once one gets f_{MEP}, the needed closure relations are obtained by evaluating the extra fluxes and production terms with f_{MEP} instead of f. At equilibrium the distribution of electrons, both in the conduction and valence band, are given by the Fermi-Dirac distribution

$$f_{FD}(\mathbf{r}, \mathbf{k}, t) = \frac{1}{1 + \exp\left(\dfrac{\mathcal{E}(\mathbf{k}) - \varepsilon_F}{k_B T_L} \right)}, \quad -\infty < \mathcal{E}(\mathbf{k}) < +\infty, \tag{5}$$

where ε_F is the Fermi energy, T_L being the lattice temperature and k_B the Boltzmann constant.

In this work we consider a 6-moment model obtained by choosing as weight functions $\{1, \mathcal{E}, \mathbf{v}, \mathcal{E}\mathbf{v}\}$ to which the following average quantities in the conduction bands (similar results hod for the valence band) correspond

$$n(\mathbf{r}, t) = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(\mathbf{r}, \mathbf{k}, t)\, d\mathbf{k} \quad \text{density}, \tag{6}$$

$$n(\mathbf{r}, t) W(\mathbf{r}, t) = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(\mathbf{r}, \mathbf{k}, t) \mathcal{E}(\mathbf{k})\, d\mathbf{k} \quad \text{energy density}, \tag{7}$$

$$n(\mathbf{r}, t) \mathbf{V}(\mathbf{r}, t) = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(\mathbf{r}, \mathbf{k}, t) \mathbf{v}(\mathbf{k})\, d\mathbf{k} \quad \text{linear momentum density}, \tag{8}$$

$$n(\mathbf{r}, t) \mathbf{S}(\mathbf{r}, t) = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(\mathbf{r}, \mathbf{k}, t) \mathcal{E}(\mathbf{k}) \mathbf{v}(\mathbf{k})\, d\mathbf{k} \quad \text{energy-flux density}. \tag{9}$$

The corresponding evolution equations, in the unipolar case, are given by

$$\begin{aligned}
&\frac{\partial}{\partial t} n + \nabla_{\mathbf{r}} (n\, \mathbf{V}) = 0, \\
&\frac{\partial}{\partial t} (n\, W) + \nabla_{\mathbf{r}} (n\, \mathbf{S}) + q\, n\, \mathbf{E} \cdot \mathbf{V} = n\, C_W, \\
&\frac{\partial}{\partial t} (n\, \mathbf{V}) + \nabla_{\mathbf{r}} \left(n\, \mathbf{F}^{(0)} \right) + q\, n\, \mathbf{G}^{(0)} : \mathbf{E} = n\, C_{\mathbf{V}}, \\
&\frac{\partial}{\partial t} (n\, \mathbf{S}) + \nabla_{\mathbf{r}} \left(n\, \mathbf{F}^{(1)} \right) + q\, n\, \mathbf{G}^{(1)} : \mathbf{E} = n\, C_{\mathbf{S}}.
\end{aligned}$$

Besides the average densities, velocities, energies and energy fluxes, additional quantities appear[1]

$$\begin{aligned}
n\, C_{\mathbf{V}} &= \frac{2}{(2\,\pi)^2} \int_{\mathbb{R}^2} \mathbf{v}(\mathbf{k})\, \mathcal{C}(\mathbf{k})\, d\,\mathbf{k}, \\
n\, C_W \ &= \frac{2}{(2\,\pi)^2} \int_{\mathbb{R}^2} \mathcal{E}(\mathbf{k})\, \mathcal{C}(\mathbf{k})\, d\,\mathbf{k}, \quad n\, C_{\mathbf{S}} = \frac{2}{(2\,\pi)^2} \int_{\mathbb{R}^2} \mathcal{E}(\mathbf{k})\, \mathbf{v}(\mathbf{k}) \mathcal{C}(\mathbf{k})\, d\,\mathbf{k} \\
n \begin{pmatrix} \mathbf{F}^{(0)} \\ \mathbf{F}^{(1)} \end{pmatrix} &= \frac{2}{(2\,\pi)^2} \int_{\mathbb{R}^2} \begin{pmatrix} 1 \\ \mathcal{E}(\mathbf{k}) \end{pmatrix} \mathbf{v}(\mathbf{k}) \otimes \mathbf{v}(\mathbf{k}) f(\mathbf{r}, \mathbf{k}, t)\, d\mathbf{k}, \\
n \begin{pmatrix} \mathbf{G}^{(0)} \\ \mathbf{G}^{(1)} \end{pmatrix} &= \frac{2}{\hbar\, (2\,\pi)^2} \int_{\mathbb{R}^2} f(\mathbf{r}, \mathbf{k}, t) \nabla_{\mathbf{k}} \begin{pmatrix} \mathbf{v}(\mathbf{k}) \\ \mathcal{E}(\mathbf{k})\, \mathbf{v}(\mathbf{k}) \end{pmatrix} d\mathbf{k},
\end{aligned}$$

that must be expressed as function of the basic variables n, $\mathbf{V}$, W, $\mathbf{S}$.

[1] The symbol $\otimes$ denotes the tensor product of vectors.

Regarding the production terms, they are given by summing the contributions arising from the different types of phonon scattering

$$C_M = C_M^{(ac)} + \sum_{s=LO,TO,K} C_M^{(s)},$$

with $M = \rho, W, \mathbf{V}, \mathbf{S}$.

The following expression of the distribution function deduced by MEP

$$f_{MEP}(\mathbf{r}, \mathbf{k}, t) = \frac{1}{1 + \exp(\lambda(\mathbf{r}, t) + \lambda_w(\mathbf{r}, t)\mathcal{E}(\mathbf{k}) + (\lambda_{\mathbf{V}}(\mathbf{r}, t) + \mathcal{E}(\mathbf{k})\lambda_{\mathbf{S}}(\mathbf{r}, t)) \cdot \mathbf{v}(\mathbf{k}))} \tag{10}$$

has been used in the linearized form[2]

$$f_{MEP}(\mathbf{r}, \mathbf{k}, t) \approx \frac{1}{1 + e^{\lambda + \lambda_w \mathcal{E}}} - \frac{e^{\lambda + \lambda_w \mathcal{E}}}{(1 + e^{\lambda + \lambda_w} \mathcal{E})^2} (\lambda_{\mathbf{V}} + \mathcal{E}\lambda_{\mathbf{S}}) \cdot \mathbf{v}. \tag{11}$$

Explicit closure relation has been obtained in [2] and [16] the crystal heating effects have been also included. Comparisons with Direct Simulation Monte Carlo [4–6, 13, 14, 22] have shown a good accuracy of the model. In the next section the general guideline for getting quantum corrections to the semiclassical hydryodynamic models will be delineated.

3 A 6-Moment Model with Quantum Corrections

To take into account quantum phenomena, the semiclassical Boltzmann equation is not enough to describe charge transport. As a starting point for deriving the quantum corrections to the semiclassical model, we consider the Wigner equation. At zero order we recover the semiclassical models developed in [2, 11, 12, 16] by exploiting the Maximum Entropy Principle (MEP). By following the idea developed in [20] for silicon, $\hbar^2$ order corrections are obtained from the scaling of high field and collision dominated regime. In the limit of high collisional frequency of the quantum correction to the collision operator, this is equivalent to determine the $\hbar^2$ order corrections with the equilibrium Wigner function, similarly to what done in [7]. The problem to find out the equilibrium Wigner function in the case of an arbitrary energy band has been discussed in [21] where the corresponding Bloch equation is written and solved for silicon in the Kane dispersion relation approximation. Here the same approach is used for graphene.

[2] In the following the explicit dependence on $\mathbf{r}, \mathbf{k}, t$ is omitted for the sake of simplifying the notation.

One important issue is related to the conical shape of the energy band around the Dirac points of the first Brillouin zone. This fact makes singular some term of the expansion if a sharp zero gap between the conduction and the valence band is assumed. However, see for example [3], from a theoretical point of view it is possible the presence, although very small, of a gap which is related to the first and second neighbour hopping energy. Therefore, around the Dirac points we employ a regularized energy band. Explicit formulas are obtained and the resulting model is given by a set of dispersive PDEs.

Of course it is also possible to try to numerically solve directly the Wigner equation but major computational difficulties arise and at the present time it seems far from being a standard feasible tool for the design of electron devices. The interested reader can see the monograph [19] and the paper [5, 18, 22] for recent advances of the algorithms in stochastic approaches.

3.1 Wigner Equation

In the proximity of the Dirac points K (K'), which are the vertices of the Brillouin zone, by choosing in the $\mathbf{k}$-space a reference frame centered in the considered Dirac point, the energy dispersion relation can be considered approximately linear with respect to the modulus of the wave-vector $\mathbf{k}$. As already mentioned above it is not clear if a small gap between the conduction and the valence band exists. Therefore we adopt the following regularization

$$\mathcal{E}(\mathbf{k}) = \pm v_F \sqrt{a^2 + p^2},$$

where $p = \hbar|\mathbf{k}|$, $v_F \simeq 1 \times 10^6$ cm/s is the Fermi velocity, $\hbar$ is the reduced Planck's constant. The sign "+" refers to the conduction band while the sign "−" refers to the valence band. a is a small parameter related to the nearest-neighbour hopping energy [3]. To derive a transport equation, we introduce the single electron Wigner quasi-distribution $w(\mathbf{x}, p, t)$, depending on the position $\mathbf{x}$, momentum p and time t. Evolution is governed by the *Wigner-Poisson* system for w and the electrostatic potential Φ

$$\begin{aligned}
&\frac{\partial w(\mathbf{x}, p, t)}{\partial t} + S[\mathcal{E}]w(\mathbf{x}, p, t) - q\theta[\mathcal{E}]w(\mathbf{x}, p, t) = C[w],\\
&\nabla \cdot (\epsilon \nabla \Phi) = -q(N_D - n),
\end{aligned}$$

where q is the elementary (positive) charge, N_D is donor carrier concentration, $C[w]$ is the collision term representing the electron-phonon scattering while $S[\mathcal{E}]$ and $\theta[\mathcal{E}]$

represent the pseudo-differential operators

$$\begin{aligned} S[\mathcal{E}]w(\mathbf{x}, p, t) =& \frac{i}{\hbar(2\pi)^2} \int_{\mathbb{R}^2_{\mathbf{x}'} \times \mathbb{R}^2_{\boldsymbol{\nu}}} [\ \mathcal{E}(p + \frac{\hbar}{2}\boldsymbol{\nu}, t) + \\ & -\ \mathcal{E}(p - \frac{\hbar}{2}\boldsymbol{\nu}, t)] w(\mathbf{x}', p, t) e^{-i(\mathbf{x}'-\mathbf{x})\cdot\boldsymbol{\nu}} d\mathbf{x}' d\boldsymbol{\nu}, \end{aligned}$$

$$\begin{aligned} \theta[\ \mathcal{E}]w(\mathbf{x}, p, t) =& \frac{i}{\hbar(2\pi)^2} \int_{\mathbb{R}^2_{p'} \times \mathbb{R}^2_{\boldsymbol{\eta}}} [\Phi(\mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta}, t) + \\ & - \Phi(\mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta}, t)] w(\mathbf{x}, p', t) e^{i(p'-p)\cdot\boldsymbol{\eta}} dp' d\boldsymbol{\eta}. \end{aligned}$$

In the semiclassical limit $\hbar \to 0$, the Wigner equation reduces to Boltzmann one.

3.2 Equilibrium Wigner Function

If we denote the density matrix at equilibrium by $\rho_{eq}(\mathbf{r}, \mathbf{s}, \beta)$, it satisfies the Bloch equation

$$\frac{\partial\, \rho_{eq}(\mathbf{r}, \mathbf{s}, \beta)}{\partial\beta} = -\frac{1}{2}[H_{\mathbf{r}}\rho_{eq}(\mathbf{r}, \mathbf{s}, \beta) + H_{\mathbf{s}}\rho_{eq}(\mathbf{r}, \mathbf{s}, \beta)].$$

Applying the Fourier transform to the Bloch equation, we get

$$\begin{aligned} \frac{\partial w_{eq}(\mathbf{x}, \mathbf{p}, \beta)}{\partial\beta} = -\frac{1}{2}\Bigg\{ & \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2_{\mathbf{x}'} \times \mathbb{R}^2_{\nu}} \mathcal{E}\left(\mathbf{p} + \frac{\hbar}{2}\nu\right) + \\ & + \mathcal{E}\left(\mathbf{p} - \frac{\hbar}{2}\nu\right) w_{eq}(\mathbf{x}', \mathbf{p}, \beta) e^{-i(\mathbf{x}'-\mathbf{x})\cdot\nu} d\mathbf{x}'\, d\nu - \\ & - \frac{q}{(2\pi)^2} \int_{\mathbb{R}^2_{\mathbf{p}'} \times \mathbb{R}^2_{\eta}} \Phi\left(\mathbf{x} + \frac{\hbar}{2}\eta\right) + \Phi\left(\mathbf{x} - \frac{\hbar}{2}\eta\right) w_{eq}(\mathbf{x}, \mathbf{p}', \beta) e^{i(\mathbf{p}'-\mathbf{p})\cdot\eta} d\mathbf{p}'\, d\eta \Bigg\}, \end{aligned}$$

where $w_{eq}(\mathbf{x}, \mathbf{p}, \beta)$ is the equilibrium Wigner function. We looked for solution of the type

$$w_{eq}(\mathbf{x}, \mathbf{p}, \beta) = w^{(0)}_{eq}(\mathbf{x}, \mathbf{p}, \beta) + \hbar^2 w^{(1)}_{eq}(\mathbf{x}, \mathbf{p}, \beta).$$

After some algebra we get the equilibrium Wigner function

$$w_{eq}(\mathbf{x}, p, \beta) = \exp(q\Phi(\mathbf{x})\beta)\exp(-\beta\mathcal{E}(p))\left\{1 + \frac{q\beta^2\hbar^2}{8}\frac{\partial^2\mathcal{E}(p)}{\partial p_i\partial p_j}\frac{\partial^2\Phi(\mathbf{x})}{\partial x_i\partial x_j} + \right.$$
$$\left. + \frac{\beta^3\hbar^2}{24}\left[q^2\frac{\partial^2\ \mathcal{E}(p)}{\partial p_i\partial p_j}\frac{\partial\Phi(\mathbf{x})}{\partial x_i}\frac{\partial\Phi(\mathbf{x})}{\partial x_j} - q\frac{\partial^2\Phi(\mathbf{x})}{\partial x_i\partial x_j}v_i v_j\right]\right\} + o(\hbar^2).$$

3.3 Structure of the Model

Supposing the expansion

$$w = w^{(0)} + \hbar^2 w^{(1)} + O(\hbar^4)$$

holds and by proceeding formally, as $\hbar \longrightarrow 0$ the Wigner equation gives the semiclassical Boltzmann equation. Therefore we identify $w^{(0)}(\mathbf{x}, p, t)$ with the semiclassical distribution which has been approximated in [11, 12] with the maximum entropy principle estimator $w^{(0)}(\mathbf{x}, p, t) \approx f_{MEP}(\mathbf{x}, p, t)$.

At first order in $\hbar^2$ one finds

$$\frac{\partial w^{(1)}(\mathbf{x}, p, t)}{\partial t} + \mathbf{v}\cdot\nabla_{\mathbf{x}} w^{(1)}(\mathbf{x}, p, t) - \frac{1}{24}\frac{\partial^3\mathcal{E}(p)}{\partial p_i\partial p_j\partial p_k}\frac{\partial^3 w^{(0)}(\mathbf{x}, p, t)}{\partial x_i\partial x_j\partial x_k} +$$
$$+ q\nabla_{\mathbf{x}}\Phi(\mathbf{x})\nabla_p w^{(1)}(\mathbf{x}, p, t) - \frac{q}{24}\frac{\partial^3\Phi(\mathbf{x})}{\partial x_i\partial x_j\partial x_k}\frac{\partial^3 w^{(0)}(\mathbf{x}, p, t)}{\partial p_i\partial p_j\partial p_k} = \mathcal{C}[w^{(1)}]$$

Hereafter, suppose $w^{(1)} = w^{(1)}_{eq}$.

As an example, consider a 6-Moment Model based on the following moments

$$\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2} w(\mathbf{x}, p, t)dp = n(\mathbf{x}, t) \quad \text{density},$$
$$\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2} w(\mathbf{x}, p, t)\mathcal{E}(p)dp = n(\mathbf{x}, t)W \quad \text{energy density},$$
$$\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2} w(\mathbf{x}, p, t)\mathbf{v}(p)dp = n(\mathbf{x}, t)\mathbf{V} \quad \text{linear momentum density},$$
$$\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2} w(\mathbf{x}, p, t)\mathcal{E}(p)\mathbf{v}(p)dp = n(\mathbf{x}, t)\mathbf{S} \quad \text{energy-flux density}.$$

The corresponding evolution equations are given by[3]

$$\frac{\partial}{\partial t}n(\mathbf{x},t)+\frac{\partial}{\partial x_i}\left(n(\mathbf{x},t)V_i-\frac{\hbar^2}{24}\frac{\partial^2\left(n(\mathbf{x},t)T^{(0)}_{ijk}\right)}{\partial x_j\partial x_k}\right)=0, \tag{12}$$

$$\frac{\partial}{\partial t}\left(n(\mathbf{x},t)W\right)+\frac{\partial}{\partial x_i}\left(n(\mathbf{x},t)S_i-\frac{\hbar^2}{24}\frac{\partial^2\left(n(\mathbf{x},t)T^{(1)}_{ijk}\right)}{\partial x_j\partial x_k}\right)-$$

$$-q\frac{\partial}{\partial x_i}\Phi(\mathbf{x})\cdot n(\mathbf{x},t)V_i+\frac{q\hbar^2}{24}\frac{\partial^3\Phi(\mathbf{x})}{\partial x_i\partial x_j\partial x_k}n(\mathbf{x},t)T^{(0)}_{ijk}=C_W[w^{(0)}], \tag{13}$$

$$\frac{\partial}{\partial t}\left(n(\mathbf{x},t)V_i\right)+\frac{\partial}{\partial x_j}\left(n(\mathbf{x},t)F^{(0)}_{ij}-\frac{\hbar^2}{24}\frac{\partial^2\left(n(\mathbf{x},t)H^{(0)}_{ijkl}\right)}{\partial x_k\partial x_l}\right)-$$

$$-q\frac{\partial}{\partial x_j}\Phi(\mathbf{x})\cdot n(\mathbf{x},t)G^{(0)}_{ij}+\frac{q\hbar^2}{24}\frac{\partial^3\Phi(\mathbf{x})}{\partial x_j\partial x_k\partial x_l}n(\mathbf{x},t)L^{(0)}_{ijkl}=C_{V_i}[w^{(0)}], \tag{14}$$

$$\frac{\partial}{\partial t}\left(n(\mathbf{x},t)S_i\right)+\frac{\partial}{\partial x_j}\left(n(\mathbf{x},t)F^{(1)}_{ij}-\frac{\hbar^2}{24}\frac{\partial^2\left(n(\mathbf{x},t)H^{(1)}_{ijkl}\right)}{\partial x_k\partial x_l}\right)-$$

$$-q\frac{\partial}{\partial x_j}\Phi(\mathbf{x})\cdot n(\mathbf{x},t)V_j\cdot n(\mathbf{x},t)G^{(1)}_{ij}+\frac{q\hbar^2}{24}\frac{\partial^3\Phi(\mathbf{x})}{\partial x_j\partial x_k\partial x_l}n(\mathbf{x},t)L^{(1)}_{ijkl}=C_{S_i}[w^{(0)}] \tag{15}$$

where V_i and S_i are the significant components of macroscopic velocity $\mathbf{V}$ and energy-flux $\mathbf{S}$ respectively.

Besides the average densities, velocities, energies and energy fluxes, additional quantities appear

$$n(\mathbf{x},t)\begin{pmatrix}T^{(0)}_{ijk}\\ T^{(1)}_{ijk}\end{pmatrix}=\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2}\begin{pmatrix}1\\ \mathcal{E}(p)\end{pmatrix}w^{(0)}(\mathbf{x},p,t)\,\frac{\partial^3\mathcal{E}(p)}{\partial p_i\partial p_j\partial p_k}\,dp,$$

$$n(\mathbf{x},t)\begin{pmatrix}H^{(0)}_{ijkl}\\ H^{(1)}_{ijkl}\end{pmatrix}=\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2}\begin{pmatrix}1\\ \mathcal{E}(p)\end{pmatrix}w^{(0)}(\mathbf{x},p,t)\,\frac{\partial^3\mathcal{E}(p)}{\partial p_i\partial p_j\partial p_k}\,v_l\,dp,$$

$$n(\mathbf{x},t)\begin{pmatrix}G^{(0)}_{ij}\\ G^{(1)}_{ij}\end{pmatrix}=\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2}\begin{pmatrix}1\\ \mathcal{E}(p)\end{pmatrix}w(\mathbf{x},p,t)\frac{\partial^2\mathcal{E}(p)}{\partial p_i\partial p_j}\,dp,$$

$$n(\mathbf{x},t)\begin{pmatrix}L^{(0)}_{ijkl}\\ L^{(1)}_{ijkl}\end{pmatrix}=\frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2}\begin{pmatrix}1\\ \mathcal{E}(p)\end{pmatrix}w^{(0)}(\mathbf{x},p,t)\frac{\partial^4\mathcal{E}(p)}{\partial p_i\partial p_j\partial p_k\partial p_l}\,dp,$$

[3] Einstein's summation convention is used.

$$n(\mathbf{x},t)\begin{pmatrix} F_{ij}^{(0)} \\ F_{ij}^{(1)} \end{pmatrix} = \frac{2}{(2\pi\hbar)^2}\int_{\mathbb{R}^2}\begin{pmatrix} 1 \\ \mathcal{E}(p) \end{pmatrix} w(\mathbf{x},p,t)v_i\, v_j\, dp.$$

that must be expressed as function of the basic variables n, W, $\mathbf{V}$, $\mathbf{S}$. Regarding the production terms, they are given by the sum of contributions arising from the different types of phonon scattering. Explicit closure relations have been obtained and numerical simulations are under current investigation.

4 Conclusions and Future Work

Initially a semi-classical hydrodynamic model for charge transport in graphene has been presented. To include quantum effects, the proposed model has been extended by incorporating the first quantum corrections. Therefore in the last section an example of quantum hydrodynamic model for charge transport in graphene has been formulated. It is composed of the semiclassical model presented in [2, 11, 12] augmented with quantum corrections at $\hbar^2$ order deduced by exploiting the equilibrium Wigner function obtained by solving the Bloch equation in the case of graphene. As $\hbar \mapsto 0$, the proposed model of course reduces to the semiclassical one which turned out to be accurate enough when comparison with DSMC results have been performed [11, 12]. Several strategies can be found in the literature for devising quantum hydrodynamic models (the interested reader is refereed to [9] for a comprehensive review) but usually strong approximations on the collision terms or on the energy bands are introduced and the semiclassical limit leads to semiclassical models whose soundness is questionable. To asses the validity of the proposed model numerical simulations are under current investigation and they will be presented in a forthcoming article.

References

1. Barletti, L.: Hydrodynamic equations for electrons in graphene obtained from the maximum entropy principle. J. Math. Phys. **55**(8), 083303 (2014)
2. Camiola, V.D., Romano, V.: Hydrodynamical model for charge transport in graphene. J. Stat. Phys. **157**, 114–1137 (2014)
3. Castro Neto, A.H., Guinea, F., Peres, N.M.R., Novoselov, K.S., Geim, A.K.: The electronic properties of graphene. Rev. Modern Phys. **81**, 109 (2009)
4. Coco, M., Romano, V.: Simulation of electron-phonon coupling and heating dynamics in suspended monolayer graphene including all the phonon branches. J. Heat Transfer **45**(7), 540–553 (2016). https://doi.org/10.1115/1.4040082
5. Coco, M., Majorana, A., Romano, V.: Cross validation of discontinuous Galerkin method and Monte Carlo simulations of charge transport in graphene on substrate. Ricerche Mat. (2016). https://doi.org/10.1007/s11587-016-0298-4

6. Coco, M., Mascali, G., Romano, V.: Monte Carlo analysis of thermal effects in monolayer graphene. J. Comput. Theor. Transp. **45**(7), 540–553 (2016)
7. Gardner, C.L.: The quantum hydrodynamic model for semiconductor devices. SIAM J. Appl. Math. **54**(2), 409–427 (1994)
8. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**, 620 (1957)
9. Jüngel, A.: Transport Equations for Semiconductors. Springer, Berlin/Heidelberg (2009)
10. Lichtenberger, P., Morandi, O., Schürrer, F.: High-field transport and optical phonon scattering in graphene. Phys. Rev. B **84**, 045406 (2011)
11. Luca, L., Romano, V.: Comparing linear and nonlinear hydrodynamical models for charge transport in graphene based on the Maximum Entropy Principle. Int. J. Non-Linear Mech. **104**, 39–58 (2018)
12. Luca, L., Romano, V.: Hydrodynamical models for charge transport in graphene based on the Maximum Entropy Principle: the case of moments based on energy powers. Atti della Accademia Peloritana dei Pericolanti **96**(S1), A5 (2018)
13. Majorana, A., Romano, V.: Numerical solutions of the spatially homogeneous Boltzmann equation for electrons in n-doped graphene on a substrate. J. Theor. Comput. Transp. (2017). https://doi.org/10.1080/23324309.2017.1311267
14. Majorana, A., Mascali, G., Romano, V.: Charge transport and mobility in monolayer graphene. J. Math. Ind. (2016). https://doi.org/10.1186/s13362-016-0027-3
15. Mascali, G., Romano, V.: Exploitation of the maximum entropy principle in mathematical modeling of charge transport in semiconductors. Entropy **19**(1), 36 (2017). https://doi.org/10.3390/e19010036 (open access article)
16. Mascali, G., Romano, V.: Charge transport in graphene including thermal effects. SIAM J. Appl. Math. **77**, 593–613 (2017)
17. Morandi, O., Schürrer, F.: Wigner model for quantum transport in graphene. J. Phys. A: Math. Theor. **44**, 265301 (2011)
18. Muscato, O., Wagner, W.: A class of stochastic algorithms for the Wigner equation. SIAM J. Sci. Comput. **38**(3), A1483–A1507 (2016). https://doi.org/10.1137/16M105798X
19. Querlioz, D., Dollfus, P.: The Wigner Monte Carlo Method for Nanoelectronic Devices. ISTE Wiley, Hoboken (2010)
20. Romano, V.: Quantum corrections to the semiclassical hydrodynamical model of semiconductors based on the maximum entropy principle. J. Math. Phys. **48**, 123504 (2007)
21. Romano, V.: The equilibrium Wigner function in the case of nonparabolic energy bands. In: Fitt, A.D., et al. (eds.) Progress in Industrial Mathematics at ECMI 2008. Mathematics in Industry, vol. 15, pp. 135–140. Springer, Berlin/Heidelberg (2010). https://doi.org/10.1007/978-3-642-12110-4
22. Romano, V., Majorana, A., Coco, M.: DSMC method consistent with the Pauli exclusion principle and comparison with deterministic solutions for charge transport in graphene. J. Comput. Phys. **302**, 267–284 (2015)
23. Wigner, E.: On the quantum correction for thermodynamic equilibrium. Phys. Rev. **40**, 749–749 (1932)

Quantum Model for the Transport of Nearly Localized Particles

Omar Morandi

Abstract A quantum model based on the Gaussian-Hermite expansion of the wave function of a system of n particles is proposed. The dynamics is described by trajectories in a configuration space. Our method is designed to provide some corrections to the classical motion of nearly localized particles. As an application of our model we describe the motion of a nearly localized particle in a 2D confining structure.

1 Introduction

During the last decade, the dynamics of molecules inside gases or nanostructures have been extensively investigated. New concepts such as deterministic and stochastic quantum trajectories have been proposed to develop new ab initio methods that go beyond the classical description of the atomic nuclei[1–12].

Such approaches are mainly based on the Bohm interpretation of the quantum mechanics where the concept of the deterministic trajectory of a particle is extended to the quantum dynamics in a rigorous way [13]. One of the first attempts to apply Bohm theory to a real nanostructure, have been proposed by Tully [14] and Wyatt [15]. Nowadays, methods based on the Bohm formalism are very popular and are at the basis of few general methods such as the multiconfiguration time-dependent Hartree method [16] and the Bohm trajectories extended to the complex plane [17]. Moreover, stochastic methods have also been employed [18]. Based on the classical concept of Brownian and Langevin dynamics, stochastic models are able to reproduce the quantum statistical properties of protons in harmonic traps [19].

O. Morandi (✉)
University of Florence, Florence, Italy
e-mail: omar.morandi@unifi.it

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_14

In this paper, we derive the motion of a system of quantum particles. We assume that every particle is described by a Gaussian-like wave packet parametrized by a set of numbers that depend on time. Our approach is designed to describe the motion of heavy particles as for example neutrons and protons. We will discuss the application of our method to the 2D motion of a particle in the presence of a confining non harmonic potential. In particular, the connection with the Bohm interpretation of the quantum mechanics will be discussed.

2 Particle Motion: Beyond the Gaussian Beam Approximation

We consider the quantum mechanical evolution of a system defined by the wave function $\psi(\mathbf{r}) \in L^2(\mathbb{R}^d)$. Our system consists of $n \leq d$ particles, where d is the dimension of the space. We focus on the physical situation in which each particle is localized around a spatial coordinate. We assume that the probability dispersion of the particle wave function is modulated by a Gaussian function centred on mean particle positions. Moreover, we assume that all the particles are identical and have mass $m = 1$. The particles move in the presence of an external potential $U(\mathbf{r})$. We discard the direct particle-particle interaction. However, $U(\mathbf{r})$ may contains some nonlinear terms that describe the particle-particle interaction at the mean field level. We develop a model that preserves the classical description of particle motion in terms of trajectories. In particular, in our model the particle motion is expressed by a system of nonlinear ODE for a set of physically relevant parameters. This is obtained by parametrizing the particle wave function by a complete set of parameters which extend the classical dynamical quantities position and momentum. Similarly to the coherent state projection technique [20–22], we project the wave function of the system on the basis set of the Harmonic oscillator shifted at the mean particle positions. The Gaussian beam approximation is a popular method used to describe nearly localized particles [23–25] Similarly to the Gaussian beam approximation, our expansion procedure is based on the projection of the solution over a set of functions which are modulated by a Gaussian whose width changes in time according to the quantum evolution equation.

In this paper we extend the results obtained in Ref. [26], where the 1D case with some extension to the 2D problem, was considered. In Ref. [26] the evolution equations for the parameters have been obtained by applying a integral approach based on the Euler-Lagrange formalism. Here, we proceed in a more direct way. We calculate the evolution equations by using the Madelung decomposition in which the particle wave function is expressed in polar coordinates. In particular, in Sect. 2.2 we calculate the expression of the projection of the potential on the Hermite basis in a closed form. This is an important step toward the implementation of a efficient numerical solver for the particle motion.

We start by considering the time dependent Schrödinger equation of a quantum system of dimension d

$$i\frac{\partial\psi}{\partial t} = \left(-\frac{1}{2}\Delta_{\mathbf{r}} + U(\mathbf{r})\right)\psi \ . \tag{1}$$

We have normalized the Planck constant $\hbar = 1$. According to Madelung, we represent the particle wave function in polar coordinates $\psi(\mathbf{r}) = \sqrt{n(\mathbf{r}-\mathbf{s})}e^{i\chi(\mathbf{r}-\mathbf{s})}$. We have introduced the parameter $\mathbf{s} \in \mathbb{R}^d$ which represents the mean particle position. In particular, s is the centre of the Hermite expansion that will be performed in the following. The evolution equation of s (see Eq. (22) below) can be interpreted as a Newton equation with quantum corrections. We assume that the density $n(x)$ is a Gaussian function modulated by the polynomial P

$$n(\mathbf{r},t) = P(\mathbf{r},t)e^{-\langle\sigma,\mathbf{r}\rangle} \ .$$

The parameter $\boldsymbol{\sigma} \in \mathbb{R}^d$ provides the width of the Gaussian packed. For the sake of compactness, we have introduced the notation $\langle\boldsymbol{\sigma},\mathbf{r}\rangle \doteq \sum_{n=1}^{d} r_i^2\sigma_i$, where the lower indexes indicate the components of the vectors along the coordinate axis. The Madelung transform leads to the well known quantum hydrodynamic equations

$$\frac{\partial n}{\partial t} = -\nabla\cdot\boldsymbol{J} + \frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t}\cdot\nabla n \tag{2}$$

$$\frac{\partial\boldsymbol{J}}{\partial t} = -\nabla\left(\frac{\boldsymbol{J}\otimes\boldsymbol{J}}{n}\right) + n\nabla\left(Q + U_{[\mathbf{s}]}\right) + \frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t}\cdot\nabla\chi \ . \tag{3}$$

Here, we have introduced the notation $U_{[\mathbf{s}]} \doteq U(\mathbf{r}+\mathbf{s})$, $J = n\nabla\chi$ is the quantum current and $Q(\mathbf{r})$ is the quantum Bohm potential

$$Q \equiv \frac{1}{2}\frac{\Delta\sqrt{n}}{\sqrt{n}} = \frac{1}{4}\left[\frac{\Delta P}{P} - \frac{|\nabla P|^2}{2P^2} - 2\sum_{i=1}^{d}\sigma_i\frac{r_i\frac{\partial P}{\partial r_i}}{P} + 2\sum_{i=1}^{d}\sigma_i(r_i^2\sigma_i - 1)\right] . \tag{4}$$

Concerning the physical meaning of the Bohm potential, we refer to [15, 26]. By introducing the auxiliary variable $R(x) = \ln(P(x))$, the previous expression becomes

$$Q = \frac{1}{4}\left[\Delta R + \frac{|\nabla R|^2}{2} - 2\sum_{i=1}^{d}\sigma_i r_i\frac{\partial R}{\partial r_i} + 2\sum_{i=1}^{d}\sigma_i(r_i^2\sigma_i - 1)\right] .$$

From Eq. (3) with some algebra we obtain to the evolution equation for the phase χ

$$\frac{\partial\chi}{\partial t} = -\frac{1}{2}|\nabla\chi|^2 + Q + U_{[\mathbf{s}]} + \frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t}\cdot\nabla\chi \ . \tag{5}$$

The main interest of our approach is to derive the evolution equation of the parameters obtained by expanding the functions χ and P on the Hermite polynomial basis set. We write the expansion as follows

$$P(\mathbf{r},t) = \sum_{\{n\}=0}^{\infty} a_{\{n\}}(t) h^{\boldsymbol{\sigma}}_{\{n\}}(\mathbf{r}) \tag{6}$$

$$\chi(\mathbf{r},t) = \sum_{\{n\}=0}^{\infty} \chi_{\{n\}}(t) h^{\boldsymbol{\sigma}}_{\{n\}}(\mathbf{r}) \ . \tag{7}$$

We have introduced the compact notation $\{n\}$ to indicate the sequence of d integers $\{n\} \doteq (n_1, n_2, \ldots, n_d)$ and

$$h^{\boldsymbol{\sigma}}_{\{n\}}(\mathbf{r}) \doteq \prod_{i=1}^{d} h^{\sigma_i}_{n_i}(r_i) \ . \tag{8}$$

The functions $h^{\sigma_i}_{n_i}(r_i)$ are the normalized Hermite functions. For the details concerning the definition of $h^{\sigma_i}_{n_i}(r_i)$ we refer to [26], Eq. (19). In order to clarify the notation, we write explicitly Eq. (7) in the 3D case

$$\chi(\mathbf{r},t) = \sum_{n_x,n_y,n_z=0}^{\infty} \chi_{n_x,n_y,n_z}(t) h^{\sigma_x}_{n_x}(x) h^{\sigma_y}_{n_y}(y) h^{\sigma_z}_{n_z}(z) \ . \tag{9}$$

By using the property of orthonormality of the Hermite functions, it is easy to invert the previous equations. As un example, we give the inversion formula for P

$$a_{\{n\}} = \int P(\mathbf{r}) h^{\boldsymbol{\sigma}}_{\{n\}}(\mathbf{r}) e^{-\langle \boldsymbol{\sigma}, \mathbf{r} \rangle} \, \mathrm{d}\mathbf{r} \ . \tag{10}$$

Hereafter, all the integrals are considered over all the space $\mathbb{R}^d$. Finally, it is useful to expand the variable $R(x) = \ln(P(x))$ in the Hermite basis set

$$R(\mathbf{r}) = \sum_{m} R_{\{m\}} h^{\boldsymbol{\sigma}}_{\{m\}}(\mathbf{r}) \ . \tag{11}$$

In order to clarify our approach, in the following expression we indicate all the parameters that we have introduced

$$\psi(\mathbf{r}) = \sqrt{n(\mathbf{r}-\mathbf{s})} e^{i\chi(\mathbf{r}-\mathbf{s})} \Rightarrow \begin{cases} a_{\{n\}} \in \ell^2(\mathbb{N}^d) & : \text{modulus} \\ \chi_{\{n\}} \in \ell^2(\mathbb{N}^d) & : \text{phase} \\ \mathbf{s} \in \mathbb{R}^d & : \text{mean position} \\ \boldsymbol{\sigma} \in \mathbb{R}^d & : \text{Gaussian width} \\ R_{\{n\}} = R_{\{n\}}(a_{\{n\}}) \in \ell^2(\mathbb{N}^d) & \end{cases} .$$

In particular, in the last line we have indicated that the parameters R_n are not independent and obtained by $a_{\{n\}}$. The details concerning this point are given in Ref. [26]. We calculate now the evolution equations for our set of parameters. We give some technical details concerning the expansion coefficients of the phase χ. The calculations proceed straightforwardly. We multiply Eq. (5) by $h^{\boldsymbol{\sigma}}_{\{n\}}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}$ and we integrate over $\mathbb{R}^d$. We obtain

$$\frac{\mathrm{d}\chi_{\{n\}}}{\mathrm{d}t} = \underbrace{-\frac{1}{2}\int |\nabla\chi|^2\, h^{\boldsymbol{\sigma}}_{\{n\}}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}\,\mathrm{d}\mathbf{r}}_{I} + \underbrace{\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t}\cdot\int \nabla\chi\, h^{\boldsymbol{\sigma}}_{\{n\}}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}\,\mathrm{d}\mathbf{r}}_{II}$$

$$\underbrace{-\sum_{i=1}^{d}\frac{\mathrm{d}\sigma_i}{\mathrm{d}t}\sum_{\{n'\}}\chi_{\{n'\}}\int h^{\boldsymbol{\sigma}}_{\{n\}}\frac{\mathrm{d}h^{\boldsymbol{\sigma}}_{\{n'\}}}{\mathrm{d}\sigma_i}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}\,\mathrm{d}\mathbf{r}}_{III} + \int \big(Q + U_{[\mathbf{s}]}\big)\, h^{\boldsymbol{\sigma}}_{\{n\}}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}\,\mathrm{d}\mathbf{r}\ .$$

For the first term we have

$$\begin{aligned} I &= -\frac{1}{2}\sum_{\{r\},\{s\}}\chi_{\{r\}}\chi_{\{s\}}\sum_{i=1}^{d}\int \frac{\mathrm{d}h^{\boldsymbol{\sigma}}_{\{r\}}}{\mathrm{d}r_i}\frac{\mathrm{d}h^{\boldsymbol{\sigma}}_{\{s\}}}{\mathrm{d}r_i}h^{\boldsymbol{\sigma}}_{\{n\}}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}\,\mathrm{d}\mathbf{r} \\ &= -\sum_{\{r\},\{s\}}\chi_{\{r\}}\chi_{\{s\}}\sum_{i=1}^{d}\sigma_i\sqrt{r_i s_i}\int h^{\boldsymbol{\sigma}}_{\{r;r_i\to r_i-1\}}h^{\boldsymbol{\sigma}}_{\{s;s_i\to s_i-1\}}h^{\boldsymbol{\sigma}}_{\{n\}}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}\,\mathrm{d}\mathbf{r} \\ &= -\sum_{\{r\},\{s\}}\chi_{\{r\}}\chi_{\{s\}}\sum_{i=1}^{d}\sigma_i\sqrt{r_i s_i}\,\mathbb{A}^{\sigma_i}_{n_i,r_i-1,s_i-1}\prod_{j\neq i}\mathbb{A}^{\sigma_j}_{n_j,r_j,s_j}\ . \end{aligned} \tag{12}$$

The matrix $\mathbb{A}$ is defined as

$$\mathbb{A}^{\sigma}_{n,r,s} = \pi^{1/4}\int_{\mathbb{R}} h^{\sigma}_{n}(x)h^{\sigma}_{r}(x)h^{\sigma}_{s}(x)e^{-x^2\sigma}\,\mathrm{d}x\ .$$

Details concerning the calculation of $\mathbb{A}^{\sigma}_{n,r,s}$ are given in [26]. We have used the following property of the Hermite functions (see Eq. (22) of Ref. [26])

$$\frac{\mathrm{d}h^{\boldsymbol{\sigma}}_{\{n\}}(\mathbf{r})}{\mathrm{d}r_i} = \sqrt{2n_i\sigma_i}\,h^{\sigma_i}_{n-1}(r_i)\prod_{j\neq i,\,j=1}^{d}h^{\sigma_j}_{n_j}(r_j) = \sqrt{2n_i\sigma_i}\,h^{\boldsymbol{\sigma}}_{\{n;n_i\to n_i-1\}}(\mathbf{r})\ . \tag{13}$$

Finally, we have introduced the following notation $\{r;\, r_i \to a\} \doteq (r_1,\ldots,r_{i-1},a,$ $r_{i+1},\ldots,r_d)$ where the i-th term is substituted by a. For the second term, we have

$$II = \sum_{\{r\}}\chi_{\{r\}}\sum_{i=1}^{d}\frac{\mathrm{d}s_i}{\mathrm{d}t}\int \frac{\mathrm{d}h^{\boldsymbol{\sigma}}_{\{r\}}}{\mathrm{d}r_i}h^{\boldsymbol{\sigma}}_{\{n\}}e^{-\langle\boldsymbol{\sigma},\mathbf{r}\rangle}\,\mathrm{d}\mathbf{r} = \sum_{i=1}^{d}\chi_{\{n;n_i\to n_i+1\}}\frac{\mathrm{d}s_i}{\mathrm{d}t}\sqrt{2(n_i+1)\sigma_i}\ . \tag{14}$$

Furthermore,

$$III = -\sum_{i=1}^{d} \frac{\mathrm{d}\sigma_i}{\mathrm{d}t} \left(\frac{2n_i+1}{4\sigma_i} \chi_{\{n\}} + \frac{\sqrt{(n_i+2)(n_i+1)}}{2\sigma_i} \chi_{\{n;n_i \to n_i+2\}} \right) . \tag{15}$$

We consider now the term that contains the Bohm potential

$$\int Q h^{\boldsymbol{\sigma}}_{\{n\}} e^{-\langle \boldsymbol{\sigma}, \mathbf{r} \rangle} \, \mathbf{dr} = \frac{1}{4} \int \left[\Delta R + \frac{|\nabla R|^2}{2} - 2 \sum_{i=1}^{d} \sigma_i r_i \frac{\partial R}{\partial r_i} + 2 \sum_{i=1}^{d} \sigma_i (r_i^2 \sigma_i - 1) \right] h^{\boldsymbol{\sigma}}_{\{n\}} e^{-\langle \boldsymbol{\sigma}, \mathbf{r} \rangle} \, \mathbf{dr}$$

By using the expansion of R we obtain the explicit form of the previous terms. After cumbersome algebra we obtain

$$\begin{aligned} \int Q h^{\boldsymbol{\sigma}}_{\{n\}} e^{-\langle \boldsymbol{\sigma}, \mathbf{r} \rangle} \, \mathbf{dr} =& \frac{1}{4} \sum_{\{r\},\{s\}} R_{\{r\}} R_{\{s\}} \sum_{i=1}^{d} \sigma_i \sqrt{r_i s_i} \mathbb{A}^{\sigma_i}_{n_i, r_i-1, s_i-1} \prod_{j \neq i}^{d} \mathbb{A}^{\sigma_j}_{n_j, r_j, s_j} - \sum_{i=1}^{d} \frac{\sigma_i}{2} n_i R_{\{n\}} \\ &+ \pi^{d/4} \frac{1}{4} \sum_{i=1}^{d} \sigma_i^{3/4} \left(\sqrt{2} \delta_{n_i,2} - \delta_{n_i,0} \right) \prod_{j \neq i}^{d} \sigma_j^{-1/4} \delta_{n_j,0} \; . \end{aligned}$$

Here, δ denotes the Kronecker's delta. The final expression of the evolution equation for χ is given in Eq. (19). The evolution equations for a, s and σ can be obtained by using the continuity equation. It is easy to verify that

$$\frac{\partial n}{\partial t} = -\nabla \cdot \boldsymbol{J} + \frac{\mathrm{d}\mathrm{s}}{\mathrm{d}t} \cdot \nabla n = \nabla \cdot \left[n \left(-\nabla \chi + \frac{\mathrm{d}\mathrm{s}}{\mathrm{d}t} \right) \right] . \tag{16}$$

Proceeding in similar way as we have done for χ, from Eq. (16) we obtain

$$\begin{aligned} \frac{\mathrm{d}a_{\{n\}}}{\mathrm{d}t} =& \sum_{i=1}^{d} \frac{\mathrm{d}\sigma_i}{\mathrm{d}t} \frac{1}{2\sigma_i} \left[a_{\{n\}} \frac{2n_i+1}{2} + a_{\{n;n_i \to n_i-2\}} \sqrt{n_i(n_i-1)} \right] \\ &- \sum_{i=1}^{d} a_{\{n;n_i \to n_i-1\}} \frac{\mathrm{d}s_i}{\mathrm{d}t} \sqrt{2 n_i \sigma_i} \\ &+ 2 \sum_{\{r\},\{s\}} a_{\{r\}} \chi_{\{s\}} \sum_{i=1}^{d} \sigma_i \sqrt{n_i s_i} \mathbb{A}^{\sigma_i}_{n_i-1, r_i, s_i-1} \prod_{j \neq i} \mathbb{A}^{\sigma_j}_{n_j, r_j, s_j} \; . \end{aligned}$$

It is easy to verify that the first coefficient $a_{\{0,\ldots,0\}}$ is directly related to the L2 norm of ψ. It can be integrated analytically. From the previous equation we have

$$\frac{\mathrm{d}a_{\{0,\ldots,0\}}}{\mathrm{d}t} = \sum_{i=1}^{d} \frac{\mathrm{d}\sigma_i}{\mathrm{d}t} \frac{1}{4\sigma_i} a_{\{0,\ldots,0\}} \; ,$$

whose solution is $a_{\{0,\ldots,0\}} = \pi^{-d/4} \prod_{i=1}^{d} \sigma_i^{1/4}(t)$. According to the discussion of Ref. [26], it is possible to verify that the number of expansion coefficients that we have introduced so far is redundant. In particular, it is always possible to fix some of the coefficients a that multiply the linear and the quadratic terms of the expansion (6). For the linear terms, the evolution equation is

$$\frac{\mathrm{d}a_{\{0,\ldots,1,,\ldots,0\}}}{\mathrm{d}t} = \sum_{j=1}^{d} \frac{\mathrm{d}\sigma_j}{\mathrm{d}t} \frac{1}{2\sigma_j} \left[a_{\{0,\ldots,1,,\ldots,0\}} \frac{2n_j+1}{2} \right] - \sum_{j=1}^{d} a_{\{n;n_j \to n_j-1\}} \frac{\mathrm{d}s_j}{\mathrm{d}t} \sqrt{2n_j\sigma_j}$$

$$+ 2 \sum_{\{r\},\{s\}} a_{\{r\}} \chi_{\{s\}} \sigma_i \sqrt{s_i} \mathbb{A}^{\sigma_i}_{0,r_i,s_i-1} \prod_{j \neq i} \mathbb{A}^{\sigma_j}_{0,r_j,s_j}$$

$$= a_{\{0,\ldots,1,,\ldots,0\}} \sum_{j=1}^{d} \frac{\mathrm{d}\sigma_j}{\mathrm{d}t} \frac{2n_j+1}{4\sigma_j} - a_{\{0,\ldots,0\}} \frac{\mathrm{d}s_i}{\mathrm{d}t} \sqrt{2\sigma_i}$$

$$+ \frac{2\sigma_i \sqrt{r_i+1} \prod_k \sigma_k^{1/4}}{\pi^{d/4}} \sum_{\{r\}} a_{\{r\}} \chi_{\{r;r_i \to r_i+1\}} \; .$$

By imposing $a_{\{0,\ldots,1,,\ldots,0\}} = 0$ we obtain the evolution equation for s

$$\frac{\mathrm{d}s_i}{\mathrm{d}t} = \sqrt{2\sigma_i} \sum_{\{m\}} a_{\{m\}} \chi_{\{m;m_i \to m_i+1\}} \sqrt{m_i+1} \; .$$

The evolution equation of the width of the Gaussian $\boldsymbol{\sigma}$ is obtained by setting the quadratic terms to zero: $a_{\{0,\ldots,2,\ldots,0\}} = 0$. From

$$\frac{\mathrm{d}a_{\{0,\ldots,2,\ldots,0\}}}{\mathrm{d}t} = \frac{\mathrm{d}\sigma_i}{\mathrm{d}t} \frac{1}{2\sigma_i} a_{\{0\}} \sqrt{2} + \sum_{i=1}^{d} \frac{\mathrm{d}\sigma_i}{\mathrm{d}t} \frac{1}{2\sigma_i} \left[a_{\{0,\ldots,2,\ldots,0\}} \frac{2n_i+1}{2} \right]$$

$$- a_{\{0,\ldots,1,\ldots,0\}} \frac{\mathrm{d}s_i}{\mathrm{d}t} 2\sqrt{\sigma_i}$$

$$+ 2 \left(\prod_j \sigma_j^{1/4} \right) \sum_{\{r\},\{s\}} a_{\{r\}} \chi_{\{s\}} \sigma_i \sqrt{2s_i} \mathbb{A}^{1}_{1,r_i,s_i-1} \prod_{j \neq i} \mathbb{A}^{1}_{0,r_j,s_j} \; ,$$

we obtain

$$\frac{\mathrm{d}\sigma_i}{\mathrm{d}t} = -\, 4\sigma_i^2 \sum_{\{m\}} a_{\{m\}} \left(\chi_{\{m\}} m_i + \chi_{\{m;m_i \to m_i+2\}} \sqrt{(m_i+1)(m_i+2)} \right) \; .$$

2.1 Evolution Equations

We give here the final form of the evolution equations for the phase $\chi_{\{n\}}$, the modulus $a_{\{n\}}$, the Gaussian width $\boldsymbol{\sigma}$ and the center of the expansion s of the particle wave function.

$$\frac{\mathrm{d}\chi_{\{n\}}}{\mathrm{d}t} = \left(\prod_{j}^{d} \sigma_j^{1/4}\right) \sum_{\{r\},\{s\}} \left(-\chi_{\{r\}}\chi_{\{s\}} + \frac{1}{4} R_{\{r\}} R_{\{s\}}\right) \sum_{i=1}^{d} \sigma_i \sqrt{r_i s_i}\, \mathbb{A}^1_{n_i, r_i-1, s_i-1} \tag{17}$$

$$\times \prod_{j\neq i} \mathbb{A}^1_{n_j, r_j, s_j} - \sum_{i=1}^{d} \frac{\sigma_i}{2} n_i R_{\{n\}} + \pi^{d/4} \frac{1}{4} \sum_{i=1}^{d} \sigma_i^{3/4} \left(\sqrt{2}\delta_{n_i,2} - \delta_{n_i,0}\right) \left(\prod_{j\neq i}^{d} \sigma_j^{-1/4} \delta_{n_j,0}\right) \tag{18}$$

$$+ \sum_{i=1}^{d} \chi_{\{n; n_i \to n_i+1\}} \frac{\mathrm{d}s_i}{\mathrm{d}t} \sqrt{2(n_i+1)\sigma_i}$$

$$+ \sum_{i} M_i \left(\frac{2n_i+1}{2} \chi_{\{n\}} + \sqrt{(n_i+2)(n_i+1)}\, \chi_{\{n; n_i \to n_i+2\}}\right) + \int U_{[\mathbf{s}]} h^{\boldsymbol{\sigma}}_{\{n\}} e^{-\langle \boldsymbol{\sigma}, \mathbf{r} \rangle}\, \mathrm{d}\mathbf{r} \tag{19}$$

$$\frac{\mathrm{d}a_{\{n\}}}{\mathrm{d}t} = \sum_{i=1}^{d} \frac{\mathrm{d}\sigma_i}{\mathrm{d}t} \frac{1}{2\sigma_i} \left[a_{\{n\}} \frac{2n_i+1}{2} + a_{\{n; n_i \to n_i-2\}} \sqrt{n_i(n_i-1)}\right] - \sum_{i=1}^{d} a_{\{n; n_i \to n_i-1\}} \frac{\mathrm{d}s_i}{\mathrm{d}t} \times$$

$$\sqrt{2n_i\sigma_i} + 2\left(\prod_{j}^{d} \sigma_j^{1/4}\right) \sum_{\{r\},\{s\}} a_{\{r\}} \chi_{\{s\}} \sum_{i=1}^{d} \sigma_i \sqrt{n_i s_i}\, \mathbb{A}^1_{n_i-1, r_i, s_i-1} \prod_{j\neq i} \mathbb{A}^1_{n_j, r_j, s_j} \tag{20}$$

$$\frac{\mathrm{d}\sigma_i}{\mathrm{d}t} = -2M_i\sigma_i \tag{21}$$

$$\frac{\mathrm{d}s_i}{\mathrm{d}t} = \sqrt{2\sigma_i} S_i\ , \tag{22}$$

where $M_i = 2\sigma_i \sum_{\{m\}} a_{\{m\}} \left(\chi_{\{m\}} m_i + \chi_{\{m; m_i \to m_i+2\}} \sqrt{(m_i+1)(m_i+2)}\right)$ and $S_i = \sum_{\{m\}} a_{\{m\}} \chi_{\{m; m_i \to m_i+1\}} \sqrt{m_i+1}$.

2.2 Projection Coefficients of the Potential

By looking at the final system of evolution equations we see that the external potential appears only in the evolution equation for the phase (19) via the term $\int U_{[\mathbf{s}]} h^{\boldsymbol{\sigma}}_{\{n\}} e^{-\langle \boldsymbol{\sigma}, \mathbf{r} \rangle}$ $\mathrm{d}\mathbf{r}$ which, from a mathematical point of view, represents the projection of the potential on the Hermite basis set. In the general case, this term can be evaluated only numerically; since its calculation can be very costly, it is convenient to consider the case in which external potential is a polynomial. In this case, the calculations can be done explicitly. We proceed by assuming that the potential U can be written in the following form which, from the point of view of the applications, is very general

$$U(\mathbf{r}) = \prod_{i}^{d} U^{i}(x_i) = \prod_{i}^{d} \sum_{n}^{N} \alpha_n^i x_i^n \ . \tag{23}$$

Here, $U^i(x_i) \doteq \sum_n^N \alpha_n^i x_i^n$, α_n^i are a set of given coefficients and N is the maximum degree at which the coordinate appears in the polynomial expansion of U. We obtain

$$\begin{aligned}\int U(\mathbf{r}+\mathbf{s}) h^{\boldsymbol{\sigma}}_{\{m\}} e^{-\langle \boldsymbol{\sigma}, \mathbf{r}\rangle} =& \sum_i \int h^{\sigma_i}_{m_i} e^{-\sigma_i r_i} U^i(x_i+s_i)\, \mathrm{d}x_i \prod_{j\neq i}^{d} \int h^{\sigma_j}_{m_j} e^{-\sigma_j r_j}\, \mathrm{d}x_j \\ =& \pi^{\frac{d-1}{4}} \sum_i \left(\prod_{j\neq i}^{d} \sigma_j^{-1/4} \right) \int h^{\sigma_i}_{m_i} U^i(x_i+s_i) e^{-\sigma_i r_i}\, \mathrm{d}x_i \ .\end{aligned}$$

The previous equation can be elaborated by inserting the polynomial expansion of Eq. (23). In particular, it is convenient to order the terms according to the degree with respect to the parameter s_i. At the end of the computation, we obtain

$$\begin{aligned}\int U_{[s]} h^{\sigma}_{2r} e^{-x^2\sigma}\, \mathrm{d}x =& \frac{\pi^{1/4} 2^r \sigma^{-1/4}}{\sqrt{(2r)!}} \left(\sum_{u=0}^{\left\lfloor \frac{N}{2} \right\rfloor - r} s^{2u} \frac{\sigma^u 2^{2u}}{(2u)!} \Gamma(u,r,\sigma) + 2\sigma^{\frac{1}{2}} \right. \\ & \left. \times \sum_{u=0}^{\left\lfloor \frac{N-1}{2} \right\rfloor - r} s^{2u+1} \frac{\sigma^u 2^{2u}}{(2u+1)!} \widetilde{\Gamma}(u,r,\sigma) \right)\end{aligned}$$

$$\begin{aligned}\int U_{[s]} h^{\sigma}_{2r+1} e^{-x^2\sigma}\, \mathrm{d}x =& \frac{\pi^{1/4} 2^r \sigma^{-3/4}}{\sqrt{2}\sqrt{(1+2r)!}} \left(\sum_{u=1}^{\left\lfloor \frac{N}{2} \right\rfloor - r} s^{2u-1} \frac{\sigma^u 2^{2u}}{(2u-1)!} \Gamma(u,r,\sigma) + 2\sigma^{\frac{1}{2}} \right. \\ & \left. \times \sum_{u=0}^{\left\lfloor \frac{N-1}{2} \right\rfloor - r} s^{2u} \frac{2^{2u}\sigma^u}{(2u)!} \widetilde{\Gamma}(u,r,\sigma) \right) \ ,\end{aligned}$$

where $r \in \mathbb{N}$, the symbol $\lfloor x \rfloor$ denotes the integer part of x and we have defined

$$\Gamma(u, r, \sigma) \doteq \sum_{n=2(u+r)+(0,2,4,\ldots)}^{N} \alpha_n \frac{n!\sigma^{-\frac{n}{2}} 2^{-n}}{(\frac{n}{2} - u - r)!}$$

$$\widetilde{\Gamma}(u, r, \sigma) \doteq \sum_{n=2(u+r)+(1,3,5,\ldots)}^{N} \alpha_n \frac{n!\sigma^{-\frac{n}{2}} 2^{-n}}{(\frac{n-1}{2} - u - r)!} .$$

The symbols under the sums indicate that the index n takes the value $2(u + r)$ plus the numbers indicated inside the parenthesis (even numbers for the first equation and odd numbers for the second equations).

3 Numerical Simulations: 2D Case

We apply the model that has been introduced in the previous sections to the motion of heavy quantum particles (typically atoms with few protons). In this section, we provide some motivations to our work and we describe the results obtained by solving the evolution equations in the case of a two-dimensional system. There exist many cases in which electrons are delocalized in a solid structure. The electron wave function may extend over various atomic cells and has typically a very complex shape. For this reason, the simulation of electron motion in solids in a realistic case requires the application of complex many body full quantum methods like the DFT [27] or the Green function approach. On the other side, nuclei are well localized quantum particles. They are typically considered as point-like particles which move along the classical trajectories obtained by solving the Newton equation. The De Broglie wave length provides a simple estimate of the degree of localization of the quantum wave function of a particle. It is well known that De Broglie wave length provides the spatial scale on which the probability to find a particle around the position expectation value decays. Since the Broglie wave length is proportional to the inverse of the particle mass, heavy particles may have very small localization length. This is the case of atoms inside a solid. For this reason, the spread of the atomic wave function around the mean particle position is typically neglected. However, recent theoretical and experimental studies reveal the existence of physical conditions for which the previous approximation is violated. They suggest that slightly bound protons and the hydrogen molecule behave as quantum particles. In particular, we focus on the study of the phase transition of ice of water and we analyse the so called VIII-X transition. The physical phenomena concerns the modification of the atomic structure of the ice in the presence of an external pressure. In particular, the VIII-X phase transition concerns the behaviour of the protons. At low pressure, every protons of the H_2O crystal remain close to one oxygen atom. By increasing the pressure a phase transition is observed. At the critical pressure, the protons migrate to the middle of the O-O bound. This behaviour is explained by assuming that the electrostatic potential between two oxygen atoms at low pressure has a double well structure. Due to this potential profile the protons are trapped inside one of the two minima of the potential. By increasing the pressure, the two wells merge together and form a single central well. The equilibrium position of the proton is thus shifted to the middle of the two atoms. Even if the qualitative behaviour of the transition has been clarified, theoretical estimations of the transition pressure are not in agreement with the experiments. Bronstein et al. have shown that in order to obtain the experimental value of the transition pressure, it is essential to include the quantum mechanical delocalization of the protons in the numerical model [28]. Recent results [29] suggest that tunneling of protons plays a essential role in this process. Classical simulations ignore the possibility to overcome thin barriers by tunneling and tend to overestimate the transition pressure. Quantum corrected model are thus essential for simulating this kind of phenomena.

In particular, in Ref. [29], ab initio calculations indicate that the shape of the electrostatic potential around two oxygen atoms can be approximated by a simple expression. The potential is fitted by a harmonic term plus a non harmonic correction which is proportional to the macroscopic pressure applied to the sample.

Motivated by these results, we have applied our method to reproduce the two-dimensional motion of a single particle in a non harmonic potential. In the simulations we use the following double well potential

$$U(x, y) = -\frac{\omega_x}{2}x^2 + V_4 x^4 + \frac{\omega_y}{2}y^2 \ . \tag{24}$$

We take the following values of the parameters $\omega_x = 1$, $\omega_y = 1$ $V_4 = 0.05$. As initial condition, we have considered a Gaussian beam localized around the left minimum of the potential profile and initial momentum $\mathbf{p} = (0, 1)$. The results of the simulations are shown in Fig. 1, where the panels refer to different times, $t = 1, 2, 3.5, 6$. In our simulation, we have solved the system of Eqs. (19)–(22) by choosing a cutoff on n_1 and n_2. More precisely, we evaluate the following parameters: a_{n_1,n_2}, χ_{n_1,n_2} with $0 \leq n_1 \leq 3$, $0 \leq n_2 \leq 3$. We plot by solid blue curves the contour of the solution. In order to follow the evolution of the particle, we have represented the trajectory of the mean particle position by a solid light blue line. In order to appreciate the difference between the classical and the quantum corrected dynamics, we have depicted the classical trajectories obtained by solving the Newton equation by red solid lines. Our simulations show the relevance of the tunneling effect on the localization of the particle inside the two well structure. We see that in the classical case the particle has not enough energy to overcome the potential barrier. The classical trajectory is confined in the left well. Quantum calculations show that the particle tends to bounce back and forth between the two potential minima. This indicate, as expected, that there exists a range of pressure in which, according to classical dynamics, the proton is trapped in a minima near one of the oxygen atoms, while, according to quantum dynamics, the proton can be found on the left or on the right well with similar probability.

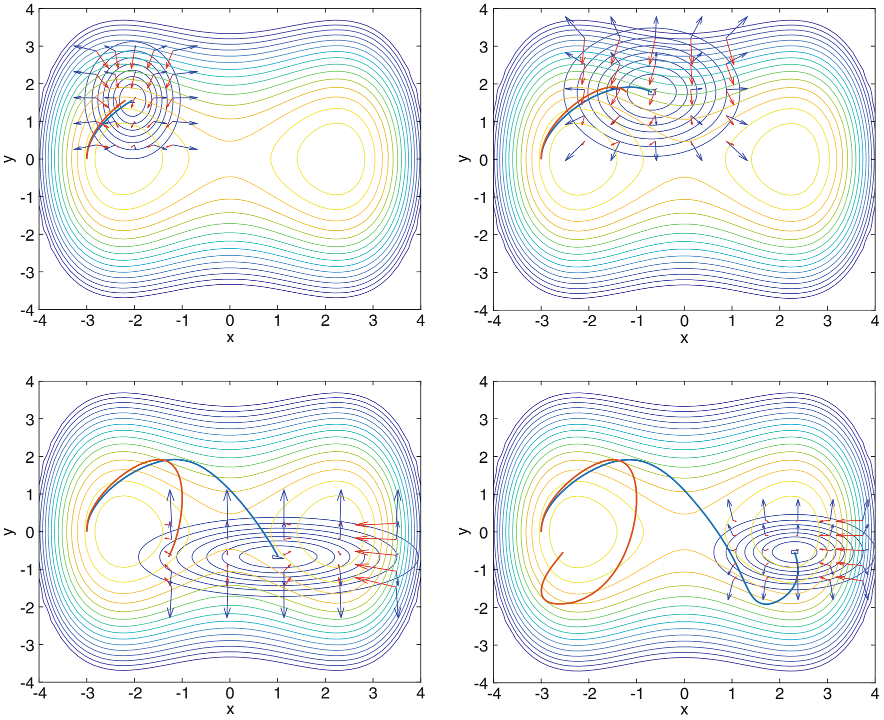


Fig. 1 Evolution of a initially localized Gaussian pulse inside the potential profile (24) (Coloured lines represent the contour plot of U). The panels refer to different times: upper-left $t = 1$ upper-right $t = 2$ bottom-left $t = 3.5$ bottom-right $t = 6$. The contour plot of the solution is depicted by blue lines, the trajectory of the centre of the wave function and the trajectory of the classical motion are depicted by, respectively, light blue and red lines. Arrows depict the classical force field (red arrows) and the Bohm force field (blue arrows)

By looking at the Bohm force we see some interesting features of the quantum behaviour of the nearly localized particles. In Fig. 1 we have represented by red arrows the classical force field obtained by the taking the gradient of the potential given in Eq. (24). According to the Bohm description of quantum mechanics, the quantum motion of a particle can be understood in terms of the evolution of a fluid whose volume elements move along integral curves of the total force field $\mathbf{F} = \nabla(U + Q)$, where Q is the Bohm potential defined in Eq. (4). This point of view provides a simple interpretation of the Bohm force field and can help to visualize the particle motion. In our plots, the gradient of the Bohm potential is represented by blue arrows. According to the Bohm framework every material point is accelerated by two force fields, the classical force and the Bohm one. In particular, the simulations show that the Bohm field is responsible of the spread of the particle wave packed, which allows the particle to overcome the potential barrier.

4 Conclusions

We have presented a quantum model for particles characterized by Gaussian wave packets. The oscillations of the wave function around the mean particle positions are represented by Hermite polynomials. The particle motion is described by a set of time dependent parameters. Our approach shows an interesting connection with the description of the particle motion provided by the Bohm theory. We have applied our method to investigate the motion of a nearly localized particle in a 2D confining structure.

Acknowledgement This work was supported by National Group of Mathematical Physics (GNFM-INdAM).

References

1. Maddox, J.B., Bittner, E.R.: J. Chem. Phys. **115**, 6309 (2001)
2. Abedi, A., Maitra, N.T., Gross, E.K.U.: Phys. Rev. Lett. **105**, 123002 (2010)
3. Sawada, S.I., Nitzan, A., Metiu, H.: Phys. Rev. B **32**(2), 851 (1985)
4. Wang, L., Zhang, Q., Xu, F., Cui, X.-D., Zheng, Y.: Int. J. Quantum Chem. **115**, 208 (2015)
5. Basile, F.E., Curchod, U.R., Tavernelli, I.: Chem. Phys. Chem. **14**, 1314 (2013)
6. Horowitz, J.M.: Phys. Rev. E **85**, 031110 (2012)
7. Poirier, B.: Trajectory-based derivation of classical and quantum mechanics. In: Hughes, K.H., Parlant G. (eds.) Quantum Trajectories, CCP6. Daresbury Laboratory, Warrington (2011)
8. Singer, K.: Mol. Phys. **85**, 701 (1995)
9. Morandi, O.: J. Phys. A: Math. Theor. **43**, 365302 (2010)
10. Morandi, O.: J. Math. Phys. **53**, 063302 (2012)
11. Sellier, J.M., Nedjalkov, M., Dimova, I.: Phys. Rep. **577**, 1 (2015)
12. Muscato, O., Wagner, W.: SIAM J. Sci. Comput. **38**(3), 1483 (2016)
13. Bohm, D.: Phys. Rev. **85**, 166 (1952)
14. Tully, J.C.: J. Chem. Phys. **93**, 1061 (1990)
15. Wyatt, R.E.: Quantum Dynamics with Trajectories: Introduction to Quantum Hydrodynamics. Springer, New York (2005)
16. Beck, M.H., Jäckle, A., Worth, G.A., Meyer, H.-D.: Phys. Rep. **324**, 1105 (2000)
17. Goldfarb, Y., Degani, I., Tannor, D.J.: J. Chem. Phys. **125**, 231103 (2006)
18. Coco, M., Mascali, G., Romano, V.: J. Comput. Theor. Transp. **45**(7), 540 (2016)
19. Ceriotti, M., Bussi, G., Parrinello, M.: Phys. Rev. Lett. **103**, 030603 (2009)
20. Glauber, R.J.: Phys. Rev. **131**, 2766 (1963)
21. Perelomov, A.: Generalized Coherent States and Their Applications. Springer, Berlin (1986)
22. Klauder, J.R., Skagerstam, B.: Coherent States. World Scientific, Singapore (1985)
23. Jin, S., Wei, D., Yin, D.: J. Comput. Appl. Math. **265**, 199 (2014)
24. Heller, E.J.: Acc. Chem. Res. **39**, 127 (2006)
25. Heller, E.J.: J. Chem. Phys. **62**, 1544 (1975)
26. Morandi, O.: J. Phys. A: Math. Theor. **51**, 255301 (2018)
27. Sprengel, M., Ciaramella, G., Borzì, A.: SIAM J. Math. Anal. **49**(3), 1681 (2017)
28. Bronstein, Y., Depondt, P., Finocchi, F., Saitta, A.M.: Phys. Rev. B **89**, 214101 (2014)
29. Bronstein, Y., Depondt, P., Bove, L.E., Gaal, R., Saitta, A.M., Finocchi, F.: Phys. Rev. B **93**, 024104 (2016)

Wigner Monte Carlo Simulation of a Double Potential Barrier

Orazio Muscato

Abstract The Wigner transport equation can be solved stochastically by Monte Carlo simulations, based on the generation and annihilation of particles. This creation mechanism has been recently understood in terms of the Markov jump process, producing new stochastic algorithms. One of this has been used to investigate the quantum transport through a double potential barrier.

1 Introduction

The Wigner transport equation is a full quantum transport model able to capture the relevant physics in next generation of quantum semiconductor devices. It is well known that the pure state Wigner equation is an equivalent phase-space reformulation of the Schrödinger equation. At the same time the Wigner equation can be augmented by a Boltzmann-like collision operator accounting for the process of decoherence. However, this equation has represented a numerically daunting task and it has raised more problems than solutions.

In literature, we find a range of proposed techniques to tackle this problem. Deterministic solvers, based on finite difference method (FDM) for time-dependent Wigner simulations, has been introduced for the first time in the mid 1980s (see [2] for a review). In order to tackle the oscillatory components introduced by the Wigner kernel and the diffusion term, recently more sophisticated solvers have been developed [1, 4, 6, 7, 13, 14, 16].

In alternative particle Monte Carlo (MC) techniques can be introduced (see [12] for a review). In this paper, we have focused in the so called *Signed Monte Carlo method* [11], where the Wigner potential is treated as a scattering source which

O. Muscato (✉)
Dipartimento di Matematica e Informatica, Università degli Studi di Catania, Catania, Italy
e-mail: orazio.muscato@unict.it

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_15

determines the electron-potential interaction, and consequently new particles with different signs are stochastically added to the system. Recently this method has been also be understood in terms of the Markov jump process theory [8–10, 15], producing a class of new stochastic algorithms. In this paper a thorough validation of one of these algorithms will be presented in the already traditional benchmark experiment of a double well potential barrier.

2 The Wigner Function Formalism

The Wigner formulation of quantum mechanics offers a description of the electron state in terms of a phase-space function $f_w(x, k, t)$, where $x \in \mathbb{R}^d$ is the particle position, $k \in \mathbb{R}^d$ the wave vector (and $\hbar k$ the momentum). The Wigner equation has the form

$$\frac{\partial f_w}{\partial t} + \frac{\hbar}{m^*} k \cdot \nabla_x f_w = \mathcal{Q}(f_w) \tag{1}$$

which includes the quantum evolution term

$$\mathcal{Q}(f_w) = \int V_w(x, k - k') f_w(x, k') \, dk' \tag{2}$$

where V_w is the Wigner potential

$$V_w(x, k) = \frac{1}{i\hbar(2\pi)^d} \int dx' \, e^{-ik\cdot x'} \left[V\left(x + \frac{x'}{2}\right) - V\left(x - \frac{x'}{2}\right) \right] \tag{3}$$

and $V(x)$ the potential energy. The Wigner potential is a non-local potential operator which is responsible of the quantum transport, is real-valued, and anti-symmetric with respect to k. The solution f_w is real valued, but not necessarily nonnegative, and it can therefore not be interpreted as a probability density, but as a *quasi-distribution* of particles. It is related to the solution of the Schrödinger equation $\psi(x, t)$ via the Wigner-Weyl transform [5] and, under some restrictions on ψ , the function f_w satisfies

$$n(x, t) = \int f_w(x, k) \, dk = |\psi(x, t)|^2 \geq 0 \tag{4}$$

where n is the mean density.

3 The Signed Particle Monte Carlo Method

Among MC solution techniques, we have considered the so called *Signed particle Monte Carlo approach* [11]. This technique is based on the observation that the quantum evolution term (2) looks like the *Gain* term of a collisional operator in which the *Loss* term is missing. But the Wigner potential (3) is not always positive and cannot be considered a scattering term. For this reason, it can be separated into a positive and negative parts V_w^+, V_w^- such that

$$V_w = V_w^+ - V_w^-, \quad V_w^+, V_w^- \geq 0 \quad . \tag{5}$$

In this way, we can define an integrated scattering probability per unit time as

$$\gamma(x) = \int dk' \, V_w^+(x, k-k') = \int dk' \, V_w^-(x, k-k') \tag{6}$$

and rewrite the quantum evolution term as the difference between *Gain* and *Loss* terms, i.e.

$$\mathcal{Q}(f_w) = \int dk' w(k', k) f_w(x, k') - \gamma(x) f_w(x, k) \tag{7}$$

$$w(k', k) = V_w^+(x, k-k') - V_w^-(x, k-k') + \gamma(x)\delta(k-k'). \tag{8}$$

The term $w(k', k)$ defines a new scattering mechanism whose effect is the production of pair of signed particles. An initial parent particle with sign u evolves on a free-flight trajectory and, according to a generation rate given by the function $\gamma(x)$, two new signed particles, one positive and one negative, are generated with one momentum state generated with a probability equal to $V_w^+(x, k)/\gamma(x)$. The main drawback of such technique is the exponential grow of the particle number.

To limit this number, a cancellation procedure must be introduced in such a way, if the total number of particles exceeds a certain bound N_{canc}, then pairs of particles with similar positions and wave vectors, but with opposite signs, are removed from the system. Recently, the previous creation process has been understood in terms of the Markov jump process theory, providing that functionals of the solution of the Wigner equation (1) are expressed in terms of the particle system [15].

Moreover a time-splitting scheme has been introduced in [9] in order to separate the transport (i.e. movement in the position space) and the generation process, as follows

1. **Transport step**

   All particles $(x_i(t), k_i(t), u_i(t))$ move according to

$$x_i \rightarrow x_i + \Delta t \frac{\hbar}{m^*} k_i \tag{9}$$

   The components u_i and k_i do not change.

2. **Generation step**
   According to probabilistic rules (to be determined below), all particles create new particles that are added to the system.
3. **Cancellation step**
   If the total number of particles exceeds a certain bound,

$$N > N_{canc} \tag{10}$$

   then pairs of particles with similar positions and wave-vectors, but with opposite signs,are removed from the system.

In the following, we shall introduce a cutoff $c > 0$ which assures finiteness of the integrals (6) with respect to the wave vector. The generation step is based on a majorant $\hat{V}_w(x,k)$ of the Wigner potential (3)

$$|V_w(x,k)| \le \hat{V}_w(x,k) \quad \forall x, k \tag{11}$$

and the creation rate

$$\hat{\gamma}(x,k) = \frac{1}{2}\int_{-c}^{c} \hat{V}_w(x,k)\, dk. \tag{12}$$

The splitting time step Δt is assumed to satisfy

$$\Delta t < \left[\sup_x \hat{\gamma}(x,c)\right]^{-1}. \tag{13}$$

This is the generation algorithm:

s0. For each j-th particle
s1. Let be $r \in U[0,1]$

$$\text{if} \quad r < 1 - \hat{\gamma}(x_j,c)\Delta t \tag{14}$$

do not create anything, next particle GOTO s0.
s2. Otherwise generate a random parameter $\tilde{k}$ uniformly on the interval $[-c,c]$.
s3. Let be $r \in U[0,1]$

$$\text{if} \quad r < 1 - \frac{|V_w(x_j,\tilde{k})|}{\hat{V}_w(x_j,\tilde{k})} \tag{15}$$

do not create anything, next particle GOTO 2.0
s4. Otherwise generate a couple of particle

$$(x_j, k_j + \tilde{k}, \tilde{u}) \quad , \quad (x_j, k_j - \tilde{k}, -\tilde{u}) \quad , \quad \tilde{u} = u_j \,\text{sign}\left[\frac{V_w(x_j,\tilde{k})}{\hat{V}_w(x_j,\tilde{k})}\right] \tag{16}$$

next particle GOTO s0.

4 The Double Potential Barrier Benchmark

Let us introduce the double barrier, which is symmetric with respect to the y-axis, the height is a (in eV), the centers in $\pm|d|$, and the amplitude b (see Fig. 1). The well potential length is

$$b_w = 2\left(|d| - \frac{b}{2}\right) \quad . \tag{17}$$

In the double barrier case we have:

$$V_w(x,k) = \frac{4a}{\hbar\pi k} \sin(2kx)\ \sin(kb)\ \cos(2k|d|) \tag{18}$$

$$\hat{V}_w(x,k) = \frac{4ab}{\hbar\pi} \quad , \quad \hat{\gamma}(x,c) = \frac{4ab}{\hbar\pi}c \quad , \quad \Delta t < \frac{\hbar\pi}{4abc} \tag{19}$$

and we have considered a double barrier with the following parameters:

$$b = 2.2\,\text{nm}, \quad |d| = 5.1\ \text{nm}, \quad b_w = 8\,\text{nm}. \tag{20}$$

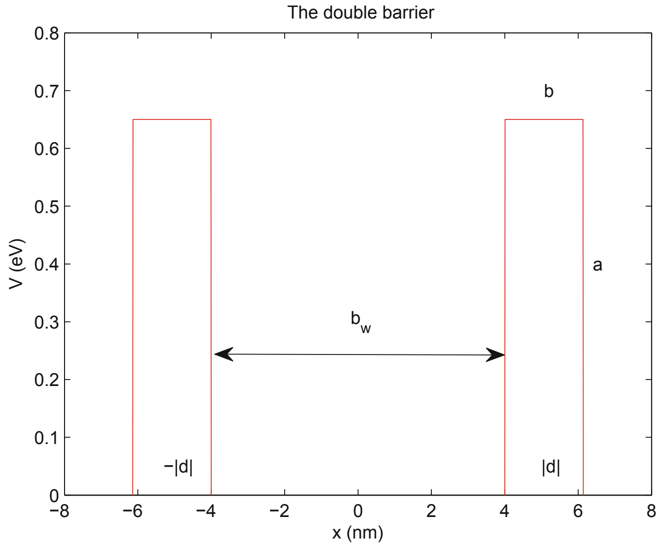


Fig. 1 The double barrier

If the barrier height a goes to infinity, the Schrödinger equation gives a well known analytic solution (see Appendix) where, the lowest energy eigenstate with $l = 1$ is

$$E_1 = \frac{\pi^2 \hbar^2}{2m^* b_w^2} = 1.8361\, 10^{-2}\text{eV} \tag{21}$$

and the corresponding eigenfunction $\psi(x,t)$ is shown in (35). According to Eq. (4), we have (for the infinite well)

$$n(x,t) = |\psi(x,t)|^2 = \frac{2C^2}{b_w} \cos^2\left(\frac{\pi x}{b_w}\right) . \tag{22}$$

The previous solution does not depend on time, consequently if the system is prepared initially in this state, then it will remain in it throughout.

Our goal is to study how the our Wigner MC scheme, for large values of the barrier height, approaches to the Schrödinger one in the case of infinite well.

In particular we want to reproduce the density solution (22) for the lowest energy eigenstate. To achieve that, we have to use the right initial condition for the Wigner equation. Since, for pure state, Schrödinger and Wigner equations are equivalent, we have taken as initial condition the Wigner function corresponding to the Schrödinger one in the lowest energy state [3].

In the x-space we have considered an uniform mesh $[-20, 20]$ (nm) with N_x = 200 grid-points; also in the k-space we have an uniform mesh $[-7.78, 7.78]$ (nm^{-1}) with $N_k = 256$. We have chosen absorption boundary conditions, i.e. if a particle is out of the mesh then it is erased. The cutoff has been fixed $c = 7.68$ nm^{-1}, the initial particle number is N_{ini} = 160,000, the cancellation parameter N_{canc} = 480,000.

We plot in Fig. 2 the density (4) at the simulation times of 20 fsec, obtained with $a = 0.3, 0.6, 1.2, 2.4, 4.8$ eV, and compared with the solution (22). The penetration into the barrier decreases as the height a increases, while the MC solution approaches to the Schrödinger one.

Due to the limitation on the time step Δt (19), the higher the barrier height a, the smaller the time step will be. Consequently a huge CPU consumption is measured, as shown in Table 1, which limits *de facto* the convergence of the MC solution to the analytic one. The results presented have been obtained using an AMD Phenom II X6 1090T 3.2 GHz and 8 Gb RAM.

Table 1 CPU and particle number at the final simulation time 20 fs

a (eV)	Δt (fs)	Final N_{part}	CPU (s)
0.3	0.05	420,000	4691
0.6	0.05	590,000	7482
1.2	0.025	749,000	16,758
2.4	0.0125	780,000	31,659
4.8	0.00625	1,200,000	90,060

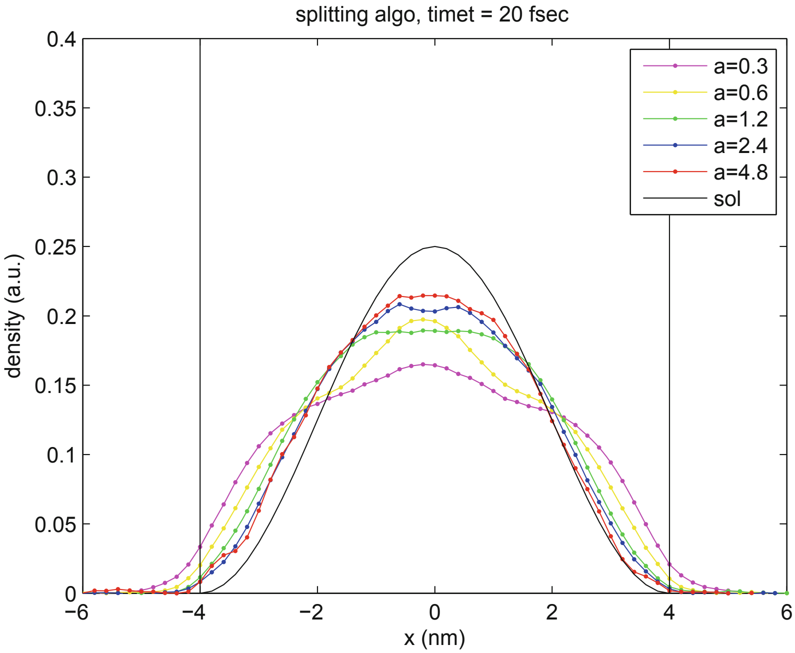


Fig. 2 The particle density (4) versus position a $t = 20\,\text{fs}$, for same values of the barrier height a and the corresponding Schrödinger solution for an infinite barrier

5 Conclusions

The Wigner equation has been solved by using the Signed particle Monte Carlo method, where new pair of particles characterized by a sign are created randomly and added to the system. This creation mechanism has been recently understood in terms of the Markov jump process, producing a class of new stochastic algorithms [9]. One of these algorithms has been implemented and applied to the double potential barrier benchmark, and compared to the corresponding analytic solution of the Schrödinger equation for the infinite well. The results show that the MC data, for large values of the barrier height, converges to the analytic solution of the infinite well, but the accuracy is limited by a huge computational effort.

Acknowledgments We acknowledge the support of the project "Modellistica, simulazione e ottimizzazione del trasporto di cariche in strutture a bassa dimensionalità", Università degli Studi di Catania—Piano della Ricerca 2016/2018 Linea di intervento 2.

Appendix

Let us consider a particle of mass m^* inside a box, surrounded by an infinite square well with potential given by

$$V(x) = \begin{cases} 0, & |x| \leq b_w/2 \\ \\ \infty & |x| > b_w/2 \end{cases} \tag{23}$$

We want to describe the motion of this particle free to move in a small space surrounded by impenetrable barriers. To do that, let us introduce Schrödinger equation

$$i\hbar \frac{\partial \psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m^*} \Delta_x \psi(x,t) + V(x)\psi(x,t) \tag{24}$$

where V the potential energy, and Δ_x the Laplace operator, and the wavefunction $\psi(x,t)$ satisfying the condition

$$\int_{\mathbb{R}} |\psi(x,t)|^2 \, dx = 1. \tag{25}$$

In the case of infinite square-well potential, the natural boundary conditions for Eq. (24) are

$$\psi(-b_w/2, t) = \psi(b_w/2, t) = 0 \tag{26}$$

and we want to solve the following problem

$$\begin{cases} i\hbar \frac{\partial \psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m^*} \Delta_x \psi(x,t) \\ \\ \psi(-b_w/2, -t) = \psi(b_w/2, t) = 0 \\ \\ V(x) = 0 \quad \forall x \in [-b_w/2, +b_w/2] \end{cases} \tag{27}$$

Since the potential does not depend on the time, we can look for solutions of the kind

$$\psi(x,t) = \phi(x)\, \chi(t) \tag{28}$$

then we have

$$\frac{i\hbar}{\chi} \frac{\partial \chi(t)}{\partial t} = -\frac{\hbar^2}{2m^*} \frac{1}{\phi} \Delta_x \phi(t,x) + V(x) = const. = E \tag{29}$$

obtaining two separate equations:

$$\frac{i\hbar}{\chi}\frac{\partial\chi(t)}{\partial t} = E, \quad -\frac{\hbar^2}{2m^*}\frac{1}{\phi}\Delta_x\phi(t,x) + V(x) = E. \tag{30}$$

The solution of the first equation is:

$$\chi(t) = C\,\exp(-iEt/\hbar) \tag{31}$$

whereas the second equation yields

$$\left[-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial x^2} + V(x)\right]\phi(x) = E\,\phi(x) \tag{32}$$

which gives the particle energy spectrum. For the square-well barrier potential (23), using the b.c. (26), one obtains a discrete energy spectrum

$$E_l = \frac{\hbar^2 k_l^2}{2m^*}, \quad k_l = \frac{\pi l}{b_w} \quad l \in \mathbb{N} \tag{33}$$

$$\phi_l(x) = \sqrt{\frac{2}{b_w}}\sin\left(\frac{\pi l x}{b_w}\right) \quad l \text{ even}, \quad \phi_l(x) = \sqrt{\frac{2}{b_w}}\cos\left(\frac{\pi l x}{b_w}\right) \quad l \text{ odd} \tag{34}$$

and, for l odd, we have:

$$\psi(x,t) = C\exp\left(-\frac{iE_l t}{\hbar}\right)\phi_l(x) = \psi_1(x,t) + i\,\psi_2(x,t) \tag{35}$$

$$\psi_1(x,t) = C\sqrt{\frac{2}{b_w}}\cos\left(\frac{\pi l x}{b_w}\right)\cos\left(\frac{\hbar\pi^2 l^2}{2m^* b_w^2}\,t\right), \tag{36}$$

$$\psi_2(x,t) = -C\sqrt{\frac{2}{b_w}}\cos\left(\frac{\pi l x}{b_w}\right)\sin\left(\frac{\hbar\pi^2 l^2}{2m^* b_w^2}\,t\right). \tag{37}$$

Finally the density is

$$n(x,t) = |\psi(x,t)|^2 = \psi_1^2(x,t) + \psi_2^2(x,t) = \frac{2C^2}{b_w}\cos^2\left(\frac{\pi l x}{b_w}\right) \quad l \text{ odd}. \tag{38}$$

References

1. Dorda, A., Schürrer, F.: A WENO-solver combined with adaptive momentum discretization for the Wigner transport equation and its application to resonant tunneling diodes. J. Comput. Electr. **284**, 95–116 (2015)
2. Kosina, H.: Wigner function approach to nano device simulation. Int. J. Comput. Sci. Eng. **2**(3–4), 100–118 (2006)
3. Lee, H.-W., Scully, M.O.: The Wigner phase-space description of collision processes. Found. Phys. **13**(1), 61–72 (1983)
4. Lee, J.-H., Shin, M.: Quantum transport simulation of nanowire resonant tunneling diodes based on a Wigner function model with spatially dependent effective masses. IEEE Trans. Nanotechnol. **16**(6), 1028–1036 (2017)
5. Markovich, P.A., Ringhofer, C.A., Schmeiser, C.: Semiconductor Equations. Springer, New York (1990)
6. Morandi, O., Demeio, L.: A Wigner-function approach to interband transitions based on the multiband-envelope-function model. Transp. Theor. Stat. Phys. **37**(5–7), 473–459 (2008)
7. Morandi, O., Schürrer, F.: Wigner model for quantum transport in graphene. J. Phys. A: Math. Theor. **26**, 265301 (2011)
8. Muscato, O.: A benchmark study of the signed-particle Monte Carlo algorithm for the Wigner equation. Commun. Appl. Ind. Math. **8**(1), 237–250 (2017)
9. Muscato, O., Wagner, W.: A class of stochastic algorithms for the Wigner equation. SIAM J. Sci. Comput. **38**(3), A1438–A1507 (2016)
10. Muscato, O., Wagner, W.: A stochastic algorithm without time discretization error for the Wigner equation. Kin. Rel. Models **12**(1), 59–77 (2019)
11. Nedjalkov, M., Kosina, H., Selberherr, S., Ringhofer, C., Ferry, D.K.: Unified particle approach to Wigner-Boltzmann transport in small semiconductor devices. Phys. Rev. B **70**, 115319 (2004)
12. Querlioz, D., Dollfus, P.: The Wigner Monte Carlo Method for Nanoelectronic Devices. Wiley, Hoboken (2010)
13. Shao, S., Lu, T., Cai, W.: Adaptive conservative cell average spectral element methods for transient Wigner equation in quantum transport. Commun. Comput. Phys. **9**(3), 711–739 (2011)
14. Van de Put, M.L., Soree, B., Magnus, W.: Efficient solution of the Wigner-Liouville equation using a spectral decomposition of the force field. J. Comput. Phys. **350**, 314–325 (2017)
15. Wagner, W.: A random cloud model for the Wigner equation. Kinet. Rel. Models **9**(1), 217–235 (2016)
16. Xiong, Y., Chen, Z., Shao, S.: An advective-spectral-mixed method for time-dependent many-body Wigner simulations. SIAM J. Sci. Comput. **38**(4), B491–B520 (2016)

Simulation of Graphene Field Effect Transistors

Giovanni Nastasi and Vittorio Romano

Abstract Field effects transistors, where the active region is constituted by a single layer of graphene, are simulated and the characteristic curves are shown. The current–voltage curves present a behaviour different from that of devices made of classical semiconductors, like Si or GaAs, because of the zero gap in monolayer graphene. The current is no longer a monotone function of the gate voltage but there exists an inversion gate voltage corresponding to which the type of majority carriers changes. Usually the considered devices are investigated by adopting reduced one dimensional models with some averaging procedure. Here a full two-dimensional simulation is presented. The model is based on a system of drift-diffusion equations for electrons and holes. The numerical method is based on the Scharfettel and Gummel scheme. A special treatment of the Poisson equation is adopted for taking into account the charge in the graphene sheet. The characteristic curves for fixed gate voltages and for fixed source-drain voltages have been obtained.

1 Introduction

In the last years an increasing interest has been devoted to graphene field effect transistors (GFETs) as potential candidates for high-speed analog electronics, where transistor current gain is more important than ratio current ON/current OFF [1]. Several types of GFETs will be studied and optimized: top-gated graphene based transistors, obtained synthesizing graphene on silicon dioxide wafer, and double gate GFETs. See [2] for a comprehensive review.

G. Nastasi (✉) · V. Romano
Department of Mathematics and Computer Science, Università degli Studi di Catania, Catania, Italy
e-mail: g.nastasi@unict.it; romano@dmi.unict.it

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_16

The current–voltage curves present a behaviour different from that of devices made of classical semiconductors, like Si or GaAs, because of the zero gap in monolayer graphene. The current is no longer a monotone function of the gate voltage but there exists an inversion gate voltage corresponding to which the type of majority carriers changes. This introduces a certain degree of uncertainty in the determination of the current-off regime which requires a rather well tuning of the gate-source voltage.

Usually the considered type of devices are investigated by adopting reduced one dimensional models with some averaging procedure [3, 4]. Here a full two-dimensional simulation is presented.

For the simulation of charge transport in graphene, drift-diffusion models are available in the literature. Recently, more accurate hydrodynamical models have been formulated [5–7] by exploiting the maximum entropy principle. Another approach is to get mobility models by simulations based on deterministic solutions of the semiclassical Boltzmann equation for electrons in graphene by using discontinuous Galerkin (DG) method or based on Monte Carlo simulations [8, 9].

Here, as first instance, we compare the results of the simulations for top-gated GFETs obtained with the drift-diffusion model suggested in [10]. The numerical method is based on the Scharfettel and Gummel scheme. A special treatment of the Poisson equation is adopted for taking into account the charge in the graphene sheet. In particular, the characteristic curves both for fixed gate voltages and for fixed source-drain voltages, are studied.

2 Mathematical Model

The mathematical model we adopt to simulate the charge transport in graphene is the bipolar drift-diffusion in 1D case,

$$\frac{\partial n}{\partial t} - \frac{1}{e}\frac{\partial}{\partial x}\left(\mu_n k_B T_L \frac{\partial n}{\partial x} - en\mu_n \frac{\partial \phi}{\partial x}\right) = 0,$$

$$\frac{\partial p}{\partial t} + \frac{1}{e}\frac{\partial}{\partial x}\left(-\mu_p k_B T_L \frac{\partial p}{\partial x} - ep\mu_p \frac{\partial \phi}{\partial x}\right) = 0,$$

where $n(t, x)$, $p(t, x)$ are the graphene electron density and hole density respectively, e is the positive elementary charge, k_B is the Boltzmann constant, T_L is the lattice temperature (kept constant), $\mu_n(x)$ and $\mu_p(x)$ are the mobility models for electrons and holes respectively and $\phi(x, y)$ is the electric potential. We adopt the mobility model proposed in Ref. [10] given by

$$\mu_s(x) = \frac{\nu_s}{[1 + (\nu_s E/v_{sat})^\gamma]^{1/\gamma}},$$

where $E = |\partial\phi/\partial x|$ is the absolute value of the x-component of the electric field, v_{sat} is the saturation velocity (we take the value 0.2 µm/ps), $\gamma \approx 2$ and

$$\nu_s(x) = \frac{\mu_0}{(1 + s/n_{ref})^{\alpha}},$$

where $\mu_0 = 0.4650$ µm^2/V ps is the graphene low field mobility, $n_{ref} = 1.1 \times 10^5$ µm^{-2} and $\alpha = 2.2$. The symbol s indicates the carrier density: $s = n$ for electrons and $s = p$ for holes.

The 2D Poisson equation for the electric potential reads

$$\nabla \cdot (\epsilon \nabla \phi) = h(x, y),$$

where

$$h(x, y) = \begin{cases} e(n(x) - p(x))/t_{gr} & \text{if} \quad y = y_{gr} \\ 0 & \text{if} \quad y \neq y_{gr} \end{cases}$$

being y_{gr} the y-coordinate of the graphene sheet (see Fig. 1), and ϵ is given by

$$\epsilon(x, y) = \begin{cases} \epsilon_{gr} & \text{if} \quad y = y_{gr} \\ \epsilon_{ox} & \text{if} \quad y \neq y_{gr} \end{cases}$$

where ϵ_{gr} and ϵ_{ox} are the dielectric constants of the graphene and oxide respectively. t_{gr} is the width of the graphene layer (the width of a single atom) which is assumed to be equal to 1 nm. The charge in the graphene layer is considered distributed in the volume enclosed by the parallelepiped of base the area of the graphene and height t_{gr}. Recall that n and p are areal densities.

3 Numerical Results

The simulated device is depicted in Fig. 1. The length is 100 nm. The width of the lower and upper oxide (SiO$_2$) is 10 nm. The source and drain contacts are long 25 nm. The two gate potentials are set as equal. At the metallic contacts the total voltage includes also the work function which depends on the specific material the contacts are made of. Different values of the work function will be considered in the simulations.

In Fig. 2 the shape of the electrical potential is plotted when the gate-source potential is 0.6 V and the source-drain-potential is 0.3 V, with a work function of 0.25 V. The impurity charge is neglected. Similar results are obtained in the other cases.

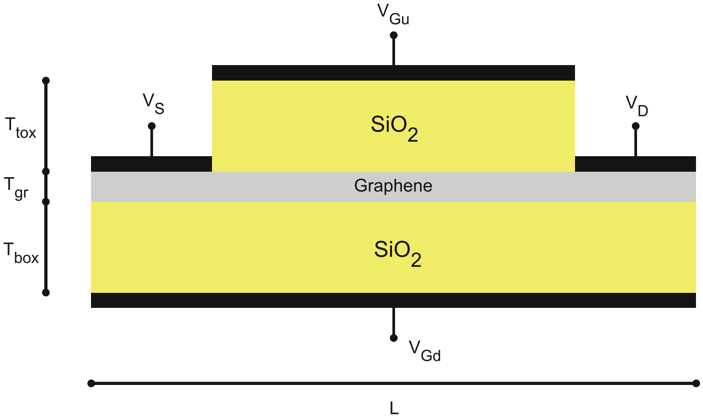


Fig. 1 Schematic representation of the simulated GFET

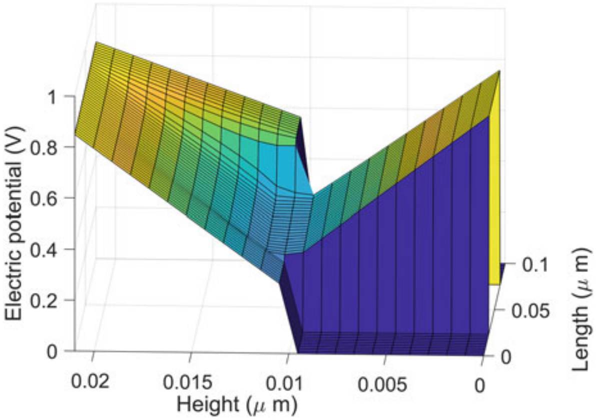


Fig. 2 Electrostatic potential when the gate-source potential is 0.6 V and the source-drain-potential is 0.3 V. The work function potential is 0.25 V

Figure 3 shows the characteristic curves of current versus gate voltage with work function equal to 0 V for several values of the source-drain voltage, neglecting the impurities. In Fig. 4 the current versus gate voltage are plotted as in Fig. 3, including also the presence of an impurity density of 3.5 $\times 10^3$ μm^{-2}, for several values of the gate voltage.

The presence of the impurities produces a small degradation of the current. The crucial issue is that the range of gate voltage where the current is off is bounded at variance with traditional semiconductors. This is due to the gapless nature of monolayer graphene. As a consequence, a fine tuning of the gate voltage is required to have an acceptable field effect transistor and in this respect accurate simulations are needed.

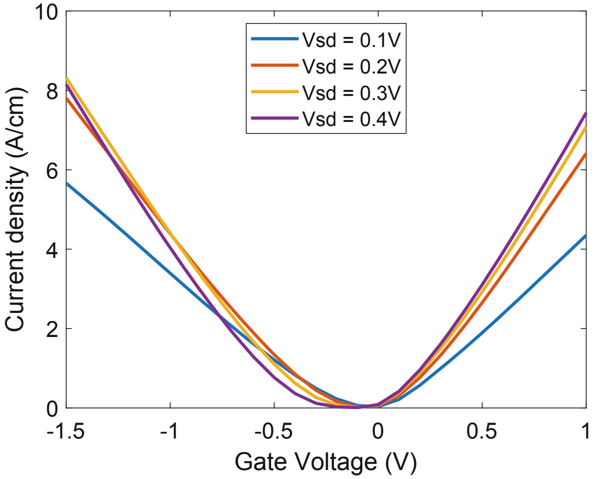


Fig. 3 Current versus gate voltage with a work function at the graphene-metal interface equal to 0 V

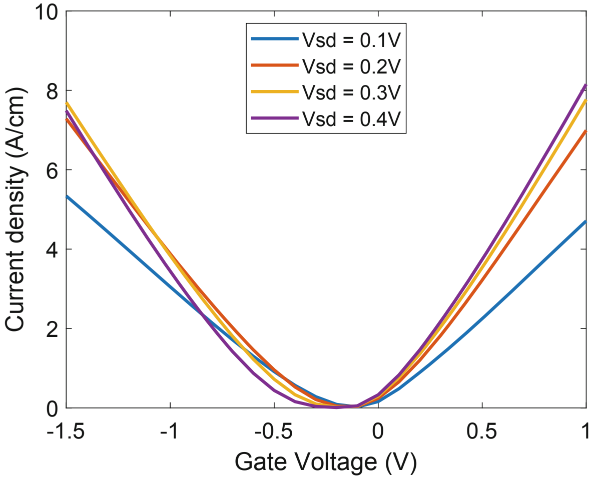


Fig. 4 As in Fig. 3 including an impurity density of $3.5 \times 10^3\ \mu\mathrm{m}^{-2}$

In Figs. 5 and 6 the previous characteristic curves are shown by considering a work function of 0.25 V. Again the presence of the impurities leads only to a small degradation of the current. Note that above the inversion voltage (about 0.25 V) the majority carriers are the electrons while below the inversion voltage the majority carriers are the holes. The behaviour of the current is very different from the traditional semiconductors like Si or GaAs on account of the zero gap in the energy band. The major issue is the difficulty of fixing the off state which requires an accurate calibration of the voltage.

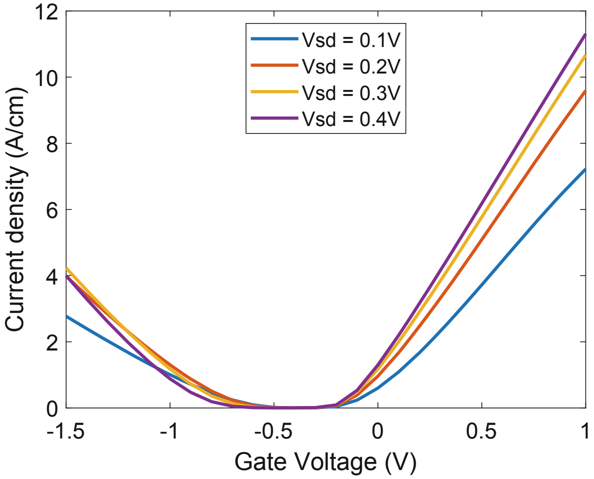


Fig. 5 Current versus gate voltage with a work function at the graphene-metal interface equal to 0.25 V

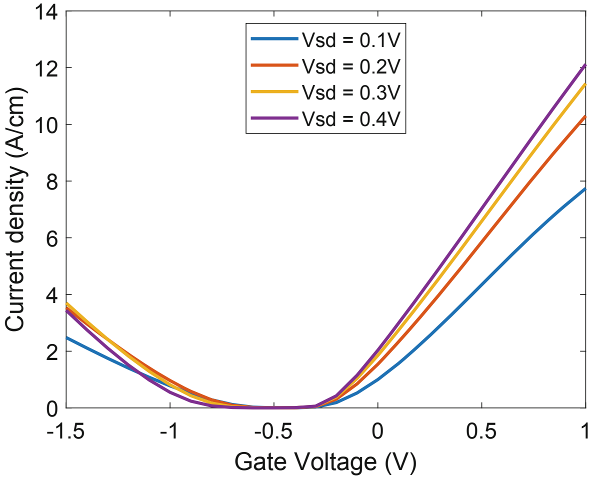


Fig. 6 As in Fig. 3 including an impurity density of 3.5 $\times 10^3$ μm^{-2}

To complete the analysis, in Figs. 7 and 8 we show the current versus the source-drain voltage for several gate-source voltages with a work function at the graphene-metal interface equal to 0.25 V.

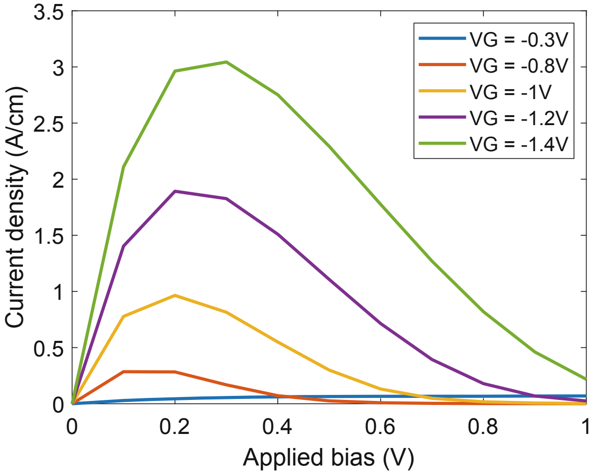


Fig. 7 Current versus source-drain voltage with a work function at the graphene-metal interface equal to 0.25 V including an impurity density of 3.5 $\times 10^3$ µm^{-2}. Negative gate-source voltages are considered

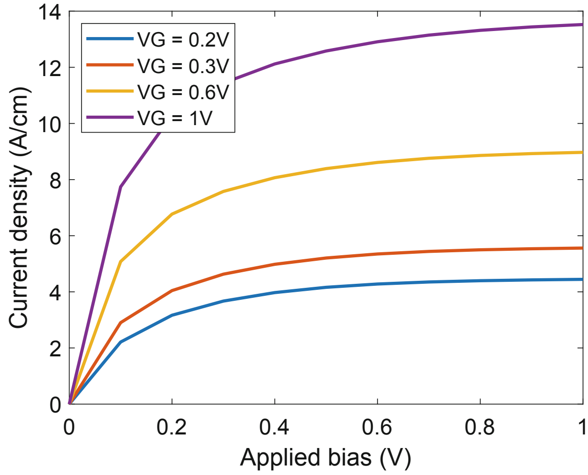


Fig. 8 As in Fig. 7 considering positive gate-source voltages

References

1. Maric, I., Han, M.Y., Young, A.F., Ozyilmaz, B., Kim, P., Shepard, K.L.: Current saturation in zero-bandgap, top-gated graphene field-effect transistors. Nat. Nanotechnol. **3**, 654–659 (2008)
2. Schwierz, F.: Graphene transistors. Nat. Nanotechnol. **5**, 487–496 (2010)
3. Jiménez, D., Moldovan, O.: Explicit drain-current model of graphene field effect transistors targeting analog and radio-frequency applications. IEEE Trans. Electron Devices **65**, 739–746 (2018)

4. Upadhyay, A.K., Kushwaha, A.K., Vishvakarma, S.K.: A unified scalable quasi-ballistic transport model of GFET for circuit simulations. IEEE Trans. Electron Devices **58**, 4049–4052 (2018)
5. Camiola, V.D., Romano, V.: Hydrodynamical model for charge transport in graphene. J. Stat. Phys. **157**, 1114–1137 (2014)
6. Mascali, G., Romano, V.: Charge transport in graphene including thermal effects. SIAM J. Appl. Math. **77**, 593–613 (2017)
7. Luca, L., Romano, V.: Comparing linear and nonlinear hydrodynamical models for charge transport in graphene based on the Maximum Entropy Principle. Int. J. Nonlinear Mech. **104**, 39–58 (2018)
8. Romano, V., Majorana, A., Coco, M.: DSMC method consistent with the Pauli exclusion principle and comparison with deterministic solutions for charge transport in graphene. J. Comput. Phys. **302**, 267–284 (2015)
9. Coco, M., Majorana, A., Romano, V.: Cross validation of discontinuous Galerkin method and Monte Carlo simulations of charge transport in graphene on substrate. Ricerche Mat. **66**, 201–220 (2016)
10. Dorgan, V.E., Bae, M.-H., Pop, E.: Mobility and saturation velocity in graphene on SiO_2. Appl. Phys. Lett. **97**, 082112 (2010)

Part III
Circuit Simulation

This part comprises three works falling in the field of circuit simulation. In the contribution *LinzFrame: A Modular Mixed-Level Simulator with Emphasis on Radio Frequency Circuits* the authors, K. Bittner et al., give an overview of the circuit and device simulator LinzFrame with an emphasis on radio frequency circuits (RF) applications. Besides SPICE-like analysis features LinzFrame offers several techniques dedicated to RF circuits. Among them are the multi-tone Harmonic Balance (HB), periodic steady state shooting method, a toolbox for autonomous circuits (oscillators), and multi-rate envelope methods. Besides transient analysis based on the BDF formulas, a toolbox for a spline-wavelet approximation has been developed. In recent time, the simulator has been extended to a circuit-device mixed-level simulator by coupling the circuit simulator to a TCAD simulator. LinzFrame permits the holistic (strong) coupling of a circuit and a device simulator, enabling either modeling of the circuit/device as lumped (concentrated) model or as a full 3D model, depending on the needed accuracy. Moreover it circumvents the prohibitive run-time of conventional transient analysis by several multi-rate techniques dedicated to RF circuits/devices.

In the electronics industry, smaller and faster electronic devices are always demanded. Full device-parasitic transient simulations of realistic circuits are time consuming or even infeasible due to a huge number of electrical components and unavoidable parasitics. In the contribution *Fast transient simulation of RC circuits with dense capacitive coupling* by N.T.K. Dang et al. a novel technique to address the problem with the presence of parasitic capacitances is devised. The selected technique activates/inactivates coupling capacitances during the transient simulation, therefore obtaining faster simulations of transients.

Modern circuit simulators predominantly use Newton-Raphson (NR) iteration to solve circuit equations. To improve NR convergence, circuit simulators use a practice called “limiting”. However, in most simulators, the implementation of limiting tends to be inflexible, non-modular, inconsistent, and confusing. In *Predictor/Corrector Newton-Raphson (PCNR): A Simple, Flexible, Scalable, Modular, and Consistent Replacement for Limiting in Circuit Simulation* by K.V. Aadithya et al. the authors propose a Predictor/Corrector Newton-Raphson (PCNR) scheme, a replacement for limiting that overcomes the mentioned disadvantages while incurring modest computational overhead. The key ideas behind PCNR are: to add each limited circuit quantity as an extra unknown to the circuit's Modified Nodal Analysis (MNA) system of equations; to split each NR iteration into a "prediction" phase followed by a "correction" phase; to mitigate the computational cost of the extra unknowns by eliminating them from all Ax = b solves using a Schur complement based technique. Example of circuits validate the robustness of the method.

LinzFrame: A Modular Mixed-Level Simulator with Emphasis on Radio Frequency Circuits

Kai Bittner, Hans Georg Brachtendorf, and Wim Schoenmaker

Abstract *Purpose*: LinzFrame is a circuit and device simulator with emphasis on radio frequency circuits (RF) applications. Slowly changing amplitudes are modulated by a carrier signal at a very high center frequency. These waveforms are referred to as multi-tone signals. RF devices are often distributed elements, i.e. their behavior cannot be adequately represented by terminal voltages and currents.
Design/Methodology: Besides SPICE-like analysis features LinzFrame offers several techniques dedicated to RF circuits. Among them are the multi-tone Harmonic Balance (HB), periodic steady state shooting method, a toolbox for autonomous circuits (oscillators), and multi-rate envelope methods. Besides transient analysis based on the BDF formulas, a toolbox for a spline-wavelet approximation has been developed. This technique combines the advantages of variable time step techniques (such as BDF) with a compact representation of signals by a set of basis functions (such as HB). In contrary to HB a spline-wavelet representation of signals with variable refinements allows for a representation of signals with sharp slew rates without the unwanted Gibbs phenomenon.
In recent time, the simulator has been extended to a circuit-device mixed-level simulator by coupling the circuit simulator to a TCAD simulator. This feature enables the co-simulation of device and circuit levels, where the critical devices are simulated and optimized in full 3D, such as distributed elements.
Originality/Value: LinzFrame enables the holistic (strong) coupling of a circuit and a device simulator, enabling either modeling of the circuit/device as lumped (concentrated) model or as a full 3D model, depending on the needed accuracy. Moreover it circumvents the prohibitive run-time of conventional transient analysis by several multi-rate techniques dedicated to RF circuits/devices.

K. Bittner · H. G. Brachtendorf (✉)
University of Applied Sciences of Upper Austria, Wels, Austria
e-mail: Kai.Bittner@fh-hagenberg.at; brachtd@fh-hagenberg.at

W. Schoenmaker
MAGWEL N.V., Leuven, Belgium
e-mail: wim.schoenmaker@magwel.com

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_17

1 The LinzFrame Circuit Simulator

The circuit simulator LinzFrame with focus on radio frequency (RF) applications [1, 2] follows a strictly modular concept as depicted in Fig. 1. The simulator kernel employs the Modified Nodal Analysis (MNA). Moreover it comprises an automatic differentiation suite [16] which simplifies the implementation of new models significantly, since partial derivatives w.r.t. the state variables required for the Jacobian calculation for Newton type methods are not coded explicitly [16]. Furthermore, model libraries for linear devices, SPICE transistor models and a stimulus library including modulated sources such as OFDM, FSK, QPSK, QAM, etc., libraries to industry relevant device models such as BSIMx, VBIC, and the Simkit library from NXP Semiconductors (MEXTRAM, MOS9, MOS11, etc.) are available. Hence, the simulator covers the majority of industry standards in circuit simulation. A Laplace model interface allows the incorporation of rational fraction transfer models obtained, e.g., from Model Order Reduction (MOR). The analysis toolbox comprises standard methods such as DC, AC and transient analysis with polynomial and trigonometric multi-step BDFx (MBDFx) methods [3] and an interface to the DASPK simulator [10] for solving higher index differential algebraic equations (DAEs). As an alternative to polynomial multi-step methods, a spline-wavelet transient simulator has been developed by the authors.

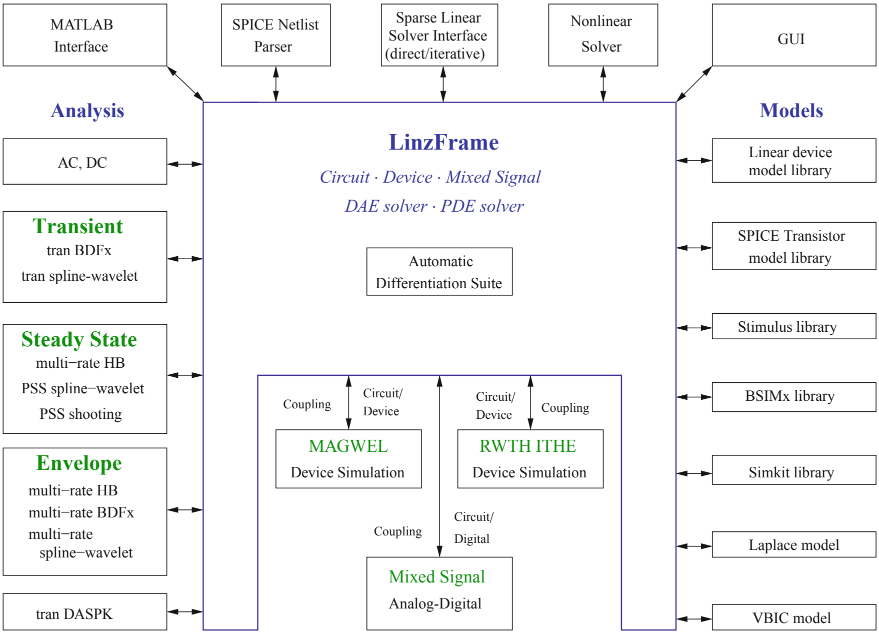


Fig. 1 Overview on the toolboxes of the circuit simulator LinzFrame

Standard transient solvers are prohibitively slow for the simulation of RF circuits. Since the time steps of multi-step integration formulas must be at least a factor of 10 smaller than the reciprocal of the highest relevant frequency, transient simulators come to their limits as the center frequencies become higher. Therefore, multi-rate simulators have been proposed to decouple the slowly varying envelope or baseband signal from the radio frequency modulation. This decoupling enables different techniques and time steps for a compact representation of the waveforms: the baseband signal can be, on the one hand, appropriately represented by multi-step integration methods whereas the periodic RF modulation on the other hand are well approximated by a trigonometric (Fourier) expansion or a spline-wavelet basis.

Several tools for multi-rate simulation [7, 8], such as Harmonic Balance (HB), BDF and spline-wavelet techniques (both algebraic and trigonometric polynomial bases) are therefore incorporated in LinzFrame. The latter technique is superior when strong nonlinearities and/or sharp transients occur, which are efficiently resolved by an adaptive mesh, whereas a trigonometric basis exhibits often the well-known Gibb's phenomenon. Periodic steady state (PSS) methods both for driven and autonomous circuits such as oscillators complete the tool [1–5].

Moreover, interfaces to numerical tools, including damped Newton solvers, homotopy methods [3], several direct sparse linear solvers (e.g. MUMPS, MA48, PARDISO, SuperLU) as well as preconditioned Krylov subspace techniques (e.g. ILUPACK) are available. For a rapid prototyping and test of novel algorithms, a MATLAB interface is at hand.

As part of the European fp7 project nanoCOPS [18, 21], the simulator has been coupled to the commercial EM/device simulator devEM from the company MAGWEL for combined EM-device/circuit simulation [21] as depicted in Fig. 2. The simulator devEM is a full 3D electro-magnetic field and device simulator, which employs as unknowns the scalar and vector potentials $(V, \mathbf{A})$. From Maxwell's equations and the device constitutive equations one obtains a system of partial differential equations (PDEs). The device simulator employs for the spartial discretization of the PDEs the Finite Integration Technique (FIT) [11, 12, 22] resulting in a huge system of ordinary DAEs. The coupling between LinzFrame and devEM is performed holistically [6, 9], that is, LinzFrame—which is the master simulator—has full excess to the Jacobian matrix stamps and the right hand side vector. This enables (damped) Newton methods with enhanced convergence properties than relaxation based techniques also reported in the literature [15]. In another ongoing DFG/FWF project, LinzFrame is coupled with a device simulator from RWTH Aachen university to study plasma oscillations in the THz range [13, 14].

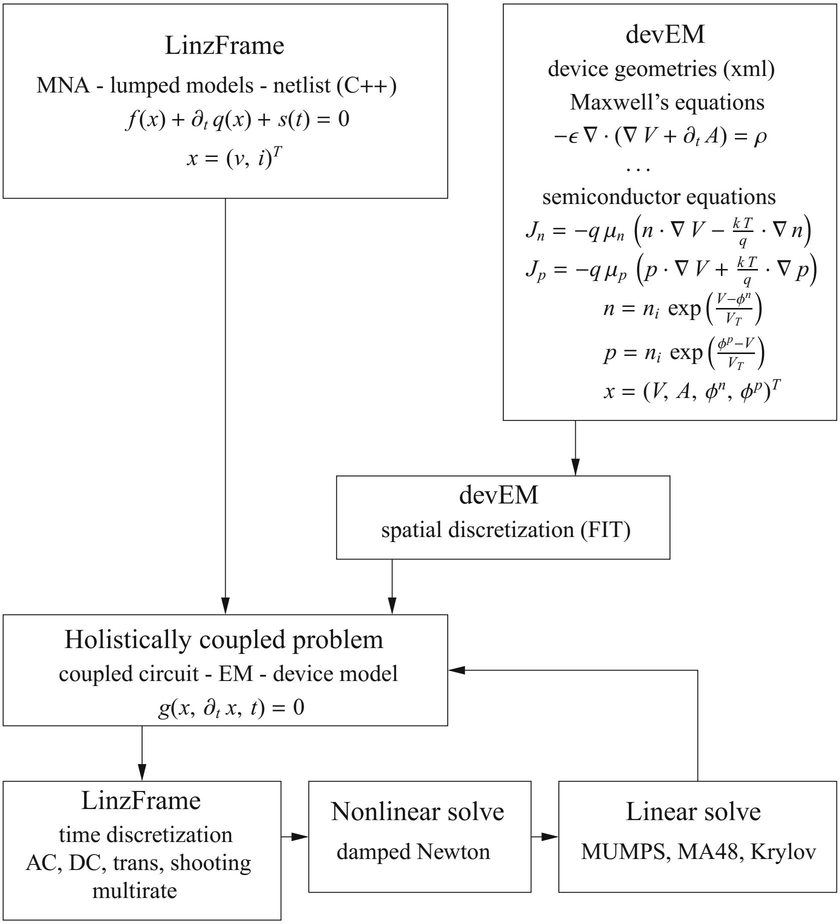

LinzFrame
MNA - lumped models - netlist (C++)
$f(x) + \partial_t q(x) + s(t) = 0$
$x = (v, i)^T$
devEM
device geometries (xml)
Maxwell's equations
$-\epsilon \nabla \cdot (\nabla V + \partial_t A) = \rho$
...
semiconductor equations
$J_n = -q \mu_n \left(n \cdot \nabla V - \frac{kT}{q} \cdot \nabla n\right)$
$J_p = -q \mu_p \left(p \cdot \nabla V + \frac{kT}{q} \cdot \nabla p\right)$
$n = n_i \exp\left(\frac{V-\phi^n}{V_T}\right)$
$p = n_i \exp\left(\frac{\phi^p-V}{V_T}\right)$
$x = (V, A, \phi^n, \phi^p)^T$
devEM
spatial discretization (FIT)
Holistically coupled problem
coupled circuit - EM - device model
$g(x, \partial_t x, t) = 0$
LinzFrame
time discretization
AC, DC, trans, shooting
multirate
Nonlinear solve
damped Newton
Linear solve
MUMPS, MA48, Krylov

Fig. 2 Concept of the holistic coupling between LinzFrame and devEM

2 Circuit-Device Simulator Coupling

From Kirchhoff's laws, the circuit topology and the device constitutive equations one obtains a system of generally nonlinear DAEs of dimension N

$$\frac{d}{dt}q\big(x(t)\big) + f\big(x(t)\big) - s(t) = 0,\ x(0) = x_0 \tag{1}$$

where $x = (v,\ i)^T$ is the vector of unknown node voltages (potentials) and some branch currents, $f : \mathbb{R}^N \to \mathbb{R}^N$ the vector sums of currents entering each node and $q : \mathbb{R}^N \to \mathbb{R}^N$ the vector sums of charges and fluxes. Moreover x_0 is the vector of initial conditions and $s(t)$ the stimulus vector, respectively. If s is independent of time, the circuit is autonomous and non-autonomous otherwise.

The electro-magnetic TCAD (EM-TCAD) simulator devEM employs both the scalar potentials V and the vector potential $\mathbf{A}$ such that the magnetic induction is $\mathbf{B} = \nabla \times \mathbf{A}$, and hence $\mathbf{E} = -(\nabla\ V + \partial_t \mathbf{A})$, where $\mathbf{E}$ is the electric field strength. Furthermore $\mathbf{D} = \epsilon\, \mathbf{E}$ is the dielectric displacement and $\mathbf{H} = \frac{1}{\mu}\, \mathbf{B}$ the magnetic field strength. To obtain systems of first-order PDEs in time, the quasi-canonical momentum $\mathbf{\Pi} = \partial_t \mathbf{A}$ is used as an additional degree-of-freedom. The simulator devEM enables both the Coulomb and Lorenz gauge (and a continuous sweep between these two). Exemplarily, the PDEs valid for semiconductors are considered next.

Let N_D, N_A be the donator/acceptor concentrations and n, p the free electron/hole concentrations, respectively. From the standard drift-diffusion equations one obtains

$$-\nabla \cdot (\epsilon\ (\nabla\ V + \mathbf{\Pi})) = \varrho,\ \varrho = q\ (p - n + N_D - N_A)$$

$$\nabla \times \left(\frac{1}{\mu}\ \nabla \times \mathbf{A}\right) = \mathbf{J}_p + \mathbf{J}_n - \epsilon\ \frac{\partial}{\partial t}\ (\nabla V + \mathbf{\Pi})$$

where $\mathbf{J}_n$, $\mathbf{J}_p$ are the currents densities of electrons/holes, given by

$$\mathbf{J}_n = -q\ \mu_n\ (n\ (\nabla\ V + \mathbf{\Pi}) - V_T \cdot \nabla\, n)$$

$$\mathbf{J}_p = -q\ \mu_p\ (p\ (\nabla\ V + \mathbf{\Pi}) + V_T \cdot \nabla p)$$

wherein q is the elementary charge, μ_n, μ_p the mobilities of electrons and holes, $V_T = \frac{k_B\, T}{q}$ the thermal voltage, k_B Boltzmann's constant and T the absolute temperature in Kelvin. The densities of electrons and holes are expressed as

$$n = n_i\ \exp\left(\frac{V - \phi^n}{V_T}\right), \quad p = n_i\ \exp\left(\frac{\phi^p - V}{V_T}\right)$$

wherein ϕ^n, ϕ^p are the quasi-Fermi potentials for electrons/holes, respectively. The continuity equation holds for the electrons and holes separately, i.e.,

$$\nabla \cdot \mathbf{J}_n - q\ \frac{\partial n}{\partial t} = -q\ U(n, p), \quad \nabla \cdot \mathbf{J}_p + q\ \frac{\partial p}{\partial t} = q\ U(n, p)$$

with net generation rate $U(n, p) = G - R$. devEM employs various generation/recombination models. The system of equations is completed with the gauge condition

$$\frac{1}{\mu}\ \nabla(\nabla \cdot \mathbf{A}) + \xi\ \epsilon \nabla\ (\partial_t V) = 0$$

For $\xi = 0$ one obtains the Coulomb and for $\xi = 1$ the Lorenz gauge as special cases. Unknowns are the scalar and vector potentials ($\mathbf{A}$, $\mathbf{\Pi} = \partial_t \mathbf{A}$) and moreover the quasi-Fermi potentials $(V,\ \mathbf{A},\ \mathbf{\Pi},\ \phi^n,\ \phi^p)^T$.

2.1 Discretization

The spatial discretization is done on an (un)structured grid using a variation of the Finite Integration Technique [12, 17, 22].

LinzFrame on the other hand is the master simulator which performs the time discretization and step size control. Besides multi-step integration formulas, specifically for radio frequency applications a multi-rate technique has been developed which decouples the slowly varying envelope or baseband signal in time scale τ and RF time scale t from the carrier signal. The underlying ordinary DAE system (1) is reformulated by a system of partial DAEs, i.e.

$$\tfrac{\partial}{\partial \tau} q\big(\hat{x}(\tau,t)\big) + \omega(\tau)\, \tfrac{\partial}{\partial t} q\big(\hat{x}(\tau,t)\big) + i\big(\hat{x}(\tau,t)\big) = \hat{s}\big(\tau,t\big)$$

where $\omega(\tau)$ is an estimate of the instantaneous frequency [19]. The signal $\hat{x}$ is assumed to be periodic in its second argument, that is $\hat{s}(\tau,t) = \hat{s}(\tau,t+P)$ with normalized period $P = 1$. The characteristic curves of the PDE are given by

$$\big(t,\ \Omega_\theta(t)\big), \quad \Omega_\theta(t) = \theta + \int_0^t \omega(s)\, ds, \quad \theta \in [0,\ P]$$

parametrized by θ. The solution of the underlying problem (1) is obtained along a specific characteristic curve through the origin, i.e. $\theta = 0$. A comprehensive documentation on the multi-rate PDE method can be found, e.g., in [1, 3, 20].

3 Results

3.1 Mixer Circuit

The mixer circuit with differential RF and oscillator inputs is depicted in Fig. 3. The input signals operate in the GHz range, whereas the center frequency of the output signal at a low intermediate frequency in the MHz range. Therefore, mixers are typical examples for which the multi-rate technique is superior compared with a classical transient analysis. Figure 4 exhibits the solution of the multi-rate PDE. The solution of the underlying ordinary DAE is obtained along a characteristic curve through the origin (not depicted in figure). One can observe sharp transients at the switching times of the mixer. Hence, an expansion of the waveforms by a trigonometric series (as in Harmonic Balance) leads to both the Gibb's phenomenon

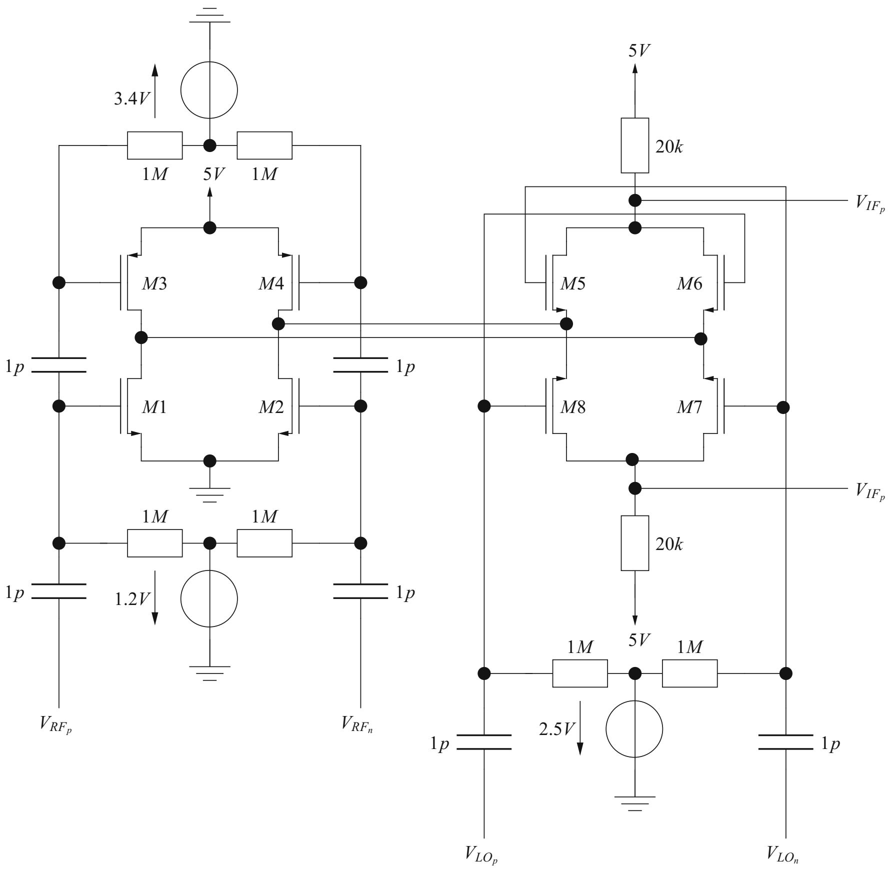


Fig. 3 Gilbert folded mixer circuit

and a large number of Fourier coefficients, making this approach inefficient. Instead, a spline-wavelet expansion based on compact basis functions are superior in capturing sharp transients.

3.2 *Coupled Circuit-Device Simulation*

Figure 5a depicts a power stage circuit with an on-chip balun for a band I application at a center frequency $f_c = 1$–9 GHz. The power stage operates in differential mode, that is all signals occur with ± signs. Since the source, e.g. the signal coming from the antenna, and output signals are single ended, a first balun, operating at a low power input signal, together with a matching circuit is required. The critical device in the design is the balun at the output of the power stage since it is driven by

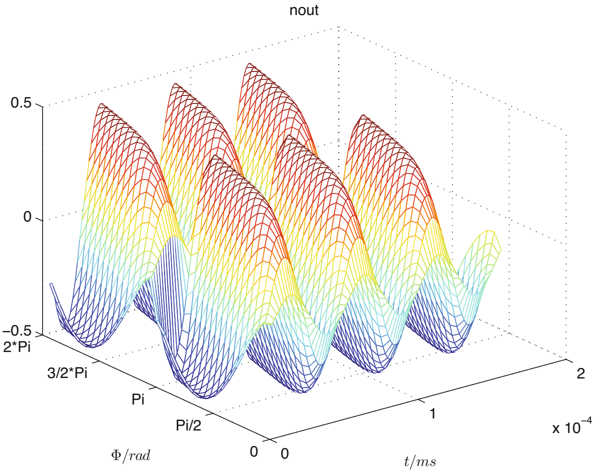
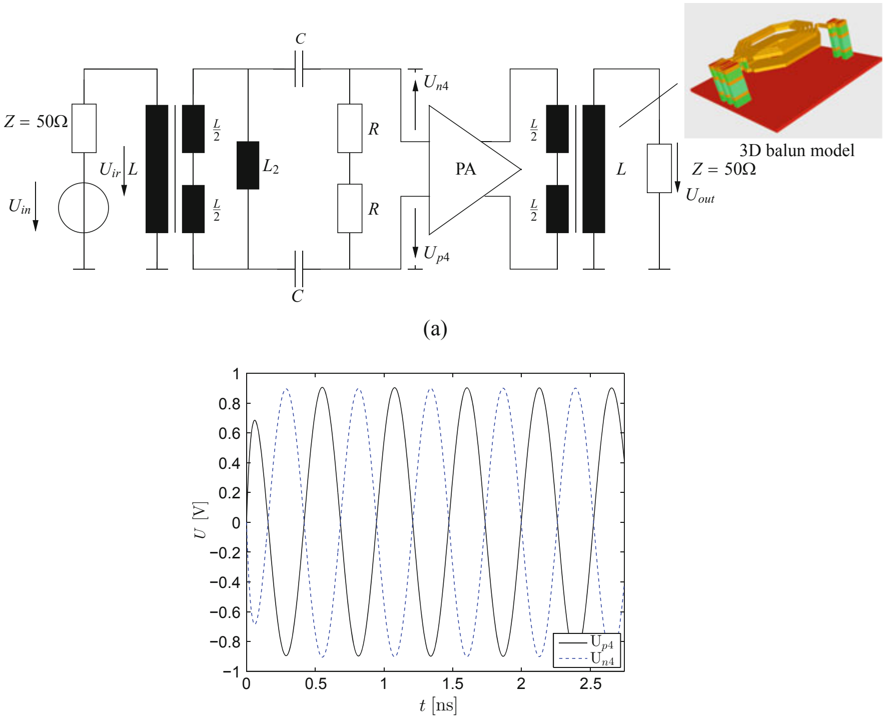


Fig. 4 PDE solution of the mixer circuit depicted in Fig. 3



(a)

(b)

Fig. 5 Balun driver circuit and differential output signal. (**a**) Circuit schematic. (**b**) PA differential output signal

a large input power. It is therefore modeled as full 3D device and simulated by the devEM TCAD solver. The remaining circuit's devices are simulated as lumped models. The balun is fabricated in bismaleimide-triazine (BT) technology with four layers. Figure 5b depicts the differential voltages waveforms at the input of the balun.

4 Conclusions

LinzFrame is a modular circuit simulator with emphasis on Radio Frequency circuits and devices. It has been holistically coupled both to the EM simulator devEM from MAGWEL NV and in an ongoing research project to the device simulator from RWTH Aachen for the development of novel devices for THz applications, enabling circuit-device mixed-level analysis.

Acknowledgement This work was partly funded by the Austrian Science Fund FWF under grant I3130-N30.

References

1. Bittner, K., Brachtendorf, H.-G.: Adaptive multi-rate wavelet method for circuit simulation. Radioengineering **23**(1), 300–307 (2014)
2. Bittner, K., Brachtendorf, H.-G.: Optimal frequency sweep method in multi-rate circuit simulation. COMPEL **33**(4), 1189–1197 (2014)
3. Bittner, K., Brachtendorf, H.-G.: Fast algorithms for grid adaptation using non-uniform biorthogonal spline wavelets. SIAM J. Sci. Comput. **37**(2), B283–B304 (2015)
4. Bittner, K., Brachtendorf, H.-G.: Latency exploitation in wavelet-based multirate circuit simulation. In: Bartel, A., Clemens, M., Günther, M., ter Maten, E.J.W. (eds.) Scientific Computing in Electrical Engineering 2014. Mathematics in Industry, pp. 13–20. Springer, New York (2016)
5. Bittner, K., Dautbegovic, E.: Adaptive wavelet-based method for simulation of electronic circuits. In: Michielsen, B., Poirier, J.-R. (eds.) Scientific Computing in Electrical Engineering 2010. Mathematics in Industry, pp. 321–328. Springer, Berlin/Heidelberg (2012)
6. Bittner, K., Brachtendorf, H.G., Schoenmaker, W., Reynier, P.: Coupled circuit/EM simulation for radio frequency. In: 54th ACM/EDAC/IEEE Design Automation Conference, Austin, TX, June 18–22, 2017, pp. 1–6 (2017)
7. Brachtendorf, H.G.: Simulation des eingeschwungenen Verhaltens elektronischer Schaltungen. Shaker, Aachen (1994)
8. Brachtendorf, H.G.: Theorie und Analyse von autonomen und quasiperiodisch angeregten elektrischen Netzwerken. Eine algorithmisch orientierte Betrachtung. PhD thesis (2001). Habilitationsschrift
9. Brachtendorf, H.G., Schoenmaker, W., Strohm, C., Bittner, K., Tischendorf, C.: Coupled circuit device simulation. In: Langer, U., Amrhein, W., Zulehner, W. (eds.) Scientific Computing in Electrical Engineering: SCEE 2016, St. Wolfgang, October 2016. Mathematics in Industry. Springer International Publishing, New York (2018)

10. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. SIAM, Philadelphia (1996)
11. Chen, Q., Schoenmaker, W., Meuris, P., Wong, N.: An effective formulation of coupled electromagnetic-tcad simulation for extremely high frequency onward. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **30**(6), 866–876 (2011)
12. Clemens, M., Weiland, T.: Discrete electromagnetism with the finite integration technique - abstract. J. Electromagn. Waves Appl. **15**(1), 79–80 (2001)
13. Jungemann, C., Bittner, K., Brachtendorf, H.G.: Simulation of plasma resonances in MOSFETs for THz-signal detection. In: 2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), January, pp. 48–51 (2016)
14. Jungemann, C., Bittner, K., Brachtendorf, H.-G.: Simulation of Plasma Resonances in MOSFETs for THz-Signal Detection. In: EUROSOI-ULIS, 2016 IEEE, pp. 48–51 (2016)
15. Mayaram, K., Pederson, D.O.: Coupling algorithms for mixed-level circuit and device simulation. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **11**(8), 1003–1012 (1992)
16. Melville, R., Moinian, S., Feldmann, P., Watson, L.: Sframe: an efficient system for detailed DC simulation of bipolar analog integrated circuits using continuation methods. Analog Integr. Circuits Signal Process. **3**(3), 163–180 (1993)
17. Meuris, P., Schoenmaker, W., Magnus, W.: Strategy for electromagnetic interconnect modeling. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **20**(6), 753–762 (2001)
18. Nanoelectronic COupled Problems Solutions, 2013–2016. FP7-ICT-2013.3.1
19. Pulch, R.: Variational methods for solving warped multirate partial differential algebraic equations. SIAM J. Sci. Comput. **31**(2), 1016–1034 (2008)
20. Roychowdhury, J.: Analyzing circuits with widely separated time scales using numerical PDE methods. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **48**(5), 578–594 (2001)
21. ter Maten, E.J.W., Brachtendorf, H.G., Pulch, R., Schoenmaker, W., De Gersem, H. (eds.) Nanoelectronic Coupled Problems Solutions. Series Mathematics in Industry. Springer (2019). https://books.google.at/books?id=P8-8DwAAQBAJ
22. Weiland, T.: A discretization method for the solution of Maxwell's equations for six-component fields. Archiv Elektronik und Uebertragungstechnik AEU **31**(3), 116–120 (1977)

Fast Transient Simulation of RC Circuits with Dense Capacitive Coupling

N. T. K. Dang, J. M. L. Maubach, J. Rommes, P. Bolcato, and W. H. A. Schilders

Abstract The main motivation of this work is lying in the acceleration of transient simulation of Analog Mixed Signal circuits. In the electronics industry, smaller and faster electronic devices are always demanded. Full device-parasitic transient simulations of realistic circuits are time consuming or even infeasible due to a huge number of electrical components and unavoidable parasitics. In this paper, we introduce a novel technique to address the problem with the presence of parasitic capacitances. The SelectC technique activates/inactivates coupling capacitances during the transient simulation, therefore, giving faster transient simulation time.

1 Introduction

Very Large Scale Integrated circuits contain numerous tiny electronic devices placed in a small flat piece of semiconductor material. In every new generation, smaller-size transistors and denser electronic devices significantly cause parasitic electromagnetic effects in transistors and interconnects to be more apparent. Simulation of such parasitic networks is too costly or unattainable. With the purpose of speedup and (or) enabling transient simulation of AMS (Analog Mixed Signal) circuits, reduced models are sought for the parasitics. Well-known MOR (Model Order Reduction) methods to deal with these large networks are Krylov subspaces [6], balanced truncation methods [8], and elimination methods [10]. However, existing MOR methods can produce dense reduced models that become more expensive to simulate than the original systems. Moreover, for multi-terminal networks MOR is inefficient because of generating large reduced models and/or dense models such as

N. T. K. Dang (✉) · J. M. L. Maubach · W. H. A. Schilders
Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: T.K.N.Dang@tue.nl; J.M.L.Maubach@tue.nl; W.H.A.Schilders@tue.nl

J. Rommes · P. Bolcato
Mentor Graphics, Grenoble, France
e-mail: Joost_Rommes@mentor.com; Pascal_Bolcato@mentor.com

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_18

when using PRIMA [6]. Recently, ReduceR [9], SparseRC [3] and TurboMOR-RC [7] are presented and share the common goals which are creating sparse reduced models and working efficiently with multi-terminal networks. The proposed method, unlike the aforementioned MOR methods, does not reduce the dimension of the system matrices, therefore, it does not face the problems related to multi-terminal networks and creating dense resulting matrices as MOR methods. We refer to it as a simplification method. We do not provide a comparison between SelectC and other MOR methods because methods such as SVDMOR [1] and ESVDMOR [4] consider and modify the transfer function, which SelectC does not do (instead, SelectC concentrates on the time-domain transient simulation). Also, ReduceR deals with R networks while SelectC modifies C networks, sparseRC [3] and TurboMOR [7] reduce the sub-nets of RC networks. A comparison would require SelectC to be combined with sub-net reduction, which is not the scope of the current research.

In this paper, we propose an efficient method for the case of RC networks with dense capacitive coupling. The basic idea is to remove small coupling capacitors between distinct resistor components having negligible current passing through. In particular, coupling capacitors between nets are selected at the current time step in order to decide whether or not to be inactive in the next time step of the transient simulation. For many time steps, sparser capacitance matrices are obtained and replace the original capacitance matrix in the system matrix, thus providing faster transient simulation. For some initial experiments to be described later, the method gives promising results.

2 Problem Formulation

The time-domain transient behaviour of a given circuit is described by the following formulation:

$$\mathbf{f}(\mathbf{x}) = \mathbf{C}\mathbf{x}'(t) + \mathbf{G}\mathbf{x}(t) + \mathbf{H}\mathbf{g}(\mathbf{x}(t)) - \mathbf{B}\mathbf{u}(t), \quad t \in (0, T] \tag{1}$$

where MNA [5] matrices $\mathbf{G}$, $\mathbf{C} \in \mathbb{R}^{N\times N}$ are here considered to be positive semi-definite, corresponding to the conductivities, capacitances, respectively, $T > 0$, $\mathbf{x} \in \mathbb{R}^N$ denotes the node voltages, $\mathbf{u} \in \mathbb{R}^M$ represents the current or voltage sources, $\mathbf{g} \in \mathbb{R}^K$ is a vector of nonlinear functions of diodes, nonlinear resistors, etc., $\mathbf{B} \in \mathbb{R}^{N\times M}$ and $\mathbf{H} \in \mathbb{R}^{N\times K}$ are the incidence matrices related to sources and nonlinear elements, respectively. N, M and K are the dimension of (1), the number of sources and the number of nonlinear elements, respectively.

The nonlinear system of equations (1) is usually solved for a given initial condition $\mathbf{x}(t_0) := \mathbf{x}_0$ where $\mathbf{x}_0$ is obtained by solving for the DC solution:

$$\mathbf{f}_0(\mathbf{x}) = \mathbf{G}\mathbf{x} + \mathbf{H}\mathbf{g}(\mathbf{x}) - \mathbf{B}\mathbf{u}_0. \tag{2}$$

Equations (2) and (1) are solved numerically by using a nonlinear iterative method such as Newton's method. At each Newton iteration, a linearized problem is solved to update the approximation:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{J}_{\mathbf{f}}^{-1}\mathbf{f}(\mathbf{x}^{(k)}), \tag{3}$$

where $\mathbf{J}_{\mathbf{f}} \in \mathbb{R}^{N\times N}$, $\mathbf{J}_{\mathbf{f}} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\big|_{\mathbf{x}=\mathbf{x}^{(k)}}$ is the Jacobian matrix of $\mathbf{f}(\mathbf{x})$.

For time domain simulation, from the initial condition the DAE (Differential Algebraic Equation) system in (1) is discretized into a set of nonlinear algebraic equations by a numerical integration method (Backward Euler, Trapezoidal). Then iterative Newton's algorithm is applied at each time point t_i and finally linear solvers are used to solve the linearized equations. Most of the computation time is spent in assembling the derivative matrix $\mathbf{J}_{\mathbf{f}}(\mathbf{x}^{(k)})$ and in solving systems with it. Therefore, increasing the sparsity of $\mathbf{C}$ and/or $\mathbf{G}$ would potentially help to accelerate the transient simulation.

3 The SelectC Method

First of all, system (1) should be made consistent by removing the equation corresponding to a reference node voltage $x_{ref} = 0$. In what follows, we use the connected components of $\mathbf{G}$, see [9]. Because SelectC deals with coupling capacitors (the capacitors between distinct $\mathbf{G}$-components), it is vital to identify those capacitors. This is done by reordering $\mathbf{C}$ by $\mathbf{G}$-components as described in the first procedure of Algorithm 1. To explain the idea of SelectC, let c_{ij} be a capacitor between nodes i and j. The current i_c passing through c_{ij} (direction $i \rightarrow j$) is

$$i_c = c_{ij}\cdot\frac{d(x_i - x_j)}{dt} \approx c_{ij}\cdot\left(\frac{x_i(t_{n+1}) - x_i(t_n)}{t_{n+1} - t_n} - \frac{x_j(t_{n+1}) - x_j(t_n)}{t_{n+1} - t_n}\right) = \tilde{i}_c. \tag{4}$$

When i_c is negligible i.e. $i_c \leq \tau_{sel}$, then c_{ij} is said to be inactive (here we assume that cumulative effects are neglected) and can be removed from the capacitance matrix. In other words, SelectC is the method of selecting active/inactive coupling capacitors based on the currents going through them. The inactive capacitors are coupling capacitors whose node voltages barely change from the previous time step to the current one. To ensure the voltage derivative part is minor and the i_c value is not affected by the c_{ij} value, the possibly inactive capacitors c_{ij} to be removed should have capacitance smaller than a chosen constant τ_{cap}. Eventually, a capacitance is removed if both

$$\begin{cases} c_{ij} \leq \tau_{cap} \\ \tilde{i}_c \leq \tau_{sel}. \end{cases} \tag{5}$$

τ_{sel} and τ_{cap} are user-defined thresholds. Coupling capacitances do not exceed τ_{cap} which satisfy related $\tilde{i}_c \leq \tau_{sel}$ are omitted. Algorithm 1 summarizes the entire flow of SelectC method.

Algorithm 1 SelectC

```
procedure USE THE G-COMPONENTS TO DETERMINE THE COUPLING CAPACITORS
    $[comps, -] = \texttt{components}(\mathbf{G})$
    $[-, \mathbf{P}] = \texttt{sort}(comps)$
    $\mathbf{G} = \mathbf{G}(\mathbf{P}, \mathbf{P})$
    $\mathbf{C} = \mathbf{C}(\mathbf{P}, \mathbf{P})$
procedure DYNAMIC SELECTION OF COUPLING CAPACITORS
    $n = 0, \quad \mathbf{C}_{sel}^{(n)} := \mathbf{C}, \quad \mathbf{x}(t_n)$ is DC solution
    while $t_n < t_{end}$ do
        $n = n + 1$
        $k = 0; \quad \mathbf{x}_{n-1}^{(k)} = \mathbf{x}(t_{n-1}) = \mathbf{x}_{n-1}$
        while $\|\mathbf{f}(\mathbf{x}_{n-1}^{(k)})\|_\infty > \varepsilon$ do                ▷ Newton iteration
            $\mathbf{x}_{n-1}^{(k+1)} = \mathbf{x}_{n-1}^{(k)} - \mathbf{J}_\mathbf{f}^{-1}\mathbf{f}(\mathbf{x}_{n-1}^{(k)})$
            $k = k + 1$
        $\mathbf{x}(t_n) = \mathbf{x}_{n-1}^{(m)}$                ▷ $\|\mathbf{f}(\mathbf{x}_{n-1}^{(m)})\|_\infty < \varepsilon$
        if $\|\mathbf{dx}(t_{n-1})\|_\infty := \|\mathbf{x}(t_n) - \mathbf{x}(t_{n-1})\|_\infty \geq \tau_{dx}$ & $\mathbf{C}_{sel}^{(n-1)} \neq \mathbf{C}$ then
            $\mathbf{C}_{sel}^{(n-1)} := \mathbf{C}$
            $n = n - 1$
        else
            build $\mathbf{C}_{sel}^{(n)}$ satisfying (5)
```

The condition of $\|\mathbf{dx}(t_{n-1})\|_\infty \geq \tau_{dx}$ and $\mathbf{C}_{sel}^{(n-1)} \neq \mathbf{C}$ is used to back up the solution $\mathbf{x}(t_n)$ incase there is a large difference between the two consecutive vector of solutions $\mathbf{x}(t_n) - \mathbf{x}(t_{n-1}) = \mathbf{dx}(t_{n-1})$. τ_{dx} is an absolute value as time scaling the linear version of (1) by $s := f \cdot t$ (frequency f) reduces (1) to $\mathbf{f}(\mathbf{x}(s)) = f \cdot \mathbf{C}\mathbf{x}'(s) + \mathbf{G}\mathbf{x}(s) + \mathbf{B}\mathbf{u}(s), s \in (0, f \cdot T]$ where instead of $\mathbf{x}'(t) = O(f)$ time the input signal, $\mathbf{x}'(s) = O(1)$ times inputs signal. This implies that with or without time scaling $\mathbf{x}(t_{n-1}) - \mathbf{x}(t_n) = dt \cdot O(f)$ and $\mathbf{x}(s_{n-1}) - \mathbf{x}(s_n) = dt \cdot 1$ where $dt \cdot O(f) = ds$.

4 Numerical Results

In this section, we present results of SelectC for large linear RC networks derived from realistic designs of very-large-scale integration (VLSI) chips (netlists 1 and 2 in Table 1) and a self-created nonlinear network (netlist 3, Fig. 5). For the linear networks the square inputs $2 \cdot square(2\pi \cdot f \cdot t)$ are injected to one external per net and for nonlinear network, the input $square(2\pi \cdot f \cdot t)$ is driven to the first external of the first net. SelectC is implemented in MATLAB ver. 9.4(R2018a). The method requires $\mathbf{G}, \mathbf{C}$ to be reordered by $\mathbf{G}$-components, as mentioned in Sect. 3.

Table 1 Numerical results of SelectC

Netlist	$\tau_{cap}(F)$	$\tau_{sel}(A)$	Type	Avg#cap	Sim. time (s)	Rel. error (V)	Time lin. solver (s)	Time newt. iter. (s)
1. $N = 18,927$	–	–	Orig.	168,309	619	–	588	31
	1e−14	1e−8	SelectC	15,090	105	6e−4	77	11
			Red. rate	91%	5.9×		7.6×	2.8×
		1e−4	SelectC	17,756	88.5	3.1e−2	66	10
			Red. rate	89.5%	7×		8.9×	3.1×
2. $N = 6887$	–	–	Orig.	14,161	53	–	48	6
	1e−14	1e−8	SelectC	2102	23	3.1e−4	15	5
			Red. rate	85%	2.3×		3.2×	1.2×
		1e−4	SelectC	611	11.6	6e−2	7	4
			Red. rate	94.7%	4.6×		6.9×	1.5×
3. $N = 32$	–	–	Orig.	23	7.7	–	0.4	4
	1e−14	1e−8	SelectC	8	7.1	3.3e−4	0.2	3.3
			Red. rate	65.2%	1.1×		2×	1.2×
		1e−4	SelectC	3	6.7	4.7e−4	0.2	3.2
			Red. rate	86.9%	1.15×		2×	1.25×

Netlists 1 and 2 are linear, netlist 3 is nonlinear

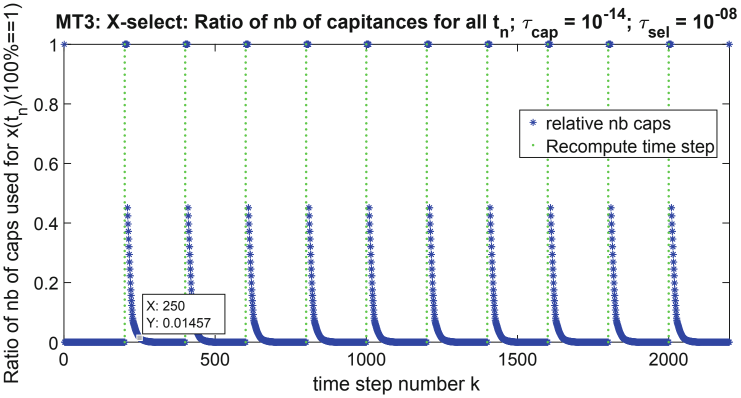


Fig. 1 The ratio of the number of active capacitors during transient simulation (star-dotted line). The time steps where the recalculation with all capacitors is required (dotted line). The time steps correlate with the simulation interval in Fig. 3 (*the top figure*)

We employed the MATLAB function `components` which is in the Boost Graph Library[2]. The numerical integration is Backward Euler with constant increment.

Figure 1 shows the ratio of the number of active capacitors during the transient simulation (of netlist 1) when $\mathbf{C}_{sel}$ is used. While full transient simulation requires full $\mathbf{C}$ elements, SelectC uses full $\mathbf{C}$ elements only 10 times (to recompute the time steps) when the condition of $\|\mathbf{dx}(t_n)\|_\infty \geq \tau_{dx}$ and $\mathbf{C}_{sel}^{(n)} \neq \mathbf{C}$ meets (because the signal suddenly rises from 0 to 2 and falls from 2 to 0, see Fig. 3 *the top figure* for

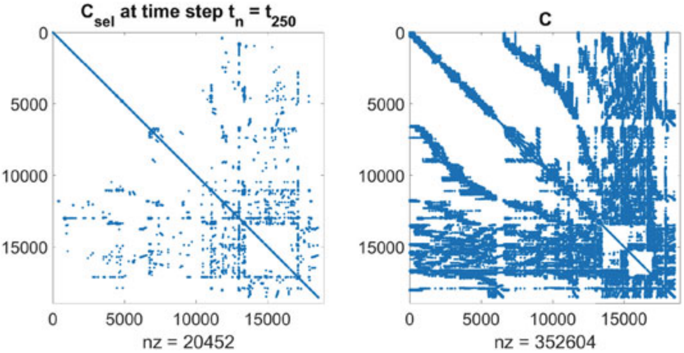


Fig. 2 $\mathbf{C}_{sel}$ (left) at time step $t_n = t_{250}$ is much sparser than the original (re-ordered) $\mathbf{C}$ (right)

the input signal-plotted in blue line). For other time steps, for instance at time step $t_n = t_{250}$, $\mathbf{C}_{sel}$ is sparser than $\mathbf{C}$ (Fig. 2).

Table 1 shows some SelectC's results for two multi-terminal netlists extracted from real chip designs and a self-created nonlinear netlist compared to the results obtained by their original problem. Herein, N denotes the number of nodes, τ_{cap} and τ_{sel} are thresholds for selecting inactive capacitors in (5), $\tau_{dx} = 1$ (linear case) and $\tau_{dx} = 0.1$ (non-linear case) indicate the condition of re-computing the time step with dramatic change in $\mathbf{dx}$, and **Avg#cap** stands for the mean number of the active capacitors during the transient simulation. Indeed, for the original problem Avg#cap is the number of capacitors in $\mathbf{C}$ network and is computed by the MATLAB command `nnz(triu(C,1))`, i.e. the number of non-zero elements of the strictly upper triangular of $\mathbf{C}$. For the SelectC system, Avg#cap stands for the division of the sum of the amount of the active capacitors per time step over the total number of the time steps. The reduction rates (**Red. rate**) are shown for the corresponding columns. For instance, the percentage reduction in Avg#cap is $\frac{(Avg\#cap_{orig} - Avg\#cap_{sel}) \cdot 100}{Avg\#cap_{orig}}$, the Red. rate in simulation time (**Sim. Time**) is $\frac{Sim.Time_{orig}}{Sim.Time_{sel}}$. Finally, **Rel. Error** displays the maximum relative error (in voltage) of all variables and is computed by $\max\limits_{i=1,2,\ldots,N} \left\{ \|x_i - x_i^{Csel}\|_\infty \right\} / \max\limits_{i=1,2,\ldots,N} \{\|x_i\|_\infty\}$ with $\|x_i\|_\infty = \max\limits_{k=1,2,\ldots,n_T} \{|x_i(t_k)|\}$, n_T is the number of time steps. **Time Newt. Iter.** and **Time Lin. Solver** are time needed for Newton iteration (without linear solving) and for solving the linear system inside, respectively.

The transient simulation time strongly depends on the sparsity of the resulting matrices (Table 1). Obviously, SelectC can reduce about 90% of the average density of the original $\mathbf{C}$, leading to faster transient simulation time (at maximum by a factor of 7 in netlist 1). Additionally, the error is acceptable for $\tau_{sel} = 10^{-8}$ (for instance in netlist 1 Fig. 3). For $\tau_{sel} = 10^{-4}$, especially in netlist 2 we gain more speed up, however, the accuracy is not acceptable (Fig. 4). Note that the Time Lin. Solver of the SelectC problem is faster than that of the original problem (up to

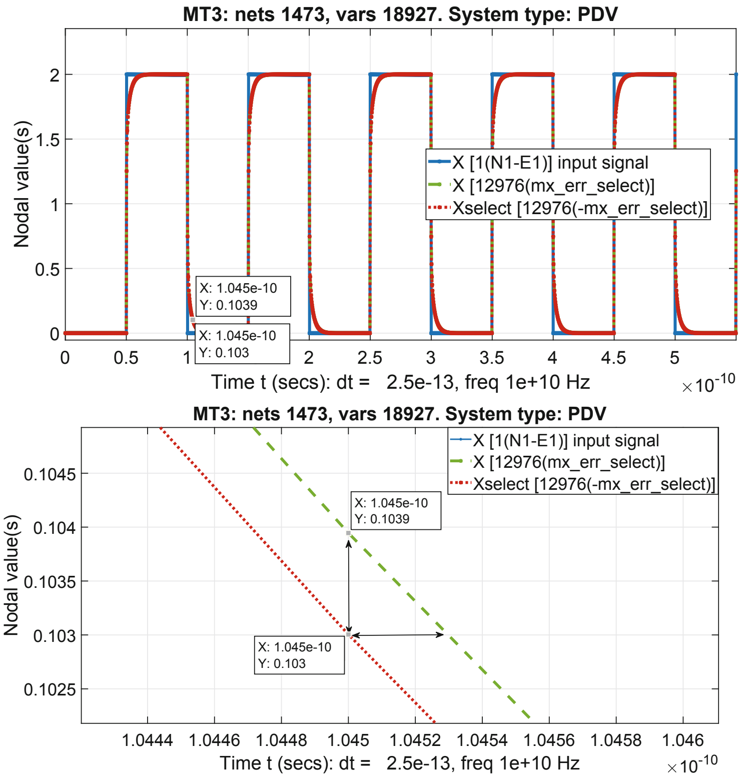


Fig. 3 *The top figure*: Transient simulation of the original system versus SelectC system (netlist 1), the output 12,976 gives maximum absolute error, inputs are square signals injected to one external/net, $\tau_{sel} = 10^{-8}$. *The bottom figure* shows the zoom in of the two marked points (from the top figure) shown the difference (about 10^{-3}) between the original problem and the modified one. The delay error (horizontal line) is about 0.03 ps and is also acceptable

8.9×). However, considering the transient simulation time, the speedup is only up to 7× since there needs the computation effort at Csel construction at every time step (Fig. 5).

5 Conclusions and Outlook

The SelectC technique provides faster transient simulations of RC networks up to the factor of 7. The method works nicely with stable signals, for instance trapezoidal, pulse and/or square signals for problems with many C-parasitic. To be investigated are the reliability and automatic detection/generation of the method (more precisely, given an error tolerance, which value of τ_{sel} and τ_{cap} we should

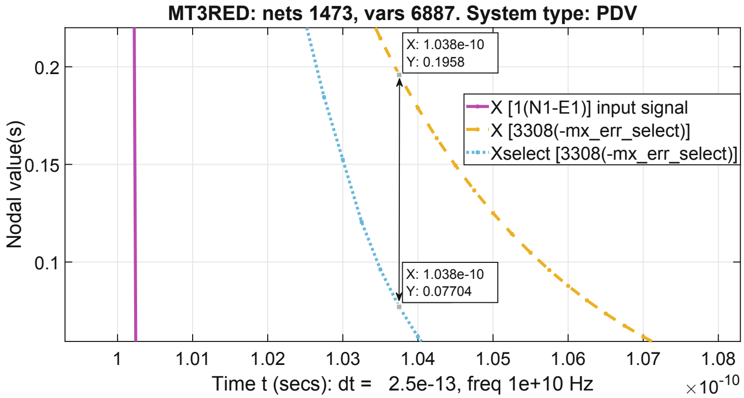


Fig. 4 The zoom in of transient simulation of the original system versus SelectC system (netlist 2), the output 3308 gives maximum absolute error, inputs are square signals injected to one external/net, $\tau_{sel} = 10^{-4}$. The two marked points shown the large difference (about 10^{-1}) between the original problem and the modified one

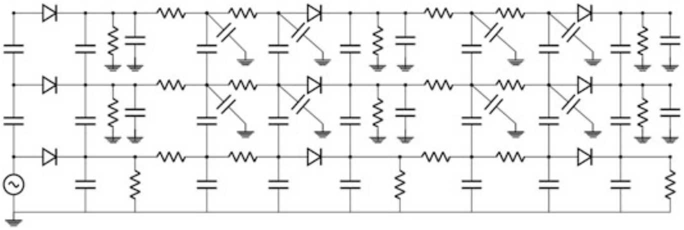


Fig. 5 An academical nonlinear network. Coupling capacitor $c_c = 10^{-15}$, capacitor connected to ground $c_g = 10^{-16}$, $r = 10^{-4}$, $r_g = 10^4$ and diode with $I_{sat} = 1.23e - 9$, $\eta = 1.73$ and $V_T = 26e - 3$

choose, and vice versa), and also the application of SelectC for general circuits including nonlinear elements.

Acknowledgments This work was supported by the European Union in the FP7-PEOPLE-2013-ITN Program (FP7 Marie Curie Action, Project ASIVA14—Analog SImulation and Variability Analysis for 14 nm designs).

References

1. Feldmann, P.: Model order reduction techniques for linear systems with large numbers of terminals. In: Proceedings Design, Automation and Test in Europe Conference and Exhibition, Paris, pp. 944–947. IEEE Computer Society, Washington (2004)
2. Gleich, D.: MatlabBGL: A Matlab Graph Library. https://www.cs.purdue.edu/homes/dgleich/packages/matlabbgl/

3. Ionutiu, R., Rommes, J., Schilders, W.H.A.: SparseRC: sparsity preserving model reduction for RC circuits with many terminals. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **30**(12), 1828–1841 (2011)
4. Liu, P., Tan, S.X.-D., Yan, B., McGaughy, B.: An efficient terminal and model order reduction algorithm. Integration **41**(2), 210–218 (2008)
5. Najm, F.N.: Circuit Simulation. Wiley, Hoboken (2010). OCLC: ocn403853209
6. Odabasioglu, A., Celik, M., Pileggi, L.: PRIMA: passive reduced-order interconnect macromodeling algorithm. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **17**(8), 645–654 (1998)
7. Oyaro, D., Triverio, P.: TurboMOR-RC: an efficient model order reduction technique for RC networks with many ports. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **35**(10), 1695–1706 (2016)
8. Reis, T., Stykel, T.: PABTEC: passivity-preserving balanced truncation for electrical circuits. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **29**(9), 1354–1367 (2010)
9. Rommes, J., Schilders, W.H.A.: Efficient methods for large resistor networks. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **29**(1), 28–39 (2010)
10. Sheehan, B.N.: Realizable reduction of RC networks. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **26**(8), 1393–1407 (2007)

*P*redictor/*C*orrector *N*ewton-*R*aphson (PCNR): A Simple, Flexible, Scalable, Modular, and Consistent Replacement for Limiting in Circuit Simulation

Karthik V. Aadithya, Eric R. Keiter, and Ting Mei

Abstract Modern circuit simulators predominantly use Newton-Raphson (NR) iteration to solve circuit equations. To improve NR convergence, circuit simulators use a practice called "limiting". This ensures that sensitive circuit quantities (such as diode voltages) do not change too much between successive NR iterations. However, in most simulators, the implementation of limiting tends to be inflexible, non-modular, inconsistent, and confusing. We therefore propose *P*redictor/*C*orrector *N*ewton-*R*aphson (PCNR), a replacement for limiting that overcomes these disadvantages while incurring modest computational overhead. The key ideas behind PCNR are, (1) to add each limited circuit quantity as an extra unknown to the circuit's Modified Nodal Analysis (MNA) system of equations, (2) to split each NR iteration into a "prediction" phase followed by a "correction" phase, and (3) to mitigate the computational cost of the extra unknowns by eliminating them from all $Ax = b$ solves using a Schur complement based technique.

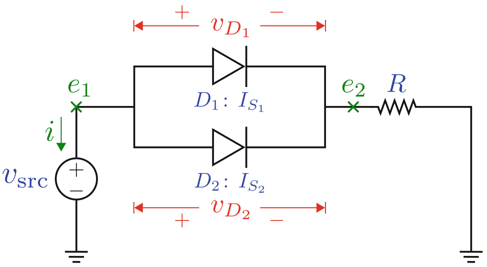
1 An Illustrative Example

Consider the circuit in Fig. 1; it contains two parallel diodes D_1 and D_2 (with saturation currents I_{S_1} and I_{S_2} respectively), in series with a resistor R, driven by a DC voltage source v_{src}. In Sect. 2, we use this circuit to highlight the problems with existing limiting implementations. Then, in Sect. 3, we use the same circuit to introduce the key ideas behind PCNR, which we then generalize into a powerful replacement for limiting that works for all circuits (as we illustrate using the family of circuits shown in Fig. 3).

K. V. Aadithya (✉) · E. R. Keiter · T. Mei
Sandia National Laboratories, Albuquerque, NM, USA
e-mail: kvaadit@sandia.gov; erkeite@sandia.gov; tmei@sandia.gov

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_19

Fig. 1 The circuit we use to illustrate key ideas in this write-up



2 Problems with Traditional Limiting

The MNA equations [1, 2] for the circuit in Fig. 1 are:

$$\mathbf{g}\left(\underbrace{\begin{bmatrix} e_1 \\ e_2 \\ i \end{bmatrix}}_{\mathbf{x}}\right) = \begin{bmatrix} i + \underbrace{I_{S_1}\left(\mathrm{e}^{\frac{e_1-e_2}{V_T}} - 1\right)}_{D_1} + \underbrace{I_{S_2}\left(\mathrm{e}^{\frac{e_1-e_2}{V_T}} - 1\right)}_{D_2} \\ -\underbrace{I_{S_1}\left(\mathrm{e}^{\frac{e_1-e_2}{V_T}} - 1\right)}_{D_1} - \underbrace{I_{S_2}\left(\mathrm{e}^{\frac{e_1-e_2}{V_T}} - 1\right)}_{D_2} + \frac{e_2}{R} \\ e_1 - v_{\mathrm{src}} \end{bmatrix} = \mathbf{0}, \tag{1}$$

where V_T is the thermal voltage ($\approx$ 26 mV at room temperature).

The equations take the form $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, where $\mathbf{x}$ is the vector of unknowns the simulator has to solve for.[1] The solution is usually obtained by NR iteration [1, 3, 4]; the simulator starts with an initial guess $\mathbf{x_0}$, and repeatedly refines it into successively better guesses ($\mathbf{x_1}, \mathbf{x_2}$, etc.), until a close enough approximate solution is found. Each iteration requires computing $\mathbf{g}$, as well as its Jacobian $d\mathbf{g}/d\mathbf{x}$, at the current guess $\mathbf{x_i}$, which are then used to update/improve the current guess.[2]

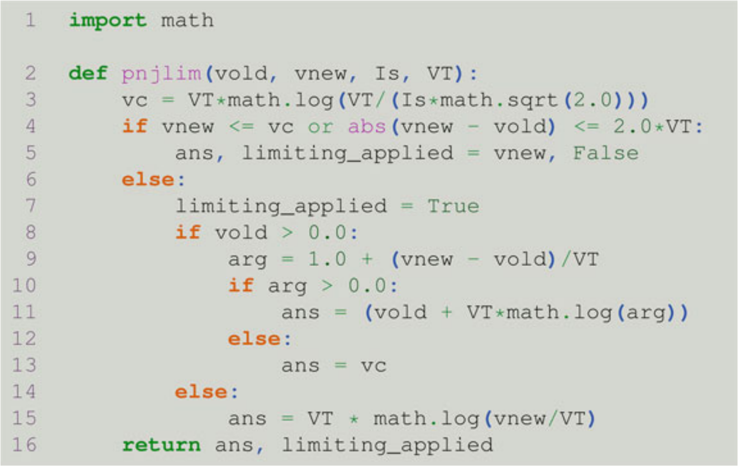
Thus, at each NR iteration, the simulator calls each device in the circuit, requesting it to calculate its branch currents and charges at the current guess $\mathbf{x_i}$, and then assembles these into vectors/matrices according to the circuit's topology [1, 2, 5]. For example, (1) shows the values calculated by the diodes D_1 and D_2, based on the exponential relationship between the current flowing through a P-N junction diode and the voltage across it [6]. The exponentials in such calculations often adversely affect NR convergence. So, it is necessary to prevent the diode

[1] PCNR works for differential-algebraic equations as well, but for simplicity, we only consider algebraic equations in this write-up.

[2] Equivalently, one can also compute the Jacobian $d\mathbf{g}/d\mathbf{x}$, and an "RHS" function given by $\mathrm{RHS}(\mathbf{x}) = \left(\frac{d\mathbf{g}}{d\mathbf{x}}\right).\mathbf{x} - \mathbf{g}(\mathbf{x})$, at each iteration $\mathbf{x_i}$, which is the approach traditionally followed by SPICE simulators.

voltages from changing too much between successive iterations. This idea is called limiting [1, 5, 7].

To implement limiting, each diode keeps track of not only the current guess $\mathbf{x_i}$, but also an *internal junction voltage* $\hat{v}_{i-1}$ that depends on all the previous guesses $\mathbf{x_0}$ through $\mathbf{x_{i-1}}$. When called by the simulator at the i^{th} iteration $\mathbf{x_i}$, instead of doing calculations at the junction voltage v_i implied by the node voltages in $\mathbf{x_i}$, each diode calculates, using a method like `pnjlim` shown in Listing 1 [1, 8], a *new* internal junction voltage $\hat{v}_i$ that lies between the previous internal junction voltage $\hat{v}_{i-1}$ and the junction voltage v_i requested by the simulator. Then, all branch current/charge calculations are done at the voltage $\hat{v}_i$, linearly extrapolated to v_i, and returned to the simulator pretending as though the calculations had indeed been done at v_i [1, 5, 7]. The simulator proceeds being none the wiser.[3]



```python
import math

def pnjlim(vold, vnew, Is, VT):
    vc = VT*math.log(VT/(Is*math.sqrt(2.0)))
    if vnew <= vc or abs(vnew - vold) <= 2.0*VT:
        ans, limiting_applied = vnew, False
    else:
        limiting_applied = True
        if vold > 0.0:
            arg = 1.0 + (vnew - vold)/VT
            if arg > 0.0:
                ans = (vold + VT*math.log(arg))
            else:
                ans = vc
        else:
            ans = VT * math.log(vnew/VT)
    return ans, limiting_applied
```

Listing 1: Python implementation of `pnjlim`, where `vold`, `vnew`, and `ans` represent $\hat{v}_{i-1}$, v_i, and $\hat{v}_i$ respectively.

There are several problems with this approach. First, $\mathbf{g}$ and its Jacobian are no longer functions of just $\mathbf{x}$; they also depend on the history of $\mathbf{x}$. This adds confusion and breaks the clean mathematical abstractions underlying circuit simulation theory [5]. For example, evaluating $\mathbf{g}$ twice on the same $\mathbf{x}$ can result in two completely different answers. Second, this approach is fundamentally inconsistent. For example, under this scheme, the diodes D_1 and D_2 above could assume very different internal voltages even though they share the *same* branch voltage $e_1 - e_2$, making both $\mathbf{g}$ and its Jacobian inconsistent, and hence difficult to analyze mathematically [1, 5]. Third, by requiring each device to know about previous

[3] In practice, each diode also returns a Boolean flag to the simulator telling it whether limiting was or was not applied, which the simulator uses to determine whether NR has truly converged or not.

iterations and other analysis-specific context information, this approach increases code complexity and reduces modularity; we believe it is the simulator's (and not the devices') responsibility to keep track of the analysis context, NR iterations, etc.

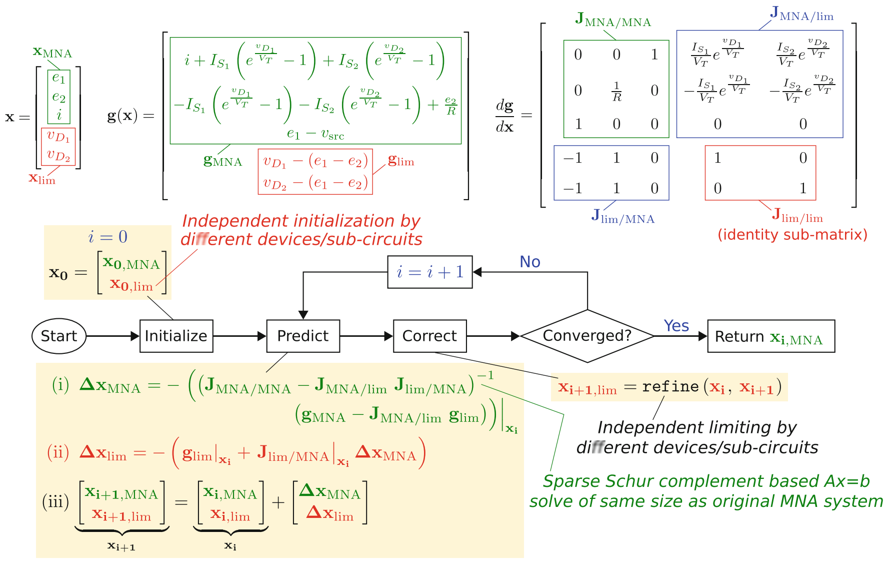
3 PCNR: Our Replacement for Limiting

In Sect. 2, we saw that *different* devices in a circuit may try to limit the *same* branch voltage, leading to inconsistencies in the evaluation of $\mathbf{g}$ and its Jacobian. For example, the diodes D_1 and D_2 from Fig. 1 both try to limit the same branch voltage $e_1 - e_2$. The root cause of this problem is that traditional MNA only treats node voltages (like e_1 and e_2), and *not* branch voltages (like v_{D_1} and v_{D_2} from Fig. 1) as unknowns [1, 2, 5]. This creates "clashes" between limiting devices that share node connections. Instead, if we treat limited quantities as unknowns in their own right, we can eliminate these clashes (and the inconsistencies they induce) by making sure that each device "owns" all the solution variables that it limits. This is our first key insight: in PCNR, we treat each limited quantity as a circuit unknown. This increases system size and hence the computational cost of solving the system, but it removes inconsistencies, increases code modularity, and provides the flexibility needed to implement new and innovative limiting schemes for modern devices and compact models—so we believe that the tradeoff is worth it.

Figure 2 shows the PCNR flow. At the top, we see that the PCNR vector $\mathbf{x}$ of unknowns contains all the original MNA unknowns ($\mathbf{x}_{\text{MNA}}$), as well as all the limited quantities ($\mathbf{x}_{\text{lim}}$). For example, when PCNR is applied to the circuit of Fig. 1, the limited branch voltages v_{D_1} and v_{D_2} become additional unknowns, owned by the diodes D_1 and D_2 respectively.

Thus, the diode D_1 (D_2), when called by the simulator, can simply do all its calculations at the current iteration's v_{D_1} (v_{D_2}), instead of having to keep track of an internal junction voltage $\hat{v}_{D_1}$ ($\hat{v}_{D_2}$) and its evolution between the previous iteration and this one. This is shown in $\mathbf{g}_{\text{MNA}}$, the top part of $\mathbf{g}$, as depicted in Fig. 2. The bottom part of $\mathbf{g}$ (denoted $\mathbf{g}_{\text{lim}}$) is obtained by simply writing equations expressing the limited quantities $\mathbf{x}_{\text{lim}}$ in terms of the original MNA unknowns $\mathbf{x}_{\text{MNA}}$; in our circuit, this means equating both the limited voltages v_{D_1} and v_{D_2} to the branch voltage $e_1 - e_2$ (Fig. 2). In similar fashion, as the top right of Fig. 2 shows, we split the Jacobian $d\mathbf{g}/d\mathbf{x}$ into four block sub-matrices, of which the bottom right sub-matrix ($\mathbf{J}_{\text{lim/lim}}$) is guaranteed to be identity. Thus, in PCNR, both $\mathbf{g}$ and its Jacobian are functions of just $\mathbf{x}$ (and not the history of $\mathbf{x}$), which ensures simplicity and consistency.

The bottom half of Fig. 2 shows a flowchart for solving circuits using PCNR. This brings us to our second key insight: in PCNR, each NR iteration is split into two phases—a "prediction" phase followed by a "correction" phase.



$\mathbf{x}_{\text{MNA}}$
$\mathbf{x} = [e_1, e_2, i, v_{D_1}, v_{D_2}]^T$
$\mathbf{x}_{\text{lim}}$
$\mathbf{g}(\mathbf{x}) =$
$i + I_{S_1}\left(e^{\frac{v_{D_1}}{V_T}} - 1\right) + I_{S_2}\left(e^{\frac{v_{D_2}}{V_T}} - 1\right)$
$-I_{S_1}\left(e^{\frac{v_{D_1}}{V_T}} - 1\right) - I_{S_2}\left(e^{\frac{v_{D_2}}{V_T}} - 1\right) + \frac{e_2}{R}$
$e_1 - v_{\text{src}}$
$\mathbf{g}_{\text{MNA}}$
$v_{D_1} - (e_1 - e_2)$
$v_{D_2} - (e_1 - e_2)$
$\mathbf{g}_{\text{lim}}$
$\frac{d\mathbf{g}}{d\mathbf{x}} =$
$\mathbf{J}_{\text{MNA/MNA}}$
$\mathbf{J}_{\text{MNA/lim}}$
0 0 1
0 $\frac{1}{R}$ 0
1 0 0
$\frac{I_{S_1}}{V_T}e^{\frac{v_{D_1}}{V_T}}$ $\frac{I_{S_2}}{V_T}e^{\frac{v_{D_2}}{V_T}}$
$-\frac{I_{S_1}}{V_T}e^{\frac{v_{D_1}}{V_T}}$ $-\frac{I_{S_2}}{V_T}e^{\frac{v_{D_2}}{V_T}}$
0 0
−1 1 0
−1 1 0
1 0
0 1
$\mathbf{J}_{\text{lim/MNA}}$
$\mathbf{J}_{\text{lim/lim}}$
(identity sub-matrix)
$i = 0$
$\mathbf{x_0} = \begin{bmatrix}\mathbf{x_{0,\text{MNA}}}\\ \mathbf{x_{0,\text{lim}}}\end{bmatrix}$
Independent initialization by different devices/sub-circuits
$i = i + 1$
No
Start
Initialize
Predict
Correct
Converged?
Yes
Return $\mathbf{x_{i,\text{MNA}}}$
(i) $\boldsymbol{\Delta}\mathbf{x}_{\text{MNA}} = -\left(\left(\mathbf{J}_{\text{MNA/MNA}} - \mathbf{J}_{\text{MNA/lim}}\,\mathbf{J}_{\text{lim/MNA}}\right)^{-1}\left(\mathbf{g}_{\text{MNA}} - \mathbf{J}_{\text{MNA/lim}}\,\mathbf{g}_{\text{lim}}\right)\right)\Big|_{\mathbf{x_i}}$
(ii) $\boldsymbol{\Delta}\mathbf{x}_{\text{lim}} = -\left(\mathbf{g}_{\text{lim}}\big|_{\mathbf{x_i}} + \mathbf{J}_{\text{lim/MNA}}\big|_{\mathbf{x_i}}\,\boldsymbol{\Delta}\mathbf{x}_{\text{MNA}}\right)$
(iii) $\underbrace{\begin{bmatrix}\mathbf{x_{i+1,\text{MNA}}}\\ \mathbf{x_{i+1,\text{lim}}}\end{bmatrix}}_{\mathbf{x_{i+1}}} = \underbrace{\begin{bmatrix}\mathbf{x_{i,\text{MNA}}}\\ \mathbf{x_{i,\text{lim}}}\end{bmatrix}}_{\mathbf{x_i}} + \begin{bmatrix}\boldsymbol{\Delta}\mathbf{x}_{\text{MNA}}\\ \boldsymbol{\Delta}\mathbf{x}_{\text{lim}}\end{bmatrix}$
$\mathbf{x_{i+1,\text{lim}}} = \texttt{refine}(\mathbf{x_i},\ \mathbf{x_{i+1}})$
Independent limiting by different devices/sub-circuits
Sparse Schur complement based Ax=b solve of same size as original MNA system

Fig. 2 The generalized PCNR flow that replaces limiting, and its application to the circuit shown in Fig. 1

The prediction phase is identical to traditional NR (without limiting); at the i^{th} iteration, the following update is applied:

$$\mathbf{x_{i+1}} = \mathbf{x_i} - \left(\left.\frac{d\mathbf{g}}{d\mathbf{x}}\right|_{\mathbf{x_i}}\right)^{-1} \mathbf{g}(\mathbf{x_i}), \quad \text{for } i \geq 0.$$

The naïve way to apply the above involves solving the following sparse $Ax = b$ problem of size $|\mathbf{x}_{\text{MNA}}| + |\mathbf{x}_{\text{lim}}|$:

$$\left.\frac{d\mathbf{g}}{d\mathbf{x}}\right|_{\mathbf{x_i}} \Delta\mathbf{x_i} = \begin{bmatrix}\mathbf{J}_{\text{MNA/MNA}} & \mathbf{J}_{\text{MNA/lim}}\\ \mathbf{J}_{\text{lim/MNA}} & \mathbf{J}_{\text{lim/lim}} = \mathbf{I}\end{bmatrix}\begin{bmatrix}\Delta\mathbf{x}_{\text{MNA}}\\ \Delta\mathbf{x}_{\text{lim}}\end{bmatrix} = -\mathbf{g}(\mathbf{x_i}) = \begin{bmatrix}-\mathbf{g}_{\text{MNA}}\\ -\mathbf{g}_{\text{lim}}\end{bmatrix}.$$

Writing the above as block matrix equations, we obtain:

$$\mathbf{J}_{\text{MNA/MNA}}\Delta\mathbf{x}_{\text{MNA}} + \mathbf{J}_{\text{MNA/lim}}\Delta\mathbf{x}_{\text{lim}} = -\mathbf{g}_{\text{MNA}}, \text{ and}$$
$$\mathbf{J}_{\text{lim/MNA}}\Delta\mathbf{x}_{\text{MNA}} + \Delta\mathbf{x}_{\text{lim}} = -\mathbf{g}_{\text{lim}}.$$

Substituting $\Delta\mathbf{x}_{\text{lim}}$ from the second equation into the first equation above, we eliminate $\Delta\mathbf{x}_{\text{lim}}$ to obtain:

$$\Delta\mathbf{x}_{\text{MNA}} = \left(\mathbf{J}_{\text{MNA/MNA}} - \mathbf{J}_{\text{MNA/lim}}\,\mathbf{J}_{\text{lim/MNA}}\right)^{-1}\left(\mathbf{J}_{\text{MNA/lim}}\,\mathbf{g}_{\text{lim}} - \mathbf{g}_{\text{MNA}}\right),$$

which represents a sparse $Ax = b$ problem of size just $|\mathbf{x}_{\text{MNA}}|$. Once this is solved for $\Delta\mathbf{x}_{\text{MNA}}$, we can obtain $\Delta\mathbf{x}_{\text{lim}}$ using the following equation:

$$\Delta\mathbf{x}_{\text{lim}} = -\mathbf{g}_{\text{lim}} - \mathbf{J}_{\text{lim/MNA}}\ \Delta\mathbf{x}_{\text{MNA}}.$$

This block based elimination is a well-known technique (that goes by the name "Schur complement reduction"[4] in the literature), and it is our third key insight; it enables us to partially mitigate the computational cost of the extra unknowns introduced by PCNR, since it allows us to solve a sparse $Ax = b$ problem of size $|\mathbf{x}_{\text{MNA}}|$, rather than one of size $|\mathbf{x}_{\text{MNA}}| + |\mathbf{x}_{\text{lim}}|$. Also, it can be shown that PCNR never takes more iterations than traditional limiting, which ensures scalability.

Finally, in the correction phase, the updates computed during the prediction phase are limited, by requesting each device/sub-circuit to limit the solution variables it owns. So, in PCNR, limiting is *explicitly* invoked by the simulator, rather than being *implicitly* done by the devices without the simulator's knowledge. This increases both modularity and flexibility.

4 Experimental Results

We now compare the number of NR iterations and the total runtime of PCNR against traditional limiting.

To do so, we run both PCNR and traditional limiting on the family of circuits shown in Fig. 3. As the figure shows, these circuits contain n repeated blocks of a subcircuit X, in addition to a voltage source v_{src} and a load resistance R_L. The subcircuit X in turn contains two parallel diodes and two resistors, as shown in Fig. 3. Thus, by increasing n, one can generate larger and larger circuits (n blocks $\implies |\mathbf{x}_{\text{MNA}}| = 2n + 2$ and $|\mathbf{x}_{\text{lim}}| = 2n$) belonging to this family, which allows us to study the asymptotic behaviour of PCNR and compare it against traditional limiting.

Table 1 shows the results. As we see from the second and third columns of the table, for a given level of accuracy, PCNR always takes the same number of NR iterations as traditional limiting, to converge to a circuit solution. However, the total runtime of PCNR tends to be higher than that of traditional limiting (typically by a factor of about 1.7). This is because PCNR iterations involve more computational effort than traditional limiting iterations, not only because PCNR iterations contain two phases ("predict" followed by "correct"), but also because of the extra unknowns introduced by the PCNR formulation (whose burden is partially but not completely offset by the Schur complement based technique described in Sect. 3). As mentioned before, we believe that the extra computational costs of PCNR are well worth paying—because the benefits of PCNR (simplicity, consistency, modularity, and flexibility) are substantial.

[4] https://en.wikipedia.org/wiki/Schur_complement.

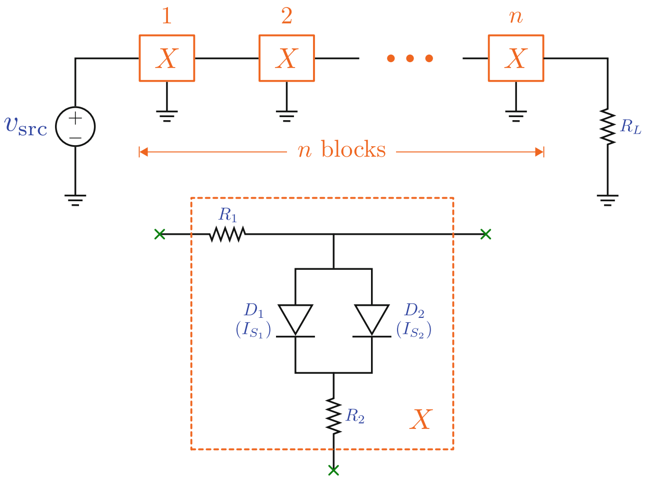


Fig. 3 The family of circuits (parameterised by n) that we use to illustrate the accuracy and computational efficiency of PCNR, compared against traditional limiting

Table 1 Computational cost of PCNR vs traditional limiting, in terms of number of NR iterations and total runtime, for the family of circuits shown in Fig. 3, as n increases from 1 to 1 million

	#NR iters		Runtime	
n	Traditional limiting	PCNR	Traditional limiting	PCNR
1	7	7	2.08 ms	7.69 ms
10	17	17	12.28 ms	29.67 ms
100	18	18	89.89 ms	162.17 ms
1000	21	21	1.00 s	1.6848 s
10,000	25	25	11.96 s	20.51 s
100,000	25	25	121.42 s	207.17 s
1,000,000	25	25	1213.00 s	2051.80 s

5 Conclusions

Thus, PCNR is a simple, scalable, and easy-to-understand replacement for limiting. It allows device models to use a stateless API to communicate with circuit simulators, and frees them from cumbersome bookkeeping, which is especially attractive for next-generation CPU + GPU architectures. Also, since PCNR reduces code complexity and increases modularity and flexibility, we believe that it can be used to rapidly develop and test robust limiting strategies at the device, sub-circuit, and circuit levels—for mainstream as well as newly emerging devices. Finally, we believe that PCNR's generic predictor/corrector flow opens the door to developing limiting-inspired heuristics to accelerate NR convergence in domains outside circuit simulation.

Acknowledgments This work was sponsored by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

1. Nagel, L.W.: SPICE2: a computer program to simulate semiconductor circuits. Ph.D. thesis, The University of California at Berkeley (1975)
2. Ho, C.W., Ruehli, A., Brennan, P.: The modified nodal approach to network analysis. IEEE Trans. Circuits Syst. **22**(6), 504–509 (1975)
3. Roychowdhury, J.: Numerical simulation and modelling of electronic and biochemical systems. Found. Trends Electron. Des. Autom. **3**(2–3), 97–303 (2009)
4. Sangiovanni-Vincentelli, A.L.: Computer design aids for VLSI circuits. In: Circuit Simulation, pp. 19–112. Springer, Dordrecht (1984)
5. Keiter, E.R., Hutchinson, S.A., Hoekstra, R.J., Russo, T.V., Waters, L.J.: Xyce® parallel electronic simulator design: mathematical formulation. Tech. Rep. SAND2004-2283, Sandia National Laboratories, Albuquerque, NM (2004)
6. Neamen, D.A.: Semiconductor Physics and Devices: Basic Principles, 4th edn. McGraw-Hill, New York (2011)
7. Kao, W.H.: Comparison of quasi-Newton methods for the DC analysis of electronic circuits. Master's thesis, The University of Illinois at Urbana-Champaign (1972)
8. Wang, T., Roychowdhury, J.: Well-posed models of memristive devices (2016). ArXiv. e-prints

Part IV
Mathematical and Computational Methods

Five papers belong to the part dedicated to Mathematical and Computational Methods.

In *GCA-H^2 matrix compression for electrostatic simulations* by S. Börm and S. Christophersen a compression method is proposed for boundary element matrices arising in the context of the computation of electrostatic fields. Green cross approximation combines an analytic approximation of the kernel function based on Green's representation formula and quadrature with an algebraic cross approximation scheme in order to obtain both the robustness of analytic methods and the efficiency of algebraic ones. One particularly attractive property of the new method is that it is well-suited for acceleration via general-purpose graphics processors (GPUs).

The contribution *On symmetry reductions of a third-order partial differential equation* by M.S. Bruzón et al. is more mathematical-oriented and it is devoted to perform symmetry reductions of a third-order partial differential equation belonging to a wide class which models many real world phenomena. In particular, many third-order partial differential equations of this class appear in different macroscopic models for semiconductors which consider quantum effects such as, for instance, quantum hydrodynamic models or models for the transmission of electrical lines.

The yield of an Integrated Circuit (IC) is commonly expressed as the fraction (in percentage) of working chips over all the manufactured chips and often interpreted as the failure probability of its analog blocks. The chapter *An Unbiased Hybrid Importance Sampling Monte Carlo Approach for Yield Estimation in Electronic Circuit Design* by A. Kumar et al. considers the Importance Sampling Monte Carlo (ISMC) as a reference method for estimating failure probabilities. For situations, where only a limited number of simulations is allowed, ISMC remains unattractive. In such cases, it is proposed an unbiased hybrid Monte Carlo approach that provides a fast estimation of the probability through a combination of a surrogate model, ISMC technique and the stratified sampling.

The contribution *Shape optimization of a PM synchronous machine under probabilistic constraints* by P. Putek et al. proposes a robust and reliability-based shape optimization method to find the optimal design of a permanent magnet

(PM) synchronous machine. Specifically, design of rotor poles and stator teeth is subjected to the shape optimization under manufacturing tolerances/imperfections and probabilistic constraints. In a forward problem, certain parameters are assumed to be random. This affects also shape optimization problems formulated in terms of a tracking-type robust cost functional where probabilistic constraints are introduced in order to attain a new, desired, robust design. The topological gradient is evaluated using the Topological Asymptotic Expansion Method, with the application of a Stochastic Collocation Method. The approach is illustrated providing the optimization results for a 2D model of a PM machine.

In magnetics, topology optimization (TO) is a tool helping to find a suitable ferromagnetic material space distribution in order to meet magnetic specifications. TO is a tool that becomes very interesting when the designer looks for new and original structures. TO is also used to design a Hall-effect thruster but the topological solutions are often not feasible. In the work *Topology Shape and Parametric Design Optimization of Hall Effect Thrusters* by R. Youness et al., in order to remedy to this, shape optimization (SO) and parametric optimization (PO) are carried out on the topological solution. SO and PO take account of the manufacturing constraints as well as the non linearity of the ferromagnetic materials.

GCA-$\mathcal{H}^2$ Matrix Compression for Electrostatic Simulations

Steffen Börm and Sven Christophersen

Abstract We consider a compression method for boundary element matrices arising in the context of the computation of electrostatic fields. Green cross approximation combines an analytic approximation of the kernel function based on Green's representation formula and quadrature with an algebraic cross approximation scheme in order to obtain both the robustness of analytic methods and the efficiency of algebraic ones. One particularly attractive property of the new method is that it is well-suited for acceleration via general-purpose graphics processors (GPUs).

1 Introduction

Boundary integral formulations are particularly useful when dealing with electrostatic exterior domain problems: we only have to construct a mesh for the boundary of the domain, and once an integral equation on this boundary has been solved, we can directly evaluate the electrostatic field in all points of the infinite domain by computing a surface integral.

Standard formulations typically lead to equations of the form

$$\int_{\partial\Omega} g(x,y)u(y)\,dy = \lambda u(x) + \int_{\partial\Omega} \frac{\partial g}{\partial n_y}(x,y)v(y)\,dy$$

for all $x \in \partial\Omega$, where $\Omega \subseteq \mathbb{R}^3$ is a domain, $\lambda \in \mathbb{R}$, u are the Neumann and v the Dirichlet boundary values, and

$$g(x,y) = \frac{1}{4\pi\,\|x-y\|}$$

S. Börm (✉) · S. Christophersen
Department of Mathematics, University of Kiel, Kiel, Germany
e-mail: boerm@math.uni-kiel.de; christophersen@math.uni-kiel.de

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_20

is the fundamental solution of Laplace's equation. Once Neumann and Dirichlet values are available, the electrostatic potential can be evaluated using Green's representation formula.

Discretization by Galerkin's method with basis functions $(\varphi_i)_{i \in \mathcal{I}}$ leads to a matrix $G \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ given by

$$g_{ij} = \int_{\partial\Omega} \varphi_i(x) \int_{\partial\Omega} g(x, y)\varphi_j(y)\, dy\, dx \qquad (1)$$

for all $i, j \in \mathcal{I}$, and all of these coefficients are typically non-zero.

Working directly with the matrix G is unattractive, since for $n := \#\mathcal{I}$ basis functions, we would have to store n^2 coefficients and quickly run out of memory.

This problem can be fixed by taking advantage of the properties of the kernel function g: analytic approximation schemes like the fast multipole method [11, 14], Taylor expansion [12], or interpolation [5, 8] replace g in suitable subdomains of the boundary $\partial\Omega$ by a short sum

$$g(x, y) \approx \sum_{\nu=1}^{k} a_\nu(x) b_\nu(y)$$

that leads to a low-rank approximation of corresponding submatrices of G, while algebraic schemes like the adaptive cross approximation (ACA) [1, 2, 17] or rank-revealing factorizations [9] directly construct low-rank approximations based on the matrix entries.

Hybrid approximation schemes like generalized fast multipole methods [10, 18] or hybrid cross approximation (HCA) [6] combine the concepts of analytic and algebraic approximation in order to obtain the near-optimal compression rates of algebraic methods while preserving the stability and robustness of analytic techniques.

Our algorithm falls into the third category: *Green cross approximation* (GCA) combines an analytic approximation based on Green's representation formula with adaptive cross approximation (ACA) to obtain low-rank approximations of submatrices. In order to improve the efficiency, we employ GCA in a recursive fashion that allows us to significantly reduce the storage requirements without losing the method's fast convergence.

While all of these technique allow us to handle the boundary integral equation more or less efficiently, analytic methods and some of the hybrid methods can also be used to speed up the subsequent evaluation of the electrostatic field in arbitrary points of Ω.

2 Green Quadrature

In order to find a data-sparse approximation of G, we consider a domain $\tau \subseteq \mathbb{R}^3$ and a superset $\omega \subseteq \mathbb{R}^3$ such that the distance from τ to the boundary $\partial\omega$ of ω is non-zero. For any $y \in \mathbb{R}^3 \setminus \overline{\omega}$, the function $x \mapsto g(x, y)$ is harmonic in ω, so we can apply Green's representation formula (also known as Green's third identity) to obtain

$$g(x, y) = \int_{\partial\omega} g(x, z)\frac{\partial g}{\partial n_z}(z, y) - \frac{\partial g}{\partial n_z}(x, z)g(z, y)\,dz$$

for all $x \in \tau$ and $y \in \mathbb{R}^3 \setminus \overline{\omega}$. If the distances between $\partial\omega$ and τ and between $\partial\omega$ and y are sufficiently large, the integrand is smooth, and we can approximate the integral by a quadrature rule to find

$$g(x, y) \approx \sum_{\nu=1}^{k} w_\nu g(x, z_\nu)\frac{\partial g}{\partial n_z}(z_\nu, y) - w_\nu\frac{\partial g}{\partial n_z}(x, z_\nu)g(z_\nu, y) \tag{2}$$

with weights w_ν and quadrature points z_ν, and in this approximation the variables x and y are separated.

This gives rise to a first low-rank approximation of G: given subsets $\widehat{\tau}, \widehat{\sigma} \subseteq \mathcal{I}$ of the index set, we can introduce axis-parallel boxes

$$\tau \supseteq \bigcup_{i\in\widehat{\tau}} \operatorname{supp}\varphi_i, \qquad \sigma \supseteq \bigcup_{j\in\widehat{\sigma}} \operatorname{supp}\varphi_j$$

containing the supports of the corresponding basis functions, and if these boxes are well-separated, we can find a superset ω of τ such that its boundary $\partial\omega$ is sufficiently far from both τ and σ. Replacing g in the definition (1) of the Galerkin matrix by the quadrature-based approximation leads to a factorized approximation

$$G|_{\widehat{\tau}\times\widehat{\sigma}} \approx A_{\tau\sigma}B_{\tau\sigma}^*,$$

with $A_{\tau\sigma} \in \mathbb{R}^{\widehat{\tau}\times 2k}$ and $B_{\tau\sigma} \in \mathbb{R}^{\widehat{\sigma}\times 2k}$. Here we use the notation $\mathbb{R}^{\widehat{\tau}\times 2k}$ to denote matrices with row indices from the set $\widehat{\tau}$ and column indices between 1 and $2k$, so the rank of the approximation $A_{\tau\sigma}B_{\tau\sigma}^*$ is bounded by $2k$.

The matrix coefficients are given by

$$a_{\tau\sigma,i\nu} = \sqrt{w_\nu}\int_{\partial\Omega} g(x, z_\nu)\varphi_i(x)\,dx,$$

$$a_{\tau\sigma,i(\nu+k)} = -d_\tau\sqrt{w_\nu}\int_{\partial\Omega} \frac{\partial g}{\partial n_z}(x, z_\nu)\varphi_i(x)\,dx,$$

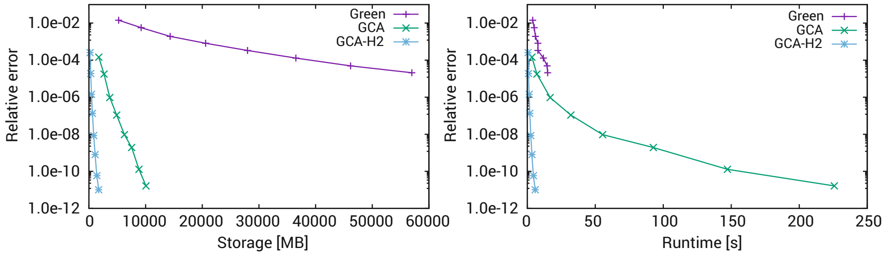


Fig. 1 Relative error of the Green quadrature approximation, GCA, and GCA-$\mathcal{H}^2$ compared to the storage requirements and setup time

$$b_{\tau\sigma,j\nu} = \sqrt{w_\nu} \int_{\partial\Omega} \frac{\partial g}{\partial n_z}(z_\nu, y)\varphi_j(y)\,dy,$$

$$b_{\tau\sigma,j(\nu+k)} = \frac{\sqrt{w_\nu}}{d_\tau} \int_{\partial\Omega} g(z_\nu, y)\varphi_j(y)\,dy,$$

where the scaling parameter $d_\tau = \operatorname{diam}(\tau)$ serves to balance the different scaling behaviour of the kernel function and its normal derivative.

We apply the approximation scheme to a polygonal approximation of the unit sphere by $n = 32{,}768$ triangles, choosing piecewise constant basis functions and the admissibility condition

$$\max\{\operatorname{diam}(\tau), \operatorname{diam}(\sigma)\} \leq 2\eta \operatorname{dist}(\tau, \sigma)$$

with a parameter $\eta \in \mathbb{R}_{>0}$ to determine whether a submatrix $G|_{\hat{\tau}\times\hat{\sigma}}$ can be approximated. The resulting approximation errors can be found labeled “Green” in Fig. 1.

3 Green Cross Approximation

In order to make the approximation more efficient, we can apply adaptive cross approximation [1] to derive the algebraic counterpart of interpolation: this technique provides us with a small subset $\tilde{\tau} \subseteq \hat{\tau}$ and a matrix $V_\tau \in \mathbb{R}^{\hat{\tau}\times\tilde{\tau}}$ such that

$$V_\tau A_{\tau\sigma}|_{\tilde{\tau}\times 2k} \approx A_{\tau\sigma},$$

i.e., we can reconstruct $A_{\tau\sigma}$ using only a few of its rows. Since $A_{\tau\sigma}$ is a thin matrix, we can afford to use reliable pivoting strategies and do not have to rely on heuristics. We conclude

$$V_\tau G|_{\tilde{\tau}\times\hat{\sigma}} \approx V_\tau A_{\tau\sigma}|_{\tilde{\tau}\times 2k} B_{\tau\sigma}^* \approx A_{\tau\sigma} B_{\tau\sigma}^* \approx G|_{\hat{\tau}\times\hat{\sigma}},$$

i.e., the algebraic interpolation can also be applied directly to the original matrix G instead of the low-rank approximation. This is called a *Green cross approximation* (GCA).

It is important to keep in mind that the matrices $A_{\tau\sigma}$ only depend on τ, but not on σ, so the cross approximation algorithm has to be performed only once for each τ and both the set $\tilde{\tau}$ and the matrix V_τ do not depend on σ.

Compared to the simple quadrature approximation, this modification has two major advantages: on one hand, the ranks are bounded by both the cardinality of $\widehat{\tau}$ and the number of quadrature points, so that the approximation can be far more efficient for small clusters. On the other hand, we can reach significantly higher accuracies, since the Green quadrature is only used to choose good "interpolation points" $\tilde{\tau}$, while the final approximation relies on the entries of the original matrix.

4 $\mathcal{H}^2$-Matrices

Since Green's formula is symmetric with respect to τ and σ, we can also apply the representation formula to a superset of σ and combine the formula with quadrature and cross approximation to obtain a subset $\tilde{\sigma} \subseteq \widehat{\sigma}$ and $V_\sigma \in \mathbb{R}^{\widehat{\sigma}\times\tilde{\sigma}}$ with

$$G|_{\widehat{\tau}\times\widehat{\sigma}} \approx G|_{\widehat{\tau}\times\tilde{\sigma}} V_\sigma^*.$$

Together with the approximation for τ introduced before, we obtain the *symmetric* factorization

$$G|_{\widehat{\tau}\times\widehat{\sigma}} \approx V_\tau G|_{\tilde{\tau}\times\tilde{\sigma}} V_\sigma^*,$$

and this turns out to be very efficient, since $G|_{\tilde{\tau}\times\tilde{\sigma}}$ is usually significantly smaller than $G|_{\widehat{\tau}\times\widehat{\sigma}}$.

We can improve the construction further by representing the *basis matrices* V_τ and V_σ in a hierarchy: assume that $\widehat{\tau}$ is subdivided into disjoint subsets $\widehat{\tau_1}$ and $\widehat{\tau_2}$ and that matrices V_{τ_1}, V_{τ_2} and subsets $\tilde{\tau}_1 \subseteq \widehat{\tau_1}$, $\tilde{\tau}_2 \subseteq \widehat{\tau_2}$ have already been constructed. We let $\tilde{\tau}_{1,2} := \tilde{\tau}_1 \cup \tilde{\tau}_2$ and observe

$$A_{\tau\sigma} = \begin{pmatrix} A_{\tau\sigma}|_{\widehat{\tau_1}\times 2k} \\ A_{\tau\sigma}|_{\widehat{\tau_2}\times 2k} \end{pmatrix} \approx \begin{pmatrix} V_{\tau_1} A_{\tau\sigma}|_{\tilde{\tau}_1\times 2k} \\ V_{\tau_2} A_{\tau\sigma}|_{\tilde{\tau}_2\times 2k} \end{pmatrix} = \begin{pmatrix} V_{\tau_1} & \\ & V_{\tau_2} \end{pmatrix} A_{\tau\sigma}|_{\tilde{\tau}_{1,2}\times 2k}.$$

If we now apply cross approximation to the right factor $\widehat{A}_{\tau\sigma} := A_{\tau\sigma}|_{\tilde{\tau}_{1,2}\times 2k}$, we obtain a subset $\tilde{\tau} \subseteq \tilde{\tau}_{1,2}$ and a matrix $\widehat{V}_\tau \in \mathbb{R}^{\tilde{\tau}_{1,2}\times\tilde{\tau}}$ with

$$\widehat{V}_\tau \widehat{A}_{\tau\sigma}|_{\tilde{\tau}\times 2k} \approx \widehat{A}_{\tau\sigma}$$

and therefore

$$A_{\tau\sigma} \approx \begin{pmatrix} V_{\tau_1} & \\ & V_{\tau_2} \end{pmatrix} \widehat{A}_{\tau\sigma} \approx \begin{pmatrix} V_{\tau_1} & \\ & V_{\tau_2} \end{pmatrix} \widehat{V}_\tau A_{\tau\sigma}|_{\tilde{\tau}\times 2k} = V_\tau A_{\tau\sigma}|_{\tilde{\tau}\times 2k},$$

where the basis matrix

$$V_\tau := \begin{pmatrix} V_{\tau_1} & \\ & V_{\tau_2} \end{pmatrix} \widehat{V}_\tau$$

can be expressed in the form

$$V_\tau = \begin{pmatrix} V_{\tau_1} E_{\tau_1} \\ V_{\tau_2} E_{\tau_2} \end{pmatrix}, \qquad E_{\tau_1} := \widehat{V}_\tau|_{\tilde{\tau}_1\times\tilde{\tau}}, \qquad E_{\tau_2} := \widehat{V}_\tau|_{\tilde{\tau}_2\times\tilde{\tau}}.$$

If we use this factorized representation of the matrices V_τ, we only have to store V_τ if $\widehat{\tau}$ has no subsets, while we use the substantially smaller *transfer matrices* E_{τ_1}, E_{τ_2} for all other index sets.

Since $\tilde{\tau}_{1,2}$ is usually significantly smaller than $\widehat{\tau}$, this construction is faster than the straightforward GCA approach, and the recursive use of transfer matrices reduces the storage requirements. The resulting approximation of G is known as an $\mathcal{H}^2$-matrix [3, 7, 13], and it can be proven to have *linear* complexity with respect to the matrix dimension n.

The resulting GCA-$\mathcal{H}^2$-matrix compression algorithm can be proven to converge exponentially and to have almost optimal complexity [4]. Indeed, Fig. 1 illustrates that the new algorithm requires only a few seconds to compute a highly accurate approximation.

5 Linear Basis Functions

So far, we have only considered piecewise constant basis functions in our experiments, since they make it particularly simple to approximate the entries of the matrix G. If the solution of the integral equation is smooth, it is generally a good idea to employ basis functions of higher order to obtain faster convergence.

One step up from piecewise constant basis functions are piecewise linear functions, and we choose *continuous* piecewise linear functions, both in order to reduce the number of unknown variables and to be able to work with integral operators that require an $H^{1/2}$-conforming trial space. The trial space is spanned by nodal basis functions φ_i: φ_i is continuous, piecewise linear on each triangle, equal to one in the i-th vertex, and equal to zero in all other vertices.

The support of φ_i consists of all triangles that contain the i-th vertex, therefore computing the entry g_{ij} of the matrix requires us to compute integrals on all pairs of triangles $t \times s$ where t belongs to the support of φ_i and s to the support of φ_j:

$$g_{ij} = \sum_{t \subseteq \operatorname{supp} \varphi_i} \sum_{s \subseteq \operatorname{supp} \varphi_j} \int_t \varphi_i(x) \int_s g(x, y) \varphi_j(y) \, dy \, dx.$$

Due to this property, the computation of one entry of the matrix with nodal basis functions can be *significantly* more computationally expensive than for a piecewise constant basis.

This problem can be somewhat mitigated by assembling the matrix triangle pair by triangle pair: we start with a zero matrix and loop over all pairs of triangles $t \times s$. For each pair, we evaluate the integrals for *all* of the triangles' vertices and add the results to the appropriate matrix coefficients. Although the final result is the same, we consider each pair of triangles only once, and this allows us to re-use the values of the kernel function g in the quadrature points for all combinations of basis functions. Since the evaluation of the transformed kernel function is the most computationally expensive part of the quadrature, this approach can make the entire construction far more efficient.

Since our compression scheme does not need *all* of the matrix entries, only entries for subsets $\tilde{\tau} \times \tilde{\sigma}$ or $\hat{\tau} \times \hat{\sigma}$, looping over *all* pairs of triangles would be a waste of time. Instead, we need an algorithm that determines only the required triangles and loops over them.

We typically store a matrix $G|_{\hat{\tau} \times \hat{\sigma}}$ by enumerating the indices $\hat{\tau} = \{i_1, \ldots, i_n\}$, $\hat{\sigma} = \{j_1, \ldots, j_m\}$ with $n = \#\hat{\tau}$, $m = \#\hat{\sigma}$, and using a matrix in $\mathbb{R}^{n \times m}$. This means that it is not enough to find which triangles have to participate in our computation, we also have to determine which index numbers correspond to the triangles' vertices.

Given a standard representation of the mesh, it is quite simple to determine for each index i the set T_i of triangles covering the support of the basis function φ_i. The challenge is to unify these sets for all basis functions corresponding to a subset $\hat{\tau}$ of indices. We use a variant of the well-known *mergesort* algorithm to handle this task: for example, assume that we have triangles

$$t_1 = (1, 2, 3), \qquad t_2 = (2, 3, 5), \qquad t_3 = (4, 1, 3),$$
$$t_4 = (6, 5, 2), \qquad t_5 = (1, 7, 4), \qquad t_6 = (7, 6, 1)$$

and are looking for the list of triangles for the vertex set $\hat{\tau} = \{1, 6, 4\}$. We have

$$T_1 = \{t_1, t_3, t_5, t_6\}, \qquad T_6 = \{t_4, t_6\}, \qquad T_4 = \{t_3, t_5\}.$$

We write the triangles for each vertex into columns of a matrix, where each column starts with the triangle, followed by three entries for its three vertices that are equal to the local index if this vertex is the current one or equal to the special symbol $\perp$ if it is not:

t_1	t_3	t_5	t_6	t_4	t_6	t_3	t_5
1	$\perp$	1	$\perp$	2	$\perp$	3	$\perp$
$\perp$	1	$\perp$	$\perp$	$\perp$	2	$\perp$	$\perp$
$\perp$	$\perp$	$\perp$	1	$\perp$	$\perp$	$\perp$	3

Now we apply the mergesort algorithm to sort the columns by the first row. If two columns have the same first row, i.e., if they correspond to the same triangle, the columns are combined: if a row has an index in one column and $\perp$ in the other, the combined column will have the index in this row. If the columns have $\perp$ in the same row, the combined column will, too. In our example, the result looks as follows:

t_1	t_3	t_4	t_5	t_6
1	3	2	1	$\perp$
$\perp$	1	$\perp$	$\perp$	2
$\perp$	$\perp$	$\perp$	3	1

Each triangle appears in exactly one column, and each column provides us with the local indices for all vertices of this triangle. The mergesort algorithm has a complexity of $\mathcal{O}(n \log n)$ if n indices with $\mathcal{O}(1)$ triangles per index are used, therefore the overhead for finding the triangles and the local indices is low compared to the computational work for the quadrature itself.

6 Curved Triangles

In our examples, linear basis function by themselves did reduce the storage requirements, but did not lead to faster convergence of the solution. Since the reason appears to be that the polygonal approximation of the smooth surface is insufficiently accurate, we consider replacing the piecewise linear parametrizations of the triangles by piecewise quadratic functions. This leads to *curved* triangles.

We implement these generalized triangles using the reference triangle $\hat{t} := \{x \in \mathbb{R}^2 \ : \ x_1, x_2 \geq 0, \ x_1 + x_2 \leq 1\}$ and quadratic parametrizations $\Phi_t, \Phi_s : \hat{t} \to \mathbb{R}^3$ such that

$$\begin{aligned}&\int_t \varphi_i(x) \int_s g(x,y)\varphi_j(y)\,dy\,dx\\&\quad = \int_{\hat{t}} \gamma_t(\hat{x})\varphi_i(\Phi_t(\hat{x})) \int_{\hat{t}} g(\Phi_t(\hat{x}), \Phi_s(\hat{y}))\gamma_s(\hat{y})\varphi_j(\Phi_s(\hat{y}))\,d\hat{y}\,d\hat{x}\end{aligned}$$

holds with the Gramians

$$\gamma_t(\hat{x}) = \left\| \frac{\partial \Phi_t}{\partial \hat{x}_1}(\hat{x}) \times \frac{\partial \Phi_t}{\partial \hat{x}_2}(\hat{x}) \right\|_2, \qquad \gamma_s(\hat{y}) = \left\| \frac{\partial \Phi_s}{\partial \hat{y}_1}(\hat{y}) \times \frac{\partial \Phi_s}{\partial \hat{y}_2}(\hat{y}) \right\|_2.$$

For the basis functions, we choose *mapped* nodal basis functions, i.e., $\varphi_i \circ \Phi_t$ and $\varphi_j \circ \Phi_s$ are nodal linear basis functions on the reference triangle $\hat{t}$, while φ_i and φ_j are not necessarily linear themselves.

We can evaluate the double integral by using Sauter's quadrature rule [15, 16], we only have to provide an efficient way of evaluating the parametrization and the Gramian in the quadrature points. For the parametrization, we simply use quadratic interpolation in the vertices and the midpoints of the edges. For the Gramian, we observe that the outer normal vector

$$n_t(\hat{x}) = \frac{\partial \Phi_t}{\partial \hat{x}_1}(\hat{x}) \times \frac{\partial \Phi_t}{\partial \hat{x}_2}(\hat{x})$$

is again a quadratic polynomial, so we can evaluate it also by interpolation once we have computed its values in the vertices and the midpoints. Once we have $n_t(\hat{x})$ at our disposal, computing $\gamma_t(\hat{x}) = \|n_t(\hat{x})\|_2$ is straightforward. If we want to evaluate the double-layer potential operator and need the *unit* outer normal vector, we can obtain it by simply dividing $n_t(\hat{x})$ by $\gamma_t(\hat{x})$.

Figure 2 shows the L^2 error of the solution compared to the storage requirements and the setup time for constant basis functions, linear basis functions on plane triangles, and linear basis functions on curved triangles. We can see that constant and linear basis functions on plane triangles converge at approximately the same rate, while curved triangles lead to a significantly improved rate of convergence and pronouncedly smaller errors for identical problem dimensions.

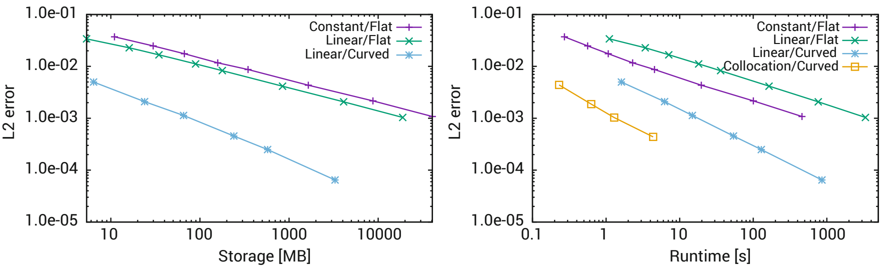


Fig. 2 L^2 error compared to the storage requirements and setup times for constant and linear basis functions as well as plane and curved triangles

7 GPU Implementation

Modern computers are frequently equipped with powerful graphics processors that are (reasonably) programmable and can therefore help with certain computational tasks. These processors are frequently called *general-purpose graphics processing units* (GPGPUs or short GPUs) and differ substantially from standard processors (CPUs). In order to use GPUs to accelerate our algorithm, we have to take the architectural differences into account.

A first important difference is the way CPUs and GPUs handle data: high-end GPUs typically are connected to dedicated high-bandwidth memory. While a current CPU may reach a memory bandwidth of 60 GBytes/s, modern GPUs provide up to 550 GBytes/s. It has to be pointed out that the higher bandwidth comes at a price: while even desktop CPUs can access 64 GBytes of RAM, with server CPUs accessing up to 1024 GBytes, current GPUs are limited to 24 GBytes of memory. In order to deal with large data sets, we have to move data between graphics memory and main memory, and these transfers are fairly slow.

The most important difference is the number of arithmetic units: while a 28-core CPU with 512-bit vector registers can perform $28 \times 16 = 448$ double-precision floating-point operations per clock, high-end GPUS offer currently up to 4 608 arithmetic units that can work in parallel. Even taking differences in clock speeds into account, the theoretical computing power of GPUs is significantly larger than that of CPUs.

On the other hand, GPUs restrict the ways the arithmetic units can be used: they are grouped together in *multiprocessors*, and all units in a multiprocessor have to either execute the same instruction or do nothing in each clock cycle. If the control flow of the program's threads diverges, i.e., if all of the threads have to execute different instructions, only one of the instructions can be executed per cycle, allowing only one of the threads to advance. Obviously, having, e.g., 63 of 64 arithmetic units idle for an extended period of time is not the best use of the available hardware.

8 GCA-$\mathcal{H}^2$ for GPUs

Let us now consider how to adapt our algorithm for execution on GPUs.

The computational work is dominated by three tasks:

- the construction of the leaf and transfer matrices V_τ and E_τ and the index sets $\tilde{\tau}$ by Green quadrature and cross approximation,
- the computation of the coupling matrices $G|_{\tilde{\tau}\times\tilde{\sigma}}$ for admissible blocks, and
- the computation of $G|_{\hat{\tau}\times\hat{\sigma}}$ for the remaining inadmissible blocks.

Although the entries of $A_{\tau\sigma}$ and $\widehat{A}_{\tau\sigma}$ involve no control-flow divergence and should therefore be well-suited for SIMT architectures, the highly adaptive nature of the cross approximation leads us to leave the first part of the algorithm to the CPU, where parallelization and vectorization can be employed to take full advantage of the available resources.

Once the sets $\tilde{\tau}$ and $\tilde{\sigma}$ are known, the computation of the entries of the matrices $G|_{\tilde{\tau}\times\tilde{\sigma}}$ and $G|_{\hat{\tau}\times\hat{\sigma}}$ requires no adaptivity whatsoever, so the second and third part of our algorithm can be expected to be ideally suited for GPUs.

Setting up the GPU to run a number of threads involves a certain amount of communication and management operations, therefore we should make sure that the number of threads is sufficiently high in order to minimize organizational overhead. Since our algorithm only works with small matrices that would not allow us to reach an adequate number of threads, we switch to an asynchronous execution model: instead of computing the entries of a matrix the moment it is encountered by our algorithm, the corresponding task is added to a list for later handling. Only once the list has grown enough to keep a sufficiently large number of threads busy, it is transferred to the GPU for execution.

This approach also allows us to handle different cases appearing in the numerical quadrature: we use Sauter's quadrature rule [15, 16] to integrate the singular kernel function on pairs of triangles. Sauter's algorithm requires different quadrature points (and even different numbers of quadrature points) depending on whether the triangles are identical, share an edge, a vertex, or are disjoint. By simply using one list for each of the four cases, we can ensure that all threads execute almost exactly the same sequence of instructions and that control-flow divergence is kept down to a minimum.

Since communication between the CPU and the GPU is slow, we should try to keep the amount of data that has to be transferred as small as possible. In our implementation, the geometrical information of the triangles is kept permanently in graphics memory, so that we only have to transfer the numbers of the triangles t and s in order to describe an integral that has to be computed.

Another important step in reducing the impact of the communication between CPU and GPU is to "hide" the communication behind computation: modern graphics cards can perform computations and memory transfers concurrently, and we use this feature in order to use the time spent by the arithmetic units on one list to transfer the results of the previous list back to main memory and the input of the next list to graphics memory.

Finally, we employ multiple threads on the CPU to fill multiple lists concurrently in order to ensure that both the memory management and the arithmetic units of the GPU are kept busy.

Figure 3 shows the runtime in seconds per degree of freedom for setting up the GCA-$\mathcal{H}^2$ matrix for different meshes approximating the unit sphere. For the CPU, we use an Intel Core i7-7820X with 8 cores and AVX 512 running at a base frequency of 3.6 GHz, providing a peak performance of 460 GFlops/s at double precision. For the GPU, we have used an AMD Vega 64 card running at 1.25 GHz with 8 GBytes of HBM2 memory and 4096 arithmetic units providing a peak performance of 791 GFlops/s at double precision (and *considerably* more for single precision). It should be pointed out that this GPU is designed for the

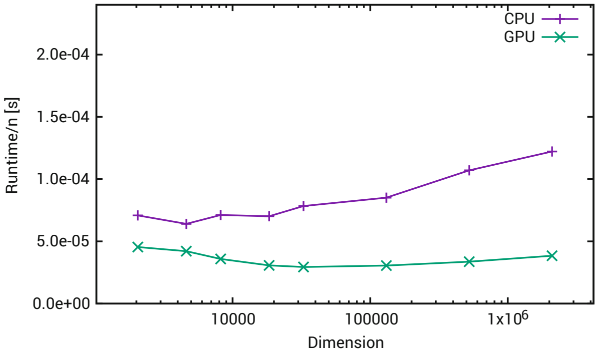


Fig. 3 Runtime of CPU and GPU setup of the GCA-$\mathcal{H}^2$-matrix

consumer market, so its double-precision performance is quite low. GPUs designed for computation like the NVIDIA Tesla P100 or V100 should provide around 5000 and 7000 GFlops/s at double precision, respectively.

Our figure suggests that both the CPU and the GPU implementation have $\mathcal{O}(n \log n)$ complexity: we use a logarithmic scale for the dimension n and a linear scale for the work per degree of freedom, and the figure shows the latter as, essentially, a line. We can also see that the slope of this line for the CPU implementation is significantly steeper than for the GPU implementation. This suggests that the GPU implementation may become increasingly more attractive for larger meshes.

References

1. Bebendorf, M.: Approximation of boundary element matrices. Numer. Math. **86**(4), 565–589 (2000)
2. Bebendorf, M., Rjasanow, S.: Adaptive low-rank approximation of collocation matrices. Computing **70**(1), 1–24 (2003)
3. Börm, S.: Efficient numerical methods for non-local operators: $\mathcal{H}^2$-matrix compression, algorithms and analysis. In: EMS Tracts in Mathematics, vol. 14. European Mathematical Society, Zürich (2010)
4. Börm, S., Christophersen, S.: Approximation of integral operators by Green quadrature and nested cross approximation. Numer. Math. **133**(3), 409–442 (2016)
5. Börm, S., Grasedyck, L.: Low-rank approximation of integral operators by interpolation. Computing **72**, 325–332 (2004)
6. Börm, S., Grasedyck, L.: Hybrid cross approximation of integral operators. Numer. Math. **101**, 221–249 (2005)
7. Börm, S., Hackbusch, W.: Data-sparse approximation by adaptive $\mathcal{H}^2$-matrices. Computing **69**, 1–35 (2002)
8. Börm, S., Löhndorf, M., Melenk, J.M.: Approximation of integral operators by variable-order interpolation. Numer. Math. **99**(4), 605–643 (2005)

9. Chandrasekaran, S., Ipsen, I.C.F.: On rank-revealing factorisations. SIAM J. Matrix Anal. Appl. **15**(2), 592–622 (1994)
10. Gimbutas, Z., Rokhlin, V.: A generalized fast multipole method for nonoscillatory kernels. SIAM J. Sci. Comput. **24**(3), 796–817 (2002)
11. Greengard, L., Rokhlin, V.: A new version of the fast multipole method for the Laplace equation in three dimensions. In: Acta Numerica 1997, pp. 229–269. Cambridge University Press, Cambridge (1997)
12. Hackbusch, W., Nowak, Z.P.: On the fast matrix multiplication in the boundary element method by panel clustering. Numer. Math. **54**(4), 463–491 (1989)
13. Hackbusch, W., Khoromskij, B.N., Sauter, S.A.: On $\mathcal{H}^2$-matrices. In: Bungartz, H., Hoppe, R., Zenger, C. (eds.) Lectures on Applied Mathematics, pp. 9–29. Springer, Berlin (2000)
14. Rokhlin, V.: Rapid solution of integral equations of classical potential theory. J. Comp. Phys. **60**, 187–207 (1985)
15. Sauter, S.A.: Cubature techniques for 3-d Galerkin BEM. In: Hackbusch, W., Wittum, G. (eds.) Boundary Elements: Implementation and Analysis of Advanced Algorithms, pp. 29–44. Vieweg-Verlag, Wiesbaden (1996)
16. Sauter, S.A., Schwab, C.: Boundary Element Methods. Springer, New York (2011)
17. Tyrtyshnikov, E.E.: Mosaic-skeleton approximation. Calcolo **33**, 47–57 (1996)
18. Ying, L., Biros, G., Zorin, D.: A kernel-independent adaptive fast multipole algorithm in two and three dimensions. J. Comp. Phys. **196**(2), 591–626 (2004)

On Symmetry Reductions of a Third-Order Partial Differential Equation

M. S. Bruzón, R. de la Rosa, M. L. Gandarias, and R. Tracinà

Abstract This work is devoted to perform symmetry reductions of a third-order partial differential equation belonging to a wide class which models many real-world phenomena. In particular, many third-order partial differential equations of this class appear in different macroscopic models for semiconductors, which consider quantum effects, for instance, quantum hydrodynamic models; or models for the transmission of electrical lines, among others. The symmetry group of the considered equation has been determined. We prove that this group constitutes a three-dimensional solvable Lie group which allows us to reduce the equation to a first-order nonlinear ODE. Furthermore, a solution of the equation is determined by quadrature in a particular case.

1 Introduction

The study of integrable equations which model real-world phenomena have drawn the attention of numerous researchers in the last years. Nevertheless, in the early period of development of integrability theory, most of the partial differential equations (PDEs) considered were constant-coefficient ones. The issue lies in the fact that variable-coefficient models describe nonlinear phenomena more realistically that constant-coefficient models.

Gandarias and Bruzón [5] considered the generalized third-order variable-coefficient equation given by

$$u_t = (g(u))_{xxx} + (f(u))_x \, , \tag{1}$$

M. S. Bruzón · R. de la Rosa (✉) · M. L. Gandarias
Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain
e-mail: m.bruzon@uca.es; rafael.delarosa@uca.es; marialuz.gandarias@uca.es

R. Tracinà
Dipartimento di Matematica e Informatica, Università di Catania, Catania, Italy
e-mail: tracina@dmi.unict.it

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_21

where $f(u)$ and $g(u)$ are arbitrary functions satisfying $f_u \neq 0$, $g_u \neq 0$. They obtained the point symmetries admitted by Eq. (1). Moreover, they determined the subclasses of family (1) which are self-adjoint and quasi self-adjoint. Then, by using a general theorem on conservation laws which does not require the existence of a classical Lagrangian, they found conservation laws for some equations belonging to class (1).

The symmetry group of a PDE is the largest transformation group which acts on dependent and independent variables of the equation in a manner that it maps solutions of the equation into other solutions. Symmetry groups have several well-known applications. Among them, we highlight local symmetries admitted by a PDE are useful for obtaining invariant solutions [3, 10]. The fundamental basis of this technique is that, when a differential equation is invariant under a Lie group of transformations, a reduction transformation exists. Furthermore, symmetry groups can also be used to determine conservation laws, to obtain exact solutions or the construction of maps between equivalent equations of the same family [1, 4, 11, 12, 16, 17]. In particular, in [15] Lie symmetry analysis was used to determine approximate solutions to a quantum drift-diffusion model for semiconductors.

We are interested to consider the following generalization of Eq. (1)

$$u_t = (g(u))_{xxx} + (f(u))_x + h(u)u_{xx}, \tag{2}$$

where $h(u)$ is an arbitrary function of u.

Several well-known equations belong to this class, as example, the Korteweg-de Vries equation, the Burgers equation and the Gardner equation. Furthermore, several third-order PDEs, many of them are part of class (2), appear in many macroscopic models for semiconductors, which consider quantum effects, for instance, quantum hydrodynamic models; or models for the transmission of electrical lines. The interested reader can refer, for instance, to [2, 6, 8, 14].

In this paper, for the sake of simplicity, we will restrict our attention to the subclass of Eq. (2) characterized by $f(u) = f_0e^{2pu} + f_1u$, $g(u) = g_0u$ and $h(u) = h_0e^{pu}$,

$$u_t = g_0u_{xxx} + h_0e^{pu}u_{xx} + \left(2pf_0e^{2pu} + f_1\right)u_x, \tag{3}$$

where $f_0 \neq 0$, f_1, $g_0 \neq 0$, h_0, $p \neq 0$ are arbitrary constants.

We notice that family (2) is preserved under the equivalence transformation given by

$$\tilde{u} \longrightarrow u + u_0, \quad u_0 \text{ constant},$$

which allows us to simplify the results achieved on point symmetries by considering in Eq. (3), without loss of generality, $f_0 = \frac{1}{2p}$. Thus, we consider the subclass of Eq. (2) given by

$$u_t = g_0 u_{xxx} + h_0 e^{p\,u} u_{xx} + \left(e^{2\,p\,u} + f_1\right) u_x. \tag{4}$$

2 Lie Algebra

For Eq. (4), a one-parameter Lie group of transformations is a transformation depending on the parameter ϵ

$$\begin{aligned} \hat{t}(t,x,u;\epsilon) &= t + \epsilon\,\tau(t,x,u) + O(\epsilon^2),\\ \hat{x}(t,x,u;\epsilon) &= x + \epsilon\,\xi(t,x,u) + O(\epsilon^2),\\ \hat{u}(t,x,u;\epsilon) &= u + \epsilon\,\eta(t,x,u) + O(\epsilon^2), \end{aligned} \tag{5}$$

such that the action of group (5) leaves the solution space of Eq. (4) invariant. A general element of the associated Lie algebra of Eq. (4) takes the form

$$X = \tau(t,x,u)\partial_t + \xi(t,x,u)\partial_x + \eta(t,x,u)\partial_u. \tag{6}$$

Therefore, point symmetries of Eq. (4) are obtained by applying the symmetry invariance condition

$$pr^{(3)}X(u_t - g_0 u_{xxx} - h_0 e^{p\,u} u_{xx} - \left(e^{2\,p\,u} + f_1\right) u_x) = 0, \tag{7}$$

when Eq. (4) is satisfied. Here, $pr^{(3)}X$ is the prolongation of generator X (6) to the space of the derivatives of the dependent variable up to third order [11].

The symmetry invariance condition (7) leads to a linear system of determining equations. By solving this system we obtain that the symmetry group of Eq. (4) is spanned by the following generators

$$X_1 = \partial_x, \quad X_2 = \partial_t, \quad X_3 = 3pt\partial_t + p(x - 2f_1 t)\partial_x - \partial_u, \tag{8}$$

with commutator structure

$$[X_1, X_3] = pX_1, \quad [X_2, X_3] = -2f_1 pX_1 + 3pX_2.$$

3 Exact Solutions

By using infinitesimal symmetries, it is well known that a PDE with two independent variables can be reduced to an ordinary differential equation (ODE). This reduction procedure can be performed taking into account the characteristic system

$$\frac{dt}{\tau} = \frac{dx}{\xi} = \frac{du}{\eta}.$$

The solutions of the characteristic system provide a similarity variable z, and a similarity solution $U(z)$. By substituting these new variables into Eq. (4) we obtain a third-order nonlinear ODE for $U(z)$.

However, it is not always evident how to solve the ODEs obtained. In fact, not all third-order nonlinear ODEs can be solved. One alternative is to prove if an n-order nonlinear ODE admits a solvable Lie group of dimension, at least, n. We recall that a sufficient condition for Eq. (4) can be reduced to quadrature is that the initial reduction comes from a point symmetry belonging to a four-dimensional solvable Lie group. Following reference [1], $\mathcal{A}^k$ is a k-dimensional solvable Lie algebra if there exists a chain of subalgebras $\mathcal{A}^{(1)} \subset \mathcal{A}^{(2)} \subset \ldots \subset \mathcal{A}^{(k-1)} \subset \mathcal{A}^{(k)} = \mathcal{A}^k$, with $\mathcal{A}^{(m)}$ an m-dimensional Lie algebra, being $\mathcal{A}^{(m-1)}$ an ideal of $\mathcal{A}^{(m)}$, $m = 1, 2, \ldots, k$. An equivalent formulation of this result more fitting for reducing differential equations by using symmetries is $\mathcal{A} \supset \mathcal{A}^{(1)} \supset \mathcal{A}^{(2)} \supset \ldots \supset \mathcal{A}^{(k)} \supset \mathcal{A}^{(k+1)} = 0$, where $\mathcal{A}^{(m)} = \left[\mathcal{A}^{(m-1)}, \mathcal{A}^{(m-1)}\right]$, $m = 1, 2, \ldots, k \leq \dim \mathcal{A}$. The three-dimensional symmetry group given by X_1, X_2 and X_3 (8) verifies

$$\mathcal{A} = \text{span}\,(X_1, X_2, X_3)\,, \quad \mathcal{A}^{(1)} = \text{span}\,(X_1, X_2)\,, \quad \mathcal{A}^{(2)} = 0.$$

Consequently, $\mathcal{A} = \text{span}\,(X_1, X_2, X_3)$ forms a three-dimensional solvable symmetry group. Taking into account $X = -f_1 X_1 + X_2$, we obtain the invariants

$$z = x + f_1 t \quad \text{and} \quad U(z) = u,$$

where $U(z)$ satisfies the third-order ODE given by

$$g_0 U''' + h_0 e^{pU} U'' + e^{2pU} U' = 0. \tag{9}$$

Equation (9) inherits the two-dimensional solvable symmetry algebra spanned by $Y = X_1$ and $Z = X_3$, which are written in the new variables as $Y = \partial_z$ and $Z = pz\partial_z - \partial_U$ satisfying $[Y, Z] = pY$. This allows us to integrate Eq. (9) as follows. From Y, we obtain the invariants

$$\omega = U \quad \text{and} \quad \chi = U',$$

therefore ODE (9) can be transformed into a second-order ODE

$$g_0\left(\chi'^2+\chi\chi''\right)+h_0e^{p\omega}\chi'+e^{2p\omega}=0. \tag{10}$$

After this point, it is necessary to distinguish two different cases:

- $h_0\neq 0$, for which Eq. (10) admits a one-dimensonal Lie algebra.
- $h_0=0$, for which Eq. (10) admits a three-dimensonal Lie algebra and can be integrated by quadrature.

3.1 $\boldsymbol{h_0\neq 0}$

It can be easily checked that Eq. (10) inherits

$$W=\mathrm{pr}^{(1)}Z\left|_{(\omega,\chi)}\right.=\partial_\omega+p\chi\partial_\chi,$$

as a symmetry. By using W, whose invariants are given by

$$\phi=e^{-p\omega}\chi \quad \text{and} \quad \gamma=e^{-p\omega}\chi',$$

Eq. (10) can be reduced to a first-order ODE for $\gamma(\phi)$

$$1+h_0\gamma+g_0\gamma^2+g_0p\phi\gamma+g_0\phi\gamma\gamma'-g_0p\phi^2\gamma'=0. \tag{11}$$

Dividing by $g_0\phi$ and regrouping terms Eq. (11) becomes

$$(\gamma-p\phi)\gamma'=-\frac{1}{\phi}\gamma^2-\left(p+\frac{h_0}{g_0\phi}\right)\gamma-\frac{1}{g_0\phi}, \tag{12}$$

which is an Abel equation of the second kind. Furthermore, by applying the substitution $\gamma(\phi)=\frac{1}{\nu(\phi)}+p\phi$, Eq. (12) is transformed into an Abel equation of the first kind

$$\nu'=\left(\frac{1}{g_0\phi}+\frac{h_0p}{g_0}+2p^2\phi\right)\nu^3+\left(4p+\frac{h_0}{g_0\phi}\right)\nu^2+\frac{1}{\phi}\nu. \tag{13}$$

Unfortunately, Eqs. (12) and (13) do not satisfy any of the integrability conditions presented, for instance, in [7, 9, 13] and references therein.

3.2 $h_0 = 0$

As in the previous case, Eq. (10) admits W as a symmetry. However, when $h_0 = 0$, Eq. (10) admits two new point symmetries

$$W_1 = -\frac{1}{\chi}\partial_\chi \quad \text{and} \quad W_2 = \frac{\omega}{\chi}\partial_\chi,$$

which are known as Type-II hydden symmetries of Eq. (9). By using W and W_1, Eq. (10) can be reduced to quadrature. To begin with, $\{W, W_1\}$ constitutes a two-dimensional solvable symmetry algebra of Eq. (10) satisfying $[W, W_1] = -2pW_1$. We have that

$$\phi = \omega \quad \text{and} \quad \gamma = \chi\chi',$$

are invariants of W_1 from which ODE (10) becomes in a first-order ODE for $\gamma(\phi)$

$$g_0\gamma'(\phi) + \frac{1}{2}e^{p\phi} = 0. \tag{14}$$

This equation admits the symmetry $V = pr^{(1)}W\,\big|_{(\phi,\gamma)} = 2\partial_\phi + 2p\gamma\partial_\gamma$. The canonical coordinates $\left\{r, s, s^1\right\}$, where $s^1 = \dfrac{ds}{dr}$, associated with V are given by

$$r = \frac{e^{p\phi}}{\gamma}, \quad s = \frac{\ln\gamma}{2p}, \quad s^1 = \frac{e^{-p\phi}\gamma\gamma'}{2p\,(p\gamma - \gamma')}.$$

Hence we obtain

$$s = -\frac{1}{2\,p}\ln(2\,p\,g_0 + r) \Longrightarrow \gamma(\phi) = \frac{c_1 - e^{p\phi}}{2\,g_0 p}.$$

After undoing the change of variables, we obtain that the general solution of Eq. (4) starting from generator X is given by

$$x + f_1 t \mp \int^{u(t,x)} \frac{g_0 p^2}{-e^{pm} + c_1 p\,m + 2\,c_2 g_0 p^2} dm + c_3 = 0, \quad c_1,\ c_2,\ c_3 \text{ constants.}$$

4 Conclusions

In this paper, by using Lie group method we have determined that Eq. (4) admits a three-dimensional symmetry group which is spanned by X_1, X_2 and X_3 (8). Furthermore, we have proved that this group is a three-dimensional solvable Lie group. Taking into account the symmetry generator $X = -f_1 X_1 + X_2$, which verifies

$$< X > \lhd < X, X_1 > \lhd < X, X_1, X_3 >,$$

we transform Eq. (4) into a third-order nonlinear ODE which inherits a two-dimensional symmetry group formed by X_1 and X_3, once they are written in the new variables. Therefore, the nonlinear PDE (4) can be transformed into a first-order nonlinear ODE (11). Nevertheless, in the case that $h_0 = 0$, the second-order equation (10) admits two Type-II hydden symmetries. Thus, the general solution of ODE (10) can be found by quadrature. Finally, by undoing the change of variables, we have determined the general solution of Eq. (4) with $h_0 = 0$ starting from generator X.

Acknowledgments R. Tracinà acknowledges Università di Catania for financial support, *Metodi gruppali e umbrali per modelli di diffusione e trasporto*. M.S. Bruzón, R. de la Rosa and M.L. Gandarias acknowledge the financial support from Junta de Andalucía group FQM-201, Universidad de Cádiz. M.S. Bruzón and M.L. Gandarias also acknowledge the support of DGICYT project MTM2009-11875 with the participation of FEDER. R. de la Rosa also expresses his sincere gratitude to the Plan Propio de Investigación de la Universidad de Cádiz.

References

1. Bluman, G.W., Anco, S.C.: Symmetry and Integration Methods for Differential Equations. Springer, New York (2002)
2. Crighton, D.C.: Applications of KdV. Acta Applicandae Mathematicae **39**, 39–67 (1995)
3. de la Rosa, R., Bruzón, M.S.: Symmetry reductions of a generalized Kuramoto-Sivashinsky equation via equivalence transformations. Commun. Nonlinear Sci. Numer. Simul. **63**, 12–20 (2018)
4. Freire, I.L., Faleiros, A.C.: Lie point symmetries and some group invariant solutions of the quasilinear equation involving the infinity Laplacian. Nonlinear Anal. TMA **74**, 3478–3486 (2011)
5. Gandarias, M.L., Bruzón, M.S.: Conservation laws for a class of quasi self-adjoint third order equations. Appl. Math. Comput. **219**, 668–678 (2012)
6. Gardner, C.L.: The classical and quantum hydrodynamic models. In: Proceedings of the International Workshop on Computational Electronics, pp. 25–36 (1993)
7. Harko, T., Lobo, F.S.N., Mak, M.K.: A Chiellini type integrability condition for the generalized first kind Abel differential equation. Univ. J. Appl. Math. **1**, 101–104 (2013)
8. Jüngel, A.: Quasi-hydrodynamic semiconductor equations. In: Progress in Nonlinear Differential Equations and Their Applications. Springer Basel AG, Basel (2001)
9. Kamke, E.: Differentialgleichungen: Lösungsmethoden und Lösungen. Springer, Stuttgart (1977)
10. Kontogiorgis, S., Sophocleous, C.: Group classification of systems of diffusion equations. Math. Meth. Appl. Sci. **40**, 1746–1756 (2017)
11. Olver, P.J.: Applications of Lie Groups to Differential Equations. Springer, New York (1986)
12. Ovsyannikov, L.V.: Group Analysis of Differential Equations. Academic, New York (1982)
13. Polyanin, A.D., Zaitsev, V.F.: Handbook of Exact Solutions for Ordinary Differential Equations. Chapman & Hall/CRC, New York (2003)

14. Romano, V.: Quantum corrections to the semiclassical hydrodynamical model of semiconductors based on the maximum entropy principle. J. Math. Phys. **48**, 123504 (2007)
15. Romano, V., Torrisi, M., Tracinà, R.: Approximate solutions to the quantum drift-diffusion model of semiconductors. J. Math. Phys. **48**, 023501 (2007)
16. Tracinà, R., Bruzón, M.S., Gandarias, M.L., Torrisi, M.: Nonlinear self-adjointness, conservation laws, exact solutions of a system of dispersive evolution equations. Commun. Nonlinear Sci. Numer. Simul. **19**, 3036–3043 (2007)
17. Tracinà, R., Bruzón, M.S., Gandarias, M.L.: On the nonlinear self-adjointness of a class of fourth-order evolution equations. Appl. Math. Comput. **275**, 299–304 (2016)

An Unbiased Hybrid Importance Sampling Monte Carlo Approach for Yield Estimation in Electronic Circuit Design

Anuj Kumar Tyagi, Xavier Jonsson, Theo Beelen, and W. H. A. Schilders

Abstract The yield of an Integrated Circuit (IC) is commonly expressed as the fraction (in %) of working chips overall manufactured chips and often interpreted as the failure probability of its analog blocks. We consider the Importance Sampling Monte Carlo (ISMC) as a reference method for estimating failure probabilities. For situations where only a limited number of simulations is allowed, ISMC remains unattractive. In such cases, we propose an unbiased hybrid Monte Carlo approach that provides a fast estimation of the probability. Hereby we use a combination of a surrogate model, ISMC technique and the stratified sampling.

1 Introduction

We study the problem of estimating yield of an Integrated Circuit (IC), through electric level simulations which are performed with the Eldo® simulator from Mentor Graphics®, under variability constraints due to the manufacturing process. To ensure a high yield of an IC, the failure probabilities of its components must be very small. For example, consider a 256 Mbit SRAM circuit, having a 256 million "identical" bitcells. To guarantee a yield of 99% of the SRAM, the failure probability of a single bitcel is required to be less than 3.9×10^{-11}. For details, see [9].

We consider Monte Carlo (MC) techniques for estimating the failure probability of a circuit. For complex systems one usually can only afford a limited number (say, hundreds) of simulations, and therefore standard MC method is not directly applicable and a variance reduction Importance Sampling MC (ISMC) technique remains unattractive. Then an effective hybrid ISMC (HISMC) approach is introduced in

A. K. Tyagi (✉) · T. Beelen · W. H. A. Schilders
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: a.k.tyagi@tue.nl; t.g.j.beelen@tue.nl; w.h.a.schilders@tue.nl

X. Jonsson
Mentor, A Siemens Business, Grenoble, France
e-mail: xavier_jonsson@mentor.com

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_22

[9]. The main drawback of HISMC is that the authors in [9] do not prove that the resulted probability estimator is unbiased. In this paper, we propose an Unbiased-HISMC (UHISMC) approach which is identical to HISMC in the exploration but different in the estimation phase using the controlled stratified sampling as in [2].

The outline of the paper is as follows: in Sect. 2, we give a brief summary of ISMC. Then we introduce our UHISMC approach in Sect. 3. The numerical results are presented in Sect. 4. Finally, we end with a conclusion in Sect. 5.

2 The Importance Sampling Monte Carlo Technique

2.1 *Basic Idea*

ISMC is a well-known variance reduction technique for rare-event simulations. For literature on ISMC techniques we refer to [1]. ISMC tunes the basic MC to an area in the parameter space from where the rare-events are generated, i.e., one seeks to sample the random variables from a different distribution, called the Importance Sampling (IS) distribution, rather than the original distribution.

Figure 1 illustrates the ISMC technique in a two dimensional space, i.e., $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2)$ is a vector of two input variables. The samples drawn from $g(\mathbf{x})$ are assumed to be distributed in the red ellipse centered at origin. In this case no failure would occur, since failure lies outside the red ellipse. On the other hand, ISMC allows us to sample from the new pdf (say $\psi(\mathbf{x})$) that generates the samples (assumed to be fallen in the green ellipse) around the limit state surface $\mathcal{L}_\gamma$.

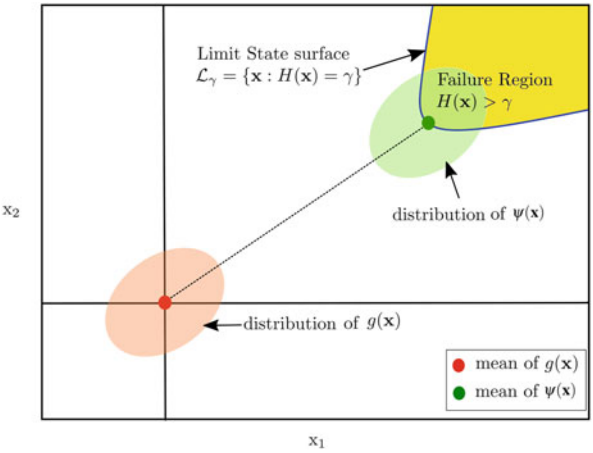

Limit State surface
$\mathcal{L}_\gamma = \{\mathbf{x} : H(\mathbf{x}) = \gamma\}$
Failure Region
$H(\mathbf{x}) > \gamma$
distribution of $\psi(\mathbf{x})$
distribution of $g(\mathbf{x})$
● mean of $g(\mathbf{x})$
● mean of $\psi(\mathbf{x})$
x_2
x_1

Fig. 1 Illustration of the ISMC technique

2.2 Mathematical Formulation

Let $\mathbf{X} \in \mathbb{R}^d$ be the input vector of the circuit under study and $\mathbf{x}$ a realization of $\mathbf{X}$, with probability density function (pdf) $g(\mathbf{x})$, and let $H(\mathbf{X})$ be the corresponding response of the circuit. Then the failure probability $p_{\text{fail}} = \mathbb{P}(H(\mathbf{X}) \geq \gamma)$ of the circuit is defined as

$$p_{\text{fail}} = \mathbb{E}_g\left[\mathbb{1}_{\{H(\mathbf{X}) \geq \gamma\}}\right] = \int \mathbb{1}_{\{H(\mathbf{x}) \geq \gamma\}} g(\mathbf{x}) d\mathbf{x} \tag{1}$$

where subscript g means that the expectation is taken with respect to the pdf $g(\mathbf{x})$, γ is a given failure threshold and $\mathbb{1}_{\{H(\mathbf{x}) \geq \gamma\}}$ is an indicator function that gives the value 1 if $H(\mathbf{x} \geq \gamma)$, 0 otherwise.

In ISMC, we sample from another distribution (from which the rare-events are generated) rather than the original. Let $\psi(\mathbf{x})$ be the new density function. Then probability equation (1) can be written as

$$p_{\text{fail}} = \int \mathbb{1}_{(H(\mathbf{x}) \geq \gamma)} g(\mathbf{x}) d\mathbf{x} = \int \mathbb{1}_{(H(\mathbf{x}) \geq \gamma)} \frac{g(\mathbf{x})}{\psi(\mathbf{x})} \psi(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_{\psi}\left[\mathbb{1}_{(H(\mathbf{Y} \geq \gamma)} \frac{g(\mathbf{Y})}{\psi(\mathbf{Y})}\right] \tag{2}$$

where $\mathbf{Y}$ is a random vector with pdf $\psi(\mathbf{x})$.

Following [7, 9], we consider a particular case where $g(\mathbf{x})$ is standard Gaussian,[1] i.e., $g(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We define $\psi(\mathbf{x}) = g^{\boldsymbol{\theta}}(\mathbf{x})$ with $g^{\boldsymbol{\theta}}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I})$ parameterized by its mean $\boldsymbol{\theta} \in \mathbb{R}^d$. Changing the density from $g(\mathbf{x})$ to $g^{\boldsymbol{\theta}}(\mathbf{x})$ implies the translation of mean-shift, then $g^{\boldsymbol{\theta}}(\mathbf{x}) = g(\mathbf{x}-\boldsymbol{\theta})$. Therefore, the ISMC estimator of (2) is given by

$$\hat{p}_{\text{fail}}^{\text{IS}} = \frac{1}{N} \sum_{j=1}^{N} J(\mathbf{X}_j) \tag{3}$$

where $J(\mathbf{X}) = \mathbb{1}_{(H(\mathbf{X}+\boldsymbol{\theta}) \geq \gamma)}\, e^{-\boldsymbol{\theta}\mathbf{X} - \frac{|\boldsymbol{\theta}|^2}{2}}$ and $\mathbf{X}_j$'s are N independent and identically distributed random vectors with the density $g(\mathbf{x})$, and thus $J(\mathbf{X}_j)$'s are N independent random numbers. For details, we refer to [7, Section 3.5.2].

It is easy to prove that $\hat{p}_{\text{fail}}^{\text{IS}}$ is an unbiased estimator for p_{fail} and its variance is given by

$$\text{Var}_{\psi}(\hat{p}_{\text{fail}}^{\text{IS}}) = \frac{\sigma_{IS}^2}{N} \tag{4}$$

[1] The approach can be extended for other distributions assuming that the original input distributions can be transformed into a standard Gaussian distribution.

where

$$\sigma_{IS}^2 = \mathrm{Var}_\psi\,[J(\mathbf{X})] = \mathbb{E}_\psi\left[(J(\mathbf{X}))^2\right] - p_{\mathrm{fail}}^2 \tag{5}$$

Notice that the accuracy of the estimator $\hat{p}_{\mathrm{fail}}^{\mathrm{IS}}$ depends on the unknown mean-shift $\boldsymbol{\theta}$ which is estimated by minimizing the variance σ_{IS}^2. Only, $\mathbb{E}_\psi\left[(J(\mathbf{X}))^2\right]$ in Eq. (5) depends on $\boldsymbol{\theta}$. Thus, we find the mean-shift $\boldsymbol{\theta}$ by minimizing

$$v(\boldsymbol{\theta}) := \mathbb{E}_\psi\left[(J(\mathbf{X}))^2\right] = \mathbb{E}_g\left[\mathbb{1}_{(H(\mathbf{X})\geq\gamma)}\, e^{-\boldsymbol{\theta}\mathbf{X}+\frac{|\boldsymbol{\theta}|^2}{2}}\right] \tag{6}$$

For details we refer to [4, 6–9].

3 An Unbiased Hybrid Importance Sampling Monte Carlo Approach

In this section, we propose the UHISMC approach which is a hybrid of ISMC, surrogate models[2] and a stratification sampling technique [4]. Moreover, it provides an unbiased probability estimator.

In UHISMC, we split the ISMC technique into an exploration and an estimation phase. In the exploration phase, we substitute the circuit response by its surrogate model, directly. However, in the estimation phase, a surrogate model is used to stratified the sampling region, to find samples that contribute to each stratum by using an accept/reject criterion. At the end, the full response is used for the accepted samples. For details, see Sect. 3.2.

3.1 The Exploration Phase

Let $\widehat{H}(\mathbf{x})$ be a surrogate prediction for the circuit response $H(\mathbf{x})$ at some input $\mathbf{x}$. Then, we can approximate $v(\boldsymbol{\theta})$ in Eq. (6) by

$$\widehat{v}(\boldsymbol{\theta}) = \mathbb{E}_g\left[\mathbb{1}_{(\widehat{H}(\mathbf{X})\geq\gamma)}\, e^{-\boldsymbol{\theta}\mathbf{X}+\frac{|\boldsymbol{\theta}|^2}{2}}\right] \tag{7}$$

[2] Surrogate models mimic the complex behaviour of (circuit) simulation model. The study of surrogate models is out of scope of this paper. A review on surrogate models is given in [3].

The estimate $\widehat{\boldsymbol{\theta}}$ of the mean shift $\boldsymbol{\theta}$ can be computed with the Newton algorithm by solving the following optimization problem

$$\widehat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \hat{v}_m(\boldsymbol{\theta}) \tag{8}$$

with

$$\widehat{v}_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{(\widehat{H}(\mathbf{X}_j)\geq\gamma)}\, e^{-\boldsymbol{\theta}\mathbf{X}_j + \frac{|\boldsymbol{\theta}|^2}{2}}, \tag{9}$$

the MC approximation of the second moment $\widehat{v}(\boldsymbol{\theta})$.

Notice that there must be at least one $\mathbf{X}_j$ such that $\mathbb{1}_{(\widehat{H}(\mathbf{X}_j)\geq\gamma)} \neq 0$ to have $\widehat{v}_m \neq 0$. However, this condition may fail in a rare event context. To overcome this problem, a multilevel approach is used. We refer for more details to [4, 5, 9].

3.2 The Estimation Phase

In the estimation phase we estimate the failure probability p_{fail} with the known $\boldsymbol{\theta}$ (from the exploration phase). Moreover, we use a so called controlled stratified sampling (CSS) technique as proposed in [2].

We partition the input space $\mathbb{R}^d$ into I mutually exclusive and exhaustive regions $D_1, \ldots, D_I$ called strata. To use stratified sampling we must know how to sample in each D_i. To this end we use a surrogate model to find the strata $D_i = \{\mathbf{X}+\boldsymbol{\theta} : \hat{h}_{\rho_{i-1}} < \widehat{H}(\mathbf{X}+\boldsymbol{\theta}) \leq \hat{h}_{\rho_i}\}$ where $\widehat{H}(\mathbf{X}+\boldsymbol{\theta})$ is a surrogate prediction and $\hat{h}_{\rho_{i-1}}$ and $\hat{h}_{\rho_i}$ are quantiles of $\widehat{H}(\mathbf{X}+\boldsymbol{\theta})$ computed for $0 \leq \rho_{i-1} < \rho_i < 1$. Now let N be the total number of simulations from the simulator. We fix an allocation[3] $N_1, \ldots, N_I$ of positive integers with $\sum_{i=1}^{I} N_i = N$. For each stratum i, we generate many realizations[4] of the r.v. $\mathbf{X} \sim g(\mathbf{x})$ and accept N_i realizations $\left(\mathbf{X}_j^{(i)}\right)_{j=1,\ldots,N_i}$ among them such that the model predictions $\hat{H}\left(\mathbf{X}_j^{(i)} + \boldsymbol{\theta}\right)$ lie in the interval $(h_{\rho_{i-1}}, h_{\rho_j}]$. The true response $H\left(\mathbf{X}_j^{(i)} + \boldsymbol{\theta}\right)$ is computed for the accepted realizations. The conditional probability $P_i = \left\{\mathbb{E}_g\left[J(\mathbf{X}^{(i)})\right] : \mathbf{X}^{(i)} + \boldsymbol{\theta} \in D_i\right\}$ is estimated for each i:

$$\hat{P}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} J(\mathbf{X}_j^{(i)}) \tag{10a}$$

[3] An allocation means the partition of N.

[4] We keep sampling from $g(\mathbf{x})$ until we have N_i accepted samples.

with its conditional variance:

$$\mathrm{Var}_{\psi}(\hat{P}_i) = \frac{\mathrm{Var}\left(J(\mathbf{X}_j^{(i)})\right)}{N_i} =: \frac{\sigma_i^2}{N_i} \tag{10b}$$

Finally, the Importance Sampling Controlled Stratified (ISCS) probability estimator of p_{fail} is given by:

$$\hat{p}_{\text{fail}}^{\text{ISCS}} = \sum_{i=1}^{I} w_i \hat{P}_i \tag{11}$$

with variance

$$\mathrm{Var}_{\psi}(\hat{p}_{\text{fail}}^{\text{ISCS}}) = \sum_{i=1}^{I} w_i^2 \frac{\sigma_i^2}{N_i} \tag{12}$$

where $w_i = \mathbb{P}(\mathbf{X} + \boldsymbol{\theta} \in D_i)$. For an optimal mean-shift $\boldsymbol{\theta}$, the estimator $\hat{p}_{\text{fail}}^{\text{ISCS}}$ is unbiased and satisfies the central limit theorem [2].

From the above it should be clear that when using the controlled stratification approach one needs to consider the following three issues carefully:

- **The choice of the number I of strata**
  To the authors knowledge, there is no exact rule to choose the number I of strata. By applying rule of thumb we use $I = 10$ "equal probable" strata. The equal probable means that all w_i are equal for $i = 1, \ldots, I$.
- **The estimation of the allocation** $N_1, \ldots, N_I$
  For choosing the allocation we first need to find the allocation fractions q_i. Next the allocation can be estimated by the relation $N_i = \lceil q_i N \rceil$ where $\lceil x \rceil$ is the smallest positive integer greater than or equal to x. Substituting this in Eq. (12), we get

$$\mathrm{Var}_{\psi}(\hat{p}_{\text{fail}}^{\text{ISCS}}) \approx \frac{1}{N} \sum_{i=1}^{I} w_i^2 \frac{\sigma_i^2}{q_i} \tag{13}$$

The fraction q_i can be found by solving the following minimization problem

$$\min_{q_i} \left\{ \sum_{i=1}^{I} \frac{w_i^2}{q_i} \sigma_i^2 \right\}, \text{ with the constraints } 0 < q_i < 1, \sum_{i=1}^{I} q_i = 1 \tag{14}$$

which gives the unique solution

$$q_i = \frac{w_i \sigma_i}{\sum_{k=1}^{I} w_k \sigma_k}, \quad i = 1, \ldots, I \tag{15}$$

In practice, the σ_i are unknown, so the optimal fractions q_i are not directly applicable. Nevertheless, it is often beneficial to use pilot runs for getting the estimate $\hat{\sigma}_i$ of σ_i. The estimated allocation fractions $\hat{q}_i$ can then be used to allocate samples to strata in a second (typically larger) set of runs. For details, we refer to [2].

4 Numerical Results

Here we present the results of two realistic circuits.[5] The first one is a VCO with 1500 stochastic input parameters and scalar response 'oscillation frequency' and the second one is a memory cell with 2096 stochastic input parameters and scalar response 'read delay'.

Note that to compare the empirical results of ISMC and HISMC methods, we need a reference probability[6] $p_{\text{fail}}^{\text{ref}}$. A simple estimation $\hat{p}_{\text{fail}}^{\text{ref}}$ of $p_{\text{fail}}^{\text{ref}}$ can be found by using ISMC estimator Eq. (3) with a small (say less than 1%) Coefficient of Variation

$$\text{CV} = z_{\alpha/2} \frac{\sqrt{\text{Var}(\hat{p}_{\text{fail}}^{\text{M}})}}{\hat{p}_{\text{fail}}^{\text{M}}}$$

where $\hat{p}_{\text{fail}}^{\text{M}}$ denotes the probability estimate of p_{fail} computed with the method M and $z_{\alpha/2} = 1.96$ for the 95% confidence interval of the estimated probability [9].

4.1 The VCO

For VCO our goal is to estimate the probability of oscillation frequency to be larger than the given threshold $\gamma = 1900$. The reference probability $\hat{p}_{\text{fail}}^{\text{ref}} = 1.10 \times 10^{-10}$ which is estimated by using the ISMC technique with $\text{CV} = 0.99\%$.

The left and right plots in Fig. 2 show the probability distributions from $N_{\text{rep}} = 100$ experiments (repetitions) of ISMC and HISMC method, respectively. The

[5] These circuits are provided by engineering team of Mentor Graphics. For more details see [4, 9].

[6] In theory, the reference probability is the true failure probability to which the simulated results to be compared. However, in practice, we do not know the true probability. Thus, an approximation (with a high accuracy) of the true probability is used.

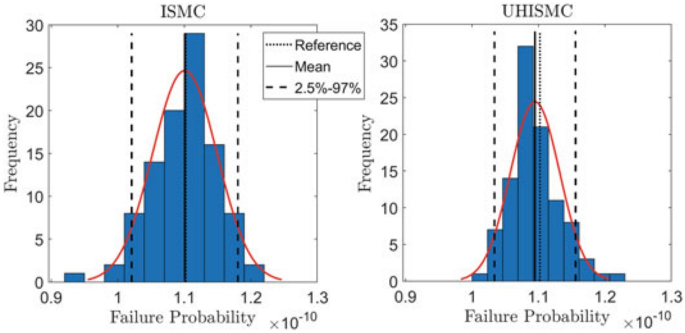


Fig. 2 VCO: Empirical probability distributions of ISMC and UHISMC from $N_{\text{rep}} = 100$ experiments with mean 1.10×10^{-10} and 1.09×10^{-10}, respectively

Table 1 VCO: ISMC versus HISMC probability estimation

Method	$\hat{p}_{\text{fail}}$	CV(%)	MSE	#Runs
ISMC	$\mathbf{1.10 \times 10^{-10}}$	**8.14**	2.34×10^{-23}	12,000
UHISMC	$\mathbf{1.09 \times 10^{-10}}$	**6.16**	1.47×10^{-23}	2978

vertical solid ‘black’ line in the center represents the mean of the distribution. The dotted line in the center is the reference probability $\hat{p}_{\text{fail}}^{\text{ref}}$. The two dashed lines around the center represent the 95% confidence interval. Notice that the distribution mean is equal (for ISMC) or close (for HISMC) to the reference probability.

The mean results of the methods are given in Table 1. ‘$\hat{p}_{\text{fail}}$’, ‘CV’ and ‘#Runs’ are the estimated probability, coefficient of variation and number of runs, respectively. ‘MSE’ is the mean squared error $\left(\sum_{i=1}^{100} (\hat{p}_{\text{fail}i} - \hat{p}_{\text{fail}}^{\text{ref}})^2\right)$. The efficiency of the UHISMC method with respect to the ISMC is computed as

$$\text{Eff(ISMC, UHISMC)} = \frac{(\text{MSE} \times \#\text{Runs}) \text{ in ISMC}}{(\text{MSE} \times \#\text{Runs}) \text{ in UHISMC}} = \frac{2.34 \times 10^{-23} \times 12000}{1.47 \times 10^{-23} \times 2978} \approx 6$$

So, ISMC requires about 6 times more simulations than UHISMC to achieve the same accuracy. Thus, UHISMC is preferred.

4.2 *The Memory Cell*

Our goal for the memory cell is to estimate the probability of the oscillation frequency to be larger than the given thresholds $\gamma = 902$. The reference probability $\hat{p}_{\text{fail}}^{\text{ref}}$ is 6.01×10^{-8} which is estimated by using the ISMC technique with CV $=$ 0.99%.

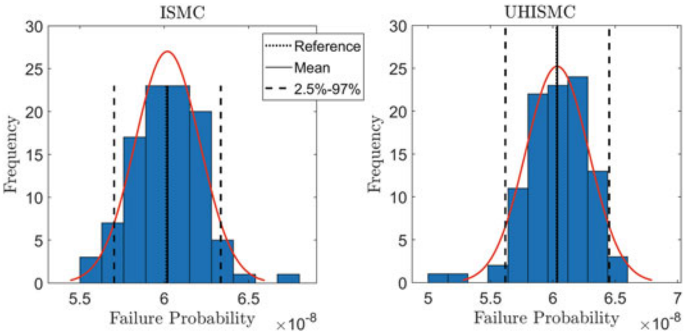


Fig. 3 Memory Cell: Empirical probability distributions of ISMC and UHISMC from $N_{\text{rep}} = 100$ experiments with mean 6.02×10^{-8} and 6.03×10^{-8}, respectively

Table 2 Memory Cell: ISMC versus HISMC probability estimation

Method	$\hat{p}_{\text{fail}}$	CV(%)	MSE	#Runs
ISMC	$\mathbf{6.02 \times 10^{-8}}$	**7.85**	3.65×10^{-18}	11,000
UHISMC	$\mathbf{6.03 \times 10^{-8}}$	**6.90**	6.5×10^{-18}	3347

Similar to the VCO, Fig. 3 shows the probability distributions and Table 2 represents the mean results with $N_{\text{rep}} = 100$ experiments of ISMC and UHISMC for the memory cell. It can be seen that the probabilities from both the methods are close to the reference probability. The efficiency of the UHISMC method with respect to the ISMC is computed as

$$\text{Eff(ISMC, UHISMC)} = \frac{3.65 \times 10^{-18} \times 11000}{6.5 \times 10^{-18} \times 3347} \approx 2$$

5 Conclusion

It is clear from both the examples above that UHISMC is more efficient than ISMC. Moreover, UHISMC provides an unbiased estimation of the probability. Thus, we prefer UHISMC over ISMC. For future work we will try to compare our UHISMC and UHISMC as in [9] in terms of efficiency and robustness.

Acknowledgement The first author is grateful to the financial support from the Marie Curie Action.

References

1. Bucklew, J.A.: Introduction to Rare Event Simulation. Springer Series in Statistics. Springer, New York (2004)
2. Cannamela, C., Garnier, J., Iooss, B.: Controlled stratification for quantile estimation. Ann. Appl. Stat. **2**(4), 1554–1580 (2008)
3. Chen, V.C.P., Tsui, K.-L., Barton, R.R., Meckesheimer, M.: A review on design modeling and applications of computer experiments. IIE Trans. **38**(4), 273–291 (2006)
4. Ciampolini, L., Lafont, J.-C., Drissi, F. T., Morin, J.-P., Turgis, D., Jonsson, X., Descléves, C., Nguyen, J.: Efficient yield estimation through generalized importance sampling with application to NBL-assisted SRAM bitcells. In: Proceedings of the 35th International Conference on Computer-Aided Design (2016). Article No. 89
5. Homem-de-Mello, T., Rubinstein, R.Y.: Estimation of rare event probabilities using cross-entropy. In: Yücesan, E., Chen, C.-H., Snowdon, J.L., Charnes, J.M. (eds.) Proceedings of the 2002 Winter Simulation Conference (2002)
6. Jourdain, B., Lelong, J.: Robust adaptive importance sampling for normal random vectors. Ann. Appl. Probab. **19**(5), 1687–1718 (2009)
7. Tyagi, A.K.: Speeding up rare-event simulations in electronic circuit design by using surrogate models. PhD thesis, Department of Mathematics and Computer Science, 10 (2018). Proefschrift
8. Tyagi, A.K., Jonsson, X., Beelen, T.G.J., Schilders, W.H.A.: Speeding up rare event simulations using Kriging models. In: Proceedings of IEEE 21st Workshop on Signal and Power Integrity (SPI). IEEE, New York (2017).
9. Tyagi, A. K., Jonsson, X., Beelen, T.G.J., Schilders, W.H.A.: Hybrid importance sampling monte carlo approach for yield estimation in circuit design. J. Math. Ind. **8**(1), 11 (2018)

Shape Optimization of a PM Synchronous Machine Under Probabilistic Constraints

Piotr Putek, Andreas Bartel, E. Jan W. ter Maten, and Michael Günther

Abstract This paper proposes a robust and reliability-based shape optimization method to find the optimal design of a permanent magnet (PM) synchronous machine. Specifically, design of rotor poles and stator teeth is subjected to the shape optimization under manufacturing tolerances/imperfections and probabilistic constraints. In a forward problem, certain parameters are assumed to be random. This affects also a shape optimization problem, which is formulated in terms of a tracking-type robust cost functional. The latter is equipped with probabilistic constraints in order to attain a new, desired, robust design. The topological gradient is evaluated using the Topological Asymptotic Expansion Method, to which we apply a Stochastic Collocation Method. In the end, to illustrate our approach, we provide the optimization results for a 2D model of the PM machine.

1 Introduction

Following with the rapid development of the performance of PM synchronous machines, they have been widely applied in various fields such as industrial automation, household applications and electric vehicles. In particular, due to several advantages including high efficiency in the whole working region and a good dynamic performance, they have become the main type of a driving motor for electric vehicles [8].

Compared with the conventional surface-mounted PM synchronous machine, the ECPSM,[1] on the one hand, has a wider speed range due to the field-weakening

[1] The Electrically Controlled Permanent Magnet Excited Synchronous Machine was investigated within the scientific project under grant no. N510 508040 (2011–2013), Poland.

P. Putek (✉) · A. Bartel · E. J. W. ter Maten · M. Günther
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: putek@math.uni-wuppertal.de; bartel@math.uni-wuppertal.de;
termaten@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_23

capability and better output torque characteristics [11]. On the other hand, it typically suffers from a considerably high level of a cogging torque (CT) because of its specific structure and a high air gap flux density. This results in the undesired torque and speed ripples as well as acoustic noise and vibrations, that influence its further application.

Yet, as a result of manufacturing processes, a design of an electric machine is strongly affected by the uncertainties in both the geometrical and material parameters [14]. Thus, to provide reliable simulations, a mathematical model with random input data needs to be considered [13]. This implies the use of the reliability analysis and the robust framework for a design assessment. The former allows for reducing the risk of a failure in an operating device and the latter results in minimizing the variability of the output performance functions.

In this paper, we formulate the shape optimization in terms of both concepts. Therefore, we combine the reliability-based and the robust approach for a re-design of the ECPSM in order to obtain a new topology which, meets both considered criteria.

2 Forward Problem with Random Input Parameters

A mathematical model of the ECPSM [13] can be described by a quasilinear curl-curl equation on a spatial domain $D \subset \mathbb{R}^2$ (with Lipschitz boundaries $\partial D \in C^2$). The domain is composed of four regions: $D = D_{\text{rot}} \cup D_{\text{sta}} \cup D_{\text{air}} \cup D_{\text{PM}}$ for rotor, stator, air and permanent magnet region, respectively. Accordingly, the deterministic reluctivity $\nu^{\text{det}} : D \times \mathbb{R}_0^+ \to \mathbb{R}^+$ models the different types of soft magnetic material as $\nu^{\text{det}} = \nu^{\text{det}}(\boldsymbol{x}, |B|^2)$ for spatial location $\boldsymbol{x}$ and magnitude of the magnetic flux density $|B|^2$. The reluctivity is monotone, bounded and differentiable with a well-defined derivative satisfying $\alpha \leq \partial_B \nu^{\text{det}}(\cdot, B^2) \leq \gamma$, cf. [2]. Furthermore, we have the magnetization as $\mathbf{M} = b_r^{\text{det}} \mathbf{T}$ with remanence flux density b_r^{det} and magnetization direction $\mathbf{T}$.

All five components, the reluctivities and the remanence flux density, suffer from uncertainty, cf. [10]. Thus, we model these $Q = 5$ quantities via random input parameters as follows. Let $\xi_1, \ldots, \xi_5$ be independent, normally distributed random variables. First, we transform random variables ξ_i to a bounded domain:

$$p_i(\xi_i) := \delta_i \cdot \Upsilon(\xi_i), \quad \text{for } i = 1, \ldots, 5, \quad \Upsilon(\cdot) := \arctan(\cdot)$$

and with fixed perturbation magnitudes δ_i. This gives a random vector $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\xi}) \in \mathbb{R}^5$ with image space $\Gamma_\rho \subset \mathbb{R}^5$ and respective probability density on Γ_ρ denoted by $\boldsymbol{\rho}$. Then, the related probability space, which is used for numerical computations, reads: $(\Gamma_\rho, \mathcal{B}^5, \rho \mathrm{d}\boldsymbol{p})$ with the 5-dimensional Borel space $\mathcal{B}^5$ and a probability measure $\rho \mathrm{d}\boldsymbol{p}$. Using $\boldsymbol{p}$, we model the random dependent reluctivities as

$$\nu(\boldsymbol{x},\ \boldsymbol{p},\ B^2) = \begin{cases} \nu_{\mathrm{Fe}}^{\mathrm{sta}}(|B|^2) \cdot [1 + p_1(\xi_1)] \, \text{for } \ \boldsymbol{x} \in D_{\mathrm{rot}}\,, \\ \nu_{\mathrm{Fe}}^{\mathrm{rot}}(|B|^2) \cdot [1 + p_2(\xi_2)] \, \text{for } \ \boldsymbol{x} \in D_{\mathrm{sta}}\,, \\ \nu_{\mathrm{air}} \cdot [1 + p_3(\xi_3)] \qquad\quad \text{for } \ \boldsymbol{x} \in D_{\mathrm{air}}\,, \\ \nu_{\mathrm{PM}} \ \cdot [1 + p_4(\xi_4)] \qquad\quad \text{for } \ \boldsymbol{x} \in D_{\mathrm{PM}}\,, \end{cases} \tag{1}$$

and the remanence flux density as:

$$b_r(\boldsymbol{p}) = b_r^{\mathrm{det}} \cdot [1 + p_5(\xi_5)]). \tag{2}$$

We remark, that our probabilistic model inherits the above mentioned properties of the reluctivities (monotonicity in $|B|^2$ is not touched, boundedness is due to the transformation of the normally distributed variables).

For the magnetic vector potential $\mathbf{A} = (0, 0, u)$ depending on $\boldsymbol{\chi} := (\boldsymbol{x}, \boldsymbol{p}) \in D \times \Gamma_\rho$, we define the space for u:

$$V_\rho = \left\{ w = w(\boldsymbol{x}, \boldsymbol{p}) \,|\, w(\boldsymbol{x}, \cdot) \in H_0^1(D),\ \ w(\cdot, \boldsymbol{p}) \in L^2(\Gamma_\rho) \right\}.$$

Then, the variational form of the curl-curl equation reads: find $u \in V_\rho$ such that

$$\int_{\Gamma_\rho} \Big(\nu(\boldsymbol{x},\ |\nabla u(\boldsymbol{\chi})|^2) \nabla u(\boldsymbol{\chi}),\ \nabla \varphi(\boldsymbol{\chi}) \Big) \rho \mathrm{d}\boldsymbol{p} = \int_{\Gamma_\rho} \Big(f,\ \varphi(\boldsymbol{\chi}) \Big) \rho \mathrm{d}\boldsymbol{p} \tag{3}$$

for all test functions $\varphi \in V_\rho$ and $(\cdot, \cdot)$ being the standard L^2 inner product $\big(w,\ w\big) := \int_{\mathrm{D}} |w|^2 dx$. Thus, we seek a numerical approximation to the exact solution (3) in V_ρ. This will be constructed by using the Polynomial Chaos Expansion (PCE) technique. Moreover, the function f is assumed to be square integrable w.r.t. Γ_ρ and it is defined by $f(\boldsymbol{\chi}) = J_i(\boldsymbol{x}) + \nu_{\mathrm{PM}}(\boldsymbol{\chi}) \nabla \cdot \mathbf{M}(\boldsymbol{\chi})$, where J_i, $\mathbf{M}$ and ν_{PM} denote the current density, the magnetization and the reluctivity of the PM, respectively.

3 Reliability and Robustness Analysis

To assess the reliability and robustness of the electric machine design w.r.t uncertain input parameters arisen from, e.g., manufacturability, we explore the stochastic collocation method (SCM) compound with the polynomial chaos expansion (PCE).

3.1 The Pseudo-Spectral Approach

For the uncertainty quantification (UQ), we used the spectral approach by [17], where the quantities of interests to be calculated are the unknown expansion coefficient of the polynomial chaos. For this reason, we consider the probabilistic Hilbert space of square integrable random functions

$$L^2_\rho(\Gamma_\rho) = \{Y(\boldsymbol{p}) : \mathit{\Gamma}_\rho \to \mathbb{R} \mid \mathbb{E}[Y(\boldsymbol{p})^2] < \infty\} \text{ with } \mathbb{E}[Z(\boldsymbol{p}(\boldsymbol{\xi}))] := \int_{\mathit{\Gamma}_\rho} Z(\mathbf{p})\,\rho(\boldsymbol{p})\,\mathrm{d}\boldsymbol{p}$$

and equipped with inner product and corresponding norm:

$$\langle Y(\boldsymbol{p}), Z(\boldsymbol{p})\rangle_\rho := \mathbb{E}\left(Y(\boldsymbol{p})Z(\boldsymbol{p})\right), \qquad \|Y(\boldsymbol{p})\|^2_{L^2_\rho} = \langle Y(\boldsymbol{p}), Y(\boldsymbol{p})\rangle_\rho .$$

Consequently, the variance of $Y(\boldsymbol{p}) \in L^2_\rho(\mathit{\Gamma}_\rho)$ read as

$$\mathrm{Var}[Y(\boldsymbol{p})] := \mathbb{E}[Y(\boldsymbol{p})^2] - \mathbb{E}[Y(\boldsymbol{p})]^2. \tag{4}$$

Now, we introduce the truncated PCE [18] for $u \in L^2_\rho(\mathit{\Gamma}_\rho)$:

$$\widetilde{u}(\boldsymbol{x}, \boldsymbol{p}) \doteq \sum_{m=0}^{M} u_m(\boldsymbol{x})\, \mathit{\Phi}_m(\boldsymbol{p}) \tag{5}$$

with a respective basis of multivariate polynomials $\mathit{\Phi}_m : \mathbb{R}^5 \to \mathbb{R}$, which correspond to the distribution of input random parameters, that is, the uniform distribution implies the Legendre polynomials, while the Hermite polynomials refers to the Gaussian-type PDF, respectively.

In Eq. (5), u_m are *a priori* unknown coefficient functions to be determined by using projections of provided solutions at quadrature points on the basis polynomials as

$$u_m(\boldsymbol{x}) = \langle \widetilde{u}(\boldsymbol{x}, \boldsymbol{p}), \mathit{\Phi}_m(\boldsymbol{p})\rangle_\rho . \tag{6}$$

Next, to approximate the probabilistic integrals of (6), we applied the Stroud formulas with a constant weight function [17] in the form

$$u_m(\boldsymbol{x}) \doteq \sum_{k=1}^{K} w_k\, \widetilde{u}\left(\boldsymbol{x}, \boldsymbol{p}^{(k)}\right) \mathit{\Phi}_m\left(\boldsymbol{p}^{(k)}\right), \tag{7}$$

in which the w_k and $\boldsymbol{p}^{(k)}$ are deterministic quadrature weights and points ($k = 1, \ldots, K$). This type of quadrature methods is exact for multivariate polynomials up to the degree d_{PC}, e.g., for $d_{\mathrm{PC}} = 3$ we need $K = 2Q = 10$, for $d_{\mathrm{PC}} = 5$

one needs $K = 2Q^2 + 1 = 51$. They seem to be highly efficient in particular for large numbers of parameters [13, 18], but their accuracy, unfortunately, is fixed and cannot be improved. Finally, the statistical moments are approximated by (including quadrature)

$$\mathbb{E}\left[\widetilde{u}\left(\boldsymbol{x}, \boldsymbol{p}\right)\right] \doteq u_0(\boldsymbol{x}), \quad \mathrm{Var}\left[\widetilde{u}\left(\boldsymbol{x}, \boldsymbol{p}\right)\right] \doteq \sum_{m=1}^{M} |u_m(\boldsymbol{x})|^2, \tag{8}$$

using $\Phi_0 = 1$ [17]. Also, other quantities such as the local sensitivity and the variance-based global sensitivity can easily be calculated based on (5) see [16, 18].

3.2 Reliability Index Approach

We use the First-Order Reliability Method (FORM) [19] to evaluate the reliability criteria. Within this methodology, a transformation $\boldsymbol{r} = T(\boldsymbol{p})$ from the physical variables $\boldsymbol{p}$ to the normalized variables $\boldsymbol{r}$ is required in order to calculate the reliability index[5]. Thus, $\boldsymbol{r}$ is given by

$$r_q = \frac{p_q - \mu_q}{\sigma_q}, \qquad q = 1, \ldots, 5$$

with the particular mean and the standard deviation denoted by μ_q and σ_q. Now, the reliability index β is found by solving the constrained optimization problem (excluding a trivial case $\boldsymbol{r} = 0$)

$$\begin{aligned} \beta^* = \min_{\boldsymbol{r}} \ & \beta(\boldsymbol{r}) = \sqrt{(\boldsymbol{r}^\top \boldsymbol{r})} \\ \text{s.t.} \quad & g(\boldsymbol{r}) = 0, \end{aligned} \tag{9}$$

where $g(\boldsymbol{r})$ is a limit state function. The failure probability is approximated by $\mathbb{P}[g(\boldsymbol{r}) \leq 0] \approx \Phi^{\mathrm{NT}}(-\beta^*)$, where Φ^{NT} is the standard normal cumulative distribution function. Consequently, the resulting normalized vector $\boldsymbol{r}$ is used to modify the random vector $\boldsymbol{p}$, which influences the UQ of model (3).

4 Shape Optimization Problem

We consider a random-dependent cost functional, defined in terms of the magnetic energy [11] as

$$F(\Omega, u) := \frac{1}{2} \int_D \nu(\boldsymbol{\chi}, |\nabla u|^2) |\nabla u|^2 d\boldsymbol{x} \tag{10}$$

with $\Omega = D$, but with different decomposition into the respective materials: $\Omega = D^{\Omega}_{\text{rot}} \cup D^{\Omega}_{\text{sta}} \cup D^{\Omega}_{\text{air}} \cup D^{\Omega}_{\text{PM}}$ and $u := u_{\Omega}(\boldsymbol{x}, \boldsymbol{p})$ being the solution of (3) given the material decomposition Ω (in $\boldsymbol{x}$ and $\boldsymbol{p}$).

Let $\mathcal{M}$ denote the set of admissible material decompositions. Correspondingly, the shape optimization problem is given by

$$\inf_{\Omega \in \mathcal{M}} \mathcal{E}\left[F\left(\Omega, u\right)\right] := \inf_{\Omega \in \mathcal{M}} \left(\mathbb{E}\left[F\left(\Omega, u\right)\right] + \iota\sqrt{\text{Var}\big[F\left(\Omega, u\right)\big]}\right) \tag{11a}$$

$$\text{s.t.} \quad \beta(\cdot) \geq \beta_t, \tag{11b}$$

$$u \text{ satisfies } \quad (3), \tag{11c}$$

with the prescribed parameters $\iota = 6$ and $\beta_t = 3.8$, where the probabilistic constraint (11b) was expressed by an equivalent form using the reliability index approach (9). For the solution of the problem (11a)–(11c), we construct an iterative scheme [13], which requires developing a topological derivative.

To minimize the functional $F(\cdot)$, the topological derivative is employed. Therefore, it is required to create infinitely small holes at certain points $\boldsymbol{x}_0 \in \Omega$, for which the topological derivative (or sensitivity) is given by $G(\boldsymbol{x}_0, \boldsymbol{p}) < 0$. To this end, let $\epsilon > 0$ be the perturbation parameter and consider any $\boldsymbol{x}_0 \in \Omega$, then we define $\Omega_\epsilon := \Omega \backslash D^{\epsilon}_{\boldsymbol{x}_0}$, where $D^{\epsilon}_{\boldsymbol{x}_0} := \{\boldsymbol{y} \in \mathbb{R}^2 \mid |\boldsymbol{y} - \boldsymbol{x}_0| \leq \epsilon\}$ (disc with center $\boldsymbol{x}_0$ and radius ϵ). The topological asymptotic expansion can be found for broad class of boundary value problems including our problem, see, e.g., [1, 7, 9]. It takes the form [3, 15]

$$F\left(\Omega_\epsilon, \cdot\right) - F\left(\Omega, \cdot\right) = f^{\text{ex}}\left(\epsilon\right) G(\boldsymbol{x}_0) + o(f(\epsilon))$$

for some function f^{ex} satisfying $\lim_{\epsilon \to 0} f^{\text{ex}}(\epsilon) = 0$, $f^{\text{ex}}(\epsilon) > 0$. Notice, F is now used for a domain $\Omega_\epsilon \subset D$. For the technicalities, we refer to e.g. [15].

Topological Derivative for the Fixed Cubature Point In practice, the probabilistic integral (10) is approximated by a quadrature scheme compound with the PCE. As a result, after generating quadrature nodes $\boldsymbol{p}^{(k)}$ (in Γ_ρ) and weights $w^{(k)}$ (with $k = 1, \ldots, K$), the function (10) is calculated at nodes $F(\Omega, u(\boldsymbol{x}, \boldsymbol{p}^{(k)})) =: F^{(k)}(\cdot)$ as well as $u^{(k)} := u^{(k)}(\boldsymbol{x}) := u(\boldsymbol{x},\ \boldsymbol{p}^{(k)})$ and so on.

To accelerate the topological derivative calculation, a dual problem for $F^{(k)}(\Omega, u^{(k)})$ with an adjoint variable $\lambda^{(k)}$ is defined as

$$a\left(\lambda^{(k)}, \psi\right) = (dF^{(k)}\left[\Omega, u^{(k)}\right], \psi),$$

in which the bilinear form $a\left(\lambda^{(k)}, \psi\right)$ is given by [2]

$$a\left(\lambda^{(k)}, \psi\right) = \left(\nu^{(k)}\left(|\nabla u|^2\right) \nabla\lambda^{(k)}, \nabla\psi\right) + 2\left(\nu^{(k)\prime}\left(|\nabla u^{(k)}|^2\right) \nabla\lambda^{(k)} \cdot \nabla u^{(k)} \nabla u^{(k)}, \nabla\psi\right).$$

Given the result in [4], completed by the respective magnetization term using the local expansion method[1, 12] $\nu_{\text{PM}}\mathbf{M}$. The approximation of the topological derivative is calculated at the k-th quadrature point ($k = 1, \dots K$) by

$$G^{(k)}(\boldsymbol{x}, \boldsymbol{p}^{(k)}) \doteq \begin{cases} \nabla u^{(k)^\top}(\boldsymbol{x})\mathcal{P}\big(\vartheta_a, \nu_0^{(k)}, \nu_1^{(k)}, \nabla u^{(k)}(\boldsymbol{x})\big)\nabla\lambda^{(k)}(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in (D_{\text{rot}}^{\Omega} \cup D_{\text{sta}}^{\Omega}), \\ 2\pi\, b_{r_2}^{(k)} \dfrac{b_{r_0} - b_{r_2}^{(k)}}{b_{r_0} + b_{r_2}^{(k)}} \mathbf{T}(\boldsymbol{x}) \cdot \nabla\lambda^{(k)}(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in D_{\text{PM}}^{\Omega} \end{cases} \tag{12}$$

with adjoint variable λ, and polarization matrix $\mathcal{P}(\cdot)$ for the unit disc ϑ_a, [4]. Here, the material parameters with the following subscripts $\square_0$, $\square_1$ and $\square_2$ are related to the air, iron and PM domain, respectively. For the sake of simplicity, a further contribution to G (in $D_{\text{rot}} \cup D_{\text{sta}}$) related to the variation of the state and adjoint variable has been neglected.

Topological Derivative for the Robust Functional Furthermore, when the SCM with the PCE is involved in the optimization procedure, we can use (5) to represent the functional $F^{(k)}$ as in (10) and also the topological derivative $G^{(k)}$ as in (12) as the truncated response surface models

$$\widetilde{F}\,(\boldsymbol{x}, \mathbf{p}) \doteq \sum_{m=0}^{M} F_m\,(\boldsymbol{x})\,\Phi_m\,(\mathbf{p})\,, \qquad \widetilde{G}\,(\boldsymbol{x}, \mathbf{p}) \doteq \sum_{m=0}^{M} G_m\,(\boldsymbol{x})\,\Phi_m\,(\mathbf{p})\,. \tag{13}$$

Next, to obtain the unknown expansion coefficient $F_m\,(\boldsymbol{x})$ and $G_m\,(\boldsymbol{x})$, the provided solutions $F^{(k)}$ and $G^{(k)}$ at the corresponding quadrature point $k = 1, \dots, K$ is projected into the polynomial basis using (7). Then, the *robust* topological derivative of the expectation value and of the variance are given by [11]

$$d\,\mathbb{E}\,[F\,(\Omega, \widetilde{u}(\mathbf{p}))] = G_0(\boldsymbol{x}), \qquad d\,\text{Var}\,[F\,(\Omega, \widetilde{u}(\mathbf{p}))] = \sum_{m=1}^{M} 2F_m\,(\boldsymbol{x})\,G_m\,(\boldsymbol{x})\,. \tag{14}$$

Then, the topological derivative of the robust functional (11a) reads as

$$d\,\mathcal{E}\,[F\,(\Omega, \widetilde{u}(\mathbf{p}))] = d\,\mathbb{E}\,[F\,(\cdot)] + 0.5\,\iota\left(\sqrt{\text{Var}\,[F\,(\cdot)]}\right)^{-1} d\,\text{Var}\,[F\,(\cdot)]\,, \tag{15}$$

where the mean and the variance are approximated by

$$\mathbb{E}\,[F\,(\Omega, \widetilde{u}(\mathbf{p}))] = F_0(\boldsymbol{x}), \qquad \text{Var}\,[F\,(\Omega, \widetilde{u}(\mathbf{p}))] = \sum_{m=1}^{M} F_m^2\,(\boldsymbol{x})\,. \tag{16}$$

Optimization Procedure for the Reliability Index To solve the constrained optimization problem (9), we use the iteration procedure ($i = 1, \dots Q$), in which a design point is defined as $\boldsymbol{r} = \beta \cdot \boldsymbol{\alpha}$. Therein, the normal vector to the function g, i.e, the gradient, takes the form

$$\alpha_i = -\partial_{r_i} g\,(\beta\,\boldsymbol{\alpha}) \cdot \left(\sum_{j=1}^{Q} [\partial_{r_i} g\,(\beta\,\boldsymbol{\alpha})]^2 \right)^{-1/2} .$$

Finally, the reliability index is defined as $g(u(\beta\,\alpha_1,\ \dots,\ \beta\,\alpha_Q)) = 0$.

Implementation Remarks In practical implementation, first, we expand a function $g(\cdot)$, e.g., the air-gap magnetic flux density (MFD), in the form of a surface model using (5). Secondly, the global sensitivity analysis [16] can be used to identify the most influential input parameters w.r.t. the variation of the function g. Then, we solve the optimization problem (9) to find the reliability index β. Next, based on the local sensitivity analysis [18] (a sign of mean values derivative w.r.t. the particular random input variables), the mean values μ_q and the standard deviations σ_q of the particular random variables defined in (1) and (2) are modified using $\boldsymbol{r}$. In the end, this algorithm can be incorporated into the sensitivity-based optimization flow in order to find the robust and reliable low cogging design of the ECPSM. Likewise, to the reliability-based topology method [6], this procedure does not contain the nested robust and reliability loops.

5 Numerical Example and Discussion

We used the proposed method to design the rotor poles as well as the base tooth in the stator of the ECPSM machine at on-load condition, i.e., the model of the ECPSM was supplied with $I_n = 15[A],\ n = 1, 2, 3$. The 2D finite elements model, which consisted of a triangular mesh with the second order Lagrange polynomials, for the A-potential formulation was built in the COMSOL 3.5a. The sensitivity-based algorithm for the topology optimization was implemented in MATLAB 7.10. The area of rotor in the initial 2D model was divided into 360 and 480 voxels for the iron and the PM poles, while the base teeth was composed of 512 voxels. For simplicity, in our work we considered as the limit state function $f_{\text{PM}}(\boldsymbol{\chi}) = \int_D \nu_{\text{PM}}(\chi)\nabla \cdot \mathbf{M}(\boldsymbol{\chi})d\boldsymbol{x} =: g(\boldsymbol{\chi})$ with $\beta_t = 3.8$, which corresponds to the the (Gaussian) failure probability $P_f = 10^{-4}$. Here, scalings in (1) are $\delta_{1-4} = 0.15$ and $\delta_5 = 0.1$. Additionally, in the postprocessing stage, the ECPSM was analyzed in the magnetoquasistatic regime with $\sigma_{\text{FE}} = 11.2\,\text{M}$ [S/m] in order to investigate the core losses, shown on Fig. 3.

6 Conclusions

We demonstrated how to efficiently incorporate the reliability analysis into the robust framework to accomplish such a design of the ECPSM, depicted in Fig. 1, which is not only resistant to input variations, but also satisfies safety criteria. The mean value and the standard deviation of both the ET and the back EMF are depicted in Fig. 2. We could observe a decrease of statistical moments for both considered quantities by 5%/7% and 21%/23%, respectively. These results extend our outcomes provided in [13] (Fig. 3).

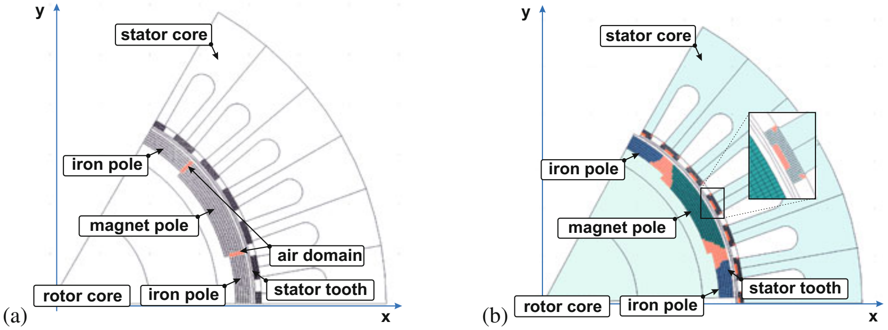


Fig. 1 ECPSM topology before and after optimization: (**a**) an initial model, (**b**) the optimized configuration found at 15^{th} iteration

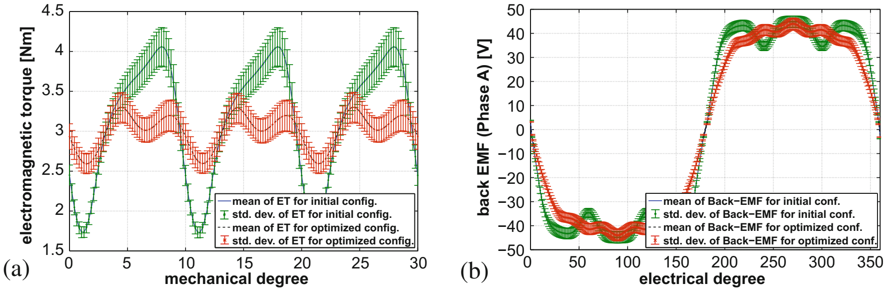


Fig. 2 Statistical moments for electromagnetic torque (ET) (**a**) and back electromotive force (EMF) (**b**)

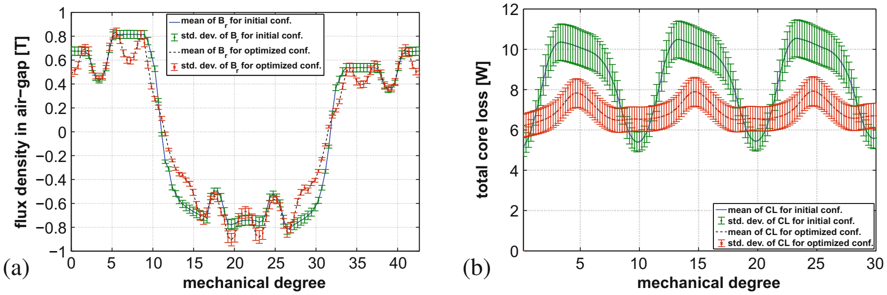


Fig. 3 Statistical moments for the MFD in the air-gap (**a**) and the core losses (CL) (**b**)

References

1. Céa, J., Garreau, S., Guillaume, P., Masmoudi, M.: The shape and topological optimizations connection. Comput. Methods Appl. Mech. Eng. **188**(4), 713–726 (2000)
2. Cimrák, I.: Material and shape derivative method for quasi-linear elliptic systems with applications in inverse electromagnetic interface problems. SIAM J. Numer. Anal. **50**(3), 1086–1110 (2012)

   Somasundaram, Chandirasekearan <C.Somasundaram@spi-global.com>; Sugavanam, Jaganathan <J.Sugavanam@spi-global.com>
3. Eschenauer, H.A., Kobelev, V.V., Schumacher, A.: Bubble method for topology and shape optimization of structures. Struct. Optim. **8**(1), 42–51 (1994)
4. Gangl, P., Amstutz, S., Langer, U.: Topology optimization of electric motor using topological derivative for nonlinear magnetostatics. IEEE Trans. Magn. **52**(3), 1–4 (2016)
5. Hasofer, A.M., Lind, N.: An exact and invariant first order reliability format. J. Eng. Mech. **100**, 01 (1974)
6. Kharmanda, G., Olhoff, N., Mohamed, A., Lemaire, M.: Reliability-based topology optimization. Struct. Multidisc. Optim. **26**(5), 295–307 (2004)
7. Masmoudi, M., Pommier, J., Samet, B.: The topological asymptotic expansion for the maxwell equations and some applications. Inverse Probl. **21**(2), 547–564 (2005)
8. Morimoto, S., Asano, Y., Kosaka, T., Enomoto, Y.: Recent technical trends in PMSM. In: 2014 IPEC-Hiroshima 2014 - ECCE ASIA, May, pp. 1997–2003 (2014)
9. Novotny, A.A., Feijóo, R.A., Taroco, E., Padra, C.: Topological sensitivity analysis. Comput. Methods Appl. Mech. Eng. **192**(7), 803–829 (2003)
10. Offermann, P., Hameyer, K.: Stochastic models for the evaluation of magnetisation faults. COMPEL **33**(1/2), 245–253 (2013)
11. Putek, P.: Nonlinear magnetoquasistatic interface problem in a permanent-magnet machine with stochastic partial differential equation constraints. Eng. Optim. **51**, 1–24 (2019)
12. Putek, P., Slodicka, M., Paplicki, P., Pałka, R.: Minimization of cogging torque in permanent magnet machines using the topological gradient and adjoint sensitivity in multi-objective design. Int. J. Appl. Electrom. **39**(1–4), 933–940 (2012)
13. Putek, P., ter Maten, E.J.W., Günther, M., Sykulski, J.K.: Variance-based robust optimization of a permanent magnet synchronous machine. IEEE Trans. Magn. **54**(3), 1–4 (2018)
14. Sergeant, P., Crevecoeur, G., Dupré, L., Van den Bossche, A.: Characterization and optimization of a permanent magnet synchronous machine. COMPEL **28**(2), 272–285 (2009)
15. Sokolowski, J., Zochowski, A.: Topological derivatives for elliptic problems. Inverse Probl. **15**(1), 123–134 (1999)

16. Sudret, B.: Global sensitivity analysis using polynomial chaos expansions. Reliab. Eng. Syst. Safe. **93**(7), 964–979 (2008)
17. Xiu, D.: Efficient collocational approach for parametric uncertainty analysis. Commun. Comput. Phys. **2**(2), 293–309 (2007)
18. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)
19. Zhao, Y.G., Ono, T.: A general procedure for first/second-order reliability method (FORM/SORM). Struct. Safe. **21**(2), 95–112 (1999)

Topology Shape and Parametric Design Optimization of Hall Effect Thrusters

Rtimi Youness, Maxime Bonnet, Frédéric Messine, and Carole Hénaux

Abstract In magnetics, topology optimization (TO) is a tool helping to find a suitable ferromagnetic material space distribution in order to meet magnetic specifications. TO is a tool that becomes very interesting when the designer looks for new and original structures. Herein, TO is used to design a Hall-effect thruster. But, the topological solutions are often not feasible. In order to remedy to this, shape optimization (SO) and parametric optimization (PO) are carried out on the topological solution. SO and PO take account of the manufacturing constraints as well as the non linearity of the ferromagnetic materials.

1 Introduction

Electric propulsion has become increasingly a solution to thrust satellites. Hall Effect thrusters (HET) is a leading technology in electric propulsion. The 3D structure of a Hall-effect thruster is provided in Fig. 1. Special magnetic field is applied near the HET exit plan to confine a bunch of electrons. In order to produce the thrust a flow of the propellant gas (Xenon in our case) is diffused via the channel. As one propellant atom collides with the confined electrons it get ionized and consequently accelerated by the electrical field between the anode and the cathode, details can be found in [1].

1.1 The Specifications and the Design Variables

In order to reduce the cpu time of the magnetostatic simulations, the axisymmetry of the HET is used in order to reduce its 3D structure to a 2D equivalent one, see

R. Youness (✉) · M. Bonnet · F. Messine · C. Hénaux
Laboratory on Plasma and Conversion of Energy , Toulouse Cedex, France
e-mail: rtimi@laplace.univ-tlse.fr; mbonnet@laplace.univ-tlse.fr; frederic.messine@laplace.univ-tlse.fr; henaux@laplace.univ-tlse.fr

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_24

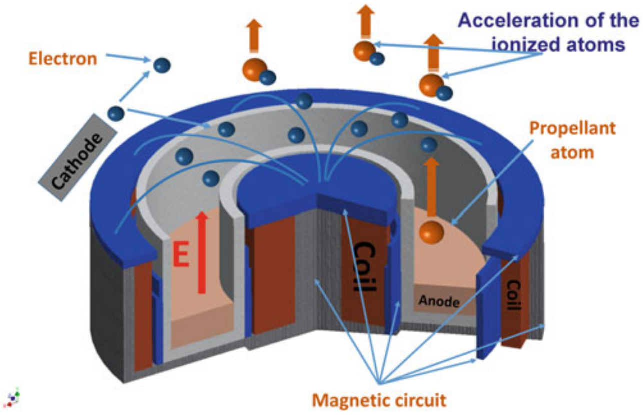


Fig. 1 The 3D structure of a Hall-effect thruster

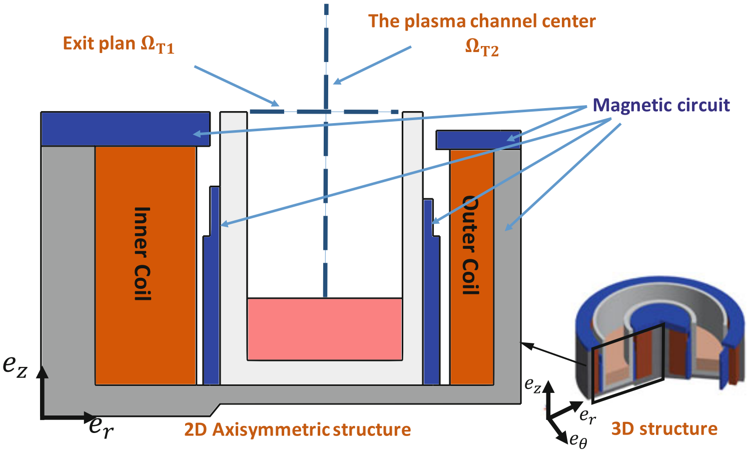


Fig. 2 The 3D model of the HET is reduced to a 2D axisymmetric one

Fig. 2. The efficiency of the HET tightly depends on the magnetic field along the plasma channel and at the exit plan (see Fig. 2). Indeed, the specifications that we are presumed to comply with are defined in two points:

- In the target region Ω_{T1} (see Fig. 2), the magnetic field should be radial and its strength should vary according to a Gaussian function given in Fig. 3.
- In the target region Ω_{T2} (see Fig. 2), the magnetic field should be radial.

The two main design variables that affect the magnetic field in the target regions Ω_{T1} and Ω_{T2} are: the magnetic circuit geometry as well as the coils current densities (see Fig. 2). The objective in this work is to develop a free initialization design method that shapes the magnetic circuit and finds the right current densities in the coils in order to obtain a desired magnetic field.

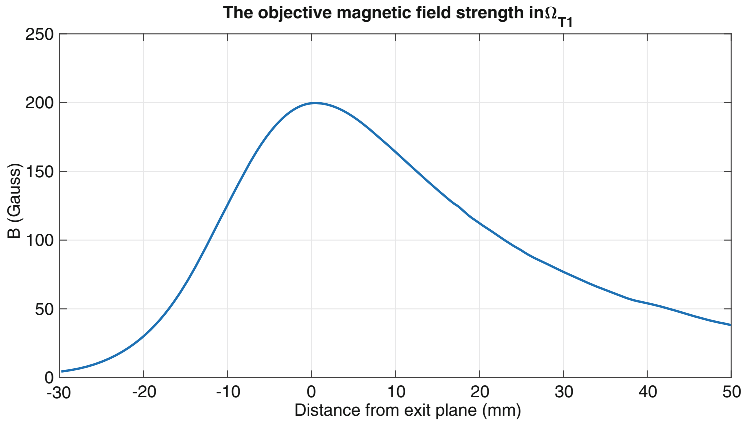


Fig. 3 The radial component of the objective magnetic field in the target region Ω_{T1}

The proposed optimization methodology consists of two steps: first, topology optimization is performed such that a first idea of the magnetic circuit shape is identified. Then, from that shape, shape and parametric optimization methods are performed to meet again the specifications while ensuring that the finale magnetic circuit shape will be feasible.

2 Topology Optimization: TO

The topology design optimization problem is formulated as follow:

$$(\mathfrak{P}) \begin{cases} \min\limits_{(X_\mu, X_J)} F_1(X_\mu, X_J) = \displaystyle\int_{\Omega_{T1}} \| B - B_0 \|^2 \, d\Omega, & (1) \\ \text{uc}: \quad F_2(X_\mu, X_J) = \displaystyle\int_{\Omega_{T2}} B_z^2 d\Omega \quad \leq \epsilon. \end{cases}$$

B is the magnetic field the solution of Maxwell equation:

$$-\nabla \times (\frac{1}{\mu_0.\mu_r} B) + J = 0, \tag{2}$$

with J is the current density, μ_r the relative permeability and μ_0 the permeability of vacuum. The design variable X_J represents the distribution of current densities on the coils section. Indeed, each coil section is meshed as shown in Fig. 4. X_J is defined as the vector of current densities $J = X_J(i)$ on each mesh element Ω_{S_i}, see Fig. 4. As fot the design variable X_μ, it represents the ferromagnetic materials layout on the magnetic circuit region. Indeed, each magnetic circuit part is

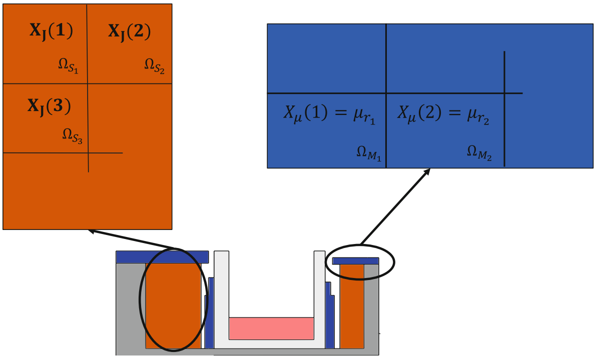


Fig. 4 The illustration of the design variables in topology optimization

meshed as shown in Fig. 4. X_μ is defined as the vector of the relative permeabilities $\mu_r = X_\mu(i)$ on each mesh element Ω_{M_i}. Meshes with $\mu_r = 1$ are considered as void meshes whereas those with $\mu_r > 1$ are considered filled with a ferromagnetic material whose relative permeability is equal μ_r. In practice, magnetic circuits are made only of one ferromagnetic material characterized by a relative permeability $\tilde{\mu_r}$. That means that after all, X_μ components are constrained to take just two possible values of μ_r: either 1 for void meshes or $\tilde{\mu_r}$ for meshes filled with material.

The cost function F_1 reflects the first specification which is that the magnetic field B along the channel center Ω_{T1} should be as close as possible to the specified magnetic field B_0. Here, B_0 is radial magnetic field whose strength varies according to a Gaussian function given in Fig. 3. The constraint function reflects the second specification which is that the magnetic field B in the exit plan Ω_{T2} should be radial or, in other words, its vertical component (z-component) should be as low as possible. The tolerance ϵ value is set based on the maximum permissible amplitude of the vertical component (z-component) of the magnetic field in the region Ω_{T2}.

The problem ($\mathfrak{P}$) is solved using gradient based optimization solver. To be able to do that the design variable X_μ, that is a binary variable (it takes just two possible values), is relaxed into a reel variable that could take all intermediate values between 1 and $\tilde{\mu_r}$. But in return, penalization techniques are integrated in the optimization procedure such that X_μ will be enforced to converge toward binary solutions (solutions with just 1 and $\tilde{\mu_r}$ values). More details about penalization techniques could be found in [2, 3]; herein, the TO results given later are obtained using arctan penalization method. In order to reduce the cpu-time needed to compute the gradients of the cost and the constraint functions, F_1 and F_2, the adjoint method is used, details are in the following section.

2.1 Sensitivity Analysis in TO

The functions F_1 and F_2 are not explicitly expressed in terms of X_J and X_μ, but rather in terms of the magnetic field B that is implicitly linked to X_J and X_μ via Maxwell equation (2). Hence, the cost and constraint functions F_1 and F_2 are implicit functions of the design variables X_J and X_μ. Thus, analytic expressions of F_1 and F_2 gradients could not be directly obtained. Herein the adjont method is used to provide the analytic expressions of F_1 and F_2 gradients but in terms of an additional variables: λ_1 and λ_2 . The adjoint method in similar context is already developed in [3]. In this paper, we just use directly the results. F_1 and F_2 gradients expressions are given by:

$$\frac{\partial F_1}{\partial X_\mu(i)} = \int_{\Omega_{M_i}} -\frac{1}{(\mu_0.X_\mu(i))^2}.B.\nabla \times \lambda_1 d\Omega, \tag{3}$$

$$\frac{\partial F_1}{\partial X_J(i)} = \int_{\Omega_{S_i}} \lambda_1 d\Omega, \tag{4}$$

$$\frac{\partial F_2}{\partial X_\mu(i)} = \int_{\Omega_{M_i}} -\frac{1}{(\mu_0.X_\mu(i))^2}.B.\nabla \times \lambda_2 d\Omega, \tag{5}$$

$$\frac{\partial F_2}{\partial X_J(i)} = \int_{\Omega_{S_i}} \lambda_2 d\Omega. \tag{6}$$

where λ_1 and λ_2 are the adjoint variables that are the solutions of the adjoint equations given by:

$$-\nabla \times (\frac{1}{\mu_0.\mu_r}\nabla \times \lambda_1 - M_1^{adj}\mathbb{1}_{\Omega_{T_1}}) = 0, \text{ with } M_1^{adj} = 2(B - B_0). \tag{7}$$

$$-\nabla \times (\frac{1}{\mu_0.\mu_r}\nabla \times \lambda_2 - M_2^{adj}\mathbb{1}_{\Omega_{T_2}}) = 0, \text{ with } M_2^{adj} = 2\begin{pmatrix} 0 \\ B_z \end{pmatrix}. \tag{8}$$

with $\mathbb{1}_{\Omega_{T_1}}$ and $\mathbb{1}_{\Omega_{T_2}}$ are indicator functions of the target regions: Ω_{T_1} and Ω_{T_2}. We can notice that both adjoint equations (7) and (8) are very similar to the following Maxwell equation given by :

$$-\nabla \times (\frac{1}{\mu_0.\mu_r}\nabla \times A - M) = 0. \tag{9}$$

with A is the potential vector and M the magnetization term. Hence, same existing solvers for Maxwell equations could be used to solve the adjoint equations (7) and (8).

A program named `ATOP`^{TO} was developed to solve topology optimization problems like ($\mathfrak{P}$). It is a `MATLAB` program based on:

- `fmincon` that is `MATLAB` non programming solver.
- `FEMM` that is finite element method (FEM) software used to solve Maxwell and adjoint equations (2), (7) and (8). In `fmincon` four gradient-based algorithms are available: `Active set` , `SQP` , `interior point` and `trust region reflective`. In our case, all this four algorithms were tried and it turns out that `interior point` algorithm is the most efficient one.

2.2 Topology Optimization Design Result

The optimized structure of the HET is given in Fig. 5. The topology of the magnetic circuit is identified by meshes that concentrate the most the magnetic field lines. The optimal current density values in the inner and outer coils are 1.6 A/mm^2 and 1.0 A/mm^2 respectively. Notice that on this particular solution, the coils are not meshed, which means that the current densities are considered homogeneous on each coil. This is only a choice (based on the manufacturer's recommendation), but it should be kept in mind that coil meshing is also a possible option.

The magnetic circuit topology, as shown in Fig. 5, is not yet the final solution for two main reasons. The first reason is that the obtained topology of the magnetic circuit is clearly not feasible. The second reason is that the non linearity of the magnetic circuit material is not taken into account. However, the topological solution is rather used to have a first idea of the shape of the magnetic circuit. Concretely, in Fig. 6, we draw in a the red line a simplified and feasible shape inspired from the topological solution. This procedure of simplifying the topological

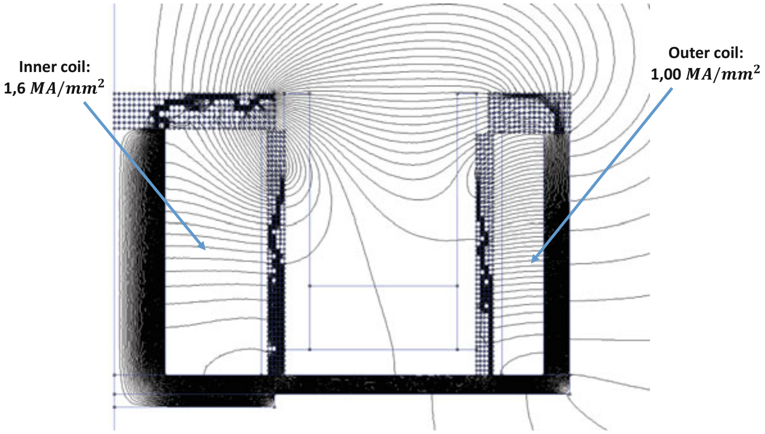
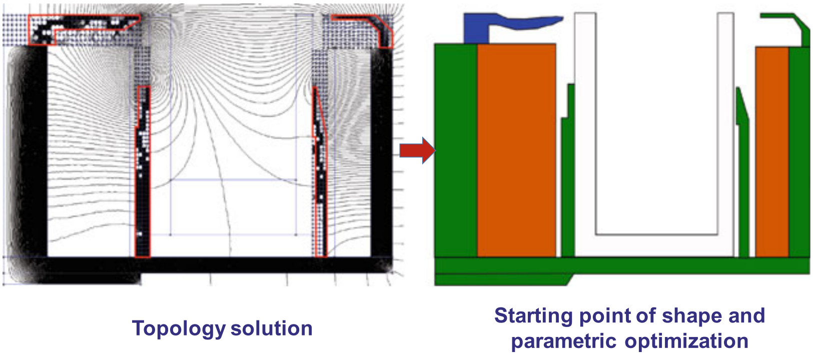


Fig. 5 The topology optimization design solution



Topology solution

Starting point of shape and parametric optimization

Fig. 6 TO design simplified in order to be used as starting point for SO and PO

solution is carried out carefully such that compliance with the specifications is not significantly degraded. In order to compensate for that deviation from the specifications (caused by the simplification procedure), the simplified structure (drawn in red line in Fig. 6) is redesigned by shape and parametric optimization (SO and PO). Moreover, in SO and PO we take into account the non linearity of the ferromagnetic material from which the magnetic circuit is made. Details are exposed in the next section.

3 Shape and Parametric Optimization: SO and PO

Shape and parametric optimization are carried out, together, on the structure given in Fig. 6. PO is performed on the green parts. As for SO, it is performed on the blue part of the structure since the details of its shape have a significant impact on the direction of the magnetic field at the exit plane. Shape and parametric design optimization problem is formulated by:

$$(\mathfrak{P})_{SP} \left\{ \begin{array}{l} \displaystyle\min_{(X_S, X_P, X_J)} F_1(X_S, X_P, X_J) = \int_{\Omega_{T1}} \| B - B_0 \|^2 \, d\Omega, \\ \text{uc :} \quad F_2(X_S, X_P, X_J) = \displaystyle\int_{\Omega_{T2}} B_z^2 d\Omega \quad \leq \epsilon. \end{array} \right. \tag{10}$$

The proleme $(\mathfrak{P})_{SP}$ is nearly similar to and $(\mathfrak{P})$, the only difference is the design variables. Here, X_J is defined of vector of two current densities in the inner and the outer coil sections $X_J = [J_{in}, J_{out}]$. X_S, X_P are respectively the shape and the parametric design variable. Indeed, the blue parts that will be designed by shape optimization are considered as Bezier curves, see Fig. 7. Hence, X_S is the vector of

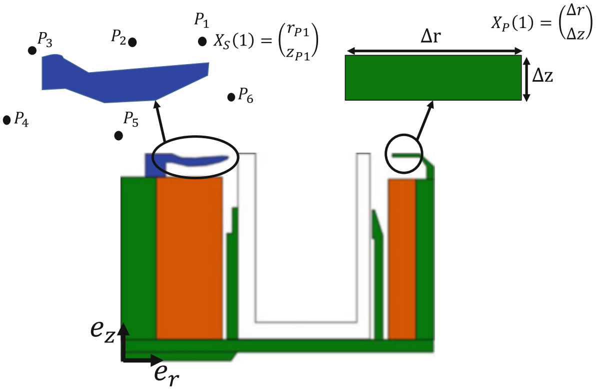

P_1
$X_S(1) = \begin{pmatrix} r_{P1} \\ z_{P1} \end{pmatrix}$
P_2
P_3
P_4
P_5
P_6
$X_P(1) = \begin{pmatrix} \Delta r \\ \Delta z \end{pmatrix}$
Δr
Δz
e_z
e_r

Fig. 7 The illustration of the design variables in shape and parametric optimization

the coordinates $\begin{pmatrix} r_{P_i} \\ y_{P_i} \end{pmatrix}$ of the control points P_i, see Fig. 7. As for the green parts, they are parametrized as shown in Fig. 7. Hence, X_P is the vector of the geometric parameters Δr and Δz of each piece of the green part of the magnetic circuit. Considering how the design variables X_S and X_P were defined, the final design of the magnetic circuit could not be unfeasible.

In order to solve $(\mathfrak{P})_{SP}$ MADS method ("Mesh Adaptive Direct Search") is used. It is a derivative-free optimization algorithm. MADS is implemented in NOMAD (Nonlinear Optimization by Mesh Adaptive Direct Search): a software application for simulation-based optimization. It consists of searching local solutions by adapting the mesh of the design variables space. Indeed, in each iteration MADS explores, on the current mesh, by evaluating the cost and the constraint functions F_1 and F_2 trying to find a better point that improves the current solution. If this does not succeed, the mesh is refined in the next iteration. More details are in [4] and [5].

The choice of a free-gradient optimization method as MADS is based on one main reason. Indeed, MADS method considers the objective and the constraint functions (that are FEM simulation based functions) as black boxes functions; this means that the MADS optimization algorithm is totally disconnected from the physics of the design problem. In particular, no further development is required to adapt the MADS algorithm in order to take into account the non-linearity of ferromagnetic materials within the FEM simulations.

But in return, we are fully aware that MADS could be expensive in terms of cpu-time. But this still acceptable for two reasons: first, in SO and PO the number the variables is very small compared to TO (30 variables in SO and PO versus 720 variables in TO) and second, the starting point given in Fig. 6 has been defined in such a way that it is not very far from compliance with the given specifications (see Sect. 2).

SO and PO Optimization Design Result The final design structure of the HET is given in Fig. 8. The magnetic field respect the specifications given in Sect. 1. Indeed, the radial component values of the magnetic field along the plasma channel

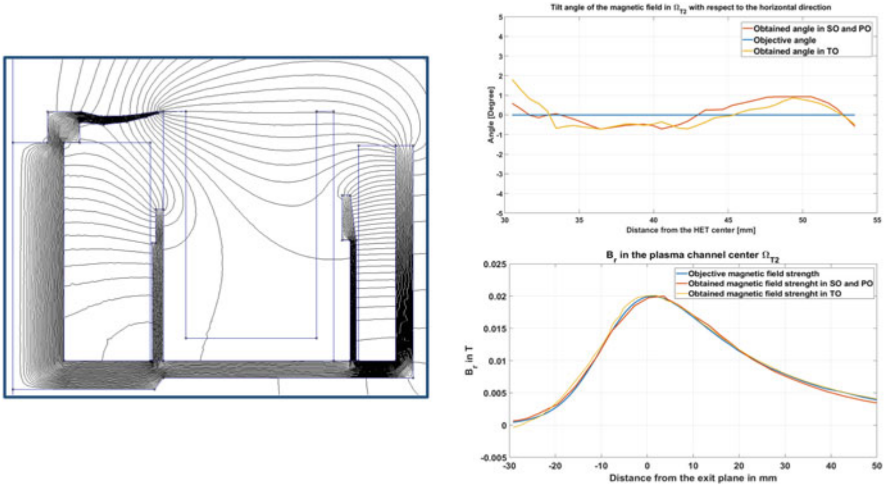


Fig. 8 Shape and parametric optimization design solution

center Ω_{T1} respect the specified Gaussian function with a maximal error of 4%. As for the exit plan region Ω_{T2}, the angle of the deviation of the magnetic field from the horizontal does not exceed 2°. The current density values are in the inner coil 2.2 A/mm^2 and 1.09 A/mm^2 in the outer coil. Special attention is paid to the fact that magnetic circuit part above the outer coil is completely erased by SO and PO. That makes the obtained structure original.

4 Conclusion

The HET so-optimized structure meets all the specifications. The magnetic circuit geometry is totally original regarding the existing ones. This was achieved using first topology optimization that gives the main idea of the magnetic circuit topology. Then parametric and shape optimization were performed in order to make the topological solution feasible. The same methodology can be used to design HETs for other purposes like: magnetic shielding or low erosion specifications.

References

1. Rossi, A., Messine, F., Hénaux, C.: Parametric optimization of a hall effect thruster magnetic circuit. In: Transactions of JSASS, Aerospace Technology, Japan, vol. 14(30), pp. 197–202 (2016)
2. Bendsøe, M.P., Sigmund, O.: Material interpolation schemes in topology optimization. Arch. Appl. Mech. **69**, 635–654 (1999)

3. Sanogo, S., Messine, F., Henaux, C., Vilamot, R.: Topology optimization for magnetic circuits dedicated to electric propulsion. IEEE Trans. Mag. **50**, 12 (2014)
4. Le Digabel, S.: Algorithm 909: NOMAD: nonlinear Optimization with the MADS Algorithm. ACM Trans. Math. Softw. **37**, 44:1–44:15 (2011)
5. Audet, C., Dennis, J.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**, 188–217 (2006)

Part V
Model Order Reduction

Two contributions form this part. In *Stability Preservation in Model Order Reduction of Linear Dynamical Systems*, the author, R. Pulch, examines projection-based model order reduction of Galerkin-type for linear dynamical systems. In the case of ordinary differential equations, a transformation of the original system guarantees that any reduced system inherits asymptotic stability. The transformation matrix satisfies a high-dimensional Lyapunov equation. It is used a frequency domain approach, where the solution of the Lyapunov equation represents a matrix-valued integral. Consequently, quadrature methods yield approximations in numerical computations. In the case of differential- algebraic equations, the stabilisation technique is applicable via a regularisation. Numerical results for a test example complete the work.

In the literature the 'Schur after MOR' method has proved successful in obtaining stable reduced piezoelectric device models. Even though the method is already used in industry, the stability preservation of 'Schur after MOR' is still mathematically unproven. In *Quasi-Schur Transformation for the Stable Compact Modeling of Piezoelectric Energy Harvester Devices* by S. Hu et al., it is shown that the involved quasi-Schur transformation indeed does efficiently restabilize the aforementioned reduced piezoelectric energy harvester models. The transformation is only quasi-Schur as the unstable reduced systems require eigenspace projection and approximation to become Schur-transformable. During the transformation, the negative eigenvalues are eliminated from the reduced stiffness matrix and the system is stabilized. Further, 'Schur after MOR' is also compared to another recently presented stabilization method, the 'MOR after Implicit Schur', and it is proven that the computational effort is significantly reduced.

Quasi-Schur Transformation for the Stable Compact Modeling of Piezoelectric Energy Harvester Devices

Siyang Hu, Chengdong Yuan, and Tamara Bechtold

Abstract The 'Schur after MOR' method has proved successful in obtaining stable reduced piezoelectric device models. Even though the method is already used in industry, the stability preservation of 'Schur after MOR' is still mathematically unproven. In this work, we show that the involved quasi-Schur transformation indeed does efficiently re-stabilize the aforementioned reduced piezoelectric energy harvester models. The transformation is only quasi-Schur as the unstable reduced systems require eigenspace projection and approximation to become Schur-transformable. During the transformation, the negative eigenvalues are eliminated from the reduced stiffness matrix and the system is stabilized. Further, we compare 'Schur after MOR' to another recently presented stabilization method: 'MOR after Implicit Schur'. We show that the computational effort is significantly reduced.

1 Introduction

Modeling and simulation-driven development has become state-of-the-art due to the increasing capacity of today's computers. However, even the power of modern computers fails to always cope with the faster growing demands of the industry. To overcome this issue, the methodology of model order reduction (MOR) has been introduced. MOR significantly reduces the computational effort required for e.g. system-level simulations by replacing the original high-dimensional model with a lower dimensional but still accurate surrogate. Novel MOR methods are mostly interpolation-based and perform well when applied to single-physical-domain models [3, 6, 7]. However, for models involving coupled physical domains, people often encounter difficulties in preserving the stable input/output behavior of the original system, e.g. in [9] and [10].

S. Hu (✉) · C. Yuan (✉) · T. Bechtold
Department of Engineering, Jade University of Applied Sciences, Wilhelmshaven, Germany

Institute for Electronic Appliances and Circuits, University of Rostock, Rostock, Germany
e-mail: siyang.hu@jade-hs.de; chengdong.yuan@jade-hs.de; tamara.bechtold@jade-hs.de

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_25

In [9], the authors introduce three different approaches to solve stability issues they have encountered when reducing piezoelectric models. However, except for ‘MOR after Schur’ in [8], none of those methods have been mathematically proven yet. In this contribution, we considered the ‘Schur after MOR’ approach, as it proved effective in a number of industrial applications. We prove that the stable input/output behavior of the original system can be re-established by the quasi-Schur transformation involved in ‘Schur after MOR’. The transformation is only quasi-Schur as an approximative pre-processing of the reduced model is required to make it Schur-transformable.

We emphasize that for the proof presented in this work, we only consider MOR for the models of piezoelectric MEMS devices created with the commercial FEM software ANSYS [1] using proportional damping. The mathematical representations of these models correspond to second order index-1 dynamical algebraic equation (DAE) systems with properties presented in Sect. 2. Further, we only consider fixed geometry, isotropic material and small deflections, which makes model linear. We do not consider different FE basis or other types of spatial discretization with e.g. finite differences or finite volume methods. For general application, one need to adapt the MOR algorithm itself. The proof will then only remain valid if all assumptions also remain intact.

Section 2 briefly introduces the piezoelectric energy harvester device. On its model, we recapture the ‘Schur after MOR’ approach in Sect. 3 while introducing some preliminaries on the way. Subsequently, we establish a link between the quasi-Schur transformation performed during ‘Schur after MOR’ and the stabilization of the reduced model. In Sect. 4, we present results of some numerical experiments. We verify our proof on two different harvester devices and show that ‘Schur after MOR’ is significantly more efficient than ‘MOR after Schur’ and its improved successor ‘MOR after Implicit Schur’, introduced in [8]. Finally, we conclude and give a brief outlook in Sect. 5.

2 Piezoelectric Energy Harvesters

Piezoelectric energy harvesters transform environmental mechanical vibration into electrical energy using piezoelectric effect [5, 9]. They can supply power to sensors deployed in harsh environmental conditions and where batteries or wires are undesirable, e.g. machine health monitoring. The mechanical part of piezoelectric energy harvesters are oscillatory systems consisting of mass and spring elements, where piezoelectric patches, the electric part, are attached to the spring elements to convert mechanical stress into electrical voltage.

The mechanical part of the vibrational energy harvester is fabricated using isotropic materials, e.g. steel or silicon. The piezoelectric patches are made of aluminum nitride (AIN) with an aluminum electrode on top and a platinum electrode on the bottom. For optimal energy harvesting, the resonance frequency of the system has to coincide with the excitation frequency. As the excitation frequency can

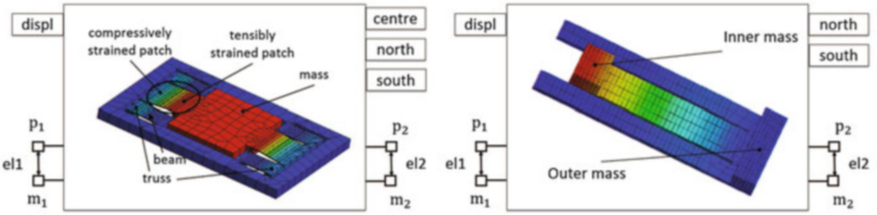


Fig. 1 Micro-structured piezoelectric energy harvester model adopted from [13] (left); tunable piezoelectric energy harvester model adopted from [8] (right)

vary depending on environmental parameters like humidity or temperature, dual frequency structures, e.g. in [13], are introduced that have a bandwidth of operating range in between the resonant frequencies. Geometric dimensions vary from a few millimeters for MEMS applications to several centimeters.

For the simulation of piezoelectric energy harvesters, the devices are usually modeled using commercial finite elements software, e.g. ANSYS [1]. This results in a linear coupled domain finite element model consisting of a mechanical and an electrical part as the geometrical and material properties are fixed and the displacement is minor. The FEM models of two different implementations of a piezoelectric energy harvester are depict in Fig. 1, where ***center***, ***north*** and ***south*** refers to the center of gravity of mass elements and ***el1***, ***el2*** refers to the electrical (voltage) ports of the system, which can be interfaced to electrical circuitries.

The mathematical representation of the mechanical part of the coupled domain finite element model reads:

$$\mathbf{M_{11}}\ddot{\mathbf{x}}_\mathbf{1} + \mathbf{D_{11}}\dot{\mathbf{x}}_\mathbf{1} + \mathbf{K_{11}}\mathbf{x_1} = \mathbf{b_1}u, \tag{1}$$

where $\mathbf{M_{11}}, \mathbf{K_{11}} \in \mathbb{R}^{n \times n}$ are the symmetric positive definite (s.p.d.) mass and stiffness matrices, respectively. $\mathbf{D_{11}} = \alpha\mathbf{M_{11}} + \beta\mathbf{K_{11}}, \alpha, \beta \in \mathbb{R}$ is the damping matrix and $\mathbf{x_1}$ is the vector of nodal displacements. The electrical part of the coupled domain finite element model reads:

$$\mathbf{K_{22}}\mathbf{x_2} = \mathbf{b_2}u, \tag{2}$$

with $\mathbf{K_{22}} \in \mathbb{R}^{k \times k}$ the electrical conductivity matrix, which is symmetric negative definite (s.n.d.) and

$$||\lambda_{\max}(\mathbf{K_{22}})|| \ll ||\lambda_{\min}(\mathbf{K_{11}})|| \tag{3}$$

holds for the respective eigenvalues. $\mathbf{x_2}$ is a vector of nodal electrical potentials. Both physical domains are coupled via piezoelectric patches, which transform vibrational stress into electric field. Thus, we have the piezoelectric coupling term $\mathbf{K_{12}} \in \mathbb{R}^{n\times k}$, such that:

$$\Sigma = \begin{cases} \underbrace{\begin{bmatrix} \mathbf{M_{11}} & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{M}} \begin{bmatrix} \ddot{\mathbf{x}}_\mathbf{1} \\ \ddot{\mathbf{x}}_\mathbf{2} \end{bmatrix} + \underbrace{\begin{bmatrix} \mathbf{D_{11}} & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \begin{bmatrix} \dot{\mathbf{x}}_\mathbf{1} \\ \dot{\mathbf{x}}_\mathbf{2} \end{bmatrix} + \underbrace{\begin{bmatrix} \mathbf{K_{11}} & \mathbf{K_{12}} \\ \mathbf{K_{12}^T} & \mathbf{K_{22}} \end{bmatrix}}_{\mathbf{K}} \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \end{bmatrix}}_{\mathbf{b}} u \\ y = \mathbf{c^T} \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{bmatrix} \end{cases} . \tag{4}$$

The input function u corresponds to the displacement imposed to the harvester structure with input distribution vector $\mathbf{b} \in \mathbb{R}^{n+k}$ chosen accordingly. The total electrical potential is gathered via the output vector $\mathbf{c} \in \mathbb{R}^{n+k}$ within the output y.

3 Schur After MOR

This section briefly reassembles the ‘Schur after MOR’ procedure introduced in [9]. For a survey on general MOR methods for this class of models, please refer to [4].

‘Schur after MOR’ method stabilizes unstable reduced order models:

$$\Sigma_r = \begin{cases} \mathbf{M_r}\ddot{\mathbf{x}}_\mathbf{r} + \mathbf{D_r}\dot{\mathbf{x}}_\mathbf{r} + \mathbf{K_r}\mathbf{x_r} = \mathbf{b_r}u \\ y = \mathbf{c_r^T}\mathbf{x_r} \end{cases} , \tag{5}$$

obtained by projective MOR: $\mathbf{V^T}\Sigma\mathbf{V}$, where:

$$\begin{gathered} \{\mathbf{M_r}, \mathbf{D_r}, \mathbf{K_r}\} = \mathbf{V^T}\{\mathbf{M}, \mathbf{D}, \mathbf{K}\}\mathbf{V}, \\ \mathbf{b_r} = \mathbf{V^T}\mathbf{b} \quad \text{and} \quad \mathbf{c_r^T} = \mathbf{c^T}\mathbf{V}. \end{gathered} \tag{6}$$

$\mathbf{V} \in \mathbb{R}^{(n+k)\times p}$, $p \ll n+k$, is chosen as an orthonormal basis of the p-dimensional second-order input Krylov subspace via the second order Arnoldi reduction (SOAR) method [2, 12]:

$$\mathcal{K}_p(-\mathbf{K^{-1}M}, -\mathbf{K^{-1}D}, -\mathbf{K^{-1}b}). \tag{7}$$

The stabilization of the reduced model is achieved by performing a quasi-Schur transformation on Σ_r, where Σ_r is approximated by a system of DAEs before being Schur transformed. The approximation involves an eigen-transformation $\widetilde{\Sigma}_r = \mathbf{T^T}\Sigma_r\mathbf{T}$, where $\mathbf{T}$ is a sorted orthonormal eigenbasis of the matrix $\mathbf{M_r}$, such that for the entries of $\widetilde{\mathbf{M}}_\mathbf{r}$, $\widetilde{m}_{r,ii} \geq \widetilde{m}_{r,jj}$ holds for all $i > j$. In the next step, we set $\widetilde{m}_{r,ii} = 0$ for all $i \geq I$ with $I \in [1, p]$, such that $\widetilde{m}_{r,(I-1)(I-1)} \gg \widetilde{m}_{r,II}$. In this way, we obtain a reduced order DAE system, which can be Schur transformed. We call the subspace spanned by all those eigenvectors corresponding to these eigenvalues $\widetilde{m}_{r,ii}, i \geq I$ quasi-algebraic.

In [12], a criteria for the stability of a second-order DAE is given.

Lemma 1 (Stability Criteria for Second-Order DAEs [12]) *A second-order DAE is stable, if* $D + D^T \succeq 0$, $M = M^T \succeq 0$ *and* $K = K^T \succ 0$.

Proof The proof is given in [12]. □

With this criteria, we can prove that the quasi-Schur transformation stabilizes the reduced order system.

Theorem 1 *The quasi-Schur transformation stabilizes the reduced model* Σ_r.

Proof As $\mathbf{M_{11}}$ is s.p.d., $\mathbf{M_r}$ has to be symmetric positive semi-definite as well. Furthermore, $\mathbf{K_r}$ must have negative eigenvalues. Otherwise, Σ_r is stable according to Lemma 1.

Since $\mathbf{M}$ and $\mathbf{K}$ can obviously be simultaneously diagonalized (e.g. with eigenbasis of the matrix pencil $\mathbf{M} - \lambda\mathbf{K}$), the system domain can be represented as a direct sum of these eigenspaces. Thus:

$$\lambda(\mathbf{K_r}) = \sum_i \nu_i \lambda_i(\mathbf{K}), \quad \sum_i \nu_i = 1, \tag{8}$$

holds for all eigenvalues of $\mathbf{K_r}$. Now, given (3) and let $P, N \subset \{1, \ldots, n+k\}$ be the set of indices corresponding to the positive and negative eigenvalues of $\mathbf{K}$. Equation (8) can only be negative if $\sum_{i \in P} \nu_i \ll \sum_{i \in N} \nu_i$. That is to say, given the structure of Σ, the subspaces of the reduced system corresponding to these negative eigenvalues has to be dominated by the electric domain. Since $\mathbf{M}$ and $\mathbf{K}$ share the same decomposition, $\lambda(\mathbf{M_r}) = \sum_{i \in P} \nu_i \lambda(\mathbf{M})_i \approx 0$ must also hold.

Finally, when $\widetilde{\Sigma}_r$ is Schur transformed, we have:

$$\mathbf{\widetilde{K}_s = \widetilde{K}_{r,(1:I,1:I)} - \widetilde{K}_{r,(1:I,I:p)} \widetilde{K}^{-1}_{r,(I:p,I:p)} \widetilde{K}_{r,(I:p,1:I)}} \tag{9}$$

which is s.p.d as $\mathbf{\widetilde{K}_{r,(1:I,1:I)}}$ is s.p.d. and $\mathbf{\widetilde{K}^{-1}_{r,(I:p,I:p)}}$ s.n.d.[1] This makes the quasi-Schur Transformed system stable according to Lemma 1. □

[1] $\mathbf{\widetilde{K}_{r,(1:I,1:I)}}$ is the submatrix consisting of the first I rows and columns of $\mathbf{\widetilde{K}_r}$, and $\mathbf{\widetilde{K}_{r,(I:p,I:p)}}$ consists of the the rows and columns I to p of $\mathbf{\widetilde{K}_r}$.

Remark 1 In industrial software, the quasi-Schur Transformation introduced in [9] is actually modified [11]. The index I is obtained by the eigen-transformation $\widehat{\mathbf{K}}_{\mathbf{r}}$=$\mathbf{T}_{\mathbf{K}}^{\mathbf{T}}\mathbf{K}_{\mathbf{r}}\mathbf{T}_{\mathbf{K}}$ and then setting I such that $\widehat{K}_{r,ii} < 0$ for all $i \geq I$. This equivalent criteria is easier to implement and more robust.

4 Numerical Experiments

For the verification of Theorem 1, the micro-structured energy harvester device from [9] (see Fig. 1 left) as well as a novel frequency tunable piezoelectric energy harvester introduced in [8] (see Fig. 1 right) are considered. Both energy harvester models are excited by a displacement input ***displ***. To test the stability of the reduced model, the displacements at further nodes (***centre***, ***north***, ***south***) are selected as outputs of the system.

According to the observation from [8, 9], the reduced piezoelectric energy harvester model computed by conventional SOAR method shows unstable behavior at the electrical ports ***el1*** and ***el2***. Therefore, quasi-Schur transformation has been performed subsequently on these two reduced models to stabilize them. The accuracy of the respective frequency responses are shown in Fig. 2. The plots also include the reduced models obtained from ‘MOR after Implicit Schur’ from [8] for comparison.

To verify Theorem 1, we compute the angle between the respective subspaces $T_{K,I}$ and $T_{M,I}$ corresponding to the negative eigenvalues of $\mathbf{K_r}$ and the near 0 elements in $\mathbf{M_r}$. Various reduced models with different dimensions (from 6 up to 240) have been tested. We found the considered subspaces to coincide (see Fig. 4, $\theta = 0°$), even when taking numeric errors into account. This verifies Theorem 1 and justifies the assumption of a quasi-algebraic subspace. Furthermore, Fig. 3 shows reduced order models different dimensions (6, 12, 18, 24 and 30 or 2–4 DOF less after Schur transformation) compared to the full model. The deviation is negligible for reduced models with more than 16 DOF.[2]

Table 1 shows the computation times of ‘Schur after MOR’ compared to ‘MOR after Implicit Schur’ from [8] for generating the reduced order model of the tunable piezoelectric energy harvester model with 24643 degrees of freedom. ‘Schur after MOR’ speeds up the computation of the reduced order model, as it avoids the implicit Schur transformation on the full model.

Remark 2 For the sake of completeness and to demonstrate the capability of MOR, we present the CPU times, full vs. reduced models of both devices, when used to compute a simple step response: The computations time at respectively 1469.5 s

[2] Expansion point for MOR is set to $s_0 = 0$ as obtaining the smallest possible accurate reduced model was not the main concern of this experiment. We rather want to verify the equivalence of considered subspaces. With the optimal choice of expansion points, one can obtain smaller models with acceptable accuracy.

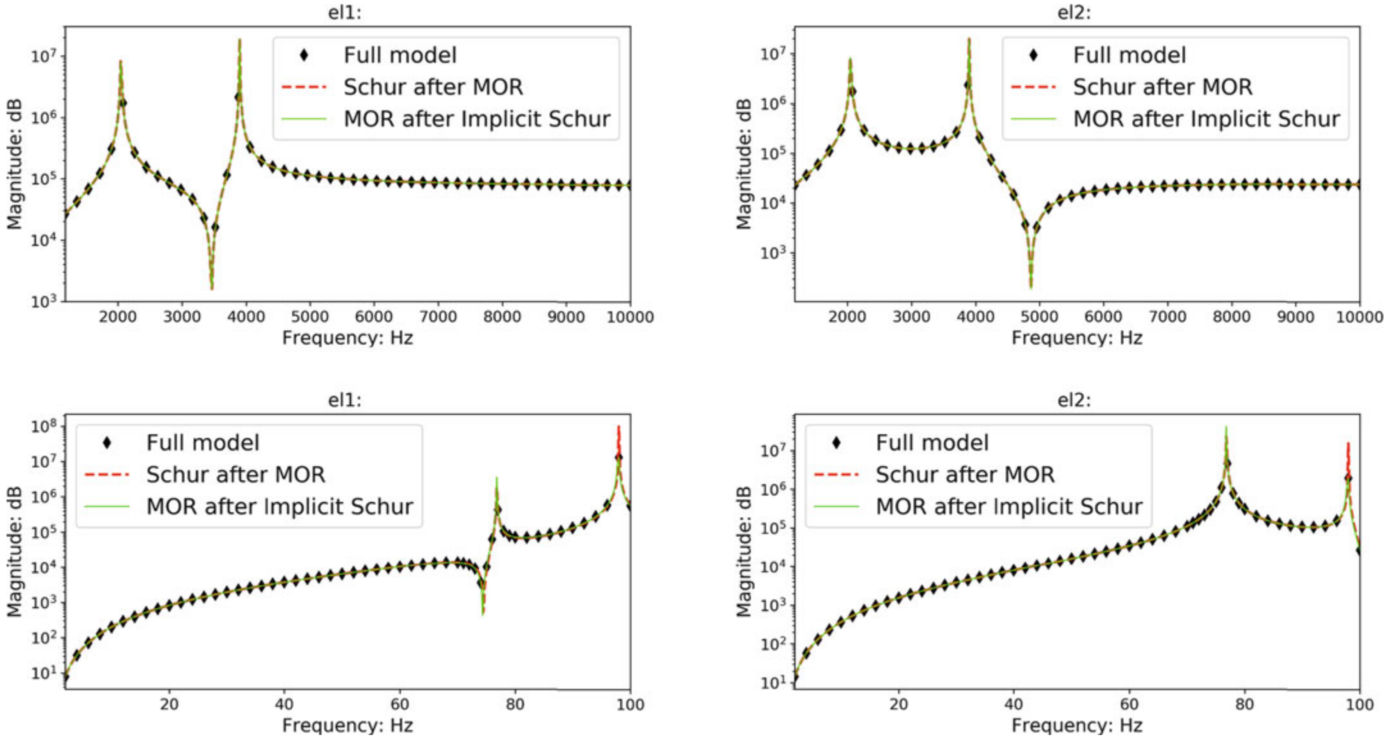


Fig. 2 Frequency response of the output voltage from ports ***el1*** and ***el2*** with displacement excitation ***displ*** of the micro-structured model (top, 48351 DOF), frequency tunable harvester model (bottom, 29392 DOF) and the respective reduced models (30 DOF (The reduced models obtained from ‘Schur after MOR’ actually have smaller DOF depending on the size of the quasi-algebraic subspace. The DOF after Schur transformation can be found in Fig. 3.)) from [9] and [8]

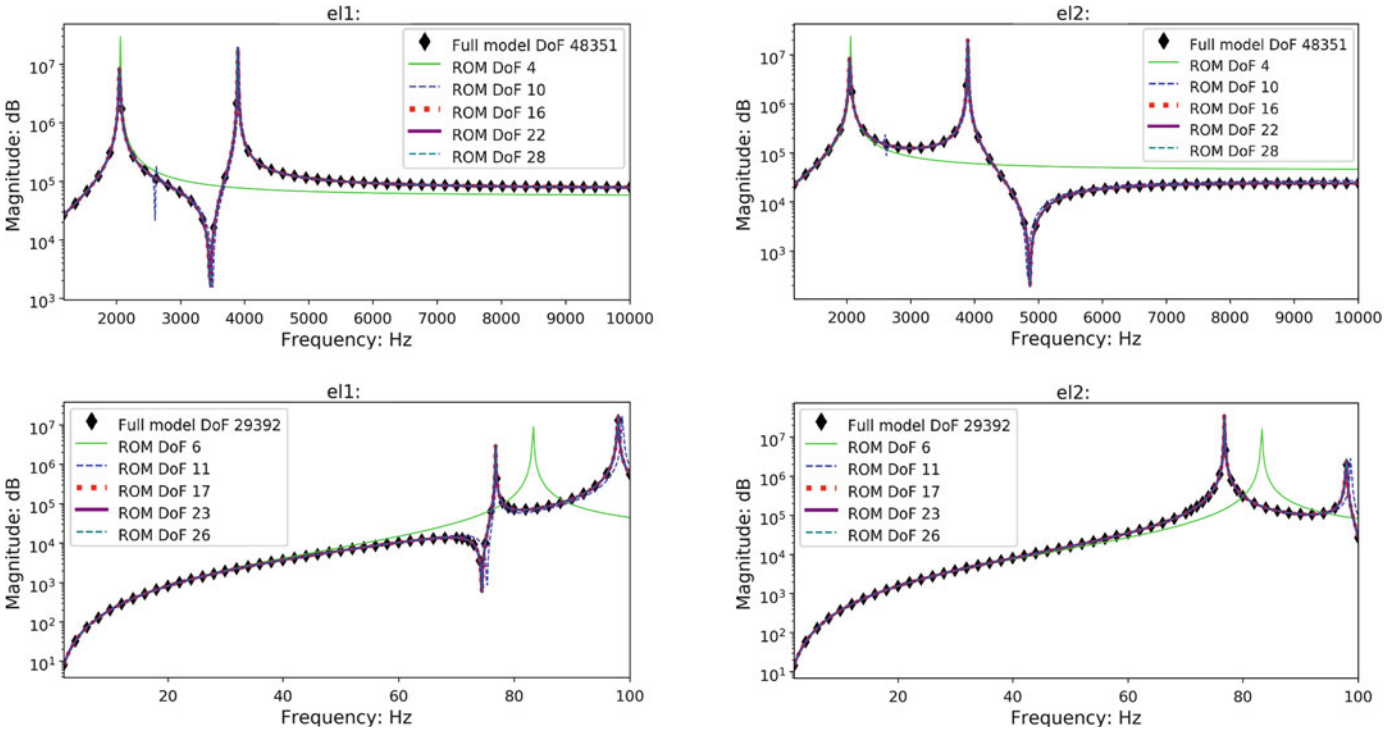


Fig. 3 Frequency response of the output voltage from ports ***el1*** and ***el2*** with displacement excitation ***displ*** of the micro-structured model (top, 48351 DOF), frequency tunable harvester model (bottom, 29392 DOF) and the respective reduced models of different DOF

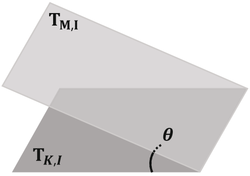


Fig. 4 Angle θ between the quasi-algebraic subspaces $\mathbf{T}_{\mathbf{M},\mathbf{I}}$ and $\mathbf{T}_{\mathbf{K},\mathbf{I}}$

Table 1 Computation time of 'Schur after MOR' vs. 'MOR after implicit Schur' for the generation of reduced order model (on Intel® Core™ i5-7600 CPU@3.5 GHz, 32 GB RAM)

Reduced order	Schur after MOR	MOR after implicit Schur
30	36.98 s	52.85 s
90	41.48 s	57.24 s
240	57.42 s	71.11 s

(full, 48351 DOF) vs. 0.08916 s (reduced, 28 DOF) for the micro-structured model, and 277.7 s (full, 29392 DOF) vs. 0.08156 s (reduced, 28 DOF) for the frequency tunable model.

5 Conclusion and Outlook

In this work, we have given a mathematical proof for stability preservation of 'Schur after MOR' method, which was initially suggested in [9]. We have shown that the quasi-Schur transformation, when applied to reduced models of piezoelectric energy harvesters obtained by projective MOR, stabilizes the models. We have shown the efficiency of the method (∼30% decrease of computation time) compared to 'MOR after Implicit Schur', which was initially suggested in [8].

In the next step, one can compare quasi-Schur transformation with conventional stabilization method, e.g. by simply truncating the unstable part of the reduced system. Finally, one can also consider comparing the performance of the whole 'Schur after MOR' procedure to structure preserving MOR, which is a third suggested stability preserving MOR method in [9].

References

1. ANSYS Inc.: ANSYS Academic Research Mechanical, Release 19.2 (2018)
2. Bai, Z., Su, Y.: Dimension reduction of large-scale second-order dynamical systems via a second-order Arnoldi method. SIAM J. Sci. Comput. **26**(5), 1692–1709 (2005)
3. Beattie, C., Gugercin, S.: Chapter 7: Model reduction by rational interpolation. In: Benner, P. et al. (eds.) Model Reduction and Approximation, pp. 297–334. SIAM, Philadelphia (2017)

4. Benner, P., Stykel, T.: Model order reduction for differential-algebraic equations: a survey. In: Surveys in Differential-Algebraic Equations IV, pp. 107–160. Springer, New York (2017)
5. Bouhedma, S., Zheng, T.H., Hohlfled, D.: Multiphysics Modeling and Simulation of a Dual Frequency Energy Harvester. In: ECMS 2018 Proceedings 2018, pp. 386–390 (2018)
6. Freund, R.W.: Krylov-subspace methods for reduced-order modeling in circuit simulation. J. Comput. Appl. Math. **123**, 395–421 (2000)
7. Gugercin, S., Stykel, T., Wyatt, S.: Model reduction of descriptor systems by interpolary projection methods SIAM J. Sci. Comput. **35**(5), B1010–B1033 (2013)
8. Hu, S.Y., Yuan, C.D., Castagnotto, A., Lohmann, B., Bouhedma, S., Hohlfeld, D., Bechtold, T.: Stable reduced order modeling of piezoelectric energy harvesting modules using implicit schur complement. Microelectron. Reliab. **85**, 148–155 (2018)
9. Kudryavtsev, M., Rudnyi, E.B., Korvink, J.G., Hohlfeld, D., Bechtold, T.: Computationally efficient and stable order reduction methods for a large-scale model of mems piezoelectric energy harvester. Microelectron. Reliab. **55**(5), 747–757 (2015)
10. Kurch, M., Entwicklung einer Simulationsumgebung für die Auslegung piezoelektrischer Energy Harvester. Ph.D. thesis, Technische Universität Darmstadt (2014)
11. Rudnyi, E.B.: MOR for ANSYS. In: System-Level Modeling of MEMS, pp. 425–438 (2013)
12. Salimbahrami, S.B.: Structure preserving order reduction of large scale second order models. Ph.D. thesis, Technische Universität München (2005)
13. Wang, Z., et al. A piezoelectric vibration harvester based on clamped-guided beams. 2012 IEEE 25th International Conference on Micro Electro Mechanical Systems (MEMS). IEEE, New York (2012)

Stability Preservation in Model Order Reduction of Linear Dynamical Systems

Roland Pulch

Abstract We examine projection-based model order reduction of Galerkin-type for linear dynamical systems. In the case of ordinary differential equations, a transformation of the original system guarantees that any reduced system inherits asymptotic stability. The transformation matrix satisfies a high-dimensional Lyapunov equation. We use a frequency domain approach, where the solution of the Lyapunov equation represents a matrix-valued integral. Consequently, quadrature methods yield approximations in numerical computations. In the case of differential-algebraic equations, the stabilization technique is applicable via a regularization. We present numerical results for a test example.

1 Introduction

Mathematical modelling of electronic circuits and devices yields typically high-dimensional dynamical systems. On the one hand, linear or nonlinear electric networks consist of a large number of basic components, which cause large systems of differential-algebraic equations (DAEs). On the other hand, a refined modelling of secondary effects implies (coupled) multiphysics problems, which include partial differential equations (PDEs). A spatial discretisation of the PDE part generates systems of ordinary differential equations (ODEs), which are often linear. Methods of model order reduction (MOR) allow for decreasing the dimensionality of the systems, see [1, 4, 16].

We consider linear dynamical systems of high dimensions, where projection-based MOR produces low-dimensional systems. The balanced truncation technique, see [1], is a Petrov-Galerkin-type MOR method, which guarantees the asymptotic stability of all reduced systems. However, this technique requires a large computational effort in comparison to Krylov subspace methods, for example. Hence we

R. Pulch (✉)
Institute of Mathematics and Computer Science, Universität Greifswald, Greifswald, Germany
e-mail: roland.pulch@uni-greifswald.de

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2_26

investigate MOR methods of Galerkin-type, which are relatively cheap. A reduced system may be unstable, even though the original system is asymptotically stable.

In the case of ODEs, a transformation yields an equivalent system, where the preservation of stability is guaranteed, see [5, 14]. A Lyapunov inequality characterises the set of admissible transformation matrices. This inequality can be satisfied by solving a high-dimensional Lyapunov equation. We use the approach from [15]. The solution also represents a matrix-valued integral in the frequency domain. Hence quadrature methods discretise the integral. In the case of DAEs, we employ a stability-preserving regularisation from [10]. Consequently, the transformation technique is applicable to the regularised system of ODEs, which guarantees a stable reduced system again, see [15].

In this paper, we apply the stabilisation technique to an example consisting of linear DAEs. A band-pass filter was already considered as test example in [13]. A modelling with random parameters followed by the stochastic Galerkin method yields a large coupled system of DAEs. We compare the direct solution of the Lyapunov equation and the frequency domain approach using a quadrature rule. The results of numerical computations are presented using MOR with the one-sided Arnoldi method.

2 Linear Dynamical Systems and Model Order Reduction

We consider linear time-invariant dynamical systems

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) \end{aligned} \tag{1}$$

with matrices $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n_{\text{in}}}$, $\mathbf{C} \in \mathbb{R}^{n_{\text{out}} \times n}$. The state variables or inner variables $\mathbf{x} : I \rightarrow \mathbb{R}^n$, the inputs $\mathbf{u} : I \rightarrow \mathbb{R}^{n_{\text{in}}}$ and the outputs $\mathbf{y} : I \rightarrow \mathbb{R}^{n_{\text{out}}}$ represent functions on a time interval $I = [t_0, t_{\text{end}}]$. Initial values $\mathbf{x}(t_0) = \mathbf{x}_0$ are predetermined. The system (1) consists of ODEs in the case of a non-singular mass matrix $\mathbf{E}$, whereas the system represents DAEs in the case of a singular mass matrix. We assume that the linear dynamical system (1) is asymptotically stable, i.e., all eigenvalues λ satisfying $\det(\lambda \mathbf{E} - \mathbf{A}) = 0$ have a negative real part.

The purpose of MOR is to construct a linear time-invariant dynamical system

$$\begin{aligned} \bar{\mathbf{E}}\dot{\bar{\mathbf{x}}}(t) &= \bar{\mathbf{A}}\bar{\mathbf{x}}(t) + \bar{\mathbf{B}}\mathbf{u}(t) \\ \bar{\mathbf{y}}(t) &= \bar{\mathbf{C}}\bar{\mathbf{x}}(t) \end{aligned} \tag{2}$$

with matrices $\bar{\mathbf{A}}, \bar{\mathbf{E}} \in \mathbb{R}^{r \times r}$, $\bar{\mathbf{B}} \in \mathbb{R}^{r \times n_{\text{in}}}$, $\bar{\mathbf{C}} \in \mathbb{R}^{n_{\text{out}} \times r}$ of much lower dimension $r \ll n$. Yet the outputs of (1) and (2) should agree, i.e., $\mathbf{y}(t) \approx \bar{\mathbf{y}}(t)$ for all $t \in I$. We call (1) a full-order model (FOM) and (2) a reduced-order model (ROM).

In projection-based MOR of Petrov-Galerkin-type, each method determines two projection matrices $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times r}$ of full rank. We assume that $\mathbf{V}$ is an orthogonal matrix, i.e., $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_r$ with the identity matrix. The reduced matrices within the system (2) read as

$$\bar{\mathbf{A}} = \mathbf{W}^\top \mathbf{A} \mathbf{V}, \quad \bar{\mathbf{B}} = \mathbf{W}^\top \mathbf{B}, \quad \bar{\mathbf{C}} = \mathbf{C} \mathbf{V}, \quad \bar{\mathbf{E}} = \mathbf{W}^\top \mathbf{E} \mathbf{V}. \tag{3}$$

A Galerkin-type projection-based MOR is characterised by the choice $\mathbf{W} = \mathbf{V}$ in (3), where just an appropriate matrix $\mathbf{V}$ has to be determined. Well-known Galerkin techniques are, for example, the one-sided Arnoldi method and proper orthogonal decomposition (POD), see [1].

The input-output mapping of a linear dynamical system (1) is described by a transfer function $\mathbf{H} \in \mathbb{C}^{n_{\mathrm{out}} \times n_{\mathrm{in}}}$, $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ for almost all $s \in \mathbb{C}$. The $\mathcal{H}_2$-norm of the Hardy space reads as

$$\|\mathbf{H}\|_{\mathcal{H}_2} = \sqrt{\frac{1}{2\pi} \int_{-\infty}^{+\infty} \|\mathbf{H}(\mathrm{i}\omega)\|_{\mathrm{F}}^2 \, \mathrm{d}\omega} \qquad \text{with} \quad \mathrm{i} = \sqrt{-1} \tag{4}$$

and the Frobenius norm $\| \cdot \|_{\mathrm{F}}$ provided that the integral exists. An asymptotically stable system of ODEs always has a finite $\mathcal{H}_2$-norm. However, the existence is not guaranteed in the case of DAEs, see [3]. Now the difference between the FOM (1) and the ROM (2) can be measured by the absolute error $\|\mathbf{H}_{\mathrm{FOM}} - \mathbf{H}_{\mathrm{ROM}}\|_{\mathcal{H}_2}$ or the relative error $\|\mathbf{H}_{\mathrm{FOM}} - \mathbf{H}_{\mathrm{ROM}}\|_{\mathcal{H}_2} / \|\mathbf{H}_{\mathrm{FOM}}\|_{\mathcal{H}_2}$ in the norm (4) if the integrals exist. We use the $\mathcal{H}_2$-norm to quantify the error of an MOR only. We do not investigate $\mathcal{H}_2$-optimal MOR techniques, cf. [6].

3 Stability Preservation and Lyapunov Equations

In many projection-based MOR methods, the ROM (2) may be unstable, although the FOM (1) is asymptotically stable. Several special techniques were constructed, which guarantee a stable reduced system, see [2, 8], for example.

Let the system (1) be linear ODEs. If both the matrix $\mathbf{E}$ is symmetric positive definite and the matrix $\mathbf{A}$ is dissipative ($\mathbf{A} + \mathbf{A}^\top$ is negative definite), then any Galerkin-type method yields an asymptotically stable system (2). Otherwise, the system (1) can be transformed into an equivalent system satisfying these properties, see [5, 14]. This approach cannot be extended to MOR methods of Petrov-Galerkin-type, because the symmetry of some matrices is essential.

Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. The transformed linear dynamical system reads as

$$\mathbf{E}^\top \mathbf{M} \mathbf{E} \dot{\mathbf{x}}(t) = \mathbf{E}^\top \mathbf{M} \mathbf{A} \mathbf{x}(t) + \mathbf{E}^\top \mathbf{M} \mathbf{B} \mathbf{u}(t), \tag{5}$$

where the mass matrix is symmetric positive definite. The involved transformation matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ has to satisfy the Lyapunov inequality

$$\mathbf{A}^\top \mathbf{M}\mathbf{E} + \mathbf{E}^\top \mathbf{M}\mathbf{A} < \mathbf{0}, \tag{6}$$

i.e., the left-hand side represents a negative definite matrix. There is an infinite set of solutions for (6). The inequality is satisfied if and only if $\mathbf{M}$ solves the (generalised) Lyapunov equation

$$\mathbf{A}^\top \mathbf{M}\mathbf{E} + \mathbf{E}^\top \mathbf{M}\mathbf{A} + \mathbf{F} = \mathbf{0} \tag{7}$$

for a symmetric positive definite matrix $\mathbf{F}$. A Galerkin-type reduction of the transformed system (5) is equivalent to a Petrov-Galerkin-type projection (3) of the original system (1) with $\mathbf{W} = \mathbf{M}\mathbf{E}\mathbf{V}$. Hence we do not require to compute the transformed system (5).

A direct solution of the Lyapunov equation (7) is not possible for high dimensions n due to a huge computational effort $O(n^3)$, see [11]. Approximate techniques often produce low-rank factorisations

$$\mathbf{M} \approx \widetilde{\mathbf{M}} = \mathbf{Z}\mathbf{Z}^\top \quad \text{with} \quad \mathbf{Z} \in \mathbb{R}^{n \times k} \tag{8}$$

and $k \ll n$. However, the approximation $\widetilde{\mathbf{M}}$ is a singular matrix, which causes problems, cf. [14]. Furthermore, iterative methods like the alternating direction implicit (ADI) algorithm, for example, require a low-rank factorisation $\mathbf{F} = \mathbf{G}\mathbf{G}^\top$ with $\mathbf{G} \in \mathbb{R}^{n \times \ell}$ satisfying $\ell \ll n$, where $\mathbf{F}$ becomes just semi-definite.

4 Frequency Domain Approach

Due to Parseval's theorem, we write the solution of the (generalised) Lyapunov equations (7) as a matrix-valued integral in the frequency domain. It holds that

$$\mathbf{M} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathbf{S}(\omega)^{-\mathsf{H}} \mathbf{F} \mathbf{S}(\omega)^{-1} \, \mathrm{d}\omega \tag{9}$$

including the matrix

$$\mathbf{S}(\omega) = \mathrm{i}\omega \mathbf{E} - \mathbf{A} \in \mathbb{C}^{n \times n} \quad \text{with} \quad \mathrm{i} = \sqrt{-1}. \tag{10}$$

Phillips and Silveira [12] apply a quadrature rule with positive weights to compute an approximation of the integral (9). This approach yields a factorisation (8) of rank $k = q\ell$ for $\mathbf{F} = \mathbf{G}\mathbf{G}^\top$, where q is the number of nodes in the quadrature and ℓ is the rank of $\mathbf{G}$.

Alternatively, the idea from [15] is to arrange the identity matrix $\mathbf{F} = \mathbf{I}_n$ in (7), which owns the trivial (high-rank) factorisation $\mathbf{F} = \mathbf{I}_n \mathbf{I}_n^\top$ ($\mathbf{G} = \mathbf{I}_n$). We assume large sparse matrices $\mathbf{A}, \mathbf{E}$. The matrix (10) inherits this sparsity. However, the inverse matrices are dense in (9) and thus we never compute them explicitly. We do not need the matrix $\mathbf{M}$ but the matrix-matrix product $\mathbf{W} = \mathbf{M}\mathbf{V}'$ with $\mathbf{V}' = \mathbf{E}\mathbf{V}$. The computational effort for the multiplication $\mathbf{E}\mathbf{V}$ is low if $\mathbf{E}$ is sparse. Now the integral (9) simplifies to $\mathbf{W} = \frac{1}{\pi}\,\mathrm{Re}[\mathbf{U}]$ with

$$\mathbf{U} = \int_0^{+\infty} \mathbf{S}(\omega)^{-\mathsf{H}} \mathbf{S}(\omega)^{-1} \mathbf{V}' \;\mathrm{d}\omega \tag{11}$$

due to a symmetry. We transform the integral (11) to the finite domain [0, 1] and obtain

$$\mathbf{U} = \int_0^1 \mathbf{S}\left(\tfrac{\xi}{1-\xi}\right)^{-\mathsf{H}} \mathbf{S}\left(\tfrac{\xi}{1-\xi}\right)^{-1} \mathbf{V}' \,\frac{\mathrm{d}\xi}{(1-\xi)^2}. \tag{12}$$

The integrand exists also in the limit case $\xi \to 1$. Hence the integrand is an analytic function in [0, 1].

We use a quadrature rule with nodes $\xi_j \in [0, 1]$ and positive weights γ_j for $j = 1, \ldots, q$. For example, Gauss-Legendre quadrature, trapezoidal rule or Simpson rule can be employed. The approximation of the integral (12) becomes the finite sum

$$\mathbf{U} \approx \sum_{j=1}^{q} \frac{\gamma_j}{(1-\xi_j)^2}\, \mathbf{S}\left(\tfrac{\xi_j}{1-\xi_j}\right)^{-\mathsf{H}} \mathbf{S}\left(\tfrac{\xi_j}{1-\xi_j}\right)^{-1} \mathbf{V}'. \tag{13}$$

Each term of the sum requires to solve complex-valued linear systems, where r right-hand sides appear with identical coefficient matrices. Just a single complex-valued LU-decomposition of a sparse matrix (10) has to be computed for each node of the quadrature rule. Thus we perform q LU-decompositions and $4qr$ forward/backward substitutions in total.

5 Method for Differential-Algebraic Equations

In the case of linear DAEs (1), the Lyapunov equation (7) has no solution for positive definite $\mathbf{F}$. We adopt a regularisation technique used by Müller [10], where the matrices change into

$$\mathbf{E}' = \mathbf{E} - \beta^2 \mathbf{A} \qquad \text{and} \qquad \mathbf{A}' = \mathbf{A} + \beta \mathbf{E} \tag{14}$$

with a regularisation parameter $\beta > 0$. This technique represents a singular perturbation. The linear dynamical system becomes an ODE, which is asymptotically stable for all $\beta \in (0, \beta_{\max})$ with some $\beta_{\max} > 0$.

Now we apply the stability-preserving technique to the linear dynamical system (1) including the matrices (14) for a fixed parameter β. The integral (9) is finite for all $\beta \in (0, \beta_{\max})$, whereas the integral does not exist in the limit case $\beta = 0$.

If a linear system of DAEs exhibits a transfer function with a finite $\mathcal{H}_2$-norm (4), then the reduction error can be measured in this norm. The existence of the norm (4) also depends on the definition of inputs and outputs in the system. The total error of the MOR is bounded by

$$\|\mathbf{H}_{\mathrm{DAE}} - \mathbf{H}_{\mathrm{ROM}}\|_{\mathcal{H}_2} \leq \|\mathbf{H}_{\mathrm{DAE}} - \mathbf{H}_{\mathrm{ODE}}\|_{\mathcal{H}_2} + \|\mathbf{H}_{\mathrm{ODE}} - \mathbf{H}_{\mathrm{ROM}}\|_{\mathcal{H}_2} . \qquad (15)$$

The first term converges to zero for $\beta \to 0$. The magnitude of the second term depends on the quality of the MOR for the regularised system. A detailed error analysis is given for the regularisation in [15].

Furthermore, a linear system of DAEs with (nilpotency) index one always has a transfer function with a finite $\mathcal{H}_\infty$-norm. There is a potential for the quantification of the regularisation error by a restriction to a compact frequency interval $[-\omega_{\max}, \omega_{\max}]$.

6 Numerical Results for Test Example

We consider the electric circuit of a band-pass filter depicted in Fig. 1. A single input voltage is supplied and a single output voltage is observed. Modified nodal analysis [7] yields a linear system of DAEs with (nilpotency) index one and dimension $n^* = 23$. In an uncertainty quantification, we replace all physical parameters (capacitances, inductances, resistances) by random variables with independent uniform probability distributions, which vary 20% around their mean values. We use a polynomial chaos expansion of degree two, which includes $m = 300$ orthogonal basis polynomials, see [17]. The stochastic Galerkin method generates a larger coupled system of the form (1) with dimension $n = mn^*$. This system is an asymptotically stable DAE of index one again. Furthermore, the DAE's transfer

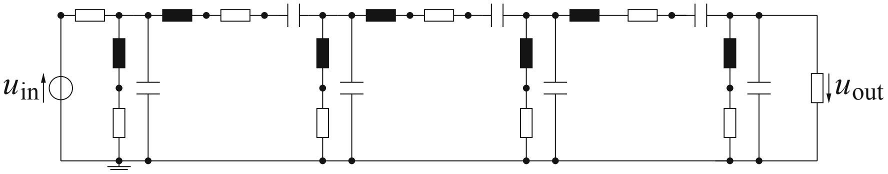


Fig. 1 Electric circuit of band-pass filter consisting of L-C-Π components

Table 1 Properties of linear dynamical system in stochastic Galerkin method

dimension n	6900	# non-zero elements in A	49,096
# inputs	1	# non-zero elements in E	20,906
# outputs	300	rank(E)	4200

Table 2 Number of stable ROMs within dimensions $r = 1, \ldots, 100$ for different systems

		# stable ROMs	# unstable ROMs
(i)	DAE	23	77
(ii)	ODE	51	49
(iii)	ODE, stabilisation, direct method	100	0
(iv)	ODE, stabilisation, quadrature with 10 nodes	81	19
(v)	ODE, stabilisation, quadrature with 1700 nodes	85	15

function is strictly proper, i.e., the $\mathcal{H}_2$-norm (4) is finite. Table 1 illustrates the properties of this linear dynamical system.

All numerical calculations were done in MATLAB [9] on a FUJITSU computer with processor Intel(R) Core(TM) i5-4570 CPU @ 3.20 GHz (four cores) and operating system Windows 7. In all dynamical systems, we perform an MOR by the one-sided Arnoldi method, see [1], using the single expansion point $s_0 = 10^6$.

We investigate five variants:

(i) Arnoldi method for system of DAEs.

(ii) Arnoldi method for regularised system (ODEs) including matrices (14) with parameter $\beta = 10^{-5}$.

(iii) Stabilisation of the reduction from (ii): We solve the Lyapunov equation (7) with $\mathbf{F} = \mathbf{I}_n$ by a direct linear algebra method. (computation time: 2.1 h)

(iv) Stabilisation of the reduction from (ii): We use the frequency domain approach including quadrature. The Gauss-Legendre rule with $q = 10$ nodes yields the approximation (13). (computation time: 1.4 s)

(v) Stabilisation of the reduction from (ii): Same as in (iv) with $q = 1700$ nodes (computation time: 177 s)

Table 2 shows the number of stable and unstable ROMs within $r = 1, 2, \ldots, 100$ in the five cases. The stabilisation technique using the direct method always generates stable ROMs. However, the stabilisation technique based on the frequency domain integral and quadrature is not always successful, because just more than 80% of the ROMs are stable. The number of stable ROMs increases for larger numbers of nodes. The remaining loss of stability indicates a critical behaviour of the approximate method in the case of small regularisation parameters.

We also compute reduction errors by an approximation of the relative $\mathcal{H}_2$-norms, see (4), for the differences between the transfer functions of the DAEs and the ROMs. Figures 2 and 3 illustrate the numerical results for the cases (i), (ii), (iii) and (v). We recognise that the errors decrease exponentially for increasing reduced dimensions. In the ODE variants, the error stagnates for $r > 80$, because the

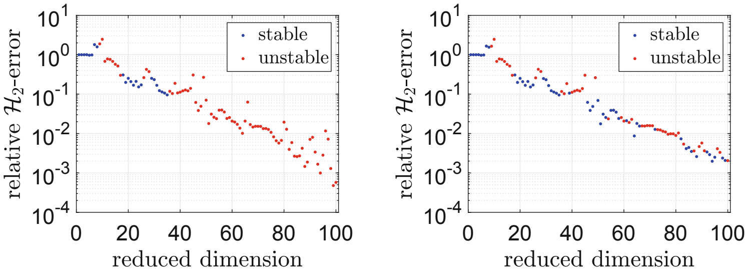


Fig. 2 Relative $\mathcal{H}_2$-errors of ROMs for dimensions $r = 1, \ldots, 100$ associated to reduction of the original DAEs (left) and ODEs from regularisation (right)

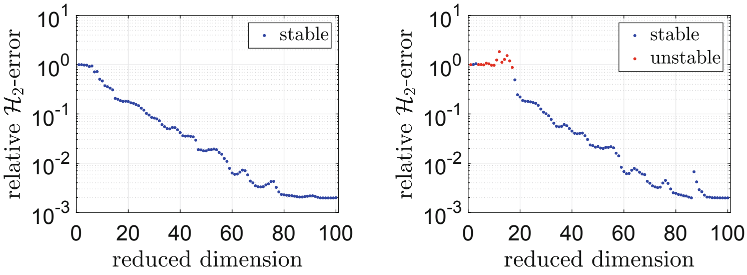


Fig. 3 Relative $\mathcal{H}_2$-errors of ROMs for dimensions $r = 1, \ldots, 100$ obtained from stabilisation technique using direct method (left) and quadrature rule with 1700 nodes (right)

regularisation error dominates the total error in agreement to the estimate (15). We also address where the loss of stability appears in the variant (v). Only ROMs of small dimensions $r \leq 17$ become unstable, which are not relevant due to a large error for small dimensions in all variants. In comparison to the case (ii), the approximate procedures (iv) and (v) are superior with respect to the generation of stable ROMs.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia (2005)
2. Bai, Z., Freund, R.: A partial Padé-via-Lanczos method for reduced order modeling. Linear Algebra Appl. **332–334**, 139–164 (2001)
3. Benner, P., Stykel, T.: Model order reduction of differential-algebraic equations: a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations IV. Differential-Algebraic Equations Forum, pp. 107–160. Springer, Cham (2017)
4. Benner, P., Hinze, M., ter Maten, E.J.W. (eds.): Model Reduction for Circuit Simulation. Springer, Dordrecht (2011)
5. Castañé Selga, R., Lohmann, B., Eid, R.: Stability preservation in projection-based model order reduction of large scale systems. Eur. J. Control **18**, 122–132 (2012)

6. Gugercin, S., Antoulas, A.C., Beattie, C.: H_2 model reduction for large-scale linear dynamical systems. SIAM J. Matrix Anal. Appl. **30**, 609–638 (2008)
7. Ho, C.W., Ruehli, A., Brennan, P.: The modified nodal approach to network analysis. IEEE Trans. Circuits Syst. **22**, 504–509 (1975)
8. Ionescu, T.C., Astolfi, A.: On moment matching with preservation of passivity and stability. In: 49th IEEE Conference on Decision and Control, pp. 6189–6194 (2010)
9. MATLAB, version 9.4.0.813654 (R2018a). The Mathworks Inc., Natick, MA (2018)
10. Müller, P.C.: Modified Lyapunov equations for LTI descriptor systems. J. Braz. Soc. Mech. Sci. Eng. **28**, 448–452 (2006)
11. Penzl, T.: Numerical solution of generalized Lyapunov equations. Adv. Comput. Math. **8**, 33–48 (1998)
12. Phillips, J.R., Silveira, L.M.: Poor man's TBR: a simple model reduction scheme. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **24**, 43–55 (2005)
13. Pulch, R.: Quadrature methods and model order reduction for sparse approximations in random linear dynamical systems. In: Langer, U., Amrhein, W., Zulehner, W. (eds.) Scientific Computing in Electrical Engineering SCEE 2016. *Mathematics in Industry*, vol. 28, pp. 203–217. Springer, Cham (2018)
14. Pulch, R.: Stability preservation in Galerkin-type projection-based model order reduction. Numer. Algebra Contr. Optim. **9**, 23–44 (2019)
15. Pulch, R.: Frequency domain integrals for stability preservation in Galerkin-type projection-based model order reduction. Int. J. Control (2019). https://doi.org/10.1080/00207179.2019.1670360
16. Schilders, W.H.A., van der Vorst, M.A., Rommes, J. (eds.): Model Order Reduction: Theory, Research Aspects and Applications. Mathematics in Industry, vol. 13. Springer, New York (2008)
17. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)

Index

G. Nicosia, V. Romano (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 32, https://doi.org/10.1007/978-3-030-44101-2