Mathematics in Industry 36
The European Consortium for Mathematics in Industry

Martijn van Beurden
Neil Budko
Wil Schilders *Editors*

Scientific Computing in Electrical Engineering

SCEE 2020, Eindhoven, The Netherlands, February 2020

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Springer

MATHEMATICS IN INDUSTRY **36**

THE EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

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 series at http://www.springer.com/series/4651

Martijn van Beurden • Neil Budko • Wil Schilders
Editors

Scientific Computing in Electrical Engineering

SCEE 2020, Eindhoven, The Netherlands, February 2020

Springer

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Editors
Martijn van Beurden
Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands

Neil Budko
Department of Applied Mathematics
Delft University of Technology
Delft, The Netherlands

Wil Schilders
Eindhoven University of Technology
Eindhoven, The Netherlands

ISSN 1612-3956 ISSN 2198-3283 (electronic)
Mathematics in Industry
The European Consortium for Mathematics in Industry
ISBN 978-3-030-84237-6 ISBN 978-3-030-84238-3 (eBook)
https://doi.org/10.1007/978-3-030-84238-3

Mathematics Subject Classification: 65Lxx, 65-06, 65Mxx, 65Nxx, 65Zxx, 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

From February 16 until February 20, 2020, the 13th International Conference on "Scientific Computing in Electrical Engineering" (SCEE) was held in Eindhoven, The Netherlands. It was jointly organized by the Centre for Analysis, Scientific Computing and Analysis (CASA) and the Electromagnetics group of Eindhoven University of Technology, and the group Numerical Analysis of Delft University of Technology.

Even though 13 is a number often associated with bad luck, this edition was actually very fortunate. Already prior to and during the conference, the world was discussing the SARS-CoV-2virus and associated problems and measures, and not long after the conference ended, there was a lockdown in many countries. We are very happy that SCEE 2020 took place, not in a virtual way, but with many face-to-face contacts, meeting our esteemed colleagues once again, having lunches and dinner together in an excellent location, the "Academisch Genootschap Eindhoven." Participants enjoyed the setting and the surroundings, as well as the opportunity to sit in the garden and discuss.

The thirteenth edition of the SCEE conference brought together some 85 participants from the fields of applied mathematics, electrical and electronic engineering, and the computer sciences as well as scientists from industry. Again, it created an excellent working atmosphere, especially due to its unique workshop character, where all talks and poster introductions were presented in plenary sessions. In addition, we had very clear and high-quality talks and poster presentations, lively and fruitful discussions, and a great deal of personal networking.

The Scientific Program Committee invited four experts to give keynote presentations on the main topics in the regular program. Keynote speakers at SCEE 2020 were (in alphabetical order):

- Liliane Borcea (University of Michigan—USA), "Reduced order model approach for inverse scattering"
- Romanus Dyczij-Edlinger (Universty of Saarland—Germany), "Reduced-order finite-element modeling and optimization of antennas"

Participants of SCEE 2020 in the garden of Academisch Genootschap Eindhoven

- Slawomir Koziel (Reykjavik University—Iceland), "Forward and Inverse Surrogate Modeling for Accelerated Design Optimization of High-Frequency Structures"
- Jasmin Smajic (ETH Zürich—Switzerland), "Numerical Analysis of Electromagnetic Transients in Power Devices"

We also re-installed a tradition, a Sunday evening speaker: Albert Ruehli (IEEE life fellow, 50 years IBM, currently at Missouri University of Science and Technology—USA) entertained us with the very nice overview talk entitled "Retrospective: 50 years of circuit and electromagnetic solutions." Another feature of this conference was the Industry Morning, where three renowned speakers from industry gave very nice presentations on urgent topics within the electronics industry:

- Rick Janssen (NXP Semiconductors), "Electromagnetic Simulation Challenges in the Semiconductor Industry"
- Frank Buijnsters (ASML), "Efficient Maxwell Solvers for Optical Semiconductor Metrology"
- Stefan Kurz (Bosch), "Newton-Kepler-Bosch: Towards the Next Level of Scientific Computing in Engineering"

The topics of the invited and industry speakers were representative of the conference's range. In addition, from Monday to Thursday, we had a total of 30 oral presentations and 30 poster presentations. The dense conference program was completed with two special sessions: a meeting of the European project (Marie-

Skłodowska-Curie EID) ROMSOC, and a meeting of the recently established ECMI Special Interest Group MSOEE. On Wednesday evening, the SCEE Standing Committee, the Program Committee, and the Local Organizing Committee also had a meeting, followed by a lavish dinner with the invited speakers. A special highlight of the SCEE 2020 conference was our conference excursion to either the Philips museum or the DAF museum. Both companies have been instrumental in building up the city of Eindhoven, and the collections of these museums were enjoyed by the participants. After this excursion, the conference dinner took place in the Academisch Genootschap Eindhoven, where many ideas and new research directions were discussed in parallel to the enjoyment of good food and wine.

The present book collects the conference outcomes as proceedings papers. All these papers have successfully passed a standard peer-review process. The contributions are divided into four parts, which reflect the main focus areas of SCEE 2020 ("coupled problems," also a traditional focus area, has been put under "Mathematical and Computational Methods"):

- Circuit Simulation and Design
- Device Simulation
- Computational Electromagnetics
- Mathematical and Computational Methods

In retrospect, we feel we have compiled a very successful and interesting collection. We wish to thank all the participants for their valued contributions to the SCEE 2020 conference and to this book, and we hope we will meet each other at future SCEE conferences!

Eindhoven, The Netherlands — Martijn van Beurden
Delft, The Netherlands — Neil Budko
Eindhoven, The Netherlands — Wil Schilders
October 2020

Acknowledgment

We would like the thank Eindhoven University of Technology, viz. the Centre for Analysis, Scientific Computing and Applications (CASA) within the Department of Mathematics and Computer Science and the Electromagnetics (EM) group within the Department of Electrical Engineering, and Delft University of Technology, Department of Applied Mathematics (DIAM), for their help and support in the organization of the SCEE 2020 Conference.

We are also grateful for the financial support by the Applied Mathematics Institute of the four Universities of Technology in The Netherlands (4TU.AMI), the mathematics cluster NDNS+ (Nonlinear Dynamics of Natural Systems), the Dutch National Organisation for Research (NWO), the European Marie-Curie-Skłodowska Industrial Doctorate Project ROMSOC (Reduced Order Modelling, Simulation and Optimization of Coupled Systems), and CST—Computer Simulation Technology AG in Darmstadt, part of Dassault Systèmes.

Last but not least, we would like to thank all the members of the Local Organizing Committee and the Scientific Committee who helped us very much in preparing and running the conference. The careful reviewing process was only possible with the help of the members of the Scientific Committee who were handling the reviewing process. The anonymous referees did a wonderful job that helped the authors to improve the quality of their contributions.

Finally, we express our gratitude to our colleagues from Springer Heidelberg for continued support and patience during the preparation of this volume.

Organization

Local Organizing Committee

Wil Schilders (Chair, TU Eindhoven, The Netherlands)
Martijn van Beurden (TU Eindhoven, The Netherlands)
Neil Budko (TU Delft, The Netherlands)
Enna van Dijk (TU Eindhoven, The Netherlands)

Program Committee

Wil Schilders (Chair, TU Eindhoven, The Netherlands)
Martijn van Beurden (TU Eindhoven, The Netherlands)
Neil Budko (TU Delft, The Netherlands)
Gabriela Ciuprina (Politehnica University of Bucharest, Romania)
Georg Denk (Infineon, Germany)
Herbert de Gersem (TU Darmstadt, Germany)
Michael Günther (University of Wuppertal, Germany)
Stefan Kurz (Bosch, Germany)
Ulrich Langer (Johannes Kepler University Linz, Austria)
Jan ter Maten (University of Wuppertal, Germany)
Jörg Ostrowski (ABB, Switzerland)
Ursula van Rienen (University of Rostock, Germany)
Vittorio Romano (University of Catania, Italy)
Ruth Vazquez Sabariego (KU Leuven, Belgium)
Sebastian Schöps (TU Darmstadt, Germany)
Caren Tischendorf (Humboldt University of Berlin, Germany)

Standing Committee

Ursula van Rienen (Chair, University of Rostock, Germany)
Gabriela Ciuprina (Secretary, Politehnica University of Bucharest, Romania)
Michael Günther (Treasurer, University of Wuppertal, Germany)
Jörg Ostrowski (ABB, Switzerland)
Wil Schilders (TU Eindhoven, The Netherlands)

Sponsors

TU Eindhoven
TU Delft
NWO
CST
4TU.AMI
NDNS+
ECMI

Contents

Part I
Circuit Simulation and Design

Efficient Model Reduction of Myelinated Compartments as Port-Hamiltonian Systems

Ruxandra Barbulescu, Gabriela Ciuprina, Tudor Ionescu, Daniel Ioan, and Luis Miguel Silveira

Abstract The information is transmitted in neurons through axons, many of whom have myelin-covered sections, whose main purpose is to increase the speed of electrical signal transmission. Modeling the myelinated axons in a realistic way, by maintaining the physical meaning of components may lead to complex systems, described by high-dimensional systems of PDEs, whose solution is computationally demanding. Analysis of larger neuronal circuits including multiple myelinated axons therefore requires the generation of equivalent low-order models to control complexity. Such models must preserve the physical interpretation and properties of the original system including its passivity and stability. The axons' port-based structure makes them suitable to be modeled as port-Hamiltonian systems. This paper uses a structure-preserving reduction method for port-Hamiltonian systems to reduce the description of a myelinated compartment into a model with comparable accuracy with the previously used vector fitting technique. The reduced system is synthesized into an equivalent passive circuit with no controlled sources and only positive elements, amenable for inclusion in standard neuronal simulators.

1 Introduction

A myelinated axon (Fig. 1) consists of myelinated sections through which the signal dissipates, which alternate with Ranvier nodes where the signal is regenerated ("saltatory conduction"). To model the transmission of signals through this chain, the phenomena occurring in the Ranvier *nodes* have to be coupled with those in the myelinated sections (*internodes*). The underlying mechanisms of a Ranvier node are well described by the Hodgkin & Huxley model [10], or its reduced versions [19].

R. Barbulescu (✉) · L. M. Silveira
INESC ID/IST Tecnico Lisboa, Universidade de Lisboa, Lisboa, Portugal
e-mail: ruxi@inesc-id.pt; lms@inesc-id.pt

G. Ciuprina · T. Ionescu · D. Ioan
Politehnica University of Bucharest, Bucharest, Romania
e-mail: gabriela@lmn.pub.ro; tudor.ionescu@acse.pub.ro; daniel@lmn.pub.ro

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_1

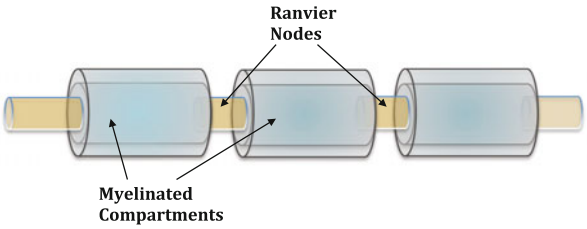


Fig. 1 Simplified geometrical model of a myelinated axon, as a chain of internode myelinated compartments and Ranvier nodes

The most popular approach to model the internodes (myelinated compartments) is the "cable model", described by parabolic 1D PDEs [13], i.e., the RC transmission line equation. In a previous work [11], the authors reduced the internode model with different methods, resulting in a hierarchical series of models of three spatial geometry classes: 2.5D, 1D and 0D and three categories of models: analytical, numerical and reduced order models. The analytical 1D model reduced with the vector fitting (VF) technique proved to be the most accurate according to our weighted error metric, as shown in the Results section. In [11] the error is computed using a weighted norm, where the weights associated to frequencies are extracted from the spectrum of the standard neuronal signal. This error is suitable to estimate the global accuracy of neuronal signals, since in the typical neuronal spectrum the low frequency components (up to a few hundred Hz) are much more significant than the high frequency ones. The accuracy of the current reduction method is compared with the results in [11].

For the simulation of the saltatory conduction in a whole axon, the internodes were replaced in [12] with the differential equation macromodel extracted from VF, so the equivalent circuit had many controlled sources. This is acceptable when there is no constraint on the reduced circuit, but in some environments dedicated to neuronal simulations, such as NEURON [9], one can only create a circuit with no controlled sources (or a small amount of controlled sources, modeled using Op-Amps) and with positive parameters. In this work we synthesize the reduced system into an equivalent circuit with only positive RCs and no controlled sources (we call this circuit **ECi+**). We start from the large discretized transmission line lumped RC model as a substitute of the PDEs (Fig. 2). All elements in this model are linear and frequency independent. The long network of RC sections has resistive parameters describing longitudinal electrical conduction phenomena through axoplasm, and capacitive and transverse conductive effects through the cell membrane.

This particular model of a myelinated compartment, as a chain of RC cells, is suitable for port-based network modeling, as in the port-Hamiltonian (pH) framework. The pH systems are widely used in modeling, analysis and control of (multi-)physical systems [5, 21]. Extensive research has been done on model order reduction targeting preservation of relevant properties and/or port-Hamiltonian structure for linear [2, 8, 15] and nonlinear systems [4, 17, 20]. Among these techniques, the moment-matching procedure is an efficient tool [1, 2, 16]. The reduced model is obtained by constructing a lower degree rational function that approximates the given transfer function and matches it at various interpolation

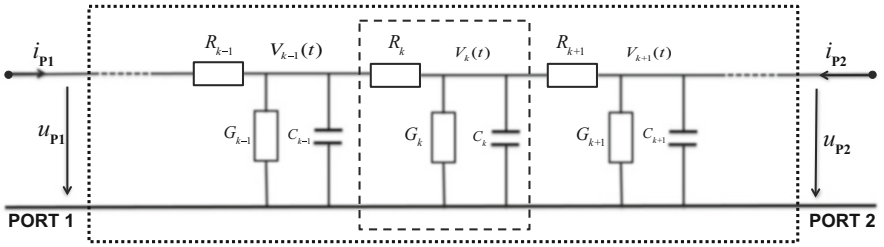


Fig. 2 The companion circuit of an internode. The network of RC (in our case identical) cells is generated by the spatial discretization with centered differences of the transmission line equation. The outer dotted line outlines the system and the inner dashed line denotes one individual RC cell

points in the complex plane. This formulation is preferred to the direct reduction of the number of cells of the segmented numerical model, which already is a reduced model.

2 Port-Hamiltonian Formulation and Reduction

The pH representation is based on the energy state space, which represents a natural state space for the equations composing the mathematical models of physical systems. The Hamiltonian gives the total stored energy of the system, whereas the system has boundary ports to interact with the environment, through the exchange of energy. The mathematical representation of a pH system is

$$\begin{cases} \dot{\mathbf{x}} = (\mathbf{J} - \mathbf{R})\nabla_{\mathrm{x}}\mathrm{H}(\mathrm{x}) + \mathbf{B}u(t) \\ \mathbf{y} = \mathbf{B}^{\mathrm{T}}\nabla_{\mathrm{x}}\mathrm{H}(\mathrm{x}) \end{cases} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the state vector; $\mathrm{H} : \mathbb{R}^n \to [0, \infty]$ is continuously differentiable – the Hamiltonian, describing the internal energy of the system as a function of state; $\mathbf{J} = -\mathbf{J}^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ is the structure matrix (skew-symmetric) describing the interconnection of energy storage elements in the system; $\mathbf{R} = \mathbf{R}^{\mathrm{T}} \geq 0$ is the dissipation matrix describing the energy loss in the system; and $\mathbf{B} \in \mathbb{R}^{n \times m}$ is the port matrix describing how energy enters and exits the system through the m terminals/ports (here $m = 2$).

Our approach is based on describing the myelinated compartment in Fig. 2 as a pH system (1) and reducing the overall model with structure-preserving moment-matching. We start from the circuit description of the original model (a SPICE netlist) and generate the pH form of this system. Next, the system is reduced by moment-matching. Finally, the equivalent reduced circuit is synthesized from the state-space representation of the reduced system.

We consider the network in Fig. 2 as a 2x2 system with input $u = \left[\frac{u_{\mathrm{P1}}(t)}{R_1} \; i_{\mathrm{P2}}(t)\right]^{\mathrm{T}}$ and output $y = [V_1(t) \; u_{\mathrm{P2}}(t)]^{\mathrm{T}}$. The state space vector consists of the charges of the capacitors $\mathbf{x} = [q_1, q_2, \ldots, q_n]^{\mathrm{T}}$, thus its derivative $\dot{\mathbf{x}} = \left[i_{C_1}, i_{C_2}, \ldots, i_{C_n}\right]^{\mathrm{T}}$ is composed of the currents through the capacitors. The Hamiltonian is defined as $\mathrm{H(x)} = \frac{1}{2}\sum_{k=1}^{n}\frac{1}{C_k}q_k^2 = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}$ and its derivative with respect to the state variables is a vector of voltages: $\nabla_{\mathrm{x}}\mathrm{H(x)} = \left[u_{C_1}, u_{C_2}, \ldots, u_{C_n}\right]^{\mathrm{T}} = \mathbf{Q}\mathbf{x}$.

In this formulation $\mathbf{Q}$ is a diagonal matrix, $\mathbf{Q} = \mathrm{diag}\left(\frac{1}{C_k}\right)$, the structure matrix $\mathbf{J} = \mathbf{0}$, and the dissipative matrix $\mathbf{R}$ is tridiagonal, having on line k the elements $-\frac{1}{R_k}$, $\frac{1}{R_k} + \frac{1}{R_{k+1}} + G_k$ and $-\frac{1}{R_{k+1}}$. The port matrix $\mathbf{B} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 \end{bmatrix}^{\mathrm{T}}$. Because $\mathbf{R}$ is positive definite and the capacitors are identical, the product $\mathbf{RQ}$ is positive definite.

3 System Reduction

The reduction is based on a moment-matching technique, part of the family of interpolatory methods [1]. A linear SISO system described by the state matrices $(\mathscr{A}, \mathscr{B}, \mathscr{C})$ has the transfer function $K(s) = \mathscr{C}(s\mathbf{I} - \mathscr{A})^{-1}\mathscr{B}$, $K : \mathbb{C} \to \mathbb{C}$. Consider a point in the complex plane that is not in the spectrum of $\mathscr{A}$. The k-order moment of the system with the transfer function K at $s^* \in \mathbb{C} - \sigma(\mathscr{A})$ is formally defined as:

$$\eta_k(s^*) = \frac{(-1)^k}{k!}\left[\frac{\mathrm{d}^k K(s)}{\mathrm{d}s^k}\right]_{s=s^*}. \tag{2}$$

For a fixed point s^*, a reduced-order system described by the transfer function $\hat{K}$ with the corresponding moments $\hat{\eta}_k(s^*)$ matches the first n^* moments of K if $\eta_k(s^*) = \hat{\eta}_k(s^*), k = \overline{1, n^*}$, which in fact means it matches the coefficients of n^* terms of the Taylor expansion of K [14]. One can either choose one point s^* and match the first moments of the two transfer functions, or choose a set of points and match the 0-order moment in all the points in the set. The selection of the interpolation points is important. Whereas selecting n^* moments at a fixed s^* may improve the approximation accuracy locally, selecting $s_1, \ldots, s_r$ points for a reduced order r and matching the 0-order moments at these points better preserves input-output behaviours (here, the * notation was dropped for readability). Customary, $s_1 = 0$ is chosen to preserve the step response of the given system.

For SISO systems the interpolation conditions are enforced pointwise, but in the MIMO case – where $\mathbf{K}(s)$ is a $m \times m$ matrix-valued rational function – full matrix interpolation would translate into $m \times m$ conditions at every interpolation point. This would result in an actual larger order of the reduced system than the

initially imposed r. Instead we only interpolate along certain directions $\mathbf{b_k}$ ("right tangential interpolation"). This relaxed notion of interpolation is adequate for an optimal approximation [7]. For an imposed reduced order r we compute the matrix

$$\mathbf{\Pi} = \left[(s_1\mathbf{I} - \mathscr{A})^{-1}\mathscr{B}\mathbf{b}_1 \;\; (s_2\mathbf{I} - \mathscr{A})^{-1}\mathscr{B}\mathbf{b}_2 \;\; \ldots \;\; (s_r\mathbf{I} - \mathscr{A})^{-1}\mathscr{B}\mathbf{b}_r \right],$$

where the vectors $\mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b_r}$ represent the tangential directions of interpolation.

Since the interpolation points s_k and the tangential directions $\mathbf{b_k}$ are dependent on the reduced model, we use an iterative process to correct the interpolation points and tangential directions until the interpolation conditions are met [4] .

The reduced matrices are computed as in [2]:

$$\mathbf{J}_r = \mathbf{\Pi}^{\mathrm{T}}\mathbf{Q}\mathbf{J}\mathbf{Q}\mathbf{\Pi} \qquad \mathbf{Q}_r = \left(\mathbf{\Pi}^{\mathrm{T}}\mathbf{Q}\mathbf{\Pi}\right)^{-1}$$

$$\mathbf{R}_r = \mathbf{\Pi}^{\mathrm{T}}\mathbf{Q}\mathbf{R}\mathbf{Q}\mathbf{\Pi} \qquad \mathbf{B}_r = \mathbf{\Pi}^{\mathrm{T}}\mathbf{Q}\mathbf{B}$$

and they are used to construct the reduced system in the port-Hamiltonian form:

$$\begin{cases} \dot{\mathbf{x}}_r = (\mathbf{J}_r - \mathbf{R}_r)\mathbf{Q}_r\mathbf{x}_r + \mathbf{B}_r u(t) \\ \mathbf{y} = \mathbf{B}_r^{\mathrm{T}}\mathbf{Q}_r\mathbf{x}_r \end{cases} . \tag{3}$$

Such reduced order system matches the 0-order moments of the original system at the chosen interpolation points [3]. The reduction procedure is structure-preserving, in the sense that the reduced system is still in the port-Hamiltonian form, but the matrices have lost some of their properties, for instance $\mathbf{Q}_r$ is not diagonal anymore, $\mathbf{R}_r$ is not tridiagonal, but is still symmetric and positive definite, $\mathbf{B}_r$ is now likely full.

4 Synthesis of Equivalent Reduced Circuit

There is extensive research on circuit realization of systems, either by direct interpretation of the mathematical model or from the state-space form or the system's transfer function [18]. However, in most approaches the resulting circuit is not guaranteed to contain only physically-meaningful elements, due to the presence of negative R, L or C elements or it has a large number of controlled sources.

The transfer function of the reduced system is actually a 2×2 symmetrical matrix of impedances. A possible circuit realization for this is a star (Fig. 3), where the impedances of the subcircuits result directly from either the transfer function components or the state-space matrices of the reduced system.

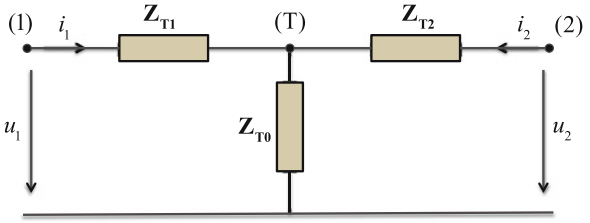


Fig. 3 The circuit realization T scheme for a 2×2 system

Each impedance $\mathbf{Z}_{T0}$, $\mathbf{Z}_{T1}$ and $\mathbf{Z}_{T2}$ can be realized through a pole-residue decomposition as the sum of the impedances of r cells connected in series, each composed of a capacitor in parallel with a conductance: $\mathbf{Z}_{xx} = \sum_{k=1}^{r} 1/(C_k s + G_k)$.

The reduced system (3) can be viewed in the form of a standard description of an RC circuit

$$\begin{cases} \mathbf{C}\dot{\mathbf{x}}_r = -\mathbf{G}\mathbf{x}_r + \mathbf{B}u(t) \\ \mathbf{y} = \mathbf{E}\mathbf{x}_r \end{cases}$$

where $\mathbf{C}$, $\mathbf{G}$, $\mathbf{B}$ and $\mathbf{E}$ are defined accordingly.

The reduced states are the capacitors' voltages in the reduced circuit. To simplify realization, each state should be involved in only one equation. To that end, matrices $\mathbf{C}$ and $\mathbf{G}$ are diagonalized to allow the equations to be separated. Their diagonalization impacts the matrices $\mathbf{B}$ and $\mathbf{E}$, which become full (likely already the case here). In the reduced system this would translate into the circuit as controlled sources. To avoid that, the two matrices are scaled so that all their values are either 1 or -1 and consequently the outputs will be algebraic sums of all the states.

The computations lead to the following relations for the components $\mathbf{Z}_{xx}$, where c_{kk} and g_{kk} ($k = \overline{1, r}$) are the diagonal values of $\mathbf{C}$ and $\mathbf{G}$ (after diagonalization) and the denominator actually represents the scaling of $\mathbf{B}$ and $\mathbf{E}$:

$$\mathbf{Z_{T0}}: \begin{cases} C_k = \frac{c_{kk}}{e_{1k} b_{k2}} \\ G_k = \frac{g_{kk}}{e_{1k} b_{k2}} \end{cases} \quad \mathbf{Z_{T1}}: \begin{cases} C_k = \frac{c_{kk}}{e_{1k}(b_{k1} - b_{k2})} \\ G_k = \frac{g_{kk}}{e_{1k}(b_{k1} - b_{k2})} \end{cases} \quad \mathbf{Z_{T2}}: \begin{cases} C_k = \frac{c_{kk}}{e_{2k}(b_{k2} - b_{k1})} \\ G_k = \frac{g_{kk}}{e_{2k}(b_{k2} - b_{k1})} \end{cases}$$

In theory the capacitance C_k and the conductance G_k of a cell may have any sign. But the CG pair signs differ only by the signs of the diagonal values of the matrices $\mathbf{C}$ and $\mathbf{G}$. Here $\mathbf{C}$ is the identity matrix, so clearly positive definite. $\mathbf{G}$ is a diagonal matrix that comes from the original system matrix $\mathbf{RQ}$, which is positive definite. Since the reduction procedure guarantees passivity, it will preserve the definiteness of the system matrix. Hence $\mathbf{G}$ has only positive values on the diagonal. This means that for every cell, C_k and G_k are either both positive or both negative.

Consider the synthesized circuit of $\mathbf{Z}_{xx}$ as in Fig. 4 (left), where the first two cells have positive values and the third has negative values. In Fig. 4 (right) the circuit is split into the "positive" and the "negative" contributions [18]. For the negative subcircuit the signs for both C_k and G_k are reversed and the same excitation is used

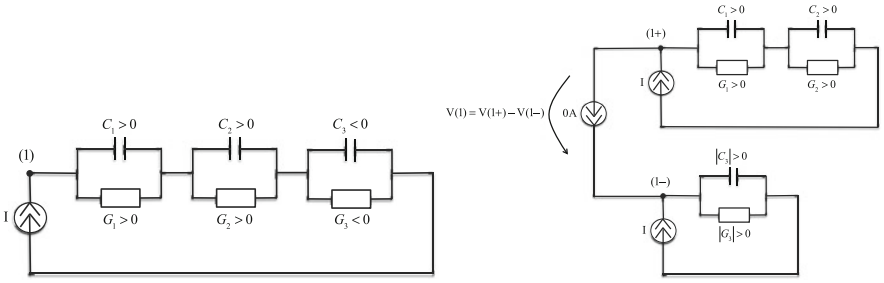


Fig. 4 Synthesis of a component $\mathbf{Z}_{\mathrm{xx}}$ of order 3. (Left): The circuit with positive and negative CG pairs, $y = \mathrm{V}(1)$. (Right): The circuit split into the "positive" and "negative" subcircuits, $y = \mathrm{V}(1+) - \mathrm{V}(1-)$ extracted as the voltage of a null current source that connects the two subcircuits

for both subcircuits. The initial circuit has the same output $y = \mathrm{V}(1)$ as the circuit after splitting, computed as the difference of two voltages $y = \mathrm{V}(1+) - \mathrm{V}(1-)$.

5 Results

Figure 5 shows the frequency responses of the two components of the original (50 cells) and reduced (order 5) systems. The response of the transfer component (1,2) is very far from the original system's, but the values are so small that this graph is in fact not relevant accuracy-wise, because the reduction procedure has an implicit minimization of the H_2 norm. This is proved by almost identical step responses.

The relative error is under 2% even for order 1 and is comparable with the one obtained with vector fitting (Fig. 6 left) with the adaptive frequency sampling (AFS) procedure described in [6]. In the interest of fairness, the errors are computed for

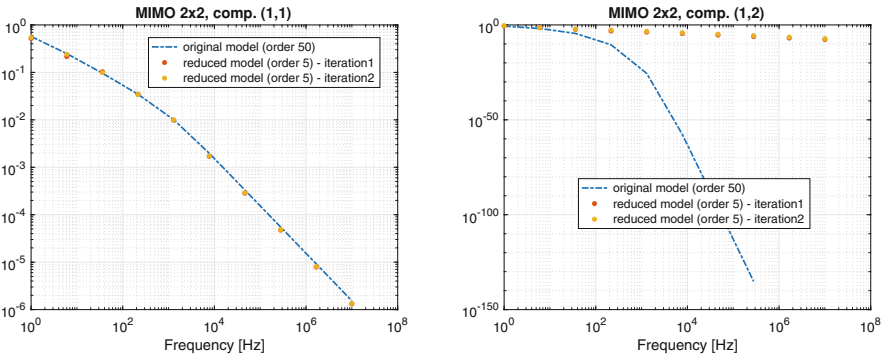


Fig. 5 The frequency responses of the original (50 cells) and reduced (order 5) systems (note the diminutive vertical scale on the graph to the right)

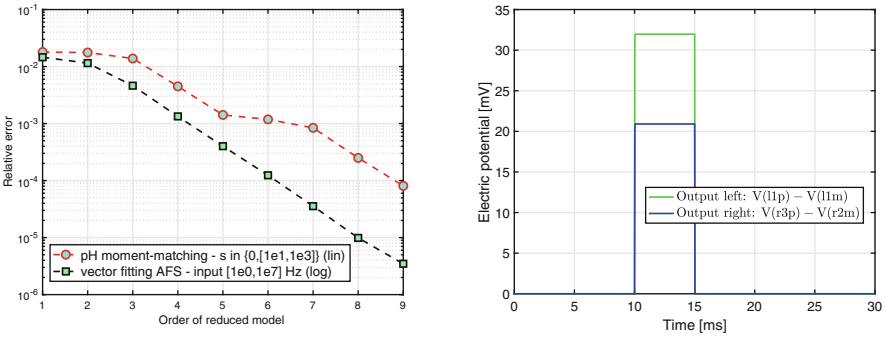


Fig. 6 (Left): The relative error vs. the order of the reduced model for pH and VF methods; $\text{err}_{\text{rel}} = \int_{f_m}^{f_M} w(f) \left\| \text{Z}_{\text{orig}}(f) - \text{Z}_{\text{red}}(f) \right\|_2 \text{d}f/\text{Z}_0$, frequency $f \in [f_m, f_M] = [10^0, 10^7]$ Hz, logarithmically spaced, $w(f)$ is the weight function, Z_0 is the d.c. impedance of the line [11]. (Right): The reproduced output of the reduced circuit (order 3) built in NEURON

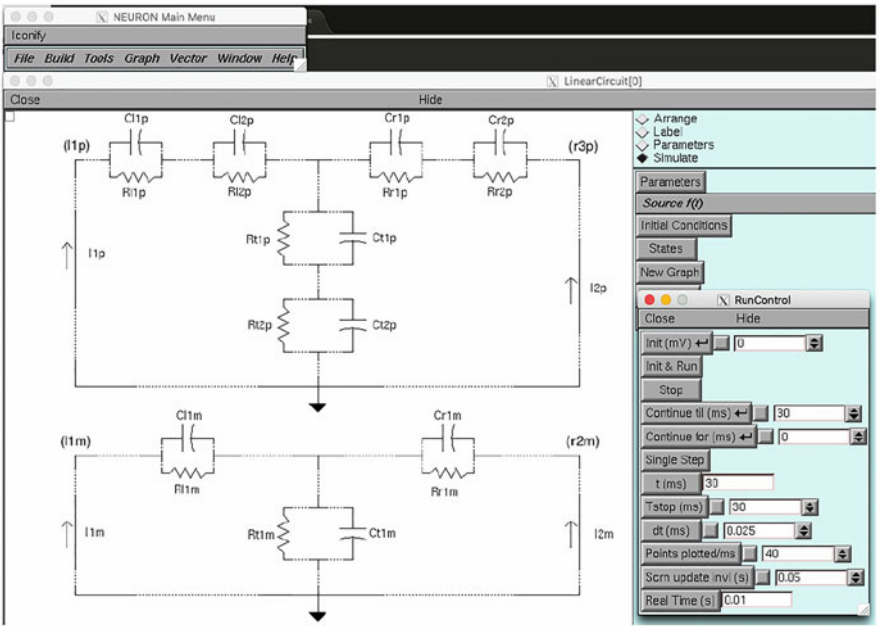


Fig. 7 The **ECi+** circuit extracted from the reduced system of order 3 and reproduced in NEURON. The output is [V(l1p) − V(l1m); V(r3p) − V(r2m)]

the best set of conditions for each of the two methods. The new method is not meant to improve accuracy but to generate a model without controlled sources and with positive elements, which is a requirement for the inclusion in NEURON.

The reduced circuit of order 3 is built in NEURON (Fig. 7) and the output is reproduced in Fig. 6 (right). The input is a rectangular pulse in the left ($i_1 = \text{I1p} =$

I1m) and open-circuit in the right ($i_2 = \text{I2p} = \text{I2m} = 0$). The corresponding outputs copy the shape of the input and the relative difference between the corresponding peaks of the original circuit (50 cells) and the reduced one is between 2% and 3%.

6 Conclusions

This paper uses a structure-preserving reduction method for pH systems to reduce a myelinated compartment in the model of a neuron. The automatic procedure starts from the netlist of the original model and generates its port-Hamiltonian form. The pH system is reduced using an interpolatory method through moment-matching, resulting in a reduced system that is still port-Hamiltonian, therefore preserving the passivity and the stability of the original model. The relative error is acceptable even for order 1 (less than 2%).

This procedure allows for a trade-off between a good approximation error and the desired structure preservation. The choice of interpolation points is a degree of freedom to be used for potentially improved accuracy in the moment matching reduction.

The state-space representation of the reduced system is subsequently synthesized into an equivalent circuit with no controlled sources and only positive RLCs (a **ECi+** circuit). This circuit can be used in neuronal simulators such as NEURON and further integrated into larger models. The current method will further prove beneficial for the reduction of the entire myelinated axon, with the nonlinear HH model of a Ranvier node included.

Acknowledgments This work was partially supported by Portuguese national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 as well as project PTDC/EEI-EEE/31140/2017.

References

1. A.C. Antoulas, Approximation of large-scale dynamical systems, in *SIAM*, vol. 6 (2005)
2. A. Astolfi, T.C. Ionescu, Moment matching for linear port hamiltonian systems, in *50th IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 7164–7169
3. A. Astolfi, T.C. Ionescu, Families of moment matching based, structure preserving approximations for linear port Hamiltonian systems. Automatica **49**(8), 2424–2434 (2013)
4. C. Beattie, S. Gugercin, S. Chaturantabut, Structure-preserving model reduction for nonlinear port-Hamiltonian systems. SIAM J. Sci. Comput. **38**(5), B837–B865 (2016)
5. C. Beattie, V. Mehrmann, H. Xu, H. Zwart, Linear port-Hamiltonian descriptor systems. Math. Control Signals Syst. **30**(4), 17 (2018)
6. G. Ciuprina et al., Vector fitting based adaptive frequency sampling for compact model extraction on HPC systems. IEEE Trans. Magn. **48**(2), 431–434 (2012)
7. S. Gugercin, C. Beattie, Model reduction by rational interpolation, in *Model Reduction and Approximation: Theory and Algorithms* (SIAM, New York, 2017), pp. 297–334

8. J.S. Hesthaven, B.M. Afkham, Structure-preserving model-reduction of dissipative Hamiltonian systems. J. Sci. Comput. 1–19 (2018). https://doi.org/10.1007/s10915-018-0653-6
9. M.L. Hines, N.T. Carnevale, *The NEURON book* (Cambridge University Press, Cambridge, 2006). https://neuron.yale.edu/neuron/
10. A.F. Huxley, A.L. Hodgkin, A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. **117**(4), 500–544 (1952)
11. D. Ioan, R. Bărbulescu, L.M. Silveira, G. Ciuprina, Reduced order models of myelinated axonal compartments. J. Comput. Neurosci. **47**(2–3), 141–166 (2019)
12. D. Ioan, G. Ciuprina, R. Barbulescu, Coupled macromodels for the simulation of the saltatory conduction. UPB Sci. Bull. Ser. C **18**(3) (2019). ISSN:2286-3540
13. K.A. Lindsay et al., An introduction to the principles of neuronal modelling, in *Modern Techniques in Neuroscience Research* (Springer, New York, 1999), pp. 213–306
14. D.D. Ling, I.M. Elfadel, A block rational Arnoldi algorithm for multipoint passive model-order reduction of multiport RLC networks. ICCAD **97**, 66–71 (1997)
15. R.V. Polyuyga, Model reduction of port-Hamiltonian systems. PhD thesis, University of Groningen, 2010
16. R.V. Polyuyga, A. van der Schaft, Structure preserving model reduction of port-Hamiltonian systems by moment matching at infinity. Automatica **46**(4), 665-672 (2010)
17. R.V. Polyuyga, A. van der Schaft, Effort-and flow-constraint reduction methods for structure preserving model reduction of port-Hamiltonian systems. Syst. Control Lett. **61**(3), 412-421 (2012)
18. L.M. Silveira, J.F. Villena, Circuit synthesis for guaranteed positive sparse realization of passive state-space models. IEEE Trans. Circ. Syst. I **64**(6), 1576–1587 (2017)
19. K.K. Sriperumbudur, U. van Rienen, R. Appali, 3d axonal network coupled to microelectrode arrays: a simulation model to study neuronal dynamics, in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, New York, 2015), pp. 4700–4704
20. A.J. van der Schaft, *L2-gain and Passivity Techniques in Nonlinear Control* (Springer, New York, 2000)
21. A. van der Schaft, Port-Hamiltonian systems: an introductory survey. *Proceedings of the International Congress of Mathematicians*, vol. 3, pp. 1339–1365. Citeseer, 2006

Towards a Parallel-in-Time Calculation of Time-Periodic Solutions with Unknown Period

Iryna Kulchytska-Ruchka and Sebastian Schöps

Abstract This paper presents a novel parallel-in-time algorithm able to compute time-periodic solutions of problems where the period is not given. Exploiting the idea of the multiple shooting method, the proposed approach calculates the initial values at each subinterval as well as the corresponding period iteratively. As in the Parareal method, parallelization in the time domain is performed using discretization on a two-level grid. A special linearization of the time-periodic system on the coarse grid is introduced to speed up the computations. The iterative algorithm is verified via its application to the Colpitts oscillator model.

1 Introduction

Steady-state analysis is a common task in electrical engineering, for example, during the initial design stages of, e.g., electric circuits or motors. Classical sequential time stepping may lead to lengthy transient computation particularly when the underlying dynamical system possesses a large time constant. Various approaches for efficient steady-state calculation are known from the literature. For instance, clever methods to choose the starting value [1] or an explicit error correction [15] could accelerate the time-domain calculation considerably.

A powerful tool for speeding up the classical time stepping is the class of parallel-in-time methods, such as the multigrid reduction in time [5] or Parareal [11]. Originating from the multiple shooting method [13], they are based on the splitting of the considered time interval into several windows and updating the solution at synchronization points iteratively. The use of coarse and fine discretizations propagates quickly low-frequency information of the solution using a cheap sequential solver followed by a very accurate result with a precise fine solver applied in parallel.

I. Kulchytska-Ruchka (✉) · S. Schöps
Computational Electromagnetics Group, Technical University of Darmstadt, Darmstadt, Germany
e-mail: kulchytska@temf.tu-darmstadt.de; schoeps@temf.tu-darmstadt.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_2

Another direction of obtaining the steady state is based on the solution of the joint space- and time-discrete time-periodic system formulated on the whole period [8]. There the initial and final values are coupled through the prescribed periodicity condition. An obstacle within the solution of the periodic problem in the time domain becomes the large size of the system matrix as well as its special block structure due to the interdependence of the solution vectors over the period. To deal with this difficulty a frequency domain approach was proposed in [2]. In case of linear problems, the method takes advantage of the block-cyclic matrix structure by applying the discrete Fourier transform. It fully decouples the variables, thereby allowing for the separate solution of each harmonic coefficient. This approach was further extended and incorporated into the Parareal framework by the authors in [10]. There, a simplified Newton-based iterative algorithm was presented together with its convergence analysis for the efficient treatment of nonlinear problems.

Solutions of time-periodic problems become much more challenging when the period is not given. Such situation occurs, e.g., when dealing with an autonomous system [3]. In contrast to a non-autonomous problem, the periodicity cannot be determined from the applied excitation. This paper proposes a numerical algorithm capable of determining an appropriate period automatically using parallelization in the time domain. Extending the idea of the multiple shooting method we include the unknown period together with multiple initial values as the sought parameters into the iterative procedure. Verification of the presented approach is illustrated through its application to the Colpitts oscillator model [9].

The paper is organized as follows. Section 2 describes the basis of the multiple shooting approach including the unknown period as an additional variable. This is further expanded to the family of the Parareal-based methods in Sect. 3. Section 4 applies the proposed parallel-in-time approach to the Colpitts oscillator model using a particular linerization on the coarse level. The paper is finally summarized in Sect. 5.

2 Multiple Shooting with Unknown Period

We consider the following time-periodic problem for a system of ordinary differential equations (ODEs)

$$\begin{aligned} \mathbf{M}\tilde{\mathbf{u}}'(t) &= \mathbf{f}(\tilde{\mathbf{u}}(t)), \quad t \in (0, T) \\ \tilde{\mathbf{u}}(0) &= \tilde{\mathbf{u}}(T), \end{aligned} \tag{1}$$

where the period $T > 0$ and the vector $\tilde{\mathbf{u}} : [0, T] \to \mathbb{R}^d$, $d \geq 1$ are sought. $\mathbf{M}$ is a given non-singular mass matrix, $\mathbf{f}$ is a bounded and Lipschitz continous right-hand side (RHS) function. Following [3] we incorporate the period T as an unknown parameter by performing the change of variables

$$[0, T] \ni t \mapsto \tau := t/T \in [0, 1]. \tag{2}$$

The problem (1) is thereby transformed into the equivalent one: find $T > 0$ and $\mathbf{u} : [0, 1] \to \mathbb{R}^d$ such that

$$\begin{aligned} \mathbf{M}\mathbf{u}'(\tau) &= T\mathbf{f}(\mathbf{u}(\tau)), \quad \tau \in (0, 1) \\ \mathbf{u}(0) &= \mathbf{u}(1). \end{aligned} \tag{3}$$

The unit interval $[0, 1]$ is then partitioned into N windows by the nodes $0 = \tau_0 < \tau_1 < \cdots < \tau_N = 1$. The n-th subinterval has length $\Delta\tau_n = \tau_n - \tau_{n-1}$, for $n = 1, \ldots, N$.

For a given discrete variable $\mathbf{U}_{n-1}$, we consider an initial value problem (IVP) on the window $(\tau_{n-1}, \tau_n]$

$$\begin{aligned} \mathbf{M}\mathbf{u}_n'(\tau) &= T\mathbf{f}(\mathbf{u}_n(\tau)), \quad \tau \in (\tau_{n-1}, \tau_n] \\ \mathbf{u}_n(\tau_{n-1}) &= \mathbf{U}_{n-1} \end{aligned} \tag{4}$$

and let $\mathscr{F}(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}, T)$ denote the solution operator of (4) for $n = 1, \ldots, N$. A sketch of the piecewise-defined solution due to the interval splitting is shown in Fig. 1. In order to eliminate the jumps at the synchronization points τ_n, $n = 1, \ldots, N-1$ as well as the difference between the initial value at τ_0 and the final one at τ_N the matching conditions:

$$\Phi(\mathbf{z}) := \begin{cases} \mathscr{F}(\tau_N, \tau_{N-1}, \mathbf{U}_{N-1}, T) - \mathbf{U}_0 = 0, \\ \mathscr{F}(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}, T) - \mathbf{U}_n = 0, \quad n = 1, \ldots, N-1 \end{cases} \tag{5}$$

have to be satisfied, where $\mathbf{z} = \left[\mathbf{U}_0^\top, \ldots, \mathbf{U}_{N-1}^\top, T\right]^\top$. System (5) represents the root-finding problem for the mapping $\Phi : \mathbb{R}^{Nd+1} \to \mathbb{R}^{Nd}$. The Jacobian of Φ is given by

$$\mathbf{J}_\Phi(\mathbf{z}) = \begin{bmatrix} -\mathbf{I} & & & \mathbf{G}_N & \mathbf{g}_N \\ \mathbf{G}_1 & -\mathbf{I} & & & \mathbf{g}_1 \\ & \ddots & \ddots & & \vdots \\ & & \mathbf{G}_{N-1} & -\mathbf{I} & \mathbf{g}_{N-1} \end{bmatrix}, \tag{6}$$

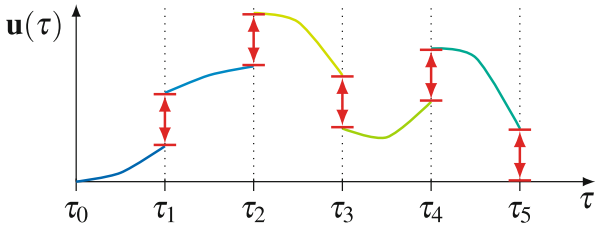


Fig. 1 Example of the interval splitting within the multiple shooting for $N = 5$. The mismatches at the synchronization points τ_n, $n = 1, \ldots, N-1$ together with the periodicity jump between the solution at τ_0 and τ_N are eliminated (up to a prescribed tolerance) by solving the root-finding problem

where

$$\mathbf{G}_n = \frac{\partial \mathscr{F}}{\partial \mathbf{U}_{n-1}}(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}, T), \qquad \mathbf{g}_n = \frac{\partial \mathscr{F}}{\partial T}(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}, T), \tag{7}$$

$n = 1, \ldots, N$ and $\mathbf{I}$ is the identity matrix. The root $\mathbf{z}$ of (5) can then be calculated using the Newton method, i.e., for a given $\mathbf{z}^{(0)}$ and $k = 0, 1, \ldots$ solution $\mathbf{z}^{(k+1)}$ at the iteration $k+1$ is updated through

$$\mathbf{J}_\Phi\big(\mathbf{z}^{(k)}\big)\Delta\mathbf{z}^{(k)} = -\Phi\big(\mathbf{z}^{(k)}\big), \tag{8}$$

$$\mathbf{z}^{(k+1)} = \Delta\mathbf{z}^{(k)} + \mathbf{z}^{(k)}. \tag{9}$$

Note that due to the introduction of the additional variable T the system of Eq. (8) is underdetermined and can be solved, e.g., by calculating the Moore-Penrose pseudoinverse. A generalized eigenvalue-based gauging as well as the corresponding theory for the Moore-Penrose pseudoinversion was presented in [12]. We note that in case when the size of (8) is large it can be condensed to a d-dimensional system with $d+1$ unknowns by block Gaussian elimination [3].

3 Periodic Time-Parallelization with Coarse Grid Correction

Inheriting the idea of the Parareal algorithm [7, 11] we approximate the derivative in $\mathbf{G}_n$ (7) in a finite difference way using a coarse propagator $\mathscr{G}$, i.e., for the iteration k and $n = 1, \ldots, N$

$$\begin{aligned}\mathbf{G}_n^{(k)}\Delta\mathbf{U}_{n-1}^{(k)} &= \frac{\partial \mathscr{F}}{\partial \mathbf{U}_{n-1}}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k)}, T^{(k)}\big)\big[\mathbf{U}_{n-1}^{(k+1)} - \mathbf{U}_{n-1}^{(k)}\big] \\ &\approx \mathscr{G}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k+1)}, T^{(k)}\big) - \mathscr{G}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k)}, T^{(k)}\big).\end{aligned} \tag{10}$$

Similar to the fine propagator $\mathscr{F}$, the operator $\mathscr{G}$ solves the IVP (4) on each time window. However, in contrast to the fine solver the coarse propagator has a considerably lower precision, e.g., it uses a lower-order time integrator or bigger time steps. Substituting (10) into (8) we obtain the periodic Parareal with periodic coarse problem [6] with unknown period (PP-PC-UP):

$$\begin{bmatrix} -\mathbf{I} & & & \mathscr{G}_N\left(\cdot, T^{(k)}\right) & \mathbf{g}_N^{(k)} \\ \mathscr{G}_1\left(\cdot, T^{(k)}\right) & -\mathbf{I} & & & \mathbf{g}_1^{(k)} \\ & \ddots & \ddots & & \vdots \\ & & \mathscr{G}_{N-1}\left(\cdot, T^{(k)}\right) & -\mathbf{I} & \mathbf{g}_{N-1}^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{U}_0^{(k+1)} \\ \mathbf{U}_1^{(k+1)} \\ \vdots \\ \mathbf{U}_{N-1}^{(k+1)} \\ T^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_N^{(k)} \\ \mathbf{b}_1^{(k)} \\ \vdots \\ \mathbf{b}_{N-1}^{(k)} \end{bmatrix}, \tag{11}$$

where we denote $\mathbf{b}_n^{(k)} := \mathbf{g}_n^{(k)} T^{(k)} + \mathscr{G}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k)}, T^{(k)}\big) - \mathscr{F}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k)}, T^{(k)}\big)$ and $\mathscr{G}_n\big(\cdot, T^{(k)}\big) := \mathscr{G}\big(\tau_n, \tau_{n-1}, \cdot, T^{(k)}\big)$ for $n = 1, \ldots, N$. Following [3] we have

$$\begin{aligned}\mathbf{g}_n^{(k)} &= \frac{\partial \mathscr{F}}{\partial T}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k)}, T^{(k)}\big) = \frac{\partial}{\partial T}\Big[\mathbf{U}_{n-1}^{(k)} + \mathbf{M}^{-1} \int_{\tau_{n-1}}^{\tau_n} T^{(k)} \mathbf{f}(\mathbf{u}(\tau)) d\tau\Big] \\ &= \mathbf{M}^{-1} \int_{\tau_{n-1}}^{\tau_n} \mathbf{f}(\mathbf{u}(\tau)) d\tau \approx \mathbf{M}^{-1} \Delta\tau_n \mathbf{f}\big(\mathscr{F}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k)}, T^{(k)}\big)\big),\end{aligned} \tag{12}$$

for $n = 1, \ldots, N$. In the general case, the system of Eq. (11) is nonlinear and implicit, which requires an additional linearization.

Building upon the ideas presented in [10], which dealt with the time-periodic problem for a known given period T, we encorporate an additive splitting of the system matrix in (11). For this let us introduce a modified coarse propagator $\bar{\mathscr{G}}$, which instead of (4) solves an approximate model with a linearized function $\bar{\mathbf{f}}(\mathbf{u}) = \mathbf{A}\mathbf{u} + \mathbf{c}$ on the RHS, i.e.,

$$\begin{aligned} \mathbf{M}\mathbf{u}_n'(\tau) &= T\bar{\mathbf{f}}(\mathbf{u}_n(\tau)) = T[\mathbf{A}_n \mathbf{u}_n(\tau) + \mathbf{c}_n], \quad \tau \in (\tau_{n-1}, \tau_n] \\ \mathbf{u}_n(\tau_{n-1}) &= \mathbf{U}_{n-1} \end{aligned} \tag{13}$$

with a given Jacobi-matrix $\mathbf{A}_n$ and a vector $\mathbf{c}$. Having the linear coarse model we construct a fixed point iteration: for $s = 0, 1, \ldots$

$$\begin{bmatrix} -\mathbf{I} & & & \bar{\mathscr{G}}_N\big(\cdot, T^{(k)}\big) & \mathbf{g}_N^{(k)} \\ \bar{\mathscr{G}}_1\big(\cdot, T^{(k)}\big) & -\mathbf{I} & & & \mathbf{g}_1^{(k)} \\ & \ddots & \ddots & & \vdots \\ & & \bar{\mathscr{G}}_{N-1}\big(\cdot, T^{(k)}\big) & -\mathbf{I} & \mathbf{g}_{N-1}^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{U}_0^{(k+1,s+1)} \\ \mathbf{U}_1^{(k+1,s+1)} \\ \vdots \\ \mathbf{U}_{N-1}^{(k+1,s+1)} \\ T^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_N^{(k+1,s)} \\ \mathbf{h}_1^{(k+1,s)} \\ \vdots \\ \mathbf{h}_{N-1}^{(k+1,s)} \end{bmatrix} \tag{14}$$

where $\mathbf{h}_n^{(k+1,s)} := \mathbf{b}_n^{(k)} + \bar{\mathscr{G}}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k+1,s)}, T^{(k)}\big) - \mathscr{G}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k+1,s)}, T^{(k)}\big)$ and $\bar{\mathscr{G}}_n\big(\cdot, T^{(k)}\big) := \bar{\mathscr{G}}\big(\tau_n, \tau_{n-1}, \cdot, T^{(k)}\big)$ for $n = 1, \ldots, N$. Assuming that $\bar{\mathscr{G}}$ solves (13) with the implicit Euler method using a single step on $(\tau_{n-1}, \tau_n]$ and that all the windows have the same length $\Delta\tau$, we have an explicit representation for the coarse solution

$$\Big[1/\Delta\tau \cdot \mathbf{M} - T^{(k)}\mathbf{A}\Big] \bar{\mathscr{G}}\big(\tau_n, \tau_{n-1}, \mathbf{U}_{n-1}^{(k+1,s)}, T^{(k)}\big) = 1/\Delta\tau \cdot \mathbf{M}\mathbf{U}_{n-1}^{(k+1,s)} + T^{(k)}\mathbf{c}_n, \tag{15}$$

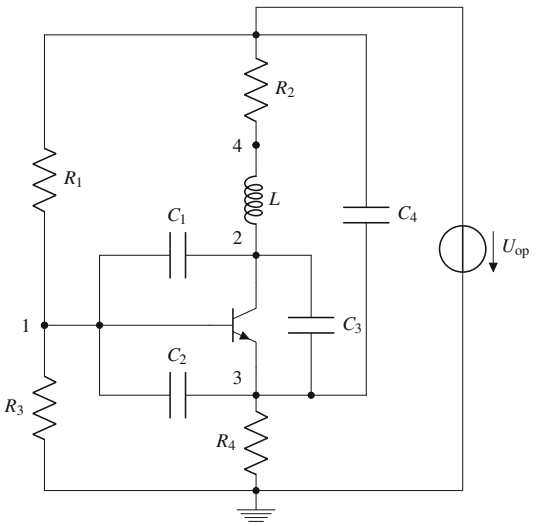
for $n = 1, \ldots, N$. Denoting by $\mathbf{C} := 1/\Delta\tau \cdot \mathbf{M}$ and $\mathbf{Q}^{(k)} := \mathbf{C} - T^{(k)}\mathbf{A}$ and plugging this into the system (13) we obtain

$$\begin{bmatrix} -\mathbf{Q}^{(k)} & & & \mathbf{C} & \mathbf{Q}^{(k)}\mathbf{g}_N^k \\ \mathbf{C} & -\mathbf{Q}^{(k)} & & & \mathbf{Q}^{(k)}\mathbf{g}_1^k \\ & \ddots & \ddots & & \vdots \\ & & \mathbf{C} & -\mathbf{Q}^{(k)} & \mathbf{Q}^{(k)}\mathbf{g}_{N-1}^k \end{bmatrix} \begin{bmatrix} \mathbf{U}_0^{(k+1,s+1)} \\ \mathbf{U}_1^{(k+1,s+1)} \\ \vdots \\ \mathbf{U}_{N-1}^{(k+1,s+1)} \\ T^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}^{(k)}\mathbf{h}_N^{(k+1,s)} - T^{(k)}\mathbf{c}_N \\ \mathbf{Q}^{(k)}\mathbf{h}_1^{(k+1,s)} - T^{(k)}\mathbf{c}_1 \\ \vdots \\ \mathbf{Q}^{(k)}\mathbf{h}_{N-1}^{(k+1,s)} - T^{(k)}\mathbf{c}_{N-1} \end{bmatrix}.$$

Remark 1 We note that when the period T is given within the problem setting (1), the corresponding block-cyclic matrix (system matrix of (3) without the last column) can be transformed into a block-diagonal using the frequency domain transformation [2]. A detailed description of the approach as well as a Newton-like linearization of the periodic system within the parallel-in-time setting is presented in [10].

4 Numerical Example

We now consider the Colpitts oscillator model presented in [9]. It is described by the circuit illustrated in Fig. 2, which consists of an inductance, a bipolar transistor, as well as of four capacitances and four resistances. The Colpitts oscillator model was exploited in the multi-rate context in [14].



$C_1 = 50\,\text{pF}$, $C_2 = 1\,\text{nF}$,
$C_3 = 50\,\text{nF}$, $C_4 = 100\,\text{nF}$,

$R_1 = 12\,\text{k}\Omega$, $R_2 = 3\,\Omega$,
$R_3 = 8.2\,\text{k}\Omega$, $R_4 = 1.5\,\text{k}\Omega$,

$L = 10\,\text{mH}$, $U_{op} = 10\,\text{V}$.

Fig. 2 Circuit of the Colpitts oscillator model [9]

The mathematical model of the circuit is given by an implicit system of ODEs [9], namely, we search for the four node voltages $\mathbf{U} = [U_1, U_2, U_3, U_4]^\top$ s.t.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & C_1 + C_3 & -C_3 & -C_1 \\ 0 & -C_3 & C_2 + C_3 + C_4 & -C_2 \\ 0 & -C_1 & -C_2 & C_1 + C_2 \end{bmatrix} \begin{bmatrix} \dot{U}_1 \\ \dot{U}_2 \\ \dot{U}_3 \\ \dot{U}_4 \end{bmatrix} = \begin{bmatrix} (U_2 - U_1)R_2/L \\ (U_{\mathrm{op}} - U_1)/R_2 + x_{\mathrm{C}} h(U_4 - U_2) - I_{\mathrm{S}} h(U_4 - U_3) \\ -U_3/R_4 + x_{\mathrm{E}} h(U_4 - U_3) - I_{\mathrm{S}} h(U_4 - U_2) \\ -U_4/R_3 + (U_{\mathrm{op}} - U_4)/R_1 - y_{\mathrm{E}} h(U_4 - U_3) - y_{\mathrm{C}} h(U_4 - U_2) \end{bmatrix}, \tag{16}$$

with the parameters $y_{\mathrm{E}} = 10\,\mu\mathrm{A}$, $x_{\mathrm{E}} = 1.01\,\mathrm{mA}$, $I_{\mathrm{S}} = 1\,\mathrm{mA}$ $y_{\mathrm{C}} = 20\,\mu\mathrm{A}$, $x_{\mathrm{C}} = 1.02\,\mathrm{mA}$, and the nonlinear function $h(x) = \exp(x/U_{\mathrm{T}}) - 1$, $U_{\mathrm{T}} = 2.585\,\mathrm{V}$, coming from the applied transistor model. Compared to the model introduced in [9], the value of U_{T} is chosen bigger to ease the convergence of PP-PC using the function h. In practice, one may need appropriate homotopy or damping strategies, see [4]. The transient behavior of the oscillator on $[0, 1.125]$ ms is shown in Fig. 3 on the left. The time step $\delta T = 0.1125\,\mu\mathrm{s}$ and the initial value at $t = 0$ is $\mathbf{u}_0 = [9.75, 1, 1, 1]^\top$ are chosen.

To find the periodic steady-state solution and the corresponding period T we apply the iteration (14). Linearization of the nonlinear periodic system on the coarse

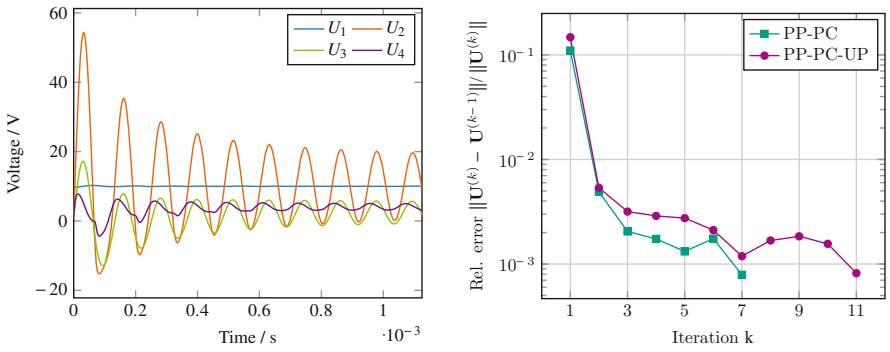


Fig. 3 Left: Transient behavior of the Colpitts oscillator until the steady state. Right: Convergence of the PP-PC approach with a linearized coarse grid problem for the case when the period T is given [10] and of PP-PC-UP when T is unknown (3)

level is performed using a surrogate linear model, i.e., $\bar{\mathscr{G}}$ solves the problem (13) with $\bar{\mathbf{f}}(\mathbf{U}) = \mathbf{A}\mathbf{U} + \mathbf{c}$ given by

$$\mathbf{A} = \begin{bmatrix} -R_2/L & R_2/L & & \\ -1/R_2 & -x_\mathrm{C}/\bar{U}_\mathrm{T} & I_\mathrm{S}/\bar{U}_\mathrm{T} & (x_\mathrm{C} - I_\mathrm{S})/\bar{U}_\mathrm{T} \\ 0 & I_\mathrm{S}/\bar{U}_\mathrm{T} & -1/R_4 - x_\mathrm{E}/\bar{U}_\mathrm{T} & (x_\mathrm{E} - I_\mathrm{S})/\bar{U}_\mathrm{T} \\ 0 & y_\mathrm{C}/\bar{U}_\mathrm{T} & y_\mathrm{E}/\bar{U}_\mathrm{T} & -1/R_3 - 1/R_1 - y_\mathrm{E}/\bar{U}_\mathrm{T} - y_\mathrm{C}/\bar{U}_\mathrm{T} \end{bmatrix}, \tag{17}$$

$$\mathbf{c} = \left[0, U_\mathrm{op}/R_2, 0, U_\mathrm{op}/R_1\right]^\top, \tag{18}$$

with $\bar{U}_\mathrm{T} = 0.2585\,\mathrm{V}$. The unit interval $[0, 1]$ is split into $N = 10$ windows. The coarse time step $\Delta\tau = 0.1$ and the fine step $\delta\tau = 10^{-4}$ were chosen within the time integration. The calculated period with the fixed point iteration (14) is $T = 0.1125\,\mathrm{ms}$. The right-hand part of Fig. 3 shows convergence of the PP-PC iteration with the linearization from [10] for a given period T as well as for an unknown period (PP-PC-UP). Both results are obtained up to the relative tolerance of 10^{-3}. One can see that in case when T is known the method required less iterations, as one would expect. When comparing the computational cost of the computations in terms of the number of linear systems solves, PP-PC and PP-PC-UP delivered the periodic solution effectively 4 and 3 times faster than the sequential time stepping, respectively.

5 Conclusions

An iterative parallel-in-time method for solving time-periodic problems where the period is not initially given is proposed in this paper. It complements the system of equations originating from the Parareal-like algorithm with an additional variable T and gives an underdetermined system of nonlinear equations. A linearization using the fixed point iteration is applied on the coarse grid. The algorithm is verified through its application to a Colpitts oscillator model.

Acknowledgments The authors thank Roland Pulch from Universität Greifswald for his assistance with implementation and for the fruitful discussions on the Colpitts oscillator model.

This research was supported by the Graduate School CE within the Centre for Computational Engineering at Technische Universität Darmstadt, as well as by DFG grant SCHO1562/1-2 and BMBF grant 05M2018RDA (PASIROM).

References

1. A. Bermúdez, D. Gómez, M. Piñeiro, P. Salgado, A novel numerical method for accelerating the computation of the steady-state in induction machines. Comput. Math. Appl. **79**(2), 274–292 (2019)
2. O. Bíró, K. Preis, An efficient time domain method for nonlinear periodic eddy current problems. IEEE Trans. Magn. **42**(4), 695–698 (2006)
3. P. Deuflhard, Computation of periodic solutions of nonlinear ODEs. BIT **24**, 456–466 (1984)
4. P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms* (Springer, Berlin, 2004)
5. R.D. Falgout, S. Friedhoff, T.V. Kolev, S.P. MacLachlan, J.B. Schroder, Parallel time integration with multigrid. SIAM J. Sci. Comput. **36**(6), C635–C661 (2014)
6. M.J. Gander, Y.-L. Jiang, B. Song, H. Zhang, Analysis of two parareal algorithms for time-periodic problems. SIAM J. Sci. Comput. **35**(5), A2393–A2415 (2013)
7. M.J. Gander, S. Vandewalle, Analysis of the parareal time-parallel time-integration method. SIAM J. Sci. Comput. **29**(2), 556–578 (2007)
8. T. Hara, T. Naito, J. Umoto, Time-periodic finite element method for nonlinear diffusion equations. IEEE Trans. Magn. **21**(6), 2261–2264 (1985)
9. W. Kampowsky, P. Rentrop, W. Schmidt, Classification and numerical simulation of electric circuits. Surv. Math. Ind. **2**(1), 23–65 (1992)
10. I. Kulchytska-Ruchka, S. Schöps, Efficient parallel-in-time solution of time-periodic problems using a multi-harmonic coarse grid correction (2019). arXiv: 1908.05245
11. J.L. Lions, Y. Maday, G. Turinici, A parareal in time discretization of PDEs. Comp. Rend. de l'Académie des Sci. – Ser. I – Math. **332**(7), 661–668 (2001)
12. R. Mirzavand Boroujeni, E.J.W. ter Maten, T.G.J. Beelen, W.H.A. Schilders, A. Abdipour, Robust periodic steady state analysis of autonomous oscillators based on generalized eigenvalues, in *Scientific Computing in Electrical Engineering SCEE 2010*, ed. by B. Michielsen, J.-R. Poirier. Mathematics in Industry, vol. 16 (2012), pp. 293–302
13. D.D. Morrison, J.D. Riley, J.F. Zancanaro, Multiple shooting method for two-point boundary value problems. Commun. ACM **5**(12), 613–614 (1962)
14. R. Pulch, Multi time scale differential equations for simulating frequency modulated signals. APNUM **53**(2–4), 421–436 (2005)
15. Y. Takahashi, T. Tokumasu, A. Kameari, H. Kaimori, M. Fujita, T. Iwashita, S. Wakao, Convergence acceleration of time-periodic electromagnetic field analysis by singularity decomposition-explicit error correction method. IEEE Trans. Magn. **46**(8), 2947–2950 (2010)

On the Exactness of Rational Polynomial Chaos Formulation for the Uncertainty Quantification of Linear Circuits in the Frequency Domain

Paolo Manfredi and Stefano Grivet-Talocia

Abstract We discuss the general form of the transfer functions of linear lumped circuits. We show that an arbitrary transfer function defined on such circuits has a functional dependence on individual circuit parameters that is rational, with multi-linear numerator and denominator. This result demonstrates that rational polynomial chaos expansions provide more suitable models than standard polynomial chaos for the uncertainty quantification of this class of circuits.

1 Introduction

The polynomial chaos expansion (PCE) method [6] has emerged in the macromodeling and model-order reduction communities because of the remarkable accuracy and efficiency in the uncertainty quantification by stochastic systems, including electric and electronic circuits [3]. Stochastic output variables of interest are approximated with a suitable polynomial model w.r.t. random input parameters, from which statistical information is inexpensively extracted. While the method was demonstrated to provide very high accuracy with a very limited expansion order in many application scenarios, the modeling of resonant and/or distributed circuits may require large orders and the accuracy of the calculated PCE coefficients may be deteriorated by the large variability of the outputs.

A rational polynomial chaos (RPC) model with tensor-product truncation was recently introduced [4] and was shown to provide better performance, compared to the conventional single PCE with total-degree truncation that is used in most engineering applications, specifically in electrical engineering [3]. In this work, we show that the general form of any transfer function defined for a linear lumped circuit is rational w.r.t. both frequency and element values. Specifically, both numerator and denominator are multi-linear functions of element values.

P. Manfredi (✉) · S. Grivet-Talocia
Politecnico di Torino, Torino, Italy
e-mail: paolo.manfredi@polito.it; stefano.grivet@polito.it

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_3

We provide a rigorous and formal proof of this fundamental theoretical result that is somewhat well-known in electrical engineering [5], but unavailable in an unambiguous and explicit form. Thanks to the theoretical findings herein presented, we are able to show that the RPC model is exact and should be the method of choice for lumped circuits.

2 Rational Polynomial Chaos Expansion

Given an arbitrary transfer function, generically denoted with Z and defined on a linear lumped electrical circuit with d uncertain elements collected into vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)$, its RPC model reads [4]

$$Z(s, \boldsymbol{\xi}) \approx \frac{\sum_{\ell=1}^{L} N_\ell(s)\varphi_\ell(\boldsymbol{\xi})}{1 + \sum_{\ell=2}^{L} D_\ell(s)\varphi_\ell(\boldsymbol{\xi})} \tag{1}$$

where s is the Laplace variable (complex frequency). In (1), the basis functions φ_ℓ are multivariate orthogonal polynomials in the uncertain variables $\boldsymbol{\xi}$, and the coefficients N_ℓ and D_ℓ are computed using a linearized and iteratively re-weighted regression. It was empirically shown [4] that, for the uncertainty quantification of electric circuits, the RPC (1) is more accurate than the standard PCE [3]. The purpose of this work is to provide a rigorous justification.

3 Transfer Functions of Linear Lumped Circuits

We review the basic modified nodal analysis (MNA) formulation [2] of lumped linear time-invariant (LTI) circuits with RGLC components. The main objective of this derivation is to reveal in explicit form the functional dependence on the individual circuit parameters of any transfer function that can be defined on such circuits.

3.1 Basic MNA Formulation for RGLC Circuits

Let us consider a lumped LTI P-port circuit with n nodes and b branches (one-port elements). The branches are split into b_R resistors with resistance R_k, b_G resistors with conductance G_k, b_L inductors with inductance L_k, and b_C capacitors with capacitance C_k, where k is an index identifying individual components. We distinguish between resistance-defined and conductance-defined resistors to allow additive variations of either parameter. In addition, the last $b_J = P$ branches are

assumed to represent the P ports of the structure. We place ideal current sources J_k providing an excitation to the circuit, with the objective of characterizing the $P \times P$ impedance matrix $\boldsymbol{Z}(s)$ in the Laplace domain by computing the corresponding port voltages as outputs.

The branch voltage and current vectors $\boldsymbol{v}, \boldsymbol{i} \in \mathbb{R}^b$ are split according to element types as

$$\boldsymbol{v} = \left(\boldsymbol{v}_R^\mathsf{T}, \boldsymbol{v}_G^\mathsf{T}, \boldsymbol{v}_L^\mathsf{T}, \boldsymbol{v}_C^\mathsf{T}, \boldsymbol{v}_J^\mathsf{T}\right)^\mathsf{T}, \quad \boldsymbol{i} = \left(\boldsymbol{i}_R^\mathsf{T}, \boldsymbol{i}_G^\mathsf{T}, \boldsymbol{i}_L^\mathsf{T}, \boldsymbol{i}_C^\mathsf{T}, \boldsymbol{i}_J^\mathsf{T}\right)^\mathsf{T},$$

where $\boldsymbol{v}_\nu, \boldsymbol{i}_\nu \in \mathbb{R}^{b_\nu}$ for $\nu \in \{R, G, L, C, J\}$, and where the passive sign convention is used for each branch, including sources. The branch characteristic equations are collectively written for each class of components as

$$\boldsymbol{v}_R = \boldsymbol{R}\,\boldsymbol{i}_R \qquad \boldsymbol{R} = \mathrm{diag}(R_1, \ldots, R_{b_R}) \tag{2a}$$

$$\boldsymbol{i}_G = \boldsymbol{G}\,\boldsymbol{v}_G \qquad \boldsymbol{G} = \mathrm{diag}(G_1, \ldots, G_{b_G}) \tag{2b}$$

$$\boldsymbol{v}_L = \boldsymbol{L}\,\frac{d}{dt}\,\boldsymbol{i}_L \qquad \boldsymbol{L} = \mathrm{diag}(L_1, \ldots, L_{b_L}) \tag{2c}$$

$$\boldsymbol{i}_C = \boldsymbol{C}\,\frac{d}{dt}\,\boldsymbol{v}_C \qquad \boldsymbol{C} = \mathrm{diag}(C_1, \ldots, C_{b_C}) \tag{2d}$$

$$\boldsymbol{i}_J = -\boldsymbol{J} \qquad \boldsymbol{J} = (J_1, \ldots, J_{b_J})^\mathsf{T}. \tag{2e}$$

Note that the current J_k of each source is incident into its positive node.

Circuit connectivity is described by the (reduced) incidence matrix $\boldsymbol{A} \in \mathbb{R}^{n-1,b}$, with the n-th node serving as reference for the definition of the set of nodal voltages $\boldsymbol{e} \in \mathbb{R}^{n-1}$. The incidence matrix columns are partitioned according to the branch classes as

$$\boldsymbol{A} = \left(\boldsymbol{A}_R, \boldsymbol{A}_G, \boldsymbol{A}_L, \boldsymbol{A}_C, \boldsymbol{A}_J\right). \tag{3}$$

Combining Kirchhoff's current law (KCL) equations $\boldsymbol{A}\boldsymbol{i} = \boldsymbol{0}$ and Kirchhoff's voltage law (KVL) equations $\boldsymbol{v}_\nu = \boldsymbol{A}_\nu^\mathsf{T}\boldsymbol{e}$ for $\nu \in \{R, G, L, C, J\}$ with the characteristics (2), leads to the system of linear differential-algebraic equations

$$\boldsymbol{\mathcal{G}}\,\boldsymbol{x} + \boldsymbol{\mathcal{C}}\,\frac{d}{dt}\,\boldsymbol{x} = \boldsymbol{\mathcal{B}}\,\boldsymbol{u} \tag{4a}$$

$$\boldsymbol{y} = \boldsymbol{\mathcal{B}}^\mathsf{T}\,\boldsymbol{x}, \tag{4b}$$

which represents the standard MNA formulation. In (4), $\boldsymbol{u} = \boldsymbol{J}$ denotes the port currents, considered as inputs, $\boldsymbol{y} = \boldsymbol{v}_J$ denotes the corresponding port voltages,

considered as outputs, vector $\boldsymbol{x} \in \mathbb{R}^m$, with $m = n - 1 + b_R + b_L$, collects the MNA variables $\boldsymbol{e}$, $\boldsymbol{i}_R$, $\boldsymbol{i}_L$, and

$$\boldsymbol{\mathcal{G}} = \begin{pmatrix} \boldsymbol{A}_G \boldsymbol{G} \boldsymbol{A}_G^\mathsf{T} & \boldsymbol{A}_R & \boldsymbol{A}_L \\ -\boldsymbol{A}_R^\mathsf{T} & \boldsymbol{R} & \boldsymbol{0} \\ -\boldsymbol{A}_L^\mathsf{T} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{\mathcal{C}} = \begin{pmatrix} \boldsymbol{A}_C \boldsymbol{C} \boldsymbol{A}_C^\mathsf{T} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{L} \end{pmatrix}, \quad \boldsymbol{\mathcal{B}} = \begin{pmatrix} \boldsymbol{A}_J \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}. \tag{5}$$

Throughout this work, we denote with $\boldsymbol{0}$ an all-zero matrix or vector, whose size is inferred from the context.

The so-called *stamps* of the individual circuit elements in the MNA system (4) are now easily characterized. A straightforward derivation shows that

$$\boldsymbol{\mathcal{G}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{G}}_0 + \sum_{k=1}^{b_a} (\boldsymbol{p}_k \boldsymbol{p}_k^\mathsf{T})\, \theta_k, \quad \boldsymbol{\mathcal{C}}(\boldsymbol{\zeta}) = \sum_{k=1}^{b_d} (\boldsymbol{q}_k \boldsymbol{q}_k^\mathsf{T})\, \zeta_k, \tag{6}$$

where

- $b_a = b_R + b_G$ is the number of adynamic components with values collected in vector $\boldsymbol{\theta} \in \mathbb{R}^{b_a}$, having elements $\{\theta_k\}_{k=1}^{b_a} = \{R_k\}_{k=1}^{b_R} \cup \{G_k\}_{k=1}^{b_G}$;
- $b_d = b_L + b_C$ is the number of dynamic components with values collected in vector $\boldsymbol{\zeta} \in \mathbb{R}^{b_d}$, having elements $\{\zeta_k\}_{k=1}^{b_d} = \{L_k\}_{k=1}^{b_L} \cup \{C_k\}_{k=1}^{b_C}$;
- the constant vectors $\boldsymbol{p}_k \in \mathbb{R}^m$ collect the sets $\{\boldsymbol{p}_k\}_{k=1}^{b_a} = \{\boldsymbol{r}_k\}_{k=1}^{b_R} \cup \{\boldsymbol{g}_k\}_{k=1}^{b_G}$ individually defined as $\boldsymbol{r}_k = \left(\boldsymbol{0}, \boldsymbol{1}_{b_R,k}^\mathsf{T}, \boldsymbol{0}\right)^\mathsf{T}$ and $\boldsymbol{g}_k = \left(\boldsymbol{a}_{G,k}^\mathsf{T}, \boldsymbol{0}, \boldsymbol{0}\right)^\mathsf{T}$, where $\boldsymbol{1}_{b_v,k}$ denotes the Euclidean basis vector in $\mathbb{R}^{b_v}$ with all vanishing elements except the k-th component equal to 1, and $\boldsymbol{a}_{G,k}$ is the k-th column of $\boldsymbol{A}_G$;
- the constant vectors $\boldsymbol{q}_k \in \mathbb{R}^m$ collect the sets $\{\boldsymbol{q}_k\}_{k=1}^{b_d} = \{\boldsymbol{l}_k\}_{k=1}^{b_L} \cup \{\boldsymbol{c}_k\}_{k=1}^{b_C}$ individually defined as $\boldsymbol{l}_k = \left(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{1}_{b_L,k}^\mathsf{T}\right)^\mathsf{T}$ and $\boldsymbol{c}_k = \left(\boldsymbol{a}_{C,k}^\mathsf{T}, \boldsymbol{0}, \boldsymbol{0}\right)^\mathsf{T}$, where $\boldsymbol{a}_{C,k}$ is the k-th column of $\boldsymbol{A}_C$;
- the constant matrix $\boldsymbol{\mathcal{G}}_0$ is defined as

$$\boldsymbol{\mathcal{G}}_0 = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{A}_R & \boldsymbol{A}_L \\ -\boldsymbol{A}_R^\mathsf{T} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{A}_L^\mathsf{T} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}. \tag{7}$$

3.2 Parameterization for Uncertainty Quantification

For the uncertainty quantification problem to be well posed, we assume that the circuit is well defined and uniquely solvable for all parameter configurations, i.e., $\exists s \in \mathbb{C}$ for which $\det(\boldsymbol{\mathcal{G}}(\boldsymbol{\theta}) + s\boldsymbol{\mathcal{C}}(\boldsymbol{\zeta})) \neq 0$. Equivalently, the pencil $(\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{C}})$ is regular for any $\boldsymbol{\theta}, \boldsymbol{\zeta}$. We further consider a nominal parameter configuration $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ and

$\boldsymbol{\zeta} = \bar{\boldsymbol{\zeta}}$. For instance, this nominal configuration can be considered as the set of expected values of the circuit element values, assumed to be stochastic variables. The initial hypothesis also implies unique solvability for this nominal parameter configuration, which is the only assumption required by the following derivations.

We introduce the variable transformation

$$\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\zeta} = \bar{\boldsymbol{\zeta}} + \boldsymbol{\delta}, \tag{8}$$

where each element of vectors $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ is a zero-mean stochastic variable, and we denote $\bar{\boldsymbol{\mathcal{G}}} = \boldsymbol{\mathcal{G}}(\bar{\boldsymbol{\theta}})$ and $\bar{\boldsymbol{\mathcal{C}}} = \boldsymbol{\mathcal{C}}(\bar{\boldsymbol{\zeta}})$. Due to linearity, (6) can now be written as

$$\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{G}}(\boldsymbol{\varepsilon}) = \bar{\boldsymbol{\mathcal{G}}} + \sum_{k=1}^{b_a} (\boldsymbol{p}_k \boldsymbol{p}_k^{\mathsf{T}})\, \varepsilon_k, \quad \boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{C}}(\boldsymbol{\delta}) = \bar{\boldsymbol{\mathcal{C}}} + \sum_{k=1}^{b_d} (\boldsymbol{q}_k \boldsymbol{q}_k^{\mathsf{T}})\, \delta_k. \tag{9}$$

We see that both the static ($\boldsymbol{\mathcal{G}}$) and the dynamic ($\boldsymbol{\mathcal{C}}$) MNA matrices are expressed as a finite sum of rank-one updates with respect to the nominal circuit formulation. Each rank-one update pertains to a single individual stochastic circuit element. The corresponding constant rank-one matrices $\boldsymbol{p}_k \boldsymbol{p}_k^{\mathsf{T}}$ and $\boldsymbol{q}_k \boldsymbol{q}_k^{\mathsf{T}}$ are recognized as the standard MNA stamps of the various circuit elements.

Let us now consider the Laplace-domain solution of (4), which in the present case corresponds to the impedance matrix of the considered P-port element and reads

$$\boldsymbol{Z}(s;\boldsymbol{\xi}) = \boldsymbol{\mathcal{B}}^{\mathsf{T}} \left[\boldsymbol{\mathcal{G}}(\boldsymbol{\varepsilon}) + s\, \boldsymbol{\mathcal{C}}(\boldsymbol{\delta})\right]^{-1} \boldsymbol{\mathcal{B}} = \frac{\mathbf{N}(s;\boldsymbol{\xi})}{\mathsf{D}(s;\boldsymbol{\xi})}, \tag{10}$$

where we have collected all stochastic parameters in a single vector $\boldsymbol{\xi}$ having elements $\{\xi_k\}_{k=1}^{d} = \{\varepsilon_k\}_{k=1}^{b_a} \cup \{\delta_k\}_{k=1}^{b_d}$, with $d = b_a + b_d$ being the total number of uncertain circuit elements, as previously defined in Sect. 2. In (10), the scalar denominator $\mathsf{D}(s;\boldsymbol{\xi})$ coincides with the determinant of the MNA matrix $\boldsymbol{\mathcal{Y}}(s;\boldsymbol{\xi}) = \boldsymbol{\mathcal{G}}(\boldsymbol{\varepsilon}) + s\, \boldsymbol{\mathcal{C}}(\boldsymbol{\delta})$, whereas each element of the numerator $\mathbf{N}(s;\boldsymbol{\xi})$ is a linear combination of the determinants of the submatrices (minors) obtained from $\boldsymbol{\mathcal{Y}}(s;\boldsymbol{\xi})$ by deleting one row and one column.

We now provide an explicit characterization of the numerator and denominator of (10). To this end, we collect all stochastic parameters in a diagonal matrix

$$\boldsymbol{\Xi} = \mathrm{diag}(\xi_1, \ldots, \xi_d), \tag{11}$$

which we use to cast the MNA matrix in the compact form, by restating (9) as

$$\boldsymbol{\mathcal{Y}}(s;\boldsymbol{\xi}) = \bar{\boldsymbol{\mathcal{Y}}}(s) + \boldsymbol{U}\, \boldsymbol{\Xi}\, \boldsymbol{S}(s). \tag{12}$$

The matrix $\bar{\mathcal{Y}}(s) = \bar{\mathcal{G}} + s\,\bar{\mathcal{C}}$ corresponds to the nominal configuration, and

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{P} & \boldsymbol{Q} \end{pmatrix}, \quad \boldsymbol{S}(s) = \begin{pmatrix} \boldsymbol{P} & s\,\boldsymbol{Q} \end{pmatrix}^{\mathsf{T}}, \tag{13}$$

where the constant matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ collect as columns all the vectors $\boldsymbol{p}_k$ and $\boldsymbol{q}_k$, respectively. From now on, we will omit the dependence on the Laplace variable s, since we are interested in the dependence on the stochastic variables $\boldsymbol{\xi}$.

We introduce two useful lemmas:

Lemma 1 *Given a square invertible matrix $\boldsymbol{X}$ and two matrices $\boldsymbol{U}$, $\boldsymbol{V}$ of compatible size, we have*

$$\det(\boldsymbol{X} + \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}) = \det(\boldsymbol{I} + \boldsymbol{V}^{\mathsf{T}}\boldsymbol{X}^{-1}\boldsymbol{U}) \cdot \det(\boldsymbol{X}).$$

The above Lemma 1 is known as *matrix determinant lemma*, see [1] for a proof.

Lemma 2 *Let a matrix $\boldsymbol{W} \in \mathbb{R}^{n,n}$ have elements in the form $W_{ij} = F_{ij} + \xi_i B_{ij}$, where F_{ij}, B_{ij} are constants for $i, j = 1, \ldots, n$, and ξ_i are independent parameters. Then,*

$$\det(\boldsymbol{W}) = \sum_k \beta_k \prod_{\ell=1}^{n} \xi_\ell^{\alpha_{k\ell}}, \tag{14}$$

where $\alpha_{k\ell} \in \{0, 1\}\ \forall k, \ell$, and β_k are real constants.

Proof We use an induction argument, noting that the statement is trivially verified for $n = 1$. Assuming that the statement holds for size $n - 1$, we evaluate $\det(\boldsymbol{W})$ for size n, for which $\boldsymbol{W}$ reads

$$\boldsymbol{W} = \begin{pmatrix} F_{11} + \xi_1 B_{11} & F_{12} + \xi_1 F_{12} & \cdots & F_{1n} + \xi_1 B_{1n} \\ F_{21} + \xi_2 B_{21} & F_{22} + \xi_2 F_{22} & \cdots & F_{2n} + \xi_2 B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ F_{n1} + \xi_n B_{n1} & F_{n2} + \xi_n F_{n2} & \cdots & F_{nn} + \xi_n B_{nn} \end{pmatrix}.$$

Expanding $\det(\boldsymbol{W})$ using Laplace's formula along the first row, we get

$$\det(\boldsymbol{W}) = \sum_{j=1}^{n} (-1)^{1+j} (F_{1j} + \xi_1 B_{1j}) M_{1j}, \tag{15}$$

where M_{1j} is the determinant of the submatrix of size $n - 1$ obtained by deleting row 1 and column j from $\boldsymbol{W}$. By the induction ansatz, we have

$$M_{1j} = \sum_k \beta_k \prod_{\ell=2}^{n} \xi_\ell^{\alpha_{k\ell}}, \quad \alpha_{k\ell} \in \{0, 1\} \quad \forall k, \ell. \tag{16}$$

Inserting (16) into (15) leads to

$$\det(\boldsymbol{W}) = \sum_{j=1}^{n}(-1)^{1+j}(F_{1j}+\xi_1 B_{1j})\sum_k \beta_k \prod_{\ell=2}^{n}\xi_\ell^{\alpha_{k\ell}}$$
$$= \sum_k \sum_{j=1}^{n}(-1)^{1+j}\left[F_{1j}\beta_k \prod_{\ell=2}^{n}\xi_\ell^{\alpha_{k\ell}} + B_{1j}\beta_k\xi_1 \prod_{\ell=2}^{n}\xi_\ell^{\alpha_{k\ell}}\right] = \sum_k \hat{\beta}_k \prod_{\ell=1}^{n}\xi_\ell^{\alpha_{k\ell}},$$

where $\alpha_{k\ell} \in \{0, 1\}$ for $\ell = 1, \ldots, n$ and $\forall k$, and $\hat{\beta}_k$ are constants. □

We are now ready to calculate the denominator $\mathsf{D}(s; \boldsymbol{\xi})$ in (10) as

$$\mathsf{D} = \det\left(\bar{\boldsymbol{\mathcal{Y}}} + \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{S}\right). \tag{17}$$

Applying Lemma 1 with $\boldsymbol{V}^{\mathsf{T}} = \boldsymbol{\Xi}\boldsymbol{S}$ and $\boldsymbol{X} = \bar{\boldsymbol{\mathcal{Y}}}$, we have

$$\mathsf{D} = \det\left(\boldsymbol{I} + \boldsymbol{\Xi}\boldsymbol{B}\right) \cdot \det\left(\bar{\boldsymbol{\mathcal{Y}}}\right), \tag{18}$$

where both $\boldsymbol{B} = \boldsymbol{S}\bar{\boldsymbol{\mathcal{Y}}}^{-1}\boldsymbol{U}$ and $\det\left(\bar{\boldsymbol{\mathcal{Y}}}\right)$ depend only on s and are thus constant with respect to the stochastic parameters $\boldsymbol{\xi}$. We have

$$\boldsymbol{I} + \boldsymbol{\Xi}\boldsymbol{B} = \begin{pmatrix} 1+\xi_1 B_{11} & \xi_1 B_{12} & \cdots & \xi_1 B_{1n} \\ \xi_2 B_{21} & 1+\xi_2 B_{22} & \cdots & \xi_2 B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_n B_{n1} & \xi_n B_{n2} & \cdots & 1+\xi_n B_{nn} \end{pmatrix}.$$

This matrix verifies the conditions of Lemma 2 with $\boldsymbol{F} = \boldsymbol{I}$. Therefore

$$\det\left(\boldsymbol{I} + \boldsymbol{\Xi}\boldsymbol{B}\right) = \sum_k \beta_k \prod_{\ell=1}^{n}\xi_\ell^{\alpha_{k\ell}} \tag{19}$$

with $\alpha_{k\ell} \in \{0, 1\}$ for all k, ℓ, which in turn implies that

$$\mathsf{D}(s; \boldsymbol{\xi}) = \sum_k \mathsf{d}_k(s) \prod_{\ell=1}^{n}\xi_\ell^{\alpha_{k\ell}}, \quad \alpha_{k\ell} \in \{0, 1\} \quad \forall k, \ell. \tag{20}$$

Due to the lumped nature of the system under consideration, the coefficients $\mathsf{d}_k(s)$ are polynomials in s of degree up to the dynamic order N of the circuit.

The same arguments used for the denominator $\mathsf{D}(s; \boldsymbol{\xi})$ can be seamlessly adopted to show that also the elements of the numerator matrix $\mathbf{N}(s; \boldsymbol{\xi})$ in (10) have the same

structural dependence on frequency s and parameters $\boldsymbol{\xi}$. Therefore, we conclude that any element (i, j) of the impedance matrix $\boldsymbol{Z}(s; \boldsymbol{\xi})$ has the following structure

$$Z_{ij}(s; \boldsymbol{\xi}) = \frac{\sum_{k=0}^{N_{ij}} a_{k;ij}(\boldsymbol{\xi}) s^k}{\sum_{k=0}^{N} b_k(\boldsymbol{\xi}) s^k}, \tag{21}$$

where all numerator and denominator coefficients $a_{k;ij}(\boldsymbol{\xi})$ and $b_k(\boldsymbol{\xi})$ have a *multi-linear* dependence in the stochastic parameters, i.e., they are multivariate polynomials in which each element of $\boldsymbol{\xi}$ appears with up to order one. In conclusion, any impedance element is a rational function of any stochastic parameter ξ_i with both numerator and denominator degrees that cannot exceed one.

Based on the above result, the RPC model (1) is exact for linear lumped circuits, provided that the polynomial basis functions φ_ℓ are multi-linear. This is readily achieved by adopting a tensor-product truncation of order one [4]. By extension, the model turns out to be more accurate also for distributed circuits and electromagnetic systems, albeit with higher-degree approximations, as was effectively and empirically demonstrated based on a number of application examples in [4].

4 An Illustrative Example

We consider the filter of Fig. 1 (left), which is designed to exhibit both a band- and a high-pass behavior. All 9 circuit elements are uncertain, with inductances and capacitances having independent Gaussian variations with a 20% standard deviation around the nominal values indicated in the schematic.

The right panel in Fig. 1 shows the variability of the insertion loss of the filter. The gray lines are a subset of random samples from a reference Monte Carlo (MC) simulation with 10,000 runs, whereas the solid blue line is the standard deviation of the MC samples. The dashed red and green lines are the standard deviations obtained with a conventional PCE having a maximum total degree of three, and

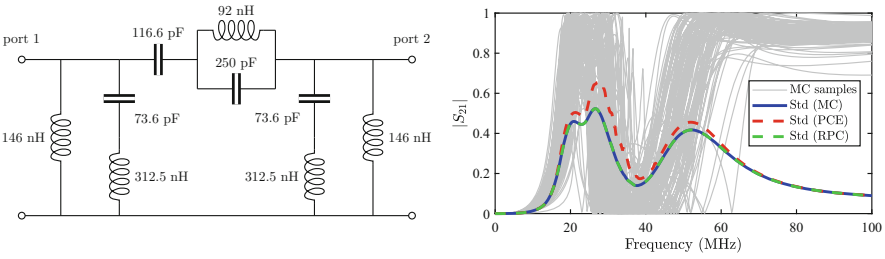


Fig. 1 Left: filter schematic. Right: variability of the insertion loss of the filter. Gray lines: MC samples; solid blue, dashed red, and dashed green lines: standard deviation obtained with MC, conventional PCE, and proposed RPC methods, respectively

with a tensor-product RPC model having a maximum degree of one, respectively. The conventional model has 220 terms in its single PCE, and the corresponding coefficients are calculated by means of an ordinary least square regression [3]. The RPC model has a total of 1023 terms (512 in the numerator and 511 in the denominator), and the coefficients are calculated with an iterative linearized least-square regression [4]. In both cases, we use a number of regression samples that is twice the number of unknowns.

As expected, the RPC provides an exact model, and the result is therefore consistent with the reference MC curve. This is further confirmed by the mean squared deviation of the two models from the MC samples, which is 3.8574×10^{-2} and 2.6264×10^{-10} for the conventional PCE and the RPC model, respectively.

5 Conclusions

This work presented a formal derivation that any frequency-domain transfer function defined on linear lumped circuits is a rational function with multi-linear dependence on the circuit element values. This results provides a rigorous motivation for using a Rational Polynomial Chaos (RPC) model for the uncertainty quantification of the frequency-domain responses of electrical circuits, and more generally of electromagnetic systems. Our findings are illustrated based on a lumped filter example.

While a first-order tensor-product truncation provides an exact model for lumped circuits, a more compact total-degree truncation (possibly of higher order) can be used to improve the efficiency, especially for applications in which the exactness no longer holds. This is the case, for example, of distributed, electromagnetic, and/or photonic systems. We are also currently investigating a compression strategy, based on principal component analysis, that avoids having to optimize the model coefficients separately for each frequency.

References

1. D.A. Harville, *Matrix Algebra From a Statistician's Perspective* (Springer, New York, 1997)
2. C.-W. Ho, A. Ruehli, P. Brennan, The modified nodal approach to network analysis. IEEE Trans. Circ. Syst. **22**(6), 504–509 (1975)
3. A. Kaintura, T. Dhaene, D. Spina, Review of polynomial chaos-based methods for uncertainty quantification in modern integrated circuits. Electronics **7**(3), 1–21 (2018)
4. P. Manfredi, S. Grivet-Talocia, Rational polynomial chaos expansions for the stochastic macromodeling of network responses. IEEE Trans. Circ. Syst. I Reg. Papers **67**(1), 225–234 (2020)
5. J. Vlach, K. Singhal, *Computer Methods for Circuit Analysis and Design* (Wiley, New York, 1983)
6. D. Xiu, Fast numerical methods for stochastic computations: a review. Commun. Comput. Phys. **5**(2–4), 242–272 (2009)

Parallel-in-Time Simulation of Power Converters Using Multirate PDEs

Andreas Pels, Iryna Kulchytska-Ruchka, and Sebastian Schöps

Abstract This paper presents a numerical algorithm for the simulation of pulse-width modulated power converters via parallelization in time domain. The method applies the multirate partial differential equation approach on the coarse grid of the (two-grid) parallel-in-time algorithm Parareal. Performance of the proposed approach is illustrated via its application to a DC-DC converter.

1 Introduction

Switch-mode power converters are devices which convert electric voltages or currents between different levels. For this purpose they use transistors to switch on and off the input voltage or current to obtain the desired average voltage or current at the output of the converter. A technique called pulse-width modulation (PWM) is often utilized to control the transistors, i.e., to generate the pulsed voltage from a given carrier and reference. An exemplary circuit of a buck converter (DC-DC converter) is depicted in Fig. 1a along with its solution in Fig. 2. It consists of fast periodically varying ripples and a slowly varying envelope. The simulation of these power converters with conventional time stepping is computationally expensive since a high number of time steps is necessary to resolve the fast variations induced by the transistor switching.

This paper proposes the simulation of power converters using a combination of two methods, namely the parallel-in-time algorithm Parareal [7] and a multirate approach based on Multirate Partial Differential Equations (MPDEs) [8]. This is accomplished via the application of the MPDE approach on the coarse grid of Parareal. It allows the coarse propagator to obtain a more precise solution given the PWM input signal, in contrast to the standard coarse propagator when using a large time step on the original system of equations.

A. Pels · I. Kulchytska-Ruchka (✉) · S. Schöps
Computational Electromagnetics Group, Technical University of Darmstadt, Darmstadt, Germany
e-mail: pels@temf.tu-darmstadt.de; kulchytska@temf.tu-darmstadt.de; schoeps@temf.tu-darmstadt.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_4

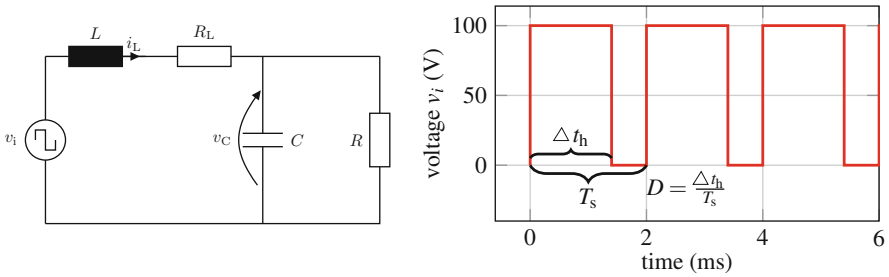


Fig. 1 Power converter model with pulsed voltage source: (**a**) Circuit of a simplified buck converter. Transistor switching is modeled as pulsed voltage source. (**b**) PWM generated pulsed voltage

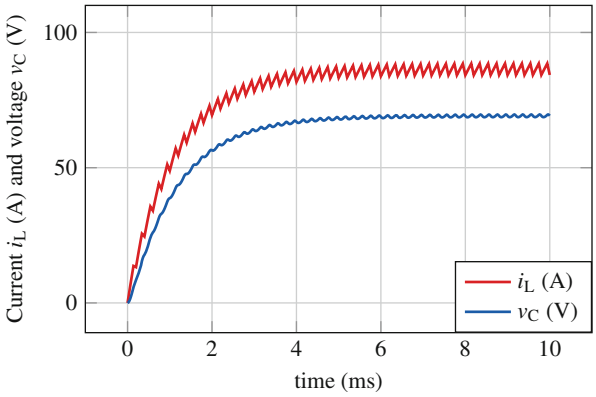


Fig. 2 Exemplary solution of the buck converter depicted in Fig. 1a. Switching frequency $f_s = 1/T_s = 5\,\text{kHz}$

The paper is organized as follows: first we introduce our model problem with pulsed excitation in Sect. 2, then in Sect. 3 the Parareal method is summarized, Sect. 4 proposes the usage of MPDEs as coarse propagators for Parareal that can deal with pulsed right-hand sides and finally Sect. 5 discusses a numerical example before concluding the paper.

2 Power Converter Model

Switch-mode power converters, which convert AC to DC, DC to AC, AC to AC, or DC to DC voltages, are frequently used devices. They use power electronic switches to periodically switch the input voltage on and off to regulate the output voltage. For example a buck converter (DC-DC converter) transforms a given voltage to a

lower output voltage. It consists of a part that generates a pulsed voltage v_i and a filter circuit. The latter is shown in Fig. 1a. The pulsed voltage, see Fig. 1b, is often generated using PWM. Important quantities defining the pulsed signal are the switching period T_s and the duty cycle D which is the relation between the "on"-time and the switching period. Given a reference signal $r(t)$ and a carrier signal $s(t)$ the pulsed voltage is generated by

$$v_i(t) = \frac{V_i}{2}\left(\operatorname{sgn}\left(r(t) - s(t)\right) + 1\right), \tag{1}$$

where sgn denotes the sign function and V_i is the amplitude. The converter circuit is mathematically described by a system of ordinary differential or differential-algebraic equations, e.g.,

$$\mathbf{A}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) + \mathbf{B}\,\mathbf{x}(t) = \mathbf{c}(t), \qquad t \in (t_0, T], \tag{2}$$

with given initial value $\mathbf{x}(t_0) = \mathbf{x}_0$, where $\mathbf{x}(t) \in \mathbb{R}^{N_\mathrm{s}}$ is the unknown solution vector consisting for example of currents and voltages, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N_\mathrm{s} \times N_\mathrm{s}}$ are matrices, and $\mathbf{c}(t) \in \mathbb{R}^{N_\mathrm{s}}$ is the right-hand side containing current and voltage sources, e.g., the pulsed voltage $v_i(t)$. The system may be assembled from lumped element descriptions based on loop or (modified) nodal analysis as described in [2]. Please note, that we focus on the linear case but the approach can be straight-forwardly generalized, e.g., considering $\mathbf{B} = \mathbf{B}(\mathbf{x})$.

The solution of power converters, as for example the one shown in Fig. 2, exhibits the multirate phenomenon: slow variations in the solution require large time intervals, i.e., a large end time point T, while the fast dynamics due to the switching enforce small time steps. This is the motivation to turn to (parallel) methods that can exploit this multirate behavior. In the following, we focus on the settling process until the steady state is reached. If one is interested only in the latter, then other methods may also be used, for example the application of Parareal for time-periodic problems is a natural generalization of this work, see, e.g., [5].

3 Parareal Algorithm

Parareal is an iterative algorithm which is able to accelerate the solution of (2) via parallelization in time. The method originates from [7] and its superlinear convergence is proven in [3]. The two main ingredients of Parareal are the fine and the coarse propagators. We denote by $\mathscr{F}(t, t_0, \mathbf{x}_0)$ and $\mathscr{G}(t, t_0, \mathbf{x}_0)$ the solutions of the initial value problem (IVP) (2) at $t \in (t_0, T]$ obtained with sequential time stepping using fine and coarse time steps, respectively.

Partitioning the time interval $t_0 = T_0 < T_1 < \cdots < T_N = T$ we write the Parareal iteration: for $k = 0, 1, \ldots$ and $n = 1, \ldots, N$ solve

$$\mathbf{x}_0^{(k+1)} = \mathbf{x}_0, \tag{3}$$

$$\mathbf{x}_n^{(k+1)} = \mathscr{F}\big(T_n, T_{n-1}, \mathbf{x}_{n-1}^{(k)}\big) + \mathscr{G}\big(T_n, T_{n-1}, \mathbf{x}_{n-1}^{(k+1)}\big) - \mathscr{G}\big(T_n, T_{n-1}, \mathbf{x}_{n-1}^{(k)}\big). \tag{4}$$

The solution operator $\mathscr{F}$ is assumed to deliver a very accurate solution (e.g., using a numerical time-integration method with small time steps δT) and can be executed in parallel, while $\mathscr{G}$ gives rough information about the solution using a cheap method (e.g., using a numerical method with large time steps $\Delta T_i = T_{i+1} - T_i$) and has to be calculated sequentially, cf. (4).

A difficulty in applying Parareal to solve problems with PWM input is that a naive implementation of a coarse propagator using a time-integrator with large time steps will not capture the high-frequency dynamics and may also fail to propagate low-frequency components. A modified Parareal algorithm which still approximately captures the high-frequency behavior was introduced in [4]. The idea is to separate the high-frequency (pulsed) components from the low-frequency components, i.e.,

$$\mathbf{A}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) + \mathbf{B}\,\mathbf{x}(t) = \underbrace{\bar{\mathbf{c}}(t) + \tilde{\mathbf{c}}(t)}_{=\mathbf{c}(t)}, \tag{5}$$

where $\bar{\mathbf{c}}$ can be given as the sum of a few low-frequency sinusoids from a (fast) Fourier transform and $\tilde{\mathbf{c}}(t) := \mathbf{c}(t) - \bar{\mathbf{c}}(t)$ is the remainder. This allows to define a reduced coarse propagator $\bar{\mathscr{G}}_{\mathrm{fft}}$ which solves

$$\mathbf{A}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) + \mathbf{B}\,\mathbf{x}(t) = \bar{\mathbf{c}}(t) \tag{6}$$

and gives rise to a modified Parareal update formula with coarse propagator $\bar{\mathscr{G}}_{\mathrm{fft}}$ in (3)–(4). This modified method converges reliably but possibly with reduced order [4]. In this paper we propose an alternative method to perform time integration by using the MPDE approach as the coarse propagator.

4 Multirate PDEs

The MPDE approach, which is used for obtaining the coarse solution in Parareal uses the MPDE concept [1]. For the given problem the solution can be conveniently decomposed into a slowly varying envelope and fast periodically varying ripples using the solution expansion [8]

$$\widehat{x}_j(t_1, t_2) \doteq \sum_{k=1}^{N_{\mathrm{p}}} y_{j,k}(t_1) w_k(\tau(t_2)) = \mathbf{w}^{\top}(\tau(t_2))\mathbf{y}_j(t_1), \tag{7}$$

where $y_{j,k}(t_1)$ are slowly varying coefficients and $w_k(\tau(t_2))$ are a finite set of basis functions ($k = 1, \ldots, N_\mathrm{p}$) whose periodicity is accounted for by the relative time $\tau(t_2) = \frac{t_2}{T_\mathrm{s}} \bmod 1$. Its application to (2) yields

$$\mathbf{A}\left(\frac{\partial \widehat{\mathbf{x}}(t_1,t_2)}{\partial t_1} + \frac{\partial \widehat{\mathbf{x}}(t_1,t_2)}{\partial t_2}\right) + \mathbf{B}\,\widehat{\mathbf{x}}(t_1,t_2) = \widehat{\mathbf{c}}(t_1,t_2), \tag{8}$$

where the relation between the original (2) and the MPDE (8) solution and right-hand side are given by

$$\widehat{\mathbf{x}}(t,t) = \mathbf{x}(t), \quad \widehat{\mathbf{c}}(t,t) = \mathbf{c}(t). \tag{9}$$

This implies that if a solution to (8) is found, the solution of (2) can be extracted from it. The specification of a suitable multitime source $\widehat{\mathbf{c}}(t_1,t_2)$ has to be supplied by the user. However, the method converges to the correct solution for any choice that fulfills (9) but it may not be more efficient than conventional time stepping. A suitable choice for PWM excitations is discussed in Sect. 5. Now, applying a Galerkin approach along the fast time scale t_2 leads to the enlarged equation system

$$\boldsymbol{\mathcal{A}}\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t_1} + \boldsymbol{\mathcal{B}}\,\mathbf{y}(t_1) \;=\; \boldsymbol{\mathcal{C}}(t_1)\,, \tag{10}$$

where the matrices are given by [8]

$$\boldsymbol{\mathcal{A}} = \mathbf{A} \otimes \boldsymbol{\mathcal{J}}, \qquad \text{with} \qquad \boldsymbol{\mathcal{J}} = T_\mathrm{s}\int_0^1 \mathbf{w}(\tau)\,\mathbf{w}^\top(\tau)\,\mathrm{d}\tau,$$

$$\boldsymbol{\mathcal{B}} = \mathbf{B} \otimes \boldsymbol{\mathcal{J}} + \mathbf{A} \otimes \boldsymbol{\mathcal{Q}}, \qquad \text{with} \qquad \boldsymbol{\mathcal{Q}} = -\int_0^1 \frac{\partial \mathbf{w}(\tau)}{\partial \tau}\,\mathbf{w}^\top(\tau)\,\mathrm{d}\tau,$$

$$\boldsymbol{\mathcal{C}} = \int_0^{T_\mathrm{s}} \widehat{\mathbf{c}}(t_1,t_2) \otimes \mathbf{w}(\tau(t_2))\,\mathrm{d}t_2\,.$$

Suitable basis functions, which can well represent the ripples in the power converter solution, are, e.g., B-Splines with suitable continuity or the PWM basis functions [6]. The latter are global polynomial ansatz functions with $w_1(\tau, D) = 1$, $w_2(\tau, D)$ piecewise linear and $w_k(\tau, D)$ is obtained recursively by integrating $w_{k-1}(\tau)$ and orthonormalizing for $3 \le k \le N_\mathrm{p}$, see Fig. 3. It has been shown in [8] that they are capable of very effectively representing the ripples in linear problems.

Finally, Eq. (10) can be time-stepped along t_1 by using much larger time steps than are needed to solve (2) since the fast variations are taken into account by the basis functions. The accuracy of the solution (reconstructed using (7)) increases with N_p. However increasing N_p also makes each time step of an implicit method more

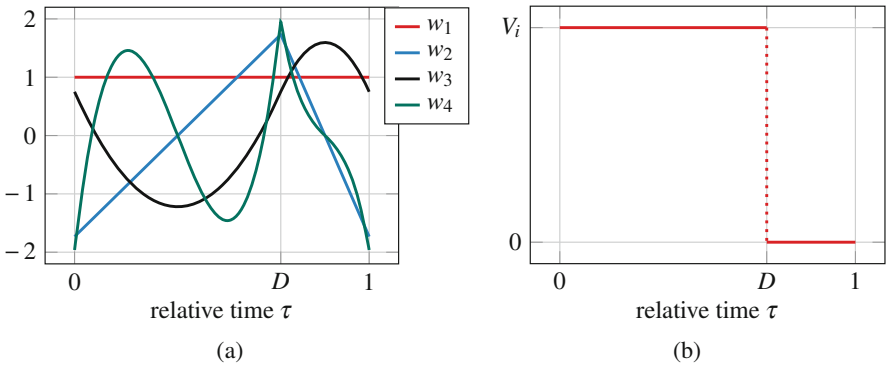


Fig. 3 Construction of basis functions with cusp at relative switching time D: (**a**) PWM basis functions on relative time interval and (**b**) right-hand side

costly since an enlarged linear equation system has to be solved. Nevertheless, even with very few basis functions the reconstructed solution can be expected to capture the main features of the exact solution. This motivates the introduction of another coarse propagator $\bar{\mathscr{G}}_{\text{mpde}}$ in Parareal which solves (10) and extracts afterwards the single-time solution according to (7).

5 Numerical Experiments

The proposed approach is applied to the academic example of the buck converter (see Fig. 1a). Its circuit is described by the IVP (2) given by

$$\mathbf{A} = \begin{bmatrix} L & 0 \\ 0 & C \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} R_{\text{L}} & 1 \\ -1 & 1/R \end{bmatrix} \quad \text{and} \quad \mathbf{c}(t) = \begin{bmatrix} v_i(t) \\ 0 \end{bmatrix}, \tag{11}$$

with inductance $L = 10^{-3}\,\text{H}$, capacitance $C = 10^{-4}\,\text{F}$, resistances $R_{\text{L}} = 10^{-2}\,\Omega$ and $R = 0.8\,\Omega$. The PWM input $v_i(t)$ has the amplitude of $V_i = 100\,\text{V}$ and is generated by a sawtooth carrier signal $s(t) = tf_{\text{s}} \bmod 1$ with switching frequency of $f_{\text{s}} = 5\,\text{kHz}$ and the reference signal $r(t) = 0.7$ according to (1). The considered time interval $[0, 12]$ ms is partitioned into $N = 40$ windows for all Parareal variants. The coarse time step size is $\Delta T = T/N = 3 \times 10^{-4}\,\text{s}$ and the fine propagator uses the time step $\delta T = 10^{-6}\,\text{s}$. All solutions are obtained with the implicit Euler method.

First, the classical Parareal method (3)–(4) is applied. It solves the original system (2) with the PWM input in both propagators, i.e., $\mathscr{G}$ and $\mathscr{F}$. It is compared to two variants where $\mathscr{G}$ is changed to: 1. $\bar{\mathscr{G}}_{\text{fft}}$ which solves system (6) containing only the DC component instead of the PWM signal on the right-hand side (modified

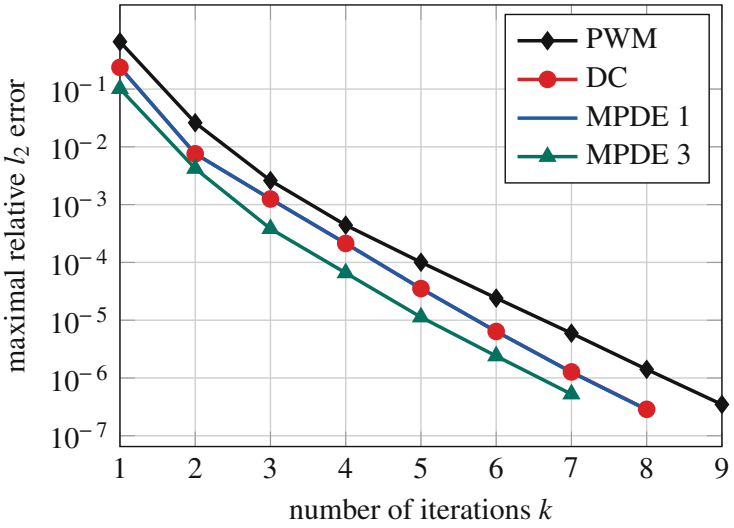


Fig. 4 Convergence of Parareal towards the sequential (reference) solution using different coarse propagators for the given example

Parareal [4]); 2. $\bar{\mathscr{G}}_{\text{mpde}}$ which solves (2) using the MPDE approach with $N_\text{p} = 1$ and $N_\text{p} = 3$ with the right-hand side $\widehat{\mathbf{c}}(t_1, t_2) = \mathbf{c}(t_2)$.

The maximal relative l_2 error w.r.t. the sequential (reference) solution $\mathbf{x}_\text{seq}(t)$

$$\max_{i \in \mathbb{N}} \frac{\|\mathbf{x}_{\star,k}(t_i) - \mathbf{x}_\text{seq}(t_i)\|_2}{\|\mathbf{x}_\text{seq}(t_i)\|_2} \quad \text{with } t_0 \le t_i := t_0 + i\delta T \le T$$

is depicted in Fig. 4 for all the considered approaches $\star \in \{\text{PWM, DC, MPDE 1, MPDE 3}\}$ at iteration k. The conventional Parareal converges for our test case (11) up to a relative error of 10^{-6} in 9 iterations which is remarkable since the time step of the coarse propagator does not resolve the dynamics of the PWM input and also violates smoothness assumption, see [4] for details.

This method requires 2 700 and 360 sequential solutions of linear algebraic systems of size $N_\text{s} = 2$ on the fine and the coarse levels, respectively, or 3 060 linear systems in total. By the number of sequential solves we mean the number of solver calls which cannot be carried out in parallel (communication costs are neglected). The approaches using the DC component and the MPDE approach with $N_\text{p} = 1$ both required 8 iterations (2 400 fine and 320 coarse solves, or in total 2 720 solutions of linear systems in 2 variables). Finally, the MPDE approach with $N_\text{p} = 3$ basis functions on the coarse level converged after 7 iterations, thereby solving 2 100 linear systems of size $N_\text{s} = 2$ on the fine level and 280 linear systems of size $N_\text{s} \times N_\text{p} = 6$ on the coarse level. One observes that the conventional coarse propagator requires always roughly 1-2 Parareal iterations more than the MPDE

approach with $N_\mathrm{p} = 3$ to obtain the same accuracy, e.g., MPDE 3 needs 3 instead of 5 iterations for an error of 4×10^{-4}.

From Fig. 4 we see that Parareal with coarse propagator $\bar{\mathscr{G}}_\mathrm{mpde}$ using a constant basis function, i.e., $N_\mathrm{p} = 1$ and the modified Parareal with $\bar{\mathscr{G}}_\mathrm{fft}$ using only the DC excitation perform very similarly (if not identically). This resemblance is not surprising since the MPDE 1 approach with $N_\mathrm{p} = 1$ computes only the envelope of the solution, which is conceptually similar to the modified Parareal with a smooth (in this case constant) coarse input. Finally, further tests show that the exploitation of more basis functions ($N_\mathrm{p} > 3$) does not improve the convergence of Parareal, they are similar to the case $N_\mathrm{p} = 3$.

6 Conclusions

In this paper we introduced a novel parallel-in-time algorithm, able to treat systems excited by pulse-width modulated signals. The method extends the two-grid Parareal algorithm by exploiting the MPDE solution approach on the coarse grid. It was applied to the time-domain simulation of a buck converter supplied by a PWM voltage source. Future research will further investigate the similarity of Parareal with the MPDE coarse propagator and the modified Parareal as well as higher order MPDE approaches as coarse propagators.

Acknowledgments The authors thank Ruth Vazquez Sabariego from KU Leuven for many fruitful discussions on the MPDE approach. This research was supported by the Graduate School CE within the Centre for Computational Engineering at Technische Universität Darmstadt, as well as by DFG grant SCHO1562/1-2 and BMBF grant 05M2018RDA.

References

1. H.G. Brachtendorf, G. Welsch, R. Laur, A. Bunse-Gerstner, Numerical steady state analysis of electronic circuits driven by multi-tone signals. Electr. Eng. (Arch. für Elektrotech.) **79**(2), 103–112 (1996)
2. D. Estévez Schwarz, C. Tischendorf, Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theory Appl. **28**(2), 131–162 (2000)
3. M.J. Gander, E. Hairer, Nonlinear convergence analysis for the parareal algorithm, in *Domain Decomposition Methods in Science and Engineering XVII*, ed. by U. Langer, M. Discacciati, D.E. Keyes, O.B. Widlund, W. Zulehner (Springer, New York, 2008), pp. 45–56
4. M.J. Gander, I. Kulchytska-Ruchka, I. Niyonzima, S. Schöps, A new parareal algorithm for problems with discontinuous sources. SIAM J. Sci. Comput. **41**(2), B375–B395 (2019)
5. M.J. Gander, I. Kulchytska-Ruchka, S. Schöps, A new parareal algorithm for time-periodic problems with discontinuous inputs, in *Domain Decomposition Methods in Science and Engineering XXV*. Lecture Notes in Computational Science and Engineering (Springer, New York, 2019)

6. J. Gyselinck, C. Martis, R.V. Sabariego, Using dedicated time-domain basis functions for the simulation of pulse-width-modulation controlled devices – application to the steady-state regime of a buck converter, in *Electromotion 2013*, 2013
7. J.L. Lions, Y. Maday, G. Turinici, A parareal in time discretization of PDEs. Comp. Rend. de l'Académie des Sci. – Ser. I – Math. **332**(7), 661–668 (2001)
8. A. Pels, J. Gyselinck, R.V. Sabariego, S. Schöps, Efficient simulation of DC-DC switch-mode power converters by multirate partial differential equations. IEEE J. Multiscale Multiphys. Comput. Tech. **4**(1), 64–75 (2019)

Part II
Device Simulation

A Maximum Principle for Drift-Diffusion Equations and the Scharfetter-Gummel Discretization

Kai Bittner, Hans Georg Brachtendorf, Tobias Linn, and Christoph Jungemann

Abstract The solution of the drift-diffusion equation does not satisfy a maximum principle in general. Here it is shown that a maximum principle can be established for the so called Slotboom variable, which permits statements on uniqueness, stability, and positivity. This maximum principle is preserved for the discretized system obtained by the Scharfetter-Gummel scheme.

1 Introduction

A maximum principle states that the solution of certain partial differential equations attains its maximum on the boundary of the solution domain. Usually, if a maximum principle holds, there follows also a corresponding minimum principle by straightforward arguments (as changing signs). The maximum principle implies several properties as uniqueness and stability of solutions. Furthermore, it often ensures positiveness of physical quantities, as e.g. electron and hole densities in semiconductors, where negative values would be non-physical.

A drift-diffusion equation, used e.g. to model the transport of electrons and holes in a semiconductor, does not exhibit a maximum principle for the densities in general. The densities in semiconductor devices may vary by orders of magnitude due to huge differences in doping concentrations. On the over hand the drift-diffusion equations ensures the positivity of densities. In a numerical scheme this positivity shall still be guaranteed, which is indeed the case for the Scharfetter-Gummel discretization [1]. However, the situation looks different if one considers the Slotboom variable instead of the density.

K. Bittner (✉) · H. G. Brachtendorf
University of Applied Sciences of Upper Austria, Hagenberg, Austria
e-mail: Kai.Bittner@fh-hagenberg.at; Hans-Georg.Brachtendorf@fh-hagenberg.at

T. Linn · C. Jungemann
Institute of Electromagnetic Theory, RWTH Aachen University, Aachen, Germany
e-mail: tl@ithe.rwth-aachen.de; cj@ithe.rwth-aachen.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_5

A maximum principle for Fermi potentials, which implies the maximum principle for the Slotboom variables was shown in [2] for static drift diffusion equations without recombination term. An estimate for the Slotboom variables in the stationary semiconductor equations (van Roosbroeck system) was established by Markowich [3, Theorem 3.2.1]. Here, the bound depends on the Dirichlet boundary values, but do not yield a Maximum principle as presented by us in Theorem 1. A discretized version of this estimate for a Scharfetter-Gummel finite volume scheme can be found in [4]. An estimate for the time dependent van Roosbroeck system is given in [5], where it is shown that the solution of a Scharfetter-Gummel finite volume scheme is bounded by constants depending only on boundary and initial data.

The results in this paper were motivated by the investigation of extended drift-diffusion equations (see e.g. [6]) containing a time derivative of the flux and the convective derivative. This equations are hyperbolic and do therefore not satisfy a maximum principle. Thus it is a challenge to ensure positivity of densities for the original equation as well as for discretizations. However, the study of the drift diffusion equation and the Scharfetter-Gummel discretization lead to a maximum principle with improved bounds for continuous solutions. This maximum principle and its implications are presented in Sect. 2. In Sect. 3 we show how this maximum principle is preserved by the Scharfetter-Gummel scheme.

2 A Maximum Principle for the Drift Diffusion Equation

We consider the drift-diffusion equation for the steady state:

$$\nabla^T \big(\nabla n(x) - n(x)\nabla\phi(x)\big) = d(n(x), x), \tag{1}$$

where n is the unknown particle density, ϕ is a potential, and d is a source term (e.g. for carrier generation and recombination in semiconductors). Here we have used without loss of generality a scaling of quantities and equations which simplifies the formulation. In the sequel $\Omega \subset \mathbb{R}^n$ will be an open domain, $\overline{\Omega}$ denotes its closure and $\partial\Omega = \overline{\Omega} \setminus \Omega$ is the boundary. We split the boundary into the closed Dirichlet boundary $\Gamma_D \subset \partial\Omega$, with $\Gamma_D \neq \emptyset$, and the Neumann boundary $\Gamma_N = \partial\Omega \setminus \Gamma_D$. We consider the Dirichlet boundary conditions

$$n(x) = g(x), \quad x \in \Gamma_D \tag{2}$$

for the density and Neumann boundary conditions

$$\nu^T \big(\nabla n(x) - n(x)\nabla\phi(x)\big) = h(x), \quad x \in \Gamma_N \tag{3}$$

for the flux, where ν denotes the outer normal vector.

The densities $n(x)$ do not satisfy a maximum principle in general. In particular, rapid changes of the potential (caused e.g. by differences of the doping concentration in semiconductors) result in changes of the density $n(x)$ such that a maximum on the boundary is not guaranteed. However, the *Slotboom variable* [7]

$$\tilde{n}(x) := e^{-\phi(x)} n(x) \tag{4}$$

takes into account changes of the potential. For the equilibrium one has e.g. a constant Slotboom variable. In particular, in terms of the Slotboom variable the flux is written as

$$\nabla n - n \nabla \phi = e^{\phi} \nabla \tilde{n}, \tag{5}$$

such that (1) becomes an elliptic problem in terms of $\tilde{n}(x)$, which allows us to formulate a maximum principle for the Slotboom variable. One formulation could be obtained from [8, Theorem A.1], however we will present sharper bounds here.

Theorem 1 *Let $n : \overline{\Omega} \to \mathbb{R}$ be a continuous solution of (1) on the open domain $\Omega \subset \mathbb{R}^n$. For the Neumann boundary conditions (3) we require $h(x) \leq 0$, $x \in \Gamma_N$. If $d(n, x) \geq 0$, $n \geq n_0(x)$, $x \in \Omega$, then the Slotboom variable $\tilde{n}(x)$ satisfies*

$$\tilde{n}(x) \leq \max\left(n_0(x) e^{-\phi(x)}, \max_{y \in \Gamma_D} \tilde{n}(y) \right), \quad x \in \overline{\Omega}. \tag{6}$$

Proof Let us first assume that the maximum is not attained in Γ_D but in $x^* \in \Omega$ and $n(x^*) > n_0(x)$. For any sufficiently small $\varepsilon > 0$ 0ne has $n(x) > n_0(x)$ for $x \in \overline{B_\varepsilon(x^*)} \subset \Omega$, where $B_\varepsilon(x^*) := \{y \in \mathbb{R}^n : \|y - x^*\| < \varepsilon\}$ is the ε-ball around x^*. Furthermore, from $\tilde{n}(x) \leq \tilde{n}(x^*)$ it follows that $\nu^T \nabla \tilde{n}(x) \leq 0$ for $x \in \partial B_\varepsilon(x^*)$ with $\varepsilon > 0$ sufficiently small. Without loss of generality we can further assume that $\nu^T \nabla \tilde{n}(x) < 0$ on a subdomain of $\partial B_\varepsilon(x^*) \subset \Omega$.[1] Then we obtain from the divergence theorem that

$$\begin{aligned} 0 \leq \int_{B_\varepsilon(x^*)} d(n(x), x)\, dx &= \int_{B_\varepsilon(x^*)} \nabla^T \left(e^{\phi(x)} \nabla \tilde{n}(x) \right) dx \\ &= \oint_{\partial B_\varepsilon(x^*)} e^{\phi(x)}\, \nu^T \nabla \tilde{n}(x)\, dS < 0, \end{aligned}$$

which is a contradiction.

[1] If $\nu^T \nabla \tilde{n}(x) = 0$, then $\tilde{n}(x)$ is constant in a neighborhood of x^* and one replaces x^* by a point from the boundary of that neighborhood. In the particular case that $\tilde{n}(x)$ is constant on the entire domain (6) follows immediately.

If we assume that the maximum is attained in $x^* \in \Gamma_N$, we choose $\varepsilon > 0$ such that $B_\varepsilon(x^*) \cap \Gamma_D = \emptyset$ is satisfied in addition to the conditions above. Now we conclude analogously

$$0 \leq \int_{B_\varepsilon(x^*)\cap\Omega} d(n(x), x)\, dx = \int_{B_\varepsilon(x^*)\cap\Omega} \nabla^T \left(e^{\phi(x)}\, \nabla\tilde{n}(x)\right) dx$$
$$= \underbrace{\oint_{\partial B_\varepsilon(x^*)\cap\Omega} e^{\phi(x)}\, \nu^T \nabla\tilde{n}(x)\, dS}_{<0} + \underbrace{\oint_{B_\varepsilon(x^*)\cap\Gamma_N} e^{\phi(x)}\, \nu^T \nabla\tilde{n}(x)\, dS}_{\leq 0} < 0,$$

which is again a contradiction. □

Remark 1 An analogous statement holds for the minimum if we have $d(n, x) \leq 0$ for $n < n_0(x)$ and $h(x) \geq 0$. For $h(x) = 0$, $x \in \Gamma_N$, we obtain upper and lower bounds. As a particular result we obtain the positivity of densities for positive $g(x)$ and non-negative $h(x)$.

Remark 2 The above assumptions reflect the simulation of semiconductors, where the drift-diffusion equation is used to describe the movement of electrons (or holes) in an electric field generated by the potential ϕ (see e.g. [9] for a detailed treatment). Dirichlet boundary conditions occur at semiconductor-metal interfaces with positive boundary values, while Neumann boundary conditions correspond to semiconductor-insulator interfaces where the normal of the current density is $h(x) = 0$.

The right hand side it then the generation-recombination rate, typically given as

$$d(n, x) = r(n, p)\,(np - N_{\text{intr}}^2),$$

where p is the hole density, N_{intr} the intrinsic density, and $r(n, p)$ is a model dependent positive factor. That is, the assumptions of Theorem1 are satisfied with $n_0(x) = \frac{N_{\text{intr}}^2}{p(x)}$.

In the complete semiconductor model n, p, and ϕ are solutions of the Rosenbroeck system, consisting of two drift diffusion equations for electrons and holes as well as the Poisson equation for the potential. However the assumption of arbitrary p and ϕ might be justified in a numerical scheme, as e.g. the Gummel iteration.

From the above theorem follows immediately a stability result

Corollary 1 *Let $n_i : \overline{\Omega} \to \mathbb{R}$, $i = 1, 2$ be continuous solutions of (1), which both fulfill Neumann boundary conditions (3) for the same arbitrary $h(x)$. If $d(n, x)$, is monotonically increasing with respect to n, then the Slotboom variables satisfy*

$$\left|\tilde{n}_1(x) - \tilde{n}_2(x)\right| \leq \max_{y\in\Gamma_D} \left|\tilde{n}_1(y) - \tilde{n}_2(y)\right|, \quad x \in \overline{\Omega}.$$

Proof Let us first assume that the maximum is attained in $x^* \in \Omega$. Without loss of generality we assume that $\tilde{n}_1(x^*) > \tilde{n}_2(x^*)$. Analogously to the proof of Theorem 1 we conclude

$$0 \leq \int_{B_\varepsilon(x^*)} d(n_1(x), x) - d(n_2(x), x)\, dx = \oint_{\partial B_\varepsilon(x^*)} e^{\phi(x)}\, \nu^T \nabla\big(\tilde{n}_1(x) - \tilde{n}_2(x)\big)\, dS < 0,$$

which is a contradiction.

If we assume that the maximum is attained in $x^* \in \Gamma_N$ the proof is analogous using that

$$\nu^T \nabla\big(\tilde{n}_1(x) - \tilde{n}_2(x)\big) = 0, \quad x \in \Gamma_N.$$

□

The above corollary states that small distortions of the Dirichlet boundary conditions will result only in small distortions of the solution. Although it is a statement for the Slotboom variable it provides also a stability statement for the densities based on relation (4). As a particular result we obtain here the uniqueness of the boundary value problem for the drift-diffusion equation.

For the time dependent drift-diffusion equation

$$\tfrac{\partial}{\partial t} n(x,t) = \nabla^T\big(\nabla n(x,t) - n(x,t)\nabla\phi(x,t)\big) - d\big(n(x,t), x, t\big), \tag{7}$$

the situation is more involved. However, we can still show the positivity of the densities, under suitable assumptions.

Theorem 2 *Let $n : \overline{\Omega} \times (0,T) \to \mathbb{R}$ be a continuous solution of (7) on the open domain $\Omega \times (0,T)$. We require positive Dirichlet (2) and nonnegative Neumann (3) boundary conditions, i.e., $g(x,t) > 0$ and $h(x,t) \geq 0$, as well as positive initial conditions $n(x,0) > 0$, $x \in \Omega$. If there is a $\delta > 0$ with $d(n,x,t) \leq 0$, $x \in \Omega \cup \Gamma_N$, $t \in (0,T)$, $n \leq \delta$, then*

$$n(x,t) > 0, \quad x \in \Omega \cup \Gamma_N,\ t \in (0,T).$$

Proof We assume there is a (x^*, t^*) such that $n(x^*, t^*) = 0$, while $n(x,t) > 0$ for $t < t^*$, $x \in \overline{\Omega}$. As in the proof of Theorem 1 we conclude

$$\int_{B_\varepsilon(x^*)\cap\Omega} \nabla^T\big(e^{\phi(x,t^*)}\, \nabla\tilde{n}(x,t^*)\big) - d(n(x,t^*), x, t^*)\, dx > 0$$

for any sufficient small $\varepsilon > 0$. Here we have assumed without loss of generality that $\tilde{n}(x,t^*)$ attains positive values in any neighborhood of x^*, which is admissible due to $g(x,t^*) > 0$. This implies that $\frac{\partial}{\partial t} n(x^*,t^*) > 0$, i.e., there is a $t < t^*$, with $n(x^*,t) < 0$, which is a contradiction. □

3 Discretization by Scharfetter-Gummel and Finite Volumes

The Scharfetter-Gummel scheme was introduced to avoid non-physical behavior resulting from conventional discretization techniques, e.g., central differences. The approach is based on the integration of the flux along the interval between two discretization nodes under the assumption of constant current density and electric field. Here, we present a different derivation which is based on midpoint finite differences and equality (5) for the Slotboom variable.

Consider the adjacent nodes x_i and x_j in a discetization grid. We denote by $h_{ij} = \|x_j - x_i\|$ their distance, the unit vector $\nu_{ij} = \frac{x_j - x_i}{h_{ij}}$ gives us the direction of the link and $z_{ij} = \frac{x_i + x_j}{2}$ is the midpoint. The flux in the direction of the link is given by

$$\nu_{ij}^T J = \nabla_\nu n - n \nabla_\nu \phi = \frac{\nabla_\nu \left(n\, e^{-\phi}\right)}{e^{-\phi}} = -\frac{\nabla_\nu \phi}{\nabla_\nu e^{-\phi}} \nabla_\nu \left(n\, e^{-\phi}\right),$$

where $\nabla_\nu := \nu_{ij}^T \nabla$ denotes the directional derivative. Using midpoint finite differences for all derivatives we obtain

$$\begin{aligned} \nu_{ij}^T J(z_{ij}) + \mathcal{O}(h_{ij}^2) &= -\tfrac{1}{h_{ij}} \frac{\phi_{ij} \left(n(x_j)\, e^{-\phi(x_j)} - n(x_i)\, e^{-\phi(x_i)}\right)}{e^{-\phi(x_j)} - e^{-\phi(x_i)}} \\ &= -\tfrac{\phi_{ij}}{h_{ij}} \left(\frac{n(x_j)}{1 - e^{\phi_{ij}}} - \frac{n(x_i)}{e^{-\phi_{ij}} - 1} \right) = \tfrac{1}{h_{ij}} \Big(n(x_j) B(\phi_{ij}) - n(x_i) B(-\phi_{ij}) \Big) =: J_{ij}, \end{aligned}$$

where $\phi_{ij} = \phi(x_j) - \phi(x_i)$ and $B(x) := \frac{x}{e^x - 1}$ is the Bernoulli function. That is, we have indeed obtained the Scharfetter-Gummel discretization (with an error of order h_{ij}^2). In terms of the Slotboom variables this becomes

$$J_{ij} = \tfrac{1}{h_{ij}} B(-\phi_{ij}) \big(n(x_j)\, e^{-\phi_{ij}} - n(x_i) \big) = \underbrace{\tfrac{1}{h_{ij}} B(-\phi_{ij})\, e^{\phi(x_i)}}_{w_{ij}} \big(\tilde{n}(x_j) - \tilde{n}(x_i) \big),$$

with

$$w_{ij} = w_{ji} = \tfrac{1}{h_{ij}} \frac{\phi(x_j) - \phi(x_i)}{e^{-\phi(x_j)} - e^{-\phi(x_i)}} > 0$$

(see also [10] for a different derivation).

To discretize (1) one uses a finite volume approach. Consider the grid nodes $\overline{X} = \{x_k : k \in \mathcal{I}\} \subset \overline{\Omega}$. We distinguish between the interior nodes $X := \Omega \cap \overline{X} = \{x_k : k \in \mathcal{I}_I\}$ and Dirichlet and Neumann boundary nodes $\partial X_D := \Gamma_D \cap \overline{X} = \{x_k : k \in \mathcal{I}_D\} \neq \emptyset$ and $\partial X_N := \Gamma_N \cap \overline{X} = \{x_k : k \in \mathcal{I}_N\}$, respectively. Here $\mathcal{I}$, $\mathcal{I}_I$, $\mathcal{I}_D$, and $\mathcal{I}_N$ are suitable finite index sets. The set of links $L = \{x_{ij} := (x_i, x_j)\} \subset X^2$

contains all pairs of adjacent nodes, with intersecting Voronoi cells ω_i and ω_j. The finite volume approach leads to

$$\int_{\omega_i} \nabla^T J(x)\, dx = \oint_{\partial\omega_i} \nu^T J(x)\big)\, dS \approx \sum_{j:x_{ij}\in L} A_{ij}\, J_{ij} + A_i^N\, h(x_i), \quad i \in \mathcal{I}_I \cup \mathcal{I}_N.$$

where A_{ij} and A_i^N are the size of $\omega_i \cap \omega_j$ and $\omega_i \cap \Gamma_N$. respectively. Note that for interior nodes $A_i^N = 0$. That is, we obtain the equations

$$\sum_{j:x_{ij}\in L} \underbrace{A_{ij}\, w_{ij}}_{W_{ij}} \big(\tilde{n}_j - \tilde{n}_i\big) = d(n_i, x_i) - A_i^N\, h(x_i), \quad i \in \mathcal{I}_I \cup \mathcal{I}_N, \tag{8}$$

to determine the approximative solution $\tilde{n}_i \approx \tilde{n}(x_i)$ for $i \in \mathcal{I}_I \cup \mathcal{I}_N$. Note that $W_{ij} = W_{ji} > 0$.

Writing the left hand side of (8) in matrix vector notation $\boldsymbol{M}\tilde{\boldsymbol{n}}$ one sees easily that $\boldsymbol{M} = (m_{ij})$ is symmetric, weakly diagonal dominant with $m_{ii} > 0$ and $m_{ij} \le 0$, $i \ne j$, as noted e.g. in [4]. From this relation we finally obtain a maximum principle for the discretized equations.

Theorem 3 *Let $\{n_i : i \in \mathcal{I}_I \cup \mathcal{I}_N\}$ be a solution of (8). If $h(x_i) \le 0$, $i \in \mathcal{I}_N$, and $d(n, x_i) \ge 0$, $n \ge n_0(x_i)$, $i \in \mathcal{I}_I \cup \mathcal{I}_N$, then*

$$\tilde{n}_i \le \max\left(n_0(x_i)e^{-\phi(x_i)},\ \max_{j\in\mathcal{I}_D} \tilde{n}_j)\right), \quad i \in \mathcal{I}.$$

Proof We use a similar argument as for the proof of Theorem 1. Let us assume that the maximum is attained for $i \in \mathcal{I}_I \cup \mathcal{I}_N$ and $n_i > n_0(x_i)$. Then

$$0 \le d(n_i, x_i) - A_i^N\, h(x_i) = \sum_{j:x_{ij}\in L} W_{ij} \big(\underbrace{\tilde{n}_j - \tilde{n}_i}_{<0}\big) < 0,$$

which is a contradiction. □

Using analogous arguments we obtain also discrete versions of Corollary 1.

Corollary 2 *Let $\{n_i : i \in \mathcal{I}_I \cup \mathcal{I}_N\}$ and $\{m_i : i \in \mathcal{I}_I \cup \mathcal{I}_N\}$ be solutions of (8), with identical $h(x)$. If $d(n, x)$ is monotonically increasing with respect to n, then the Slotboom variables satisfy*

$$\big|\tilde{n}_i - \tilde{m}_i\big| \le \max_{j\in\mathcal{I}_D} \big|\tilde{n}_j - \tilde{m}_j\big|, \quad i \in \mathcal{I}.$$

For the semi-discretized version of (7) we obtain also a discrete version of Theorem 2 for the positivity of the densities.

Theorem 4 *Let* $\{n_i(t) : i \in \mathcal{I}_I \cup \mathcal{I}_N\}$ *be a solution of*

$$\frac{\partial}{\partial t} n_i(t) = \sum_{j:x_{ij}\in L} A_{ij}\Big(n_j(t)B(\phi_{ij}(t)) - n_i(t)B(-\phi_{ij}(t))\Big) + A_i^N\, h(x_i) - d(n_i(t), x_i, t), \quad i \in \mathcal{I}_I \cup \mathcal{I}_N,$$

where $n_i(t) > 0$, $i \in \mathcal{I}_D$; $h(x_i) \geq 0$, $i \in \mathcal{I}_N$, *and* $n_i(0) > 0$, $i \in \mathcal{I}_I \cup \mathcal{I}_N$.

If there is a $\delta > 0$ *with* $d(n, x_i, t) \leq 0$, $i \in \mathcal{I}_I \cup \mathcal{I}_N$, $t \in (0, T)$, $n \leq \delta$, *then*

$$n_i(t) > 0, \quad i \in \mathcal{I},\ t > 0.$$

Acknowledgments This project has been funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Ref. No.: JU406/14-1 and the Austrian Science Fund (FWF): I3130-N30.

References

1. D.L. Scharfetter, H.K. Gummel, Large-signal analysis of a silicon read diode oscillator. IEEE Trans. Electron Devices **16**(3), 391–415 (1969)
2. J. Jerome, *Analysis of Charge Transport: A Mathematical Study of Semiconductor Devices* (Springer, Berlin, Heidelberg, 1996)
3. P. Markowich, *The Stationary Semiconductor Device Equations*. Computational Microelectronics (Springer, New York, 1986)
4. K. Gärtner, Existence of bounded discrete steady-state solutions of the van Roosbroeck system on boundary conforming Delaunay grids. SIAM J. Sci. Comput. **31**(2), 1347–1362 (2009)
5. M. Bessemoulin-Chatard, C. Chainais-Hillairet, A. Jüngel, Uniform L^∞ estimates for approximate solutions of the bipolar drift-diffusion system. in Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects (Springer International Publishing, Cham, 2017), pp. 381–389
6. Z. Kargar, T. Linn, D. Ruić, C. Jungemann, Investigation of transport modeling for plasma waves in THz devices. IEEE Trans. Electron Dev. **63**(11), 4402–4408 (2016)
7. J.W. Slotboom, Computer-aided two-dimensional analysis of bipolar transistors. IEEE Trans. Electron Dev. **20**(8), 669–679 (1973)
8. J. Jerome, T. Kerkhoven, L^∞ stability of finite element approximations to elliptic gradient equations. Numer. Math. **57**(6–7), 561–576 (1990)
9. P. Farrell, N. Rotundo, D.H. Doan, M. Kantner, J. Fuhrmann, T. Koprucki, Drift-diffusion models, in *Handbook of Optoelectronic Device Modeling and Simulation*, ed. by J. Piprek. Series in Optics and Optoelectronics, vol. 2 (CRC Press, Boca Raton, 2017), pp. 733–771
10. R.E. Bank, D.J. Rose, W. Fichtner, Numerical methods for semiconductor device simulation. IEEE Trans. Electron Dev. **30**(9), 1031–1041 (1983)

Numerical Calculation of Electronic Properties of Transition Metal-Doped mWS_2 via DFT

Chieh-Yang Chen and Yiming Li

Abstract In this work, we use the spin-polarized density functional theory (DFT) to study the atomic structures of transition metal-doped monolayer WS_2 (mWS_2). The structures of doped mWS_2 are simulated via atomic relaxation which moves the ions according to the interactive force between electrons and the ions until converge condition is reached, where the Kohn-Sham equation is solved numerically. We do reveal not only simulation flow but also the accuracy examination for the explored mWS_2. The estimated physical properties are further described and discussed.

1 Introduction

Two-dimensional materials, the monolayer transition metal dichalcogenide disulfide (TMD), feature a high on/off ratio, low power consumption, and thermal stability, especially the direct energy band gap of monolayer structure becoming eye-catching study issues. Our recent study revealed the key steps for the stability of doping sites for discussing electronic properties of TMD materials [1–3]. For WS_2, although some doping techniques on monolayer tungsten disulfide have been reported, they only focused on the certain doping material [4, 5]; thus, in this work, we analyze the doping sites, formation energy, work function, and charge transfer of mWS_2 with $3d$ transition metals doped mWS_2. The considered doping materials consist of scandium (Sc), titanium (Ti), vanadium (V), chromium (Cr), manganese (Mn), iron (Fe), cobalt (Co), nickel (Ni), copper (Cu), and zinc (Zn). We use symbols 4×4 and 2×2 as the dimension of the supercell containing a dopant, and the aforementioned effective doping concentrations are 2.04% and 7.69% in the atomic percentage, corresponding to 7.13×10^{13} and 2.85×10^{14} cm^{-2}, respectively.

The doped mWS_2 structure can be obtained via atomic relaxation process which is an iteration of ion moving steps according to electron charge density. The ions are

C.-Y. Chen · Y. Li (✉)
Parallel and Scientific Computing Laboratory, Institute of Communications Engineering, National Yang Ming Chiao Tung University, Hsinchu City, Taiwan
e-mail: ymli@faculty.nctu.edu.tw

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_6

moved by conjugate gradient algorithm, where the direction decided by calculated force and stress tensor. Then the electron distribution is updated according to new ion locations. By solving the Kohn-Sham equation constructed from electrons' and ions' relation, the total energy, force, and stress tensor are updated. These steps are repeated until the force and stress tensor reach the convergence condition. The exchange-potential term in the Kohn-Sham equation [6] is critical in DFT approach, and it can include other corrections such as van der Waals correction. Deriving from an approximated Schrödinger equation, the Kohn-Sham equation is solved by blocked Davidson algorithm [7]. Notably, the suitable correlation-exchange function is first examined by experimental values and the sampling k-points are tested.

2 The Computational Model

We are interested in the electronic properties of advanced material such as mWS_2. To achieve the correct doped mWS_2, the spin-polarized DFT is used for atomic level structure relaxation process and studying electronic properties. The first step of simulation flow is to construct the initial atomic geometry based on the periodic boundary condition of a given system. In the second step, we calculate the wavefunction and electron density in the ground state by self-consistent electronic structure calculation. However, these electronic properties are not in the equilibrium condition because of the Coulomb interaction between ions. We relax the ions configuration to its equilibrium state based on the atomic forces calculated by Hellmann-Feynman theorem using the ground-state wavefunction. Once the atomic forces between arbitrary two atoms are smaller than a tolerance, we recalculate the electronic properties (wavefunction distribution, electron density, band structure, and density of states) based on the relaxed atomic structure and given wavevector.

Starting from the electronic Schrödinger equation and the Hartree–Fock approximation [6] based on the molecular orbital theory, the Kohn-Sham equation in (1) is a simplified form which mainly constructed from electron density. The Hohenberg-Kohn theorem [8] states that the ground-state energy of a given potential distribution, i.e., atomic configuration, is a unique functional of electron density.

$$\begin{aligned}&(T_e + V_{nuclear} + V_{Hatree} + V_{xc})\phi_i \\ =&(\frac{p_i^2}{2m_i} + \sum_I \frac{-Z_I e^2}{\left|r_i - \vec{R_I}\right|} + e^2 \int \frac{n(r')}{\left|r - \vec{r'}\right|} d\vec{r'} + V_{xc})\phi_i \\ =&\varepsilon_i \phi_i, i = 1, 2, \ldots, N,\end{aligned} \tag{1}$$

where T_e, $V_{nuclear}$, V_{Hatree}, and V_{xc} are the kinetic energy operator of electron i, nucleus-electron potential energy, Hartree potential energy, and exchange-correlation potential energy, respectively. p_i, m_i, I, Z_I, e, r_i, R_I, ε_i, $\phi_i(\vec{r})$, and

N are the momentum operator, electron mass, the index of nucleus, the charges of nucleus, the charge of electron, the position vector of electron, the position vector of nucleus, the orbital energy, the Kohn-Sham orbital, and N non-interacting electrons, respectively. The number of these equations depends on the number of valence electrons which come all of the simulated atoms. The $V_{xc}(\vec{r})$ term can be obtained by different approaches; for example, Perdew-Burke-Ernzerhof (PBE) [9] is used as an exchange-correlation function after our intensive accuracy test. The aforementioned equations are construct for single k-point. For sampling of Brillouin zone, Monkhorst-pack [10] k-points centered at the Γ point (0, 0, 0) is generated:

$$\vec{k} = \sum_{i=1}^{3} \vec{b_i} \frac{n_i + 1/2}{N_i}, n_i = 0, \ldots, N_i - 1, N_i \text{is even}, \tag{2}$$

where $\vec{k}$, $\vec{b_i}$, and N_i are the vector in k-space, the basis of reciprocal lattice, and the mesh numbers for $\vec{b_1}$, $\vec{b_2}$, and $\vec{b_3}$ directions.

To solve the Kohn-Sham equation, the blocked Davidson algorithm[7] has been considered in the numerical calculation. It consists of five steps: basis initialization, subspace construction, residual vector calculation, correction vector calculation, and subspace expansion. In the basis initialization, a set of orthonormal basis $\psi_i, i = 1, 2, \ldots, m, \forall m \geq n$ for the lowest n states are guessed and built. In the subspace construction, the full-size Hamiltonian matrix $\vec{H}$ is projected on a set of sub-matrices $\{\tilde{\vec{H}}_{ij} = \psi_i^T \vec{H} \psi_i\}$ and solved for the eigenpairs in the subspace $(\tilde{\vec{H}} \varphi^k = \varepsilon^k \varphi^k, \forall k = 1, 2, \ldots, n)$. Next, the calculated eigenpairs are used for residual vector calculation. The residual vector is defined as $\vec{r}_i = (\vec{H} - \varepsilon_i \vec{I})\varphi_i$, while ε_i and φ_i are the eigenvalue and eigenvector of sub-matrices. We check the individual element in the residual vector for the convergence. If the elements are larger than tolerance, we calculate the correction vector based on the residual vectors and the eigenpair of sub-matrices. However, several ways evolve to calculate the correction vector and result in different branches of Davidson algorithm.

The correction vectors $\{g^k, k = 1, 2, \ldots, n\}$ are given by $g_I^k = (\varepsilon^k - \vec{H}_{II})^{-1} r_I^k, I = 1, 2, \ldots, N$ and normalized, while N is the number of determinants of $\vec{H}$ and $r^k = \sum_{i=1}^{m} \varphi_i^k (\vec{H} - \varepsilon^k)\psi_i$. In the final step, the correction vectors $\{g^k, k = 1, 2, \ldots, n\}$ orthonormalized against the set $\{\psi_i, i = 1, 2, \ldots, m\}$ using Gram-Schmidt process and appended in the set $\{\psi_i\}$ if the orthonormalized norm value is larger than a threshold said 10^{-3}. The resulting number of basis might increase by a, while $1 \leq a \leq n$. The whole process returns to subspace construction using the updated orthonormal basis $\{\psi_i, i = 1, 2, \ldots, m', m' = m + a\}$ until the residual vectors reach convergence.

Under the Kohn-Sham formalism, the DFT was developed based on the local and semi-local functionals, such as local density and generalized-gradient approximation. However, it may not work well in describing the long-range charge dynamics such as van der Waals interaction. This may cause a significant inaccuracy in delineating the system energy, lattice constant, and electronic prop-

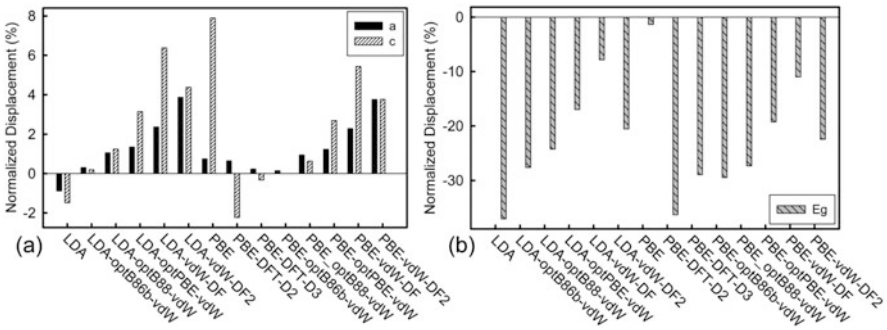


Fig. 1 The comparison between experimental and simulation values from different van der Waals models. (**a**) The displacement of lattice parameters, $\vec{a}$ and $\vec{c}$. (**b**) The displacement of bandgap, E_g

erty of two-dimensional materials because it is the van der Waals interaction that forms stacked structure in two-dimensional materials. Thus, for different materials, the examination and comparison to experimental value are necessary. For bulk WS_2, the comparisons between experimental[11] and simulation values with different exchange-correlation functions are shown in Fig. 1. We examine the exchange-correlation functions combining local-density approximation (LDA), Perdew-Burke-Ernzerhof (PBE), and van der Waals correlation. The discussed van der Waals model includes DFT-D2, DFT-D3, optB86b-vdW, optB88b-vdW, optPBE-vdW[12], vdW-DF[13], and vdW-DF2[14]. Due to abundant exchange-correlation functions have been established there for different atoms, Eq. (1) with different van der Waals correlations are solved by using Vienna *ab initio* Simulation Package (VASP)[15].

We normalize the displacement between simulation and experimental value: $\text{Displacement}(\%) = (X_{sim} - X_{exp})/X_{exp} \times 100\%$, where X_{sim} and X_{exp} are values from simulation and experiment, respectively. Figure 1a shows that LDA-optB86b-vdW, PBE-DFT-D3, and PBE-optB86b-vdW can achieve very small displacement of lattice parameters compared with the experimental data[11]. Among all lattice parameters, PBE has the largest displacement of the lattice parameter $\vec{c}$, but it is merely about 7.89% overestimation. Notably, a relatively small displacement along the lattice parameter $\vec{a}$ implies that the strain along horizontal direction can be properly described. Figure 1b shows that these functions all underestimate the bandgap compared with the experimental data[11]. PBE function shows the smallest displacement of bandgap which only has 1.34% underestimation while other exchange-correlation functions have large underestimation. Note that the vdW-DF and vdW-DF2 corrections result in incorrect indirect bandgap position for the bulk WS_2. Our calculation indicates that the bandgap of bulk WS_2 mainly depends on the strain of $\vec{a}$- and $\vec{b}$-direction instead of it of $\vec{c}$-direction.

The previous examination of the different exchange-correlation functions is with a large number of Brillouin zone sampling which is performed by Monkhorst-pack k-points. To find proper numbers of k-points which can shorten computation time

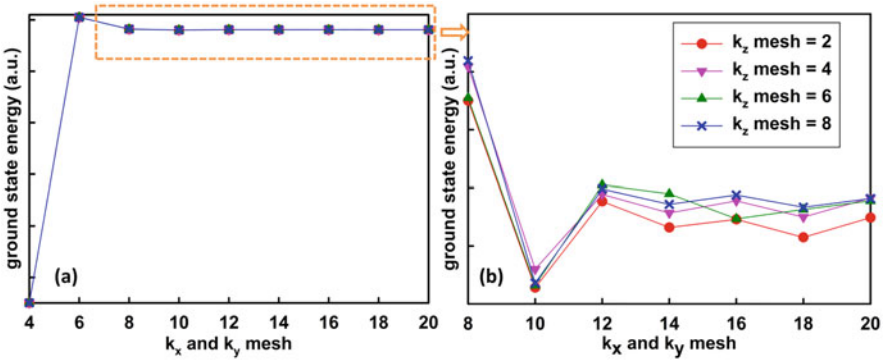


Fig. 2 (**a**) The ground state energies of bulk WS_2 with different examined numbers of k-points. Because bulk WS_2 have the same length of lattice parameters $\vec{a}$ and $\vec{b}$, the numbers of k_x and k_y are set equal and it of k_z varies. (**b**) A zoom-in plot of (**a**)

and preserve good accuracy, we fix the bulk WS_2 structure cell and alter the number of k-points. Since the lattice parameters $\vec{a}$ and $\vec{b}$ of bulk WS_2 have the same length, the numbers of k_x and k_y are set to be equal. Figure 2 shows the ground state energies of bulk WS_2 versus the different numbers of k-points. The k-points of k_x and k_y became stable when they are larger than 8. Notably, they are sensitive due to the nature of the two-dimensional materials, where the k_z of 12 is the most insensitive to k_x and k_y numbers. The numbers of suitable k-points is related to the lattice parameter; lengths of $\vec{a}$ and $\vec{c}$ are 3.1532 and 12.323 angstrom from the experiment [11], respectively. The length of $\vec{c}$ is nearly 4 times to $\vec{a}$, thus the suitable k-points of k_z will be around one-fourth to it of k_x. Since the monolayer WS_2 has an additional vacuum layer and similar lattice parameter $\vec{a}$ and $\vec{c}$ compared to bulk WS_2, the numbers of k_x and k_y are the same as bulk WS_2 but less for k_z. The number of k_z of monolayer WS_2 is set to 1 due to its long lattice parameter $\vec{c}$. Finally, $12 \times 12 \times 4$ and $12 \times 12 \times 1$ k-points for bulk and monolayer WS_2 are concluded, respectively.

Our structure relaxation flow of monolayer WS_2 is shown in Fig. 3a. First, we obtain the reliable bulk WS_2 from previous examined settings, and the band structure is shown in Fig. 3b. From the results of exchange-correlation function and k-points examinations, they indicate the importance of correct structure relaxation along $\vec{a}$ and $\vec{b}$; thus, the atomic force along these two directions should be preserved. For this reason, monolayer WS_2 is built from relaxed bulk WS_2 which adds a 10 angstrom thick vacuum layer along $\vec{c}$ and fixes the simulation cell. Figure 3c shows verified calculated band structure of monolayer WS_2. Confidently, the calculated energy bandgaps from this method are in agreement with the experiment [11]. Then, we continuously analyze the TM-doped mWS_2. The discussed characteristics of TM-doped mWS_2 include formation energy, work function, and band structure. As

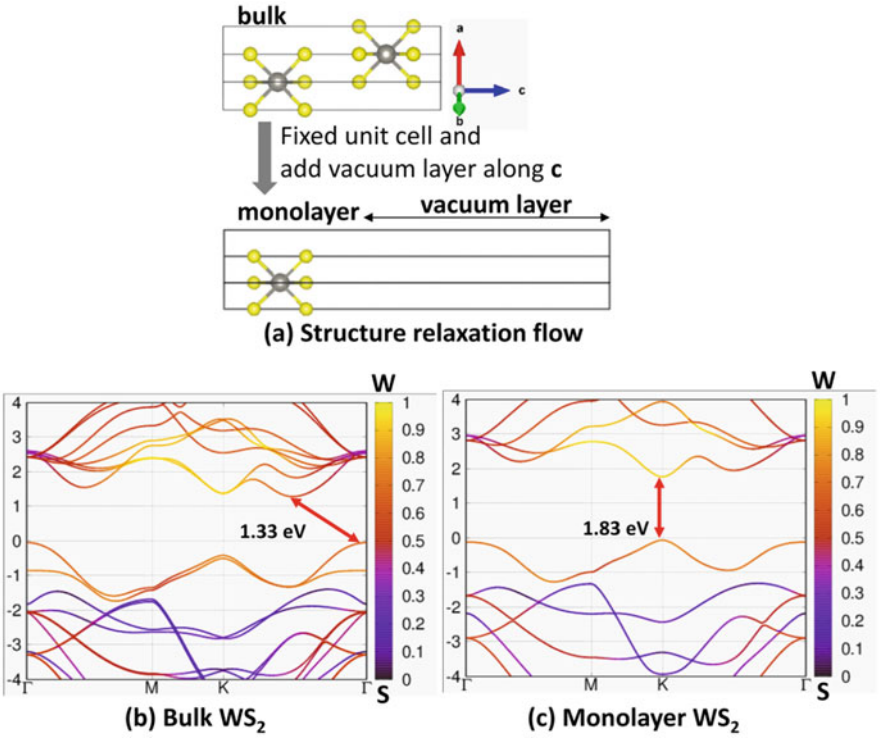


Fig. 3 (**a**) The structure relaxation flow of monolayer WS_2. The monolayer structure relaxation is complete under fixed unit cell for maintaining same atomic force along $\vec{a}$ and $\vec{b}$ directions. The cutoff kinetic energy is 500 eV; the force acting on each atom of relaxed structure is smaller than 0.01 eV/angstrom; the energy difference is less than 10^{-6} eV per atom. From the simulation, the atom-projected band structure of (**b**) bulk WS_2 and (**c**) monolayer WS_2 are obtained. The color bars indicate the weighting of band dominated by tungsten atoms [1]

shown in Fig. 4a, because the new structure is built from different materials, we consider the formation energy formula:

$$E_{form} = E_{doped,mWS2} - E_{mWS2} + \sum n_i \mu_i, \tag{3}$$

where $E_{doped,mWS2}$ and E_{mWS2} are the total energies of the doped mWS_2 system and the pristine mWS_2, n_i and μ_i are the number of atom i added (-1) or removed (+1) and the corresponding chemical potential, respectively. The four possible doping sites are discussed, as illustrated in Fig. 4b. To study the effect result from two different doping concentrations, 4×4 and 2×2 supercells with one doping atom were built from pristine mWS_2. The definition of work function is the external energy exciting an electron from the surface of solid material into the vacuum space. It can be calculated from the energy difference between the simulated vacuum energy and the Fermi level, i.e., $E_{vacuum} - E_{fermi}$.

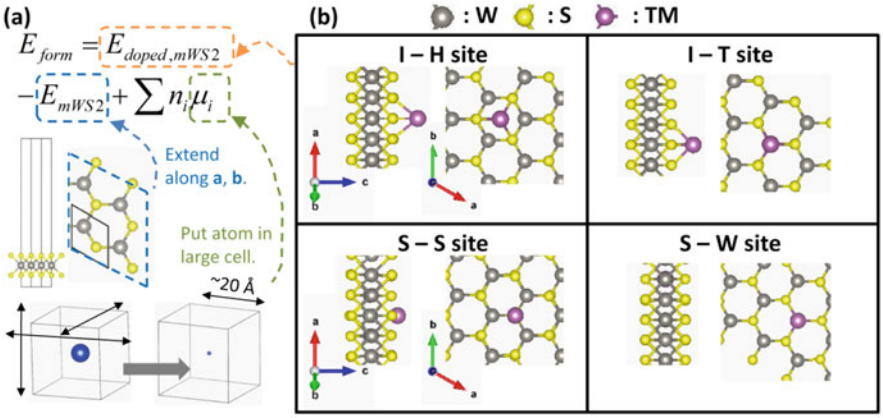


Fig. 4 (**a**) The formula of the formation energy calculation. The $E_{doped,mWS2}$ and E_{mWS2} are total energies of relaxed doped- and undoped-monolayer WS_2, respectively. For the calculation of chemical potential, μ_i, an atom is located in a large cell. The potential can be obtained until the value is stable by keeping enlarging the cell size. (**b**) The structures of four possible doping sites. The "TM" means the atom of doping material. The notations "I-" and "S-" mean interstitial and substitutional sites, respectively. The structures are built from repeatedly extended monolayer cell and add doping atom for different doping concentrations

3 Results and Discussion

From the atomic relaxed structures, we can plot the band structure according to the solved eigen energies, as shown in Fig. 5a and b. Both two plots are shifted so that the simulated Fermi energy is located at zero. Since our simulated Fermi energy is located at the band which is occupied by the last valence electron, i.e., the band contributed by the doping atom. For example, comparing to Fig. 5b and a show that the Sc doping contributes additional bands between the original conduction and valence band. The calculated formation energy of discussed doping sites with two different concentration are summarized in Fig. 5c. The formation energies of discussed interstitial sites are lower than that of the substitutional sites, it indicates the structure stability. Figure 5d plots the work function of pristine and TM-doped mWS_2 with respect to two concentrations. For simplicity, only the results of I-T doping site are shown. The titanium (Ti)-doped mWS_2 has the lowest work function with higher concentration while zinc (Zn)-doped has the highest one with both concentrations. The Sc possesses the largest range of modulation of work function, 1.63 eV, among discussed doping species and concentrations. It implies that there is high flexibility in tuning work function of mWS_2 which is promising for the design and fabrication of future advanced nano-devices.

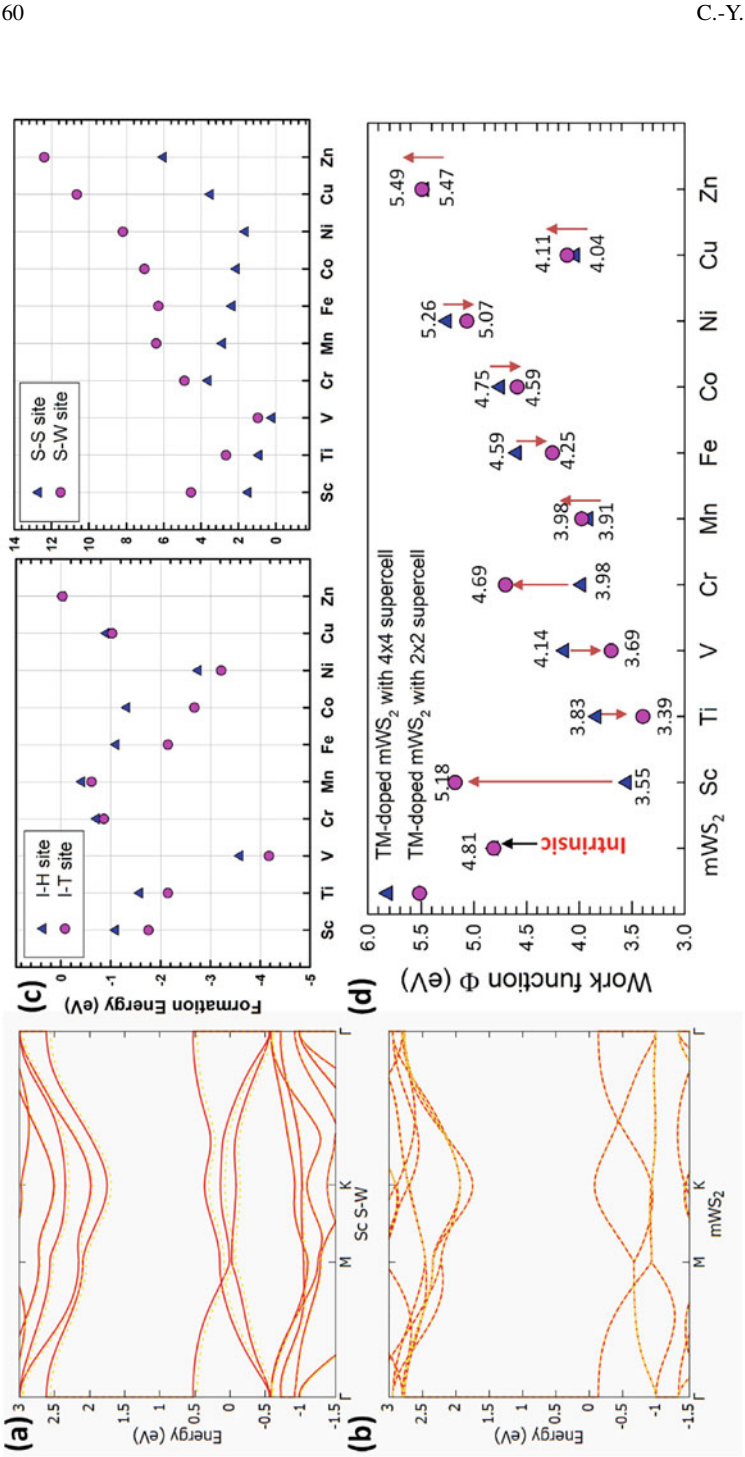


Fig. 5 The band structure of (**a**) a Sc substitutional doping which a tungsten atom is replaced and (**b**) the undoped monolayer WS_2. The solid red lines and dotted orange lines are of spin-up and spin-down states, respectively. (**c**) The calculated formation energy of discussed doping sites. (**d**) The work functions of TM-doped mWS_2 with respect to different TM materials and concentrations. Here only shows the results of the doping site with the lowest formation energy. The arrows indicate how the work function changes as the doping concentration increases

4 Conclusions

In this work, the numerical method and simulation flow for studying doped monolayer tungsten disulfide have been described. The studies are completed with the examined exchange potential model and k-points sampling. The key simulated results indicate the values of work function of Sc- and Cr-doped mWS_2 have relatively large flexibility for work function modulation via doping technology.

Acknowledgments This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-009-008, Grant MOST 108-3017-F-009-001, Grant MOST 109-2221-E-009-033, Grant MOST-109-2634-F-009-030, and Grant MOST 110-2221-E-A49-139, and in part by the "Center for mmWave Smart Radar Systems and Technologies"İ under the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

References

1. C.-Y. Chen, Y. Li, On electronic structure of monolayer tungsten disulfide doped by 3*d* transition metals, in *Proceedings of International Conference on Solid State Devices and Materials SSDM 2019*, Nagoya, September 2–5 (2019), pp. 161–162
2. Y.-C. Tsai, Y. Li, On electronic structure and geometry of MoX_2 (X = S, Se, Te) and black phosphorus by ab initio Simulation with various Van der Waals corrections, in *Proceedings of IEEE Int. Conf. Simul. Semicond. Process. Devices SISPAD 2017*, Kamakura, September 7–9, 2017, pp. 169–172
3. Y.-C. Tsai, Y. Li, Impact of doping concentration on electronic properties of transition metal-doped monolayer molybdenum disulfide. IEEE Trans. Electron Devices **65**, 733–738 (2018)
4. S. Sasaki et al., Growth and optical properties of Nb-doped WS_2 monolayers. Appl. Phys. Express **9**, 71201 (2016)
5. C. Sun et al., N-doped WS_2 nanosheets: a high-performance electrocatalyst for the hydrogen evolution reaction. J. Mater. Chem. A **4**, 11234–11238 (2016)
6. W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects. Phys. Rev. **140**, A1133–A1138 (1965)
7. E.R. Davidson, Matrix eigenvector methods, in Methods in Computational Molecular Physics, ed. by G.H.F. Diercksen, S. Wilson (Springer, Dordrecht, 1983), pp. 95–113
8. P. Hohenberg, W. Kohn, Inhomogeneous electron gas. Phys. Rev. **136**, B864–B871 (1964)
9. J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. Phys. Rev. Lett. **77**, 3865–3868 (1996)
10. H.J. Monkhorst, J.D. Pack, Special points for Brillouin-zone integrations. Phys. Rev. B **13**, 5188–5192 (1976)
11. R. Lv et al., Transition metal dichalcogenides and beyond: synthesis, properties, and applications of single- and few-layer nanosheets. Acc. Chem. Res. **48**, 56–64 (2015)
12. J. Klimeš, D.R. Bowler, A. Michaelides, Chemical accuracy for the van der Waals density functional. J. Phys. Condens. Matter **22**, 022201 (2010)
13. G. Román-Pérez, J.M. Soler, Efficient implementation of a van der Waals density functional: application to double-wall carbon nanotubes. Phys. Rev. Lett. **103**, 96102 (2009)
14. K. Lee, Ë.D. Murray, L. Kong, B.I. Lundqvist, D.C. Langreth, Higher-accuracy van der Waals density functional. Phys. Rev. B **82**, 022201 (2010)
15. G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals. Phys. Rev. B **47**, 558–561 (1993)

Numerical Simulation of Thermal Conductivity of Silicon Nanowires

Min-Hui Chuang and Yiming Li

Abstract To provide the sufficient power of trillion sensors in the era of internet-of-things, the thermoelectric materials and devices have been of great interest recently. In this paper, we construct a model for the periodic silicon nanowires (SiNWs) embedded in $Si_{0.7}Ge_{0.3}$ (SiNWs-$Si_{0.7}Ge_{0.3}$ composite) and propose a simulation flow for the calculation of its thermoelectric properties. The electron band structure and phonon energy dispersion of SiNWs-$Si_{0.7}Ge_{0.3}$ composite are simulated by using the effective mass Schrödinger equation formulated by the Bloch theorem and the elastodynamic wave equation, respectively. The aforementioned equations are discretized by using the finite element method and the corresponding eigenvalue problems are solved by the implicitly restarted Arnoldi method. Then, the thermoelectric properties of SiNWs-$Si_{0.7}Ge_{0.3}$ composite are estimated by Landauer approach.

1 Introduction

For thermoelectric (TE) energy conversion materials in solid-state power generation, the dimensionless figure of merit (FOM) given by

$$ZT = \frac{S^2 \sigma T}{\kappa_{ph} + \kappa_{el}} \tag{1}$$

is used to indicate the TE performance, where S is the Seebeck coefficient, σ is the electrical conductivity, κ_{ph} is the thermal conductivity from phonon, and κ_{el} is the thermal conductivity from electron. To increase the value of ZT, researchers are either focus on reducing the thermal conductivity or enhancing the power factor,

M.-H. Chuang · Y. Li (✉)
Parallel and Scientific Computing Laboratory, Institute of Communications Engineering, Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
e-mail: ymli@faculty.nctu.edu.tw

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_7

$S^2\sigma$. The Landauer approach can be used to calculate the TE properties; however, these properties are determined by the electron and phonon energy dispersions but not independently controlled. Recent experiments have shown that the lattice thermal conductivity would be significantly reduced without suffering from the loss of power factor by using nanostructures [1–3]. With these nanostructures, the lattice thermal conductivity can be suppressed due to the phonon boundary scattering. Furthermore, alloys shows a better TE performance with respect to nonalloys because of the reduction of the lattice thermal conductivity [4]. Studies on nanocomposite TEs show that proper nanostructures in SiGe nanocomposite materials can lead to a reduction in the thermal conductivity [5].

In this paper, the conductive matrix material of $Si_{0.7}Ge_{0.3}$ is studied [6]. This composite film consists of the SiNWs embedded in $Si_{0.7}Ge_{0.3}$ and is modelled as a periodic superlattice for the following simulation. The electron and phonon energy dispersions are solved from two eigenvalue problems. To solve the Schrödinger equation, researches give several approaches in which the finite-element discrete variable representation plays an important role to get an efficient and highly accurate result [7]. This paper is organized as follows. In Sect. 2, we show the simulation structure and the physical setting. In Sect. 3, we show the numerical approach for the calculation of the band profile and TE properties. A part of the simulation results and discussions are illustrated in Sect. 4. Finally, we draw the conclusions in Sect. 5.

2 Computational Structure and Models

As shown in Fig. 1, the direction of the carrier and phonon transports is parallel to the x-y plane so that the nanowires can play as interface for the phonon transport and reduce the thermal conductivity. To simplify the simulation structure, we assume that the nanowires are periodic with the radius r, the space s between the closest two nanowires, and the height h, as shown in Fig. 1b. For the band profile calculation,

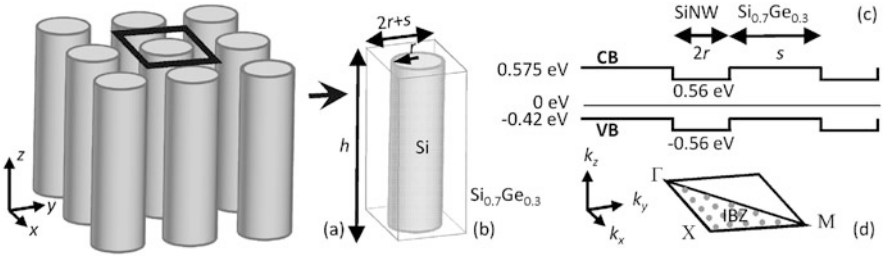


Fig. 1 (**a**) The three-dimensional (3D) schematic structure of the periodic nanowires. (**b**) The geometry parameter of the simulation structure. (**c**) The CB and VB of SiNWs-$Si_{0.7}Ge_{0.3}$ composite [10]. (**d**) The definition of irreducible Brillouin zone (IBZ), where Γ, X, and M are in the x-y plane of the k-space. Γ is the original point

Table 1 The adopted physical parameters. m_l^* and m_t^* are the effective masses of electrons in the longitudinal and transverse directions, respectively. They are used for the quantized energy band calculation in the CB. Similarly, m_{hh}^* and m_{lh}^* are the effective masses of heavy holes and light holes, respectively. They are used for quantized energy band calculation in the VB [10, 13, 14]

Material	Bandgap	Electron mass		Hole mass		Elastic constant		
		m_l^*	m_t^*	m_{hh}^*	m_{lh}^*	C_{11}	C_{12}	C_{44}
Si	1.12	0.98	0.19	0.49	0.16	165.8	63.9	79.6
$Si_{0.7}Ge_{0.3}$	1.00	1.14	0.12	0.41	0.10	154.6	59.2	75.8

we simplily consider the undoped situation, where the conduction and valence bands are plotted in Fig. 1c. For an electron or a hole in a periodic potential, the Bloch theorem [8] is used to describe the phase change; thus, the corresponding Schrödinger equation with effective mass approximation is given by [9]

$$\nabla[\frac{-\hbar^2}{2m^*}\nabla\vec{u}_k] - \frac{i\hbar^2}{m^*}\vec{k}\cdot\nabla\vec{u}_k(\vec{r}) + [V(\vec{r}) + \frac{\hbar^2k^2}{2m^*}]\vec{u}_k(\vec{r}) = E_{n,k}\vec{u}_k(\vec{r}), \tag{2}$$

where $\hbar$, m^*, $V(\vec{r})$, $E_{n,k}$, and $\vec{u}_k(\vec{r})$ are the reduced Plank's constant, the effective mass, the position-dependent potential energy, quantum energy levels, and the corresponding wave function, respectively. Notably, $V(\vec{r})$ is equal to the conduction band (CB) or valence band (VB) for electrons and holes [10], respectively. In addition, the phononic band structure is calculated by the elastodynamic wave equation [11]

$$\nabla\cdot[\vec{C}\nabla\vec{u}(\vec{r})] = \rho\omega^2\vec{u}(\vec{r}), \tag{3}$$

where $\vec{C}$ is the elastic constant matrix, $\vec{u}$ is the Fourier transform of the displacement vector [12], ρ is the mass density, and ω is the eigenfrequency, respectively. The elastic constant matrix $\vec{C}$ describes second-order strain energy density. Since Si has cubic symmetry, the number of independent elastic constants can be reduced to C_{11}, C_{12}, and C_{44} [13]. The elastic constants of $Si_{0.7}Ge_{0.3}$ are decided by the linear interpolation of the values from Si and Ge, as listed in Table 1.

3 Simulation Techniques

To estimate the FOM in (1), by considering the physical transparency and the computational efficiency of the solution method, the Landauer approach has been implemented on the TE region [15–17]. In situations close to equilibrium, the Landauer approach is mathematically equivalent to the Boltzmann transport

equation under the relaxation time approximation if the mean-free-path (MFP) for backscattering is

$$\langle\langle \lambda(E) \rangle\rangle = \frac{2\langle v_x^2 \tau \rangle}{\langle |v_x| \rangle}, \tag{4}$$

where v is the group velocity, τ is the momentum relaxation time, and the subscription x represents the transport direction of the carriers or phonons [16]. Within the Landauer approach, the number of modes and the MFP for backscattering are two important physically parameters. The calculation flows for the TE properties related to the electrons and phonons are listed in Algorithms 1 and 2, respectively.

In Algorithm 1, the differential conductivity is given by [15]

$$\sigma'(E) = \frac{2q^2}{h} \lambda_e(E) \frac{M_e(E)}{A} (-\frac{\partial f_0}{\partial E}), \tag{5}$$

where $2q^2/h$ is the quantum of the conductance, λ_e is the MFP of the electron transport, M_e/A is the number of modes per area A, and $\partial f_0/\partial E$ is the window function. f_0 is the equilibrium Fermi-Dirac function which is related to the band structure in the CB. The differential lattice thermal conductivity in Algorithm 2 is given by [15]

$$\kappa'_{ph}(\hbar\omega) = \frac{\pi^2 k_B^2 T}{3h} \lambda_{ph}(\hbar\omega) \frac{M_{ph}(\hbar\omega)}{A} (\frac{3}{\pi^2})(\frac{E}{k_B T})^2 (-\frac{\partial n_0}{\partial \hbar\omega}), \tag{6}$$

where $\pi^2 k_B^2 T/3h$ is the quantum of the thermal conductance, λ_{ph} is the MFP of the phonon transport, M_{ph}/A is the number of modes per area, and $(\frac{3}{\pi^2})(\frac{E}{k_B T})^2(-\frac{\partial n_0}{\partial \hbar\omega})$ is the window function. n_0 is the Bose-Einstein distribution which is related to the phonon dispersion. The carriers are under the diffusion transport with the MFP for back-scattering calculated by the Matthiessen rule, where the average MFP for back-scattering without nanowire structure is extracted by setting the thermal conductivity of 150 W/m-K from the measured data of bulk silicon [18].

For the calculation of the electronic and phononic band structures, the boundary conditions of (2) and (3) are set periodically owing to highly periodical array of SiNWs [6]. We solve these two discretized eigenvalue problems by the implicitly restarted Arnoldi method [19]. The finite element method with Lagrange elements is implemented to discretize the Schrödinger and elastodynamic equations. A finite element is a triple including a geometry domain, a space function in this domain, and a set of linear functionals (so-called the degree of freedom) [20]. The band structure is calculated by sampling in k-space, more specifically, in the irreducible Brillouin zone (IBZ) [21], as shown in Fig. 1d. The calculation flow to get the band diagrams is listed in Algorithm 3.

Algorithm 1: The Landauer approach for electronic TE proprieties

Require: E-k relationship and the energy upper limitation E_m
Ensure: Calculate S, σ, and κ_e
1: **while** energy $E < E_m$ **do**
2: Calculate the differential electrical conductivity σ'
3: Conductivity $\sigma += \sigma'$
4: $\kappa_0 += (E - E_F)^2\sigma'$
5: Seebeck coefficient $S += (E - E_F)\sigma'$
6: **end while**
7: $\kappa_0 = \kappa_0/q^2T$
8: $S = -S/qT\sigma$
9: $\kappa_e = \kappa_0 - T\sigma S$

Algorithm 2: The Landauer approach for phononic TE proprieties

Require: $\hbar\omega$-q relationship and the frequency upper limitation ω_m
Ensure: Calculate κ_{ph}
1: **while** frequency $\omega < \omega_m$ **do**
2: Calculate the differential lattice thermal conductivity κ'_{ph}
3: $\kappa_{ph} += \kappa'_{ph}$
4: **end while**

Algorithm 3: The band diagram calculation by sampling some specific points in IBZ

Require: effective masses, the geometry structure, and the sampling points
Ensure: eigenvalues or eigenfrequencies
1: **while** There is at least one point in IBZ which has not been solved. **do**
2: Assign the value of $\vec{k}$
3: Solve the eigenvalue or eigenfrequency problem at $\vec{k}$
4: Record the eigenvalues or eigenfrequencies at $\vec{k}$
5: **end while**
6: Connect the i^{th} eigenvalue or eigenfrequency at all sampling points
7: Output the band structure

4 Results and Discussion

Because h is relatively ten times larger than r, we forcus on the 2D simulation. The first ten lowest energies of electrons and light holes calculated from (2) are shown in Fig. 2. When $s = 2$ nm, the barrier of the CB ($Si_{0.7}Ge_{0.3}$) is with a small width and the electrons are easy to tunneling. Thus, the ground state (E_0) is with the smallest energy compared with Figs. 2b and c. As the space between each nanowire increases, the electrons are more localized in the finite energy well, which leads to an increased ground state energy. For example, in Γ-valley, the ground state energies are 0.565, 0.569, and 0.570 eV with s =2, 15, and 50 nm, respectively. Similarly, the band structure of light holes is clustered and the ground state energy of light holes in the VB rise as the barrier becomes widers. As the space decreases, as shown in

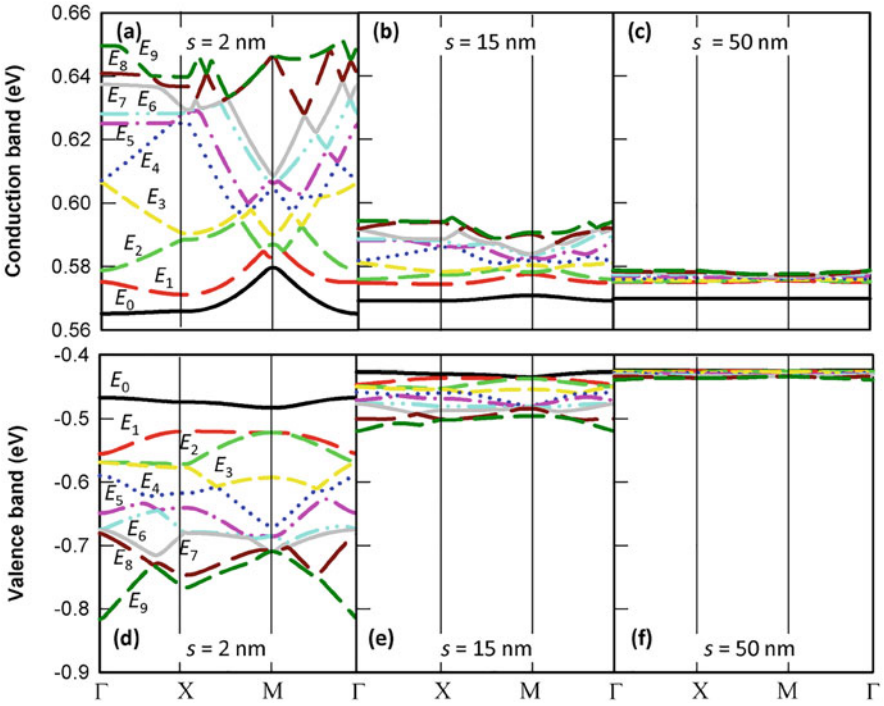


Fig. 2 (**a**)–(**c**) The first ten energies of electrons in the CB. (**d**)–(**f**) The first ten energies of light holes in the VB. E_i represents the ith energy state. For electrons, E_0 is the ground state and E_1 is the first excited state

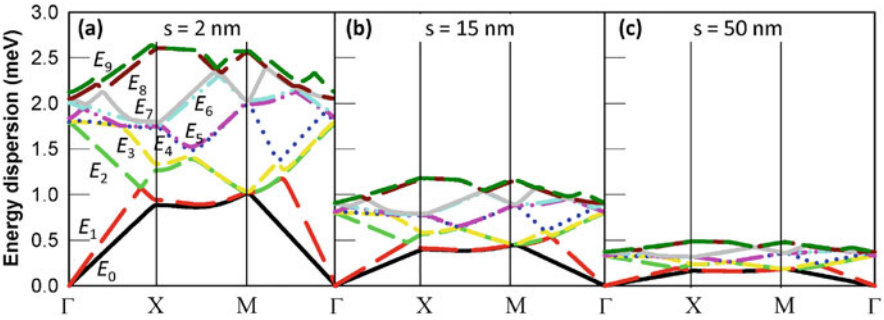


Fig. 3 (**a**)–(**c**) The energy dispersion of phonons in the SiNWs-$Si_{0.7}Ge_{0.3}$ composite with s = 2, 15, and 50 nm, respectively

Fig. 3, the phonons meet more interfaces and the scattering rate is huge. In room temperature, the low energy phonon plays an important role to carry heat. Thus, the thermal conductivity can be expected to be decreased as s decreases.

Notably, the number of eigenvalues will influence the accuracy of the results. Thus, to find an optimal sampling number with a minimal time cost, Fig. 4a shows

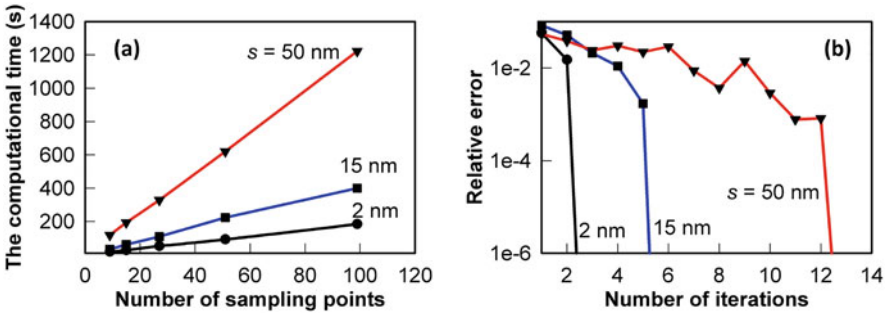


Fig. 4 (**a**) The computational time versus the number of sampling points and (**b**) the relative error versus the numbers of iterations with respect to different s, where the results are solved from the Schrödinger equation. The lines are the results with s =2, 15, and 50 nm, respectively. The tested PC is with an Intel® Core™ i7–7500 CPU and the RAM is 16 GB

the computational time when solving (2) versus the sampling numbers with respect to different s. For each s, we solve the first ten eigenvalues for electrons in the CB. There are 894, 1126, and 1344 elements in our simulation with s=2, 15, and 50 nm, respectively; the sizes of corresponding matrices are 1856, 2329, and 2765, respectively. The computational time will increase as the sampling points increase linearly. Figure 4b shows the relative error between iterations of the implicitly restarted Arnoldi method with respect to different s. The stopping criterion is the relative error $<10^{-6}$. Notably, the accuracy of the computed eigenenergy is almost the same when the matrix size increases from several thousands to ten thousands; however, the time cost increases significantly. Not shown here, we have the similar numerical tests when solving (3), where both the computational time and convergence behavior are faster than that of (2).

By considering the doping effect, TE properties calculated via the Landauer approach will vary as the Fermi level. The calculated lattice thermal conductivity is about 2.2 W/mK in Algorithm 2, which is close to the experimentally measured data of 3.5 W/mK [6] when the density of SiNWs is 1.6×10^{11} cm^{-3} ($r = 5$ nm and $s = 15$ nm). For the SiNWs-$Si_{0.7}Ge_{0.3}$ composite with p-type doping of 1.16×10^{15} cm^{-3}, ZT calculated from (1) is about 1.5×10^{-4} at room temperature.

5 Conclusions

In this paper, we have applied the numerical method to calculate the electronic and phononic band structures of the silicon nanowires embedded in $Si_{0.7}Ge_{0.3}$ by solving the Schrödinger equation and the elastodynamic wave equation. The Landauer approach is used for the calcultion of thermoelectric properties, respectively. The simulated thermal conductivity is close to the measurement.

Acknowledgments This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-009-008, Grant MOST 108-3017-F-009-001, Grant MOST 109-2221-E-009-033, Grant MOST-109-2634-F-009-030, and Grant MOST 110-2221-E-A49-139, and in part by the "Center for mmWave Smart Radar Systems and Technologies" under the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

References

1. Y.-M. Lin, X. Sun, M.S. Dresselhaus, Theoretical investigation of thermoelectric transport properties of cylindrical Bi nanowires. Phys. Rev. B **62**, 4610 (2000)
2. A.I. Hochbaum, R. Chen, R.D. Delgado, W. Liang, E.C. Garnett, M. Najarian, A. Majumdar, P. Yang, Enhanced thermoelectric performance of rough silicon nanowires. Nature **451**, 163–167 (2008)
3. P. Martin, Z. Aksamija, E. Pop, U. Ravaioli, Impact of phonon-surface roughness scattering on thermal conductivity of thin Si nanowires. Phys. Rev. Lett. **102**, 125503 (2009)
4. C. Bera, N. Mingo, S. Volz, Marked effects of alloying on the thermal conductivity of nanoporous materials. PRL **104**, 115502 (2010)
5. M.S. Dresselhaus, G. Chen, M.Y. Tang, R. Yang, H. Lee, D. Wang, Z. Ren, J.-P. Fleurial, P. Gogna, New directions for low-dimensional thermoelectric materials. Adv. Mater. **19**, 1043–1053 (2007)
6. A. Kikuchi, A. Yao, I. Mori, T. Ono, S. Samukawa, Composite films of highly ordered Si nanowires embedded in $SiGe_{0.3}$ for thermoelectric applications. J. Appl. Phys. **122**, 165302 (2017)
7. B.I. Schneider, Parallel solver for the time-dependent linear and nonlinear Schrödinger equation. Phys. Rev. E **73**, 036708 (2006)
8. P. Kratzer, J. Neugebauer, The basics of electronic structure theory for periodic systems. Front. Chem. **7**, 1–18 (2019)
9. M.-Y. Lee, Y. Li, S. Samukawa, Miniband calculation of 3-D nanostructure array for solar cell applications. IEEE Trans. Electron Dev. **62**, 3709–3714 (2015)
10. F. Schaffler, Silion-germanium, in *Properties of Advanced Semiconductor Materials: GaN, AlN, InN, BN, SiC, SiGe*, ed. by M.E. Levinshtein, S.L. Rumyantsev, M.S. Shur (Wiley, New York, 2001), pp. 149–188
11. R. Anufriev, M. Nomura, Thermal conductance boost in phononic crystal nanostructures. Phys. Rev. B, **91**, 245417 (2015)
12. G. Mascali, V. Romano, A hierarchy of macroscopic models for phonon transport in graphene. Physica A **548**, 124489 (2020)
13. W.-W. Zhang, H. Yu, S.-Y. Lei, Q.-A. Huang, Modelling of the elastic properties of crystalline silicon using lattice dynamics. J. Phys. D **44**, 335401 (2011)
14. Y. Shiraki, N. Usami, *Silicon–Germanium (SiGe) Nanostructures: Production, Properties and Applications in Electronics* (Woodhead Publishing, Cambridge, 2011)
15. J. Maassen, M. Lundstrom, The Landauer approach to electron and phonon transport. ECS Trans. **69**, 23–36 (2015)
16. C. Jeong, R. Kim, M. Luisier, S. Datta, M. Lundstrom, On Landauer versus Boltzmann and full Band versus effective mass evaluation of thermoelectric transport coefficients. J. Appl. Phys. **107**, 023707 (2010)
17. L. Musl, E. Flage-Larsen, Thermoelectric transport calculations using the Landauer approach, ballistic quantum transport simulations, and the Buttiker approximation. Comput. Mater. Sci. **132**, 146–157 (2017)

18. A.I. Hochbaum, R. Chen, R.D. Delgado, W. Liang, E.C. Garnett, M. Najarian, A. Majumdar, P. Yang, Enhanced thermoelectric performance of rough silicon nanowires. Nature **451**, 163–167 (2008)
19. R.B. Lehoucq, D.C. Sorensen, C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods* (Society for Industrial and Applied Mathematics, Philadelphia, 1998)
20. J. Sun, A. Zhou, *Finite Element Methods for Eigenvalue Problems* (CRC Press, New York, 2017)
21. M.S. Kushwaha, P. Halevi, L. Dobrzynski, B. Djafari-Rouhani, Acoustic band structure of periodic elastic composites. Phys. Rev. Lett. **71**, 2022 (1993)

A Novel Surface Mesh Simplification Method for Flux-Dependent Topography Simulations of Semiconductor Fabrication Processes

Christoph Lenz, Alexander Scharinger, Paul Manstetten, Andreas Hössinger, and Josef Weinbub

Abstract In etching and deposition simulations of a semiconductor fabrication process the calculation of the surface rates of particles is an essential but also the computationally most demanding step. A promising approach is to preprocess the simulation domain by simplifying the surface. We thus propose a new surface mesh simplification method that takes advantage of geometric domain-specific surface properties that are prevalent in topography simulations. We compare our method to a suitable reference algorithm and show that our method maintains higher geometric accuracy and accordingly maintains the original geometry in great detail. Furthermore, the evaluation of the simplified meshes show an enhanced performance of the particle surface rate calculation.

1 Introduction

Process technology computer-aided design (TCAD) tools are used to simulate fabrication processes of semiconductor devices. One important branch of process TCAD is the evolution of the topography during etching and deposition processes. In each time step the three essential computational tasks are: (a) the calculation of the particle flux on the surface, which is used to (b) calculate the surface velocity according to a surface model and (c) the calculation of the new position of the surface using the surface velocities [1]. In Process TCAD the surface can be represented implicitly using the level-set method where the domain is discretized on

C. Lenz (✉) · A. Scharinger · P. Manstetten · J. Weinbub
Christian Doppler Laboratory for High Performance TCAD, Institute for Microelectronics, TU Wien, Wien, Austria
e-mail: lenz@iue.tuwien.ac.at; scharinger@iue.tuwien.ac.at; manstetten@iue.tuwien.ac.at; weinbub@iue.tuwien.ac.at

A. Hössinger
Silvaco Europe Ltd., St. Ives, UK
e-mail: andreas.hoessinger@silvaco.com

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_8

a regular grid. This approach is attractive due to the robust handling of topographical changes in a level set framework [2]. The particle flux on the semiconductor surface denotes the number of particles interacting on the surface. One possible numerical method for calculating the surface flux is Monte Carlo ray tracing [3]. At practically relevant surface resolutions the flux calculation dominates the overall execution time of an etching or deposition simulation [1]. It is thus useful to investigate approaches that speed up the flux calculation. One promising approach is to use temporary explicit surface meshes as there exists a large body of knowledge about ray tracing on explicit surfaces. The *marching cubes algorithm* [4] is commonly used to extract an explicit surface from the level set. However, the resulting surface meshes typically contain very narrow and long triangles (needles) or small triangles in flat regions that contain no geometric variation. Therefore eliminating those surface elements reduces the total surface element count which speeds up the ray tracing tasks, further underlining the attractiveness of an explicit surface mesh approach.

There exist several algorithms that reduce the resolution of surface meshes with respect to a given metric; several metrics have been proposed in literature [5–8]. However, some of these algorithms try to simplify the geometry homogeneously [5, 6] or use computationally expensive metrics [7, 8]. The latter is particularly relevant when considering the entire etching or deposition workflow where the mesh simplification has to be conducted at every single time step. Mesh simplification, or more general domain simplification, is a commonly used approach in process TCAD simulations [9, 10]. In particular, in [11] the authors evaluate the flux on a mesh by sampling only a sparse set of surface elements to accelerate the simulation.

In this paper we introduce a flexible and computationally lightweight simplification method based on the local surface curvature. We evaluate the impact of our mesh simplification method on typical process TCAD topography simulations by using the high performance ray tracing library Embree [12] by conducting a ray tracing performance analysis. Specifically we compare the flux calculation time for surfaces obtained with the presented method, with the flux calculation time obtained for surfaces generated by the reference Lindstrom-Turk algorithm [5], by comparing the execution time of the simplification process and the performance of the flux calculation using Monte Carlo ray tracing.

2 Surface Mesh Simplification

The simplification method presented in this work is based on the Lindstrom-Turk algorithm [5]. This algorithm uses an *Edge Collapse* procedure to simplify the surface mesh. It offers a relatively low computational complexity and takes the quality of triangles into account: The latter is particularly important for process TCAD simulations, as the mesh quality directly influences subsequent procedures.

Our method uses the mean curvature of each vertex to partition the mesh into regions. This allows us to adjust the amount of simplification according to the local

geometric properties in each region. This simplification method has been designed to simplify regions of the mesh offering negligible geometric variation (e.g. flat areas) to a higher degree, thus allowing to maintain a higher resolution in regions of the mesh with high geometric variation. Furthermore, our method is not limited to the Lindstrom-Turk simplification algorithm, hence other simplification algorithms [6] can be used in combination with our method.

2.1 Feature Detection

The first step in our simplification method is the detection of geometric features in the mesh: We use the absolute mean curvature of each vertex and calculate it via a discrete approximation of the *Laplace-Beltrami operator* [13] in the vertex $\mathbf{x}_i$

$$|H(\mathbf{x}_i)| = \frac{\| \sum_{j \in N_1(i)} (\cot \alpha_{ij} - \cot \beta_{ij})(\mathbf{x}_i - \mathbf{x}_j) \|}{4A_{\text{avg}}}, \tag{1}$$

where $H(\mathbf{x}_i)$ denotes the mean curvature in the vertex $\mathbf{x}_i$ and $N_1(i)$ is the set of all vertices adjacent to $\mathbf{x}_i$. The angles α_{ij}, β_{ij} are the angles of the triangles that share the edge between $\mathbf{x}_i$ and $\mathbf{x}_j$, which are opposite to this edge and A_{avg} is the average area of the triangles surrounding the vertex $\mathbf{x}_i$. The mean curvature is used to categorize each vertex to be either a *flat* or a *feature* vertex. In particular, an empirical threshold is used to identify vertices with small curvature (numerical artifacts), which are considered to be flat.

2.2 Mesh Partitioning and Movement of Regions

The Mesh is partitioned into the *feature regions* and the *transition regions* according to the metrics above. The feature region encompasses the triangles of the mesh with significant geometric variation. The transition region contains the triangles that do not hold information about the geometric variation. This partition of the mesh allows to simplify the transition region to a greater extent, which reduces the overall number of mesh elements without loosing information about the geometric variation. Furthermore, this approach allows to keep a high resolution in regions of the mesh with high geometric variation by simultaneously limiting the overall mesh size in terms of number of triangles. However, simplifying the flat region to a higher degree than the feature region leads to low quality triangles (e.g. needles).

To prevent the formation of low quality elements the transition region is simplified with linearly increasing parameters, thus creating a reasonable mesh grading. Figure 1 schematically depicts two steps of the discussed process. At first the whole mesh, including the feature region, is simplified until the smallest edge has an edge length of l_0. If the feature region should not be simplified l_0 is set

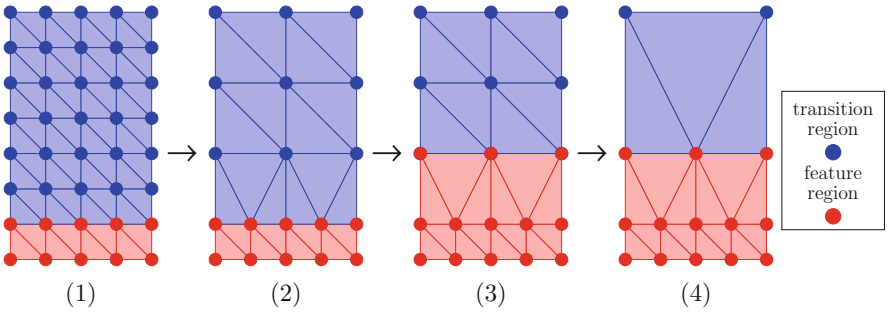


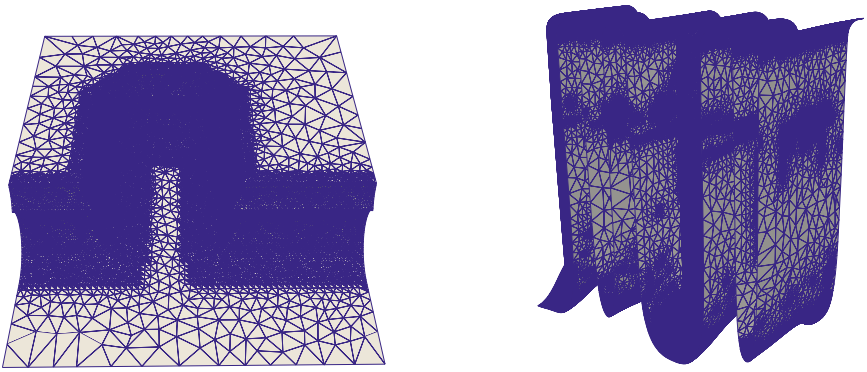
Fig. 1 Example of the simplification process: (1) shows the mesh after it has been divided into regions. (2) shows the simplification of the feature region. (3) shows the extension of the feature region. (4) shows again the simplification of the transition region with an increased edge length

to 0. After this initial simplification step the transition region is simplified until the smallest edge has an edge length of $l_1 = l_0 + sl$, where sl denotes the step length. Next, the feature region is expanded into the transition region. Afterwards the now smaller transition region is simplified until the smallest edge has an edge length of $l_{i+1} = l_i + sl$ with $i \in \{0, 1, \ldots, n \in \mathbb{N}\}$. These last two steps continue until the feature region cannot move any further into the transition region, and thus terminates the simplification process. To avoid unwanted side effects of the potentially large edge lengths produced by our iterative scheme, another parameter $l_{\max}$ is used to terminate the refinement once the edge length l_i in the transition region has reached $l_{\max}$.

The parameter for the simplification of the feature region l_0, when using the level set method, can be connected to the level-set and is chosen in concordance with the minimal grid size Δ_t. When using meshes not originating from a level-set, this parameter can be chosen by averaging the edge length of all feature vertices. We have empirically determined that the step length sl should be approximately the edge length of the feature region after the simplification with the parameter l_0 stops. A bigger step size increases the amount of edges that are removed. However, the bigger the difference between the edge length of the feature region and the step length, the worse the triangle quality of the mesh.

3 Results

The simplification method has been evaluated in the context of process TCAD in three ways: geometric distance to the original geometry, execution time of the simplification method, and the execution time of a subsequent surface flux calculation by ray tracing. In this study two example geometries have been analyzed and each example geometry has been simplified applying eight different degrees of simplification, resulting in a reduction of vertices from 20–90%. Figure 2 shows the



(a) (b)

Fig. 2 Process TCAD surface meshes simplified with our method. (**a**) Surface 1 with 78% of the vertices of the original mesh removed by our simplification method. (**b**) Surface 2 with 52% of the vertices of the original mesh removed by our simplification method

two surface meshes after they have been simplified with our method. The original surface meshes of Surface 1 and Surface 2 have 70,831 and 175,550 vertices, respectively. The performance benchmarks presented in the following are based on a serial C++ implementation of our method executed on a 64bit GNU/Linux platform equipped with an Intel Devil's Canyon CPU.

3.1 Distance to Original Geometry

Surface mesh simplification introduces geometric distortions into the simplified mesh. To measure the error introduced by the simplification process we use the *Hausdorff distance* [14] between the original and the simplified mesh. The Hausdorff distance is measured from each vertex of the original mesh to the simplified mesh. Figure 3 shows the results for one test case of our analysis. The distance to the original mesh is smaller when using our simplification method. On average our simplification method has 20–40% lower Hausdorff distance to the original geometry than using the Lindstrom-Turk algorithm. The reason for the significantly improved Hausdorff distance is our method which allows to use more vertices in areas of high geometric variation, allowing to represent the overall geometry better.

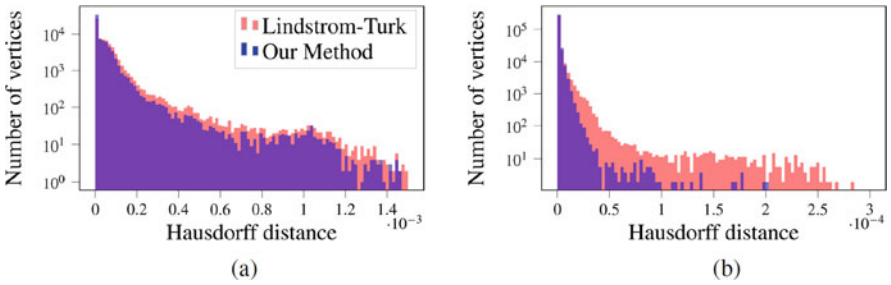


Fig. 3 Hausdorff distance from each vertex of the original mesh to a mesh simplified with our method and a mesh simplified using the Lindstrom-Turk algorithm. (**a**) Surface 1 with 78% of the vertices of the original mesh removed by our simplification method. (**b**) Surface 2 with 52% of the vertices of the original mesh removed by our simplification method

3.2 Time Spent on Simplification

The simplification method presented in this work introduces an overhead to the simplification process. This overhead consists primarily of the feature detection, at the start of the simplification process, and the movement of the feature regions. As can be seen in Fig. 4 our simplification method takes on average 17% longer than the Lindstrom-Turk algorithm.

3.3 Flux Calculation and Monte Carlo Ray Tracing

A common approach to compute the surface flux in a Process TCAD application is to use a Monte Carlo simulation [15]. This is a randomized procedure and the results of the Monte Carlo method are of stochastic nature. To compute the trajectories

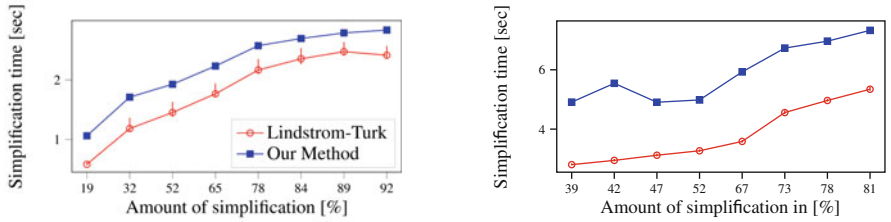


Fig. 4 Average simplification time of our method and the Lindstrom-Turk algorithm. The amount of simplification denotes the number of vertices which have been removed from the original mesh. (**a**) Surface 1. (**b**) Surface 2

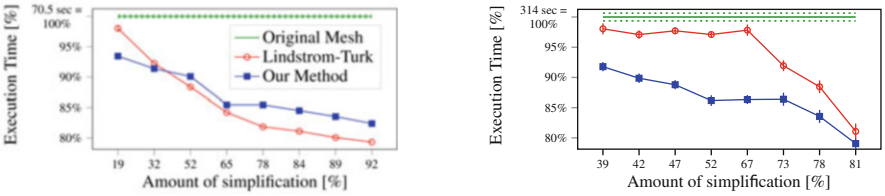


Fig. 5 Execution time of Monte Carlo ray tracing using 10^8 rays. The amount of simplification denotes the number of vertices which have been removed from the original mesh. (**a**) Surface 1. (**b**) Surface 2

on which particles move through the simulation domain (modeling the surface flux) we use the Embree ray tracing library [12]. In Embree a bounding volume hierarchy data structure [3, 12] is used to efficiently compute the paths on which the particles move through space. The internal structure of the bounding volume hierarchy depends eminently on the structure and the coarseness of the surface mesh.

Figure 5 shows the execution times measured to perform the Monte Carlo ray tracing on the meshes with different degrees of simplification. As the simplified meshes contain less triangles the bounding volume hierarchy data structure used for ray tracing will have less elements than the data structure for the original mesh. As the size of the data structure is decreased the memory footprint is reduced and this leads to faster flux calculations because less data has to be processed and the caches of the processor are used more effectively. Figure 5a and b show that the empirical speedup in flux calculation depends on the shape of the surface mesh. When tracing Surface 1, the meshes of both simplification methods perform approximately the same and are faster than the original mesh. When tracing Surface 2, the meshes generated by our simplification method clearly outperform the meshes simplified with the Lindstrom-Turk algorithm and the original geometry. Surface 2 contains deep trenches and the rays of the tracing algorithm need to travel towards the bottom of these trenches. As the walls of the trenches do not have high curvature the bounding volume hierarchy data structure created from the mesh simplified with our method will be less complex within the deep trenches and hence, the traces of the rays down the trench can be computed by performing less operations. Also, the rays which travel towards the bottom of the trench usually reflect off the surface many times, which makes the difference in computational effort for using a bounding volume hierarchy from a mesh simplified with our method even more evident. Figure 5b for Surface 2 shows a speedup of about 12% compared to the Lindstrom-Turk algorithm for simplification levels of 52 and 67%.

4 Summary

We introduce a new surface mesh simplification method that uses the curvature of the surface mesh to identify regions which can be simplified with different sets of parameters depending on the local surface properties. Our approach is well suited for meshes that are common in flux-dependent process TCAD simulations since such meshes often contain large flat regions with high resolutions from the originating regular grid. We have evaluated our method with respect to geometric distances and execution times for simplification and subsequent computations of flux estimates. The geometric distances in the experiments have improved in comparison to the reference algorithm. In particular, the average Hausdorff distance of the investigated geometries has improved by 20–40%. The ray tracing time in all our experiments has been improved on average by 15%, furthermore, demanding real world geometries from process TCAD have shown a compelling improvement of 12% of time spent on ray tracing. The execution time of our simplification method is on average 17% slower than the reference algorithm. When considering entire topography simulations, the accelerated ray tracing significantly exceeds the additional time spent on our simplification method.

Acknowledgments The financial support by the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development is gratefully acknowledged.

References

1. P. Manstetten, Efficient Flux Calculations for Topography Simulation. Doctoral Dissertation, TU Wien (2018). http://www.iue.tuwien.ac.at/phd/manstetten/
2. J.A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science* (Cambridge University Press, 1999)
3. A.S. Glassner, *An Introduction to Ray Tracing* (Elsevier, 1989)
4. W.E. Lorensen, H.E. Cline Marching cubes: A high resolution 3D surface construction algorithm. ACM SIGGRAPH Comput. Graph. **21**, 163–169 (1987)
5. P. Lindstrom, G. Turk, Fast and memory efficient polygonal simplification, in *Proceedings of the IEEE Visualization Conference*, pp. 279–286 (1998)
6. M. Garland, P.S. Heckbert, Surface simplification using quadric error metrics, in *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, pp. 209–216 (1997)
7. H. Borouchaki, P.J. Frey, Simplification of surface mesh using Hausdorff envelope. Comput. Methods Appl. Mech. Eng. **194**, 4864–4884 (2005)
8. S.J. Kim, C.H. Kim, D. Levin, Surface simplification using a discrete curvature norm. Comput. Graph. **26**, 657–663 (2002)
9. F. Rudolf, Symmetry- and Similarity-Aware Volumetric Meshing. Doctoral Dissertation, TU Wien (2016). https://www.iue.tuwien.ac.at/phd/rudolf/
10. L. Gnam, High Performance Mesh Adaptation for Technology Computer-Aided Design. Doctoral Dissertation, TU Wien (2019). https://www.iue.tuwien.ac.at/phd/gnam/

11. L. Gnam, P. Manstetten, A. Hössinger, S. Selberherr, J. Weinbub, Accelerating flux calculations using sparse sampling. Micromachines **9**, 1–17 (2018)
12. I. Wald, S. Woop, C. Benthin, G.S. Johnson, M. Ernst, Embree: A kernel framework for efficient CPU ray tracing. ACM Trans. Graph. **33**, 143:1–143:8 (2014)
13. M. Meyer, M. Desbrun, P. Schröder, A.H. Barr, Discrete differential-geometry operators for triangulated 2-manifolds, in *Visualization and Mathematics III*, pp. 35–57 (2003)
14. R. Straub, Exact computation of the Hausdorff distance between triangular meshes, in *Proceedings of the Conference of the European Association for Computer Graphics*, pp. 17–20 (2007)
15. R.Y. Rubinstein, D.P. Kroese, *Simulation and the Monte Carlo Method* (Wiley, 2016)

Simulations of a Novel DG-GFET

Giovanni Nastasi and Vittorio Romano

Abstract A peculiar geometry for a graphene double gate field effect transistor is proposed. It allows us to overcome the problems encountered for a standard MOSFET geometry due to the zero gap in monolayer graphene. It is found that for a wide range of the gate voltage the current is in an off state with a ratio current-on over current-off of about 10^4.

1 Introduction

As quoted in [1] "Graphene has changed from being the exclusive domain of condensed-matter physicists to being explored by those in the electron-device community. In particular, graphene-based transistors have developed rapidly and are now considered an option for post-silicon electronics. However, many details about the potential performance of graphene transistors in real applications remain unclear."

Device engineers devote considerable effort for developing transistor designs in which short-channel effects are suppressed and series resistances are minimized. Scaling theory predicts that a FET with a thin barrier and a thin gate-controlled region will be robust against short-channel effects down to very short gate lengths. The possibility of having channels that are just one atomic layer thick is perhaps the most attractive feature of graphene for its use in transistors. Main drawback of a large-area monolayer graphene is the zero gap. This has the consequence that the current versus the gate voltage is no longer a monotone function and the off region is very narrow (see [2]), making graphene not usable in a straightforward way for transistors. Moreover, graphene on substrate suffers also from the degradation of the mobility because of the additional interaction with the phonons of the oxide.

G. Nastasi (✉) · V. Romano
Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: g.nastasi@unict.it; romano@dmi.unict.it

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_9

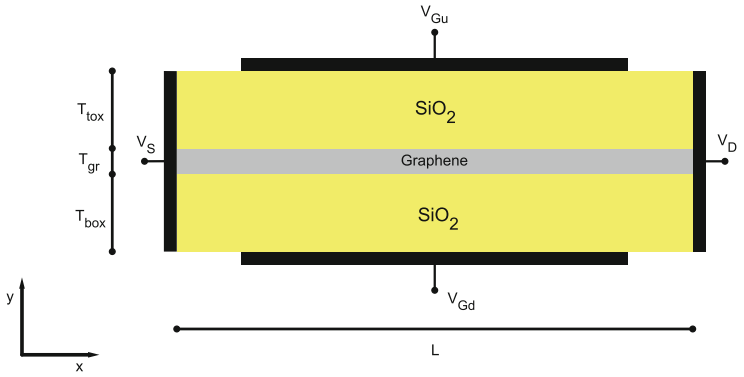


Fig. 1 Schematic representation of the investigated DG-GFET

Here we propose a special geometry for a double gate graphene FET (DG-GFET) which overcomes the problem related to the zero gap as will be shown by the numerical simulations. The devices is depicted in Fig. 1. The active area is made of just one graphene layer.

Usually the GFETs are investigated by adopting reduced one dimensional models of the Poisson equation with some averaging procedure [3, 4]. Here a full two-dimensional simulation is presented based on a drift-diffusion-Poisson system with the mobilities proposed in [5].

Other approaches are based on hydrodynamical models, e.g. those deduced with the maximum entropy principle [6–9], or the direct solution of the Boltzmann equation [10–14] or Monte Carlo methods [15]. Thermal effects can also be included [16–21]. Here the crystal lattice will be kept at a constant temperature and considered as a thermal bath. For the inclusion of quantum effects the interest reader is referred to [22, 23].

2 Mathematical Model

The mathematical model we adopt to simulate the charge transport in the graphene layer of the DG-GFET 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 [5] (for other models the interested reader is referred to [2, 24, 25]) given by

$$\mu_s(x) = \frac{v_s}{[1+(v_s E/v_{sat})^\gamma]^{1/\gamma}},$$

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

$$v_s(x) = \frac{\mu_0}{(1+s/n_{ref})^\alpha},$$

where $\mu_0 = 0.4650\ \mu\text{m}^2/\text{V ps}$ is the low field mobility, $n_{ref} = 1.1\times 10^5\ \mu\text{m}^{-2}$ and $\alpha = 2.2$. The symbol s indicates the carrier density: $s = n$ for electrons and $s = p$ for holes.

In order to determine the electric potential a 2D Poisson equation is coupled to the drift-diffusion system

$$\nabla\cdot(\epsilon\nabla\phi) = h(x,y),$$

where

$$h(x,y) = \begin{cases} e(n(x)-p(x)-N_{imp})/t_{gr} & \text{if} \quad y = y_{gr} \\ 0 & \text{if} \quad y \neq y_{gr} \end{cases}$$

being y_{gr} the y-coordinate (see Fig. 1), $N_{imp} = 3.5\times 10^3 \mu\text{m}^{-2}$ the impurity density due to the SiO_2, t_{gr} the distance between the two layers of oxide which is assumed to be equal to 1 nm. We remark that 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. Moreover 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. The source and drain contacts are assumed to be thermal bath charge reservoirs.

3 Numerical Results

Here some numerical results are presented in order to show that the proposed DG-GFET is able to perform as a transistor. The length is 100 nm. The width of both the oxide layers (SiO_2) is 10 nm. The source and drain contacts are positioned in the direction transversal with respect the graphene sheet and they occupy all the device height (21 nm). The two gate potentials are set as equal. At the metallic contacts the total voltage including the work function is considered equal to 0.25 V plus the bias voltage, which is zero at source. Indeed the work function depends on the specific material the contacts are made of.

A full 2D discretization of the Poisson equation is adopted in the whole device by standard central differencing enforced with a Gummel iteration, while the drift-diffusion equation is solved only in the graphene sheet as a 1D problem with a Scharfetter-Gummel method (indeed only one row of grid points is used by considering a kind of average in the y direction). The interested reader is referred to [2] for the details. By numerical experiments a good resolution is already obtained with 41×23 grid points.

In Figs. 2, 3 the shape of the electrical potential is plotted when the source-drain-potential is 0.3 V and the gate-source potential is -1 V and 1 V respectively. In the first case the device is off while in the second case is on. The Fig. 4 shows the characteristic curve current versus gate voltage with source drain voltage equal to 0.2 V while Fig. 5 shows the same but in a logarithmic scale.

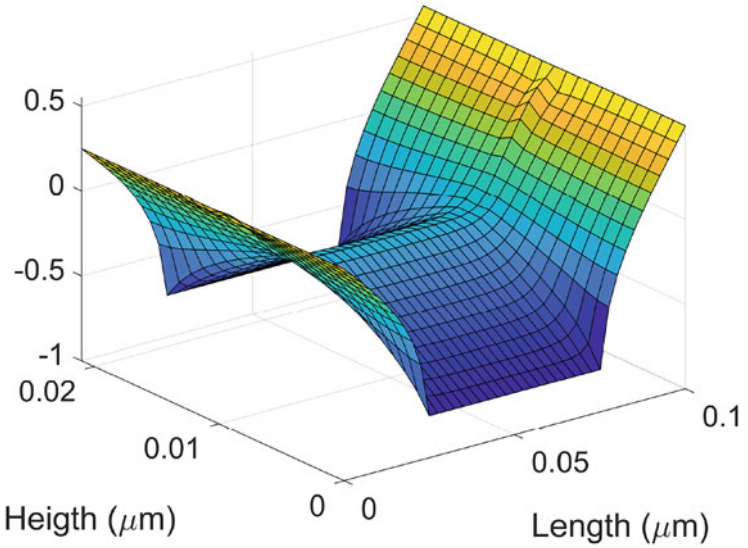


Fig. 2 Electrostatic potential when the gate-source potential is -1 V and the source-drain-potential is 0.3 V

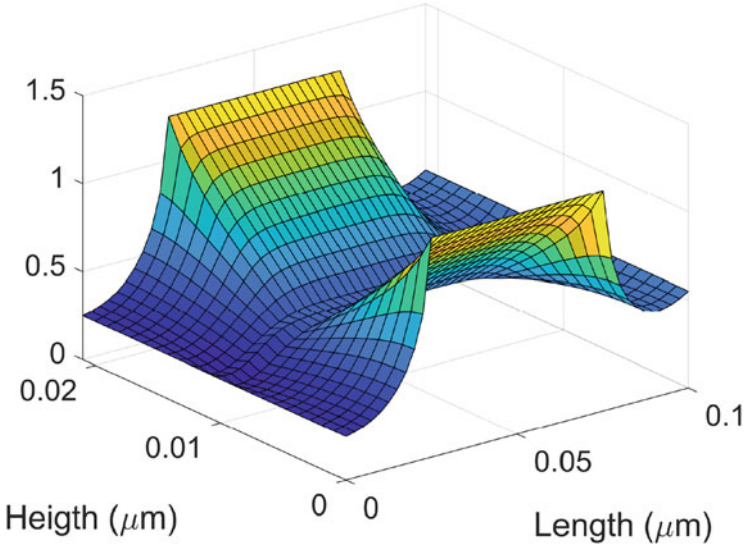


Fig. 3 Electrostatic potential when the gate-source potential is 1 V and the source-drain-potential is 0.3 V

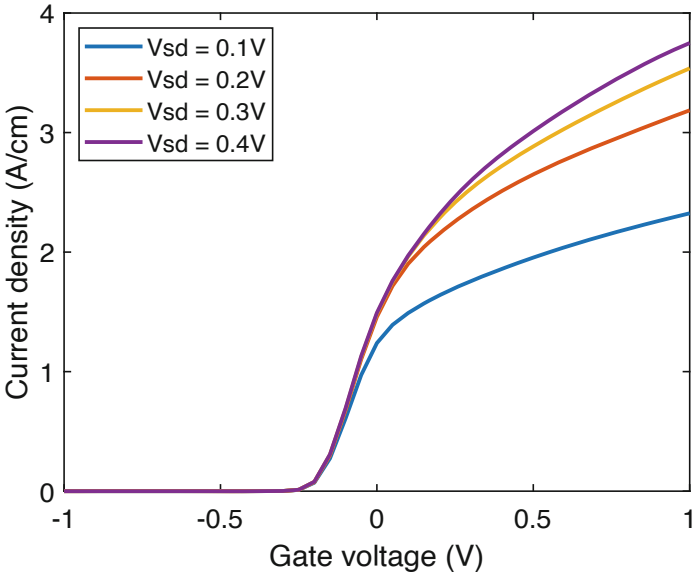


Fig. 4 Current versus gate voltage for several values of the bias voltage

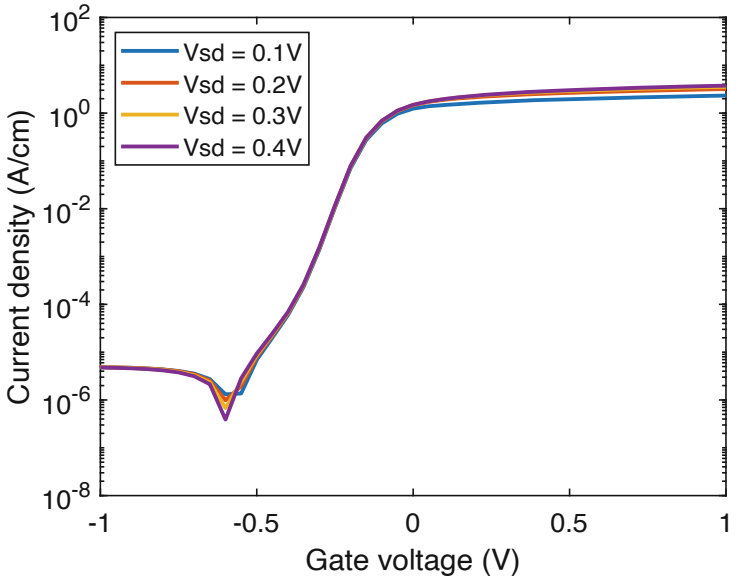


Fig. 5 Current versus gate voltage for several values of the bias voltage in logarithmic scale

It is evident that the peculiar geometry of the DG-FET we have investigated leads to characteristic curves that are appropriate for field effect transistors mainly because there exists a clear and wide off region similar to the case of tradition semiconductors like Si or GaAs. With the particular set up of gate-source and source-drain voltage the current of the minor charges is not triggered and transport remain mainly unipolar (Figs. 6 and 7). This at variance with the case of standard MOSFET configurations where minority charges are triggered for sufficiently gate-drain voltage limiting the current-off zone to a short range [2, 26].

The current-on over current-off ratio is about of four orders of magnitude which is acceptable for electrical engineering purposes. This is clearly shown by the figure in logarithmic scale.

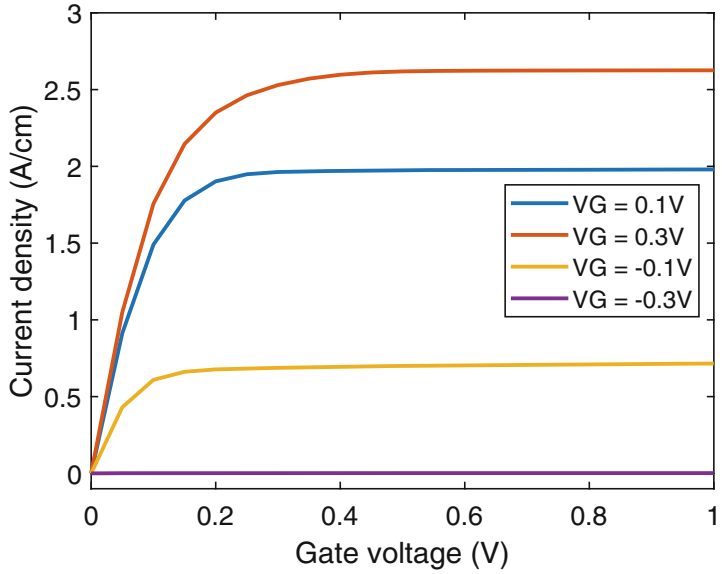


Fig. 6 Current versus bias voltage for several values of the gate voltage

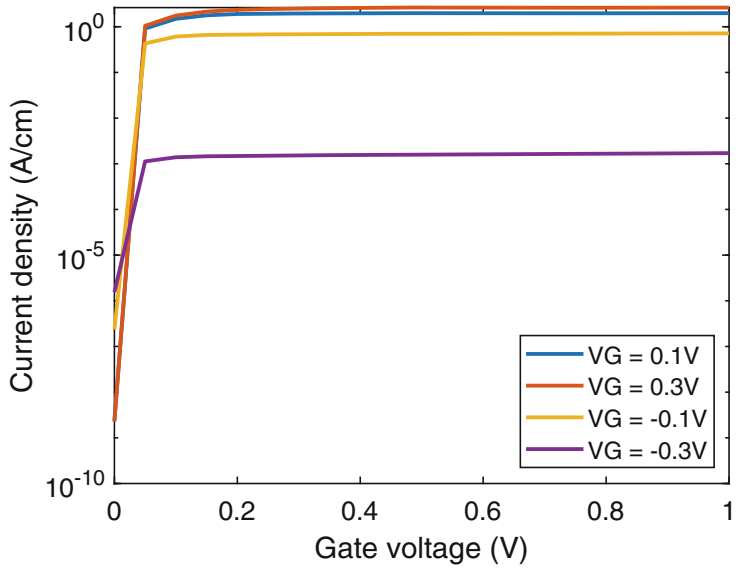


Fig. 7 Current versus bias voltage for several values of the gate voltage in logarithmic scale

4 Conclusions

A novel geometry for a double gate FET with active area made of a single layer of graphene has been proposed and simulated with a drift-diffusion model by solving a full 2D Poisson equation for the electrostatic potential. The results are rather encouraging because a good transistor effect is obtained at variance with of other GFETs proposed in the literature. The simulation based on more sophisticated models is currently under investigation by the authors in order to get a further validation of the devised device.

Acknowledgments The authors acknowledge the support from INdAM (GNFM). The author G.N. acknowledges the support from Progetto Giovani GNFM 2019 *Modelli matematici, numerici e simulazione del trasporto di cariche e fononi nel grafene*. This work has been also supported by the Università degli Studi di Catania, *Piano della Ricerca 2018/2020 Linea di intervento 2*.

References

1. F. Schwierz, Graphene transistors. Nat. Nanotechnol. **5**, 487–496 (2010)
2. G. Nastasi, V. Romano, A full coupled drift-diffusion-Poisson simulation of a GFET. Commun. Nonlinear Sci. Numer. Simul. **87**, 105300 (2020)
3. D. Jiménez, O. Moldovan, Explicit drain-current model of graphene field effect transistors targeting analog and radio-frequency applications. IEEE Trans. Electron. Dev. **58**(11), 4049–4052 (2011)
4. A.K. Upadhyay, A.K. Kushwaha, S.K. Vishvakarma, A unified scalable quasi-ballistic transport model of GFET for circuit simulations. IEEE Trans. Electron. Dev. **65**(2), 739–746 (2018)
5. V.E. Dorgan, M.-H. Bae, E. Pop, Mobility and saturation velocity in graphene on SiO_2. Appl. Phys. Lett. **97**(8), 082112 (2010)
6. V.D. Camiola, V. Romano, Hydrodynamical model for charge transport in graphene. J. Stat. Phys. **157**, 1114–1137 (2014)
7. L. Luca, V. Romano, Hydrodynamical models for charge transport in graphene based on the maximum entropy principle: the case of moments based on energy powers. Atti Accad. Pelorit. Pericol. Cl. Sci. Fis. Mat. Nat. **96**(S1), A5 (2018)
8. V.D. Camiola, G. Mascali, V. Romano, *Charge Transport in Low Dimensional Semiconductor Structures - The Maximum Entropy Approach*. Springer International Publishing (2020)
9. L. Luca, V. Romano, Comparing linear and nonlinear hydrodynamical models for charge transport in graphene based on the Maximum Entropy Principle. Int. J. Nonlin. Mech. **104**, 39–58 (2018)
10. A. Majorana, G. Nastasi, V. Romano, Simulation of bipolar charge transport in graphene by using a discontinuous Galerkin method. Comm. Comp. Phys., **26**(1), 114–134 (2019)
11. P. Lichtenberger, O. Morandi, F. Schürrer, High-field transport and optical phonon scattering in graphene. Phys. Rev. B **84**(4), 045406 (2011)
12. V. Romano, A. Majorana, M. Coco, 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)
13. M. Coco, A. Majorana, G. Nastasi, V. Romano, High-field mobility in graphene on substrate with a proper inclusion of the Pauli exclusion principle. Atti Accad. Pelorit. Pericol. Cl. Sci. Fis. Mat. Nat., **97**(S1), A6 (2019)

14. M. Coco, G. Nastasi, Simulation of bipolar charge transport in graphene on h-BN. COMPEL **39**(2), 449–465 (2020)
15. M. Coco, A. Majorana, V. Romano, Cross validation of discontinuous Galerkin method and Monte Carlo simulations of charge transport in graphene on substrate. Ricerche mat. **66**, 201–220 (2017)
16. M. Coco, G. Mascali, V. Romano, Monte Carlo analysis of thermal effects in monolayer graphene. J. Comput. Theor. Trans. **45**(7), 540–553 (2016)
17. M. Coco, V. Romano, Simulation of electron–phonon coupling and heating dynamics in suspended monolayer graphene including all the phonon branches. J. Heat Transfer. **140**(9), 092404 (2018)
18. M. Coco, V. Romano, Assessment of the constant phonon relaxation time approximation in electron-phonon coupling in graphene. J. Comput. Theor. Trans. **7**(1–3), 246–266 (2018)
19. G. Mascali, V. Romano, Charge transport in graphene including thermal effects. SIAM J. Appl. Math. **77**, 593–613 (2017)
20. G. Mascali, V. Romano, Exploitation of the maximum entropy principle in mathematical modeling of charge transport in semiconductors. Entropy **19**(1), 36 (2017)
21. G. Mascali, V. Romano, A hierarchy of macroscopic models for phonon transport in graphene. Physica A **548**, 124489 (2020)
22. O. Morandi, F. Schürrer, Wigner model for quantum transport in graphene. J. Phys. A Math. Theor. **44**(26), 265301 (2011)
23. L. Luca, V. Romano, Quantum corrected hydrodynamic models for charge transport in graphene. Ann. Phys. **406**, 30–53 (2019)
24. A. Majorana, G. Mascali, V. Romano, Charge transport and mobility in monolayer graphene. J. Math. Industry **7**(4), 4 (2016).
25. G. Nastasi, V. Romano, Improved mobility models for charge transport in graphene. Commun. Appl. Ind. Math. **10**(1), 41–52 (2019)
26. G. Nastasi, V. Romano, Simulation of graphene field effect transistors, in *Scientific Computing in Electrical Engineering, SCEE 2018*, Taormina, September 23–27, ed. by G. Nicosia, V. Romano. Mathematics in Industry, vol. 32 (Springer Nature, Switzerland AG, 2020), pp. 171–178

Part III
Computational Electromagnetics

Electric Circuit Element Boundary Conditions in the Finite Element Method for Full-Wave Frequency Domain Passive Devices

Gabriela Ciuprina, Daniel Ioan, Mihai Popescu, and Sorin Lup

Abstract A natural coupling of a circuit with an electromagnetic (EM) device is possible if special boundary conditions, called Electric Circuit Element (ECE), are used for the EM field formulation. This contribution shows how these ECE boundary conditions can be implemented into the finite element method for the solving of coupled full-wave EM field-circuit problems in the frequency domain. The implementation is based on a weak formulation that uses the electric field strength strictly inside the domain and a scalar potential defined solely on the boundary. Edge elements are used inside the three-dimensional domain and nodal elements are used on its two-dimensional boundary surface. The weak formulation is given and its discrete form is validated on a 2D example, with known analytic solution.

1 Motivation

Many EM devices with distributed parameters and field effects specific to full-wave (FW) or Magneto-Quasi-Static (MQS) EM field regime are connected to circuits with lumped parameters (e.g. in measuring and control applications). For this, the EM devices need boundary conditions compatible with external circuits (Fig. 1,left).

By definition, an isolated electric circuit has a finite number of components connected to common terminals. Each terminal is characterized by its voltage with respect to the ground. A non-isolated circuit, i.e. a sub-circuit with m terminal nodes has each of these terminals characterized by a pair of scalar quantities, a current i_k

G. Ciuprina (✉) · D. Ioan · M. Popescu · S. Lup
Politehnica University of Bucharest, Bucharest, Romania
e-mail: gabriela.ciuprina@upb.ro; daniel.ioan@upb.ro; mihai.popescu@upb.ro; sorin.lup@upb.ro

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_10

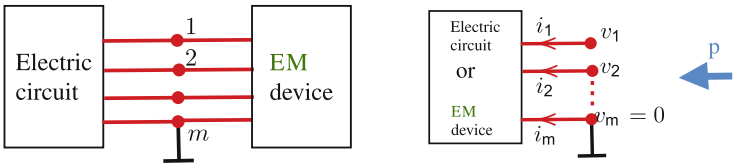


Fig. 1 Left: Coupling of electric circuits and EM device models are naturally ensured by means of terminals. Right: To ensure the coupling, "node voltages" (potentials) and electric currents of non-isolated circuits must have a correspondent in the EM device model

entering into the sub-circuit and a "node voltage" (potential) v_k (Fig. 1,right). The power transferred to it is

$$P = \sum_{k=1}^{m} i_k v_k = \sum_{k=1}^{m-1} i_k (v_k - v_m) = \sum_{k=1}^{m-1} i_k v_k \tag{1}$$

if i_m is expressed according to Kirchhoff current law for a cutset and the terminal m is connected to ground. This power expression shows that the state of a m-terminal circuit is characterized by $2(m-1)$ independent quantities: $m-1$ currents and $m-1$ voltages. The assumption $v_m = 0$ is not a restriction for the purpose of this paper, which is stated at the end of Sect. 2. A natural coupling of this sub-circuit with an EM device is possible if some connecting surfaces are defined on the device's boundary, for which currents and potentials are defined, in order to satisfy Kirchhoff relationships and provide the same transmitted power formula (1) as subcircuits do. The conditions that satisfy these requirements are the ones proposed in [10], used in [4, 8] and called Electric Circuit Element (ECE) boundary conditions.

The ECE boundary conditions, combined with current excited terminals, are the "realistic boundary conditions" used in [1] to solve eddy current problems with the finite element method (FEM) using a formulation in $\vec{H}$ and an ungauged $\vec{T} - \varphi, \varphi$ one in [2]. Similar conditions, although with a different definition for the terminal voltages are proposed in [5] and used for $\vec{A}, V$ eddy current formulations [7].

The use of ECE in MQS problems for inductance extraction with an $\mathbf{A}, V$ formulation is discussed in [9]. Our aim is to use ECE boundary conditions to solve full-wave (FW) problems with FEM. We have successfully used ECE to model passive on-chip components such as resistors, inductors, capacitors, interconnects or RF-MEMS switches in FW [3], with the Finite Integration Technique as numerical method. According to our knowledge, the ECE conditions are not available in FEM codes which implement the formulation of microwave ports for FW. Theoretical studies exists, e.g. in [4], based on an $\vec{E}, V$ formulation for the whole domain. In this paper we use $\vec{E}$ strictly inside the domain and V solely on the boundary. During the reviewing process of this paper, Hiptmair and Ostrowski released a relevant report [6], proving the interest for this subject.

2 ECE Boundary Conditions

Assume a simply connected domain Ω, with a Lipschitz boundary $\partial\Omega$ that includes m disjoint parts $S_k, k = 1, 2, \ldots, m$ (device's terminals), so that conditions (ECE1), (ECE2) and (ECE3) are satisfied (Fig. 2):

- **(ECE1)** there is no magnetic coupling with the exterior: $\vec{n} \cdot \frac{\partial \vec{B}(\mathbf{r},t)}{\partial t} = 0, \quad \forall \mathbf{r} \in \partial\Omega$;
- **(ECE2)** the electric coupling is carried out only through the terminals: $\vec{n} \cdot (\nabla \times \vec{H}(\mathbf{r}, t)) = 0, \quad \forall \mathbf{r} \in \partial\Omega - \cup_{k=1}^{m} S_k$;
- **(ECE3)** the terminals are equipotential: $\vec{n} \times \vec{E}(\mathbf{r}, t) = \vec{0}, \quad \forall \mathbf{r} \in S_k, \ k = 1, \ldots, m$.

According to Faraday's law, (ECE3) implies (ECE1) for the terminals, the inclusion of the terminals in (ECE1) is kept only for emphasizing the physical meaning.

By definition, the currents and potentials of any terminal are:

$$i_k(t) = \oint_{\partial S_k} \vec{H} \cdot \vec{dl} = -\int_{S_k} \left(\vec{J} + \frac{\partial \vec{D}}{\partial t} \right) \cdot \vec{n} \, \mathrm{d}s, \qquad v_k(t) = \int_{C_k \subset \partial\Omega} \vec{E} \cdot \vec{dl}, \tag{2}$$

where, in order to ensure conservation, each terminal current is the total current (conductive and displacement) and the potential is properly defined as the voltage between this terminal and the reference one, along a path C_k included in the domain boundary. Due to (ECE1) the voltage between two points placed on the boundary surface is independent of the path of the integration line connecting these points, provided that this path is included in the surface. Thus, the potential on the surface is well defined, although this is not the case in a general time-varying EM field. Under these conditions, (1) holds for the EM device, where i_k and v_k are given

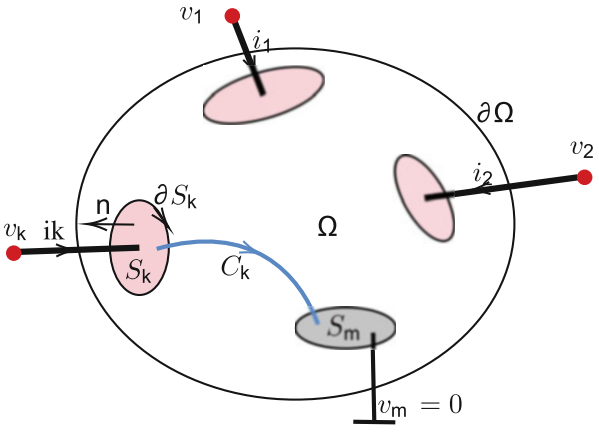


Fig. 2 Electric terminals are disjoint surfaces on the domain's boundary. The non-grounded terminals can be either voltage excited (its potential is given) or current excited (its total current is given)

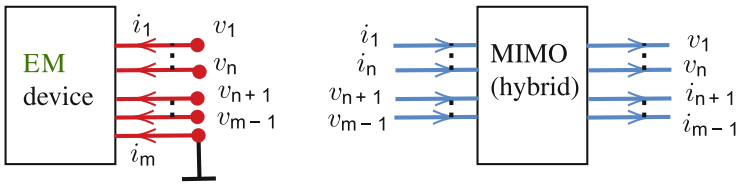


Fig. 3 Each non-grounded terminal of the EM device with ECE boundary conditions can be either current excited or voltage excited. Its hybrid transfer matrix is obtained after computing voltages of the current excited terminals and currents of the voltage excited terminals in linear problems

by (2), and thus the ECE boundary conditions are perfectly compatible with the power transferred through its terminals by a multipolar circuit [8, 10].

If we assume that the terminals have known potentials, then it can be proved that the problem of EM field analysis in a linear domain with ECE boundary conditions has a unique solution. Consequently, the terminal currents are output signals and are obtained by solving the field problem [10]. As the domain is linear, so are the equations, hence the device with ECE conditions is a linear system, defining a multiple input multiple output (MIMO) type dynamic system with $m-1$ inputs and $m-1$ outputs (Fig. 3).

In the frequency domain, the input-output relationship is expressed as:

$$\left[\underline{V}_1 \cdots \underline{V}_n \; \underline{I}_{n+1} \cdots \underline{I}_{m-1}\right]^T = \begin{bmatrix} \underline{\mathbf{Z}} & \underline{\mathbf{A}} \\ \underline{\mathbf{B}} & \underline{\mathbf{Y}} \end{bmatrix} \left[\underline{I}_1 \cdots \underline{I}_n \; \underline{V}_{n+1} \cdots \underline{V}_{m-1}\right]^T. \tag{3}$$

The problem to be solved is: "Find $\begin{bmatrix} \underline{\mathbf{Z}}(f) & \underline{\mathbf{A}}(f) \\ \underline{\mathbf{B}}(f) & \underline{\mathbf{Y}}(f) \end{bmatrix}$, where f is the frequency in a given frequency range of interest, defined by its minimum and maximum values $f_{\min}$ and $f_{\max}$ $f \in [f_{\min}, f_{\max}]$, from the EM field solution." If this hybrid matrix is known, then the "field" element can be realized with common circuit elements and included in any circuit simulator.

3 ECE in FEM

It is useful to recall the formulation in **E** with classical boundary conditions, since the newly proposed formulation inherits a part of it.

Strong Formulation of PDE for E with Classical Boundary Conditions
The well known FW Maxwell equations in the frequency domain, for linear media and no internal field sources are: $\nabla \times \underline{\vec{E}} = -\mathrm{j}\omega\mu\underline{\vec{H}}$, $\nabla \times \underline{\vec{H}} = \sigma\vec{E} + \mathrm{j}\omega\varepsilon\underline{\vec{E}}$, $\nabla \cdot (\mu\underline{\vec{H}}) = 0$, $\nabla \cdot (\varepsilon\underline{\vec{E}}) = \rho$, where permittivity ε, permeability μ and conductivity σ are positive, space dependent material parameters. The reluctivity $\nu = 1/\mu$ might be used instead of μ. The solution of these equations is unique if in any point of

$\partial\Omega$, either exclusively $\underline{\mathbf{E}}_t$ or $\underline{\mathbf{H}}_t$ are known (given). The subscript t indicates the tangential component of the vector on the surface. It is useful to denote a disjoint partition of the boundary: $\partial\Omega = S_E \cup S_H$, $S_E \cap S_H = \emptyset$, and thus $\underline{\mathbf{E}}_t : S_E \to \mathbb{C}^2$, $\underline{\mathbf{H}}_t : S_H \to \mathbb{C}^2$. The imposed boundary conditions are: $\underline{\mathbf{E}}_t(\vec{r}) = \vec{n} \times (\vec{E}(\vec{r}) \times \vec{n})$, for $\vec{r} \in S_E$ and $\underline{\mathbf{H}}_t(\vec{r}) = \vec{n} \times (\vec{H}(\vec{r}) \times \vec{n})$, for $\vec{r} \in S_H$. In what follows we will name them ***classical boundary conditions***. The uniqueness of the field solution can be proven on the basis of the complex form of the Poynting's theorem that gives the expression of the transmitted power (assuming a linear field domain, with no moving parts):

$$-\oint_{\partial\Omega} (\underline{\mathbf{E}}_t \times \underline{\mathbf{H}}_t^*) \cdot \mathbf{n}\, \mathrm{d}s = \int_{\Omega} \underline{\vec{E}} \cdot \underline{\vec{J}}^* + 2\mathrm{j}\omega \int_{\Omega} \left(\frac{\vec{B} \cdot \vec{H}^*}{2} - \frac{\vec{E} \cdot \vec{D}^*}{2} \right). \tag{4}$$

The proof assumes that there exist two such fields that satisfy the same boundary conditions. This means that the Poynting theorem in complex form is valid for the difference field, which satisfies Maxwell's equations (due to linearity) and zero boundary conditions. This implies that the real part is zero which conduces to zero difference electric field (conductivity of the domain is assumed non-zero everywhere) and the imaginary part is zero with conduces to zero difference magnetic field.

The second order equation is:

$$\nabla \times (\nu \nabla \times \underline{\vec{E}}) + \mathrm{j}\omega(\sigma + \mathrm{j}\omega\varepsilon)\underline{\vec{E}} = \vec{0}. \tag{5}$$

Weak Formulation in E with Classical Boundary Conditions

In general, solving of (5) implies a numerical approach, e.g. FEM, which is based on weak formulations. The needed functionals result by projecting (5) onto a set of test functions $\underline{\vec{E}}'$, then integrating by parts and applying Gauss-Ostrogradski formula:

$$\int_{\Omega} \left[(\nu \nabla \times \underline{\vec{E}}) \cdot (\nabla \times \underline{\vec{E}}') + \mathrm{j}\omega(\sigma + \mathrm{j}\omega\varepsilon)\underline{\vec{E}} \cdot \underline{\vec{E}}' \right] \mathrm{d}x = -\oint_{\partial\Omega} \left[(\nu \nabla \times \underline{\vec{E}}) \times \underline{\vec{E}}' \right] \cdot \vec{n}\, \mathrm{d}s$$

Replacing the expression of the magnetic field strength in the right hand side we get

$$\int_{\Omega} \left[(\nu \nabla \times \underline{\vec{E}}) \cdot (\nabla \times \underline{\vec{E}}') + \mathrm{j}\omega(\sigma + \mathrm{j}\omega\varepsilon)\underline{\vec{E}} \cdot \underline{\vec{E}}' \right] \mathrm{d}x = \mathrm{j}\omega \oint_{\partial\Omega} \left(\underline{\vec{H}} \times \underline{\vec{E}}' \right) \cdot \vec{n}\, \mathrm{d}s. \tag{6}$$

With classical boundary conditions, the right hand side is equal to $\int_{S_E} \left(\underline{\vec{E}}_t' \times \vec{n} \right) \cdot \underline{\vec{H}}\, \mathrm{d}s + \int_{S_H} \left(\vec{n} \times \underline{\vec{H}}_t \right) \cdot \underline{\vec{E}}'\, \mathrm{d}s$. $\underline{\vec{E}}_t$ are essential boundary conditions that is why the

test functions are chosen so that $\underline{\vec{E}}_t'$ is zero on S_E. Thus, the weak equation for the trial functions $\underline{\vec{E}}$ is:

$$\int_\Omega \left[(\nu\nabla \times \underline{\vec{E}}) \cdot (\nabla \times \underline{\vec{E}}') + \mathrm{j}\omega(\sigma + \mathrm{j}\omega\varepsilon)\underline{\vec{E}} \cdot \underline{\vec{E}}' \right] \mathrm{d}x = \mathrm{j}\omega \int_{S_H} \left(\vec{n} \times \underline{\vec{H}}_t \right) \cdot \underline{\vec{E}}' \, \mathrm{d}s. \tag{7}$$

The boundary conditions $\underline{\vec{H}}_t$ are natural, they appear in the functional equation.

In conclusion, the weak formulation in $\underline{\vec{E}}$ with classical boundary conditions is: Find $\underline{\vec{E}}$ in $\mathscr{H}$, such that $a(\underline{\vec{E}}, \underline{\vec{E}}') = f(\underline{\vec{E}}'), \quad \forall \underline{\vec{E}}' \in \mathscr{H}_0$ where

$$a((\underline{\vec{E}}, \underline{\vec{E}}') = \int_\Omega \left[(\nu\nabla \times \underline{\vec{E}}) \cdot (\nabla \times \underline{\vec{E}}') + \mathrm{j}\omega(\sigma + \mathrm{j}\omega\varepsilon)\underline{\vec{E}} \cdot \underline{\vec{E}}' \right] \mathrm{d}x, \tag{8}$$

$$f(\underline{\vec{E}}') = \mathrm{j}\omega \int_{S_H} \left(\vec{n} \times \underline{\vec{H}}_t \right) \cdot \underline{\vec{E}}' \, \mathrm{d}s, \tag{9}$$

$$\mathscr{H} = \left\{ \vec{u} \in \mathscr{H}(\mathrm{curl}, \Omega) | \vec{n} \times (\vec{u} \times \vec{n}) = \vec{E}_t \quad \text{on} \quad S_E \right\}, \tag{10}$$

$$\mathscr{H}_0 = \left\{ \vec{u} \in \mathscr{H}(\mathrm{curl}, \Omega) | \vec{n} \times (\vec{u} \times \vec{n}) = \vec{0} \quad S_E \right\}. \tag{11}$$

Discrete Formulation in E with Classical Boundary Conditions

Assume a simplicial mesh (tetrahedrons in 3D, triangles in 2D), numerical test functions $\vec{N}_k$ that correspond to edge elements of order (0,1), and degrees of freedom that represent the complex representations of voltages $\underline{U}_k$ along the edges. The numerical solution is approximated as $\underline{\vec{E}} = \sum_{j=1}^{\mathrm{Ne}} \underline{U}_j \vec{N}_j$, where Ne is the total number of edges in the domain. For any cell, the sum involves 6 terms in 3D and 3 terms in 2D. By substituting the approximation of the numerical solution in (6), choosing the test function $\underline{\vec{E}}' = \vec{N}_i$ and rearranging the sums we obtain a relationship that reveals how the matrices assembling has to be done for all $i = 1, \ldots, \mathrm{Ne}$:

$$\sum_{j=1}^{\mathrm{Ne}} \left\{ \int_\Omega \left[(\nu\nabla \times \vec{N}_j) \cdot (\nabla \times \vec{N}_i) + \mathrm{j}\omega(\sigma + \mathrm{j}\omega\varepsilon)\vec{N}_j \cdot \vec{N}_i \right] \mathrm{d}x \right\} \underline{U}_j = \mathrm{j}\omega \int_{S_H} \left(\vec{n} \times \underline{\vec{H}}_t \right) \cdot \vec{N}_i \, \mathrm{d}s. \tag{12}$$

The initial assembling is carried out for all the edges in the domain. The next step refers to the boundary conditions. Assume that the edges were numbered in the following order: first—the inner edges, second—the edges on the boundary S_H and finally, the edges on the boundary S_E. This leads to the following partitioning:

$$\begin{bmatrix} \mathbf{A}_{\mathrm{in-in}} & \mathbf{A}_{\mathrm{in-SH}} & \mathbf{A}_{\mathrm{in-SE}} \\ \mathbf{A}_{\mathrm{SH-in}} & \mathbf{A}_{\mathrm{SH-SH}} & \mathbf{A}_{\mathrm{SH-SE}} \\ \mathbf{A}_{\mathrm{SE-in}} & \mathbf{A}_{\mathrm{SE-SH}} & \mathbf{A}_{\mathrm{SE-SE}} \end{bmatrix} \begin{bmatrix} \underline{\mathbf{U}}_{\mathrm{in}} \\ \underline{\mathbf{U}}_{\mathrm{SH}} \\ \underline{\mathbf{U}}_{\mathrm{SE}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_{\mathrm{SH}} \\ \mathbf{0} \end{bmatrix} \tag{13}$$

The group of equations that correspond to edges on the S_E boundary is deleted and the essential boundary conditions $\underline{\vec{E}}_t$ are translated into imposed values of electric voltages along edges on the S_E boundary. The system to be solved is

$$\begin{bmatrix} \mathbf{A}_{\text{in-in}} & \mathbf{A}_{\text{in-SH}} \\ \mathbf{A}_{\text{SH-in}} & \mathbf{A}_{\text{SH-SH}} \end{bmatrix} \begin{bmatrix} \underline{\mathbf{U}}_{\text{in}} \\ \underline{\mathbf{U}}_{\text{SH}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_{\text{SH}} \end{bmatrix} - \begin{bmatrix} \mathbf{A}_{\text{in-SE}} \\ \mathbf{A}_{\text{SH-SE}} \end{bmatrix} \left[\underline{\mathbf{U}}_{\text{SE}} \right], \tag{14}$$

the coefficient matrix being symmetric and positive defined.

Weak Formulation in E, V with ECE Boundary Conditions
If we use ECE boundary conditions, the unknowns are the electric field inside the domain and an electric scalar potential solely defined on $\partial\Omega$. That is why the formulation is still named **E**, V, but is different from other formulations, such as the **E**, V in [4] where V is defined also inside the domain. An **E**, V interpretation of the ECE boundary conditions (ECE 1,2,3)) is:

- **(ECE1b)** $\oint_\Gamma \underline{\vec{E}} \cdot \vec{dl} = 0, \quad \forall \Gamma \in \partial\Omega$;
- **(ECE2b)** $\vec{n} \cdot \underline{\vec{E}}(\mathbf{r}) = 0, \quad \forall \mathbf{r} \in \partial\Omega - \cup_{k=1}^{m} S_k$;
- **(ECE3b)** $\underline{\vec{E}}_t(\mathbf{r}) = \vec{0} \quad \forall \mathbf{r} \in S_k, \ k = 1, \ldots, m.$

From (ECE1b) an electric scalar potential $\underline{V}$ can be defined on the boundary $\partial\Omega$, such that $\underline{\vec{E}}_t = -\nabla_2 \underline{V}$. Condition (ECE3b) requires that the electric terminals are equipotential. For uniqueness reasons, one terminal has to be defined by any value. Without lack of generality we can assume it is grounded in what follows. For the other terminals the uniqueness implies that, exclusively, either their voltages or currents are known.

Using (5) we get the weak equation for $\vec{E}$:

$$\int_\Omega \left[(\nu \nabla \times \underline{\mathbf{E}}) \cdot (\nabla \times \underline{\mathbf{E}}') + \mathrm{j}\omega(\sigma + \mathrm{j}\omega\varepsilon) \underline{\mathbf{E}} \cdot \underline{\mathbf{E}}' \right] \mathrm{d}x = \mathrm{j}\omega \sum_{k \in \mathscr{I}_c} \underline{V}'_k \underline{I}_k, \tag{15}$$

where $\mathscr{I}_c$ is the set of indices of current excited terminals. Similarly, we will denote by $\mathscr{I}_v$ is the set of indices of voltage excited terminals. We need an equation for the electric potential on the boundary, as well. Let's denote the normal component of the total current density in any point on the boundary as $\underline{J}_n \overset{\text{not}}{=} (\nabla \times \underline{\mathbf{H}}) \cdot \mathbf{n}$. We will project $\underline{J}_n$ onto a set of scalar test functions $\underline{V}'$:

$$\oint_{\partial\Omega} (\nabla \times \underline{\mathbf{H}}) \cdot \mathbf{n} \, \underline{V}' \, \mathrm{d}s = \oint_{\partial\Omega} \underline{J}_n \underline{V}' \, \mathrm{d}s \overset{\text{(ECE2)}}{=} \sum_{k=1}^{m} \int_{S_k} \underline{J}_n \underline{V}' \, \mathrm{d}s = \sum_{k \in \mathscr{I}_c} \underline{V}'_k \underline{I}_k$$

The integrand of the left hand side can be further computed by using the integration by parts formula that involves the surface differential operators and the

substitution of the magnetic field with its expression with respect to the electric field, as it follows from Faraday's law:

$$\oint_{\partial\Omega}(\nabla\times\underline{\mathbf{H}})\cdot\mathbf{n}\,\underline{V}'\,\mathrm{d}s=\oint_{\partial\Omega}\underline{V}'\,\mathbf{n}\cdot\mathrm{curl}\,\underline{\mathbf{H}}\,\mathrm{d}s\overset{\mathrm{def}}{=}\oint_{\partial\Omega}\underline{V}'\,\mathrm{div}_2\,(\underline{\mathbf{H}})\,\mathrm{d}s=$$
$$=\int_{\partial(\partial\Omega)}\underline{V}'(\mathbf{n}\times\mathbf{H})\,\mathrm{d}s-\oint_{\partial\Omega}\underline{\mathbf{H}}\cdot\mathrm{grad}_2\underline{V}'\,\mathrm{d}s=\oint_{\partial\Omega}\frac{\nu}{j\omega}\mathrm{curl}\,\underline{\mathbf{E}}\cdot\mathrm{grad}_2\underline{V}'\,\mathrm{d}s$$

Consequently it follows that the weak form of the equation on the boundary is

$$\oint_{\partial\Omega}(\nu\nabla\times\underline{\mathbf{E}})\cdot\nabla_2\underline{V}'\,\mathrm{d}s=j\omega\sum_{k\in\mathscr{I}_c}\underline{V}'_k\,\underline{I}_k \tag{16}$$

Finally, we get the weak formulation in $\underline{\vec{E}}$, V with ECE boundary conditions. Find $\underline{\mathbf{E}}\in\mathscr{H}_E$, $\underline{V}\in\mathscr{H}_V$, such that

$$a(\underline{\mathbf{E}},\underline{\mathbf{E}}')=f(\underline{\mathbf{E}}'),\quad\forall\underline{\mathbf{E}}'\in\mathscr{H}_{E,0};\qquad b(\underline{\mathbf{E}},\underline{V}')=g(\underline{V}'),\quad\forall\underline{V}'\in\mathscr{H}_{V,0}$$
$$\oint_{\partial S_k}\mathbf{H}\cdot\mathbf{dl}=\underline{I}_k,\quad k\in\mathscr{I}_c;\qquad\qquad\underline{\mathbf{E}}_t=-\nabla_2 V,\quad\text{on}\quad\partial\Omega,$$

where

$$a(\underline{\mathbf{E}},\underline{\mathbf{E}}')=\int_\Omega\left[(\nu\nabla\times\underline{\mathbf{E}})\cdot(\nabla\times\underline{\mathbf{E}}')+\mathrm{j}\omega(\sigma+\mathrm{j}\omega\varepsilon)\underline{\mathbf{E}}\cdot\underline{\mathbf{E}}'\right]\mathrm{d}x,\qquad f(\underline{\mathbf{E}}')=\mathrm{j}\omega\sum_{k\in\mathscr{I}_c}\underline{V}'_k\underline{I}_k;$$
$$b(\underline{\mathbf{E}},\underline{V}')=\oint_{\partial\Omega}(\nu\nabla\times\underline{\mathbf{E}})\cdot\nabla_2\underline{V}'\,\mathrm{d}s,\qquad g(\underline{V}')=\mathrm{j}\omega\sum_{k\in\mathscr{I}_c}\underline{V}'_k\underline{I}_k;$$

where $\underline{\mathbf{E}}'_t=-\nabla_2 V'$.

$$\mathscr{H}_E=\{\mathbf{u}\in\mathscr{H}(\mathrm{curl},\Omega)|\,\mathbf{n}\times(\mathbf{u}\times\mathbf{n})=-\nabla_2\underline{V}'\ \text{on}\ \partial\Omega,\quad\underline{V}'\in\mathscr{H}_V$$
$$\mathbf{n}\times(\mathbf{u}\times\mathbf{n})=\mathbf{0}\ \text{on}\ \cup_{k=1}^m S_k\}$$
$$\mathscr{H}_{E,0}=\{\mathbf{u}\in\mathscr{H}(\mathrm{curl},\Omega)|\,\mathbf{n}\times(\mathbf{u}\times\mathbf{n})=-\nabla_2\underline{V}'\ \text{on}\ \partial\Omega,\quad\underline{V}'\in\mathscr{H}_{V,0}$$
$$\mathbf{n}\times(\mathbf{u}\times\mathbf{n})=\mathbf{0}\ \text{on}\ \cup_{k=1}^m S_k\}$$
$$\mathscr{H}_V=\{u\in\mathscr{H}(\mathrm{grad},\partial\Omega)|\ u=\underline{V}_k\ \text{on}\ S_k,\ k\in\mathscr{I}_v,$$
$$u=\text{constant(unkown) on}\ S_k,\ k\in\mathscr{I}_c\}$$

$$\mathscr{H}_{V,0}=\{u\in\mathscr{H}(\mathrm{grad},\partial\Omega)|\ u=0\ \text{on}\ S_k,\ k\in\mathscr{I}_v$$
$$u=\text{constant(unkown) on}\ S_k,\ k\in\mathscr{I}_c\}$$

Note: We have investigated two other formulations for the boundary equations for which $b(\mathbf{E}, V') = 0$. In one version $b(\mathbf{E}, V') = \oint_{\partial\Omega}(\sigma + \mathrm{j}\omega\varepsilon)(\nabla_2 \underline{V}) \cdot (\nabla_2 \underline{V'})\,\mathrm{d}s + \oint_{\partial\Omega} \frac{\partial}{\partial n}\left[(\sigma + \mathrm{j}\omega\varepsilon)\mathbf{E} \cdot \mathbf{n}\right] \underline{V'}\,\mathrm{d}s$ and another version is $b(\mathbf{E}, V') = \oint_{\partial\Omega}(\sigma + j\omega\varepsilon)\mathbf{n} \cdot \mathbf{E}V'\,\mathrm{d}s$. Due to lack of space we will not present them here.

Formulation in E, V with ECE Boundary Conditions—Algorithm in FEM

Step 1 We start with the discrete form of classical BC, given by (14), written for all the edges) $\begin{bmatrix} \mathbf{A}_{u,u} & \mathbf{A}_{u,u_b} \\ \mathbf{A}_{u_b,u} & \mathbf{A}_{u_b,u_b} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{u}_b \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_b \end{bmatrix}$. Only the first block row of equations, corresponding to the inner edges, is kept.

Step 2 Write the discrete form of the Eq. (16) on the 2D surface boundary mesh. $\sum_{j=1}^{Ne} \left[\oint_{\partial\Omega}(\nu\nabla \times \mathbf{N}_j) \cdot (\nabla_2 \underline{\varphi_i'})\,\mathrm{d}s\right] \underline{U_j} = j\omega \underline{I_i}$, where $\underline{\varphi_i'}$ is the nodal element i. This is written for all the nodes on the boundary and will be placed together with the discrete equation obtained at step 1: $\begin{bmatrix} \mathbf{A}_{u,u} & \mathbf{A}_{u,u_b} \\ \mathbf{A}_{V_b,u} & \mathbf{A}_{V_b,u_b} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{u}_b \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_b' \end{bmatrix}$.

Step 3 On the boundary, the variables are changed, from electric voltages to electric potentials, by expressing $\mathbf{u}_b$ as potential differences. The system becomes $\begin{bmatrix} \mathbf{A}_{u,u} & \mathbf{A}_{u,V_b} \\ \mathbf{A}_{V_b,u} & \mathbf{A}_{V_b,V_b} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{V}_b \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_b' \end{bmatrix}$.

Step 4 Finally, $\mathbf{V}_b$ has to be split in three ($\mathbf{V}$-for nodes that are not on terminals, $\mathbf{V}_{t,c}$-voltages of current excited terminals, $\mathbf{V}_{t,v}$-voltages of voltage excited terminals), in order to impose the rest of the natural conditions (potentials for voltage excited, or currents for current excited terminals): Finally, the system to solve is

$$\begin{bmatrix} \mathbf{A}_{u,u} & \mathbf{A}_{u,V} & \mathbf{A}_{u,V_{t,c}} \\ \mathbf{A}_{V,u} & \mathbf{A}_{V,V} & \mathbf{A}_{V,V_{t,c}} \\ \mathbf{A}_{V_{t,c},u} & \mathbf{A}_{V_{t,c},V} & \mathbf{A}_{V_{t,c},V_{t,c}} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{V} \\ \mathbf{V}_{t,c} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ j\omega\mathbf{I}_{t,c} \end{bmatrix} - \begin{bmatrix} \mathbf{A}_{u,V_{t,v}} \\ \mathbf{A}_{V,V_{t,v}} \\ \mathbf{A}_{V_{t,c},V_{t,v}} \end{bmatrix} \left[\mathbf{V}_{t,v}\right].$$

After solving, we get the unknown potentials $\mathbf{V}$ and $\mathbf{V_{t,c}}$. The currents through the terminals in $\mathscr{I}_v$ can be computed as a postprocessing step.

3.1 Numerical Results

Figure 4 shows a quantitative validation for a 2D simple case, with two terminals and with analytical solution. It is a single input single output (SISO) system, both current and voltage excitations give accurate results. The domain is a brick that occupies the space $x \in [-a, a]$, $y \in [0, l]$ and $x \in [0, h]$. One excited terminal (in voltage or in current) is on the $z = 0$ boundary and the grounded terminal is on the $z = h$ boundary. The material inside is assumed homogeneous with ε, μ, σ. The analytic solution can be obtained by solving the Helmholtz equations and considering the current excited terminal ($\underline{I}$). The complex power absorbed by this

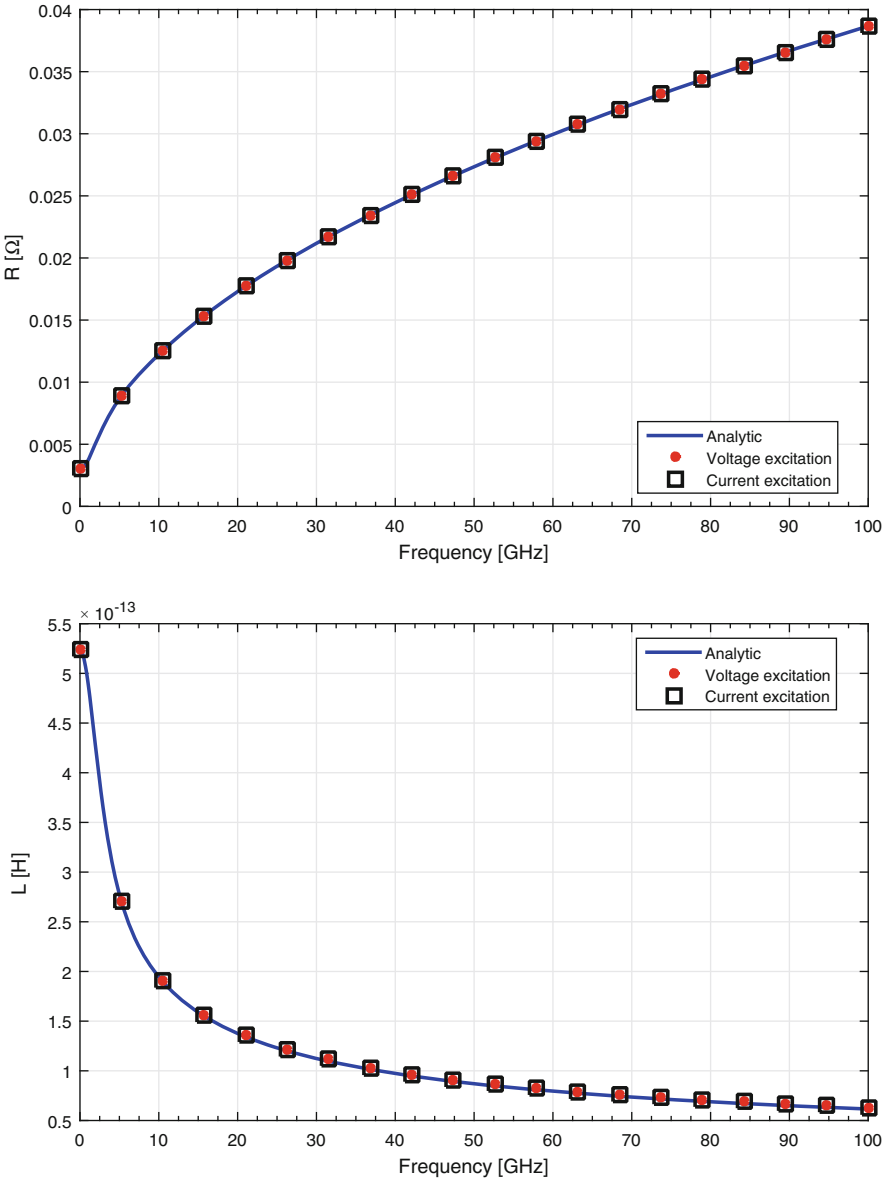


Fig. 4 Quantitative validation of the implementation for a 2D case with analytical solution. The problem is a rectangle with two opposite terminals, consequently the system is SISO. Both voltage and current excitations lead to relative errors less than 2% for the whole frequency range

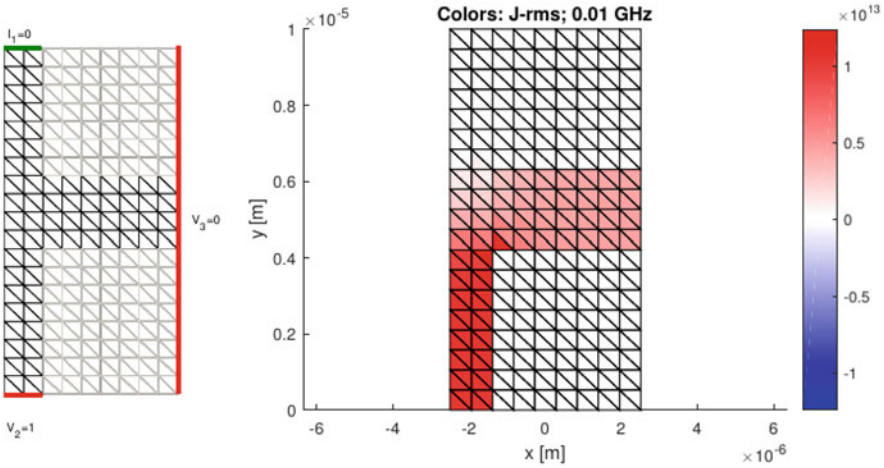


Fig. 5 Qualitative validation for a 2D case, MIMO (3 terminals), hybrid excitation (one terminal grounded, one is voltage excited and one is current excited)

domain is $\underline{P} = 2\underline{E}_y \, \underline{H}_z^* \, lh$, where $\underline{E}_y = \underline{\gamma}/(\sigma + j\omega\varepsilon) \cosh(\underline{\gamma} a)/\sinh(\underline{\gamma} a) \, \underline{I}/(2h)$ and $\underline{H}_z = \underline{I}/(2h)$. The extracted complex impedance is $\underline{Z} = \underline{P}/|\underline{I}|^2$ and its components shown in Fig. 4 are $R = \text{real}\,\underline{P}$ and $L = \text{real}\,\underline{P}/\omega$ for $a = 2.5\,\mu\text{m}$, $l = 10\,\mu\text{m}$, $h = 10\,\mu\text{m}$, $\sigma = 6.6 \cdot 10^7\,\text{S/m}$, $\mu = \mu_0$, $\varepsilon = \varepsilon_0$, $f_{min} = 0.01\,\text{GHz}$, $f_{max} = 100\,\text{GHz}$.

Figure 5 shows a qualitative validation for a MIMO test. The rectangular domain is occupied by a T-shape conductor of high conductivity, having 3 terminals, out of which the one at the right hand side of the figure is grounded.

4 Conclusions

The advantages of ECE BC for Maxwell equations are that the ports are clearly and well defined, without ambiguity, fully compatible with the circuit terminals. There is no restriction on the field regime (full wave, nonlinear). For MIMO systems, the hybrid excitation is obtained in a natural way. This paper proposed a FEM algorithm for ECE, which **E** strictly inside the domain and V on the boundary. The degrees of freedom are the electric voltages on the inner edges and the potentials of the floating nodes on the boundary (nodes outside terminals and current excited terminals). Our next research will compare the 3 mentioned formulations.

Acknowledgments S. Lup acknowledges the support of the Operational Programme Human Capital of the Ministry of European Funds through 51675/09.07.2019, SMIS code 125125.

References

1. A. Bermudez, R. Rodriguez, P. Salgado, Numerical solution of eddy current problems in bounded domains using realistic boundary conditions. Comput. Methods Appl. Mech. Eng. **194**(2), 411–426 (2005)
2. A. Bermudez, M. Pineiro, R. Rodriguez, P. Salgado, Analysis of an ungauged T,φ-φ formulation of the eddy current problem with currents and voltage excitations. ESAIM Math. Modell. Numer. Anal. (ESAIM: M2AN) **51**(6), 2487–2509 (2017)
3. G. Ciuprina, A.S. Lup, C.B. Diţă, D. Ioan, A. Stefanescu, Extraction of TL-lumped RF macromodels MEMS switches. Nr. Elmag. Multiphy. Mod. and Opt., NEMO, Canada (2015)
4. I.F. Hantila, D. Ioan, Voltage-current relation of circuit elements with field effects. Rev. Roumaine des Sci. Tech. Serie Electrotechnique et Energetique **39**(3), 405–416 (1994)
5. R. Hiptmair, O. Sterz, Current and voltage excitations for the eddy current model. Int. J. Numer. Modell. Electron. Networks Dev. Fields **18**(1), 1–21 (2005)
6. R. Hiptmair, J. Ostrowski, Electromagnetic port boundary conditions: Topological and variational perspective. Technical Report 2020–27, Seminar for Applied Mathematics, ETH Zürich, 2020. Int. J. Numer. Modell. **34**(3), 1–23 (2020)
7. R. Hiptmair, F. Kramer, J. Ostrowski, A robust Maxwell formulation for all frequencies. IEEE Trans. Magnet. **44**(6), 682–685 (2008)
8. D. Ioan, I. Munteanu, Missing link rediscovered: The electromagnetic circuit element concept. JSAEM Stud. Appl. Electromagn. Mech. **8**, 302–320 (1999)
9. S. Kurz, Some remarks about flux linkage and inductance. Adv. Radio Sci. **2**, 39–44 (2004)
10. R. Radulet et al., Introduction des parametres transitoires dans l'etude des circuits electrique lineaires ayant des elements non filiformes et avec pertes suplimen-taires. Rev. Roum. Sci Techn. Electrotech. et Energ **11**(4), 565–639 (1966)

A Convolution Quadrature Method for Maxwell's Equations in Dispersive Media

Jürgen Dölz, Herbert Egger, and Vsevolod Shashkov

Abstract We study the systematic numerical approximation of Maxwell's equations in dispersive media. Two discretization strategies are considered, one based on the traditional leapfrog time integration method and the other based on convolution quadrature. The two schemes are proven to be equivalent and to preserve the underlying energy-dissipation structure of the problem. The second approach, however, is independent of the number of internal states and in principle allows to handle rather general dispersive materials. Using ideas of fast-and-oblivious convolution quadrature, the method can be implemented efficiently.

1 Introduction

We consider electromagnetic wave propagation through linear dispersive media. The underlying physics are described by Maxwell's equations

$$\partial_t \mathbf{d} = \operatorname{curl} \mathbf{h}, \qquad \partial_t \mathbf{b} = -\operatorname{curl} \mathbf{e} \tag{1}$$

with $\mathbf{e}$, $\mathbf{h}$ and $\mathbf{d}$, $\mathbf{b}$ denoting the electric and magnetic fields and fluxes, respectively, which are mutually related by the constitutive relations

$$\mathbf{b} = \mu_0 \mathbf{h}, \qquad \mathbf{d} = \epsilon_0 \epsilon_\infty \mathbf{e} + \mathbf{p}. \tag{2}$$

Here ϵ_0, μ_0 are the permittivity and permeability of vacuum, and $\epsilon_\infty = 1 + \epsilon'_\infty$ is the high frequency limit of the relative permittivity. Further, $\mathbf{p}$ denotes the memory

J. Dölz
University of Twente, Enschede, Netherlands
e-mail: j.dolz@utwente.nl

H. Egger (✉) · V. Shashkov
TU Darmstadt, Darmstadt, Germany
e-mail: herbert.egger@tu-darmstadt.de; shashkov@mathematik.tu-darmstadt.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_11

part of the polarization $\mathbf{p}_{tot} = \epsilon_0 \epsilon'_\infty \mathbf{e} + \mathbf{p}$, which is described in frequency domain by

$$\hat{\mathbf{p}}(s) = \epsilon_0 \hat{\chi}(s) \hat{\mathbf{e}}(s). \tag{3}$$

The system is complemented by appropriate boundary and initial conditions. For ease of presentation, we assume that $\mathbf{e}(0) = \mathbf{p}(0) = 0$ in the following. By the convolution theorem for the Laplace-transform, see e.g. [21, Ch 12], the polarization can then be expressed in time domain by

$$\mathbf{p}(t) = \epsilon_0 \int_0^t \chi(t-r)\mathbf{e}(r)dr. \tag{4}$$

We further assume throughout the paper that the susceptibility kernel χ can be written as a superposition of simple Debye functions [4], i.e.,

$$\hat{\chi}(s) = \sum\nolimits_i \hat{\chi}_i(s) \qquad \text{with} \qquad \hat{\chi}_i(s) = \frac{\epsilon_{i,s} - \epsilon'_{i,\infty}}{1 + s\tau_i}, \tag{5}$$

where τ_i denotes the relaxation time and $\epsilon_{i,s} > \epsilon'_{i,\infty}$ are the static and high-frequency limits of the electric susceptibility of the ith component with $\sum_i \epsilon'_{i,\infty} = \epsilon'_\infty$. Such multipole Debye models have been used, e.g., for the modeling of the dielectric response of biological tissue; see [2, 5] and the references given there. In general, the summation in (5) may be over infinitely many terms.

One of the key features of the multipole Debye model is its provable passivity, which follows from the energy–dissipation principle [1, 12]

$$\frac{d}{dt}\mathscr{E} = -\sum\nolimits_i \| \sqrt{\tfrac{\tau_i}{\epsilon_0(\epsilon_{i,s} - \epsilon'_{i,\infty})}} \partial_t \mathbf{p}_i \|^2, \tag{6}$$

valid for any sufficiently smooth solution of (1)–(3) with homogeneous or periodic boundary conditions. Here $\| \cdot \|$ is the L^2-norm, further $\mathbf{p} = \sum_i \mathbf{p}_i$ is the decomposition of the memory part of the polarization into its components according to (5), and

$$\mathscr{E} = \frac{1}{2}\left(\| \sqrt{\mu_0}\mathbf{h} \|^2 + \| \sqrt{\epsilon_0 \epsilon_\infty}\mathbf{e} \|^2 + \sum\nolimits_i \| \tfrac{1}{\sqrt{\epsilon_0(\epsilon_i - \epsilon'_\infty)}} \mathbf{p}_i \|^2 \right) \tag{7}$$

denotes the electromagnetic energy of the system. Due to the rational structure of the transfer functions $\hat{\chi}_i$, the individual polarizations $\mathbf{p}_i$ can be characterized equivalently by the differential equations

$$\tau_i \partial_t \mathbf{p}_i + \mathbf{p}_i = \epsilon_0(\epsilon_{i,s} - \epsilon'_{i,\infty})\mathbf{e}, \tag{8}$$

with initial values $\mathbf{p}_i(0) = 0$, which is the basis for various simulation methods. Corresponding finite difference and finite element schemes have been considered, for instance, in [1, 6, 9, 10, 12, 13, 17, 20]. Let us note that with increasing number of internal states $\mathbf{p}_i$, all methods become computationally more and more expensive.

In this paper, we consider a different approach for the numerical solution of (1)–(3), which allows us to compute the time evolution of $\mathbf{e}$, $\mathbf{h}$, and $\mathbf{p}$ without explicitly computing the internal states $\mathbf{p}_i$. As indicated in [8], this can be accomplished through discretization of the integral (4) by means of appropriate convolution quadratures [14, 16], instead of integrating (8) with time-differencing schemes. The complexity of every time step is then independent of the number of internal states $\mathbf{p}_i$. Moreover, using ideas of [18, 19], the additional memory cost for storing the history of the field $\mathbf{e}$ can be reduced to the logarithm of the number of time steps.

2 Structure Preserving Discretization

After space discretization by appropriate finite-difference or finite-element methods and time-discretization by the leapfrog scheme, the system (1)–(2) with polarization components defined by (8) can be written in matrix–vector notation as

$$\mathsf{M}_\mathsf{h}\, d_\tau\, \mathbf{h}^n + \mathsf{C}\, \mathbf{e}^n = 0, \tag{9}$$

$$\mathsf{M}_\mathsf{e}\, d_\tau\, \mathbf{e}^{n+1/2} + \sum\nolimits_i d_\tau\, \mathbf{p}_i^{n+1/2} - \mathsf{C}^\mathsf{T}\, \overline{\mathbf{h}}^{n+1/2} = 0, \tag{10}$$

$$\mathsf{M}_{\mathsf{d},\mathsf{i}}\, d_\tau\, \mathbf{p}_i^{n+1/2} + \mathsf{M}_{\mathsf{p},\mathsf{i}}\, \overline{\mathbf{p}}_i^{n+1/2} = \overline{\mathbf{e}}^{n+1/2}, \qquad i \geq 1. \tag{11}$$

The equations hold for all $n \geq 0$ and are complemented by appropriate initial conditions. Note that $\mathbf{e}^n$ and $\mathbf{h}^{n+1/2}$ are the approximations for $\mathbf{e}(t^n)$ and $\mathbf{h}(t^{n+1/2})$ at staggered grid points $t^r = r\tau$ with τ denoting the time step size. Furthermore, $d_\tau\, \mathbf{e}^{n+1/2} = \frac{1}{\tau}(\mathbf{e}^{n+1} - \mathbf{e}^n)$ and $d_\tau\, \mathbf{h}^n = \frac{1}{\tau}(\mathbf{h}^{n+1/2} - \mathbf{h}^{n-1/2})$ are the central difference quotients, and $\overline{\mathbf{e}}^{n+1/2} = \frac{1}{2}(\mathbf{e}^{n+1} + \mathbf{e}^n)$, $\overline{\mathbf{h}}^{n+1/2} = \frac{1}{2}(\mathbf{h}^{n+1} + \mathbf{h}^n)$ and $\mathbf{p}^{n+1/2} = \frac{1}{2}(\mathbf{p}^{n+1} + \mathbf{p}^n)$ are the averages of two consecutive steps. Further note that Eq. (11) was obtained from (8) after dividing by $\epsilon_0(\epsilon_{i,s} - \epsilon'_{i,\infty})$.

For appropriate space discretization schemes, the mass matrices M_h, M_e are symmetric, positive-definite, and diagonal or block-diagonal [3, 7], such that (9)–(11) amounts to an explicit time-stepping scheme. Moreover, the method satisfies the following discrete equivalent of the underlying energy–dissipation identity.

Lemma 1 *Set* $\|a\|_\mathsf{M}^2 = (a, a)_\mathsf{M}$ *and* $(a, b)_\mathsf{M} = b^\mathsf{T}\mathsf{M}a$, *and denote by*

$$\mathscr{E}^n = \frac{1}{2}\left((\mathbf{h}^{n+1/2}, \mathbf{h}^{n-1/2})_{\mathsf{M}_\mathsf{h}} + \|\mathbf{e}^n\|_{\mathsf{M}_\mathsf{e}}^2 + \sum\nolimits_i \|\mathbf{p}_i^n\|_{\mathsf{M}_{\mathsf{p},\mathsf{i}}}^2\right)$$

the discrete energy at time step $t^n = n\tau$. Then any solution of (9)–(11) satisfies

$$d_\tau \mathscr{E}^{n+1/2} = -\sum\nolimits_i \|d_\tau \mathbf{p}_i^{n+1/2}\|^2_{\mathsf{M}_{\mathsf{d},\mathsf{i}}}, \qquad n \geq 0.$$

Note that a CFL condition is required to ensure $\mathscr{E}^n \geq 0$; see Remark 1 below.

Proof By elementary computations, one can verify that

$$d_\tau \mathscr{E}^{n+1/2} = \frac{1}{2}(d_\tau \mathbf{h}^{n+1} + d_\tau \mathbf{h}^n, \mathbf{h}^{n+1/2})_{\mathsf{M}_\mathsf{h}} + (d_\tau \mathbf{e}^{n+1/2}, \overline{\mathbf{e}}^{n+1/2})_{\mathsf{M}_\mathsf{e}} + \sum\nolimits_i (d_\tau \mathbf{p}_i^{n+1/2}, \overline{\mathbf{p}}_i^{n+1/2})_{\mathsf{M}_{\mathsf{p},\mathsf{i}}}.$$

Note that $(a, b)_M = (Ma, b) = (Mb, a)$ where $(\cdot, \cdot)$ denotes the Euclidean scalar product. We then test equation (10) with $\overline{\mathbf{e}}^{n+1/2}$ and (11) with $d_\tau \mathbf{p}^{n+1/2}$. Moreover, we test the average of Eq. (9) for step n and $n+1$ with $\mathbf{h}^{n+1/2}$. This allows to replace all terms on the right hand side of the above formula and leads to

$$d_\tau \mathscr{E}^{n+1/2} = -(\mathsf{C}\overline{\mathbf{e}}^{n+1/2}, \mathbf{h}^{n+1/2}) + (\mathsf{C}^\mathsf{T}\mathbf{h}^{n+1/2} - \sum\nolimits_i d_\tau \mathbf{p}_i^{n+1/2}, \overline{\mathbf{e}}^{n+1/2}) + \sum\nolimits_i (\overline{\mathbf{e}}^{n+1/2} - \mathsf{M}_{\mathsf{d},\mathsf{i}} d_\tau \mathbf{p}_i^{n+1/2}, d_\tau \mathbf{p}_i^{n+1/2}).$$

Using that $(\mathsf{C}a, b) = (\mathsf{C}^\mathsf{T}b, a)$, one can see that most of the terms drop out and we obtain the assertion of the lemma. □

Remark 1 Method (9)–(11) automatically inherits the energy-dissipation principle of the continuous problem. We therefore call it a *structure-preserving* discretization scheme. The first term in the energy $\mathscr{E}$ can be estimated from below by

$$\begin{aligned}(\mathbf{h}^{k+1/2}, \mathbf{h}^{k-1/2})_{\mathsf{M}_\mathsf{h}} &= \|\mathbf{h}^{k+1/2}\|^2_{\mathsf{M}_\mathsf{h}} + \tau(\mathbf{h}^{k+1/2}, d_\tau \mathbf{h}^k)_{\mathsf{M}_\mathsf{h}} \\ &= \|\mathbf{h}^{k+1/2}\|^2_{\mathsf{M}_\mathsf{h}} - \tau(\mathsf{C}\mathbf{e}^k, \mathbf{h}^k) \geq \frac{1}{2}\|\mathbf{h}^{k+1/2}\|^2_{\mathsf{M}_\mathsf{h}} - \frac{\tau^2}{2}\|\mathsf{C}\mathbf{e}^k\|^2_{\mathsf{M}_\mathsf{h}^{-1}},\end{aligned}$$

and the last term can be further bounded from below under the assumption that

$$\tau^2 \|\mathsf{C}\mathbf{v}\|^2_{\mathsf{M}_\mathsf{h}^{-1}} \leq \|\mathbf{v}\|^2_{\mathsf{M}_\mathsf{e}} \qquad \text{for all vectors } \mathbf{v}. \tag{12}$$

This CFL condition, restricting the time step τ in dependence of the space discretization, implies stability of the scheme and allows to show that the energy $\mathscr{E}^n$ is a positive and symmetric quadratic functional and thus induces a norm on the space of state vectors $(\mathbf{h}^{n+1/2}, \mathbf{e}^n, \mathbf{p}_1^n, \mathbf{p}_2^n, \ldots)$. Together with Lemma 1, this is the basis for the error analysis of method (9)–(11); we refer to [11] for details.

3 A Convolution Quadrature Approach

The dimension of the state space and hence also the computational cost for computing one time step of method (9)–(11) obviously increases with increasing number of internal states $\mathbf{p}_i$. We will now show that $\mathbf{e}$, $\mathbf{h}$, and $\mathbf{p} = \sum_i \mathbf{p}_i$ can be computed without explicit reference to the internal states $\mathbf{p}_i$, which results in an algorithm that is *independent of the number of internal states*. Instead of using Eq. (8), we directly discretize the integral (4) by a convolution sum

$$\mathbf{p}^n = \sum\nolimits_{k=0}^{n} \omega_{n-k} \mathbf{e}^k. \tag{13}$$

This is the field of convolution quadrature, and we refer to [14, 16] for details on the mathematical background. As illustrated in [8], a proper choice of the convolution weights $\{\omega_n\}_{n\geq 0}$ allows to obtain the following equivalence statement.

Lemma 2 *Let $\{\omega_n\}_{n\geq 0}$ be the coefficients of the power series*

$$\epsilon_0 \hat{\chi}\left(\tfrac{2(1-\xi)}{\tau(1+\xi)}\right) = \sum_{n=0}^{\infty} \omega_n \xi^n. \tag{14}$$

Then the solution $\{\mathbf{h}^{n+1/2}, \mathbf{e}^n, \mathbf{p}^n\}_{n\geq 0}$ of the scheme (9)–(11) with $\mathbf{e}^0 = \mathbf{p}_i^0 = 0$ coincides with the solution of the convolution-quadrature method (9)–(10) and (13).

Proof For convenience of the reader, we briefly summarize the basic ideas of the proof, which closely follows the arguments presented in [8]. We start by multiplying equations (11) with ξ^n and sum over all $n \geq 0$ to obtain

$$\sum\nolimits_{n\geq 0} \mathsf{M}_{\mathsf{d},\mathsf{i}} \left(\tfrac{1}{\xi} - 1\right) \mathbf{p}_i^n \xi^n + \sum\nolimits_{n\geq 0} \mathsf{M}_{\mathsf{p},\mathsf{i}} \left(\tfrac{1}{2\xi} + \tfrac{1}{2}\right) \mathbf{p}_i^n \xi^n = \sum\nolimits_{n\geq 0} \left(\tfrac{1}{2\xi} + \tfrac{1}{2}\right) \mathbf{e}^n \xi^n.$$

An appropriate rearrangement of terms then further leads to

$$\sum\nolimits_{n\geq 0} \mathbf{p}_i^n \xi^n = \hat{\chi}_i \left(\tfrac{2(1-\xi)}{\tau(1+\xi)}\right) \sum\nolimits_{n\geq 0} \mathbf{e}^n \xi^n,$$

with transfer function $\hat{\chi}_i$ as defined in (5). Summation over all i and using $\mathbf{p}^n = \sum_i \mathbf{p}_i^n$ and the definition of the weights ω_n then yields the assertion. □

Remark 2 According to the above lemma, the convolution quadrature (CQ) method defined by (9)–(10) and (13)–(14) has the same passivity and stability properties as the underlying difference scheme (9)–(11). Let us note that instead of the internal states $\{\mathbf{p}_i^n\}_{i\geq 0}$, the CQ approach utilizes the history $\{\mathbf{e}^k\}_{k\leq n}$ of the electric field values to compute the memory part $\mathbf{p}^n$ of the polarization.

Before closing this section, we briefly comment on the practical computation of the weights $\{\omega_n\}_{n\geq 0}$ and the efficient realization of the proposed CQ approach.

Remark 3 Following [14, 15], also see [8], the convolution weights $\{\omega_n\}_{n\geq 0}$ can be computed with high accuracy using fast Fourier transforms, i.e.,

$$\omega_n \approx \tfrac{1}{L\rho^n} \sum_{\ell=0}^{L-1} \hat{\chi}\left(\tfrac{2}{\tau}\tfrac{1-\rho e^{\mathrm{i}\phi_\ell}}{1+\rho e^{\mathrm{i}\phi_\ell}}\right) e^{-\mathrm{i}n\phi_\ell}, \qquad \phi_\ell = 2\pi\ell/L,$$

and the quadrature error can be controlled by appropriate choice of the parameters L and ρ; see [14–16] for details. The computation of all weights $\{\omega_n\}_{n=0}^N$ with machine precision requires $O(N)$ evaluations of $\hat{\chi}$. If the material parameters are inhomogeneous, then the weights ω_n will also depend on the spatial variable.

Remark 4 A direct implementation of the CQ approach requires the storage of the complete history $\{\mathbf{e}^k\}_{k\leq n}$ to compute the polarization $\mathbf{p}^n$ via (13), and a naive computation of the N convolution sums $\{\mathbf{p}^n\}_{n\leq N}$ is of $O(N^2)$ complexity; this can be reduced to $O(N\log^2 N)$ by FFT [19]. Using *fast and oblivious convolution quadrature* (FOCQ), the required storage can be reduced to $O(\log N)$ field vectors [18, 19]. The basic idea of these approaches is to split the sum (13) into subsums with exponentially growing number of summands

$$\sum_{k=0}^{n} \omega_k \mathbf{e}^{n-k} = \sum_{\ell=0}^{L} \sum_{k=B^{\ell-1}}^{B^\ell-1} \omega_k \mathbf{e}^{n-k} =: \sum_{\ell=0}^{L} U_n^\ell,$$

where $B > 1$ is an integer; we set $B^{-1} := 0$ and further assumed for simplicity that $n+1 = B^L$ is a power of the basis B. Under certain regularity assumptions on $\hat{\chi}$, each subsum U_n^ℓ can be approximated efficiently using interpolation [18] or contour integration [19]. In comparison to standard CQ, which requires $O(N)$ historic field values and $O(N)$ evaluations of the transfer function $\hat{\chi}$ to compute all weights ω_n, the FOCQ algorithm only requires $O(\log N)$ historic field vectors and $O(\log N)$ evaluations of the transfer function $\hat{\chi}$, which also improves the setup cost substantially.

4 Numerical Illustration

In our test problem, we consider the propagation of an electromagnetic pulse across the interface between air and human tissue. The dielectric response of the tissue is characterized by a five-pole Debye model which was taken from [6]. Using the notation of Sect. 1, the total polarization in this model is prescribed in frequency domain by $\hat{\mathbf{p}}_{tot}(s) = \epsilon_0(\epsilon'_\infty + \hat{\chi}(s))\hat{\mathbf{e}}(s)$ with $\epsilon'_\infty = 3.3$ and

$$\hat{\chi}(j\omega) = \frac{8.5\cdot 10^5}{1+j\omega/(138\pi)} + \frac{8.19\cdot 10^3}{1+j\omega/(86\pi\cdot 10^3)} + \frac{1.19\cdot 10^3}{1+j\omega/(1.34\pi\cdot 10^6)} + \frac{32}{1+j\omega/(460\pi\cdot 10^6)} + \frac{45.8}{1+j\omega/(40\pi\cdot 10^9)}.$$

For our computational tests, we consider a plane wave setting, in which the fields are of the form $\mathbf{e} = (e_x, 0, 0)$, $\mathbf{h} = (0, h_y, 0)$, and $\mathbf{p}_i = (p_{x,i}, 0, 0)$, and only depend time t and the propagation direction z. Then (1)–(4) leads to a one–dimensional wave propagation problem for unknown fields e_x, p_x and h_y. As computational domain, we consider the interval $(-1, 1)$ and we impose periodic boundary conditions for the electric and magnetic field. The initial values are described by $e_{x,0}(z) = p_{x,i,0}(z) = 0$ and $h_{y,0}(z) = 10e^{-10z^2}$. All quantities are given in SI-units.

For the spatial discretization, we utilize piecewise linear finite elements for e_x and $p_{x,i}$, and piecewise constants to represent h_y. Numerical integration by the vertex rule is used for the assembling of the mass matrices M_e, $\mathsf{M}_{\mathsf{p},\mathsf{i}}$, and $\mathsf{M}_{\mathsf{d},\mathsf{i}}$, which leads to a diagonal structure, and the matrix M_h is diagonal automatically. In the case of piecewise constant material properties only one scalar convolution weight ω_n has to be stored per time step n and per subdomain covered by a dispersive material.

In Fig. 1, we display the magnetic field component h_y for the two schemes presented in Sects. 2 and 3 for some selected time steps. As predicted, the numerical solutions cannot be distinguished by visual inspection; in our computations, the maximal difference, caused by inexact computation of the weights ω_n, was in the order of 10^{-12}, and thus much smaller than the discretization errors. We tested both, the classical CQ with $O(N^2)$ complexity and the FOCQ approach with $O(N \log N)$

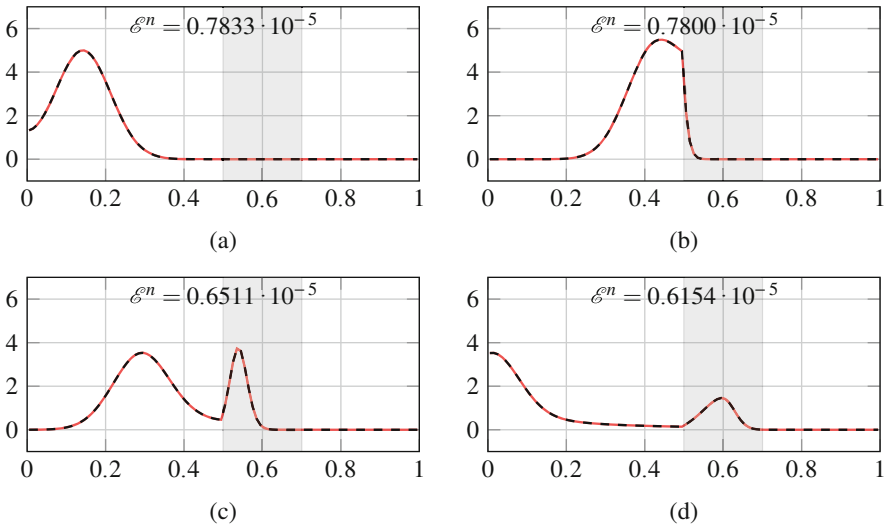


Fig. 1 Snapshots of the component h_y of the numerical solution restricted to the interval [0, 1] at different time steps. The solution of the leapfrog method (9)–(11) is drawn in red while that of the convolution-quadrature method (9)–(10) and (13) is depicted in black. The gray area indicates the location of the dispersive medium. (**a**) $t = 0.977 \cdot 10^{-9}$. (**b**) $t = 2.930 \cdot 10^{-9}$. (**c**) $t = 4.883 \cdot 10^{-9}$. (**d**) $t = 6.836 \cdot 10^{-9}$

cost; see Remark 4. Both approaches lead to almost identical results. The latter was however substantially faster, in particular for a large number N of time steps.

From the results in Fig. 1, one can also recognize the basic physical behavior: In the initial phase, the pulse propagates through air and the total energy of the system is conserved exactly. When impinging on the air-tissue interface, a part of the pulse gets reflected and the rest penetrates into the dispersive medium. Propagation in the medium is substantially slower and, moreover, energy is dissipated according to Lemma 1. We were able to reproduce this energy balance up machine precision.

5 Summary

We presented two discretization strategies for simulating Maxwell's equation in dispersive media, which were proven to be equivalent for certain classes of problems and to comply with the underlying energy–dissipation structure of the problem. The second scheme, which is based on a convolution quadrature approach, is independent of the number of internal states or relaxation times, and can be applied to dispersive media with rather general memory kernels. This might become particularly useful also in the context of uncertainty quantification.

Acknowledgments The authors are grateful for support by the German Research Foundation (DFG) via grants TRR 146 project C03, TRR 154, project C04, and Eg-331/1-1 and through grant Center for Computational Engineering at TU Darmstadt.

References

1. V.A. Bokil, N.L. Gibson, Convergence analysis of Yee schemes for Maxwell's equations in Debye and Lorentz dispersive media. Int. J. Numer. Anal. Model. **11**, 657–687 (2014)
2. J. Clegg, M. Robinson, A genetic algorithm for optimizing multi-pole Debye models of tissue dielectric properties. Phys. Med. Biol. **57**, 6227–43 (2012)
3. G. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations* (Springer, Berlin, 2002)
4. P. Debye, *Polar Molecules* (Chemical Catalogue Company, New York, 1929)
5. S. Gabriel, R.W. Lau, C. Gabriel, The dielectric properties of biological tissues: III. Parametric models for the dielectric spectrum of tissues. Phys. Med. Biol. **41**, 2271–93 (1996)
6. O.P. Gandhi, B.-Q. Gao, J.-Y. Chen, A frequency-dependent finite-difference time-domain formulation for general dispersive media. IEEE Trans. Microw. Theory Tech. **41**44–4, 658–665 (1993)
7. H. Egger, B. Radu, A mass-lumped mixed finite element method for Maxwell's equations (2018). arXiv:1810.06243. to appear in Proceedings of SCEE 2018
8. H. Egger, K. Schmidt, V. Shashkov, Multistep and Runge–Kutta convolution quadrature methods for coupled dynamical systems. J. Comput. Appl. Math. **387**, 112618 (2020)
9. D. Jiao, J.-M. Jin, Time-domain finite-element modeling of dispersive media. IEEE Microw. Wireless Components Lett. **11**, 220–222 (2001)

10. M.J. Jenkinson, J.W. Banks, High-order accurate FDTD schemes for dispersive Maxwell's equations in second-order form using recursive convolutions. J. Comput. Appl. Math. **336**, 192–218 (2018)
11. P. Joly, Variational methods for time-dependent wave propagation problems, in *Topics in Computational Wave Propagation*. LNCSE, vol. 31 (Springer, Berlin, 2003), pp. 201–264
12. S. Lanteri, C. Scheid, Convergence of a discontinuous Galerkin scheme for the mixed time-domain Maxwell's equations in dispersive media. IMANUM **33**, 432–459 (2012)
13. J. Li, Error analysis of finite element methods for 3-D Maxwell's equations in dispersive media. J. Comput. Appl. Math. **188**, 107–120 (2006)
14. C. Lubich, Convolution quadrature and discretized operational calculus. I. Numer. Math. **52**, 129–145 (1988)
15. C. Lubich, Convolution quadrature and discretized operational calculus. II. Numer. Math. **52**(4), 413–425 (1988)
16. C. Lubich, A. Ostermann, Runge-Kutta methods for parabolic equations and convolution quadrature. Math. Comp. **60**(201), 105–131 (1993)
17. R. Luebbers, F.P. Hunbserger, K.S. Kunz, R.B. Standler, M. Schneider, A frequency-dependent finite-difference time-domain formulation for dispersive materials. IEEE Trans. Electromag. Compat. **32**, 222–227 (1990)
18. J. Roychowdhury, Reduced-order modeling of time-varying systems. IEEE Trans. Circuits Syst. II **46**, 1273–1288 (1999)
19. A. Schädle, M. López-Fernandez, C. Lubich, Fast and oblivious convolution quadrature. SIAM J. Sci. Comput. **28**, 421–438 (2006)
20. S. Shaw, Finite element approximation of Maxwell's equations with Debye memory. Adv. Numer. Anal. **2010**, 923832 (2010)
21. W.A. Strauss, *Partial Differential Equations. An Introduction* (Wiley, New York, 1992)

On the Stability of Harmonic Coupling Methods with Application to Electric Machines

H. Egger, M. Harutyunyan, M. Merkel, and S. Schöps

Abstract Harmonic stator-rotor coupling offers a promising approach for the interconnection of rotating subsystems in the simulation of electric machines. This paper studies the stability of discretization schemes based on harmonic coupling in the framework of mortar methods for Poisson-like problems. A general criterion is derived that allows to ensure the relevant inf-sup stability condition for a variety of specific discretization approaches, including finite-element methods and isogeometric analysis with harmonic mortar coupling. The validity and sharpness of the theoretical results is demonstrated by numerical tests.

1 Introduction

Electric drives naturally consist of different subdomains, i.e. the stator and rotor, which move relative to each other. The time-varying geometry and nonlinearities caused by saturation effects formally require a time-domain analysis, which is often realized by solving a sequence of quasi-stationary problems at different working points. Several strategies have been proposed for the simulation of the corresponding equations of magnetostatics and, in particular, for the coupling of the fields across the air gap between stator and rotor. As it is common practice, see e.g. [13, 14, 17], we consider a two dimensional regime, in which the unknown fields are described by the axial component of the magnetic vector potential. The governing system then consists of two Poisson-like problems for the stator and the rotor, which can be coupled via Lagrange multipliers. Such domain decompositions of mortar

H. Egger (✉)
Numerical analysis and Scientific Computing, Technische Universität Darmstadt, Darmstadt, Germany
e-mail: herbert.egger@tu-darmstadt.de

M. Harutyunyan · M. Merkel · S. Schöps
Computational Electromagnetics, Technische Universität Darmstadt, Darmstadt, Germany
e-mail: mane.harutyunyan@tu-darmstadt.de; melina.merkel@tu-darmstadt.de; sebastian.schoeps@tu-darmstadt.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_12

methods, which couple subdomains via Lagrange multipliers, have been investigated intensively in the literature [2, 3, 5, 20]; see [8, 13] for results concerning electric machines. It is well-known that a careful choice of approximation spaces is required to obtain stable discretization schemes for underlying saddlepoint problems [6, 18]; appropriate stabilization [15] could be used as an alternative approach.

In this paper, we investigate the stability of mortar discretizations using trigonometric functions as Lagrange multipliers, called harmonic coupling methods in [4, 13]. We discuss in detail the discrete inf-sup condition which is necessary and sufficient to guarantee the stability of such approximations. We provide a simple criterion for the maximal number of harmonics used as Lagrange multipliers depending on the mesh size and polynomial degree of the subdomain discretizations which guarantees the stability of the scheme. Our analysis applies to the harmonic coupling of various discretization methods, e.g. obtained by isogeometric analysis (IGA) [4, 7, 16], and can in principle be extended to other Lagrange multiplier spaces.

The remainder of this note is organized as follows: In Sect. 2, we introduce the model problem to be considered and we summarize some well-known results about its analysis and discretization. In Sect. 3, we then turn to the harmonic stator-rotor coupling, and we state and prove our main results. Section 4 is concerned with numerical tests, in which we demonstrate the validity of our stability criterion for low and high order discretizations based on IGA.

2 Model Problem

We consider a typical geometric setup that consists of two subdomains Ω_1, Ω_2 representing, respectively, the stator and rotor, separated by a small air gap which contains the interface $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$; see Fig. 1. Let $\Sigma_\ell = \partial\Omega_\ell \setminus \Gamma$, $\ell = 1, 2$, be the remaining parts of the subdomain boundaries and $f_\ell = f|_{\Omega_\ell}$ denote the restriction of a function f defined on $\Omega_1 \cup \Omega_2$ to the subdomain Ω_ℓ. We then consider the following elliptic interface problem: Inside the two subdomains, we require

$$-\mathrm{div}(\nu_\ell \nabla u_\ell) = j_\ell, \qquad \text{in } \Omega_\ell, \tag{1}$$

$$u_\ell = 0, \qquad \text{on } \Sigma_\ell, \tag{2}$$

where u denotes the z-component of the magnetic vector potential, ν the magnetic reluctivity, and $j = j_s + \mathrm{div} m^\perp$ a generalized current density with j_s denoting the z-component of the source currents and $m^\perp = (m_y, -m_x)$ the rotated magnetization vector of the permanent magnet. The corresponding in-plane components of the magnetic flux density and field strength are given by $b = (\partial_y u, -\partial_x u) = \nabla^\perp u$

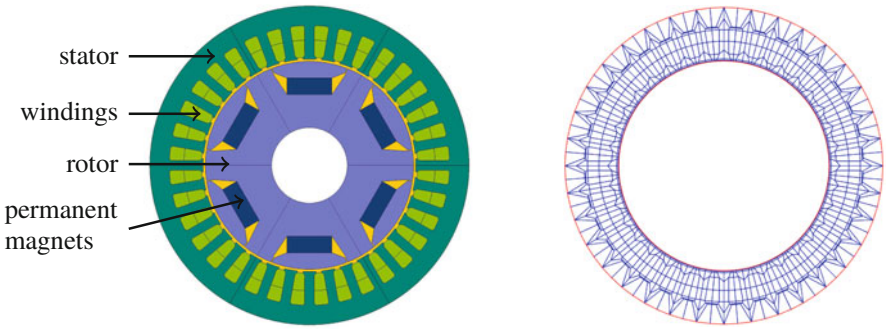


Fig. 1 Typical structure of a 6-pole permanent magnet synchronous machine (left) and the coarsest mesh of the stator domain as used in our numerical tests (right)

and $h = \nu b$, respectively. The coupling of the fields across the interface Γ is accomplished by the conditions

$$u_1 = u_2, \qquad \text{on } \Gamma, \tag{3}$$

$$n \cdot (\nu_1 \nabla u_1) = n \cdot (\nu_2 \nabla u_2), \qquad \text{on } \Gamma, \tag{4}$$

which correspond to the conditions for the normal continuity of b and the tangential continuity of h, respectively; see e.g. [4, 13]. Here $n = n_2$ is the unit normal vector at Γ pointing from Ω_2 to Ω_1. In the context of electric machines, it is natural to assume that Ω_ℓ are bounded domains with piecewise smooth boundaries Σ_ℓ and Γ, having non-zero measure. Moreover, we can assume that ν is bounded from above and below by positive constants $\underline{\nu}, \overline{\nu}$, i.e. $\underline{\nu} \le \nu(x) \le \overline{\nu}$ for all $x \in \Omega_1 \cup \Omega_2$.

The weak formulation of the interface problem (1)–(4) then reads as follows: Find $u \in V = \{v \in H^1(\Omega_1 \cup \Omega_2) : v|_{\Sigma_\ell} = 0\}$ and $\lambda \in M = H^{-1/2}(\Gamma)$ such that

$$(\nu \nabla u, \nabla v)_{\Omega_1 \cup \Omega_2} + \langle \lambda, [v] \rangle_\Gamma = \langle j, v \rangle_{\Omega_1 \cup \Omega_2} \qquad \forall v \in V, \tag{5}$$

$$\langle [u], \mu \rangle_\Gamma = 0 \qquad \forall \mu \in M. \tag{6}$$

Here $(a, b)_{\Omega_1 \cup \Omega_2} = \int_{\Omega_1} a \cdot b \, dx + \int_{\Omega_2} a \cdot b \, dx$ is the usual scalar product of functions $a, b \in L^2(\Omega_1 \cup \Omega_2)$, while $\langle a, b \rangle_{\Omega_1 \cup \Omega_2}$, $\langle a, b \rangle_\Gamma$ are the duality products on $V \times V'$ and $M \times M'$, respectively, with V', M' denoting the dual spaces of V and M. Furthermore, $H^1(\Omega_1 \cup \Omega_2)$ denotes the space of piecewise smooth functions v with restrictions $v_\ell = v|_{\Omega_\ell} \in H^1(\Omega_\ell)$ for $\ell = 1, 2$ and $[v] = v_1 - v_2$ denotes the jump of such functions across the interface Γ, and $H^{-1/2}(\Gamma)$ can be obtained by mapping the space of 2π-periodic functions $u(\phi) = a_0 + \sum_{n \ge 1} a_n \cos(n\phi) + b_n \sin(n\phi)$ whose Fourier coefficients satisfy $a_0^2 + \sum_{n \ge 1} (1 + n^2)^{-1/2} (a_n^2 + b_n^2) < \infty$.

Lemma 1 *For any $j_s \in L^2(\Omega)$ and $m \in L^2(\Omega)^2$, the variational problem (5)–(6) with $j = j_s + \mathrm{div} m^\perp$ has a unique solution $(u, \lambda) \in V \times M$ and there holds*

$$\|u\|_{H^1(\Omega_1\cup\Omega_2)} + \|\lambda\|_{H^{-1/2}(\Gamma)} \le C\big(\|j_s\|_{L^2(\Omega_1\cup\Omega_2)} + \|m\|_{L^2(\Omega_1\cup\Omega_2)}\big)$$

where the constant C does not depend on u, λ, j_s or m. Moreover, u is the unique weak solution of (1)–(4) and $\lambda = n \cdot (\nu \nabla u_\ell)$ the associated tangential component of the magnetic field strength h at the interface.

Remark 1 The result is well-known and a similar assertion can already be found in the work of Babuska [1]. Using Brezzi's theory for saddlepoint problems [6], the essential ingredient turns out to be the inf-sup stability condition

$$\inf_{\mu\in M} \sup_{v\in V} \frac{\langle \mu, [v]\rangle_\Gamma}{\|\mu\|_{H^{-1/2}(\Gamma)} \|v\|_{H^1(\Omega_1\cup\Omega_2)}} \ge \beta > 0. \tag{7}$$

Following [18], condition (7) can be proven as follows: Let $z_1 \in H^1(\Omega_1)$ be the weak solution of the mixed boundary value problem

$$-\Delta z_1 = 0 \quad \text{in } \Omega_1 \quad \text{with} \quad z_1 = 0 \quad \text{on } \Sigma_1 \quad \text{and} \quad \partial_n z_1 = \mu \quad \text{on } \Gamma. \tag{8}$$

Then by standard arguments for elliptic problems [1, 2], one can show that

$$\|z_1\|_{H^1(\Omega_1)} \le c_2 \|\mu\|_{H^{-1/2}(\Gamma)} \quad \text{and} \quad \langle \mu, z_1\rangle_\Gamma \ge c_1 \|\mu\|^2_{H^{-1/2}(\Gamma)},$$

with positive constants c_1, c_2 only depending on Ω_1, Σ_1, and Γ.

Now define $z \in H^1(\Omega_1 \cup \Omega_2)$ by $z = z_1$ on Ω_1 and $z = 0$ on Ω_2. Then

$$\langle \mu, [z]\rangle_\Gamma = \langle \mu, z_1\rangle_\Gamma \ge c_1 \|\mu\|^2_{H^{-1/2}(\Gamma)} \ge \frac{c_1}{c_2} \|\mu\|_{H^{-1/2}(\Gamma)} \|z_1\|_{H^1(\Omega_1)}.$$

The result now follows with $\beta = \frac{c_1}{c_2}$ by noting that $\|z\|_{H^1(\Omega_1\cup\Omega_2)} = \|z_1\|_{H^1(\Omega_1)}$. □

Remark 2 The solution z_1 of the auxiliary problem (8) suffices to prove the inf-sup stability conditions but does not yield the supremum in (7). The simplicity of the auxiliary problem however allows us to calculate z_1 analytically later on and thus to obtain a computable bound $\beta > 0$ depending only on the geometric setting.

Discretization As a next step, we now consider Galerkin approximations of the weak formulation (5)–(6): Find $u_h \in V_h \subset V$ and $\lambda_N \in M_N \subset M$ such that

$$(\nu \nabla u_h, \nabla v_h)_{\Omega_1\cup\Omega_2} + \langle \lambda_N, [v_h]\rangle_\Gamma = \langle j, v_h\rangle_{\Omega_1\cup\Omega_2} \qquad \forall v_h \in V_h, \tag{9}$$

$$\langle [u_h], \mu_N\rangle_\Gamma = 0 \qquad \forall \mu_N \in M_N. \tag{10}$$

Following the usual convention, we assume that V_h and M_N are finite dimensional.

Lemma 2 *Let the conditions of Lemma 1 be valid and assume that*

$$\inf_{\mu_N \in M_N} \sup_{v_h \in V_h} \frac{\langle \mu_N, [v_h] \rangle_\Gamma}{\|\mu_N\|_{H^{-1/2}(\Gamma)} \|v_h\|_{H^1(\Omega_1 \cup \Omega_2)}} \geq \beta' > 0 \tag{11}$$

Then problem (9)–(10) has a unique solution $u_h \in V_h$, $\lambda_N \in M_N$. Furthermore

$$\begin{aligned} &\|u - u_h\|_{H^1(\Omega_1 \cup \Omega_2)} + \|\lambda - \lambda_N\|_{H^{-1/2}(\Gamma)} \\ &\quad \leq C \Big(\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega_1 \cup \Omega_2)} + \inf_{\mu_N \in M_N} \|\lambda - \mu_N\|_{H^{-1/2}(\Gamma)} \Big) \end{aligned} \tag{12}$$

where (u, λ) is the solution of the continuous variational problem (5)–(6) and the constant C depends only on β' in (11), the bounds for ν, and the geometry.

Remark 3 All conditions required for the proof of the corresponding result on the continuous level, except the inf-sup stability condition, are inherited by the Galerkin approximation. The existence of a unique solution can thus again be deduced from Brezzi's saddlepoint theory [6]. The error estimate (12) follows from Galerkin orthogonality and standard arguments; we refer to [5, 6] for details. Hence any choice of approximation spaces V_h, M_N that allows to prove the discrete inf-sup stability condition (11) will lead to a well-posed discrete problem.

3 Harmonic Stator-Rotor Coupling

We now consider a particular class of Galerkin approximations (9)–(10) in which V_h is constructed by piecewise polynomials, while the Lagrange multiplier space M_N is defined by trigonometric polynomials. Our analysis in particular also covers the harmonic-coupling of the methods considered in [4, 13].

Using polar coordinates, the computational domain $\Omega = \Omega_1 \cup \Omega_2$ can be represented as the image of a rectangle $\widehat{\Omega}$ under a mapping $F : \widehat{\Omega} \to \Omega$; see Fig. 2. Now let $\widehat{T}_h$ denote a shape-regular partition of $\widehat{\Omega}_1 \cup \widehat{\Omega}_2$ into triangles and /or rectangles of size h. The meshes of the two sub-domains are assumed to be geometrically conforming, but they may be non-matching across the interface. We denote by $P_k(\widehat{T}_h)$ the space of piecewise polynomials over $\widehat{T}_h$ of degree $\leq k$ and by $\widehat{M}_N = \text{span}\{\sin(n\pi\xi), \cos(n\pi\xi) : 0 \leq n \leq N\}$ the spaces of trigonometric polynomials of degree $\leq N$. We then choose the approximation spaces V_h, M_N s.t.

$$M_N = F(\widehat{M}_N) \qquad \text{and} \qquad V_h = V_h|_{\Omega_1} \cup V_h|_{\Omega_2} \subset F(P^k(T_h)) \cap V. \tag{13}$$

By the condition $V_h = V_h|_{\Omega_1} \cup V_h|_{\Omega_2}$ we mean that discrete functions, when restricted to one of the sub-domains and extended by zero to the other still belong to the approximation space V_h. The basic assumption for the discrete inf-sup stability condition (7) of the corresponding Galerkin approximation (9)–(10) is the following.

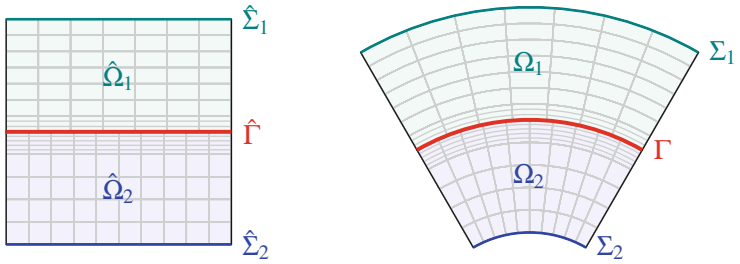


Fig. 2 Sketch of a subset of the rectangular reference domain $\hat{\Omega}$ and its mesh (left) and the physical domain $\Omega = F(\hat{\Omega})$ and mesh obtained after mapping. The boundaries on the left and right are only introduced for the illustration but not present in our application

Theorem 1 *Assume that there exists a linear operator $\Pi_h : V|_{\Omega_1} \to V_h|_{\Omega_1}$ such that*

$$\|\Pi_h v_1\|_{H^1(\Omega_1)} \le c_3 \|v_1\|_{H^1(\Omega_1)}, \tag{14}$$

$$\Pi_h v_1 = \pi_h v_1 \quad \text{on } \Gamma, \tag{15}$$

$$\|v - \pi_h v\|_{H^{-1/2}(\Gamma)} \le c_4 \frac{h}{k} \|v\|_{H^{1/2}(\Gamma)}, \tag{16}$$

where $\pi_h : L^2(\Gamma) \to V_h|_{\Omega_1 \cap \Gamma}$ denotes the L^2-projection on Γ. Then there exists a constant $0 < \epsilon < 1$, depending only on c_3, c_4 in (14)–(16), such that the discrete inf-sup condition (11) holds with $\beta' = \beta'(\epsilon)$ whenever N and h are chosen such that

$$Nh/k \le 1 - \epsilon. \tag{17}$$

Proof As an immediate consequence of the continuous inf-sup condition (7), we can find for $\mu = \pi_h \lambda_N$ a function $z \in V$ with $z = 0$ on Ω_2, such that

$$\langle \pi_h \lambda_N, [z] \rangle_\Gamma \ge \beta \|z\|_{H^1(\Omega_1 \cup \Omega_2)} \|\pi_h \lambda_N\|_{H^{-1/2}(\Gamma)}.$$

We then define $z_h = \Pi_h z_1$ on Ω_1 and $z_h = 0$ on Ω_2, and observe that

$$\langle \pi_h \lambda_N, [z_h] \rangle_\Gamma = \langle \pi_h \lambda_N, [\Pi_h z] \rangle_\Gamma = \langle \pi_h \lambda_N, \pi_h [z] \rangle_\Gamma = \langle \pi_h \lambda_N, [z] \rangle_\Gamma,$$

where we used property (15) and the orthogonality of the L^2-projection π_h. Together with the previous estimate and employing condition (14), we thus obtain

$$\langle \pi_h \lambda_N, [z_h] \rangle_\Gamma \ge \beta \|z\|_{H^1(\Omega_1 \cup \Omega_2)} \|\pi_h \lambda_N\|_{H^{-1/2}(\Gamma)} \ge \frac{\beta}{c_3} \|z_h\|_{H^1(\Omega_1 \cup \Omega_2)} \|\pi_h \lambda_N\|_{H^{-1/2}(\Gamma)}.$$

Using the triangle inequality, we can further estimate

$$\|\pi_h \lambda_N\|_{H^{-1/2}(\Gamma)} \geq \|\lambda_N\|_{H^{-1/2}(\Gamma)} - \|\lambda_N - \pi_h \lambda_N\|_{H^{-1/2}(\Gamma)},$$

and the last term can be bounded with the approximation error estimate (16) by

$$\|\lambda_N - \pi_h \lambda_N\|_{H^{-1/2}(\Gamma)} \leq c_4 \frac{h}{k} \|\lambda_N\|_{H^{1/2}(\Gamma)} \leq C' \frac{h}{k} N \|\lambda_N\|_{H^{-1/2}(\Gamma)}.$$

In the second estimate, we here used an inverse inequality for the finite dimensional Lagrange multiplier space M_N. In summary, we thus obtain

$$\langle \pi_h \lambda_N, [z_h] \rangle_\Gamma \geq \beta (1 - C' N h/k) \|z_h\|_{H^1(\Omega_1 \cup \Omega_2)} \|\lambda_N\|_{H^{-1/2}(\Gamma)},$$

from which the assertion of the theorem follows immediately.

Remark 4 The conditions of the theorem hold for a variety of discretization methods, e.g. FEM or IGA. The projection operator Π_h can here be constructed following the ideas of [11, 19] or [9] and the approximation property (16) for π_h is well-known; details will be given in a forthcoming publication. The resulting harmonic-coupling mortar methods are thus stable, if the number of degrees of freedom $n \sim k/h$ located at the interface exceeds the number of coupling modes N to some extent, cf. (17). Our main arguments may be applied to other problems and discretization strategies.

4 Numerical Results

We now illustrate the theoretical results of Theorem 1 by some numerical tests using an IGA discretization [16] as implemented in GeoPDEs [12]. We here only report about results concerning the discrete inf-sup condition; numerical results concerning the magnetic fields can be found e.g. in [4, 13]. The geometry used in our computations is depicted in Fig. 1. Following the arguments given in Remark 1 and underlying the proof of Theorem 1, we have

$$\sup_{v \in V} \frac{\langle \mu, v \rangle_\Gamma}{\|v\|_{H^1(\Omega_1 \cup \Omega_2)}} \geq \sup_{z_1 \in V_1} \frac{\langle \mu, z_1 \rangle_\Gamma}{\|z_1\|_{H^1(\Omega_1)}} \geq \beta \|\mu\|_{H^{-1/2}(\Gamma)}, \tag{18}$$

where $V = \{v \in H^1(\Omega_1 \cup \Omega_2) : v|_{\Sigma_1 \cup \Sigma_2} = 0\}$ and $V_1 = \{v \in V : v|_{\Omega_2} = 0\} \subset V$. Due to the Dirichlet boundary conditions on Σ_1, we can choose $\|v_1\|_{H^1(\Omega_1)} = \|\nabla v_1\|_{L^2(\Omega_1)}$ as the norm on V_1. One can then show that the second supremum in (18) is attained by the solution z_1 of the mixed boundary value problem (8). For our model problem, $\Omega_1 = \{x \in \mathbb{R}^2 : R_1 < |x| < R_2\}$ is a simple annulus with radii $R_1 = 0.0447$ and $R_2 = 0.0675$, and the solution of the above problem can be computed analytically in the form of a Fourier series, and we define $S_1 \mu := v_1|_\Gamma$.

Table 1 Discrete inf-sup constants obtained for n gridpoints at the interface Γ and harmonic order $N = cn$ of the Lagrange multipliers for different refinement levels ℓ and scaling parameters c

$c \setminus \ell$	1	2	3	4
1/4	0.135237	0.135556	0.135676	0.135693
1/3	0.135237	0.135556	0.135661	0.135684
3/8	0.135237	0.135536	0.135611	0.135684
1/2	3.526e–08	2.532e–08	2.401e–08	2.401e–08

Table 2 Discrete inf-sup constants for n spline degrees of freedom on the interface Γ and harmonic order $N = cn$ of the Lagrange multipliers for polynomial degree k and scaling parameter c

$c \setminus k$	2	3	4	5
1/4	0.135721	0.135723	0.135723	0.135723
1/3	0.135721	0.135722	0.135723	0.135723
3/8	0.135720	0.135723	0.135723	0.135723
1/2	3.652e–08	0	8.082e–08	1.825e–08

The largest possible constant β such that the second estimate of (18) remains true for all $\mu \in H^{-1/2}(\Gamma)$ can then be characterized by the minimal eigenvalue of

$$\langle \mu, S_1 \tilde{\mu} \rangle_{H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)} = \beta^2 (\mu, \tilde{\mu})^2_{H^{-1/2}(\Gamma)} \qquad \forall \tilde{\mu} \in H^{-1/2}(\Gamma).$$

For the problem under consideration, the solution can be computed explicitly which gives $\beta = \sqrt{R_1 \ln(R_2/R_1)} \approx 0.13573$. The discrete inf-sup constant is evaluated by numerically solving the corresponding discretized eigenvalue problem; see e.g. [10].

In the first series of tests, we utilize the lowest order approximation and consider a sequence of uniformly refined meshes. The discrete inf-sup constant is computed as outlined above. The results of these computations are depicted in Table 1. The coarsest mesh has $n = 144$ vertices at the interface Γ and is doubled in every refinement step; see Fig. 1. For $c = 1/2$, we have $\dim(M_N) = 2N + 1 = n + 1 > n$, and the discrete inf-sup stability condition is violated. The results of Table 1 thus perfectly agree with the theoretical predictions of Theorem 1. In a second sequence of tests, we study the dependence of the inf-sup constant on the polynomial degree k of the spline approximation on the mesh with refinement level 2. The corresponding results are summarized in Table 2. For the choice $c = 1/2$, the number of Lagrange multipliers $2N + 1 = n + 1 > n$ again exceeds the number of the spline degrees at the interface Γ and the discrete inf-sup stability fails. The computational results are again in perfect agreement with the theoretical predictions.

Acknowledgments This work is supported by the 'Excellence Initiative' of the German Federal and State Governments and by the Graduate School of Computational Engineering at Technische Universität Darmstadt and the grants TRR 154 project C04 and TRR 146 project C03.

References

1. I. Babuška, The finite element method with Lagrangian multipliers. Numer. Math. **20**, 179–192 (1973)
2. F. Ben Belgacem, The mortar finite element method with Lagrange multipliers. Numer. Math. **84**, 173–197 (1999)
3. C. Bernardi, Y. Maday, A.T. Patera, A new nonconforming approach to domain decomposition: the mortar element method, in *Nonlinear Partial Differential Equations and their Applications*. Pitman Research Notes in Mathematics Series, vol. 299 (1994), pp. 13–51
4. Z. Bontinck, J. Corno, S. Schöps, H. De Gersem, Isogeometric analysis and harmonic stator-rotor coupling for simulating electric machines. Comput. Meth. Appl. Mech. Engrg. **334**, 40–55 (2018)
5. D. Braess, W. Dahmen, C. Wieners, A multigrid algorithm for the mortar finite element method. SIAM J. Numer. Anal. **37**, 48–69 (1999)
6. F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. RAIRO Anal. Numer. **8**, 129–151 (1974)
7. E. Brivadis, A. Buffa, B. Wohlmuth, L. Wunderlich, Isogeometric mortar methods. Comput. Methods Appl. Mech. Engrg. **284**, 292–319 (2015)
8. A. Buffa, Y. Maday, F. Rapetti, A sliding mesh-mortar method for a two dimensional currents model of electric engines. ESAIM Math. Model. Numer. Anal. **35**, 191–228 (2001)
9. A. Buffa, E.M. Garau, C. Giannelli, G. Sangalli, On quasi-interpolation operators in spline spaces, in *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*. Lecture Notes in Computational Science and Engineering, vol. 114 (Springer, Berlin, 2016), pp. 73–91
10. D. Chapelle, K.J. Bathe, The inf-sup test. Comput. Struct. **47**(4/5), 537–545 (1993)
11. P. Clément, Approximation by finite element functions using local regularization. RAIRO **9**, 77–84 (1975)
12. C. de Falco, A. Reali, R. Vázquez, GeoPDEs: a research tool for Isogeometric Analysis of PDEs. Adv. Eng. Softw. **42**, 1020–1034 (2011)
13. H. De Gersem, T. Weiland, Harmonic weighting functions at the sliding interface of a finite-element machine model incorporating angular displacement. IEEE Trans. Magn. **40**, 545–548 (2004)
14. G. Gyselinck, L. Vandevelde, P. Dular, C. Geuzaine, A general method for the frequency domain FE modeling of rotating electromagnetic devices. IEEE Trans. Magn. **39**, 1147–1150 (2003)
15. P. Hansbo, C. Lovadina, I. Perugia, G. Sangalli, A Lagrange multiplier method for the finite element solution of elliptic interface problems using non-matching meshes. Numer. Math. **100**, 91–115 (2005)
16. T.J.R. Hughes J.A. Cottrell, T. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput. Meth. Appl. Mech. Eng. **194**, 4135–4195 (2005)
17. E. Lange, F. Henrotte, K. Hameyer, A variational formulation for nonconforming sliding interfaces in finite element analysis of electric machines. IEEE Trans. Magn. **46**, 2755–2758 (2010)
18. P.A. Raviart, J.M. Thomas, Primal hybrid finite element methods of 2nd order elliptic equations. Math. Comp. **31**, 391–413 (1977)
19. L.R. Scott, S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comp. **54**, 483–493 (1990)
20. B.I. Wohlmuth, A mortar finite element method using dual spaces for the Lagrange multiplier. SIAM J. Numer. Anal. **38**, 989–1012 (2000)

Multifidelity Uncertainty Quantification for Optical Structures

Niklas Georg, Christian Lehmann, Ulrich Römer, and Rolf Schuhmann

Abstract This work addresses uncertainty quantification for optical structures. We decouple the propagation of uncertainties by combining local surrogate models with a scattering matrix approach, which is then embedded into a multifidelity Monte Carlo framework. The so obtained multifidelity method provides highly efficient estimators of statistical quantities jointly using different models of different fidelity and can handle many uncertain input parameters as well as large uncertainties. We address quasi-periodic optical structures and propose the efficient construction of low-fidelity models by polynomial surrogate modeling applied to unit cells. We recall the main notions of the multifidelity algorithm and illustrate it with a split ring resonator array simulation, serving as a benchmark for the study of optical structures. The numerical tests show speedups by orders of magnitude with respect to the standard Monte Carlo method.

1 Introduction

Manufacturing on the nanometer-scale exhibits strong variability in the finally built structures which should be addressed in a simulation based design approach. The field of uncertainty quantification provides suitable tools to model and quantify

N. Georg (✉)
Institut für Dynamik und Schwingungen, Technische Universität Braunschweig, Braunschweig, Germany

Centre for Computational Engineering, Technische Universität Darmstadt, Darmstadt, Germany
e-mail: n.georg@tu-braunschweig.de

C. Lehmann · R. Schuhmann
Theoretische Elektrotechnik, Technische Universität Berlin, Berlin, Germany
e-mail: lehmann@tet.tu-berlin.de; rolf.schuhmann@tu-berlin.de

U. Römer
Institut für Dynamik und Schwingungen, Technische Universität Braunschweig, Braunschweig, Germany
e-mail: u.roemer@tu-braunschweig.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_13

uncertainties in the geometrical and material constitutive parameters. In this contribution, we focus on the benchmark example of a split ring resonator (SRR) array with random input data. In particular, we model uncertainties in the SRR geometry with random variables and quantify the implied variation in the frequency response of the system. Such quasi-periodic optical structures may feature a large number of uncertain parameters which makes the application of many methods, such as standard or even sparse Polynomial Chaos difficult, see [1] for instance. We present a remedy by applying spectral polynomial expansions on the unit cell level in the framework of the Scattering Matrix Approach (SMA) [2], which yields a significant reduction of the computational effort. Since the coupled surrogate model may be biased, we use a Multifidelity Monte Carlo (MFMC) method [3], which combines different numerical models with different fidelity, to obtain efficient statistical estimators. In particular, through limited recurrences to a high-fidelity simulation of the entire structure, the MFMC method then corrects for possible approximation errors in the low-fidelity data.

2 Decoupled Uncertainty Propagation with Scattering Matrices

Our benchmark application is a simplified model of an array of coupled SRRs, motivated by the research on optical metamaterials [4, 5]. It consists of a periodic, but finite-size array of metallic SRR structures on a nanometer-scale, each of which can be interpreted as a realization of a resonance circuit, with the ring and the small gap acting as inductance and capacitance, respectively. More details on the geometry and setup will be given in Sect. 4. Due to the unavoidable tolerances in manufacturing of such small structures the geometric properties of each SRR will slightly vary, and the periodicity of the array of coupled resonator will not be perfect (see Figs. 54 and 57 in [5] for an illustration). Thus, we introduce a parameter vector $\mathbf{y}_{\text{cell},j} \in \Xi \subset \mathbb{R}^P$, which models variations in the geometry or material of the structure in cell j. The full input vector is then given as $\mathbf{y} = (\mathbf{y}_{\text{cell},1}^T, \ldots, \mathbf{y}_{\text{cell},N}^T)^T \subset \mathbb{R}^{N \cdot P}$, and all results of the forward model depend on this input vector.

The structure is excited by a plane wave and the reflection and transmission coefficients are evaluated. Translated into the language of dispersion analysis, the array is expected to feature a number of bandgaps, i.e. intervals on the frequency (or wavelength) axis where no transmission through the structure is possible. Both the finite size of the arrays (in our case up to seven unit cells) and the parameter variation in each SRR will have some influence on the corresponding limit frequencies.

The electromagnetic treatment of this application example requires the solution of the wave equation with an appropriate excitation at the ports. From the field solutions the amplitudes a_i and b_i of properly normalized incoming and outgoing waves are determined. They are coupled by the scattering matrix $\mathbf{S}(j\omega)$,

$$(\ldots b_i(j\omega)\ldots)^T = \mathbf{S}(j\omega)\ (\ldots a_i(j\omega)\ldots)^T,$$

with $r(j\omega) = S_{11}(j\omega)$ the reflection coefficient at the input port. Note that we omit the frequency dependency of $\mathbf{S}$ in the following to enhance the readability.

For the discretization we apply the efficient Finite Integration Technique (FIT) time-domain algorithm [6]. It relies on a three-dimensional Cartesian mesh and allows calculating broadband results with single transient simulation runs (using Discrete Fourier Transform on the signals). The calculation of scattering parameters proceeds in two steps: First, the two-dimensional eigenvalue problem of the port apertures is analyzed to obtain the field patterns and cutoff-frequencies of the so-called waveguide modes. Note that a lossfree model is considered here, and the array of SRRs is transversally terminated by perfect electric and magnetic boundary conditions. Second, these mode patterns and their well-known orthogonality properties are used to both excite the three-dimensional structure and to extract the amplitudes of the out-going waves at the ports. From one simulation run, one column of the scattering matrix can be obtained. For further details on the FIT we refer to the literature.

A technique to reduce the computational cost in the analysis of periodic structures is to decompose the SRR array into its single unit cells and to calculate separate scattering matrices $\mathbf{S}^{(i)}$ for each of them. The final concatenation of these single-cell results can be accomplished by switching to the transfer matrices $\mathbf{T}^{(i)}$ which map the wave amplitudes of the right hand side of each cell to the left hand side (rather than from input to output quantities as with $\mathbf{S}$). For a system with 2 ports:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \mathbf{S} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad \leftrightarrow \quad \begin{pmatrix} b_1 \\ a_1 \end{pmatrix} = \mathbf{T} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \quad \text{with } \mathbf{T} = \begin{pmatrix} S_{12} - S_{11}S_{21}^{-1}S_{22} & S_{11}S_{21}^{-1} \\ -S_{21}^{-1}S_{22} & S_{21}^{-1} \end{pmatrix}.$$

Extended formulas for larger $\mathbf{S}$, $\mathbf{T}$, which take several port-modes into account, can easily be derived. Using transfer matrices, the total system behavior of N cells is simply given by a matrix multiplication $\mathbf{T} = \mathbf{T}^{(1)} \cdot \ldots \cdot \mathbf{T}^{(N)}$. This approach has been used previously in [2, 7] and is referred to as SMA.

This procedure has the intrinsic weakness that the coupling between the unit cells is not governed by a single waveguide mode alone, but an unknown number of higher modes may contribute. Of course, the coupling of modes at frequencies below their cutoff-frequency decreases rapidly with increasing spatial distance of the single SRRs. However, especially if there are resonances within the frequency range of interest (which clearly is the case for the SRRs as one of their working principles), this systematic error may become significant. In theory an extension of the SMA to an arbitrary number of coupling modes is straight-forward. However, the required number (and/or selection) of modes is sometimes hard to estimate a-priori, and the calculation of the extended transfer matrix increases the computational cost. Our approach removes any possible systematic error introduced in the coupling, by treating the SMA-based predictions as low-fidelity data and by correcting them with a couple of time domain solutions of the entire structure.

Non-intrusive Uncertainty Quantification (UQ) usually requires the repeated evaluation of the scattering matrices $\mathbf{S}(\mathbf{y})$ for different values of the inputs $\mathbf{y}$. Even with SMA the computational cost to evaluate a large number of sample points using the FIT might become prohibitive. Hence, we propose to construct a surrogate model for a unit cell of the periodic structure. In particular, we use a spectral collocation method, i.e. an approximation

$$\mathbf{S}^{(j)}(\mathbf{y}_{\text{cell},j}) \approx \mathbf{S}_{\text{uc};C}(\mathbf{y}_{\text{cell},j}) := \sum_{i=1}^{C} \mathbf{S}^{(j)}\big(\mathbf{y}_{\text{cell},j}^{(i)}\big)\Psi_i(\mathbf{y}_{\text{cell},j}) \tag{1}$$

where uc is short for unit cell and $j = 1, \ldots, N$ refers to an arbitrary unit cell of the structure. Also, $\{\mathbf{y}_{\text{cell},j}^{(i)}\}_{i=1}^{C} \subset \Xi$ denotes a set of collocation points, e.g. Chebyshev nodes, and Ψ_i denote the corresponding barycentric Lagrange polynomials. We emphasize that the same surrogate model is employed for all cells. It can be straightforwardly employed to obtain a surrogate of the full structure based on the SMA as (after transformation into $\mathbf{T}$ matrices)

$$\mathbf{T}(\mathbf{y}) \approx \mathbf{T}_C(\mathbf{y}) := \mathbf{T}_{\text{uc};C}(\mathbf{y}_{\text{cell},1}) \cdot \ldots \cdot \mathbf{T}_{\text{uc};C}(\mathbf{y}_{\text{cell},N}). \tag{2}$$

We also emphasize that (2) can be evaluated with negligible computational cost. In order to highlight the efficiency of the proposed combination of SMA and spectral surrogates for the unit cell, we give a few comments on the alternative approach, i.e. spectral approximation of the full structure. Due to spectral convergence properties, global polynomial approximations can be highly efficient, even up to a moderately large number of parameters (e.g., up to 10–20) using adaptive sparse approximations, see e.g. [1]. However, these methods still suffer from the so-called curse-of-dimensionality, i.e. the rapid growth of computational cost w.r.t. the number of parameters. As the full structure has a significant larger number of parameters, i.e. by a factor of N, this would quickly result in a very large number of simulation runs. Additionally, the computational cost for each model evaluation would also be significantly larger, when the full structure is considered instead of a single unit cell.

3 Multifidelity Monte Carlo

MFMC generalizes the multilevel Monte Carlo approach, which was recently used in [8] for a high-frequency application. MFMC simulation combines low-fidelity models of different kinds, without quantified model errors, into an efficient sampling framework. By sampling the high-fidelity model at least one time, the MFMC approach provides an unbiased estimator. Moreover, a low variance and hence, a low root-mean-square error, is realized through optimal model management and the resulting estimator is typically much more efficient than the standard Monte Carlo

(MC) estimator. The MFMC methodology was introduced in a series of papers [3, 9] and is now well-established. Hence, in the following we limit ourselves to the key aspects and refer to the literature for a more complete introduction into the field.

We adopt a probabilistic approach to represent uncertainty, where $\mathbf{y}$ represents a realization of a random vector $\mathbf{Y}$. Let $g(\mathbf{Y})$ denote an output quantity derived from the simulated frequency response. MC simulation is then based on a sample $\{\mathbf{Y}_i, g(\mathbf{Y}_i)\}_{i=1}^K$, which can be used to estimate for instance the mean value of the model output. The mean value approximation and its mean-square error read

$$\mathbb{E}[g(\mathbf{Y})] \approx \hat{g}_K := \frac{1}{K}\sum_{i=1}^{K} g(\mathbf{Y}_i), \qquad \mathbb{E}[|\mathbb{E}[g(\mathbf{Y})] - \hat{g}_K|^2] = \frac{\mathbb{V}[g(\mathbf{Y})]}{K}. \tag{3}$$

Following [9], we consider a model family $\{g^{(i)}\}_{i=1}^M$, where $g^{(1)}$ represents the high-fidelity model, and $g^{(i)}$ for $i \geq 2$ represent low-fidelity models, obtained for instance by SMA in combination with surrogate modeling. The MFMC estimator samples all models and combines the results into a single estimator as

$$\mathbb{E}[g] \approx \hat{g}_{\mathrm{MFMC}} = \hat{g}^{(1)}_{K^{(1)}} + \sum_{i=2}^{M} \alpha_i \left(\hat{g}^{(i)}_{K^{(i)}} - \hat{g}^{(i)}_{K^{(i-1)}}\right),$$

where $\hat{g}^{(i)}_{K^{(i)}}$ denotes the standard MC estimator based on the sample $\{\mathbf{Y}_j, g^{(i)}(\mathbf{Y}_j)\}_{j=1}^{K^{(i)}}$ and $0 < K^{(1)} \leq K^{(2)} \leq \ldots \leq K^{(M)}$.

In place of low-fidelity error control, the model management of MFMC employs the Pearson correlation coefficient $\rho_{1,i}$ between the high-fidelity model $g^{(1)}$ and the low-fidelity model $g^{(i)}$. In particular, low-fidelity models with a high $\rho_{1,j}$ and a low computational cost w_i are sampled extensively. For a given computational budget $\mathscr{B}$, MFMC minimizes the mean-square error by appropriately choosing α_i, $K^{(i)}$, see [9] for details. With $\sigma_i = \mathbb{V}[g^{(i)}(\mathbf{Y})]^{1/2}$, the resulting estimator is unbiased with a mean-square-error of

$$\mathbb{E}[|\hat{g}_{\mathrm{MFMC}} - \mathbb{E}[g(\mathbf{Y})]|^2] = \frac{\sigma_1^2}{K^{(1)}} + \sum_{i=2}^{M}\left(\frac{1}{K^{(i-1)}} - \frac{1}{K^{(i)}}\right)(\alpha_i^2\sigma_i^2 - 2\alpha_i\rho_{1,i}\sigma_1\sigma_i). \tag{4}$$

4 Numerical Examples

We apply the UQ methods presented in the previous section to the benchmark problem of an SRR array introduced in the beginning of Sect. 2. First, we give some details on the considered numerical models, before investigating the performance of the proposed UQ methodology.

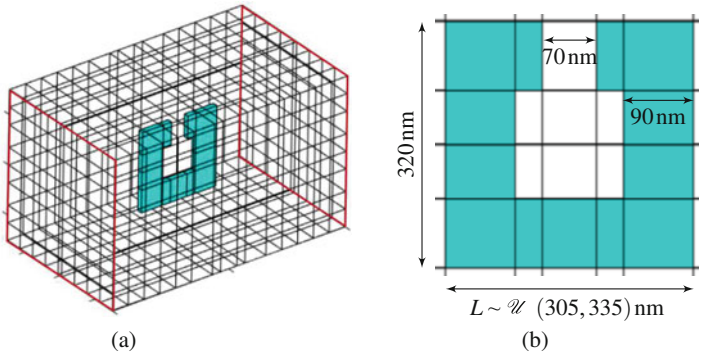


Fig. 1 Numerical model of SRR array. Depicted is only one cell out of seven. (**a**) Unit cell of size 1 µm × 0.6 µm × 0.6 µm. Red boundaries indicate the ports. (**b**) Geometry specification. Thickness: 20 nm. Uncertain longitudinal length L of SRR element

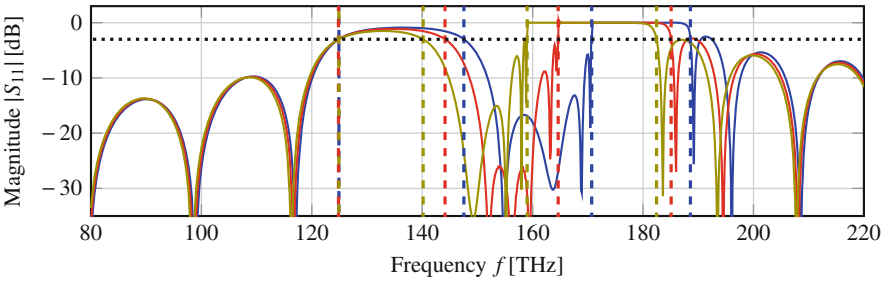


Fig. 2 Broadband scattering parameter for different realizations of SRR array. Dashed vertical lines indicate respective limit frequencies of considered bandgaps. Dotted line refers to −3 dB line

We consider a SRR array with $N = 7$ cells. The employed Cartesian grid as well as the geometric dimensions (taken from [4], except for the enlarged cell size) are presented in Fig. 1, where we consider an uncertain longitudinal length $L^{(j)}$ of each SRR element in the range of 320 nm ± 15 nm. Hence, the random vector $\mathbf{Y}$ is given as $(L^{(1)}, \ldots, L^{(N)})^T$, where $L^{(j)}$, $j = 1, \ldots, N$ are assumed to be independent and identically uniformly distributed. Figure 2 presents a broadband scattering parameter, in particular the fundamental reflection coefficient $|S_{11}|$, for different realizations of the structure. Two bandgaps can be observed, which can be defined by their limit frequencies, where the scattering parameter drops below −3 dB. The corresponding bandwidths b_i and center frequencies $f_{c,i}$, where $i \in \{1, 2\}$ refers to the first or second bandgap, can be computed from S_{11} in a post-processing step. For brevity, we restrict ourselves to the computation of the mean value of the center frequencies $\mathbb{E}[f_{c,i}]$ in the following. However, very similar findings hold for the bandwidths b_i as well. We further note that for some parameter sample points some additional resonances within the second bandgap appear which are due to the slightly detuned resonances in the series of SRRs. This effect is ignored in

Table 1 Employed numerical models of SRR array for MFMC study. The last two columns show the estimated correlation coefficients for both bandgaps

Symbol	Model	Cost w_i	$\rho_{1,i}$ for $f_{c,1}$	$\rho_{1,i}$ for $f_{c,2}$
$g^{(1)}$	Full model (FIT, $2 \cdot 10^5$ time-steps)	197.50 s	1.000000	1.000000
$g^{(2)}$	Full model (FIT, $2 \cdot 10^4$ time-steps)	11.25 s	0.999236	0.998035
$g^{(3)}$	SMA (FIT, 1 port-mode)	9.64 s	0.999943	0.968376
$g^{(4)}$	SMA (FIT, 2 port-modes)	115.47 s	0.999998	0.999998
$g^{(5)}$	SMA + unit cell surrogate (1 port-mode)	0.006 s	0.999943	0.967540
$g^{(6)}$	SMA + unit cell surrogate (2 port-modes)	0.026 s	0.999998	0.999886

the following evaluation of the MLMC algorithm and only the outer limits of this bandgap are considered.

An overview of the employed numerical models as well as the corresponding computational costs (measured in computation time for an in-house MATLAB implementation on a standard workstation) is given in Table 1. For the full FIT model $g^{(1)}$ we terminate the time stepping procedure if either the energy decays to -120 dB or a maximum number of $2 \cdot 10^5$ time-steps is reached. The low-fidelity model $g^{(2)}$ is obtained by restricting the maximum number of time-steps to $2 \cdot 10^4$. The low-fidelity models $g^{(3)}$ and $g^{(4)}$ are obtained by the SMA approach. For $g^{(3)}$ only the propagating fundamental TEM mode is considered, while $g^{(4)}$ additionally takes the evanescent first TM mode into account. The selection of suitable models is based on a pilot run (with a small sample) and model selection techniques, see also [3, 9].

The construction of the respective unit cell surrogate models for $g^{(5)}$ and $g^{(6)}$ in the offline-phase is based on $C = 7$ Chebyshev nodes, which are well-established non-equidistant interpolation nodes. Note that other choices are equally feasible, Gauss-Legendre nodes for instance. Surrogate modeling requires some additional computational effort, which, however, only needs to be invested once. Also, in this case, even a single model evaluation of $g^{(1)}$ requires a larger computational effort than constructing the surrogate models. Hence, we will neglect this cost here, for simplicity. We further note that the evaluation times of all models scale approximately linear w.r.t. to an eventually increased number of cells N, while the offline-cost for the surrogate models is independent of N. Accordingly, similar MFMC results, as presented in the following for $N = 7$, are also expected for SRR arrays with a different number of cells. Exemplarily, this has been confirmed for $N = 14$ numerically. However, we note that for larger models some care has to be taken regarding the concatenation within the SMA, since the multiplication of transfer matrices can become numerically unstable.

In order to evaluate the performance of the proposed methodology for the considered benchmark problem, we draw an input sample $\{\mathbf{Y}_i\}_{i=1}^{\tilde{K}}$ of size $\tilde{K} = 500$ and employ each model $g^{(j)}$ to compute the corresponding output samples $\{g^{(j)}(\mathbf{Y}_i)\}_{i=1}^{\tilde{K}}$, $j = 1, \ldots, 6$. The correlation coefficients with the high-fidelity

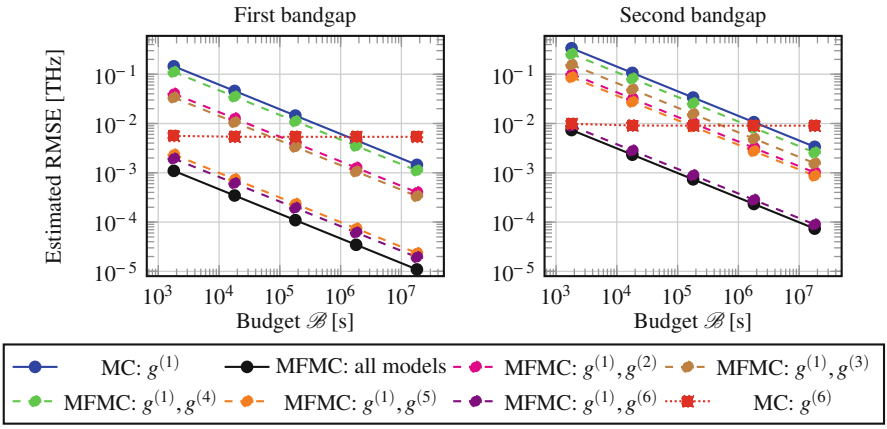


Fig. 3 Estimated RMSE for different MC and MFMC variants, see Table 1

model $g^{(1)}$ are then estimated as shown in Table 1. It can be observed that all low-fidelity models show a strong correlation with the high-fidelity model.

We employ an MFMC implementation which is based on the open-source Matlab library github.com/pehersto/mfmc, see [9]. In the following, we will compare the root-mean-square-errors (RMSEs) of MC and MFMC for given computational budgets $\mathscr{B}$, which can both be accurately estimated based on the samples $\{g^{(j)}(\mathbf{Y}_i)\}_{i=1}^{\tilde{K}}$, as explained in the following. The RMSE of standard MC on the high-fidelity model $g^{(1)}$ is obtained by (3), where K is given by $\frac{\mathscr{B}}{w_1}$ and the variance is replaced by the MC estimate for the variance using $\{g^{(1)}(\mathbf{Y}_i)\}_{i=1}^{\tilde{K}}$. This is shown in Fig. 3 in blue color. Similarly, the RMSE of MFMC can be estimated according to (4), as shown in black color in Fig. 3. We note that the proposed approach yields speedups by several orders of magnitude w.r.t. standard MC (for a fixed accuracy).

We note that the MFMC algorithm sorts out some models, as, for example, $g^{(2)}$ and $g^{(3)}$ have a smaller correlation with the high-fidelity model than the surrogate model $g^{(6)}$ but a higher computational cost. For completeness, we additionally show the convergence of MFMC using only $g^{(1)}$ and $g^{(j)}$, $j \in \{2, \ldots, 6\}$ with dashed lines in Fig. 3. As expected, in all cases this approach performs better than MC but worse than the combination of models chosen by the MFMC algorithm. It can be observed that, for both bandgaps, mainly the proposed unit cell surrogate models lead to the tremendous efficiency gains. While for the first bandgap considering only one port-mode could also be sufficient, for the second bandgap it is clearly necessary to consider two port-modes for the SMA. This is expected as the first bandgap is mainly governed by the fundamental resonance of the SRRs itself, whereas for the second one the mutual coupling between the cells play a larger role.

Finally, we show that the high-fidelity model evaluations within the MFMC framework are indeed required to remove the biasing error. If one would apply a standard MC method on the surrogate model $g^{(6)}$ solely (instead of $g^{(1)}$)

the associated error is represented by the dotted red line in Fig. 3. Both error contributions, the sampling and the biasing error, are estimated again with a Monte Carlo sample.

5 Conclusions

We have presented an uncertainty propagation technique for quasi-periodic optical structures with random influences, which combines surrogate modeling of unit cells, SMA and MFMC. The resulting multifidelity approach can significantly improve the efficiency of Monte Carlo sampling. In particular speedups by orders of magnitude were obtained for a split ring resonator. The proposed method exhaustively samples unit cell models which are combined through the scattering matrix approach and hence, can be evaluated efficiently. Only a single unit cell surrogate was required which significantly reduced the number of uncertain parameters and hence, the computational complexity. The surrogate-SMA data was then corrected with a few time domain simulations of the entire structure to obtain unbiased estimates of the bandgap properties.

Acknowledgments The work of Niklas Georg is supported by the DFG grant RO4937/1-1, the *Excellence Initiative* of the German Federal and State Governments and the Graduate School of Computational Engineering at TU Darmstadt. Christian Lehmann's work is funded by the DFG grant SCHU1157/11-1.

References

1. N. Georg, D. Loukrezis, U. Römer, S. Schöps, Enhanced adaptive surrogate models with applications in uncertainty quantification for nanoplasmonics. Int. J. Uncertain. Quan. **10**(2), 165–193 (2020)
2. B. Bandlow, R. Schuhmann, G. Lubkowski, T. Weiland, Analysis of single-cell modeling of periodic metamaterial structures. IEEE Trans. Magn. **44**(6), 1662–1665 (2008)
3. B. Peherstorfer, K. Willcox, M. Gunzburger, Survey of multifidelity methods in uncertainty propagation, inference, and optimization. SIAM Rev. **60**(3), 550–591 (2018)
4. S. Linden, C. Enkrich, M. Wegener, J. Zhou, T. Koschny, C. Soukoulis, Magnetic response of metamaterials at 100 Terahertz. Science **306**, 1351–1353 (2004)
5. K. Busch, G. von Freymann, S. Linden, S. Mingaleev, L. Tkeshelashvili, M. Wegener, Periodic nanostructures for photonics. Phys. Rep. **444**(3–6), 101–202 (2007)
6. T. Weiland, Time domain electromagnetic field computation with finite difference methods. Int. J. Numer. Model. El. **9**(4), 295–319 (1996)
7. H. Glock, K. Rothemund, U. van Rienen, CSC - A procedure for coupled S-parameter calculations. IEEE Trans. Magn. **38**(2), 1173–1176 (2002)
8. A. Litvinenko, A. Yucel, H. Bagci, J. Oppelstrup, E. Michielssen, R. Tempone, Computation of electromagnetic fields scattered from objects with uncertain shapes using multilevel Monte Carlo method. IEEE J. Multiscale Multiphys. Comput. Tech. **4**, 37–50 (2019)
9. B. Peherstorfer, K. Willcox, M. Gunzburger, Optimal model management for multifidelity Monte Carlo estimation. SIAM J. Sci. Comput. **38**(5), A3163–A3194 (2016)

Dielectric Breakdown Prediction with GPU-Accelerated BEM

Cedric Münger, Steffen Börm, and Jörg Ostrowski

Abstract The prediction of a dielectric breakdown in a high-voltage device is based on criteria that evaluate the electric field along possible breakdown paths. For this purpose it is necessary to efficiently compute the electric field at arbitrary points in space. A boundary element method (BEM) based on an indirect formulation, realized with MPI-parallel collocation, has proven to cope very well with this requirement. It deploys surface meshes only, which are easy to generate even for complex industrial geometries. The assembly of the large dense BEM-matrix, as well as the iterative solution of the resulting system, and the evaluation along the field-lines all require to carry out the same type of calculation many times. Graphical Processing Units (GPUs) promise to be more efficient than Central Processing Units (CPUs) when it is possible to do the same type of calculations for large blocks of data in parallel. In this paper we therefore investigate if GPU acceleration is a measure to further speed up the established CPU-parallel BEM solver.

1 Introduction

Every high-voltage device has to pass dielectric type tests, in which a large voltage is applied to the device. The test is passed if no dielectric breakdown occurs. A breakdown usually starts from an electrode-surface with high dielectric stress, and then propagates through the volume along a field-line of the electric field **E** towards the opposite electrode, see Fig. 1. The propagation stops if the electric field along

C. Münger (✉)
Seminar for Applied Mathematics, ETH Zürich, Zürich, Switzerland
e-mail: cedric.muenger@alumni.ethz.ch

S. Börm
Math. Seminar, Christian-Albrechts-Univ. Kiel, Kiel, Germany
e-mail: boerm@math.uni-kiel.de

J. Ostrowski
ABB Corporate Research, Baden, Switzerland
e-mail: joerg.ostrowski@ch.abb.com

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_14

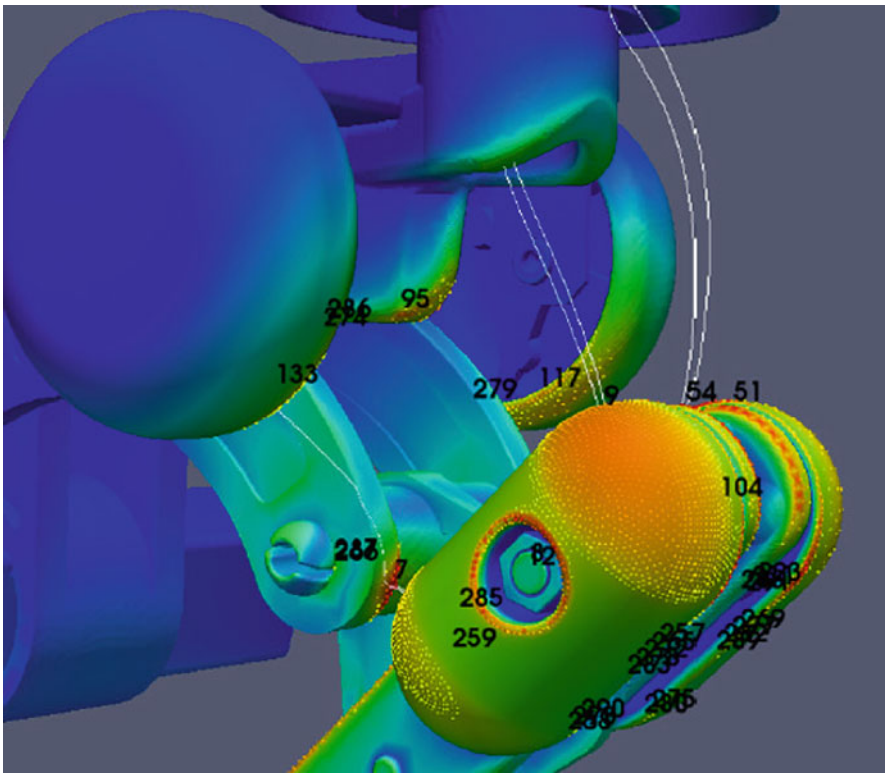


Fig. 1 The electric field strength on the surface of a disconnector and possible breakdown paths along the field lines

this breakdown path γ is not strong enough. For details see [1]. An inception of a streamer, i.e. the initial state of a breakdown only occurs if the criterion

$$\int_{\gamma} \alpha_{\text{eff}}(|\mathbf{E}|)ds > K_{\text{str}} \tag{1}$$

is fulfilled. Here α_{eff} is the effective ionization function that depends on the strength of the electric field $|\mathbf{E}|$, and K_{str} is the (gas-specific) streamer constant. The prediction of a dielectric breakdown during a type test relies on the evaluation of this criterion along the most probable breakdown paths. It requires the computation of the electric field at all surface points and along field lines in the volume.

Simulation-based dielectric design became a standard procedure because a user-friendly, i.e., fast, robust, reliable, and easy-to-use computational method was developed, see [2]. In the following we will first describe this boundary-element-based method and then introduce how general-purpose graphics processing units (GPGPUs) can be used to massively reduce computing times.

2 BEM Formulation

In this section we derive the BEM-formulation as it is in use since many years at ABB, see [2, 3]. The device consists of the subdomains $\Omega_0, \ldots, \Omega_{m-1}$, and the unbounded exterior subdomain is $\Omega_m = \mathbb{R}^3 \setminus \bigcup_{k=\{0..m-1\}} \Omega_k$, see the example in Fig. 2. The electric field $\mathbf{E} = -\,\mathbf{grad}\,\varphi$ is calculated by solving the Laplace equation

$$\operatorname{div} \epsilon\, \mathbf{grad}\, \varphi = 0 \tag{2}$$

for the electric scalar potential φ in each of these subdomains. The permittivity is denoted by ϵ. We use an indirect formulation with a single-layer potential

$$\varphi(\mathbf{x}) = \Psi_{SL}[\sigma](\mathbf{x}) = \int_{\partial\Omega} \frac{\sigma(\mathbf{y})}{4\pi\,|\mathbf{x}-\mathbf{y}|}\, dS_y\,, \tag{3}$$

and search for the unknown scalar virtual surface charge density σ that is related to the physical surface charge density σ_s. Each **conductor**, i.e. each separated conducting part with electrical conductivity $\sigma_{el} > 0$, is on a constant electrical potential. If a conductor is connected to an electric potential V_0, like Ω_0 in Fig. 2, then it holds

$$\varphi(\mathbf{x}) = V_0 \quad \forall \mathbf{x} \in \Omega_0. \tag{4}$$

The electric potential V of **floating conductors** like Ω_1 in Fig. 2 is unknown

$$\varphi(\mathbf{x}) = V \quad \forall \mathbf{x} \in \Omega_1, \tag{5}$$

and is to be determined by a charge neutrality condition, see [4]. The total charge Q of the floating conductor can be derived from the Gauss law as

$$Q = \int_{\partial\Omega_1} \sigma_s\, dS = \int_{\partial\Omega_1} \mathbf{D}\cdot\mathbf{n}\, dS. \tag{6}$$

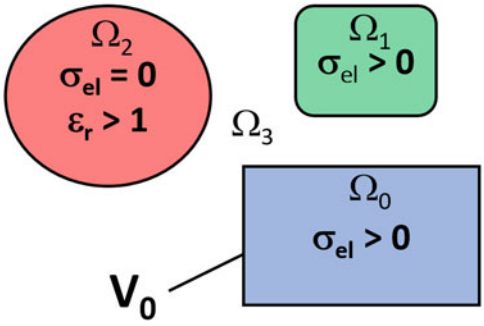


Fig. 2 Example of a device that consists of a conductor Ω_0 that is connected to a prescribed electric potential V_0, a floating conductor Ω_1, an insulator Ω_2, and the unbounded exterior domain Ω_3

The normal component of the displacement field $\mathbf{D}$ can be expressed as

$$\mathbf{D} \cdot \mathbf{n} = \varepsilon^+ \mathbf{E} \cdot \mathbf{n} = -\varepsilon^+ \, \mathbf{grad}\, \varphi \cdot \mathbf{n} = -\varepsilon^+ \, \mathbf{grad}\, \Psi_{SL}[\sigma] \cdot \mathbf{n}. \tag{7}$$

Here ε^+ denotes the permittivity of the exterior domain. The Neumann trace of the single layer potential and can be expressed with help of the adjoint double layer K'

$$\mathbf{grad}\, \Psi_{SL}[\sigma] \cdot \mathbf{n} = \frac{1}{2}\sigma + K'\sigma \text{ , with} \tag{8}$$

$$K'(\sigma)(\mathbf{x}) = \int_{\partial\Omega} \frac{\mathbf{x}-\mathbf{y}}{4\pi|\mathbf{x}-\mathbf{y}|^3} \cdot \mathbf{n}(\mathbf{x})\sigma(\mathbf{y}) dS_y \,. \tag{9}$$

Combining the Eqs. (6)–(8) with $Q = 0$ due to charge neutrality yields

$$\int_{\partial\Omega_1} \frac{1}{2}\varepsilon^+\sigma(\mathbf{y}) + \varepsilon^+(K'\sigma)(\mathbf{y}) dS_y = 0. \tag{10}$$

We model thin floating conductive sheets only by a single surface. Then the electric fields from both sides ($\pm$) need to be considered for charge neutrality, since

$$\sigma_s = \mathbf{n} \cdot (\mathbf{D}^+ - \mathbf{D}^-) \implies \tag{11}$$

$$\int_{\partial\Omega_1} \frac{1}{2}(\varepsilon^+ + \varepsilon^-)\sigma(\mathbf{y}) + (\varepsilon^+ - \varepsilon^-)(K'\sigma)(\mathbf{y}) dS_y = 0. \tag{12}$$

There is no surface charge on **non-conductors**: $\sigma_s = 0$ on $\partial\Omega_2$

$$\frac{1}{2}(\varepsilon^+ + \varepsilon^-)\sigma(\mathbf{x}) + (\varepsilon^+ - \varepsilon^-)(K'\sigma)(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \partial\Omega_2. \tag{13}$$

So for our simple but quite general example of Fig. 2 we have to solve the following set of equations:

$$\int_{\partial\Omega} \frac{\sigma(\mathbf{y})}{4\pi|\mathbf{x}-\mathbf{y}|} dS_y = V_0 \quad \forall \mathbf{x} \in \partial\Omega_0 \tag{14}$$

$$\int_{\partial\Omega} \frac{\sigma(\mathbf{y})}{4\pi|\mathbf{x}-\mathbf{y}|} dS_y - V = 0 \quad \forall \mathbf{x} \in \partial\Omega_1 \tag{15}$$

$$\int_{\partial\Omega_1} \frac{1}{2}\varepsilon^+\sigma(\mathbf{y}) + \varepsilon^+(K'\sigma)(\mathbf{y}) dS_y = 0 \tag{16}$$

$$\frac{1}{2}(\varepsilon^+ + \varepsilon^-)\sigma(\mathbf{x}) + (\varepsilon^+ - \varepsilon^-)(K'\sigma)(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \partial\Omega_2 \tag{17}$$

The solution of the system of Eqs. (14)–(16) yields the virtual surface charge distribution from which the electric field can be compute at any point in space as

$$\mathbf{E}(\mathbf{x}) = \int_{\partial\Omega} \frac{\mathbf{x}-\mathbf{y}}{4\pi|\mathbf{x}-\mathbf{y}|^3}\sigma(\mathbf{y})dS_y \quad \forall \mathbf{x} \in \mathbb{R}^3. \tag{18}$$

3 Discretization

We use a collocation boundary element approach: the surface $\partial\Omega$ is represented by a collection of triangles $\tau_1, \ldots, \tau_N$ with vertices $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The unknown function σ

$$\sigma_h(\mathbf{y}) = \sum_{j=1}^{n} u_j \psi_j(\mathbf{y}),$$

is approximated by suitable basis functions $\psi_1, \ldots, \psi_n$. This approach is to be inserted into (14)–(16) and this set of equations is only required to hold in the *collocation points* $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This is an $n + N_{fl}$-dimensional system of linear equations, with N_{fl} being the number of floating conductors. The implementation of this approach poses a number of challenges:

- High-voltage devices have smooth curved surfaces in order to avoid field enhancements. We use piecewise quadratic parametrizations with (curved) triangles $\tau_1, \ldots, \tau_N$ to minimize the geometrical discretization error.
- The entries of the matrix corresponding to the linear system have to be computed. In the established CPU-based method we employ an MPI-parallel implementation of suitable quadrature rules.
- The system of linear equations has to be solved. We use Krylov subspace methods, since these methods only need matrix-vector multiplications, which can easily provided in the established method by the MPI-based distributed representation of the matrix.

4 GPGPU Quadrature

Even on modern processors of parallel computers it is time-consuming to compute the matrix entries v_{ij} for complex industrial geometries. It is advantageous to calculate the surface integral for pairs of collocation points and triangles.

$$v_{ij} = \int_{\partial\Omega} \frac{\psi_j(\mathbf{y})}{4\pi|\mathbf{x}_i-\mathbf{y}|}\, dS_y \qquad \text{for all } i, j \in \{1, \ldots, n\} \tag{19}$$

Matrix assembly is however ideally suited for parallelization on SIMD-type of processors, because the integrations require us to carry out mostly identical operations for *all* matrix entries. The currently most common processors are general-purpose graphics processing units (GPGPUs) like the NVidia Tesla™ or AMD Radeon Instinct™ cards that contain thousands of floating-point arithmetic units and offer teraflop-level performance. Porting the quadratures to GPGPUs poses challenges:

- The most powerful GPGPU models are equipped with fast local memory. We have to ensure that geometrical data is efficiently transferred to the local memory.
- We are using piecewise (mapped) linear basis functions and multiple triangles contribute to the same matrix entry. Thus we have to avoid collisions between different triangles that may try to simultaneously update the same matrix entry.

For the parallelization it is beneficial that each collocation point corresponds to one row in the matrix. In our implementation we take advantage of this fact by assigning each collocation point to a thread. Then we consecutively iterate through all triangles of the discretization and simultaneously compute the contribution of a triangle to all collocation points, see Fig. 3. This way collisions can be avoided. For good performance it is required that all threads execute exactly the same operations. This can be ensured by grouping the collocation points depending on whether they are on the surface of a conductor or part of a dielectric interface.

For the quadrature we distinguish between regular, near-singular, and singular integration of the pairs of triangles and collocation points. If the collocation point is a node of the triangle then it is a singular pair. When the distance D between the circumcenter of the triangle and the collocation point is larger than a threshold that scales with the circumradius R of the triangle like

$$D > \eta \cdot R \tag{20}$$

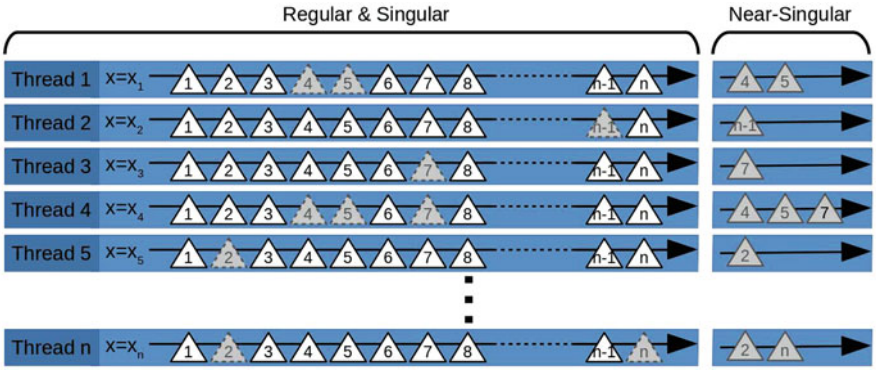


Fig. 3 Near-singular assembly: The gray triangles are near-singular and are computed after the regular and singular triangles

then it is a regular pair. We used the scaling parameter $\eta = 1.2$. All other pairs are near-singular. Circumcenters and circumradii of the triangles can be precomputed.

All three types of pairs can be integrated by using the well-known Duffy-transformation, see [5]. This is straightforward for regular and singular pairs, only the near-singular pairs need to be integrated adaptively for accuracy. They frequently occur, e.g. in cases with narrow gaps in the geometry, or during the postprocessing, when a point near the surface is to be evaluated. In this near-singular case we first compute the point of the triangle that is closest to the collocation point. Next we subdivide the triangle into smaller triangles such that this closest point is a corner-node of a subdividing triangle. Then we again employ the Duffy-transformation to integrate over all smaller triangles. This yields an adaptive quadrature with increased accuracy around the closest point. The adaptivity can impact the performance of a GPU-computation badly if no attention is paid, because then there is divergence in the control flow on the GPU. In order to minimize this divergence, we first compute the regular and singular pairs, and deal with the near-singular integrals later.

The categorization into regular, near-singular and singular pairs is carried out on the fly during the iteration through the triangles. If a pair is marked as near-singular, then it will be marked as not processed, see Fig. 3. They are computed in parallel by using the subdivision method after the regular and singular cases have been completed. This strategy allows that all three types of integrations are carried out in parallel, without the need to mix operations.

The full matrix may not fit in the memory of one GPU for larger problems. Therefore, but also to speed up the computation we use multiple GPUs. Due to the independence of matrix rows we split the matrix into multiple blocks of rows that can be computed and stored independently on different GPUs.

5 Numerical Experiments

In this chapter we show some examples that were computed with the novel GPU-implementation that is based on the H2Lib package, see [6]. We first validated our implementation for an axial-symmetric case. We compared the results of the H2Lib with the results of the already existing simulation tools Polopt (3D) see [3], and Elfi (2D) see [2]. Next we compared the performance of the new GPU-parallel H2Lib implementation with the performance of the existing MPI-parallel Polopt tool.

5.1 Validation

The benchmark problem that is used to validate the GPU-implementation in H2Lib is a bushing, see Fig. 4. It consists of an insulator that is wrapped around a conductor on 100 kV high-voltage. Five thin conducting sheets are embedded in this insulator. They accomplish the field grading. The outermost is grounded, and the potentials

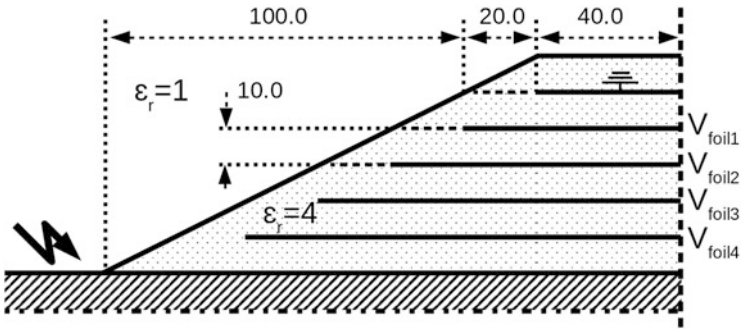


Fig. 4 Cross-section of the bushing, including dimensions in *mm*

Table 1 Potentials of the conducting sheets of the bushing

	ELFI(2D)	POLOPT	H2LIB
V_{foil1}	70.8 kV	70.15 kV	70.22 kV
V_{foil2}	51.4 kV	50.47 kV	50.50 kV
V_{foil3}	35.0 kV	34.00 kV	34.02 kV
V_{foil4}	18.9 kV	18.02 kV	18.02 kV

of the other (floating) sheets are unknown and are to be determined. The sheets are treated as single surfaces according to Eq. (12). Their potentials were computed with all three solvers. POLOPT and H2Lib use the same mesh with 3'526 nodes. The results agree very well, see Table 1. The small remaining differences are due to the use of different quadratures.

5.2 GPU-Acceleration

After the successful validation of the H2Lib implementation, we compared the computing times for the GPU-parallel H2Lib and the CPU-parallel Polopt. In both cases we assembled the dense BEM matrix and solved the system with an iterative GMRES with diagonal preconditioner. So the expected numerical work is quadratically depending on the degrees of freedom, i.e. here the number of nodes (# Nodes). Figure 5 shows the times that it took for matrix-assembly, iterative solution, and computation of the electric field at the surfaces of a realistic high-voltage device for different mesh-sizes. POLOPT used 180 CPU cores distributed over 5 nodes with two 18-core CPUs each. The CPU cluster was optimized for these calculations because it is in use by ABB product designers. For H2Lib we used a total of 12 NVidia GTX 1080Ti, distributed over multiple nodes. For POLOPT we clearly recognize the quadratic scaling. The H2Lib also scales quadratically, see Fig. 6, the proportionality constant seems however to be much better than the one for Polopt.

We computed another larger example see Table 2. We used 180 cores for POLOPT for all meshes except the largest one, where we used 360 cores. The

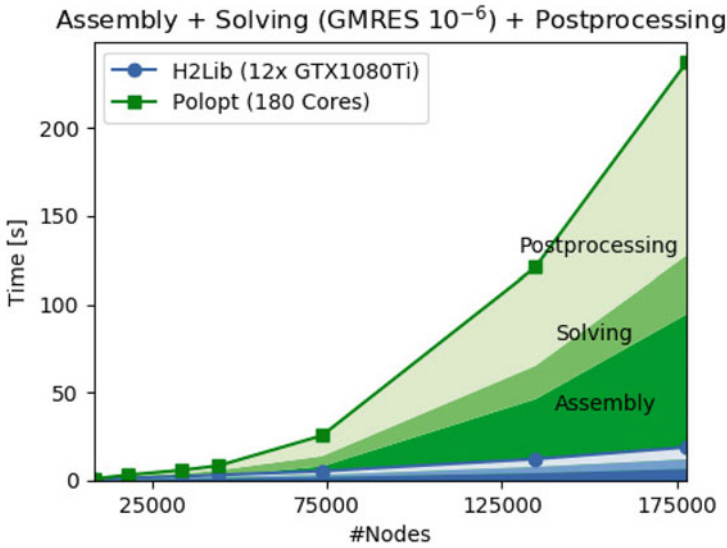


Fig. 5 Cumulative times for assembly, solving and surface electric field computation for POLOPT and H2LIB

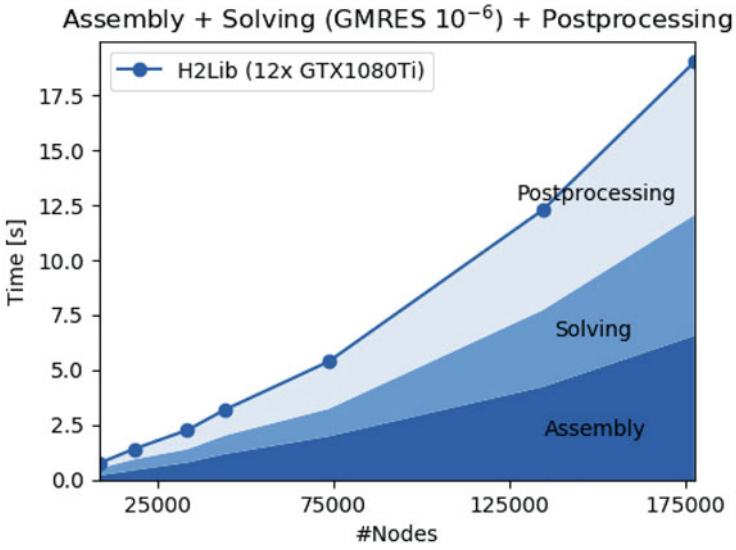


Fig. 6 Cumulative times for assembly, solving and surface electric field computation for the H2Lib only

number of GPUs were chosen such that the matrix fitted in the combined memory of all GPUs in single precision for H2Lib. Again the GPU-parallel implementation clearly outperforms the CPU-parallel version. We also evaluated the breakdown

Table 2 Calculation times for different meshes. H2Lib used multiple GTX1080Ti, POLOPT is executed on 180 CPU-Cores

#Nodes	#GPUs	H2Lib	POLOPT
68'218	2	18 s	33 s
140'183	8	19 s	130 s
232'029	20	26 s	371 s
330'706	40	34 s	702 s
432'084	72	41 s	732 s[a]

[a]360 Cores

Table 3 Time for computation and fieldline evaluation. Model size: 68'218 Nodes. Polopt in serial for the evaluation

	POLOPT 180 Cores	H2Lib 4xTesla P100
Assembly	11 s	2 s
Solving	11 s	2 s
E surface	11 s	2 s
Total	33 s	6 s
Evaluation	102 s for 17 fieldlines	67 s for 9219 fieldlines

criterion along the most probable breakdown paths (field lines), see Table 3. The acceleration seems even higher, however the evaluation is only implemented in serial in Polopt.

6 Conclusions and Outlook

The usage of GPUs drastically accelerates the computation of electrostatic fields, as well as their evaluation with respect to breakdown inception. This opens the door not only for the computation of larger problems, but also for the inclusion of additional physical models (e.g. for surface charging), or for optimization. Most promising is the combination of the GPU-acceleration with a compression technique, see [7, 8]. Another strong acceleration is to be expected in this case. Moreover the asymptotical behavior will no longer depend quadratically on the degrees of freedom.

References

1. A. Blaszczyk, J. Ekeberg, S. Pasncheshnyi, M. Saxegaard, Virtual High Voltage Lab , in *Scientific Computing in Electrical Engineering*. Mathematics in Industry (Springer Book, Cham, 2016)
2. A. Blaszczyk, H. Steinbigler, Region oriented charge simulation. IEEE Trans. Magn. **30**(5), 2924–2927 (1994)
3. N. De Kock, M. Mendik, Z. Andjelic, A. Blaszczyk, Application of 3D boundary element method in the design of the EHV GIS components. IEEE Magn. Electr. Insul. **14**(3), 17–22 (1998)

4. D. Amann, A. Blaszczyk, G. Of, O. Steinbach, Simulation of floating potentials in industrial applications by boundary element methods. J. Math. Ind. **4**(1), 1–15 (2014)
5. M.G. Duffy, Quadrature over a pyramid or cube of integrands with a singularity at a vertex. SIAM J. Num. Anal. **19**(6), 1260–1262 (1982)
6. S. Börm, S. Christophersen, et al., H2Lib, a software library for $\mathscr{H}$- and $\mathscr{H}^2$-matrices. Open source, available at http://www.h2lib.org
7. S. Börm, S. Christophersen, Approximation of integral operators by Green quadrature and nested cross approximation. Num. Math. **133**(3), 409–552 (2016)
8. S. Börm, S. Christophersen, GCA-H^2 matrix compression for electrostatic simulations, in *International Conference on Scientific Computing in Electrical Engineering*, Taormina (2018)

Empirical Analysis of a Coaxial Microwave Structure with Finite Transmission Zero

K. Papke, F. Gerigk, and U. van Rienen

Abstract Empirical studies are presented on a certain radio frequency (RF) structure that has not yet been well understood. The coaxial structure provides almost ideal conditions to approximate high-pass filter functions. It has been investigated by the aid of numerical simulations accompanied by the search for appropriate equivalent microwave networks. A particular feature is a finite transmission zero which allows not only for maximally flat and Chebyshev approximations but also the synthesis of elliptic filter functions. The synthesis is drawn by means of two examples taking into account the topology of the equivalent circuit.

1 Introduction

Coaxial microwave filters have been applied for decades to damp unwanted resonant modes in accelerating and deflecting type cavities operating at tens of megahertz up to few gigahertz while the extracted power may reach the level of 1 kW in particular cases [1–4]. These so-called higher-order mode couplers are essentially high-pass or pseudo-high-pass filters consisting of coaxial lines and certain discontinuities in between. Early design procedures were focused on the implementation of narrow-band band-pass filters using reactance-coupled $\lambda/2$ resonators [5, pp. 528]. However, such semi-analytical approaches provide only rough estimates for the

K. Papke (✉)
CERN, Geneva, Switzerland

University of Rostock, Rostock, Germany
e-mail: kai.papke@cern.ch; kai.papke@uni-rostock.de

F. Gerigk
CERN, Geneva, Switzerland
e-mail: frank.gerigk@cern.ch

U. van Rienen
University of Rostock, Rostock, Germany
e-mail: ursula.van-rienen@uni-rostock.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_15

geometrical parameters of coaxial microwave filters. Since the 1990s, the filter design was more and more based on numerical simulations which permit the precise evaluation of scattering properties associated with arbitrarily shaped microwave structures and their systematic adaptation according to individual requirements. Still, the selection of a suitable topology, as chosen prior to the numerical analyses, is very much in the realm of intuition and experience [1]. Even for a specified topology, the applied numerical optimization scheme may not be able to converge against the best solution, given a certain set of requirements, as too many variables may be involved.

This paper proposes a generally applicable procedure to systematically design coaxial microwave filters on the basis of filter or transfer functions; this further implies the most suitable topology for the given problem. The synthesis of a filter function rests on the idea that scattering properties of discontinuities in coaxial guides are well described by lumped elements within the interesting frequency range, i.e. by equivalent circuits. A large variety of microwave structures with certain filter characteristics and appropriate equivalent circuits has been worked out already until the 1950s [5, 6]. Still, scattering properties of coaxial microwave structures with multiple discontinuities being relatively close to each other, so that evanescent modes may interact, are partially unexplored by means of equivalent circuits. The structure sketched in Fig. 1 is such an example. In the limit of vanishing coaxial lines, it may equivalently be described by a canonical network realization of third-order high-pass filter functions. A particular feature is the transmission zero at finite, non-vanishing frequency which allows not only for maximally flat and Chebyshev approximations but also the synthesis of elliptic filter functions. The equivalent circuit can be considered as surrogate system, significantly cheaper to evaluate than the three-dimensional field problem, and with excellent approximation properties within a certain frequency range. This together with the fact that equivalent circuit parameters allow for large adjustment ranges as shown

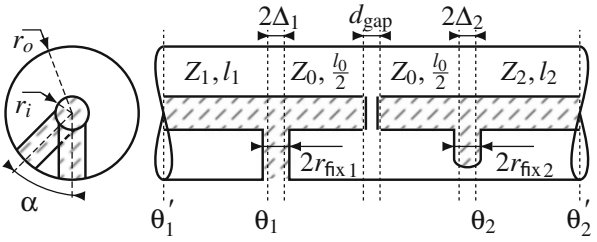


Fig. 1 Cross-sectional and side view of a coaxial structure with two cylindrical fixings of radii $r_{\text{fix}\,1}, r_{\text{fix}\,2}$ between the inner and outer conductor of radii r_i and r_o, respectively. Both fixings are rotated against each other in the transverse plane by the angle α. The parameters Δ_1 and Δ_2 represent the electric thicknesses of the corresponding fixings. The inner conductor is interrupted at the center by a distance d_{gap}. The sections of coaxial guides are described by the lengths l_ν and characteristic impedances Z_ν with $\nu = 0, 1, 2$. Terminal planes are denoted as θ_μ or θ'_μ with $\mu = 1, 2$

in the following, makes the microwave structure in Fig. 1 particularly interesting for the synthesis of high-pass filter functions with finite transmission zeros.

After introducing the equivalent network in Sect. 2, individual circuit elements are further investigated in Sect. 3, some of which provide unexpected and qualitatively new behavior for the microwave circuit theory. Finally, the synthesis is drawn in Sect. 4 by means of two examples.

2 Equivalent Circuit

The propagation of the transverse electromagnetic mode shall be considered so that the microwave structure can be represented by a two-port. Figure 2 sketches an equivalent circuit suitable to approximate the scattering matrix $\mathbf{S}$ of the structure shown in Fig. 1 between the terminal planes θ_1 and θ_2 over a considerable frequency range. It is the result of an intense systematic research. A typical fit of simulated scattering functions is shown in Fig. 3. Remarkable is the transmission zero at finite, non-vanishing frequency accounted for by the parallel LC resonator. Its resonant frequency $\omega_0 = 1/\sqrt{L_0 C_0}$ is seen from the transmission power gain in Fig. 3a.

To derive the network parameters of Fig. 2, at first, the scattering matrix $\mathbf{S}'$ of the microwave structure between the terminal planes θ_1' and θ_2' in Fig. 1 is obtained by numerical simulations.[1] Let the matrix $\mathbf{S}'$ be defined by the elements s_{ij}'; $i, j = 1, 2$ which are functions of the frequency ω. A subsequent transformation provides the scattering matrix $\mathbf{S}$ at the terminal planes of interest, θ_1 and θ_2 [9, pp. 184]

$$\mathbf{S} = \mathbf{w}\,\mathbf{S}'\mathbf{w} \tag{1}$$

where the diagonal matrix $\mathbf{w} = \mathrm{diag}\{\mathrm{e}^{\mathrm{i}\beta l_1}, \mathrm{e}^{\mathrm{i}\beta l_2}\}$ invokes the inward phase shifts along each terminal translation given the propagation constant β and lengths l_1 and l_2 which are per se not known due to the finite thickness of the obstacles in the structure. Consequently, the elements s_{ij}; $i, j = 1, 2$ of the scattering matrix $\mathbf{S}$ are considered as functions of these lengths and the frequency.

Let $l_1 + \Delta_1$ be the distance from the terminal plane θ_1' to the center of the left fixing in Fig. 1 which is a priori known. Similar, let $l_2 + \Delta_2$ be the distance from the terminal plane θ_2' to the center of the right fixing. Since Δ_1 and Δ_2 define half of the electric "thickness" of the individual fixing, the sum $l_0 + d_{\mathrm{gap}} + \Delta_1 + \Delta_2$ must correspond to the distance between the centers of both fixings. Linear transforms $\{\Delta_1, \Delta_2\} \mapsto \{l_0, l_1, l_2\}$ are introduced to reduce the number of length variables as well as to confine their variation to the vicinity of the corresponding fixing.

[1] Numerical simulations are mostly carried out using CST STUDIO SUITE® software [7] for the present work. In part, they are verified with COMSOL Multiphysics [8].

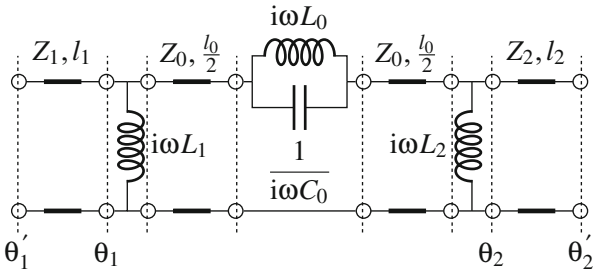


Fig. 2 Equivalent circuit model composed of lumped elements and transmission lines

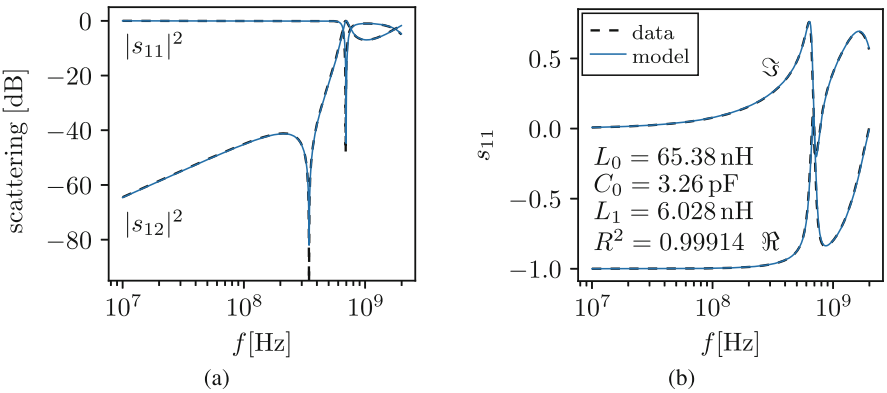


Fig. 3 Approximation of the numerically simulated RF reflection and transmission with respect to the terminal planes θ_1 and θ_1 by means of the equivalent circuit model according to Figs. 1 and 2. The structure is assumed to be symmetric, hence, $Z_1 = Z_2$, $l_1 = l_2$, and $r_{\text{fix}1} = r_{\text{fix}2}$, with $r_i = 5$ mm, $r_o = 22.5$ mm, and $r_{\text{fix}1} = 3$ mm. The cross section of the coaxial guide in between the fixings is identical to those of the input and output regions. The fixings are separated by a distance of $d = 22.5$ mm while the inner conductor is separated by a distance of $d_{\text{gap}} = 0.3$ mm. Circuit parameters L_0, C_0, L_1 are derived from the minimization problem (6). (**a**) Transmission and reflection power gains $|s_{12}|^2$, $|s_{11}|^2$. (**b**) Real and imaginary part of the reflection at the terminal plane θ_1. The R^2 value reveals very good approximation in the considered frequency range $f \leq 2$ GHz

The equivalent circuit in Fig. 2 admits a transmission matrix $\mathbf{T}_{\text{model}}$ between the terminal planes θ_1 and θ_2 whose elements t_{ij}; $i, j = 1, 2$ are given by

$$t_{11} = \frac{t_{12}}{\mathrm{i}\omega L_2} + \cos\beta l_0 - \frac{1}{2Z_0C_0}\frac{\omega}{\omega_0^2 - \omega^2}\sin\beta l_0, \tag{2}$$

$$t_{12} = \mathrm{i}\frac{1}{C_0}\frac{\omega}{\omega_0^2 - \omega^2}\cos^2\frac{\beta l_0}{2} + \mathrm{i}Z_0\sin\beta l_0, \tag{3}$$

$$t_{22} = \frac{t_{12}}{\mathrm{i}\omega L_1} + \cos\beta l_0 - \frac{1}{2Z_0C_0}\frac{\omega}{\omega_0^2 - \omega^2}\sin\beta l_0, \tag{4}$$

$$t_{11}t_{22} - t_{12}t_{21} = 1. \tag{5}$$

The necessary condition for the frequency response of the microwave structure being approximated by the equivalent circuit can be formulated as

$$\min_{\Delta_1, L_1, \Delta_2, L_2, Z_0, C_0} \sum_k \|\mathbf{T}(\omega_k) - \mathbf{T}_{\text{model}}(\omega_k)\|^2 , \tag{6}$$

where $\mathbf{T}$ results from the simulated and phase shifted scattering matrix $\mathbf{S}$ sampled at the frequencies ω_k. The relationship between the scattering matrix $\mathbf{S}$ and transmission matrix $\mathbf{T}$ can be found in [9, p. 192]. The sufficient condition requires the residual to become small and, thus, defines the applicable frequency range for the model. Since the resonant frequency of the LC resonator is directly obtained from the simulated transmission power gain $|s_{12}|^2$ only one parameter, either L_0 or C_0 is involved in the nonlinear least-square problem (6). It can be solved by iterative minimization schemes, such as a constrained BFGS algorithm.[2]

3 Analyses

The dependency of circuit parameters has been studied by successively changing the geometry. Notable characteristics are related to the LC resonator causing the transmission zero at finite, non-vanishing frequency. For simplicity, the input and output waveguide regions as well as cylindrical fixings in Fig. 1 are respectively chosen to be identically, hence $Z_1 = Z_2$ and $L_1 = L_2$. For all geometric variations, the frequency response of the coaxial structure is well approximated for $f \leq 3\,\text{GHz}$ with residuals in the order of 10^{-2} by solving (6). The convergence error of the underlying numerical simulations is significantly smaller. Consistent and smooth variations of $\{\Delta_1, L_1, \Delta_2, L_2, Z_0, C_0, L_0\}$ further justify the circuit in Fig. 2 being equivalent.

Figure 4 shows the capacitance and inductance of the LC resonator as functions of the rotating angle α and distance d between the centers of both fixings. To illustrate the influence of fringe fields, the capacitance C_0 is normalized to $C_0^* = \varepsilon_0 \pi r_i^2 / d_{\text{gap}}$, where ε_0 is the permittivity constant of vacuum. Moreover, L_0 is normalized to the shunt inductance L_1 which only marginally varies with d. Significant alterations with the angle, in particular, at small distances d provide evidence for the parallel resonator being related to evanescent higher order multipole modes in the coaxial guide. It is remarkable that the frequency response of such a problem is fully characterized by a simple resonator in between two transmission lines. Furthermore, parameter variations over orders of magnitude in Fig. 4b allow for any reasonable inductance L_0 while the impact of transmission lines can be reduced as desired.

[2] Broyden-Fletcher-Goldfarb-Shanno algorithm.

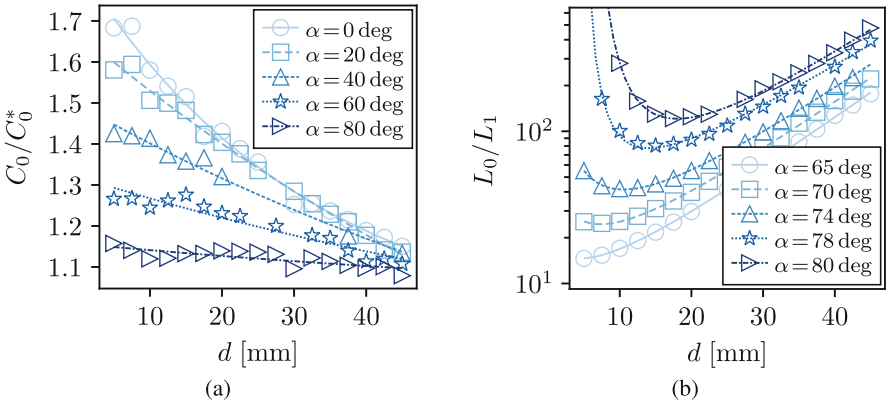


Fig. 4 (**a**) Capacitance of the parallel LC resonator normalized to the definition $C_0^* = \varepsilon_0 \pi r_i^2 / d_{\text{gap}}$ and (**b**) ratio of series and shunt inductances, both as functions of the distance between the fixings, d. Furthermore, it is r_i=5 mm, r_o=22.5 mm, $r_{\text{fix}\,1}$=$r_{\text{fix}\,2}$=3 mm, d_{gap}=0.3 mm

4 Application

Given the squared magnitude of a rational transfer function according to

$$H(\mathrm{i}\omega) H^*(\mathrm{i}\omega) = \frac{c_0}{1 + \varepsilon D_n(\mathrm{i}\omega) D_n^*(\mathrm{i}\omega)}, \tag{7}$$

where c_0 and ε are scalars and $D_n(\mathrm{i}\omega)$ is the filter function, the systematic approach to derive a microwave circuit being able to approximate this frequency response is defined as synthesis. In accordance to the equivalent circuit in Fig. 2, any rational third-order high-pass filter function may be considered. Particularly interesting are elliptic filter functions as they yield the steepest transition between passband and stopband given certain attenuation limits in both frequency bands [10, pp. 207]. The synthesis of elliptic filters is drawn in the following.

Consider the frequency map $f : \Omega \mapsto \omega$. It maps a normalized frequency space associated with a low-pass to the frequency space of a high-pass according to the definition $\omega = \sqrt{\omega_p \omega_s}/\Omega$, with the passband and stopband edges ω_p and ω_s, respectively. The filter function of a normalized elliptic low-pass of odd order n is defined as [11]

$$D_n(\mathrm{i}\Omega) = c_1 \mathrm{i}\Omega \prod_{\nu}^{(n-1)/2} \frac{\Omega^2 - \Omega_{0\nu}^2}{\Omega_{0\nu}^2 \Omega^2 - 1}, \tag{8}$$

where $n = 3, 5, 7, \cdots$. The zeros are calculated by Jacobian elliptic sine functions as $\Omega_{0\nu} = k\,\mathrm{sn}\,(2\nu K/n, k)$, where K is the complete elliptic integral of the first kind with the modulus $k = \Omega_s^{-2}$. The factor c_1 in (8) normalizes the maximum deviation.

The transfer function H is directly related to $|s_{12}|^2$ taking into account the impedance normalization to Z_1 and Z_2 at the corresponding terminal planes in Fig. 1 [10, pp. 163]. Lossless two-ports admit unitary scattering matrices, hence, $\mathbf{S}^H\mathbf{S} = \mathbf{I}$. They further fulfill reciprocity, so that $s_{12} = s_{21}$. Both properties are used to derive $\mathbf{S}$ from $|s_{12}|^2$, only. The corresponding impedance matrix is obtained via

$$\mathbf{Z} = \mathbf{P}^{\frac{1}{2}} \left[\mathbf{I} + \mathbf{S}\right] \left[\mathbf{I} - \mathbf{S}\right]^{-1} \mathbf{P}^{\frac{1}{2}}, \tag{9}$$

where $\mathbf{P} = \mathrm{diag}\{Z_1, Z_2\}$ accounts for the impedance normalization, and $\mathbf{I}$ is the identity matrix. The impedances and admittances in the series, or respectively, shunt arms of the ladder network are derived by continued fraction expansion of z_{11} and z_{22} about $\Omega = \infty$ [10, p. 165]. Finally, the map $\Omega \mapsto \omega$ is applied.

The insertion loss method described above does not account for distributed elements such as transmission lines. The procedure provides an initial set of values for the lumped elements assuming that the transmission lines are not present. The structure is simulated repeatedly, with the geometry being gradually changed so that the fitted circuit parameters in (2)–(4) approximately match the desired ones. Table 1 lists the parameter values for a third-order elliptic filter with passband and stopband edges of $f_p = 1.19\,\mathrm{GHz}$ and $f_s = 0.79\,\mathrm{GHz}$, respectively. For the same example, the geometry and insertion loss [9, p. 63] are shown in Fig. 5. There is a notable influence by transmission lines despite a relative short distance between the

Table 1 Parameters of a third-order elliptic filter with $f_s = 0.79\,\mathrm{GHz}$ and $f_p = 1.19\,\mathrm{GHz}$

Parameter	Lumped circuit	Approximation by coaxial guide
L_0	18.493 nH	18.412 nH
C_0	2.739 pF	2.751 pF
L_1, L_2	4.201 nH	4.227 nH

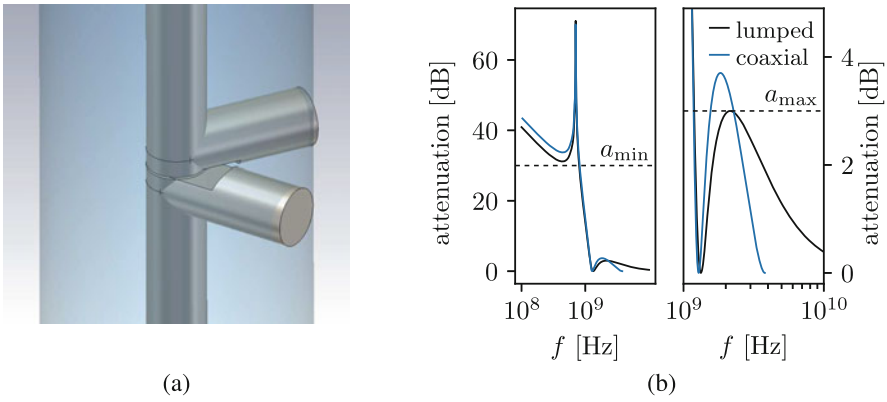
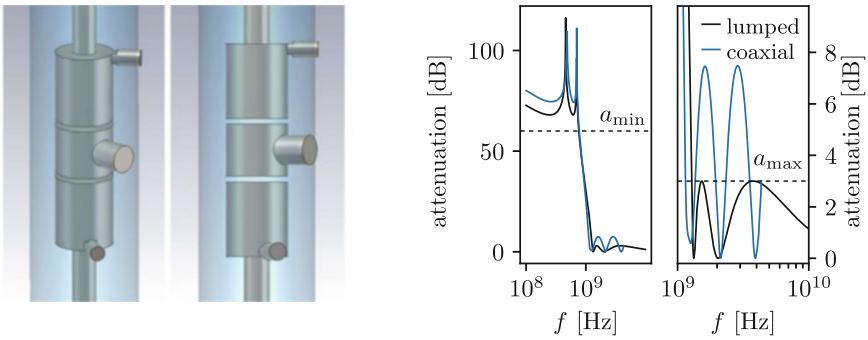


Fig. 5 (**a**) Optimized microwave structure to approximate a third-order elliptic high-pass filter characteristics. (**b**) Insertion loss as given by the predefined transfer function $H(\mathrm{i}\omega)$ in black and the approximation by the optimized structure in blue. Deviations are caused by transmission lines



(a) (b)

Fig. 6 (**a**) Optimized microwave structure to approximate a fifth-order elliptic high-pass filter characteristics. (**b**) Insertion loss as given by the predefined transfer function $H(\mathrm{i}\omega)$ in black and the approximation by the optimized structure in blue. Deviations are caused by transmission lines

fixings. Another example using the same passband and stopband edges but higher order is shown in Fig. 6. It results from a cascade of two structures each adjusted as a third-order filter. The subsequent connection requires significant changes of the rotation angles in order to achieve the attenuation curve shown in Fig. 6b.

5 Conclusions

To the authors' best knowledge, this work contains three new scientific contributions. First, the systematic design of coaxial microwave filters on the basis of abstract filter or transfer functions was demonstrated. It enables both the design of coaxial high-order mode couplers under completely new aspects and fundamental predictions about the topology prior to any computational refinement. The synthesis is based on equivalent circuit models based on a finite cascade of lumped, lossless two-ports and transmission lines whose parameters are fitted according to simulated scattering functions. One structure was elaborated and is particularly suitable for the synthesis of rational high-pass filter functions. A second important finding is that the ladder network topology for the equivalent circuit remains valid even in the presence of evanescent mode coupling between adjacent discontinuities of the coaxial guide. Finally, the empirical studies on the considered microwave structure, i.e. the nature of its transmission zero at finite, non-vanishing frequency, open up new research topics for the microwave circuit theory and await field theoretical analyses.

References

1. E. Haebel, Couplers, tutorial and update. Part. Accel. **40**, 141–159 (1992)
2. Q. Wu, B. Sergey, I. Ben-Zvi, et. al., Operation of the 56 MHz superconducting rf cavity in RHIC with higher order mode damper. APS **22**, 102001 (2019)
3. J. Mitchel, Higher Order Modes and Dampers for the LHC Double Quarter Wave Crab Cavity. Ph.D. thesis, University of Lancaster, 2019
4. K. Papke, F. Gerigk, U. van Rienen, Comparison of coaxial higher order mode couplers for the CERN Superconducting Proton Linac study. APS **20**, 060401 (2017)
5. G.L. Matthaei, L. Young, E.M.T. Jones, *Microwave Filters, Impedance-matching Networks, and Coupling Structures* (Artech House, Norwood, 1980)
6. N. Marcuwitz, *Waveguide Handbook* (McGraw-Hill Book Company, New York, 1951)
7. CST - Computer Simulation Technology Ver. 2016. CST AG, Darmstadt, Germany (2016)
8. COMSOL Multiphysics Ver. 5.3. COMSOL Multiphysics GmbH, Stockholm, Sweden (2017)
9. D.M. Pozar, *Microwave Engineering*, 4th edn. (Wiley, New York, 2012)
10. O. Wing, *Classical Circuit Theory* (Springer, New York, 2008)
11. R. Saal, E. Ulbrich, On the design of filters by synthesis. IRE Trans. Circ. Theory **5**, 284–327 (1958)

Frequency-Domain Non-intrusive Greedy Model Order Reduction Based on Minimal Rational Approximation

Davide Pradovera and Fabio Nobile

Abstract We present a technique for Model Order Reduction (MOR) of frequency-domain problems relying on rational interpolation of vector-valued functions. The selection of the sample points is carried out adaptively according to a greedy procedure. We describe several options for the choice of a posteriori error indicators, which are used to drive the greedy algorithm and define its termination condition. Namely, we illustrate a tradeoff between each indicator's accuracy and its "intrusiveness", i.e. how much information on the underlying high-fidelity model needs to be available. We test numerically the effectiveness of this technique in solving a non-Hermitian eigenproblem and a microwave frequency response analysis.

1 Introduction

Consider the function $\vec{u} : \mathbb{C} \ni \mu \mapsto \vec{u}(\mu) \in \mathbb{C}^n$ implicitly defined as the solution of the linear parametric problem with a single parameter

$$\mathrm{A}(\mu)\vec{u}(\mu) = \vec{f}(\mu), \tag{1}$$

with A and $\vec{f}$ smooth functions taking values in $\mathbb{C}^{n\times n}$ and $\mathbb{C}^n$, respectively. To be more specific, we consider here parametric problems (1) arising from spatial discretization (e.g., by FEM [3]) of frequency domain problems, with the parameter μ representing the frequency. For most such problems, $\mathrm{A}(\mu)$ depends at most quadratically on μ:

$$\mathrm{A}(\mu) = \mathrm{A}_0 + \mu\mathrm{A}_1 + \mu^2\mathrm{A}_2,$$

whereas $\vec{f}$ is usually of the form $\vec{f}(\mu) = \theta_0(\mu)\vec{f}_0$, with $\theta_0 : \mathbb{C} \to \mathbb{C}$.

D. Pradovera (✉) · F. Nobile
CSQI, Institute of Mathematics, EPFL, Lausanne, Switzerland
e-mail: davide.pradovera@epfl.ch; fabio.nobile@epfl.ch

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_16

In applications, it is often computationally unfeasible to solve (1) as many times as needed for a frequency response analysis. In recent years, this issue has been solved through MOR, whose main purpose is the construction of a surrogate $\widetilde{\vec{u}}(\mu)$ for $\vec{u}(\mu)$, much cheaper to evaluate at any given μ than solving (1), and with good approximation properties.

2 Available MOR Strategies

A plethora of MOR techniques have been employed to compute surrogates for frequency response problems; some notable ones include:

- projective techniques, e.g. the Reduced Basis (RB) and multi-moment-matching methods [2], are extremely powerful, but require knowledge of, and access to, the specific structure of A and $\vec{f}$;
- strategies based on rational approximation, e.g. the Löwner framework [4] or the Vector Fitting (VF) algorithm, are *non-intrusive*, i.e. they rely only on evaluations of $\vec{u}$ at few frequencies, which we will refer to as *snapshots* or *samples*; in particular, there is no need for any information on (nor access to) the specific structure of A and $\vec{f}$ in (1); the price to pay for this additional flexibility is a reduced accuracy of the method for a given number of snapshots.

More recently, the minimal rational interpolation (MRI) technique was proposed [7], trying to achieve non-intrusiveness and optimal snapshot management at the same time. We summarize here a practical scheme for MRI:

1. fix a set of sample points $\mu_1, \dots, \mu_S \in \mathbb{C}$, and a polynomial basis $\{\psi_i\}_{i=0}^{S-1} \subset \mathbb{P}^{S-1}(\mathbb{C})$ (e.g., one could choose monomials, or Chebyshev polynomials); also, let $\{\ell_j\}_{j=1}^{S} \subset \mathbb{P}^{S-1}(\mathbb{C})$ be the Lagrangian basis associated to the sample points;
2. build the Vandermonde matrix $\mathrm{V} \in \mathbb{C}^{S\times S}$ and the diagonal weight matrix D:

$$(\mathrm{V})_{ij} = \psi_j(\mu_i) \quad \text{and} \quad \mathrm{D} = \operatorname{diag}\left(\left[\frac{\mathrm{d}^{S-1}\ell_1}{\mathrm{d}\mu^{S-1}}, \dots, \frac{\mathrm{d}^{S-1}\ell_S}{\mathrm{d}\mu^{S-1}}\right]\right) \in \mathbb{C}^{S\times S};$$

3. compute the snapshots $\vec{u}(\mu_1), \dots, \vec{u}(\mu_S)$ and assemble a QR decomposition of the snapshot matrix

$$\left[\vec{u}(\mu_1)\middle|\vec{u}(\mu_2)\middle|\cdots\middle|\vec{u}(\mu_S)\right] = \mathrm{WR}, \quad \text{with } \mathrm{W} \in \mathbb{C}^{n\times S},\ \mathrm{R} \in \mathbb{C}^{S\times S}; \tag{2}$$

4. compute a minimal eigenvector $\vec{q} \in \mathbb{C}^S$ of the positive semidefinite (Gramian) matrix $(\mathrm{RDV})^H\,\mathrm{RDV}$, and define the surrogate denominator as

$$Q \in \mathbb{P}^{S-1}(\mathbb{C}), \quad Q(\mu) = \sum_{i=0}^{S-1} (\vec{q})_i \psi_i(\mu);$$

5. define the *reduced* minimal rational approximation $\mathring{\vec{u}}$ as

$$\mathbb{C} \ni \mu \mapsto \mathring{\vec{u}}(\mu) = \frac{\mathrm{R}\,\mathrm{diag}\left([Q(\mu_1), \ldots, Q(\mu_S)]\right)}{Q(\mu)} \sum_{i=0}^{S-1} \left(V^{-\top}\right)_{:i} \psi_i(\mu) \in \mathbb{C}^S,$$

where $(\mathrm{A})_{:i}$ denotes the i-th column of matrix A; then the *full* minimal rational approximation $\widetilde{\vec{u}} \approx \vec{u}$ can be found as $\widetilde{\vec{u}}(\mu) = \mathrm{W}\mathring{\vec{u}}(\mu)$.

2.1 Greedy Approach

A common feature of all the techniques cited above is that a "sufficiently large" number of samples is needed to guarantee the accuracy of the surrogate model; in the particular case of frequency-domain problems, there exist lower bounds [3] for the number of samples required to achieve reasonable accuracy. Unfortunately, such number depends on the unknown spectral properties of A, and on the approximability of $\vec{f}$. For RB and MRI, one can identify adaptively the correct number of samples by relying on the so-called *greedy* algorithm, which can be summarized as follows:

1. Initialize a set $V = \{\vec{u}_1, \ldots, \vec{u}_{S_0}\}$ with some preliminary snapshots at $\mu_1, \ldots, \mu_{S_0}$.
2. Build a surrogate model (e.g., by MRI) based on V.
3. Choose a measure $r(\mu)$ of the discrepancy between exact and surrogate solution, and find its maximal point $\widehat{\mu}$: $r = r(\mu) \leq r(\widehat{\mu})$ for all μ.
4. If $r(\widehat{\mu})$ is smaller than a prescribed tolerance, terminate.
5. Compute a snapshot at $\widehat{\mu}$, add it to V, and go to 2.

The main difficulty in setting up the greedy algorithm is choosing a good r. Given the presence of resonances, it is standard [3] to use as a posteriori estimator the residual of (1), namely, given some suitable norm $\|\cdot\|_\star$,

$$r(\mu) = \|\mathrm{A}(\mu)\widetilde{\vec{u}}(\mu) - \vec{f}(\mu)\|_\star. \tag{3}$$

2.2 A Posteriori Indicators

In an intrusive framework, an efficient way to compute (3) has been known in the RB literature for quite a while, see e.g. [2], assuming $\vec{f}(\mu)$ to depend affinely on μ, i.e.

$$\vec{f}(\mu) = \sum_{i=0}^{N_{\vec{f}}-1} \theta_i(\mu)\, \vec{f}_i,$$

with $\vec{f}_i \in \mathbb{C}^n$ and $\theta_i : \mathbb{C} \to \mathbb{C}$ for all i. Then, as long as the matrices A_i, the vectors $\vec{f}_i$, the weights $\theta_i(\mu)$, and the reduced surrogate solution $\mathring{\vec{u}}(\mu)$ are available, we can evaluate the residual at μ as

$$r(\mu)^2 = \sum_{i,j=0}^{N_{\vec{f}}-1} \overline{\theta_i(\mu)}\theta_j(\mu)\langle \vec{f}_j, \vec{f}_i\rangle_\star + \mathring{\vec{u}}(\mu)^* \left(\sum_{i,j=0}^{2} \overline{\mu}^i \mu^j \langle \mathrm{A}_j \mathrm{W}, \mathrm{A}_i \mathrm{W}\rangle_\star \right) \mathring{\vec{u}}(\mu) - 2\mathrm{Re}\left(\left(\sum_{i=0}^{N_{\vec{f}}-1} \sum_{j=0}^{2} \overline{\theta_i(\mu)}\mu^j \langle \mathrm{A}_j \mathrm{W}, \vec{f}_i\rangle_\star \right) \mathring{\vec{u}}(\mu) \right) \tag{I}$$

in $O((S + N_{\vec{f}})^2)$ operations. This idea can be employed in MRI as well, at the cost of making the procedure intrusive. However, we propose here some alternatives.

In [7] it was observed that, if $\widetilde{\vec{u}}$ is the MRI of $\vec{u}$ with samples at $\{\mu_j\}_{j=1}^S$ and denominator Q, and both $\mathrm{A}(\mu)$ and $\vec{f}(\mu)$ depend at most linearly on μ (i.e. $\mathrm{A}_2 = 0$ and $\vec{f}(\mu) = \vec{f}_0 + \mu\vec{f}_1$) or μ^2 (i.e. $\mathrm{A}_1 = 0$ and $\vec{f}(\mu) = \vec{f}_0 + \mu^2\vec{f}_2$), then

$$r(\mu) = \frac{c}{|Q(\mu)|} \prod_{j=1}^{S} |\mu - \mu_j|, \tag{4}$$

with $c = c(\mu_1, \ldots, \mu_S, \mathrm{A}, \vec{f})$ independent of μ. In particular, since the location of the maximum of r does not depend on c, see (4), $\widehat{\mu}$ can be found even without knowing c. In order to determine the value of c non-intrusively, it is enough to compute r using (3) at a single new point μ' (in practice, we take $\mu' = \widehat{\mu}$):

$$r(\mu) = r(\mu') \left| \frac{Q(\mu')}{Q(\mu)} \right| \prod_{j=1}^{S} \left| \frac{\mu - \mu_j}{\mu' - \mu_j} \right|. \tag{R}$$

In certain situations, however, a direct evaluation of the quantity $r(\mu')$ within each greedy iteration might be impossible (e.g. if the solver used to evaluate $\vec{u}$ is a black-box that does not allow residual evaluation) or too computationally expensive. In order to have a viable greedy loop, we need to design an alternative termination condition in step 4. In this case, we propose to employ some heuristic indicator based on snapshot collinearity [3]: more explicitly, let W be the Q-factor in the QR factorization of the current snapshot matrix (2), and assume that the sample at $\widehat{\mu}$ has been computed; we opt to terminate the greedy algorithm if

$$\left\| \vec{u}(\widehat{\mu}) - \mathrm{W}\mathrm{W}^* \vec{u}(\widehat{\mu}) \right\| < \mathrm{tol} \, \|\vec{u}(\widehat{\mu})\| . \tag{C}$$

For this last indicator, as long as the greedy iterations continue, the extra snapshot does not go wasted, since it is precisely the one which gets added to V in step 5: on

the whole, this procedure computes only one "extra" snapshot, at the final greedy iteration, with respect to the two previous versions of the algorithm. Actually, one can adjust the greedy algorithm so as to employ even the extra snapshot in the final surrogate model: it suffices to build an updated MRI using all the samples, including the last one, once the termination condition has been satisfied.

We remark that the last two strategies rely on (4), which is valid only under some strong assumptions (linear dependence on the parameter) on A and $\vec{f}$. However, (4) can still be used for general parametric problems (1), and will give a reasonable estimation of the residual as long as $\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\mathrm{A}$ and $\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\vec{f}$ are small.

3 Numerical Examples

Here, through two practical examples, we showcase the usefulness of the greedy MRI procedure, as well as the effectiveness of the three termination strategies based on (I), (R), and (C).

3.1 An Eigenproblem in Magneto-Hydrodynamics

Take the generalized eigenproblem from [5]: $\mathrm{K}\vec{v} = \mu\mathrm{M}\vec{v}$ in $\mathbb{C}^n$, with $n = 4800$; it stems from modal analysis of a FE discretization of a dissipative problem in MHD. Here, we restrict our interest to the part of the spectrum with positive imaginary part: the eigenvalues are located on 3 so-called Alfvén branches around the branch point $\mu_b \approx -0.082 + 0.613\mathrm{i}$. Our aim is to approximate the number and location of eigenvalues around $\mu_0 = -0.175 + 0.5\mathrm{i}$; more precisely, we focus on the disk $D = \{\mu \in \mathbb{C} : |\mu - \mu_0| \leq 0.175\}$.

In order to cast this problem in the form (1), let $\vec{f} \in \mathbb{C}^n$ be a (normal Gaussian) random vector: we define the non-homogeneous problem

$$\text{find } \vec{u} : \mathbb{C} \to \mathbb{C}^n \quad \text{s.t.} \quad (\mathrm{K} - \mu\mathrm{M})\vec{u}(\mu) = \vec{f},$$

and build a surrogate for $\vec{u}$ using MRI. Then, our estimates for the eigenvalues will be the roots of the MRI denominator Q.

As a first MOR method, we apply greedy MRI: in particular, we employ indicator (I) with relative tolerance 10^{-2}, and the initial snapshots are at the $S_0 = 30$ shifted roots of unity[1] $\{\mu_0 - 0.175\mathrm{e}^{2\mathrm{i}\pi j/S_0}\}_{j=1}^{S_0}$.

[1] The shifted roots of unity are chosen as sample points because they allow for very stable and efficient interpolation schemes, relying on Fast Fourier Transform. We refer to [1] for a more detailed discussion of their properties.

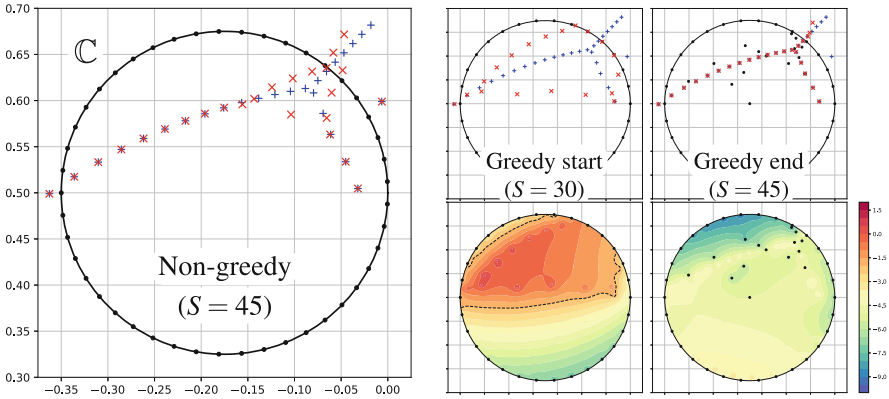


Fig. 1 Results of standard (left) and greedy (right) MOR. The exact and approximate eigenvalues are pluses and crosses, respectively, whereas the sample points are full dots. The contour plots show the logarithm of the greedy residual indicator; the dashed line represents the locus $\{\mu : r(\mu) = \text{tol}\}$, i.e. the boundary of the set where the prescribed tolerance is not satisfied

Additionally, we consider a non-greedy approach as a reference: we build an MRI starting from $S = 45$ samples at shifted roots of unity $\{\mu_0 - 0.175\mathrm{e}^{2\mathrm{i}\pi j/S}\}_{j=1}^{S}$. The number of samples is chosen so that, overall, the two methods employ exactly the same number of snapshots: the only difference is *where* the samples are taken.

The results are shown in Fig. 1. At the beginning of the greedy procedure (which corresponds to a standard MRI with $S = 30$), the spectrum is approximated quite poorly; this is correctly identified by the a posteriori indicator, which shows that the prescribed tolerance is not satisfied over a large portion of D. After 15 iterations of the greedy procedure, the residual is globally below the tolerance, and the algorithm ends. We can verify that all the eigenvalues in D are well captured.

In the standard approach with $S = 45$, most of the eigenvalues are well identified, but the quality of the approximation deteriorates around μ_b. In particular, among the two surrogates obtained with 45 snapshots, the greedy one is clearly superior. However, the improved accuracy of the greedy approach is accompanied by some risks:

- The locations of the greedy snapshots are close to the exact eigenvalues, since, according to the residual indicator (4), sampling there yields "the most information" for the surrogate model. However, sampling close to an eigenvalue may require solving numerically an ill-conditioned or even singular linear system.
- While the Vandermonde matrix for samples at the roots of unity has optimal condition number, adding new sample points at arbitrary locations is guaranteed to hinder the well-conditioning of the interpolation problem. Indeed, the residual indicator at the end of the greedy iterations shows a slightly unstable behavior near the bottom of D.

Table 1 Timing results of greedy MOR (average over 3 simulations with the same parameters for each method). All simulations were carried out on a single node of the Fidis cluster at EPFL [6]

Method	No. of snapshots	Average time per iteration		
		State solve	Indicator	Surrogate update
RB+(I)	24	97.1 s	15.0 s	3.2 s
MRI+(I)	23		4.31 s	2.9 s
MRI+(R)	23		1.04 s	
MRI+(C)	22(+1)		1.08 s	

3.2 *Frequency Response of a Waveguide Diplexer*

We consider a frequency response problem involving the FE discretization ($n = 90{,}258$) of a waveguide diplexer [3], for frequencies μ between 9.5 and 11 GHz. We are interested in approximating the scattering parameters

$$\mathrm{S} : \mathbb{C} \ni \mu \mapsto \mathrm{I} - 2\left(\mathrm{I} + \mathrm{i}\mu\sqrt{\frac{1-(\mu_c/\mu_0)^2}{1-(\mu_c/\mu)^2}}\mathrm{F}^* \underbrace{(\mathrm{K}-\mu^2\mathrm{M})^{-1}\mathrm{F}}_{\mathrm{U}(\mu)\in\mathbb{C}^{90{,}258\times 3}}\right)^{-1} \in \mathbb{C}^{3\times 3}, \qquad (5)$$

where we set $\mu_c = 6.56\,\mathrm{GHz}$, $\mu_0 = 10\,\mathrm{GHz}$, and the state matrix $\mathrm{U}(\mu)$ has one column for each port of the waveguide.

We build a surrogate for U using greedy RB and MRI, employing indicators (I) and (R) with relative tolerance 10^{-2}, and (C) with tol $= 10^{-4}$. The reduced tolerance for (C) can be justified by the considerably different nature of the indicator. To obtain an approximation of S, we just replace the exact state with the surrogate one in (5).

The results are summarized in Table 1 and visually depicted in Fig. 2. We remark that, by construction, the snapshot "history" of MRI is independent of the indicator, i.e. the parameter value $\widehat{\mu}$ which is selected at a given iteration is the same: the only effect of the choice of the indicator, besides timing, is the number of greedy iterations which are carried out before termination. In this regard, we observe that MRI+(I) and MRI+(R) yield exactly the same indicator, whereas MRI+(C) terminates one snapshot sooner,[2] causing some slight instability in the approximations of the scattering parameters for low frequencies, noticeable mostly in S_{13}.

A comparison of the surrogate S obtained by MRI+(I) and RB+(I) shows that the two methods yield very similar approximations, and reconstruct well the exact values. In fact, the approximated scattering parameters for RB are not included in Fig. 2, as they are almost indistinguishable from those obtained with MRI+(I).

[2] Here we are discarding the final extra snapshot used to check the termination condition (C). If it had been included, we would have recovered the same surrogate model as MRI+(I)/(R).

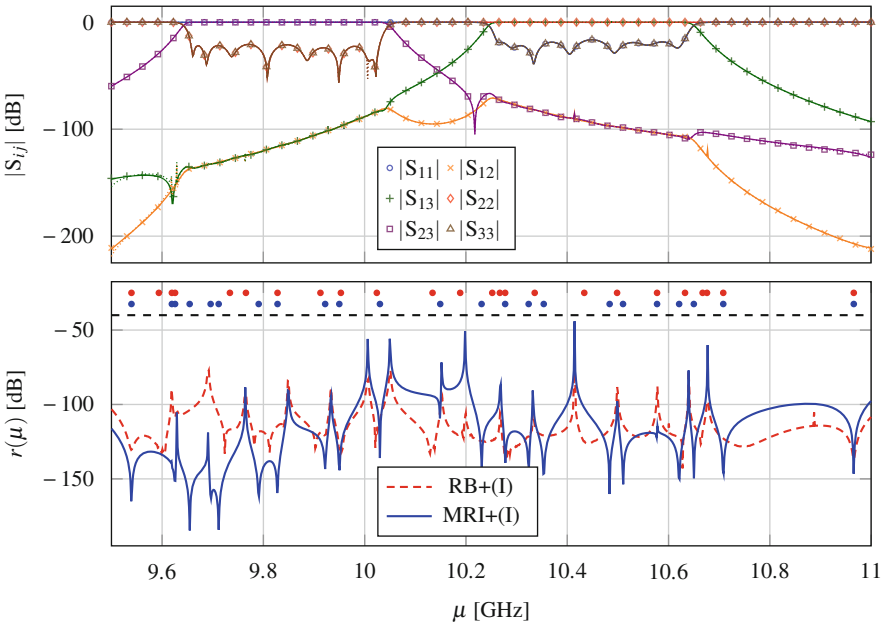


Fig. 2 Results of greedy MOR. On top the surrogate scattering parameters: the points are measurements from the original problem (5), whereas full and dotted lines are the surrogates obtained with MRI+(I) and MRI+(C), respectively. On the bottom the relative residual at the end of the greedy iterations for RB+(I) and MRI+(I); the points indicate the snapshot positions

However, RB requires one more snapshot, and the locations of the snapshots (and the residual profiles) for RB and MRI are quite different.

In terms of computing time, the efficiency of MRI seems quite remarkable, particularly for the two "least intrusive" indicators: the overhead time needed for the evaluation of indicators (R) and (C) is just a fraction of the time required for computation of (I) in RB.

4 Conclusions

We have presented several a posteriori indicators which can be employed in the greedy MRI algorithm, characterized by different degrees of intrusiveness and applicability. Good approximation properties, as well as a substantial speed-up in residual computation with respect to classical methods, have been observed in two numerical examples.

Acknowledgments This work has been funded by the Swiss National Science Foundation through FNS Research Project number 182236.

References

1. A.P. Austin, P. Kravanja, L.N. Trefethen, Numerical algorithms based on analytic function values at roots of unity. SIAM J. Numer. Anal. **52**, 1795–1821 (2014)
2. P. Benner, S. Gugercin, K. Willcox, A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Rev. **57**, 483–531 (2015)
3. V. De La Rubia, M. Mrozowski, A compact basis for reliable fast frequency sweep via the reduced-basis method. IEEE Trans. Microw. Theory Tech. **66**, 4367–4382 (2018)
4. A.C. Ionita, A.C. Antoulas, Data-driven parametrized model reduction in the loewner framework. SIAM J. Sci. Comput. **36**, A984–A1007 (2014)
5. M.N. Kooper, H.A. Van Der Vorst, S. Poedts, J.P. Goedbloed, Application of the implicitly updated Arnoldi method with a complex shift-and-invert strategy in MHD. J. Comput. Phys. **118**, 320–328 (1995)
6. SCITAS: EPFL Fidis cluster webpage. Online document (2020). http://www.epfl.ch/research/facilities/scitas/hardware/fidis. Cited 11 Feb 2020
7. D. Pradovera, Interpolatory rational model order reduction of parametric problems lacking uniform inf-sup stability. SIAM J. Numer. Analy. **58**(4), 2265–2293 (2020)

A Comparison Between Different Formulations for Solving Axisymmetric Time-Harmonic Electromagnetic Wave Problems

Erik Schnaubelt, Nicolas Marsic, and Herbert De Gersem

Abstract In many time-harmonic electromagnetic wave problems, the considered geometry exhibits an axial symmetry. In this case, by exploiting a Fourier expansion along the azimuthal direction, fully three-dimensional (3D) calculations can be carried out on a two-dimensional (2D) angular cross section of the problem, thus significantly reducing the computational effort. However, the transition from a full 3D problem to a 2D analysis introduces additional difficulties such as, among others, a singularity in the variational formulation. In this work, we compare and discuss different finite element formulations to deal with these obstacles. Particular attention is paid to spurious modes and to the convergence behavior when using high-order elements.

1 Introduction

When treating a problem exhibiting axial symmetry, a Fourier expansion along the azimuthal direction can be exploited in order to restrict the computation to a two-dimensional (2D) angular cross section of the geometry, while still considering a fully three-dimensional (3D) solution [1]. Therefore, these methods are also referred to as *quasi-3D* or *2.5D* methods. Let us consider a cylindrical coordinate system (r, φ, z), and let us expand the electric field $\vec{e}(r, \varphi, z)$ into a Fourier series along φ:

$$\vec{e}(r,\varphi,z) = \begin{bmatrix} e_r^0(r,z) \\ e_\varphi^0(r,z) \\ e_z^0(r,z) \end{bmatrix} + \sum_{m=1}^{\infty} \left(\begin{bmatrix} e_r^m(r,z)\cos(m\varphi) \\ e_\varphi^m(r,z)\sin(m\varphi) \\ e_z^m(r,z)\cos(m\varphi) \end{bmatrix} + \begin{bmatrix} e_r^{-m}(r,z)\sin(m\varphi) \\ e_\varphi^{-m}(r,z)\cos(m\varphi) \\ e_z^{-m}(r,z)\sin(m\varphi) \end{bmatrix} \right),$$

E. Schnaubelt (✉) · N. Marsic · H. De Gersem
Technische Universität Darmstadt, Institut für Teilchenbeschleunigung und Elektromagnetische Felder (TEMF), Darmstadt, Germany
e-mail: erik.schnaubelt@cern.ch; marsic@temf.tu-darmstadt.de; degersem@temf.tu-darmstadt.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_17

where the Fourier coefficients $\vec{e}^n(r,z) = \left[e_r^n, e_\varphi^n, e_z^n\right]^T$ with $n \in \mathbb{Z}$ are functions of the radial and axial coordinates *only*. Furthermore, by exploiting the orthogonality of the trigonometric functions, we can write the Maxwell eigenvalue problem for an axisymmetric cavity V with perfect electric conducting boundaries as [1]:

$$\begin{cases} \text{For a given mode } n \in \mathbb{Z}, \text{ find the eigenpairs } (\vec{e}^n, \omega^2) \text{ with } \vec{e}^n \in \mathscr{S}^n(\Omega): \\ \displaystyle\int_\Omega \mu_r^{-1}\, \mathbf{curl}_n\, \vec{e}^n \cdot \mathbf{curl}_n\, \vec{e}^{n'}\, \mathrm{d}\Omega - \frac{\omega^2}{c_0^2} \int_\Omega \varepsilon_r\, \vec{e}^n \cdot \vec{e}^{n'}\, \mathrm{d}\Omega = 0 \quad \forall \vec{e}^{n'} \in \mathscr{S}^n(\Omega), \end{cases} \tag{1}$$

with $\varepsilon_r = \varepsilon_r(r,z)$ and $\mu_r = \mu_r(r,z)$ the scalar relative electric permittivity and magnetic permeability of the medium, Ω a 2D angular cross section of V, $\mathrm{d}\Omega = r\,\mathrm{d}r\,\mathrm{d}z$, ω the angular frequency, c_0 the speed of light in vacuum, $\mathscr{S}^n(\Omega)$ the function space of the nth Fourier coefficient and

$$\mathbf{curl}_n\, \vec{e}^n = \begin{bmatrix} -r^{-1}\left(n e_z^n + \partial_z (r e_\varphi^n)\right) \\ \partial_z e_r^n - \partial_r e_z^n \\ +r^{-1}\left(n e_r^n + \partial_r (r e_\varphi^n)\right) \end{bmatrix}.$$

2 Well-Posed Variational Formulation

In order to construct an appropriate subspace of $\mathscr{S}^n(\Omega)$ and in order to account for the singular behavior of $\mathbf{curl}_n$ at $r = 0$, two strategies have been proposed in the literature.

2.1 Non-classical Conditions Along the Symmetry Axis

A first approach consists in taking the unknown fields $e_\varphi^{\star,n} = r e_\varphi^n \in H^1(\Omega)$ and $\vec{e}_{rz}^n = \left[e_r^n, e_z^n\right]^T \in H(\mathbf{curl}, \Omega)$ [2] together with *non-classical discrete conditions at the symmetry axis* [3, Section 4.4]. By following this strategy, all integrals are well-posed but exhibit singular integrands, hence requiring either *i)* a classical Gaussian quadrature with a large number of quadrature points or *ii)* specialized quadrature rules [3, Section 5.1] which differ from element to element, thus preventing the use of fast assembly techniques [4]. In what follows, this approach will be further referred to as transformation "TA".

2.2 Direct Construction of a Subspace of $\mathscr{S}^n(\Omega)$

Another approach consists in *directly constructing an appropriate subspace of* $\mathscr{S}^n(\Omega)$ such that the variational formulation is guaranteed to be always well-posed, as shown in [1, 5–7] for instance, thus avoiding the need for non-classical conditions on the symmetry axis. To this end, the unknowns e_φ^n and $\vec{e}_{rz}^{\,n}$ are transformed into $u^n \in H^1(\Omega)$ and $\vec{U}^n \in H(\mathbf{curl}, \Omega)$ by following the methodology shown in Table 1, with $\mathbf{grad}_{rz}\, e_\varphi^n = \left[\partial_r e_\varphi^n, \partial_z e_\varphi^n\right]^T$ and $\hat{r}$ the unit vector along the r-axis.

The parameters α and β in TC(α, β) must satisfy, according to [6], the constraints shown in Table 2. Furthermore, some transformations need an additional *homogeneous Dirichlet condition* at $r = 0$, as shown in Table 3. Finally, for appropriate choices of α and β, TC(α, β) leads to *polynomial* integrands (see Sect. 3.2 for more details). This property is also met by TB for $n \neq 0$ and TD for $n = \pm 1$.

Table 1 Different transformations for constructing a subspace of $\mathscr{S}^n(\Omega)$

Mode	Transf. TB [7]	Transf. TC(α, β) [6, Section 1.3]	Transf. TD [5]
$n = 0$	$u^0 = e_\varphi^0$	$r^\beta u^0 = r e_\varphi^0$	$u^0 = e_\varphi^0$
	$\vec{U}^0 = \vec{e}_{rz}^{\,0}$	$\vec{U}^0 = \vec{e}_{rz}^{\,0}$	$\vec{U}^0 = \vec{e}_{rz}^{\,0}$
$n = \pm 1$	$u^{\pm 1} = e_\varphi^{\pm 1}$	$r^\beta u^{\pm 1} = r e_\varphi^{\pm 1}$	$u^{\pm 1} = e_\varphi^{\pm 1}$
	$\vec{U}^{\pm 1} = \frac{n}{r}\vec{e}_{rz}^{\,\pm 1} + \frac{e_\varphi^{\pm 1}}{r}\hat{r}$	$r^\alpha \vec{U}^{\pm 1} = \pm\vec{e}_{rz}^{\,\pm 1} + \mathbf{grad}_{rz}(r e_\varphi^{\pm 1})$	$\vec{U}^{\pm 1} = \frac{n}{r}\vec{e}_{rz}^{\,\pm 1} + \frac{e_\varphi^{\pm 1}}{r}\hat{r}$
$\lvert n\rvert > 1$	$u^n = e_\varphi^n$	$r^\beta u^n = r e_\varphi^n$	$u^n = e_\varphi^n$
	$\vec{U}^n = \frac{n}{r}\vec{e}_{rz}^{\,n} + \frac{e_\varphi^n}{r}\hat{r}$	$r^\alpha \vec{U}^n = n\vec{e}_{rz}^{\,n} + \mathbf{grad}_{rz}(r e_\varphi^n)$	$\vec{U}^n = \frac{n}{r}\vec{e}_{rz}^{\,n}$

Table 2 Constraints on α and β for TC(α, β) according to [6, Section 1.5]

$n = 0$	$n = \pm 1$	$\lvert n\rvert > 1$
$\beta \geq 0.5$	$\alpha \geq 0.5$ and $\beta = 1$	$\alpha \geq 0.5$ and $\beta > 0$

Table 3 Conditions on the symmetry axis

Mode	Transf. TB [7]	Transf. TC(α, β) [6, Section 1.5]	Transf. TD [5]
$n = 0$	$u^0 = 0$	$u^0 = 0$ if $\beta \in [0.5, 1.5[$, none otherwise	$u^0 = 0$
$n = \pm 1$	None	None	None
$\lvert n\rvert > 1$	$u^n = 0$	$u^n = 0$ if $\beta \in]0, 1]$, none otherwise	$u^n = 0$

3 Comparison and Discussion of the Quasi-3D Methods

As already stated, this work compares the aforementioned transformations to treat the eigenvalue problem (1). To this end, they were implemented in a homemade high-order finite element (FE) code.[1] All following numerical experiments are performed on a pillbox cavity with radius $r = 150\,\text{mm}$, height $h = 294\,\text{mm}$, and $\mu_r = \varepsilon_r = 1$, for which closed-form solutions are well-known [8, Sections 1.13 and 1.14]. A structured triangular mesh is used which is refined by uniformly splitting each triangle into four subtriangles.

3.1 Spurious Modes and High-Order FE Discretizations

Let us start our comparison by determining if the methods discussed previously can avoid spurious modes. As we search the azimuthal unknown ($e_\varphi^{*,n}$ for TA and u^n for TB, TC and TD) in a *finite* subspace[2] of $H^1(\Omega)$ of polynomial order q and the in-plane unknown ($\vec{e}_{rz}$ for TA and $\vec{U}^n$ for TB, TC and TD) in a *finite* subspace (see footnote 2) of $H(\mathbf{curl}, \Omega)$ of polynomial order p, the dimension of each subspace must be selected with care. In particular, in order to satisfy the *exactness of the discrete de Rham sequence* [10], one must impose that $q = p + 1$ [11].

In order to validate this choice, we ran multiple numerical tests with the different transformations, different modes n and different values for p and q. As a result, we observed that, *apart from TD*, all eigenspectra were free of spurious modes when $q = p + 1$. Interestingly, we also observed no spurious modes when $q > p + 1$. On the other hand, spurious modes were systematically observed when $q < p + 1$, and when transformation TD was used with $|n| > 1$ (for all possible values of p and q). For this reason, TD will not be investigated further. As an illustration, Fig. 1 shows a part of the numerical spectrum of a pillbox cavity for $n = 1$ and different mesh densities. It was computed with TB, once for $q = 3$, $p = 2$ and once for $q = p = 2$.

When $n = 0$, the in-plane and azimuthal unknowns are *decoupled* from each other [6, Section 1.6]. Therefore, q and p can be chosen *independently*.

3.2 Convergence Results for Higher Order Finite Elements

In this subsection, the convergence behavior of the different formulations in combination with high-order basis functions will be compared. As an example, the

[1] See https://gitlab.onelab.info/gmsh/small_fem/blob/master/simulation/Quasi3D.cpp.

[2] In this paper, a *finite* subspace of $H^1(\Omega)$ (resp. $H(\mathbf{curl}, \Omega)$) is built using *grad-conforming* (resp. *curl-conforming*) finite elements from [9, Chapter 4.5]. In particular, we consider complete subspaces of $H(\mathbf{curl}, \Omega)$ with both irrotational and rotational functions.

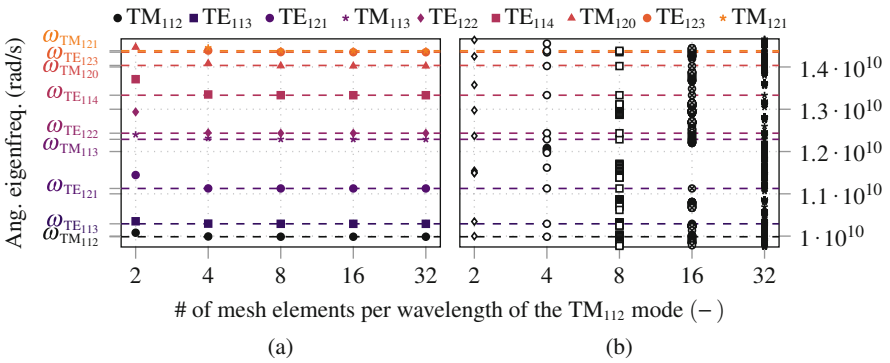


Fig. 1 Part of the spectrum of a pillbox cavity obtained with TB and $n = 1$. (**a**) Polynomial order $q = 3$, $p = 2$. (**b**) Polynomial order $q = 3$, $p = 3$

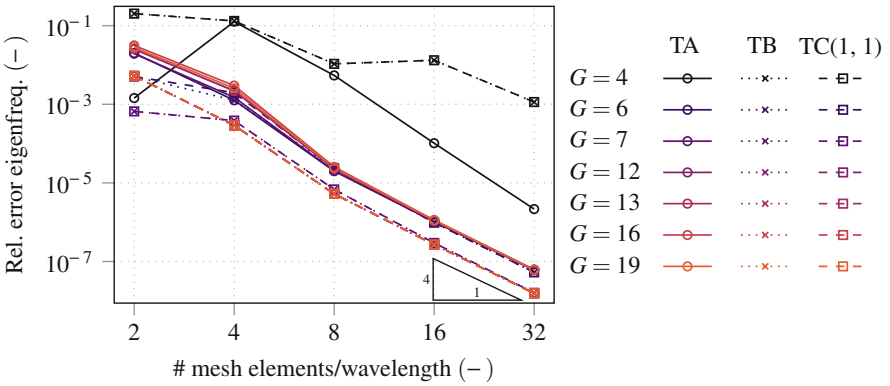


Fig. 2 Convergence results when computing the eigenfrequency of the TE_{111} mode of a pillbox cavity with TA, TB and TC(1, 1), using $q = 3$, $p = 2$ and G Gauss-Legendre quadrature points

evolution of the relative error between the numerically computed eigenfrequency and its analytical counterpart is shown in Fig. 2 for different mesh densities and different numbers of Gauss-Legendre quadrature points G. A second-order FE method with $q = 3$, $p = 2$ is used, hence resulting in an expected convergence slope of 4 [12]. This slope is indeed achieved for TA, TB and TC(1, 1) once the number of quadrature points passes a certain threshold. Again, this is not an isolated case but can be observed for all n and for different element orders.

Let us also stress that TB and TC(1, 1) *(i)* yield a lower relative error than TA and *(ii)* depend less on G than TA. This last observation can be easily explained: as TB and TC(1, 1) lead to *polynomial integrands*, the final solution is independent of G, at least for a G sufficiently large to integrate a polynomial of the given order *exactly*.

The parameters of TC(α, β) need to meet the following criteria to yield polynomial integrands in the variational formulation (1): *(i)* α and β must be *multiples*

of 0.5, *(ii)* their sum must be an *integer* and *(iii)* $\beta \geq 1.5$ for $n = 0$ and $\alpha \geq 0.5, \beta \geq 0.5$ for $n \neq 0$.

3.3 *Influence of α and β on the Convergence Behavior*

Let us now focus on the transformation TC(α, β), and let us carry out a convergence test similar to the previous section. However, now, the influence of the parameters α and β (chosen according to Table 2) on the convergence rate will be investigated. The results of this numerical experiment are displayed in Fig. 3. As it can be observed directly, while all choices converge towards the sought eigenvalue, only particular pairs (α, β) exhibit the expected convergence rate. This behavior has been observed for other choices of (n, p, q) with $q = p + 1$ as well.

This behavior can be easily explained if we assume that $\vec{e}^n \in C^\infty$ in the vicinity of the symmetry axis. This assumption is of course restrictive, but applies to the pillbox cavity [8], and gives already a good insight into the underlying numerical mechanisms. In what follows, only the case $n = \pm 1$ will be discussed, but the same methodology applies to the other cases.

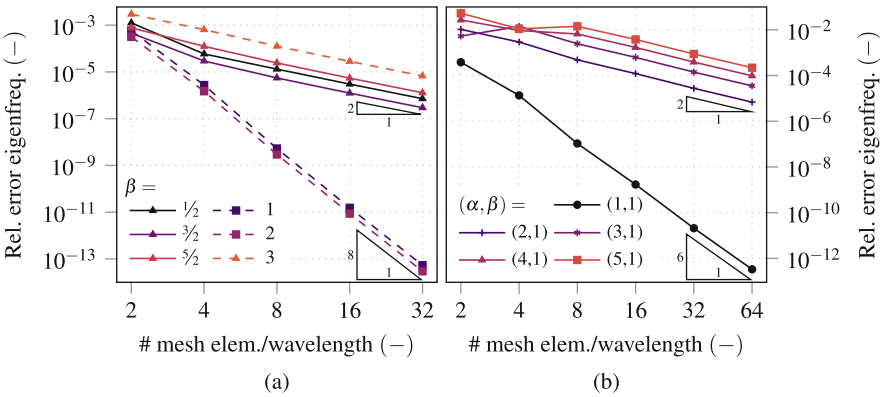


Fig. 3 Convergence rate of TC(α, β) for different values of α and β, as allowed by Table 2. The symbol "$*$" in (**a**) indicates that the result is independent of p due to the decoupling of the in-plane and azimuthal unknowns (see Sect. 3.1). (**a**) TM_{111} mode with $q = 4$, $p = 3$. (**b**) TE_{022} mode with $q = 4$, $p = *$

Let us start by expanding $\vec{e}^{\pm 1}$ into a Taylor series in the vicinity of $r = 0$ and $z = z_0$. As $e_z^{\pm 1} = 0$ at $r = 0$ (see [1]), we have:

$$\begin{cases} e_\varphi^{\pm 1}(r,z) = a_0 + a_1^r r + a_1^z (z - z_0) + a_2^{rr} r^2 + a_2^{zz} (z - z_0)^2 + 2a_2^{rz} r (z - z_0) + \dots, \\ e_z^{\pm 1}(r,z) = b_1^r r + b_2^{rr} r^2 + 2b_2^{rz} r (z - z_0) + \dots, \\ e_r^{\pm 1}(r,z) = c_0 + c_1^r r + c_1^z (z - z_0) + c_2^{rr} r^2 + c_2^{zz} (z - z_0)^2 + 2c_2^{rz} r (z - z_0) + \dots . \end{cases} \tag{2}$$

Then, by exploiting the definition of $\vec{U}^{\pm 1}$ (see Table 1), we can write:

$$\begin{cases} r^\alpha U_r^{\pm 1} = \pm e_r^{\pm 1} + e_\varphi^{\pm 1} + r(a_1^r + 2a_2^{rr} r + 2a_2^{rz}(z - z_0) + \dots), \\ r^\alpha U_z^{\pm 1} = \pm e_z^{\pm 1} + \ 0 \ + r(a_1^z + 2a_2^{zz}(z - z_0) + 2a_2^{rz} r + \dots). \end{cases} \tag{3}$$

Restricting the further analysis to the axial component, we then have that:

$$\begin{aligned} U_z^{\pm 1} &\overset{(3)}{=} r^{-\alpha} \Big[\pm e_z^{\pm 1} + r(a_1^z + 2a_2^{zz}(z - z_0) + 2a_2^{rz} r + \dots) \Big], \\ &\overset{(2)}{=} r^{-\alpha} \Big[r \left(b_1^r + b_2^{rr} r + 2b_2^{rz}(z - z_0) + \dots + a_1^z + 2a_2^{zz}(z - z_0) + 2a_2^{rz} r + \dots \right) \Big], \\ &\overset{\text{def}}{=} r^{1-\alpha} f(r,z), \end{aligned}$$

where f is a polynomial function of r and z. As $r \to 0$, we have that $f(r,z) \to f(0,z)$. Thus, the unknown $U_z^{\pm 1} \sim r^{1-\alpha}$ as $r \to 0$. Analogously, using $e_r^{\pm 1} \pm e_\varphi^{\pm 1} = 0$ at $r = 0$ [1], we find that $U_r^{\pm 1} \sim r^{1-\alpha}$ as $r \to 0$. Therefore, for the sought FE solution to be differentiable at $r = 0$, and because of the constraint enforced by Table 2 ($\alpha \geq 0.5$), *we need to impose that* $\alpha = 1$. Any other choice (in accordance with Table 2) will lead to a non-differentiable $\vec{U}^n$, jeopardizing thus the convergence of the FE scheme. By applying the same strategy to the cases $n \neq \pm 1$, and by taking into account the restrictions imposed in Table 2, the set of allowed couples (α, β) must be further narrowed, as shown in Table 4. To the best of our knowledge, these last results have been derived for the first time.

For the integrands of the variational formulation (1) to be polynomial, we further need to impose the following restriction: $\beta = 2$ when $n = 0$ (see Sect. 3.2). Moreover, as the number of quadrature points depends on the order of the integrands, it is preferable to select the smallest acceptable (α, β) couple. Therefore, the values given in Table 5 are recommended. Let us finally note that, apart from the case $n = 0$, these recommendations are in accordance with [6, Section 1.6], where the best (α, β) couples were determined on the basis of numerical experiments.

Table 4 Values of α and β for TC(α, β) leading to a high FE convergence rate

$n = 0$	$n = \pm 1$	$\lvert n \rvert > 1$
$\beta \in \{1, 2\}$	$\alpha = 1$ and $\beta = 1$	$\alpha \in \{1, 2\}$ and $\beta \in \{1, 2\}$

Table 5 Recommended choice of α and β for TC(α, β)

$n = 0$	$n = \pm 1$	$\lvert n\rvert > 1$
$\beta = 2$	$\alpha = 1$ and $\beta = 1$	$\alpha = 1$ and $\beta = 1$

Concerning the case $n = 0$, the choice $\beta = 1$ leads to slightly more accurate results for a given mesh density in a numerical experiment carried out in [6, Section 1.6]. However, our numerical experiments do not confirm these results as the choice $\beta = 2$ leads to a slightly lower relative error for the same mesh density (see Fig. 3a). This behavior is supported by the fact that the integrands are polynomial for the latter choice $\beta = 2$.

4 Conclusion

This paper compared four different transformations to treat three-dimensional time-harmonic electromagnetic wave problems in axisymmetric geometries proposed in the literature. We first determined numerically that the transformations TA, TB and TC(α, β) lead to eigenvalue problems which are free of spurious modes, while the transformation TD exhibits spurious modes when $|n| > 1$. We then compared numerically the accuracy of TA, TB and TC(α, β), and found that TB and TC(α, β) produce the most accurate results for a given mesh density. Finally, we analyzed theoretically the convergence rate of TC(α, β) for different values of α and β in a high-order FE context, and determined new restrictions on α and β.

Acknowledgments The authors would like to express their gratitude to Abele Simona for his valuable advice and the fruitful discussions on axisymmetric problems.

References

1. P. Lacoste, Solution of Maxwell equation in axisymmetric geometry by Fourier series decompostion and by use of H(rot) conforming finite element. Numer. Math. **84**(4), 577–609 (2000)
2. J.F. Lee, G.M. Wilkins, R. Mitra, Finite-element analysis of axisymmetric cavity resonator using a hybrid edge element technique. IEEE Trans. Microwave Theory Techn. **41**(11), 1981–1987 (1993)
3. O. Chinellato, The complex-symmetric Jacobi-Davidson algorithm and its application to the computation of some resonance frequencies of anisotropic lossy axisymmetric cavities. Ph.D. Thesis, Eidgenössische Technische Hochschule Zürich, 2005
4. N. Marsic, C. Geuzaine, Efficient finite element assembly of high order Whitney forms. IET Sci. Meas. Technol. **9**(2), 204–210 (2015)
5. E.A. Dunn, J.K. Byun, E.D. Branch, J.M. Jin, Numerical simulation of BOR scattering and radiation using a higher order FEM. IEEE Trans. Antennas Propag. **54**(3), 945–952 (2006)

6. S. Cambon, Méthode d'éléments finis d'ordre élevé et d'équations intégrales pour la résolution de problème de furtivité radar d'objets à symétrie de révolution. Ph.D. Thesis, Institut National des Sciences Appliquées de Toulouse, 2012
7. M. Oh, de Rham complexes arising from Fourier finite element methods in axisymmetric domains. Comput. Math. Appl. **70**(8), 2063–2073 (2015)
8. T.P. Wangler, *RF Linear Accelerators*. 2nd, Completely Revised and Enlarged edition. (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2008)
9. S. Zaglmayr, High Order Finite Element Methods for Electromagnetic Field Computation. Ph.D. Thesis, Johannes Kepler Universität Linz, 2006
10. A. Simona, L. Bonaventura, C. de Falco, S. Schöps, Isogeometric approximations for electromagnetic problems in axisymmetric domains (2019). arXiv preprint arXiv:1912.08570
11. L. Demkowicz, P. Monk, L. Vardapetyan, W. Rachowicz, de Rham diagram for hp finite element spaces. Comput. Math. Appli. **39**(7–8), 29–38 (2000)
12. S. Sauter, hp-finite elements for elliptic eigenvalue problems: error estimates which are explicit with respect to λ, h, and p. SIAM J. Numer. Analy. **48**(1), 95–108 (2010)

The Magnetization Analysis of Motor Magnet and Its Influence on Cogging Torque

Chenxi Wang, Matthias Willig, Stefan Kurz, and Kevin Gutmann

Abstract In this article, a simulation procedure of a BLDC motor is described that includes the magnetization process of the magnet poles. This leads to more realistic magnetic field distributions in the motor during the cogging torque analysis. The process of the magnetization calculation as well as the handling of the material properties is explained and the influence of isotropic and anisotropic material definition on the results is shown.

1 Introduction

The acoustic performance is one of the most important indicators to evaluate the comfort of automobile, therefore, the acoustics of electric machines as a common device in vehicles is a critical point that needs to be considered in the machine design process.

However, the increasing precision requirements for the accuracy of prediction of noise-exciting forces of electric machines pose a significant challenge to the assumptions and idealizations applied in motor design process today. In order to meet higher precision requirements, it is necessary to adapt influences from statistical geometry variations, material fluctuations and the manufacturing process of the machine. As an increasingly used electric machine in automobile, the acoustic and vibration performance of the BLDC motor (Brushless Direct Current Motor) is an important quality indicator of the machine. In the no load operation, it is mainly determined by its cogging torque, which is highly influenced by the motor geometry and magnetic field in the air gap [1–3]. Hence, instead of using an idealized magnetic field, a more accurate and realistic description of the magnetic field of

C. Wang · S. Kurz
Technical University Darmstadt, Darmstadt, Hessen, Germany
e-mail: stefan.m.kurz@jyu.fi

M. Willig · K. Gutmann (✉)
Robert Bosch GmbH, Buehl, Baden-Wuerttemberg, Germany
e-mail: Matthias.Willig@de.bosch.com; Kevin.Gutmann@de.bosch.com

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_18

magnet needs to be taken into consideration, in order to achieve a higher accuracy of the cogging torque calculation [1, 4].

2 BLDC Motor and Cogging Torque

The motor considered in this analysis is a BLDC motor in outer rotor configuration. The topology of the motor is 12/8 i.e. the stator has 12 slots and the rotor has 8 magnet poles. In the motor used for this analysis, these 8 poles are magnetized on a magnet ring as indicated by the vectors of the magnetic flux density B in Fig. 1.

The generation of cogging torque in the motor is determined by the interaction of the magnet poles of the rotor and the slots of the stator. Based on the co-energy in the system, the calculation of the motor cogging torque is given by the following equation, which is the fundament of the torque calculation in the FEA-tool [2, 3].

$$T = \frac{\mathrm{d}W}{\mathrm{d}\theta} = \frac{\partial}{\partial\theta}\left[\int_V \left(\int_0^{H_B} B(H)\mathrm{d}H\right)\mathrm{d}V\right], \tag{1}$$

where W is the magnetic coenergy, θ is the rotor position, H the magnetic field, B is the flux density, H_B is the magnetic field in operating point, V the integration volume and T the calculated torque. For the applied method of virtual work, the change in the coenergy of the system (and therefore the virtual torque) is given by the change in the coenergy of the virtually distorted finite elements.

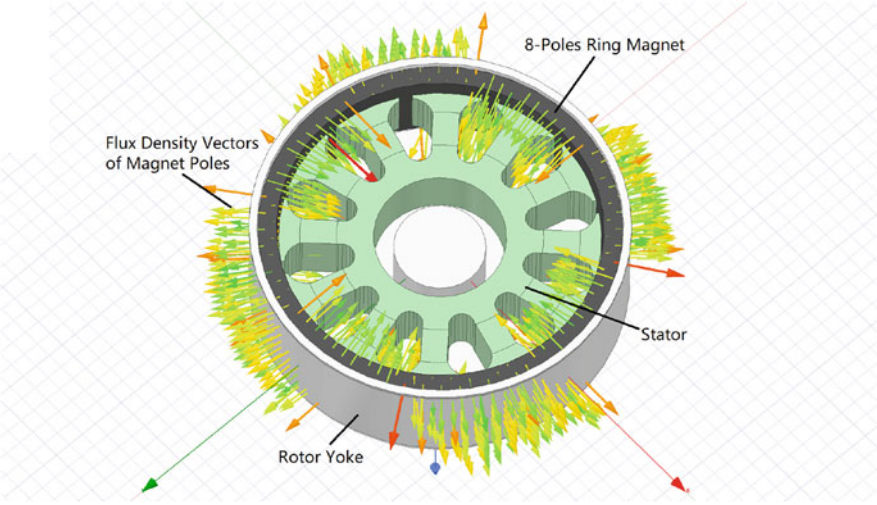


Fig. 1 Model structure of the BLDC motor

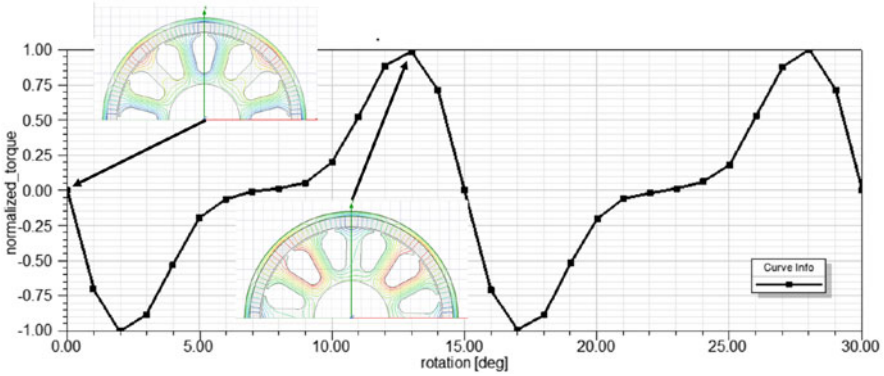


Fig. 2 The cogging torque of BLDC motor with 12 slots and 8 poles

The fundamental of the cogging torque of a 12/8 motor topology is given by the least common multiple [2, 5] of 12 and 8 which is the 24th mechanical order. Figure 2 shows a typical cogging torque curve of the motor. The distribution of magnetic flux for minimum and maximum of rotor position dependent no-load torque is also shown.

One can observe that the no load torque at different rotor positions depends on the flux distribution in the motor when there is a magnetic field imposed by the permanent magnet poles. This flux distribution is mainly determined by the magnetization of the permanent magnet poles and the shape of the magnetic circuit. A more realistic representation of the magnetization leads to a more precise prediction of cogging torque of the electric machine.

3 Definition of the Magnet Material in Analysis

3.1 Magnetization Curve of the Magnet

By using the 3D FEA-tool to simulate the magnetization process, the input data required for this method is the magnetization curve in the first quadrant of $B - H$ coordinate system, as shown in Fig. 3. This curve was measured using a standard Permagraph measurement method and extrapolated to the point of the maximum excitation field.

Technical ferrites are usually manufactured to have a main axis for the preferred direction of the magnetic flux. In this case the magnetization process not only depends on the magnetizing field but also on the axis in which the field is active. Those effects of anisotropy and isotropy on the simulation results will be shown later in this paper.

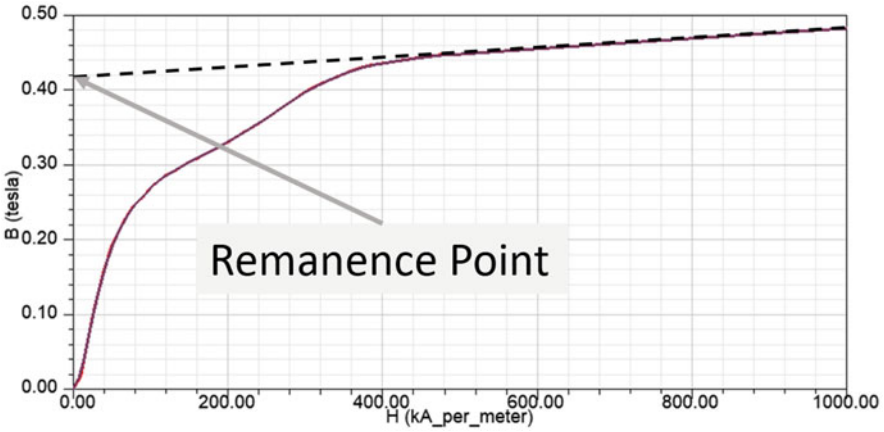


Fig. 3 The measured virgin curve of the analyzed magnet

3.2 Approximation Method to Describe the Hysteresis Effect

With the defined magnetization curve, the hysteresis effect of the magnet material needs to be introduced in analysis, in order to evaluate the remanence of the magnet after magnetization. However due to the limitations of the FEA-tool used, the commonly used hysteresis models like Preisach model and Jiles Atherton model are not supported in the simulation. Instead, the software supports method, which is called "classic approach" or "linear approach" to describe the hysteresis [6].

The linear approach is based on the approximated linearity in the descending branch of the hysteresis loop of magnet. When the magnetization field is removed, the magnetic polarization J descends with constant slope, which is identical to the slope in the saturated region [7], as indicated by the dashed line in the Fig. 3. The intersection of the descending line and vertical curve is the remanence point, which is around 415mT in the analyzed magnet. The slope of the descending part is equal to the saturated permeability, which is $\mu_0\mu_m$ for the curve of the magnetic flux density B, and $\mu_0(\mu_m - 1)$ for the curve of magnetic polarization J. For the ferrite magnet used in this motor, the relative permeability in saturated region is $\mu_m \approx 1.05$.

In addition, because of the linearity of the descending branch, only the operating point with maximum excitation field is necessary to be simulated. Hence, the magnetization analysis in the FEA-tool can be significantly simplified to a single magnetostatic simulation with the maximum excitation field.

4 Modelling of the Magnetization Device and Motor

Both the magnetization and cogging torque analyses are executed in the 3D environment in FEA-tool. The ring magnet is magnetized in the magnetization analysis first and then its remanent field will be transmitted into the motor model to calculate the cogging torque.

A full 3D model of the magnetization unit is built and the magnet inserted as shown in Fig. 4, where a 90° slice of the whole model is depicted.

The excitation current in the windings is generated by the discharge of a connected capacitor. However, in the analysis only the peak value of the current impulse is needed when the linear approach is used. Moreover, the possible effects of eddy currents due to the transient current impulse are not considered in this analysis.

Figure 5 shows a 2D magnetic field distribution in the middle cross section of the model, with the maximum excitation current applied. It is obvious that the magnet is not uniformly magnetized due to the field distribution imposed by the magnetizing unit. There are different areas in the magnet (particularly areas close to the pole transition zones) that reach different parts of the virgin curve and therefore will have different remanent inductions after the magnetizing field is removed.

After the magnetization analysis the magnetized magnet is available for cogging torque analysis of the motor via an internal datalink in the FEA software. The motor model is built up as in Fig. 1. The magnetization of the magnet in the motor is identical to the magnetization achieved in the magnetization calculation.

Since the cogging torque is evaluated at no load conditions the coils of the motor are omitted in the analysis to reduce the number of finite elements and therefore computing time. Moreover, due to the symmetry of the motor model, only a section of 30° is necessary to be analyzed.

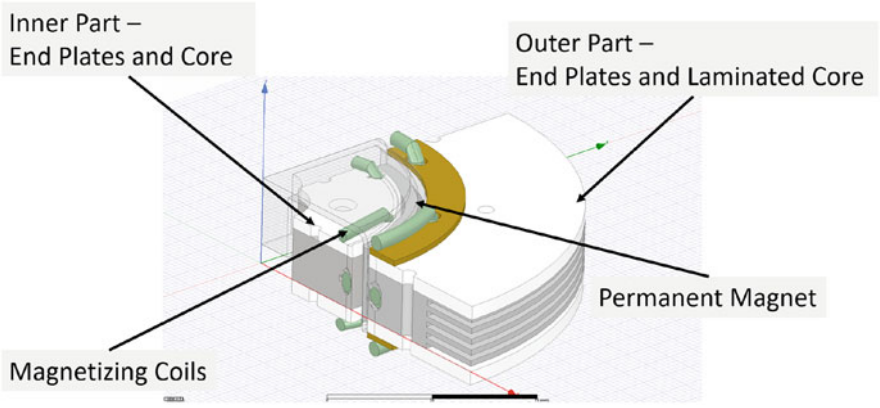


Fig. 4 Model structure of the magnetization device

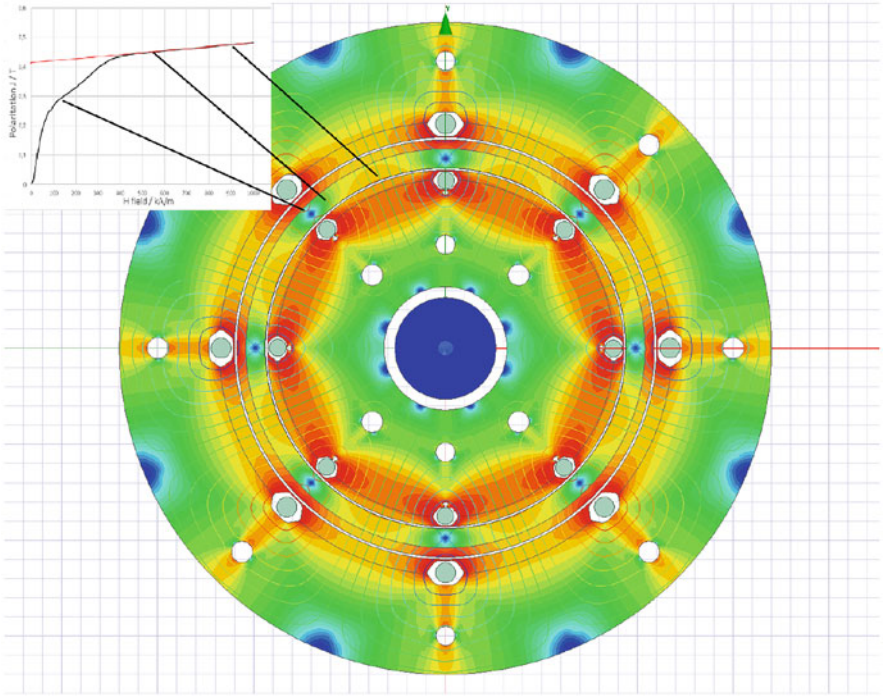


Fig. 5 The B field distribution in middle section with maximum excitation current

5 Result of Analysis

The motor cogging torque was calculated for three different magnet models:

1. A calculated magnetization profile based on isotropic magnet material.
2. A calculated magnetization profile based on anisotropic magnet material.
3. An ideal magnetization profile.

The 3D FEA solver of the tool applied uses tetrahedral mesh elements. A fine mesh was applied to the magnet in the magnetization as well as the motor analysis. In each model a mesh of approximately 60k tetrahedra for the magnet was achieved and a maximum energy error of 0.5% for the whole system was set as solver criterion.

In the isotropic magnet model, the pre-defined magnetization curve is valid for all magnetization directions, but only valid in the radial direction for the anisotropic model. The response of the materials to a magnetizing field in the simulation can be seen in Fig. 6. Whereas the resulting magnetization of the isotropic material is a vector that is parallel to the applied magnetizing field, the resulting magnetization of the anisotropic material is a vector representing the component of the applied magnetizing field in the preferred direction of the material.

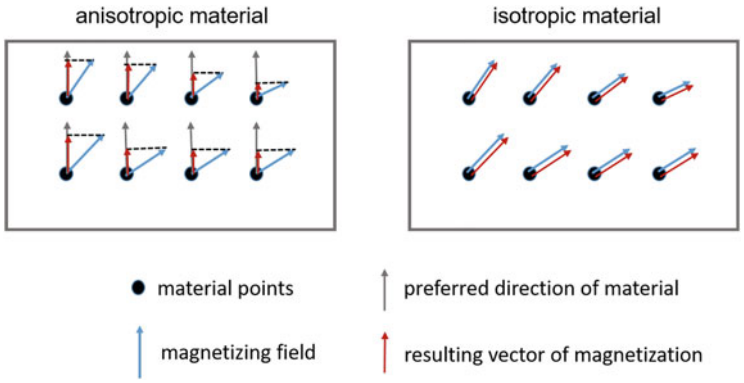


Fig. 6 Material response to magnetizing field in simulation

Compared to the calculated magnetization fields in the last section, the ideal field is defined as a purely radial and homogeneous magnetic field with constant magnitude. The magnetic fields of these three cases are shown in Fig. 7.

Results of the cogging torque analysis for the three cases mentioned above are shown in Fig. 8

From the comparison, it can be seen that the difference among the curves for all different magnet settings is slight. The curve of the anisotropic magnet is closer to the curve under ideal condition, the reason is the similarity between both magnetic fields. Both, the anisotropic field and ideal field only have a radial component of the field vector, the only difference between both fields is the magnitude in the transition zone.

For a detailed analysis, the curves can be transformed into frequency domain by using FFT analysis. The result is depicted in Fig. 9.

From this figure, it can be seen that the differences appear mainly in the 24th and 48th harmonics, which are up to 25% difference in the amplitude of order 48. Compared to the torque curves in time domain, the difference in frequency domain is much more prominent. The 24th harmonic is the main order of the torque curve. It is caused by the interaction between the first harmonic of the magnetic field in the air gap and the stator teeth.

Consequently, using the magnetic field from the magnetization analysis can improve the accuracy of the motor optimization, because the cogging torque fundamental and harmonics are mainly responsible for the coast down noise.

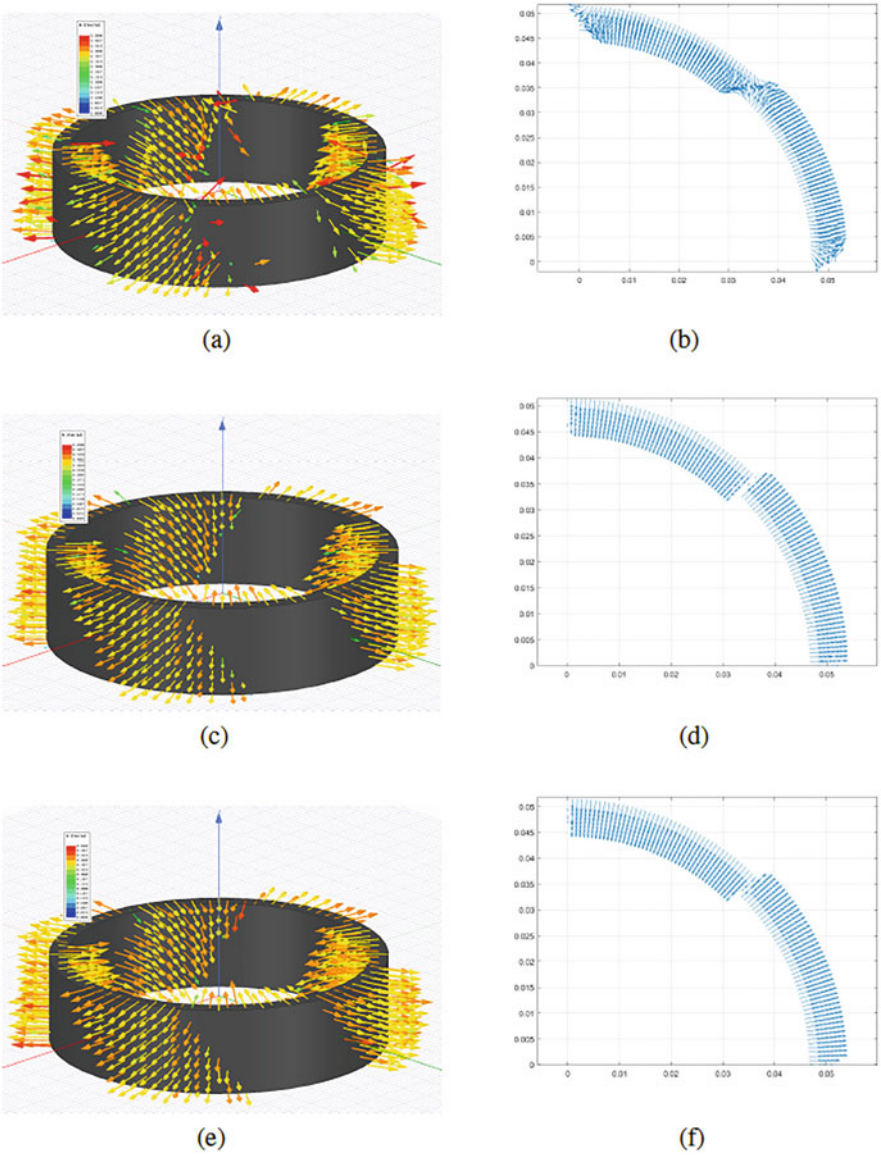


Fig. 7 3D and 2D figures of the remanence field of all three cases. (**a**) 3D Field of the isotropic magnet. (**b**) 2D Field of the isotropic magnet (Z=0, $0 \leq \varphi \leq 90°$). (**c**) 3D Field of the anisotropic magnet. (**d**) 2D Field of the anisotropic magnet (Z=0, $0 \leq \varphi \leq 90°$). (**e**) 3D Field of the ideal magnetization. (**f**) 2D Field of the ideal magnetization (Z=0, $0 \leq \varphi \leq 90°$)

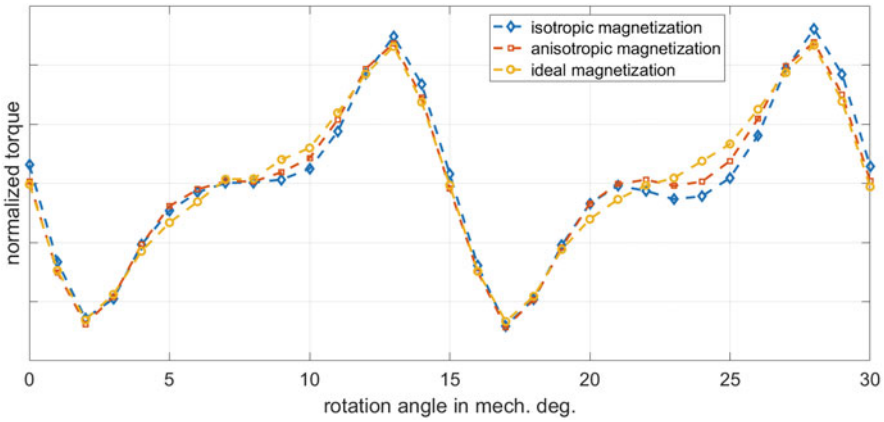


Fig. 8 Comparison of the cogging torque curves

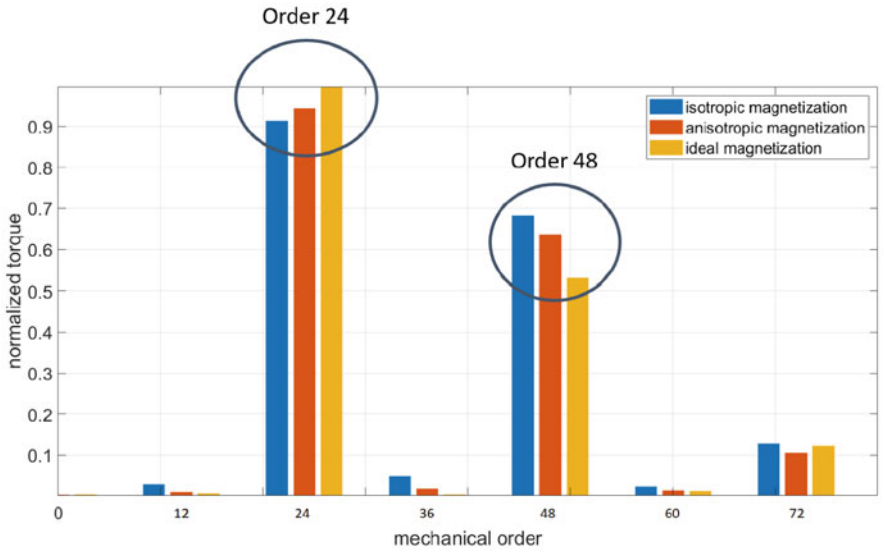


Fig. 9 Comparison the cogging torque in frequency domain

6 Summary

The calculation of the magnetization of the permanent magnet poles of an outer rotor BLDC motor results in a more realistic field distribution in the motor simulation and therefore allows a more accurate prediction of the cogging torque of the motor. This supports the overall design and optimization process of these machines. For proprietary reasons in this paper, the method is described using a ring magnet motor design. With real motor samples, an improved agreement of prediction (3D

field simulation) and measurement was found. It has to be pointed out that the ferrites used are usually manufactured be anisotropic i.e. to have a main axis and a lateral axis that respond differently to a magnetizing field. However due to material imperfections and variances in the manufacturing process the material is not perfectly anisotropic but has isotropic regions as well.

Therefore, to further improve the accuracy of the cogging torque analysis, an improved material definition that accounts for anisotropic as well as isotropic material properties needs to be developed.

References

1. I. Coenen, M. van der Giet, K. Hameyer, Manufacturing tolerances: estimation and prediction of cogging torque influenced by magnetization faults. IEEE Trans. Mag. **48**(5), 1932–1936 (2012)
2. M. Flankl, A. Tüysüz, J.W. Kolar, Cogging torque shape optimization of an integrated generator for electromechanical energy harvesting. IEEE Trans. Ind. Electron. **64**(12), 9806–9814 (2017)
3. C. Breton, J. Bartolome, J.A. Benito, G. Tassinario, I. Flotats, C.W. Lu, B.J. Chalmers, Influence of machine symmetry on reduction of cogging torque in permanent-magnet brushless motors. IEEE Trans. Magn. **36**(5), 3819–3823 (2000)
4. Z.Q. Zhu, D. Howe, C.C. Chan, Improved analytical model for predicting the magnetic field distribution in brushless permanent-magnet machines. IEEE Trans. Mag. **38**(1), 229–238 (2002)
5. J.R. Hendershot, T.J.E. Miller, *Design of Brushless Permanent-Magnet Machines* (Motor Design Books, Venice, 2010). ISBN 0-98406-870-8, 9-780-98406-870-8
6. ANSYS Maxwell support: Using the Hysteresis Model-Based Magnetization Approach
7. ANSYS Maxwell support: Compute Remanent Br from B-H curve

Part IV
Mathematical and Computational Methods

A Combination of Model Order Reduction and Multirate Techniques for Coupled Dynamical Systems

M. W. F. M. Bannenberg, A. Ciccazzo, and M. Günther

Abstract Coupled dynamical systems are often encountered in the field of circuit simulation. To drastically reduce the simulation cost of these systems a coupling of model order reduction and multirate techniques is applied. The subject of this method is a nonlinear coupled thermal-electrical system. By applying a combination of the slowest first multirate technique with the nonlinear proper orthogonal decomposition model order reduction the system is solved. Results yield a decrease in simulation time whilst maintaining accuracy.

1 Introduction

Building an integrated circuit, for instance a microchip, is a complex process. For the construction of such integrated circuits, first the silicon components are made and then connected through a conducting metal. Before these circuits can be constructed they need to be designed and tested by a computer. To this end the integrated circuits are described in mathematical form by differential-algebraic equations (DAEs). However, in the simulation of these circuits there are many more phenomena to consider besides only the connections and interactions between the components. Including these other phenomena occurring inside a microchip, like thermal or electromagnetic coupling, greatly improves the accuracy of the simulation. These considerations lead to a system of Partial Differential-Algebraic

M. W. F. M. Bannenberg (✉)
Bergische Universität Wuppertal, Wuppertal, Germany

STMicroelectronics, Catania, Italy
e-mail: bannenberg@uni-wuppertal.de

A. Ciccazzo
STMicroelectronics, Catania, Italy

M. Günther
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: guenther@uni-wuppertal.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_19

Equations (PDAEs). Where DAEs and partial differential equations, describing the spatially distributed elements and effects, are coupled via source terms or boundary conditions. Both physical and structural characteristics of these PDAEs can be exploited to increase the simulation efficiency. For instance by applying techniques such as Multirate (MR) time integration and Model Order Reduction (MOR), as will be presented in the following sections. Circuit simulation has been a driving force for the application of MOR and MR techniques, see for instance [3, 10]. Much less attention has been given to the combination of these two techniques, [13], and only with respect to linear model order reduction. In this paper a twofold approach is presented in which the PDAEs are integrated using MR time integration and parts of the system are reduced. This is done to increase the computational efficiency, whilst maintaining accuracy. In Sect. 2, the mathematical methodology is formulated for the circuit simulation and the multirate and MOR techniques are described. Section 3 presents the experimental setup and the numerical results obtained from the implementations of the previous two sections. Conclusions are drawn and an outlook is given in Sect. 4.

2 Methodology

In this section the different mathematical concepts and techniques that are needed for the simulation of electronic circuits are presented. Although most equations are purposely stated in their most general form, some of them will be restricted by assumptions with the specifics combination of MR and MOR in mind.

2.1 Mathematical Modelling

The Modified Nodal Analysis (MNA) approach for modelling electronic circuits yield time-dependent systems of DAEs,

$$\vec{A}_C \frac{d}{dt}\vec{q} + \vec{A}_R \vec{r}(\vec{A}_R^T \vec{e}) + \vec{A}_L \vec{i}_L + \vec{A}_V \vec{i}_V + \vec{A}_I i(t) = \vec{0}, \tag{1}$$

$$\frac{d}{dt}\vec{\phi} - \vec{A}_L^T \vec{e} = \vec{0}, \tag{2}$$

$$\vec{v}(t) - \vec{A}_V^T \vec{e} = \vec{0} \tag{3}$$

$$\vec{q} - \vec{q}_C(\vec{A}_C^T) = \vec{0}, \tag{4}$$

$$\vec{\phi} - \vec{\phi}_L(\vec{i}_L) = \vec{0}. \tag{5}$$

Where $\vec{e}$, $\vec{i}_L$, $\vec{V}$ are the node voltages and branch currents through inductors and voltage sources, and the charges and fluxes $\vec{q}$, $\vec{\phi}$. The functions $\vec{r}$, $\vec{q}_C$ and $\vec{\phi}_L$ are predetermined. Independent current sources $\vec{i}_I$ and voltage sources $\vec{v}_V$ may appear. The incidence matrices A_C, A_L, A_R, A_V, A_I follow from the topology of the circuit. This system can be written in the general semi-explicit DAE form, [7, 12].

$$\vec{f} : \mathbb{R}^n \times \mathbb{R}^m \times I \to \mathbb{R}^n, \quad \vec{g} : \mathbb{R}^n \times \mathbb{R}^m \times I \to \mathbb{R}^m. \tag{6}$$

$$\dot{\vec{y}} = \vec{f}(\vec{y}, \vec{z}, t), \quad \vec{y}(0) = \vec{y}_0, \tag{7}$$

$$0 = \vec{g}(\vec{y}, \vec{z}, t), \quad \vec{z}(0) = \vec{z}_0. \tag{8}$$

With $\vec{y} : I \to \mathbb{R}^n$ and $\vec{z} : I \to \mathbb{R}^m$ denoting the differential and algebraic solutions on time-interval $[t_0, t_1]$, respectively. Furthermore $\vec{y}_0$ and $\vec{z}_0$ need to be consistent initial conditions. Secondly, other phenomena can be included via PDEs, which are denoted in general form given by

$$\mathscr{L} : D \times I \times V \to \mathbb{R}^m, \quad \mathscr{L}(\vec{x}, t, \vec{u}) = 0. \tag{9}$$

where $\mathscr{L}$ is a differential operator, $D \subset \mathbb{R}^d$, with $d \in \{1, 2, 3\}$ the spatial domain and V a function space to which $\vec{u} : D \times I \to \mathbb{R}^m$ belongs. These two systems consisting of DAEs and PDEs can be coupled together in PDAEs. After applying a suitable space discretization to the PDAEs the following initial value problem of semi-explicit DAEs is obtained

$$\dot{\vec{y}} = \vec{f}(\vec{y}, \vec{z}, \vec{u}, t), \quad \vec{y}(0) = \vec{y}_0, \tag{10}$$

$$0 = \vec{g}(\vec{y}, \vec{z}, \vec{u}, t), \quad \vec{z}(0) = \vec{z}_0, \tag{11}$$

$$\dot{\vec{u}} = \vec{h}(\vec{y}, \vec{z}, \vec{u}, t), \quad \vec{u}(t_0) = \vec{u}_0. \tag{12}$$

2.2 Multirate

Since the coupled system (10)–(12) is constructed by the combination of two different processes it can be assumed that they act within different time scales. To exploit this characteristic, the total system is partitioned into fast and slow subsystems, with $x_F = y$, $x_S = u$ and $z_F = z$,

$$\dot{\vec{x}}_F = \vec{f}_F(\vec{x}_F, \vec{z}_F, \vec{x}_S), \quad \vec{x}_F(0) = \vec{x}_{F,0}, \tag{13}$$

$$\dot{\vec{x}}_S = \vec{f}_S(\vec{x}_F, \vec{z}_F, \vec{x}_S), \quad \vec{x}_S(0) = \vec{x}_{S,0}, \tag{14}$$

$$0 = \vec{g}_F(\vec{x}_F, \vec{z}_F, \vec{x}_S), \quad \vec{z}_F(0) = \vec{z}_{F,0}. \tag{15}$$

With differential variables $\vec{x}_F \in \mathbb{R}^{n_F}$, $\vec{x}_S \in \mathbb{R}^{n_S}$ and algebraic variables $\vec{z}_F \in \mathbb{R}^{n_Z}$, subscripts $\{F, S\}$ indicating fast or slow dynamics, for $t \in [t_0, t_1]$ with consistent

initial conditions. The system is guaranteed to be of index-1 by assuming that the Jacobian

$$\vec{g}_{F_{\vec{z}}}(\vec{x}_F, \vec{z}_F, \vec{x}_S) \text{ is invertible} \tag{16}$$

in a neighbourhood of the solution of the system (13-15). The algebraic constraints are partitioned into the fast subsystem. This type of coupling lets us consider electrical circuits with a differential index up to 1, coupled with slower ODE systems. The total index-1 system can be integrated with the stiffly accurate implicit Euler method. To exploit the assumed different time scales, a multirate integration method is proposed. This approach is analogous to [14] but with the algebraic constraint in the fast subsystem, taking the subsequent MOR into account. The integration of the coupled system (13-15) for one macro-step $t_n \to t_{n+1} = t_n + H$ is defined as

$$\vec{x}_{F,n+(l+1)/m} = \vec{x}_{F,n+l/m} + h\vec{f}_F(\vec{x}_{F,n+(l+1)/m}, \vec{z}_{F,n+(l+1)/m}, \bar{\vec{x}}_{S,n+(l+1)/m}), \tag{17}$$

$$\vec{x}_{S,n+1} = \vec{x}_{S,n} + H\vec{f}_S(\bar{\vec{x}}_{F,n+1}, \bar{\vec{z}}_{F,n+1}, \vec{x}_{S,n+1}), \tag{18}$$

$$0 = \vec{g}_F(\vec{x}_{F,n+(l+1)/m}, \vec{z}_{F,n+(l+1)/m}, \bar{\vec{x}}_{S,n+(l+1)/m}). \tag{19}$$

With $l = 0, \ldots, m-1$ for the micro grid and the coupling variables denoted by $\bar{\vec{x}}_F$, $\bar{\vec{z}}_F$, $\bar{\vec{x}}_s$. The coupling strategy is chosen to be the *Coupled-Slowest-First* approach as this is shown to have a consistency of order 1 for the problem posed in [14]. First the whole system is solved for the macro-step.

$$\vec{x}^*_{F,n+1} = \vec{x}_{F,n} + H\vec{f}_F(\vec{x}^*_{F,n+1}, \vec{z}^*_{F,n+1}, \vec{x}_{S,n+1}), \tag{20}$$

$$\vec{x}_{S,n+1} = \vec{x}_{S,n} + H\vec{f}_S(\vec{x}^*_{F,n+1}, \vec{z}^*_{F,n+1}, \vec{x}_{S,n+1}), \tag{21}$$

$$0 = \vec{g}_F(\vec{x}^*_{F,n+1}, \vec{z}^*_{F,n+1}, \vec{x}_{S,n+1}). \tag{22}$$

Where the step size H is chosen according to the slow dynamics. From this it follows that the fast solutions, $\vec{x}^*_{F,n+1}$ and $\vec{z}^*_{F,n+1}$, are not accurate and discarded. Following the micro-step integration the fast solutions are computed for $l = 0, \ldots, m-1$, using linearly interpolated values for the slow variables.

2.3 Model Order Reduction

Applying a spatial discretization to the PDE can result in large nonlinear ODE systems. To reduce the computational effort needed in each time step to solve this system MOR techniques are used. Due to the nonlinearity of the ODE most conventional MOR techniques can be discarded as they are only applicable to linear

systems. Hence the chosen method for this system is a reduction by a Galerkin projection, with a basis constructed by Proper Orthogonal Decomposition (POD), [4]. This is then extended by the application of the Discrete Empirical Interpolation Method (DEIM), [5], using a QR selection procedure (Q-DEIM), [6]. By using a Galerkin projection a reduced model is constructed, [6]. Let $\mathscr{V}_r$ denote an r-dimensional subspace spanned by the columns of $V \in \mathbb{R}^{n_S \times r}$. The full state of the slow subsystem $\vec{x}_S$ is then approximated by $\vec{x}_S \approx V\vec{x}_{S,r}$ using model reduction basis V. The reduced model of (13)–(15) is then defined by

$$\dot{\vec{x}}_F = \vec{f}_F(\vec{x}_F, \vec{z}_F, V\vec{x}_{S,r}), \qquad \vec{x}_F(0) = \vec{x}_{F,0}, \tag{23}$$

$$\dot{\vec{x}}_{S,r} = \vec{f}_{S,r}(\vec{x}_F, \vec{z}_F, \vec{x}_{S,r}), \qquad \vec{x}_{S,r}(0) = \vec{x}_{S,r,0}, \tag{24}$$

$$0 = \vec{g}_F(\vec{x}_F, \vec{z}_F, V\vec{x}_{S,r}), \qquad \vec{z}_F(0) = \vec{z}_{F,0}. \tag{25}$$

With $\vec{f}_{S,r}(\vec{x}_F, \vec{z}_F, \vec{x}_{S,r}) = V^T \vec{f}_S(\vec{x}_F, \vec{z}_F, V\vec{x}_{S,r})$. The reduced basis V is constructed through POD. First a numerical simulation of the full system is performed. From the numerical results of this simulation *snapshots* $x_1, x_i, \ldots, x_{N_S}$ are obtained, with $x_i = x(t_i) \in \mathbb{R}^{n_S}$ for $i = 1, \ldots, N_S$. Then the POD *snapshot matrix* is

$$\mathbb{X} = [x_1, \ldots, x_{N_S}] \in \mathbb{R}^{n_S \times N_S}. \tag{26}$$

From this the thin Singular Value Decomposition (SVD) is computed

$$\mathbb{X} = Z\Sigma Y^T, \tag{27}$$

where $Z \in \mathbb{R}^{n_S \times k}$, $Y \in \mathbb{R}^{N_S \times k}$ are orthogonal and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k \times k}$ with $k = \min(n_S, N_S)$. Now a reduction basis V is constructed by taking the leading r singular vectors of Z corresponding to the r largest singular values. However there is a problem with the Galerkin projection that causes computational inefficiencies. The reduced term $\vec{f}_{S,r}(\vec{x}_F, \vec{z}_F, \vec{x}_{S,r})$ has a computational complexity that depends on the non-reduced full order size n_S. To reduce the number of evaluations Q-DEIM is applied. Consider the nonlinear function $\vec{f}_S : \mathscr{T} \to \mathbb{R}^{n_S}$ with $\mathscr{T} \subset \mathbb{R}^{n_S}$, the coupled terms are dropped from the notation as these are not reduced, and matrix $U \in \mathbb{R}^{n_S \times m}$ of rank m. Then the DEIM approximation of $\vec{f}_S$ is defined by, [5, Definition 3.1],

$$\hat{\vec{f}}_S(\tau) = U(\mathbb{S}U)^{-1}\mathbb{S}^T \vec{f}_S(\tau). \tag{28}$$

where $\mathbb{S}$ is a selection matrix of size $n_S \times m$ by selecting columns of identity matrix $\mathbb{I}$ of size $n_S \times n_S$. Then the reduced nonlinear function $\vec{f}_{S,r}$ is approximated with the QDEIM approach by

$$\vec{f}_{S,r}(\vec{x}_{S,r}) \approx V^T U(\mathbb{S}U)^{-1}\mathbb{S}^T \vec{f}_S(V\vec{x}_{S,r}). \tag{29}$$

Using the interpolation of general nonlinear functions, outlined in Sect. 3.5 of [5], a general nonlinear function can be represented as

$$[\vec{F}(\vec{y})]_i = F_i(\vec{y}) = \vec{F}_i(\vec{y}_{j_1^i}, \vec{y}_{j_2^i}, \ldots, \vec{y}_{j_{n_i}^i}) = F_i(\vec{y}(\vec{j}_i)), \tag{30}$$

where $F_i : \mathscr{Y}_i \to \mathbb{R}$, $\mathscr{Y}_i \subset \mathbb{R}^{n_i}$, with integer vector $\vec{j}_i = [j_1^i, j_2^i, \ldots, j_{n_i}^i]$ denoting the indices of the components required to evaluate F_i. The numerical implementation of this allows to compute (29) without the full evaluation of $\vec{f}_S$. Depending on the underlying nonlinear function it might even be possible to compute (29) without lifting $x_{S,r}$ to the full dimension n_S. This however depends on the dependencies of the individual functions.

2.4 Combining MR and MOR

To maximise the effectiveness of the MR and MOR combination the following steps are taken:

- Perform a benchmark simulation using a very large number of time steps to obtain a very accurate snapshot matrix $\mathbb{X}$.
- The reduced bases V and U are then constructed by taking the appropriate columns of Z obtained through POD, and selection matrix $\mathbb{S}$ is constructed by the Q-DEIM approach.
- Using the reduced bases, the reduced order system is integrated through time using the *Coupled-Slowest-First* MR approach.

The computational approach of this is done by first using the scheme of (17)–(19) with $\vec{f}_S$ replaced by $\vec{f}_{S,r}$, as in (23)–(25), and then incorporating the *Coupled-Slowest-First* approach. The coupling for the fast intermediate time-step is done by using linear interpolated values. As these values don't change during the Newton iteration of solving the faster subsystem, computation time can be saved. By computing the coupling values once for the first and last value and interpolating between these values, expensive function evaluation of the lifted state vector can be avoided.

3 Results

In this section the previously described MR-MOR integration scheme is implemented and applied to the thermal-electric test problem of [2]. For the simulation the algorithm has been implemented in C++ using various packages, GSL 2.6 [8], LAPACK [1] and Eigen [9]. Visualisation of the results is done by using MATLAB 2019b under the Academic Student License.

3.1 Experimental Setup

To test the accuracy and convergence of the MR-MOR integration scheme the system is simulated in three settings:

- The full system with singlerate time integration.
- The full system with multirate time integration.
- The system with a reduced slow part and multirate time integration.

This is done for with intermediate micro-steps $m = 5$. Furthermore the POD-QDEIM reduction factors r and g are chosen to be equal to the number of largest singular values with $\sigma_i > 1e-15$. The step sizes are obtained by integrating the system with $N_t = [8\ 16\ 32\ 64\ 128\ 256\ 512\ 1024]$. For the simulation a thermodynamic discretisation is chosen to have $N = 101$. The circuit is simulated over a time interval from $t_0 = 0$ to $t_N = 0.01125$ seconds. The input signal $v(t)$ is set to $\sin(\frac{\pi t}{2.5e-3})mV$. The rest of the circuit and thermal settings are set to the values as described in [2]. The reference solution is obtained from an SR integration with $N = 32,000$.

In Fig. 1 we see the difference between the reference solution and the simulated solution in the final time-step. This is done for the output node u_3 of the thermal-electrical circuit. It clearly shows that the MR scheme outperforms the SR, as for the same order of error the MR approach has a slower computation time.

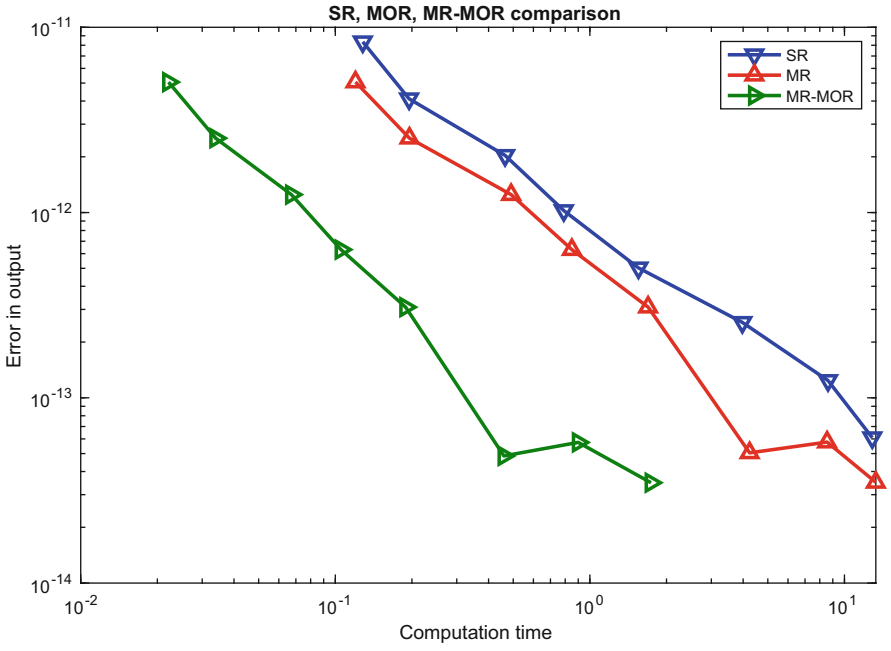


Fig. 1 Accuracy of SR, MOR and MR-MOR

Incorporating the POD-QDEIM MOR method results in an even further reduction of the computational integration effort. However, there needs to be done a snapshot simulation before this can be used. Thus the speedup has a one time extra cost.

4 Conclusion

From the numerical results it shows that the multirate implicit Euler scheme combined with POD/Q-DEIM model order reduction results in an accurate solution with a reduced computation time. The approximation errors seem to converge along with the MR errors. An expected positive result is that for similar computation times the application of MR improves the accuracy of the solution. Furthermore, the additional application of POD/Q-DEIM reduction has a trivial impact on the approximation error whilst even further reducing the computation time. Although these results are positive a side note should be made. The reduction in computation time of the POD/Q-DEIM reduction shows to be decreasing for smaller time steps with much larger systems. This is likely due to the coupling structure of the test problem but further investigation is needed. Other next steps will focus on numerical analysis for a proof of convergence for the MR-MOR scheme and the extension to general integration schemes. Besides the further investigation of DAE-ODE coupled systems, first steps have been made towards a MR-MOR scheme for a DAE-DAE coupled system.

Acknowledgments The authors are indebted to the funding given by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 765374.

References

1. E. Anderson et al., *LAPACK Users' Guide*, vol. 9 (SIAM, New York, 1999)
2. A. Bartel, M. Günther, From SOI to abstract electric-thermal-1D multiscale modeling for first order thermal effects. Math. Comput. Modell. Dynam. Syst. **9**(1), 25–44 (2003)
3. P. Benner, M. Hinze, E.J.W. Ter Maten (eds.) *Model Reduction for Circuit Simulation* (Springer, Berlin, 2011)
4. G. Berkooz, P. Holmes, J.L. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows. Annu. Rev. Fluid Mech. **25**(1), 539–575 (1993)
5. S. Chaturantabut, D.C. Sorensen, Nonlinear model reduction via discrete empirical interpolation. SIAM J. Sci. Comput. **32**(5), 2737–2764 (2010)
6. Z. Drmac, S. Gugercin, A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions. SIAM J. Sci. Comput. **38**(2), A631–A648 (2016)
7. D. Estévez Schwarz, C. Tischendorf, Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theory Appl. **28**(2), 131–162 (2000)
8. M. Galassi, GNU scientific library (2002). http://www.gnu.org/software/gsl/

9. G. Guennebaud, B. Jacob, Eigen: a c++ linear algebra library (2014). http://eigen.tuxfamily.org. Accessed 22 November 2014
10. M. Günther (ed.) *Coupled Multiscale Simulation and Optimization in Nanoelectronics* (Springer, Berlin, Heidelberg, 2015)
11. M. Günther, U. Feldmann, CAD-based electric-circuit modeling in industry I: mathematical structure and index of network equations. Surv. Math. Ind. **8**(2), 97–130 (1999)
12. M. Günther, U. Feldmann, J. ter Maten, Modelling and discretization of circuit problems, in *Handbook of Numerical Analysis*, vol. 13 (Elsevier, Amsterdam, 2005), pp. 523–659
13. C. Hachtel et al., Multirate DAE/ODE-simulation and model order reduction for coupled field-circuit systems, in *Scientific Computing in Electrical Engineering* (Springer, Cham, 2018), pp. 91–100
14. C. Hachtel et al., Multirate implicit Euler schemes for a class of differential–algebraic equations of index-1. J. Comput. Appl. Math. **2019**, 112499 (2019)

Waveform Relaxation for Low Frequency Coupled Field/Circuit Differential-Algebraic Models of Index 2

Idoia Cortes Garcia, Jonas Pade, Sebastian Schöps, and Caren Tischendorf

Abstract Motivated by the task to design quench protection systems for superconducting magnets in particle accelerators we address a coupled field/circuit simulation based on a magneto-quasistatic field modeling. We investigate how a waveform relaxation of Gauß-Seidel type performs for a coupled simulation when circuit solving packages are used that describe the circuit by the modified nodal analysis. We present sufficient convergence criteria for the coupled simulation of FEM discretised field models and circuit models formed by a differential-algebraic equation (DAE) system of index 2. In particular, we demonstrate by a simple benchmark system the drastic influence of the circuit topology on the convergence behavior of the coupled simulation.

1 Introduction

Lumped circuit models, such as modified nodal analysis (MNA), are well-established in electrical engineering. However, they neglect the spatial dimension and therefore distributed phenomena like the skin effect. For certain devices, this may lead to inaccuracies of unacceptable magnitude in the simulation, e.g. for electric machines [14] or the quench protection system of superconducting magnets in particle accelerators [1]. These cases call for field/circuit coupling [2, 16]. To solve such coupled systems, it is often advisable to use waveform relaxation (WR) [7], since this iterative method allows for dedicated step sizes and suitable solvers for the different subsystems, and even for the use of proprietary blackbox solvers. The coupled field/circuit model considered here is a DAE in the time domain after space discretisation of the field system. It is well-known that WR

I. C. Garcia · S. Schöps
Technical University of Darmstadt, CEM Group, Darmstadt, Germany
e-mail: idoia.cortes@tu-darmstadt.de; sebastian.schoeps@tu-darmstadt.de

J. Pade (✉) · C. Tischendorf
Department of Mathematics, Humboldt University of Berlin, Berlin, Germany
e-mail: jonas.pade@math.hu-berlin.de; tischendorf@math.hu-berlin.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_20

can suffer from instabilities for DAEs unless an additional contraction criterion is satisfied [7, 12]. This work presents coupled field/circuit models, which are DAEs of index 2 [5], for the case where WR is convergent and the case where it diverges. Furthermore, generalizing a convergence criterion of [12], a topological and easy-to-check criterion is provided. Finally, we present numerical simulations verifying the topological convergence criterion.

2 Field/Circuit Model

To describe the electromagnetic (EM) field part, we consider a magnetoquasistatic approximation of Maxwell's equations in a reduced magnetic vector potential formulation [4]. This leads to the curl-curl eddy current partial differential equation (PDE). The circuit side is formulated with the MNA [6]. For the numerical simulation of the coupled system, the method of lines is used with a finite element (FE) discretisation. Altogether, this leads to a time-dependent coupled system of DAE initial value problems (IVPs), described by

$$M\dot{a} + K(a)a - Xi_m = 0, \qquad X^\top \dot{a} = v_c, \tag{1}$$

$$E(x)\dot{x} + f(t,x) = Pi_m, \qquad P^\top x - v_c = 0. \tag{2}$$

The first Eq. (1) represents the space-discrete field model based on the matrices

$$(M)_{ij} = \int_\Omega \sigma \omega_i \cdot \omega_j \,\mathrm{dV}, \quad (K(a))_{ij} = \int_\Omega \nu(a) \nabla \times \omega_i \cdot \nabla \times \omega_j \,\mathrm{dV}\,, \tag{3}$$

which follow from the Ritz-Galerkin approach using a finite set of Nédélec basis functions ω_i [10] defined on the domain Ω; σ denotes the space-dependent electric conductivity and $\nu(a)$ the magnetic reluctivity that can additionally depend nonlinearly on the unknown magnetic vector potential a. The current through the field device is described by i_m. The excitation matrix is computed from a winding density function χ_j modelling the j-th stranded conductor [15] as

$$(X)_{ij} = \int_\Omega \chi_j \cdot \omega_i \,\mathrm{dV}\,. \tag{4}$$

Definition 1 A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is *strongly monotone* and a square matrix $M(x)$ is *uniformly positive definite*, if

$$\exists \mu_f : \quad (x_2 - x_1)^\top (f(x_2) - f(x_1)) \geq \mu_f \|x_2 - x_1\|^2, \qquad \forall x_1, x_2 \in \mathbb{R}^n,$$

$$\exists \mu_M : \quad y^\top M(x) y \geq \mu_M \|y\|^2, \qquad \forall x \in \mathbb{R}^n, y \in \mathbb{R}^m.$$

The space-discretization is supposed to meet the following properties.

Assumption 2 *It holds (a) M is symmetric, (b) the matrix pencil $\lambda M + K$ is symmetric and positive definite for $\lambda > 0$, (c) X has full column rank and (d) the function $a \mapsto K(a)a$ is strongly monotone.*

The assumptions are in agreement with previous formulation in the literature, e.g. [3, 15]. The first Assumption 2a follows naturally if a Ritz-Galerkin formulation (3) is chosen. The second Assumption 2b will be guaranteed by appropriate boundary and gauging conditions. Thirdly, the full column rank Assumption 2c follows from the fact that the columns are discretisations of different coils that are located in spatially disjoint subdomains. Finally, the monotonicity Assumption 2d follows from the strong monotonicity of the underlying nonlinear material law, i.e. the BH-curve [13]. In general, the field model is a multiport element such that the circuit coupling is established via multiple currents and voltages, i.e., vector-valued i_m and v_c. However, for simplicity of notation we assume a two-terminal device in the following.

The circuit Eq. (2) can be expanded into

$$E(x) = \begin{pmatrix} \mathscr{L}_C(e) & 0 & 0 \\ 0 & -L(i_L) & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad f(t,x) = \begin{pmatrix} g_R(e) + A_L i_L + A_V i_V + q_i(t) \\ A_L^\top e \\ A_V^\top e - q_v(t) \end{pmatrix}, \quad P = \begin{pmatrix} A_m \\ 0 \\ 0 \end{pmatrix} \tag{5}$$

using the definitions $\mathscr{L}_C(e) := A_C C(A_C^\top e) A_C^\top$, $g_R(e) := A_R g(A_R^\top e)$ and $x = (e, i_L, i_V)$ where $A_\star$ are the usual incidence matrices and $L(\cdot)$, $C(\cdot)$ are state-dependent square matrices describing inductances and capacitances. The position of the field device in the circuit is described by A_m, and the voltage over the device by v_c. The function $g(\cdot)$ describes the voltage–current relation of resistive elements and q_i, q_v are the input current and voltage. Finally, x collects all node potentials e, currents through branches with voltage sources i_V and inductors i_L. The circuit system shall fulfill the following properties:

Assumption 3 *It holds (a) g, C and L are Lipschitz continuous, g is strongly monotone and C, L are uniformly positive definite, (b) q_i and q_v are continuously differentiable, (c) A_V has full column rank and $\begin{pmatrix} A_C & A_V & A_R & A_L \end{pmatrix}$ has full row rank.*

Assumption 3a reflects the global passivity of the respective elements [8]. Considering well-known relations between incidence matrices and circuit topology, Assumption 3c excludes the electrically forbidden configurations of loops of voltage sources and cutsets of current sources [5].

3 Waveform Relaxation and Convergence

We consider the Gauß-Seidel WR method. Applied to the coupled system (1)–(2) for given consistent initial values, this yields the scheme

$$M\dot{a}^k + K(a^k)a^k - Xi_m^k = 0, \qquad X^\top \dot{a}^k = v_c^{k-1}, \tag{6}$$

$$E(x^k)\dot{x}^k + f(t, x^k) = Pi_m^k, \qquad P^\top x^k - v_c^k = 0. \tag{7}$$

The coupling variables are the current through and the voltage over the field device i_m and v_c, where i_m^k is computed in (6) and is then given to (7) as input, and vice versa for v_c^k. The superscript k denotes the iteration index. A common choice for the initial guess v_c^0 is constant extrapolation of the initial value.

We shall proceed as follows:

1. Lemmata 4 and 6 provide a DAE-decoupling of the EM field DAE (1) and the MNA DAE (2), respectively.
2. Definition 5 introduces the concept of parallel CVR paths. Assuming their existence and exploiting the previous decoupling Lemmata, Lemma 7 yields a DAE-decoupling of the coupled WR iteration (6)–(7). Notably, it reveals the structure of its inherent ODE, given by ϕ in Eq. (11).
3. The convergence Theorem 8 is a simple consequence of the previous Lemmata; it shows that the existence of parallel CVR paths guarantees convergence of the WR scheme (6)–(7).

For visual reasons, we shall write column vectors as (a, b, c).

Lemma 4 *Let Assumption 2 hold. Then, for a given source term v_c, there exists a coordinate transformation $(w, u) = T^{-1}a$ and a system of the form*

$$\dot{u} + A_1 u = A_2 v_c, \qquad w = Bu, \qquad i_m = G_1 u + G_2 v_c \tag{8}$$

such that (a, i_m) solves Eq. (1) if and only if (u, w, i_m) solves Eq. (8).

Proof For better readability and shortness we present the proof only for the slightly more restrictive case where $X^\top M = 0$, which is usually satisfied.

We equivalently transform the field DAE with new coordinates $T\alpha = a$:

$$\begin{aligned} T^\top M T\dot{\alpha} + T^\top K(T\alpha)T\alpha - T^\top X i_m = 0, \\ X^\top T\dot{\alpha} = v_c. \end{aligned} \tag{9}$$

The transformation matrix $T := (T_{\ker}\ X\ T_\perp)$ is constructed such that the columns of $T_{\ker}$ and $T_\perp$ form a basis of $\ker M \cap \ker X^\top$ and $(\ker M)^\perp$, respectively. It is nonsingular indeed, since its construction and Assumption 2 combined with $X^T M = 0$ guarantee that $\mathrm{im}X \perp \mathrm{im}T_{\ker}$ and $\mathrm{im}T_\perp \perp \mathrm{im}(T_{\ker}\ X)$.

With $\alpha = (w, u)$ and $u = (u_1, u_2)$, the transformed DAE (9) has the detailed form

$$\begin{aligned} \underline{T_{\ker}^\top K(T\alpha) T_{\ker}} w + T_{\ker}^\top K(T\alpha)(X \;\; T_\perp) u &= 0, \\ X^\top K(T\alpha) T\alpha - \underline{X^\top X} i_m &= 0, \\ \underline{T_\perp^\top M T_\perp} \dot{u}_2 + T_\perp^\top K(T\alpha) T\alpha &= 0, \\ \underline{X^\top X} \dot{u}_1 &= v_c. \end{aligned}$$

The underlined matrices are nonsingular due to Assumption 2, and Eq. (8) is obtained by inversion and insertion.

Definition 5 A *CVR path* in a circuit is a path which consists of only capacitances, voltages sources and resistances. An element has a *parallel CVR path*, if its incident nodes are connnected by a CVR path.

Lemma 6 *Let Assumption 3 hold. Then, for a given source term i_m, there exists a coordinate transformation $(y, z_1, z_2) = T^{-1}x$ and a system of the form*

$$\dot{y} = f_0(t, y, z, z_2, u), \quad z_1 = g_1(t, y, z_2, \dot{z}_2, u), \quad z_2 = g_2(t) + QPi_m, \tag{10a}$$

$$v_c = P^\top T(y, z_1, z_2) \tag{10b}$$

with f_0, g_1, g_2 uniformly globally Lipschitz continuous $\forall t$ and $g_2 \in C^1$ such that

1. *(x, v_c) solves Eq. (2) if and only if (y, z_1, z_2, v_c) solves Eq. (10),*
2. *$QP = 0$ if each EM field element has a parallel CVR-path.*

A detailed proof can be found in [11], where Q is shown to have the form $(Q_1 \; * \; *)$ with $\operatorname{im} Q_1 = \ker(A_C \; A_V \; A_R)^\top$. Hence, if each field element has a parallel CVR-path, each column of A_m can be written as a sum of columns of $(A_C \; A_V \; A_R)$ and it follows $Q_1 A_m = 0$, thus $QP = 0$.

Lemma 7 *Let Assumptions 2 and 3 hold. If each EM field element has a parallel CVR path, then there exists a coordinate transformation $(r, s) = T^{-1}(a, x)$ and a system of the form*

$$\dot{s}^k = \phi(t, s^k, s^{k-1}), \qquad r^k = \varphi(t, s^k) \tag{11}$$

with ϕ uniformly globally Lipschitz continuous $\forall t$ and ϕ, φ continuous such that (a^k, i_m^k, x^k, v_c^k) solves Eqs. (6)–(7) if and only if (s^k, r^k) solves Eq. (11).

Proof We apply Lemmata 4, 6 to the iterated subsystems (6), (7). This yields an equivalent system

$$\dot{u}^k = -A_1 u^k + A_2 v_c^{k-1}, \quad w^k = Bu^k, \quad i_m^k = G_1 u^k + G_2 v_c^{k-1}, \tag{12}$$

$$\dot{y}^k = f_0(t, y^k, z^k, z_2^k, u^k), \quad z_1^k = g_1(t, y^k, z_2^k, \dot{z}_2^k, u^k), \quad z_2^k = g_2(t), \tag{13}$$

$$v_c^k = P^\top T(y^k, z_1^k, z_2^k). \tag{14}$$

Since each field element has a parallel CVR path, $z_2^k = g(t)$ does not depend on u^k anymore.

We insert $v_c^{k-1} = P^\top x^{k-1} = P^\top T(y^{k-1}, z_1^{k-1}, z_2^{k-1})$ and z_1^{k-1} and z_2^{k-1} therein to obtain, with $\tilde{g}_1(t, y^{k-1}, u^{k-1}) = g_1(t, y^{k-1}, g_2(t), \dot{g}_2(t), u^{k-1})$,

$$\dot{u}^k = \phi_2(t, u^k, y^k, u^{k-1}, y^{k-1}) \quad := -A_1 u^k + A_2 P^\top T(y^{k-1}, \tilde{g}_1(t, y^{k-1}, u^{k-1}), g_2(t)).$$

Insertion of $z_1^k, z_2^k, \dot{z}_2^k$ into f_0 yields

$$\dot{y}^k = \phi_1(t, u^k, y^k) \quad := f_0(t, y^k, g_1(t, y^k, g_2(t), \dot{g}_2(t), u^k), g_2(t), u^k).$$

Hence, defining $s^k := (u^k, y^k)$ and $\phi := (\phi_1, \phi_2)$, the sequence (u^k, y^k) is given implicitly by an ODE recursion of the form $\dot{s}^k = \phi(t, s^k, s^{k-1})$.

The algebraic constraint of Eq. (11) is obtained with $r^k = (w^k, i^k, z_1^k, z_2^k, v_c^k)$, $s^k = (u^k, y^k)$ and

$$\varphi(t, s) = (Bu, Gu, g_1(t, y, g_2(t), \dot{g}_2(t), u), g_2(t)).$$

Clearly, (s^k, r^k) solves Eq. (11) if and only if $\tilde{\alpha}^k := (u^k, w^k, i_m^k, y^k, z_1^k, z_2^k, v_c^k)$ solves Eqs. (12)–(14), and $\tilde{\alpha}^k$ solves (12)–(14) if and only if (a^k, i_m^k, x^k, v_c^k) solves Eqs. (6)–(7).

We deduce the main result of this work:

Theorem 8 *If each EM field element of the coupled system* (1)–(2) *has a parallel CVR path, then the WR scheme* (6)–(7) *is uniformly convergent to the exact solution of* (1)–(2).

Proof The ODE part of Eq. (11) is a WR scheme for ODEs with Lipschitz continuous vector field ϕ. It is well-known that such schemes are unconditionally convergent on bounded time intervals [7]. The convergence of s^k clearly implies the convergence of (s^k, r^k) defined by (11). Due to the equivalence provided by Lemma 7, it follows that the original scheme (6)–(7) is convergent.

Remark 9 The convergence result holds for arbitrary continuous initial guesses x^0 and for bounded intervals of arbitrary size, see e.g. [7, 11].

Remark 10 The MNA decoupling given in Lemma 6 shows that g_1 depends on z_2 and the derivative $\dot{z}_2$. Hence, the system is most sensitive to perturbations of z_2. The input of the EM field subsystem in the WR scheme is in fact a perturbation. Therefore, the condition $QP = 0$ from Lemma 6 is crucial to derive Theorem 8. If at least one EM field element has no parallel CVR path, then $QP \neq 0$. Then, analogously to Lemma 7 and its proof, we find $\dot{s}^k = \phi(t, s^k, s^{k-1}, \dot{s}^{k-1})$, which is guaranteed to converge only if ϕ is contractive in $\dot{s}^{k-1}$, see [7, 11].

4 Numerical Examples

To illustrate the convergence behaviour of the WR scheme according to the derived criteria, we consider the toy example circuits in Fig. 2a and b. Both are described with MNA (2) and the (arbitrary) parameters $R = 1\Omega$, $L = 5\text{H}$, $C = 1\text{F}$, $i_\text{s}(t) = \sin(2t) + 5\sin(20t)$ and $v_\text{s}(t) = \sin(t) + \sin(20t)$ are set. The eddy current Eq. (1) is solved on the single phase isolation transformer shown in Fig. 1. For simplicity, a zero current is imposed on the secondary coil (dark orange) and only the primary coil is coupled to the circuit.

The WR algorithm is applied on the simulation time window $\mathscr{I} = [0\ 0.8]\,\text{s}$ and the internal time integration is performed with the implicit Euler scheme with time step size $\delta t = 10^{-2}\,\text{s}$. The theoretical result is illustrated by the successful simulation, see Fig. 3a, of the model shown in Fig. 2a which satisfies the convergence criterion of Theorem 8. However, numerical simulations of the model shown in Fig. 2b show that WR can diverge indeed if the criterion is not satisfied (Fig. 3).

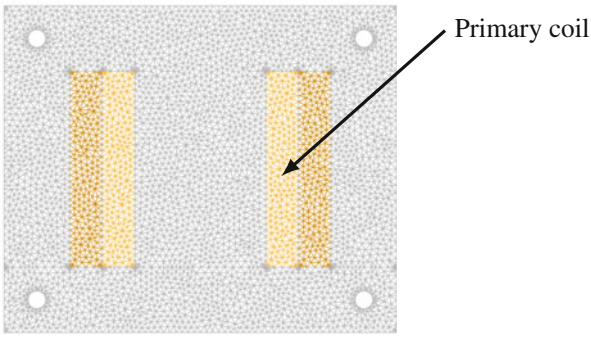


Fig. 1 Single phase isolation transformer ('MyTransformer'), see [9]

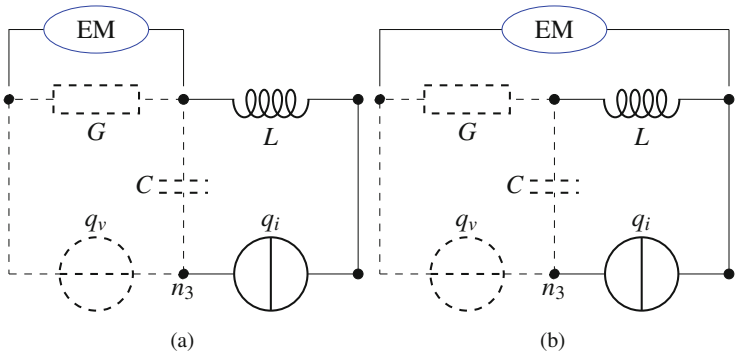


Fig. 2 Field/circuit coupling with model from Fig. 1 (CVR path is dashed). (**a**) Convergent case. (**b**) Divergent case

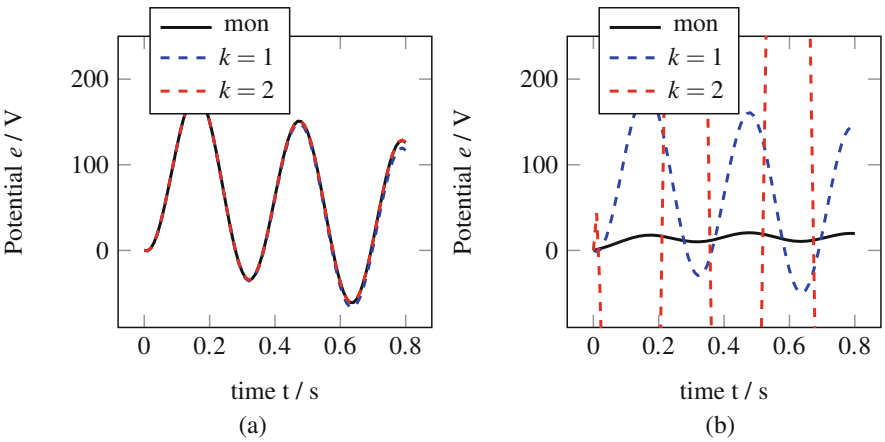


Fig. 3 Monolithic ("mon") and WR solution for $k = 1, 2$ iterations. (**a**) Convergent case. (**b**) Divergent case

5 Conclusions

In this work, we have presented a space-discretised coupled field/circuit model, which is a DAE of index 2, and a simulation of this model by means of WR. Furthermore, we have provided an easy-to-check topological convergence criterion for a class of coupled DAE/DAE systems of index 2.

Acknowledgments This work is supported by the 'Excellence Initiative' of the German Federal and State Governments, the Graduate School of CE at TU Darmstadt and DFG grant SCHO1562/1-2. Further, we acknowledge financial support under BMWi grant 0324019E and by DFG under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, ID 390685689).

References

1. L. Bortot et al., STEAM: a hierarchical co-simulation framework for superconducting accelerator magnet circuits. IEEE Trans. Appl. Super. **28**(3) (2018). 4900706
2. I. Cortes Garcia et al., Optimized field/circuit coupling for the simulation of quenches in superconducting magnets. IEEE J. Multiscale Multiphys. Comput. Tech. **2**(1), 97–104 (2017)
3. I. Cortes Garcia, H. De Gersem, S. Schöps, A structural analysis of field/circuit coupled problems based on a generalised circuit element. Numer. Algorithm. **83**(1), 373–394 (2020)
4. C.R.I. Emson, C.W. Trowbridge, Transient 3d eddy currents using modified magnetic vector potentials and magnetic scalar potentials. IEEE Trans. Magn. **24**(1), 86–89 (1988)
5. D. Estévez Schwarz, C. Tischendorf, Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl. **28**(2), 131–162 (2000)
6. C.-W. Ho, A.E. Ruehli, P.A. Brennan, The modified nodal approach to network analysis. IEEE Trans. Circ. Syst. **22**(6), 504–509 (1975)
7. E. Lelarasmee, A.E. Ruehli, A.L. Sangiovanni-Vincentelli, The waveform relaxation method for time-domain analysis of large scale integrated circuits. IEEE Trans. Comput. Aided Des. Integr. Circ. Syst. **1**(3), 131–145 (1982)
8. M. Matthes, Numerical Analysis of Nonlinear Partial Differential-Algebraic Equations: A Coupled and an Abstract Systems Approach. Dissertation, Universität zu Köln, 2012
9. D. Meeker, Finite Element Method Magnetics, version 4.2 (25feb2018 build) edition, 2018
10. P. Monk, *Finite Element Methods for Maxwell's Equations* (Oxford University Press, Oxford, 2003)
11. J. Pade, Analysis and waveform relaxation for a differential-algebraic electrical circuit model. Dissertation, Humboldt University of Berlin, 2021
12. J. Pade, C. Tischendorf, Waveform relaxation: a convergence criterion for differential-algebraic equations. Numer. Algorithm. **81**, 1327–1342 (2019)
13. C. Pechstein, Multigrid-Newton-methods for nonlinear-magnetostatic problems. Master's thesis, Universität Linz, 2004
14. S.J. Salon, *Finite Element Analysis of Electrical Machines* (Kluwer, Boston, 1995)
15. S. Schöps, *Multiscale Modeling and Multirate Time-Integration of Field/Circuit Coupled Problems* (VDI Verlag. Fortschritt-Berichte VDI, Reihe, 2011)
16. S Schöps, H. De Gersem, A. Bartel, A cosimulation framework for multirate time-integration of field/circuit coupled problems. IEEE Trans. Magn. **46**(8), 3233–3236 (2010)

Splitting Methods for Linear Circuit DAEs of Index 1 in port-Hamiltonian Form

Malak Diab and Caren Tischendorf

Abstract Operator splitting is a powerful method for numerical investigation of complex models. This method was successfully used for ordinary and partial differential equations (ODEs and PDEs). In constrained dynamical problems as electric circuits or energy transport networks, differential-algebraic equations (DAEs) arise. The constraints prevent a simple transfer of operator splitting from ODEs to DAEs. Here, we present an approach for splitting linear circuit DAEs of index 1 based on a port-Hamiltonian modeling that we derive from loop and cutset equations by a topological decoupling. Finally, we present convergence results for the proposed DAE operator splitting.

1 Introduction

The idea of operator splitting methods is based on the splitting of a complex problem into a sequence of simpler sub-problems. Usually, one exploits some structural properties of the separated operators belonging to the sub-problems, for example, the linear behavior, the symmetric behavior or the stiff behavior that allows the application of efficient integration methods to the sub-problems, see for instance [8, 10–12]. For dynamical problems like ODEs or parabolic PDEs, additive operator splitting are well established and appropriate. However, for constrained problems an additive operator splitting method would usually fail. This becomes obvious when comparing the simple problems

$$u' = Au = A_1u + A_2u \qquad \text{and} \qquad Ax = A_1x + A_2x = b.$$

M. Diab (✉) · C. Tischendorf
Humboldt Universität zu Berlin, Institute of Mathematics, Berlin, Germany
e-mail: m.diab@stimulate-ejd.eu; tischendorf@math.hu-berlin.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_21

Solving $u' = A_1 u$ and afterwards using its solution as an initial condition to solve $u' = A_2 u$ yields an approximate solution of $u' = Au$, while solving $Ax = b$ in the same manner, does not make sense. Here, a multiplicative splitting $Ax = A_1 A_2 x$ would be appropriate to solve $Ax = b$ by $A_1 y = b$ and afterwards $A_2 x = y$. It shows us that there is no simple extension of operator splitting for ODEs to DAEs. One has to adapt the operator splitting for DAEs to the different nature of inherent DAE parts.

The next section describes the branch oriented circuit modeling. It provides a natural decoupling of the circuit DAEs that can be exploited for a suitable operator splitting. Section 3 describes our operator splitting approach for linear circuit DAEs of index 1 [15]. It includes a convergence analysis and a discussion of some structural properties of the subsystems (for instance Hamiltonian structure of the first subsystem). Finally we demonstrate numerical results for a benchmark circuit in Sect. 4.

2 Circuit Modeling

In contrast to standard circuit modeling using the modified nodal analysis [9] we consider the branch oriented loop-cutset modeling [3, 4]. It allows us to split the operators in a natural way exploiting physical properties.

For a given circuit graph $\mathscr{G}$ with n node and b branches, select any tree and remove all its links. Then replace each link once at a time, it will form a loop that is called as fundamental loop. We select an orientation of the loop to coincide with that of the link completing it. On the other hand, a fundamental cutset with reference to a tree is a cutset formed with one tree branch and remaining links. The orientation of a cutset is the same of that of the tree branch.

Definition 1 The fundamental loop matrix $B \in \mathbb{R}^{b-(n-1)\times b}$ is defined by its entries

$$b_{ij} = \begin{cases} 1, & \text{if the branch } j \text{ has the same orientation of fundamental loop } i \\ -1, & \text{if the branch } j \text{ has the opposite orientation of fundamental loop } i \\ 0, & \text{else.} \end{cases}$$

Lemma 1 (Loop Equations, KVL [3]) *Let v be the vector of branch voltages in an electric network, then we have*

$$Bv = 0 \tag{1}$$

In general, matrix B is arranged such that the first columns correspond to entries of links and the columns correspond to entries of tree branches, therefore

$$B = \begin{pmatrix} B_l & B_t \end{pmatrix} = \begin{pmatrix} I & B_t \end{pmatrix}$$

Definition 2 The fundamental cutset matrix $Q \in \mathbb{R}^{(n-1)\times b}$ is defined by its entries

$$q_{ij} = \begin{cases} 1, & \text{if the branch } j \text{ has the same orientation of cutset } i \\ -1, & \text{if the branch } j \text{ has the opposite orientation of cutset } i \\ 0, & \text{else.} \end{cases}$$

Lemma 2 (Cut-set Equations, KCL [3]) *Let i be the vector of branch currents in an electric network, then we have*

$$Qi = 0 \tag{2}$$

and similar to the columns re-arrangement of B, we get $Q = \begin{pmatrix} Q_l & Q_t \end{pmatrix} = \begin{pmatrix} Q_l & I \end{pmatrix}$.

Theorem 1 (Orthogonality Relation [3]) *For a given connected graph $\mathcal{G}$, the orthogonality relation between the fundamental loop matrix B and the fundamental cutset matrix Q is given by $BQ^\top = 0$.*

The circuit equations consist of the loop equations (1) and cutset equations (2) reflecting the Kirchhoff's laws together with elements constitutive equations

$$i_C = Cv_C', \quad v_L = Li_L', \quad i_G = Gv_G, \quad v_R = Ri_R, \quad i_I = i_s(t), \quad v_V = v_s(t). \tag{3}$$

For simplicity, we consider only RLC circuits since our focus is to demonstrate the new splitting approach. We assume that all resistances, conductances, capacitances and inductances show a globally passive behavior, i.e. their corresponding matrices R, G, C and L are positive definite. In addition, the independent functions v_s and i_s for voltage and current sources are assumed to be continuously differentiable. Notice that, we used in our approach the conductive description for all resistances that belong to the tree and the resistive description for all resistances that does not belong to the tree, see below.

An index-1 circuit DAE models a circuit network that does neither have an LI-cutset nor a CV-loop, see [5]. Then we can construct a tree as follows [14]:

1. All capacitive elements and voltage sources belong to the tree.
2. All inductive elements and current sources do not belong to the tree.
3. Split resistors in such a way that all G-resistances belong to the tree and all R-resistances do not belong to the tree.

Then, the loop and cutset equations have the form

$$\begin{pmatrix} v_L \\ v_R \\ v_I \end{pmatrix} + B_t \begin{pmatrix} v_C \\ v_G \\ v_V \end{pmatrix} = 0, \qquad Q_l \begin{pmatrix} i_L \\ i_R \\ i_I \end{pmatrix} + \begin{pmatrix} i_C \\ i_G \\ i_V \end{pmatrix} = 0.$$

Inserting the element constitutive equations we get the DAE system

$$\begin{pmatrix} Li_L' \\ Ri_R \\ v_I \end{pmatrix} + B_t \begin{pmatrix} v_C \\ v_G \\ v_s(t) \end{pmatrix} = 0, \qquad Q_l \begin{pmatrix} i_L \\ i_R \\ i_s(t) \end{pmatrix} + \begin{pmatrix} Cv_C' \\ Gv_G \\ i_V \end{pmatrix} = 0.$$

Notice that $Q_l = -B_t^\top$ due to Lemmas 1, 2 and Theorem 1. Introducing

$$B_t =: \begin{pmatrix} B_{LC} & B_{LG} & B_{LV} \\ B_{RC} & B_{RG} & B_{RV} \\ B_{IC} & B_{IG} & B_{IV} \end{pmatrix}, \quad Q_l =: \begin{pmatrix} Q_{CL} & Q_{CR} & Q_{CI} \\ Q_{GL} & Q_{GR} & Q_{GI} \\ Q_{VL} & Q_{VR} & Q_{VI} \end{pmatrix}$$

and reordering the equations, we obtain a system of the port-Hamiltonian form

$$Dx'(t) + Jx(t) + My(t) = r_x(t) \tag{4a}$$

$$-M^T x(t) + Sy(t) = r_y(t) \tag{4b}$$

$$z(t) + K_x x(t) + K_y y(t) = r_z(t) \tag{4c}$$

with $x = \left(\begin{smallmatrix} i_L \\ v_C \end{smallmatrix}\right)$, $y = \left(\begin{smallmatrix} i_R \\ v_G \end{smallmatrix}\right)$, $z = \left(\begin{smallmatrix} v_I \\ i_V \end{smallmatrix}\right)$,

$$D = \left(\begin{smallmatrix} L & 0 \\ 0 & C \end{smallmatrix}\right), \quad J = \left(\begin{smallmatrix} 0 & B_{LC} \\ Q_{CL} & 0 \end{smallmatrix}\right), \quad M = \left(\begin{smallmatrix} 0 & B_{LG} \\ Q_{CR} & 0 \end{smallmatrix}\right), \quad S = \left(\begin{smallmatrix} R & B_{RG} \\ Q_{GR} & G \end{smallmatrix}\right)$$

and

$$K_x = \left(\begin{smallmatrix} 0 & B_{IC} \\ Q_{VL} & 0 \end{smallmatrix}\right), K_y = \left(\begin{smallmatrix} 0 & B_{IG} \\ Q_{VR} & 0 \end{smallmatrix}\right), r_x = -\left(\begin{smallmatrix} B_{LV} v_s \\ Q_{CI} i_s \end{smallmatrix}\right), r_y = -\left(\begin{smallmatrix} B_{RV} v_s \\ Q_{GI} i_s \end{smallmatrix}\right), r_z = -\left(\begin{smallmatrix} B_{IV} v_s \\ Q_{VI} i_s \end{smallmatrix}\right).$$

Notice that J is skew-symmetric since $B_{LC} = -Q_{CL}^\top$. Furthermore,

$$S = S_1 + S_2 := \begin{pmatrix} R & 0 \\ 0 & G \end{pmatrix} + \begin{pmatrix} 0 & B_{RG} \\ Q_{GR} & 0 \end{pmatrix}$$

with the positive definite diagonal matrix S_1 and the skew-symmetric matrix S_2 since $B_{RG} = -Q_{GR}^\top$. Consequently, S is not symmetric (unless $B_{RG} = 0$) but positive definite and hence non-singular. Furthermore, we see that system (4) is a port-Hamiltonian DAE in the sense of the definitions given in [6] and [13]. For [13], one can choose $\tilde{x} = (x, y)$, $\tilde{z}(\tilde{x}) = \tilde{x}$, $\tilde{y} := -z$ and $\tilde{u} = (i_s, v_s)$ where the

tilde notation refers to the variables in [13]. For [6], one can choose the space $\mathscr{V} = \{(x, y, z) : z + K_x x + K_y y = r_z\}$ with $\bar{x} = (x, y, z)$, $\bar{z}(\bar{x}) = (x, y)$, $\bar{y} := \bar{B}^\top \bar{z}(\bar{x})$ and $\bar{u} = (i_s, v_s)$, where the bar notation refers to the variables in [6]. Since (4c) can be interpreted as output equation for z, we consider only the reduced DAE system (4a)–(4b) in the following.

3 Operator Splitting for Index-1 Circuit DAEs

Regarding the fact that additive splitting makes no sense for solving the constraints (4b), we propose a splitting approach based on the inherent ODE. Therefore, we rewrite the DAE system (4a)–(4b) equivalently as

$$Dx' + Jx + MS^{-1}M^\top x = r_x(t) - MS^{-1}r_y(t) \tag{5a}$$

$$y = S^{-1}\big(r_y(t) + M^\top x\big). \tag{5b}$$

We split (5a), using Lie-Trotter splitting, into the subsystems $Dx' + Jx = 0$ and

$$Dx' + MS^{-1}M^T x = r_x(t) - MS^{-1}r_y(t). \tag{6}$$

Next, we reformulate (6) with (5b) back as DAE and obtain the following splitting approach (SADAE) for circuit index-1 DAEs.

1. Initialize $x_2(t_0) := x_0$ and $n = 0$.
2. Solve on $[t_n, t_{n+1}]$ the first subsystem

$$Dx_1' + Jx_1 = 0, \quad x_1(t_n) = x_2(t_n) \tag{splitDAE 1}$$

3. Solve on $[t_n, t_{n+1}]$ the second subsystem

$$Dx_2'(t) + My(t) = r_x(t), \quad x_2(t_n) = x_1(t_{n+1}) \tag{splitDAE 2a}$$

$$-M^T x_2(t) + Sy(t) = r_y(t). \tag{splitDAE 2b}$$

4. Set $n = n + 1$ and go to 2. unless t_n is the final time point.

3.1 Subsystem Properties

The first subsystem (splitDAE 1) is in fact a Hamiltonian ODE system with the Hamiltonian

$$H(x) = \frac{1}{2}x^\top Dx = \frac{1}{2}v_C^\top C v_C + \frac{1}{2}i_L^\top L i_L =: H(v_C, i_L) \tag{7}$$

describing the total energy stored in the capacitors and inductors. We have

$$\frac{d}{dt}H(x) = x^\top Dx' = -x^\top Jx = 0$$

since J is skew-symmetric. Obviously, H is a quadratic form. Consequently, we can apply symplectic numerical methods to (splitDAE 1). They have the advantage to preserve the total energy H stored in the capacitors and inductors [7].

The second subsystem (splitDAE 2a)–(splitDAE 2b) leads to non-symmetric but positive definite linear systems after time discretization that allows the exploitation of suitable iterative methods [2].

3.2 Convergence Analysis

In order to verify the convergence of DAE operator splitting method, one has to rely on the convergence of the ODE operator splitting method. For this reason, we define the non-homogeneous Cauchy problem

$$u'(t) = A_1u(t) + A_2u(t) + r(t), \quad u(t_0) = u_0 \tag{8}$$

where the initial condition u_0 and the source function r are bounded. Let Δt denotes the time step such that the following stability condition is satisfied

$$||e^{\Delta t(A_1+A_2)}|| \leq 1, \quad ||e^{\Delta tA_1}|| \leq 1, \quad \text{and} \quad ||e^{\Delta tA_2}|| \leq 1$$

After time discretization, apply the following operator splitting algorithm (OSA)

$$\begin{cases} u_1'(t) = A_1u_1(t), & t \in [t_n, t_{n+1}] \quad \text{and} \quad u_1(t_n) = u_{sp}^n \\ u_2'(t) = A_2u_2(t) + r(t), & t \in [t_n, t_{n+1}] \quad \text{and} \quad u_2(t_n) = u_1(t_{n+1}) \end{cases}$$

where $u_{sp}^0 = u_0$, and the splitting solution at $t = t_{n+1}$ is $u_{sp}^{n+1} = u_2(t_{n+1})$.

Theorem 2 (See [1]) *Under the boundedness and stability conditions formulated above, the approximated splitting solution obtained from the operator splitting algorithm (OSA) converges to the exact solution of the ODE* (8).

If we denote by $T(t_n)$ the solution operator of (8) at the n-th time step, and by $T_s(t_n)$ the splitting solution operator, then we have: $||T(t_n)u_0 - T_s(t_n)u_0|| \to 0$ as $\Delta t \to 0$. Regarding the equivalence of the DAE system (splitDAE 2a)–(splitDAE 2b) to the system

$$Dx_2' + MS^{-1}M^\top x_2 = r_x(t) - MS^{-1}r_y(t) \tag{9a}$$

$$y = S^{-1}\big(r_y(t) + M^\top x_2\big). \tag{9b}$$

we may directly conclude the following theorem from Theorem 2 (choosing $A_1 = -D^{-1}J$ and $A_2 = -D^{-1}MS^{-1}M^\top$ and $r = D^{-1}(r_x - MS^{-1}r_y)$).

Theorem 3 *Let the time stepsize Δt be sufficiently small, the initial currents and voltages as well as the source functions of current and voltage sources be bounded. Then, the approximated solution of the circuit DAE operator splitting approach (SADAE) on page 215 converges to the exact solution of the DAE (4a)–(4b).*

4 Numerical Simulation

We use a small RLC circuit example in order to demonstrate the operator splitting approach for DAEs. It operates in a GHz regime as often used in chip design.

Using the tree in Fig. 1, we get for the circuit DAE system (4a)–(4c) the matrices

$$D = \begin{pmatrix} L_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & L_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & L_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & C_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & C_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & C_3 \end{pmatrix}, \quad J = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad r_x = \begin{pmatrix} -v_s \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

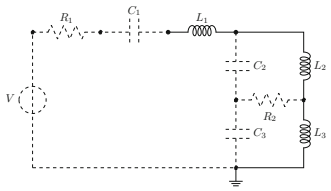
and

$$S = \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix}, \quad K_x = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad K_y = 0, \quad r_y = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad r_z = 0.$$

For comparison, we consider the following three variants of numerical simulation of the circuit:

1. Solve (4a)–(4c) by implicit Euler method.
2. Solve (splitDAE 1) and (splitDAE 2a)–(splitDAE 2b) by implicit Euler method
3. Solve (splitDAE 1) by symplectic Euler and (splitDAE 2a)–(splitDAE 2b) by implicit Euler method

In Fig. 2 we see the reference solution computed by time stepsize $h = 1e - 13$ and the error between the numerical solution for the three simulation variants with



$C_1 = 10^{-12}$ F, $C_2 = 5 \cdot 10^{-13}$ F, $C_3 = 10^{-12}$ F
$L_1 = 5 \cdot 10^{-7}$ H, $L_2 = 5 \cdot 10^{-7}$ H, $L_3 = 5 \cdot 10^{-7}$ H
$G_1 = 50\ \Omega$, $G_2 = 50\ \Omega$, $v_s(t) = \sin(10^9 t)$

Fig. 1 Benchmark RLC-circuit. The dashed branches form the tree considered for the model equations

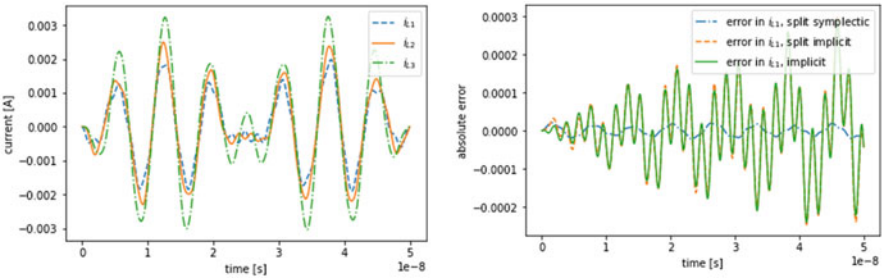


Fig. 2 Reference solution for inductive currents for circuit in Fig. 1(left). Error for numerical solution of the three simulation variants with time stepsize $h = 1e - 11$ (right)

time stepsize $h = 1e - 11$ and the reference solution. The results show that the solution of the DAE splitting approach (variant 2) is almost the same as for the non-splitted solution (variant 1). It means that the error caused by splitting is neglectable in comparison with the numerical discretization error. The use of the DAE splitting approach with the symplectic Euler method (variant 3) gives the best results and is even faster than the other variants since the symplectic Euler method for the first subsystem (5a) is an explicit method.

5 Conclusions and Outlook

In this paper, we extended the operator splitting method from ODEs to circuit linear DAEs. Followed by the topological decoupling of circuit DAEs of index 1 in loop-cutset formulation, we were able to construct a suitable decomposition of the matrices so that a natural port-Hamiltonian DAE structure is visible and can be exploited for a convergent splitting approach that is explicit and energy preserving in the dynamic part.

Acknowledgments This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 76504. Furthermore, we acknowledge financial support by DFG under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, ID 390685689).

References

1. M. Bjórhus, Operator splitting for abstract cauchy problems. IMA J. Numer. Anal. **18**, 419–443 (1988)
2. A.T. Chronopoulos, s-step iterative methods for (non)symmetric (in)definite linear systems. SIAM J. Numer. Anal. **28**(6), 1776–1789 (1991)

3. L.O. Chua, C.A. Desoer, E.S. Kuh, *Linear and Nonlinear Circuits* (McGraw-Hill, Singapore, 1987)
4. C.A. Desoer, E.S. Kuh, Basic Circuit Theory. International student edition (McGraw-Hill, New York, 1984)
5. D. Estévez Schwarz, C. Tischendorf, Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theory Appl. **28**(2), 131–162 (2000).
6. M. Günther, A. Bartel, B. Jacob, T. Reis, Dynamic iteration schemes and port-Hamiltonian formulation in coupled circuit simulation, 2020. arXiv:2004.12951v1
7. E. Hairer, G. Wanner, C. Lubich, *Geometric Numerical Integration*, vol. 31. Springer Series in Computational Mathematics (Springer, Berlin, Heidelberg, 2002)
8. E. Hansen, A. Ostermann, Dimension splitting for quasilinear parabolic equations. IMA J. Numer. Anal. **30**(3), 857–869 (2010)
9. C.-W. Ho, A. Ruehli, P. Brennan, The modified nodal approach to network analysis. IEEE Trans. Circ. Syst. **22**(6), 504–509 (1975)
10. M. Hochbruck, T. Jahnke, R. Schnaubelt, Convergence of an adi splitting for maxwell's equations. Numer. Math. **129**(3), 535–561 (2015)
11. H. Holden, K. Karlsen, K. Lie, N. Risebro, *Splitting Methods for Partial Differential Equations with Rough Solutions* (European Mathematical Society, Zürich, 2010)
12. W. Hundsdorfer, J.G. Verwer, A note on splitting errors for advection-reaction equations. Appl. Numer. Math. **18**(1), 191 – 199 (1995)
13. V. Mehrmann, R. Morandin, Structure-preserving discretization for port-Hamiltonian descriptor systems (2019) arXiv:1903.10451v1
14. R. Riaza, *Differential-Algebraic Systems: Analytical Aspects and Circuit Applications* (World Scientifc, Singapore, 2008)
15. C. Tischendorf, R. Lamour, R. März, *Differential-Algebraic Equations. A Projector Based Analysis* (Springer, Hamburg, 2012)

Reduced Order Modelling for Wafer Heating with the Method of Freezing

E. J. I. Hoeijmakers, H. Bansal, T. M. van Opstal, and P. A. Bobbert

Abstract Accurate and real-time temperature control for wafer heating is one of the main challenges in semiconductor manufacturing processes. With reduced-order modelling (ROM), the computational complexity of the mathematical model can be decreased in order to solve the model quickly at a low computational cost, while still maintaining the computational accuracy. However, the translating temperature profile, due to moving sources, render the standard reduction approaches to be ineffective. We propose to invoke the concept of the "Method of Freezing" and use it in conjunction with the standard ROM approaches to obtain an effective low-complexity model. We finally assess the effectiveness of the proposed approach on the 2-dimensional heat equation with moving heat loads. Numerical results clearly show the potential of the proposed approach over the standard one in terms of computational accuracy and the dimension of the resulting reduced-order model.

1 Introduction

In photolithography, feature sizes are decreasing in effort for manufacturers to keep up with Moore's law. This has prompted the use of higher energy lasers, leading to more wafer heating and, therefore, more thermal expansion. Accurate and real-time prediction of the temperature distribution around the moving laser beam is a necessity as this facilitates to correct the laser beam trajectory and to create

E. J. I. Hoeijmakers (✉) · P. A. Bobbert
Department of Applied Physics, Technical University of Eindhoven, Eindhoven, Netherlands
e-mail: p.a.bobbert@tue.nl

H. Bansal
Department of Mathematics and Computer Science, Technical University of Eindhoven, Eindhoven, Netherlands
e-mail: h.bansal@tue.nl

T. M. van Opstal
Sioux Mathware, Eindhoven, Netherlands
e-mail: timo.van.opstal@sioux.eu

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_22

the desired temperature at every place on the wafer [1]. However, this remains a challenge since standard numerical methods take a lot of computational time, and the increased resolution requirements due to the reduced feature sizes slow the model down.

Reduced-order modelling (ROM) reduces the model complexity and aids in real-time prediction of the quantity of interest. Translating temperature profiles, due to moving sources, render the standard ROM approaches ineffective [2]. Hence, we propose to invoke the "Method of Freezing" along with standard ROM approaches in order to obtain an effective low-complexity model computable in real-time.

The concept of the "Method of Freezing" has been applied on parabolic and hyperbolic problems in the past [3]. However, [4] is the only work which so far exploits the "Method of Freezing" for non-linear reduced basis approximations. This work considers a numerical experiment, which falls in the realm of hyperbolic problems, namely the parameterized Burgers-type problem in 2D (without source terms).

The main contribution of this work is to use the "Method of Freezing" in conjunction with standard ROM approaches to facilitate accurate and real-time prediction of the temperature. The "Method of Freezing" relies on an ansatz that decomposes the original dynamics into shape and travelling dynamics. The resulting shape dynamics is amenable for an efficient basis generation. We then use these generated bases to apply Proper Orthogonal Decomposition (POD) on the shape dynamics and, ultimately, obtain a reduced-order model. We finally assess the performance of the combined approach of the "Method of Freezing" and reduced basis approximations on a test-case of practical relevance, and discuss the computational merits of the proposed ROM approach over the standard one.

The paper is organized as follows. In Sect. 2.1, we introduce the 2-dimensional heat equation and discuss the numerical method for its discretization. We invoke the idea of the "Method of Freezing", reformulate the model problem and present the corresponding discretized representation in Sect. 2.2. A Galerkin-type projection-based ROM is performed on a semi-discrete model representation in Sect. 3. A numerical case-study is presented in Sect. 4 to showcase the effectiveness of the proposed approach. Finally, Sect. 5 ends with conclusions and future works.

2 Theory

In this section, we first introduce the model and the numerical method employed for the spatio-temporal discretization. We then introduce the idea of the "Method of Freezing" and present a model reformulation and its discrete representation.

2.1 Model Introduction

To model the wafer heating, we use the well-known heat equation in two-dimensions. As the height of the wafer is one order of magnitude less than the length and the width of the wafer, the temperature gradient along the thickness of the wafer is very small. This makes the 2-dimensional heat equation a good approximation of the real situation. The 2-dimensional heat equation is governed by:

$$\frac{\partial u}{\partial t} - \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) = Q(x, y, t), \quad (x, y) \in \Omega, \quad t \in [0, t_f], \tag{1}$$

$$u(x, y, t = 0) = u_0, \tag{2}$$

$$n_x \frac{\partial u}{\partial x} + n_y \frac{\partial u}{\partial y} = 0 \text{ on } \partial\Omega, \tag{3}$$

where u represents the wafer temperature, u_0 stands for a constant initial temperature, Ω stands for the spatial domain of interest, $\boldsymbol{n} = (n_x, n_y)$ denotes the normal to the boundary $\partial\Omega$, t_f indicates the final simulation time, and α is the thermal diffusivity constant. The thermal diffusivity constant can be expressed with the thermal conductivity k, the specific heat capacity C_p and the density ρ of the wafer in the form $\alpha = \frac{k}{\rho C_p}$. Here, a moving heat load $Q(x, y, t)$ is assumed to be of the non-affine form:

$$Q(x, y, t) = e^{-\frac{1}{2}\left(\frac{x - c_x t}{\sigma_x}\right)^2 - \frac{1}{2}\left(\frac{y - c_y t}{\sigma_y}\right)^2}, \tag{4}$$

where c_x and c_y are the speeds of the heat load in the x- and y-direction, respectively and, the variance of the Gaussian distribution along the x- and the y-direction is given by σ_x^2 and σ_y^2, respectively.

After multiplying (1) by a smooth test-function w, integrating over the domain and invoking Green's theorem, a weak formulation of the 2-dimensional heat equation can be constructed, resulting in:

$$\int_\Omega \frac{\partial u}{\partial t} w dA + \alpha \left(\int_\Omega \frac{\partial u}{\partial x} \cdot \frac{\partial w}{\partial x} dA + \int_\Omega \frac{\partial u}{\partial y} \cdot \frac{\partial w}{\partial y} dA \right) - \alpha \int_{\partial\Omega} \frac{\partial u}{\partial n} w ds = \int_\Omega Q w dA, \tag{5}$$

where $dA = dxdy$ and ds is a boundary surface element. Using (3), the fourth term on the left-hand-side of (5) cancels out[5].

In order to solve (5) numerically, discretization in space and time is necessary. We discretize the domain such that the structured mesh aligns with the orientation of the features which need to be printed. We then employ a finite element method to discretize in space. We approximate the solution with a summation over B-spline

basis-functions ϕ_i, $u = \sum_{i=1}^{N} u_i(t)\phi_i(x)$ [6]. Here, N is the number of finite elements used in the domain discretization and u_i is the weight of every basis function. To discretize in time, the first-order backwards Euler method is applied as is also used in Chap. 8 of [5]. Discretizing in both space and time results in the following equation:

$$Mu^{k+1} + \Delta t \alpha D u^{k+1} - \Delta t \tilde{Q}^{k+1} = Mu^{k}, \tag{6}$$

where M is the mass matrix, D is the diffusion matrix, $\tilde{Q}$ is the source vector representative of the moving heat loads and Δt indicates the time-step. Equation (6) needs to be solved for every time instant $k + 1$.

The numerical solution will be at most first-order accurate if the first-order backwards Euler method is applied in conjunction with the higher-order spatial discretization. However, in this paper, we are not concerned about the order of accuracy of the numerical solution, but intend to show the potential of the "Method of Freezing". To this end, the first-order temporal discretization is representative enough for quantifying the numerical performance, while being simple to implement. The implementation of a higher-order temporal discretization is deferred to future works.

We will now discuss a change of coordinates or so-called "Method of Freezing" that we propose to use in conjunction with standard ROM techniques to obtain an effective complexity reduction for problems with moving heat load(s).

2.2 Model Reformulation: Method of Freezing

The "Method of Freezing" maps all symmetry-related solutions to a single class of solutions. This method separates the dynamics in the group direction from the dynamics in the remaining directions of the phase space. The general idea of this method is to perform a coordinate transformation of the form:

$$u(x, y, t) = v(x - c_x t, y - c_y t, t) = v(\xi_x, \xi_y, t), \tag{7}$$

Incorporating (7) in (1) results in the following modified heat equation:

$$\frac{\partial v}{\partial t} - c_x \frac{\partial v}{\partial \xi_x} - c_y \frac{\partial v}{\partial \xi_y} - \alpha \left(\frac{\partial^2 v}{\partial \xi_x^2} + \frac{\partial^2 v}{\partial \xi_y^2} \right) = Q(\xi_x, \xi_y). \tag{8}$$

This modified heat equation is quite similar to the original equation given in (1), except the additional second and third term on the left-hand side which represent an

extra convection term. The weak formulation of (8) under zero Neumann boundary conditions is given by:

$$\int_{\Sigma} \frac{\partial v}{\partial t} w d\xi_x d\xi_y - c_x \int_{\Sigma} \frac{\partial v}{\partial \xi_x} w d\xi_x d\xi_y - c_y \int_{\Sigma} \frac{\partial v}{\partial \xi_y} w d\xi_x d\xi_y$$
$$- \alpha \left(\int_{\Sigma} \frac{\partial v}{\partial \xi_x} \frac{\partial w}{\partial \xi_x} d\xi_x d\xi_y + \int_{\Sigma} \frac{\partial v}{\partial \xi_y} \frac{\partial w}{\partial \xi_y} d\xi_x d\xi_y \right) = \int_{\Sigma} Q(\xi_x, \xi_y) w d\xi_x d\xi_y, \tag{9}$$

where Σ represents the transformed domain as per the coordinate transformation.

Discretizing (9) in space and time yields:

$$M v^{k+1} + \alpha \Delta t D v^{k+1} - \Delta t C v^{k+1} - \Delta t \tilde{Q}^{k+1} = M v^k, \tag{10}$$

where M and D are, respectively, the mass and diffusion matrix, and C is the convection matrix.

Although we consider constant c_x and c_y, the "Method of Freezing" can handle time-dependent speeds by adding an ingredient known as phase conditions; see [3].

3 Reduced Order Modelling

In this section, we build a reduced-order model both via the standard and the proposed ROM approach. The standard and the proposed ROM approach, built upon a Galerkin type projection-based ROM methodology [7], is discussed in Sects. 3.1 and 3.2, respectively.

3.1 Standard Reduced Order Modelling Approach

The numerical solution of the 2-dimensional heat equation can be written as a u-snapshot matrix, where every column k represents the solution at the k-th time-step. Upon performing singular value decomposition (SVD) on the snapshot matrix composed of u, a projector $P^T : U_h \rightarrow U_r$ is obtained and further used to build a reduced-order model. Here, U_h is a h-dimensional high-fidelity space and U_r is a r-dimensional reduced space spanned by the functions obtained from a truncated singular value decomposition of the u snapshot matrix. The standard reduced-order model is given by:

$$M_{red} u_{red}^{k+1} + \alpha \Delta t D_{red} u_{red}^{k+1} - P^T \Delta t \tilde{Q}^{k+1} = M_{red} u_{red}^k, \tag{11}$$

where $D_{red} = P^T D P$ and $M_{red} = P^T M P$ are the reduced diffusion and mass matrices, respectively.

3.2 Proposed Reduced Order Modelling Approach

The proposed novel ROM approach employs the "Method of Freezing" in conjunction with standard projection-based reduction techniques. We again employ SVD. However, in this proposed framework, the SVD is performed on the v snapshot matrix, instead of the u snapshot matrix. We now obtain a projector $L^T : V_h \rightarrow V_r$ where V_h is a h-dimensional high-fidelity space and V_r is a r-dimensional reduced space spanned by the functions obtained from a truncated singular value decomposition of the v snapshot matrix. Finally, the proposed (frozen) reduced-order model is:

$$M_{red,p} v_{red}^{k+1} + \alpha \Delta t D_{red,p} v_{red}^{k+1} - \Delta t C_{red,p} v_{red}^{k+1} - L^T \Delta t \tilde{Q}^{k+1} = M_{red,p} v_{red}^{k}, \tag{12}$$

where $C_{red,p} = L^T C L$ represents the reduced matrix corresponding to the extra convection term, and, $M_{red,p} = L^T M L$ and $D_{red,p} = L^T D L$, respectively, represent the reduced mass and diffusion matrices.

4 Numerical Results

In this section, we numerically test the proposed (Freezing-POD) approach and show its effectiveness as a reduced-order modelling technique.

Let the domain of the wafer be given by $\Omega_d = [-0.01, 0.02]\text{m} \times [-0.02, 0.04]\text{m}$. The wafer is subdivided into 9 smaller rectangular sub-domains and each sub-domain has the dimensions 1 by 2 cm. The heat load will move around one of these sub-domains in practice. In order to not consider the boundary conditions close to the boundary edges of the wafer, we consider that the laser only moves over the middle sub-domain Ω, i.e., $\Omega = [0, 0.01]\text{m} \times [0, 0.02]\text{m}$. Motivated by the application, we consider u_0 in (2) to be equal to the room temperature, i.e., $u_0 = 298K$. Furthermore, we spatially discretize a rectangular sub-domain by a 20×20 mesh, i.e., 400 finite-elements. Moreover, we consider a silicon wafer with thermal diffusivity constant $\alpha = 8.8 \cdot 10^{-5}$ m^2/s [8]. Additionally, we assume that the laser has a surface of approximately 2 by 20 mm. As a result, the variance in the x-direction, σ_x^2, is 0.002 m, and the variance in the y-direction, σ_y^2, is 0.02 m. We take 50 steps in time for the scenario under consideration, i.e., $t \in [0, 0.05]$s with a time step of 0.001s. A laser is considered to move along the x-direction with a speed of 0.2 m/s for first 25 time steps, i.e., $c_x = 0.2$ m/s and $c_y = 0$ m/s for $t = [0, 0.025]$s and along the y-direction with a speed of 0.2 m/s for next 25 time steps, i.e., $c_x = 0$ m/s and $c_y = 0.2$ m/s for $t = (0.025, 0.05]$s.

We build the snapshot matrix composed of solution u obtained in (6) and another snapshot matrix composed of shape dynamics v obtained in (10). We then perform SVD on these snapshot matrices to obtain the corresponding singular values, whose

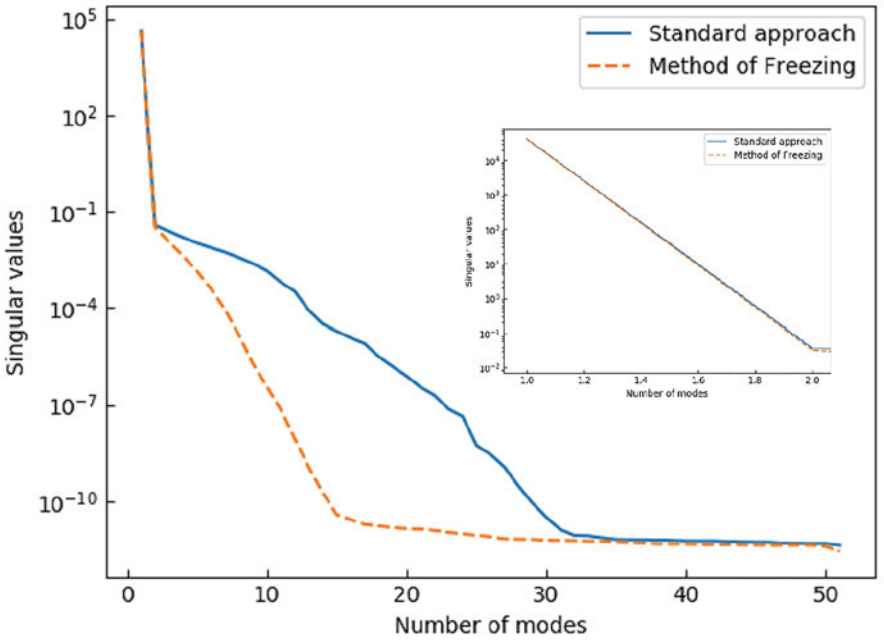


Fig. 1 Singular value decay behavior for the proposed and the standard approach

decay behavior is known to give a good expectation about the possible reduction in the dimensionality of the full-order model. In Fig. 1, the singular value decay behavior for the proposed and the standard ROM approach is shown. It can be observed that incrementing the number of POD modes by one yields a sharp initial decrease in the singular values both for the proposed and the standard ROM approach. However, post the sharp decay, we can see that the singular values corresponding to the proposed approach decay faster than the one corresponding to the standard approach. An initial sharp decrease is attributed to the fact that only a single mode is representative enough to capture the mean temperature on the silicon wafer. Other modes are required to accurately determine the change (with respect to the mean) in the temperature due to the moving heat loads. The observed decay behavior clearly indicates a possibility of an effective dimensionality reduction if the "Method of Freezing" is used together with the standard ROM techniques.

Further computational benefits of the proposed approach over the standard one can be clearly seen in Fig. 2, which shows the behavior of the reduced-order modelling (ROM) error for increasing dimensions of the reduced-order model. We assess the error of the standard and proposed approaches in the (absolute) L^2-norm in space and time. The error via the standard approach corresponds to the difference between the finite-element based numerical solution u and the reconstructed solution obtained by lifting the standard reduced-order solution u_{red}, obtained in (11), to the high-dimensional problem space. And, the error via

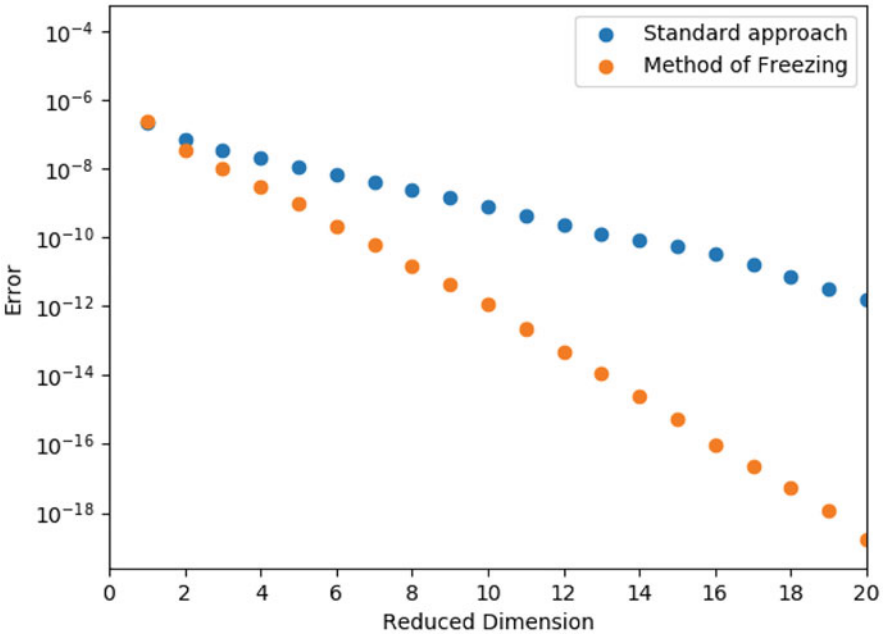


Fig. 2 ROM error for the proposed and the standard approach versus varying dimensions of the reduced-order model

the proposed approach corresponds to the difference between the finite-element based numerical solution u and the reconstructed solution obtained by lifting and shifting the reduced-order solution v_{red} obtained in (12). It is clearly observable that the (absolute) ROM error incurred upon using the proposed approach for varying dimensions of the reduced-order model is lower than the error incurred while using the standard approach. The proposed approach is expected to give a lower ROM error as the shape dynamics is essentially localized around the initial configuration. In principle, the ROM error is a function of the spatial and the temporal discretization error. Given the fact that the shape dynamics is essentially localized in the proposed approach, the amount of temporal discretization error is significantly less than that obtained in the standard approach. We also claim that the larger time-step size can be used to advance the reduced-order model built using the "Method of Freezing" compared to the admissible time-step size in the standard ROM framework. This claim is supported by the fact that the time-step size is generally controlled by the CFL restrictions, which are dictated by the time-scale of the problem. As a consequence of localized shape dynamics, the time-step size is not too severely restricted in the proposed approach as in the standard approach, which also eventually aids in temporal complexity reduction. As a result, the dimension of the reduced-order model obtained by using the proposed approach will be much

smaller than the counterpart obtained using the standard reduction approach in order to have the same accuracy.

5 Conclusion and Future Outlook

We have proposed to employ the concept of the "Method of Freezing" in conjunction with a Galerkin-type projection based methodology in order to overcome the limitation of the standard projection-based reduced-order modelling (ROM) techniques in dealing with moving heat loads. We have demonstrated the performance of the proposed approach on a test-case of practical relevance that encompasses the movement of the laser beam along both dimensions of the wafer.

This work focused on reproducing the results of the time-dependent heat equation via standard and proposed ROM approaches. This *reproduction* step is essential before attempting to develop a parametric reduced-order model as we cannot hope to have an effective low-complexity reduced-order model if the numerical approach does not fare well in the *reproduction* step. Furthermore, it should be emphasized that the considered model is non-affine due to the nature of the moving heat load(s), and that the projection alone is not sufficient to reduce the costs of constructing a reduced-order model for such non-affine (and non-linear) problems. Moreover, there might be other sources of non-affine and/or non-linear nature, such as radiative heat fluxes, temperature-dependence of parameters, etc. These non-affine and non-linear problems can be effectively dealt with the proposed ROM approach by using an additional concept of hyper-reduction introduced in [9].

Future works will deal with a modification to the idea of the "Method of Freezing" to eventually obtain a suitable decomposition ansatz that accounts for the physical boundary conditions. In addition, the effectiveness of the proposed approach will be investigated in terms of the computational speed-up. Moreover, the "Method of Freezing" in conjunction with standard projection-based ROM approaches and hyper-reduction will be used to develop a framework for parametric ROM.

Acknowledgments G. van Zwieten, J. van Zwieten, C. Verhoosel, E. Fonn, T. van Opstal, & W. Hoitinga. (2019, June 11). Nutils (Version 5.0). Zenodo. https://doi.org/10.5281/zenodo.3243447.

References

1. M. Rabus, A.T. Fiory, N.M. Ravindra, P. Frisella, A. Agarwal, T. Sorsch, J. Miner, E. Ferry, F. Klemens, R. Cirelli et al., Rapid thermal processing of silicon wafers with emissivity patterns. J. Electron. Mater. **35**(5), 877–891 (2006)
2. M. Ohlberger, S. Rave, Reduced basis methods: Success, limitations and future challenges, in *Proceedings of the Conference Algoritmy* (2016), pp. 1–12

3. W.J. Beyn, V. Thummler, Freezing solutions of equivariant evolution equations. SIAM J. Appl. Dynam. Syst. **3**(2), 85–116 (2004)
4. M. Ohlberger, S. Rave, Nonlinear reduced basis approximation of parameterized evolution equations via the method of freezing. Comptes Rendus Mathematique **351**(23–24), 901–906 (2013)
5. T.J.R. Hughes, *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis* (Courier Corporation, North Chelmsford, 2012)
6. L. Piegl, W. Tiller, Curve and surface constructions using rational B-splines. Comput.-Aided Des. **19**(9), 485–498 (1987)
7. P. Benner, W.H.A. Schilders, S. Grivet-Talocia, A. Quarteroni, G. Rozza, M. Silveira Luís, *Snapshot-Based Methods and Algorithms. Model Order Reduction*, vol. 2 (De Gruyter, Berlin, Boston, 2020)
8. R. Hull, Properties of crystalline silicon. No. 20. IET (1999)
9. M. Barrault, Y. Maday, N.C. Nguyen, A.T. Patera, An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. Comptes Rendus Math. **339**(9), 667–672 (2004)

Multirate DAE-Simulation and Its Application in System Simulation Software for the Development of Electric Vehicles

Michael Kolmbauer, Günter Offner, Ralf Uwe Pfau, and Bernhard Pöchtrager

Abstract This work is devoted to the efficient simulation of large multi-physical networks stemming from automated modeling processes in system simulation software. The simulation of hybrid, battery and fuel cell electric vehicle applications requires the coupling of electric, mechanic, fluid and thermal networks. Each network is established by combining the connection structure of a graph with physical equations of elementary components and resulting in a differential algebraic equation (DAE). In order to speed up the simulation a non-iterative multirate time integration co-simulation method for the system of coupled DAEs is established. The power of the multirate method is shown via two representative examples of a battery powered electric vehicle with a cooling system for the battery pack and a three phase inverter with a cooling system.

M. Kolmbauer (✉)
MathConsult GmbH, Linz, Austria
e-mail: michael.kolmbauer@mathconsult.co.at

G. Offner · R. U. Pfau
AVL List GmbH, Graz, Austria
e-mail: guenter.offner@avl.com; ralf-uwe.pfau@avl.com

B. Pöchtrager
Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz, Austria
e-mail: bernhard.poechtrager@ricam.oeaw.ac.at

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_23

1 Background and Introduction

State-of-the-art modeling and simulation packages such as AVL CRUISE™M[1], Dymola[2], or Amesim[3] offer many concepts for the automatic generation of dynamic system models. Modeling is done in a modularized way, based on a network of subsystems which again consists of simple standardized subcomponents. For instance, in case of HEVs (hybrid electric vehicles), BEVs (battery electric vehicles) and FCEVs (fuel cell electric vehicles) these can be the vehicle chassis, the drive line, the air path of the ICE (internal combustion engine) including combustion and exhaust aftertreatment, the cooling and lubrication system of the ICE and battery packs, the electrical propulsion system including the engine and a battery pack, the air conditioning and passenger cabin models, waste heat recovery and finally according control systems. Due to the complex interaction of the subsystems, the challenges in the development of future power trains do not only lie in the design of individual components but in the assessment of the power train as a whole. On a system engineering level it is required to optimize individual components globally and to balance the interaction of different subsystems. Due to the increasing complexity of the models, the systems exhibit largely varying time scales and are difficult in the numerical handle. A mainly automatized multirate approach is a promising way to decrease the computational effort.

The structure of the work is the following: In Sect. 2 the individual physical networks are introduced and the coupling conditions are stated in order to obtain a fully coupled system of network DAEs. The multirate time integration technique for the coupled system of network DAEs is described in Sect. 3 and the corresponding numerical results are presented in Sects. 4 and 5. Finally we conclude in Sect. 6.

2 Problem Formulation

We consider a network that is composed of multi-physical elements. The network elements describing the electric contribution are given by current sources, voltage sources, nodes, ground, resistors, capacitors and inductors. The fluid network consists of pipes, pumps, demands, junctions and reservoirs. The electro-thermal coupling is established by lumped mass elements representing the pipe wall and the masses from the battery and heat transfer connections. The individual components are assembled to a network $\mathcal{N}$, which is represented by a linear directed graph. The graph structure is described by an incidence matrix A, which can be used for the model descriptions, cf. [8]. In the following we state the DAEs for the three main involved physical networks.

[1] https://www.avl.com/de/cruise-m.

[2] http://www.dynasim.com.

[3] http://www.plm.automation.siemens.com.

Electric Network

We consider an electric network $\mathcal{N}_E = \{R, C, L, V, I, N, G, B\}$ that is composed of resistors R, capacitors C, inductors L, voltage sources V, current sources I, nodes N, grounds G and batteries B. The DAE for the network in $\mathcal{N}_E$ in input-output form is given by: For given continuous inputs $(u_R^T, u_C^T, u_B^T)^T$ find the potentials $e = (e_N^T, e_G^T)^T$, the currents $j = (j_R^T, j_C^T, j_L^T, j_V^T, j_B^T)^T$ and the outputs $y = (y_R^T)^T$, such that

$$\begin{aligned} A_R j_R + A_C j_C + A_L j_L + A_V j_V + A_V j_B + A_I \bar{j}_I &= 0 \\ r(u_R) j_R - A_R^T e &= 0 \\ j_C - \frac{d(c(u_C) A_C^T e)}{dt} &= 0 \\ l \frac{d j_L}{dt} - A_L^T e &= 0 \\ A_V^T e &= \bar{v}_V \\ A_B^T e &= \bar{v}_B(j_B, u_B) \\ y_R &= j_R A_R^T e \end{aligned} \tag{1}$$

for given boundary conditions $e_G = 0$ and given resistance r, capacitance c and inductance l as well as prescribed currents $\bar{j}_I$ and prescribed voltages $\bar{v}_V$ and $\bar{v}_B$. The coupling variables are expressed as temperature of the resistor u_R, the capacitor u_C and the battery u_B as well as the energy flux of the resistor y_R.

Solid Network

We consider a solid network $\mathcal{N}_S = \{SW, LW, HT, HS, TB\}$ that is composed of solid walls SW, lumped walls LW, heat transfers HT, heat sources HS and temperature boundaries TB. The DAE for the network $\mathcal{N}_S$ in input-output form is given by: For given continuous inputs $(u_{Hs_S}^T, u_{Tb_S}^T)^T$, find the temperatures $(T_{Sw}^T, T_{Lw}^T)^T$, the heat fluxes $(H_{Ht_S}^T)^T$ and the outputs $(y_{Sw}^T, y_{Lw}^T, y_{Ht_S}^T)^T$, such that

$$\begin{aligned} m_{Sw} c_{p,Sw} \frac{dT_{Sw}}{dt} &= A_{Sw,Ht_S} H_{Ht_S} + A_{Sw,Hs} H_{Hs} + A_{Sw,Hs_u} u_{Hs_S} \\ 0 &= A_{Lw,Ht_S} H_{Ht_S} + A_{Lw,Hs} H_{Hs} + A_{Lw,Hs_u} u_{Hs_S} \\ H_{Ht_S} &= c_{Ht_S} \left(A_{Sw,Ht_S}^T T_{Sw} + A_{Lw,Ht_S}^T T_{Lw} + A_{Tb,Ht_S}^T T_{Tb} + A_{Tb_u,Ht_S}^T u_{Tb_S} \right) \\ y_{Sw} &= |(A_{Sw,Hs_u}^T + A_{Tb_u,Ht_S} A_{Sw,Ht_S}^T)| T_{Sw} \\ y_{Lw} &= |(A_{Lw,Hs_u}^T + A_{Tb_u,Ht_S} A_{Lw,Ht_S}^T)| T_{Lw} \\ y_{Ht_S} &= A_{Tb_u,Ht_S} H_{Ht_S} \end{aligned} \tag{2}$$

for given boundary conditions $H_{Hs} = \bar{H}_{Hs}$ and $T_{Tb} = \bar{T}_{Tb}$ and given positive definite coefficient matrices m_{Sw}, $c_{p,Sw}$ and c_{Ht_S}. The coupling variables are expressed as the energy fluxes u_{Hs_S} and u_{Tb_S} and the temperatures y_{Sw}, y_{Lw} and y_{Ht_S}.

Fluid Network

We consider a fluid network $\mathcal{N}_F = \{PI, PU, DE, VJ, LJ, RE, HT, TB\}$ that is composed of pipes PI, pumps PU, demands DE, volume junctions VJ, lumped junctions LJ, reservoirs RE, heat transfers HT and temperature boundaries TB. The DAE for the network $\mathcal{N}_F$ in input-output form is given by: For given continuous inputs $(u_{Hs_F}^T, u_{Tb_F}^T)^T$, find the pressures $(p_{Lj}^T, p_{Vj}^T)^T$ the mass flows $(q_{Pi}^T, q_{Pu}^T)^T$, the temperatures $(T_{Vj}^T, T_{Lj}^T)^T$, the heat fluxes $(H_{Ht_F}^T, H_{Pu}^T, H_{Pi}^T)^T$ and the outputs $(y_{Vj}^T, y_{Lj}^T, y_{Ht_F}^T)^T$, such that

$$
\begin{aligned}
\frac{dq_{Pi}}{dt} &= c_{1,Pi}\left(A_{Jc,Pi}^T p_{Jc} + A_{Re,Pi}^T p_{Re}\right) + c_{2,Pi}\mathrm{diag}\left(|q_{Pi}|\right) q_{Pi} + c_{3,Pi} \\
f_{Pu}(q_{Pu}) &= A_{Jc,Pu}^T p_{Jc} + A_{Re,Pu}^T p_{Re} \\
0 &= A_{Jc,Pi} q_{Pi} + A_{Jc,Pu} q_{Pu} + A_{Jc,De} q_{De} \\
m_{Vj} c_{p,Vj} \frac{dT_{Vj}}{dt} &= A_{Vj,Pi} H_{Pi} + A_{Vj,Pu} H_{Pu} \\
&\quad + A_{Vj,De} H_{De} + A_{Vj,Ht_F} H_{Ht_F} + A_{Vj,Hs_u} u_{Hs_F} \\
0 &= A_{Lj,Pi} H_{Pi} + A_{Lj,Pu} H_{Pu} \\
&\quad + A_{Lj,De} H_{De} + A_{Lj,Ht_F} H_{Ht_F} + A_{Lj,Hs_u} u_{Hs_F} \\
H_{Pi} &= B_{Jc}(q_{Pi}) T_{Vj} + B_{Jc}(q_{Pi}) T_{Lj} + B_{Jc}(q_{Pi}) T_{Re} \\
H_{Pu} &= B_{Jc}(q_{Pu}) T_{Vj} + B_{Jc}(q_{Pu}) T_{Lj} + B_{Jc}(q_{Pu}) T_{Re} \\
H_{Ht_F} &= c_{Ht_F}\left(A_{Vj,Ht_F}^T T_{Vj} + A_{Lj,Ht_F}^T T_{Lj} + A_{Tb_u,Ht_F}^T u_{Tb_F}\right) \\
y_{Vj} &= |(A_{Vj,Hs_u}^T + A_{Tb_u,Ht_F} A_{Vj,Ht_F}^T)| T_{Vj} \\
y_{Lj} &= |(A_{Lj,Hs_u}^T + A_{Tb_u,Ht_F} A_{Lj,Ht_F}^T)| T_{Lj} \\
y_{Ht_F} &= (A_{Tb_u,Ht_F} + A_{Lj,Hs_u}^T A_{Lj,Ht_F}^T + A_{Vj,Hs_u}^T A_{Vj,Ht_F}^T) H_{Ht_F}
\end{aligned}
\tag{3}
$$

for given boundary conditions $q_{De} = \bar{q}_{De}$, $H_{De} = \bar{H}_{De}$, $p_{Re} = \bar{p}_{Re}$ and $T_{Re} = \bar{T}_{Re}$ and given coefficients $c_{1,Pi}$, $c_{2,Pi}$, $c_{1,Pi}$, m_{Vj}, $c_{p,Vj}$ and c_{Ht_F} as well as given functions f_{Pu}. The function B_{Jc} checks for the sign of the mass flow q_{Pi}, cf. [4]. The coupling variables are expressed as the temperatures u_{Hs_F} and u_{Tb_F} and the energy fluxes y_{Vj}, y_{Lj} and y_{Ht_F}.

Multi-Physical Model

The multi-physical model is derived by combining (1), (2) and (3) with appropriate coupling conditions. The coupling conditions describe the relation between the inputs and outputs of the individual models. For the model used in Sects. 4 and 5, the following coupling conditions are used, see e.g. [9].

$$\begin{pmatrix} u_R \\ u_C \\ u_B \\ u_{Hs_S} \\ u_{Tb_S} \\ u_{Hs_F} \\ u_{Tb_F} \end{pmatrix} = \begin{pmatrix} 0 & C_{R,Sw} & 0 & 0 & 0 & 0 & 0 \\ 0 & C_{C,Sw} & 0 & 0 & 0 & 0 & 0 \\ 0 & C_{B,Sw} & 0 & 0 & 0 & 0 & 0 \\ C_{Hs_S,R} & 0 & 0 & 0 & 0 & 0 & C_{Hs_S,Ht_F} \\ 0 & 0 & 0 & 0 & C_{Tb_S,Vj} & 0 & 0 \\ 0 & 0 & 0 & C_{Hs_F,Vj} & 0 & 0 & 0 \\ 0 & C_{Tb_F,Sw} & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_R \\ y_{Sw} \\ y_{Lw} \\ y_{Ht_S} \\ y_{Vj} \\ y_{Lj} \\ y_{Ht_F} \end{pmatrix} \tag{4}$$

The connectivity equation (4) represents the electro-thermal coupling of the electric network and the cooling systems. Combining all subsystems and their connectivity equations (4) yields a DAE:

Find

$$z := (u_R, u_C, u_B, u_{Hs_S}, u_{Tb_S}, u_{Hs_F}, u_{Tb_F}, e_N, e_G, j_R, j_C, j_L, j_V, j_B, T_{Sw}, T_{Lw}, H_{Ht_S}, p_{Lj}, p_{Vj} q_{Pi}, q_{Pu}, T_{Vj}, T_{Lj}, H_{Ht_F}, H_{Pu}, H_{Pi}, y_R, y_{Sw}, y_{Lw}, y_{Ht_S}, y_{Vj}, y_{Lj}, y_{Ht_F}),$$

$$\dot{z} := \left(\frac{d(c(u_C) A_C^T e)}{dt}, \frac{dj_L}{dt}, \frac{dT_{Sw}}{dt}, \frac{dq_{Pi}}{dt}, \frac{dT_{Vj}}{dt} \right),$$

such that

$$F(\dot{z}, z, t) = 0. \tag{5}$$

DAEs resulting from automated modeling software typically obtain a structure with (differential) index greater 1, cf. [4–6] and hence are not suitable for a direct simulation with standard solvers. In the setup of multiple physical networks it is not sufficient, that the full DAE (5) can be reduced to a d-index (differential index) 1. Additionally, each subsystem, to which a solver is applied, has to fulfill d-index 1 conditions as well, cf. [1, 3]. In our applications an automatic index reduction is performed if the electric or the fluid system happens to be of d-index 2.

3 Multirate Integration for Coupled Network DAEs

In our multirate approach the full DAE (5) is partitioned due to the physical background to $n \in \mathbb{N}$ subsystems (typically $n \gg 2$). Each subsystem is index reduced according to the available literature, cf. [4–6]. Since in the global network the individual subsystems are interacting with each other, i.e. inputs and outputs are connected according the connectivity equation (4), it is necessary to put it into an input-output form. For this purpose, each subsystem $i = 1, \ldots, n$ classifies its inputs u_i, state variables x_i, algebraic variables a_i and outputs y_i. To conclude, this approach yields a coupled system of n semi-explicit DAEs in input-output form of (differential) index 1. For inputs u_i given by Eq. (4), find x_i, $\dot{x}_i$, a_i and y_i, such that

$$\begin{aligned} \dot{x}_i &= f_i(x_i, a_i, u_i, t) \\ 0 &= r_i(x_i, a_i, u_i, t) \\ y_i &= g_i(x_i, a_i, u_i, t) \end{aligned} \tag{6}$$

for $i = 1, \ldots, n$. A careful choice of the connectivity matrix given in (4) guarantees that the coupled system obtains (differential) index 1 as well, cf. [3, 9]. E.g. one possible choice is the usage of differential states, which are not involved in any index reduction, as coupling variables.

For each subsystem (6) an arbitrary Runge-Kutta method with micro-step sizes h_i is used, cf. Figure 1. The choice of the actual integration technique depends on the properties of the underlying system and can be explicit, implicit, fixed or adaptive. The whole system (5) is integrated via a non-iterative co-simulation technique with macro-step size $H = \max(h_i)$. All systems are updated at the end of each macro-step. This principle relates to synchronous communication and we refer to these points in time as synchronization times, cf. Fig. 1. The evaluation of each macro-step of the subsystems is done in a sequential Gauss-Seidel-approach. The values u_i are handled with appropriate interpolation or extrapolation techniques, depending on the slow or active characteristic of the interacting subsystems. Due to $n \gg 2$ *slow-first* or *fast-first* strategies (cf. [2]) have been extended to strategies, that can be used for an arbitrary number n of components.

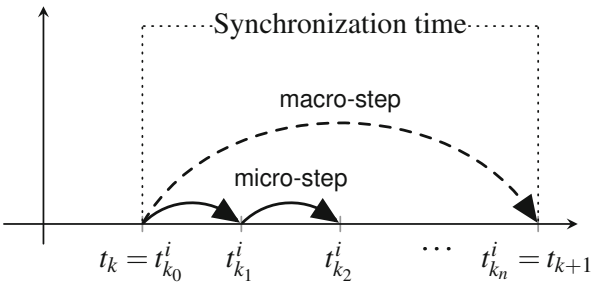


Fig. 1 Macro-step of the i-th system from synchronization time t_k to t_{k+1}.

4 Simulation of a BEV with Cooling System

We consider a BEV that demonstrates the modeling of an electrical system coupled to the required cooling system, cf. Fig. 2. The model consists of an electrical propulsion and two cooling circuits. An oil circuit is used for cooling of the electric machine and a coolant circuit is used for cooling of the battery pack, inverter and low voltage DC-DC converter. The involved subsystem of the coupled electro-thermal model can be reduced to DAEs of (differential) index 1.

The multirate approach presented in Sect. 3 is put into context with the reference solution of a single solver approach (both sequential/single CPU). In this example eight thermal circuits, three mechanic circuits, an electric circuit, 14 gas circuits and two fluid circuits are present which represent in total 461 equations. The solvers for both, the single solver approach and the multirate approach are all

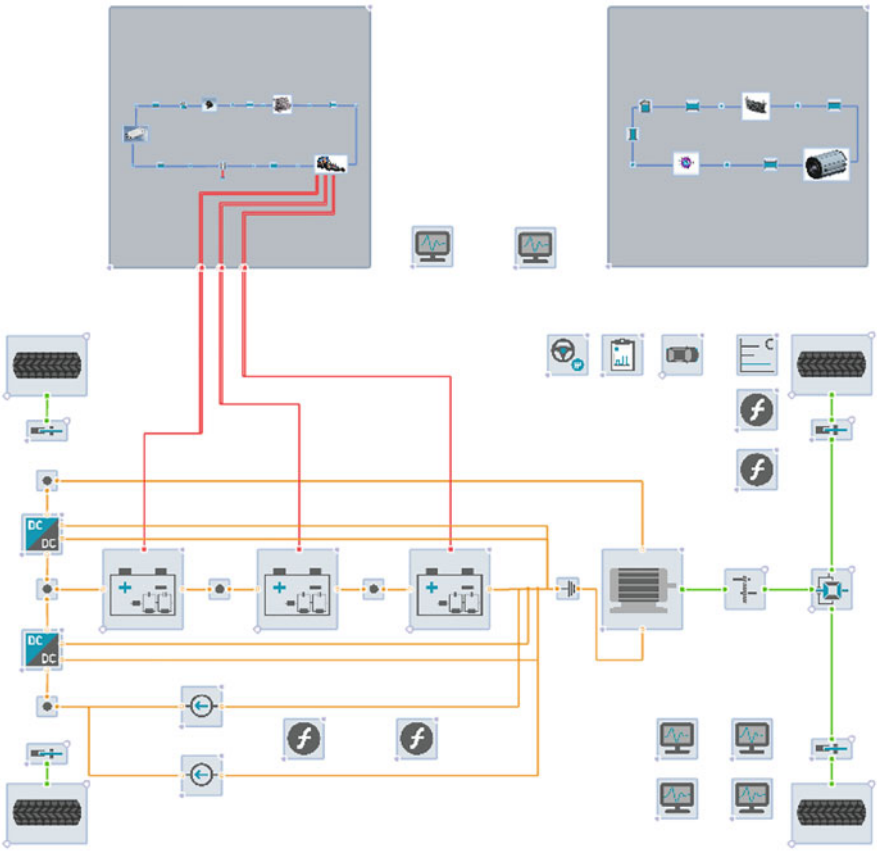


Fig. 2 Schematic representation of a BEV with cooling system in AVL CRUISE™M. The corresponding results are displayed in Fig. 3 and Table 1

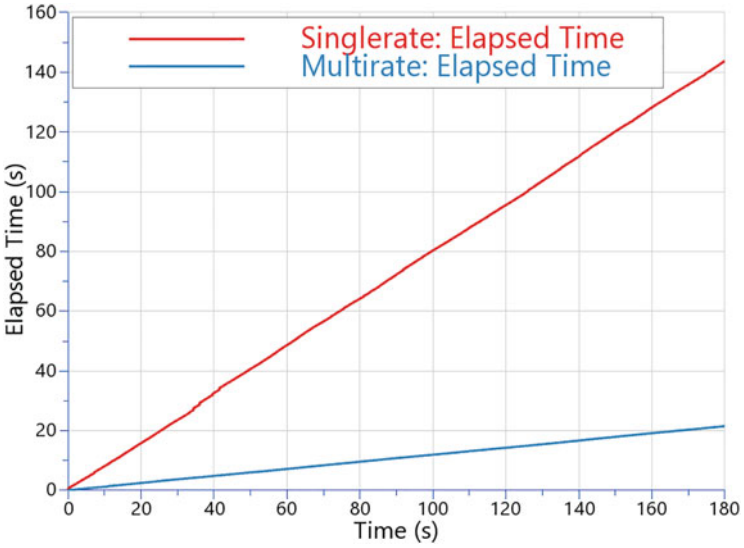


Fig. 3 Comparison of elapsed time of a multirate case against a single solver case

Table 1 Comparison of singlerate and multirate approach corresponding CPU-time and average real time factor (RTF)

Case	CPU-time	Avg RTF
Singlerate	144.98	0.805447
Multirate	21.03	0.116853

adaptive explicit solvers [7]. Hence the step size of the single solver is limited to the minimum step size of all subdomains, while the multirate approach is limited to the synchronization time or to the characteristic of its own domain. Here the synchronization times are after each macro-step of 20 ms.

The simulation time of a singlerate case (in red) is compared with those of a multirate case (in blue) using AVL CRUISE™ M, cf. Fig. 3. A significant speed up in the calculation time can be achieved, while the accuracy of the solution is still sufficiently high due to the adaptivity of the individual solvers (Table 1).

5 Simulation of a Three Phase Inverter with Cooling System

We consider a detailed physical model of an inverter with switches/transistors, an RC (resistor-capacitor) filter as well as a 3 phase ohmic load. The inverter is used to convert a DC (direct current) voltage through timed switching of the six transistors into a PWM (pulse width modulation) signal. The RC filter then averages the PWM and thus creates a 3 phase AC (alternating current) voltage.

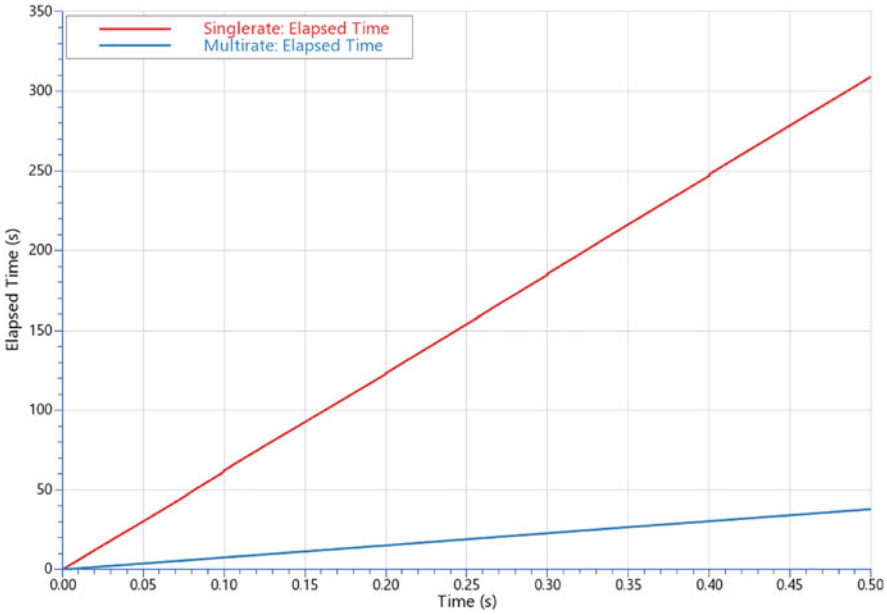


Fig. 4 Comparison of elapsed time of a multirate case against a single solver case

Table 2 Comparison of singlerate and multirate approach corresponding CPU-time and average real time factor

Case	CPU-time	Avg RTF
Singlerate	303.38	606.76734
Multirate	37.37	74.73045

In total this example consists of 178 equations which are spread over 20 solvers. A fluid circuit, seven gas circuits and eleven thermal circuits are responsible for modeling the cooling. In the multirate scheme each circuit is solved individually with one scheme. For all of them an explicit fixed step method with a step size of $1ms$ is used. On the other hand the electric network is solved by its own scheme as well. Again an explicit fixed step method is used, whereby the chosen step size is now $1\mu s$. The information exchange takes place after each macro-step of $1ms$. This model is of special interest, since the electric network and the fluid network run on completely different time scales (of order $\mathcal{O}(1000)$). Again significant speed up in the calculation time can be achieved (Fig. 4 and Table 2).

6 Conclusion

As shown, the multirate approach offers a possibility to reduce computation time considerably. In order to ensure a stable simulation, automatic index reduction of the physical networks, appropriate solver settings for each subsystem and an adequate coupling procedure, play a decisive role. For the correct choice a significant speed up can be achieved, while conserving the accuracy criteria.

Acknowledgments Part of this work has been supported by the government of Upper Austria within the programme Innovatives Oberösterreich.

References

1. A. Bartel, M. Günther, PDAEs in refined electrical network modeling. SIAM Rev. **60**, 56–91 (2018)
2. A. Bartel, M. Günther, Multirate Schemes - An Answer of Numerical Analysis to a Demand from Applications. IMACM Preprint, No. 2019–12, University of Wuppertal (2019)
3. A. Bartel, M. Günther, Inter/extrapolation-based multirate schemes – a dynamic-iteration perspective (2020). Available at https://arxiv.org/abs/2001.02310
4. A.-K. Baum, M. Kolmbauer, G. Offner, Topological solvability and DAE-index conditions for mass flow controlled pumps in liquid flow networks. Electr. Trans. Num. Anal. **46**, 395–423 (2017)
5. D. Estévez Schwarz, C. Tischendorf, Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl. **28**, 131–162 (2000)
6. S. Grundel, L. Jansen, N. Hornung, T. Clees, C. Tischendorf, P. Benner, Model order reduction of differential algebraic equations arising from the simulation of gas transport networks, in *Progress in Differential-Algebraic Equations* (Springer, Berlin, 2014), pp. 183–205
7. A. Hindmarsh, P. Brown, K. Grant, S. Lee, R. Serban, D. Shumaker, C. Woodward, SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. ACM Trans. Math. Softw. **31**, 363–396 (2005)
8. L. Jansen, C. Tischendorf, A unified (P)DAE modeling approach for flow networks, in *Progress in Differential-Algebraic Equations* (Springer, New York, 2014), pp. 127–151
9. M. Kolmbauer, G. Offner, B. Pöchtrager, Topological index analysis and its application to multi-physical systems in system simulation software (2020). Available at https://www.ricam.oeaw.ac.at/files/reports/20/rep20-22.pdf

A Hysteresis Loss Model for Tellinen's Scalar Hysteresis Model

Jan Kühn, Andreas Bartel, and Piotr Putek

Abstract A particularly well-suited hysteresis loss model for Tellinen's scalar hysteresis model is defined. Important basic properties are discussed. The model is based on the enclosed area of simple hysteresis loops as a measure of the energy loss. It can be applied for any simple excitation in a nearly steady state condition. The losses can be computed on-the-fly, during the field computations.

1 Introduction

Hysteresis is an important phenomenon for the simulation of magnetic fields in ferromagnetic materials. There are several hysteresis models already available in the literature, see e.g. [4]. One example is Tellinen's scalar hysteresis model, which was introduced in [8]. It is physically motivated but quite simple to evaluate. A recent investigation by Steentjes et al. [6] showed that it is still quite competitive to other, even more complex hysteresis models.

Moreover, there exists a thermal extension of Tellinen's model, which models the temperature depends of hysteresis in a Tellinen-like fashion [5]. Now, the work at hand aims at constructing a dedicated loss model, which inherits the computational benefits of Tellinen's model. For this new development, we currently assume that the magnetic fields are nearly in steady state.

Tellinen [8] provides a model to describe bh-curves. The thermal extension [5] deals with the influence of temperature. This paper presents a loss model. Then, in fact, the application of this model is based on a coupled problem, which consists of the curl-curl equation (magnetic field) and a heat conduction (temperature). This coupled system can be solved using distributed simulation techniques.

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

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_24

The outline of this work reads: First, we introduce Tellinen's model [8]. Then, we define the loss model for a steady state and apply it as an approximation for any nearly steady state. Based on properties of the model, we justify the use of steady state approximation also with numerical results. At the end, conclusions and an outlook are given.

2 Tellinen's Scalar Hysteresis Model

Original Model First, we report the isothermal hysteresis model introduced by Tellinen [8]. To this end, we assume that the temperature has a fixed value and the fields are described by the scalar magnetic field strength h and the scalar magnetic flux density b. Then, any ferromagnetic material has a specific limiting saturation curve. That is, for a transition from very small values of $h \ll 0$ (i.e., full saturation) to very large values $h \gg 0$, we have that the relation $b = b(h)$ is given by the specific function $B^+_{\mathrm{sat}} = B^+_{\mathrm{sat}}(h) \in C^1$. The reverse direction can be defined by $B^-_{\mathrm{sat}}(h) := -B^+_{\mathrm{sat}}(-h)$, see Fig. 1. To have a physically compliant model, $B^\pm_{\mathrm{sat}}$ is requested to form a loop:

$$B^+_{\mathrm{sat}}(h) < B^-_{\mathrm{sat}}(h)\,, \qquad \lim_{|h|\to\infty} \left(B^-_{\mathrm{sat}}(h) - B^+_{\mathrm{sat}}(h)\right) = 0 \tag{1}$$

and the derivative is bounded by the vacuum permeability μ_0 from below:

$$\frac{d}{dh} B^+_{\mathrm{sat}}(h) \geq \mu_0 > 0\,, \qquad \lim_{|h|\to\infty} \tfrac{d}{dh} B^+_{\mathrm{sat}}(h) = \mu_0\,. \tag{2}$$

Any current state of the material (h_0, b_0) has to belong to the loop region I

$$I = \left\{ (h, b) \in \mathbb{R}^2 \;\middle|\; B^+_{\mathrm{sat}}(h) \leq b \leq B^-_{\mathrm{sat}}(h) \right\}. \tag{3}$$

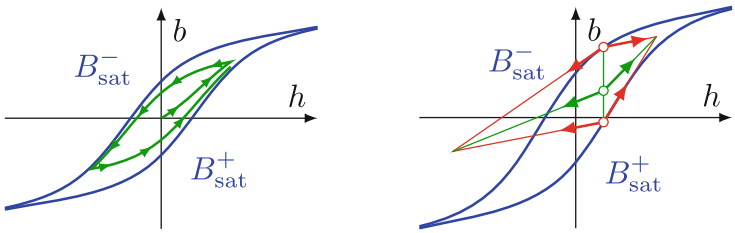


Fig. 1 Left: Example of $B^\pm_{\mathrm{sat}}$ and path starting from demagnetized state. Right: Schematic of Tellinen's model. Defining the values on the boundary and interpolate in between

Now, Tellinen's model defines the differential reluctivity $\mu_{\text{diff}} = \frac{\mathrm{d}b}{\mathrm{d}h}$ at (h_0, b_0) via vertical interpolation of the slopes on the saturation curves (see Fig. 1b) (notice that on B^-_{sat} the slope in the forward direction (i.e., increasing h) is given by μ_0) by

$$\mu_{\text{diff}} = \begin{cases} \mu^+_{\text{diff}} = \lambda \frac{\mathrm{d}B^+_{\text{sat}}(h)}{\mathrm{d}h} + (1-\lambda)\mu_0 & \text{if } h \text{ is increasing,} \\ \mu^-_{\text{diff}} = \lambda\mu_0 + (1-\lambda)\frac{\mathrm{d}B^-_{\text{sat}}(h)}{\mathrm{d}h} & \text{if } h \text{ is decreasing,} \end{cases} \tag{4}$$

where the interpolation parameter λ is given by

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

We note the following: (i) If h is regarded as the independent variable, the corresponding b can be determined as the solution of the ordinary differential equation (ODE) $\frac{\mathrm{d}b}{\mathrm{d}h} = \mu_{\text{diff}}$ with μ_{diff} given by (4). (ii) An analytical solution (h, b) of this ODE with initial value $(h_0, b_0) \in I$ stays inside the loop area I. (iii) The intrinsic induction is defined by $b_{\text{i}}(h) = b(h) - h\mu_0$. The definition can be transformed to the usage of b_{i}.

Partition of States in Tellinen's Model We introduce a partition of possible states I. To this end, we observe that for each h there exists a unique value b^{eq} with $B^+_{\text{sat}}(h) < b^{\text{eq}} < B^-_{\text{sat}}(h)$, such that holds (for λ, which correspond to that specific b^{eq}):

$$\mu^+_{\text{diff}} = \lambda \frac{dB^+_{\text{sat}}(h)}{dh} + (1-\lambda)\mu_0 = \lambda\mu_0 + (1-\lambda)\frac{dB^-_{\text{sat}}(h)}{dh} = \mu^-_{\text{diff}}. \tag{6}$$

Solving this equation for λ, we obtain

$$\lambda^{\text{eq}}(h) = \lambda = \frac{\frac{dB^-_{\text{sat}}(h)}{dh} - \mu_0}{\frac{dB^+_{\text{sat}}(h)}{dh} + \frac{dB^-_{\text{sat}}(h)}{dh} - 2\mu_0} = \frac{\frac{dB^-_{\text{i,sat}}(h)}{dh}}{\frac{dB^+_{\text{i,sat}}(h)}{dh} + \frac{dB^-_{\text{i,sat}}(h)}{dh}}, \tag{7}$$

which gives in turn the value of b^{eq} as

$$b^{\text{eq}}(h) = B^-_{\text{sat}}(h) - \frac{\frac{d}{dh}B^-_{\text{i,sat}}(h)}{\frac{d}{dh}B^-_{\text{i,sat}}(h) + \frac{d}{dh}B^+_{\text{i,sat}}(h)} \left(B^-_{\text{sat}}(h) - B^+_{\text{sat}}(h)\right). \tag{8}$$

Now, the set of all possible states I can be split into disjoint sets $I^=$, $I^<$ and $I^>$ with

$$\begin{aligned} I^\square &= \{(h,b) \in I \;\big|\; \mu^+_{\text{diff}}(h,b) \,\square\, \mu^-_{\text{diff}}(h,b)\} = \{(h,b) \in I \;\big|\; \lambda \,\square\, \lambda^{\text{eq}}\} \\ &= \{(h,b) \in I \;\big|\; b^{\text{eq}}(h) \,\square\, b\} \qquad \text{for all } \square \in \{=, <, >\}. \end{aligned} \tag{9}$$

See Fig. 2 for an example.

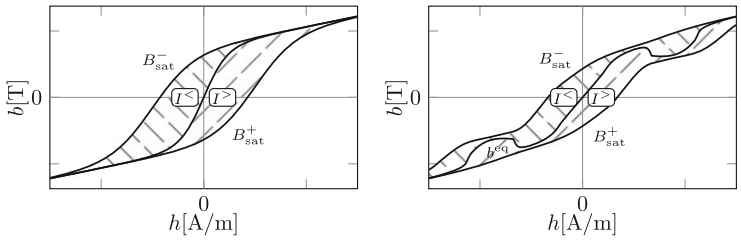


Fig. 2 Two examples with different saturation curves $B^{\pm}_{\mathrm{sat}}$ and the resulting equilibrium curve b^{eq} as well as the sets $I^{\lessgtr}$. We remark that the right example is a very academic version

3 Adapted Hysteresis Loss Model

There exists several hysteresis loss models for different kinds of hysteresis descriptions [1–3, 7]. A prediction of the losses is already presented in the original Tellinen model [8]. But the used method for the hysteresis losses is based upon a posteriori evaluation of the simulated fields b, h. Our model differs in this respect and provides a method for calculating losses at runtime. An overview and classification of different hysteresis loss models is e.g. given in [4]. In principle, the loss model presented below will work for other hysteresis models, too. However, it is particularly well suited for Tellinen's model [8] with the respective thermal extension [5] due to its structure and properties. First, we define the loss model for the steady state and then we extend it to almost steady state situations.

Idea We follow the approach of distributed simulation (Co-Simulation). A simulation of the heat equation describes the behavior of the temperature. The presented loss model provides corresponding source terms. Often in applications, e.g. electric machines, the rate of changes in the magnetic fields are several orders of magnitude faster than changes in temperature. In this setting, the assumption of a constant-temperature while handling magnetic fields is often exploited in distributed simulation techniques.

For a simple, closed loop in the bh-plane, the enclosed area represents an energy density ($\mathrm{J/m^3}$) and the material specific volumetric heat capacity c_{V} ($\mathrm{J/(m^3K)}$) provides the conversion into a temperature change ΔT (K). In a steady state, the material periodically passes through the same phases over and over again and, for this reason, runs on a closed bh curve. For memory reasons and the fact, that a priori the stable loop is unknown, we do not want to save the complete history of the curve, but calculate it from within the simulation on-the-fly, i.e., at runtime. To this end, we reverse the hysteresis model computation to predict the return path of the curve from a turning point at the same time as we compute the forward, see e.g. Fig. 4 right, where the curves p^+ (forward) and p^- (backward) will be computed at the same time.

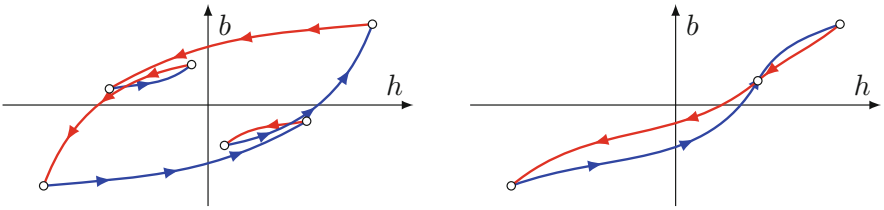


Fig. 3 Left: An example with minor hysteresis loops. This case is excluded by simple excitation. Right: An example with an intermediate intersection

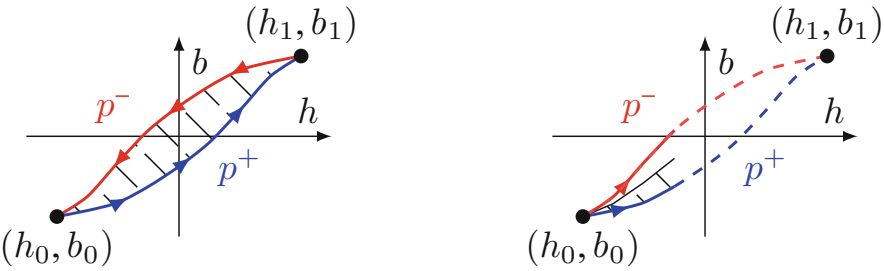


Fig. 4 Left: paths p^+ and p^- form a simple loop without intersection. Right: an incomplete cycle and its incremental section of the area. Notice, p^- is traversed in reverse direction

Prerequisites First, we consider a simple excitation such that the magnetic field strength is monotonously increased from h_0 to h_1 and then monotonously decreased back to h_0. This ensures that there is no minor hysteresis loop (see left of Fig. 3). Still, this is not sufficient to ensure that the bh-loop has no (intermediate) intersections (see Fig. 3, right). Below, sufficient conditions are presented.

Now, let (h_0, b_0) and (h_1, b_1) denote the turning points of a simple loop (cf. Fig. 4). Then, the loop can be split into two paths $p^+,\ p^- : [h_0,\ h_1] \to \mathbb{R}$ with

$$p^+(h) < p^-(h) \quad \text{for all } h \in (h_0,\ h_1)\,, \quad p^+(h_k) = p^-(h_k) = b_k \quad k \in \{0,\ 1\}. \tag{10}$$

On-the-Fly Algorithm To initiate the loss model, we assume that the current state (h_0, b_0) is a turning point of a simple hysteresis loop, where h is (wlog) increased. Thus, the simulation will follow the curve p^+ (using μ^+_{diff}), see Fig. 4. Now, a second computation is simultaneously performed based on μ^-_{diff} to follow the reverse direction, which results in a prediction of the return path p^-. Both simulations might use e.g. an ODE solver. For discrete steps, the resulting trapezoids (see Fig. 5) can be summed up to approximate the loop area. The incorporated halving of the area takes into account that there is a forward and backward phase. This continues until the second reversal point is reached. The procedure is then restarted from this point.

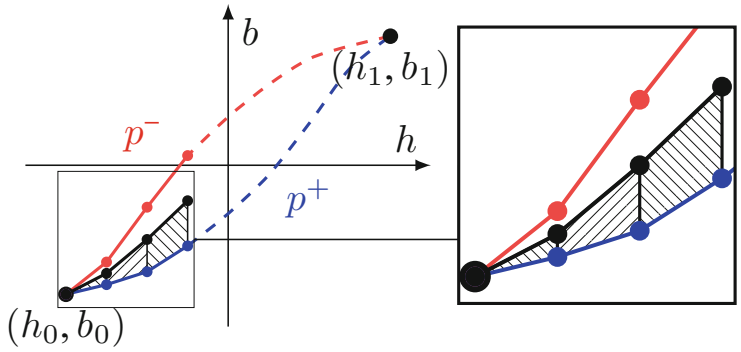


Fig. 5 For discrete points, the model results in trapezoids

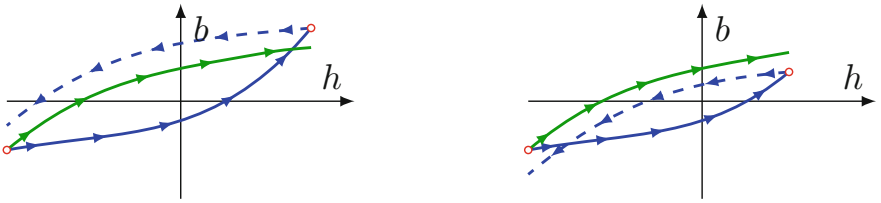


Fig. 6 Nearly steady state: actual turning point after (left) and before (right) the intersection of the curves p^+ and p^-

Non-steady State Loss Computation For simple closed bh-loops (steady state) the proposed method is accurate up to numerical precision of the employed solver. Now, if the turning point (h_1, b_1) cannot be determined accurately by the intersection of p^+ and the predicted curve p^- (as depicted in Fig. 6), we are not in steady state. We can prove (via some fixed-point argument) that this model exhibits convergent behavior for simple periodic inputs. Due to this, non-steady states are converging to the steady state. Numerical examples are presented in Sect. 4. If the difference between the actual and predicted reversal point is small enough (criteria set by user), we consider our model as a valid approximation and say it is nearly steady state.

Analytical Results Next, we develop criteria that guarantee the existence of at least one further intersection point (h_1, b_1) based on a turning point (h_0, b_0) and the corresponding paths p^+ and p^-. In a second step we then investigate when exactly only the intersection points (h_0, b_0) and (h_1, b_1) exist. It is then shown that both the Tellinen's model [8] and the thermal extension [5] converge towards the steady state.

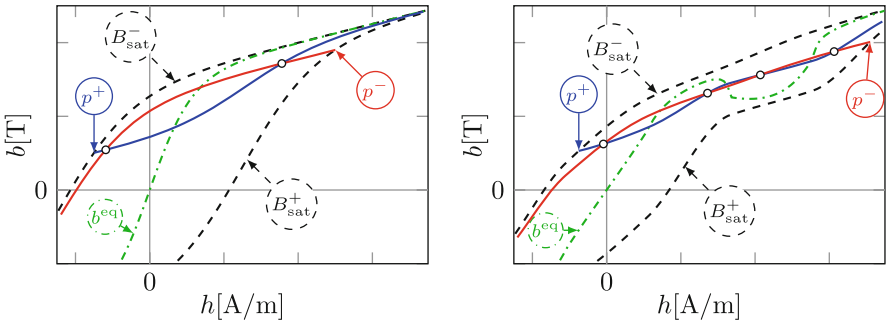


Fig. 7 Example with the same $B^{\pm}_{\text{sat}}$ as in Fig. 2, two paths p^+ and p^- and b^{eq}. Left: Only two intersections of $p^{\pm}$. Right: More than two intersections of $p^{\pm}$

As seen in Fig. 4 the intersection (h_0, b_0) fulfills $\mu^+_{\text{diff}} < \mu^-_{\text{diff}}$, while the intersection (h_1, b_1) holds $\mu^+_{\text{diff}} > \mu^-_{\text{diff}}$. In general, we have

Lemma 1 *Assuming that the paths p^+ and p^- have intersections and no contact points, i.e., the common points are not on b^{eq}. Any sequence (h_i, b_i) with $i = 0, \ldots, k$ and $h_0 < \ldots < h_k$ of all intersections of the paths p^+ and p^- must alternating fulfill $\mu^+_{diff} < \mu^-_{diff}$ and $\mu^+_{diff} > \mu^-_{diff}$, i.e., are alternating element of $I^<$ and $I^>$ (see (9)), starting with the former and ending with the latter.*

Lemma 2 (Existence of Intersection) *For any $(h_0, b_0) \in I^<$ with $B^-_{sat}(h_0) \neq b_0$, there exists at least one other point $(h_1, b_1) \in I^>$ with $h_0 < h_1$, such that the curves p^+ and p^- intersect at this point.*

The proof is mainly based on the consideration of the limit $h \to \infty$ for the paths $p^{\pm}(h)$ see Fig. 7 (left). As seen in Fig. 7 (right), there could be more than two intersections of $p^{\pm}$. These points are alternately above and below of b^{eq}.

Lemma 3 (Uniqueness of Intersection) *Given (h_0, b_0) as in Lemma 2 and*

$$\frac{db^{eq}(h)}{dh} > \mu^{eq}_{diff}(h) = \mu^{\pm}_{diff}(h, b^{eq}(h)) \quad \text{for all } h \tag{11}$$

we have exactly one further intersection, i.e., (h_1, b_1).

Proof We can restrict the paths $p^{\pm}$ to only go through b^{eq} from above to below. This effectively limits the number of possible intersections of $p^{\pm}$ to two and therefore makes (h_1, b_1) unique. □

Lemma 4 *If we assume $B^+_{sat} \in C^2$, then the uniqueness restriction (11) can be formulated as follows:*

$$\frac{\frac{dB^-_{i,sat}(h)}{dh}}{\frac{dB^+_{i,sat}(h)}{dh}} \left(\frac{dB^+_{sat}(h)}{dh} + \frac{dB^-_{sat}(h)}{dh} \right) + \left(B^-_{sat}(h) - B^+_{sat}(h) \right) \frac{d}{dh} \left(\frac{\frac{dB^+_{i,sat}(h)}{dh}}{\frac{dB^-_{i,sat}(h)}{dh}} \right) > 0 \,. \tag{12}$$

Lemma 5 *Let B^+_{sat} fulfil (12) and the temperature T be constant. We start from operation point $(h_0, b_0) \in I$ and h varies periodically between h_0 and h_1 with $h_0 < h_1$. The sequence of b-values B_k ($k \in \mathbb{N}$) at the turning point given by h_0 (computed by Tellinen's model) is convergent for $k \to \infty$. The resulting stable loop is unique and depends only on the choice of h_0 and h_1, but not on b_0.*

Proof *(sketch)* First, we define one iteration. To this end, let $b^+(h)$ be the solution of the ODE $\frac{\mathrm{d}b^+}{\mathrm{d}h} = \mu^+_{\mathrm{diff}}(h, b^+)$ with $b^+(h_0) = b_0$ and μ^+_{diff} as in (4). At $h = h_1$, we have

$$b_1 = b^+(h_1) = b_0 + \int_{h_0}^{h_1} \mu^+_{\mathrm{diff}}(h, b^+(h))\mathrm{d}h\,. \tag{13}$$

The reverse direction is the same. Let $b^-(h)$ be the solution of $\frac{\mathrm{d}b^-}{\mathrm{d}h} = \mu^-_{\mathrm{diff}}(h, b^-(h))$ with $b^-(h_1) = b_1$. Evaluated at h_0, this results in $b_2 := b^-(h_0)$ (analog to (13)).

Let $\varphi : [B^+_{\mathrm{sat}}(h_0), B^-_{\mathrm{sat}}(h_0)] \to [B^+_{\mathrm{sat}}(h_0), B^-_{\mathrm{sat}}(h_0)]$ denote the resulting b-value at h_0 after one iteration (starting from $\bar{b}$), i.e.,

$$\varphi(\bar{b}) = \bar{b} + \int_{h_0}^{h_1} \mu^+_{\mathrm{diff}}(h, b^+(h))\mathrm{d}h + \int_{h_1}^{h_0} \mu^-_{\mathrm{diff}}(h, b^-(h))\mathrm{d}h \tag{14}$$

with b^+ defined w.r.t. $(h_0, \bar{b})$ (13) and b^- to $(h_1, b^+(h_1))$. Now, we construct a sequence B_k via $B_{k+1} = \varphi(B_k)$ and $B_0 = b_0$. If $\varphi(\bar{b}) = \bar{b}$ holds, the resulting loop would be closed, i.e., it is stable.

$\mu^+_{\mathrm{diff}}(h, b)$ and $\mu^-_{\mathrm{diff}}(h, b)$ are strictly monotone in the second component. This causes two ODE solutions of $\frac{\mathrm{d}b}{\mathrm{d}b} = \mu^+_{\mathrm{diff}}$ with different initial values to become closer to each other. An analogous statement can be made for μ^-_{diff}.

Given $(h_0, b_0), (h_0, b_1) \in I$, we can prove, that $q \in [0, 1)$ exists, such that

$$|\varphi(b_1) - \varphi(b_0)| \leq q|b_1 - b_0| \tag{15}$$

holds. So φ is a contraction and Banach's fixed-point theorem applies. This ensure convergence and a unique fixed-point, i.e., a stable loop.

A numerical example is given in Fig. 8. We conclude that under these assumptions, the model tends towards the steady state. If we now assume that the temperature changes only slowly in comparison to the magnetic fields (thermal extension of Tellinen's model [5]), then the steady state is disturbed by this, but it is still approximately maintained. Similar convergence considerations for the thermal extension [5] also show the convergence towards a stable state. This is a kind of confirmation of our model's assumption and that after a certain initial phase we are permanently in an almost steady state and our loss model is therefore applicable.

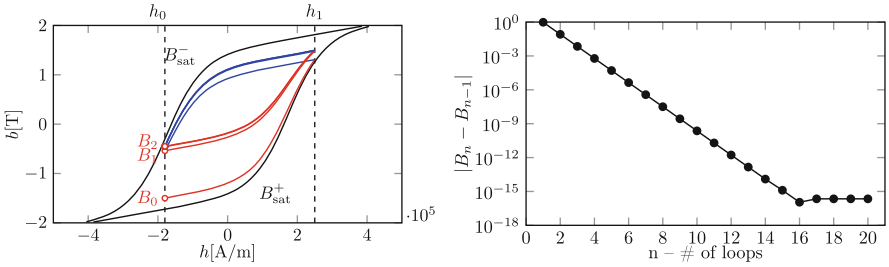


Fig. 8 Example of convergence. A random starting point $(h_0, b_0) \in I$ and $h_1 > h_0$ is chosen. Periodically and monotonously alternating between h_0 and h_1 converges to a stable loop

4 Numerical Results

As an academic example, a material is defined by $B^{\pm}_{\text{sat}}$ depicted Fig. 8, left. We choose $h_0 = -1.8\text{e}5\,\text{A/m}$, $h_1 = 2.5\text{e}5\,\text{A/m}$ and $b_0 = -1.5\,\text{T}$. Starting at $(h_0, b_0) \in I$, the ordinary differential equation $\frac{db}{dh} = \mu^+_{\text{diff}}(h, b)$ is solved numerically on the interval $[h_0, h_1]$ by a Runge-Kutta method. Then, $\frac{db}{dh} = \mu^-_{\text{diff}}(h, b)$ is solved backward from h_1 to h_0, where the initial value is the final value of the previous computation. This procedure is repeated n times. This results in the sequence of b-values at h_0: $B_0, \ldots, B_n$. As seen in Fig. 8, right, the absolute difference $|B_n - B_{n-1}|$ converges. Even a low number of loops n results in a nearly steady state and thus, would allow us to apply the loss model presented above.

5 Conclusion and Outlook

We have proposed a loss model with on-the-fly computation for Tellinen's hysteresis model. In steady state, it results in a precise computation of the model respective hysteresis loss. Thus it is a valid approximation for nearly steady state. Moreover, the convergence towards a stable bh-loop is proven. We note that our model is not suited for complex waveforms or rotating fields, cf. [4]. As a model feature, we stress that this hysteresis loss model employs only rough material data and needs small computational and memory cost. Moreover, this model can be combined with other than Tellinen's hysteresis model, if the value of b can be computed for changes in h. But the properties of the Tellinen's model make it an almost optimal candidate.

Our next step will be the integration of this loss model into a finite element magnetoquasistatic field simulation and its analysis. Special attention will be paid to whether this model can be applied per node.

References

1. A.P.S. Baghel, S. Kulkarni, Dynamic loss inclusion in the Jiles–Atherton (JA) hysteresis model using the original JA approach and the field separation approach. IEEE Trans. Magn. **50**, 369–372 (2014)
2. L.R. Dupre, R. Van Keer, J.A.A. Melkebeek, An iron loss model for electrical machines using the Preisach theory. IEEE Trans. Magn. **33**(5), 4158–4160 (1997)
3. G. Friedman, I.D. Mayergoyz, Hysteretic energy losses in media described by vector Preisach model. IEEE Trans. Magn. **34**(4), 1270–1272 (1998)
4. A. Krings, J. Soulard, Overview and comparison of iron loss models for electrical machines. J. Electr. Eng. **10**, 162–169 (2010)
5. J. Kühn, A. Bartel, P. Putek, A thermal extension of Tellinen's scalar hysteresis model, in *Proceedings of the SCEE 2018*, ed. by G. Nicosia, V. Romano. Springer (2020), pp. 55–63
6. S. Steentjes, K. Hameyer, D. Dolinar, M. Petrun, Iron-loss and magnetic hysteresis under arbitrary waveforms in no electrical steel: A comparative study of hysteresis models. IEEE Trans. Ind. Electron. **64**(3), 2511–2521 (2017)
7. C.P. Steinmetz, On the law of hysteresis. Proc. IEEE **72**(2), 197–221 (1984)
8. J. Tellinen, A simple scalar model for magnetic hysteresis. IEEE Trans. Magn. **34**(4), 2200–2206 (1998)

Hybrid Modeling: Towards the Next Level of Scientific Computing in Engineering

Stefan Kurz

Abstract The integration of machine learning (Keplerian paradigm) and more general artificial intelligence technologies with physical modeling based on first principles (Newtonian paradigm) will impact scientific computing in engineering in fundamental ways. Such hybrid models combine first principle-based models with data-based models into a joint architecture. This paper will give some background, explain trends and showcase recent achievements from an applied mathematics and industrial perspective. Examples include characterization of superconducting accelerator magnets by blending data with physics, data-driven magnetostatic field simulation without an explicit model of the constitutive law, and Bayesian free-shape optimization of a trace pair with bend on a printed circuit board.

1 Introduction: What Is Hybrid Modeling?

If we take a look at Gartner's 2018 Hype Cycle of Emerging Technologies [1], Deep Learning has been at the *top of inflated expectations* and should now be moving towards the *plateau of productivity*. Indeed, machine learning and more general artificial intelligence technologies recently have spurred a lot of interest in the applied mathematics and industrial communities, see for instance [2–4], and [5] for an introduction.

The idea of combining physics with data has a long history. Following [3, p. 57], we call modeling based on first principles the *Newtonian paradigm*. Newton's laws of motion provided (within their range of validity) "for the first time a unified quantitative explanation for a wide range of observations" [6]. Conversely, Johannes Kepler started from astronomer Tycho Brahe's and own measurement data, and worked towards a mathematical description to fit the measured data.

S. Kurz (✉)
Bosch Center for Artificial Intelligence, Renningen, Germany

Centre for Computational Engineering, Technical University of Darmstadt, Darmstadt, Germany
e-mail: stefan.kurz2@de.bosch.com

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_25

Following again [3], we call this approach the *Keplerian paradigm*. Both paradigms complement each other. For a simple enough model system, Kepler's laws can be derived from Newton's theory. Conversely, starting from a two-body model system, actual trajectories of celestial bodies can be modeled by Newton's laws plus data-driven terms that correct for perturbations due to effects that are not present in the model.

In modern terms, we call this complementary approach *hybrid modeling*.

Definition *Hybrid models* combine first principle-based models with data-based models into a joint architecture, supporting enhanced model qualities, such as robustness and explainability.

First principles express formalized domain knowledge. For the purpose of this paper the domain knowledge results from physics. But there are other possibilities, such as statistics (e.g., probabilistic graphical models [7, Ch. 8]) or discourse (e.g., ontologies [8]). Data may be obtained from any source, in particular from observation or simulation. We find also the somewhat narrower terms *scientific machine learning* [9], *physics-based machine learning* [10] and *predictive data science* [4], respectively.

Consider a high-dimensional manifold that contains some big data. It might be that a submanifold can be identified, which is dictated to us by the laws of physics, e.g. regarding admissible system dynamics. Learning algorithms can then be used to project the data into this submanifold. In other words, the structure of submanifold embeds physical constraints. A classical example is *Kálmán filtering* [11], and an example is presented in Sect. 2. Kálmán filtering was actually an enabling technology for the moon landing in 1969, where the goal was landing within ≈500 m after ≈400,000 km of travel. More preference is given to physics or data, depending on the level of uncertainty. *Ensemble Kálmán filtering* is used in weather forecasting centres worldwide [12]. They have to deal with about 10^6 incoming data points per hour, and mathematical models with about 10^9 states. Ensemble Kálmán filtering can be recognized as Gaussian hidden Markov model [13]. This use case is similar to digital twinning, since data from the field is acquired and used to update the models. Citing [4, p. 39]: "Learning from data through the lens of models is a way to exploit structure in an otherwise intractable problem."

Looking closer into engineering, we notice that a large class of physics models can be decomposed into conservation laws and constitutive laws [14, Ch. 1.3], [15]. The conservation laws are of topological nature and can therefore be discretized easily, leaving little room for data-driven techniques. The situation is different for the constitutive relations, which are of metric nature, and encode phenomenological material properties. Except for simple media (local, linear) there are many potential complications (non-local, hysteretic, non-linear, multi-scale, multi-physics, etc.). Here, data-driven models can be useful, provided that the models fulfil certain admissibility criteria, which can often be expressed in terms of invariance with respect to symmetry groups (orthogonal group, Lorentz group, etc.). This is showcased in Sect. 3.

To sum up, hybrid modeling has the potential to improve the Pareto tradeoff between simulation accuracy and simulation cost significantly, and therefore bring scientifc computing in engineering to the next level. In the remainder of the paper we will showcase this by some recent achievements.

2 Blending Data with Physics [16]

This section deals with the characterization of superconducting accelerator magnets of CERN's Large Hadron Collider. The magnets are characterized by measurements and the field in their aperture is modeled by the Boundary Element Method (BEM). Different methods for local magnetic field measurements are available. There are mapper systems, based on 3D magnetic Hall probes, and coil-based systems. The latter feature plane translating or saddle-shaped rotating coils. The coil-based systems benefit from straightforward calibration and linear transfer function, while the mapper systems offer a small active area ($\approx$0.01 mm^2), a high positioning resolution (μm scale), and admit measuring of all three field components simultaneously.

Figure 1 shows field quality maps of a dipole field in a rectangular magnet cross section. The colours ranging from blue to red indicate the deviation from an ideal dipole field, in logarithmic scale. The top image shows a field quality map that can be obtained by interpolation from measurement data, while the bottom image was obtained by reconstruction from a BEM model. In the sequel, the field is represented by a double layer potential. The dipole layer is located on the boundary of the rectangle. This reconstruction method enjoys a smoothing property. Moreover, the reconstructed field is locally an exact magnetostatic solution. The double layer

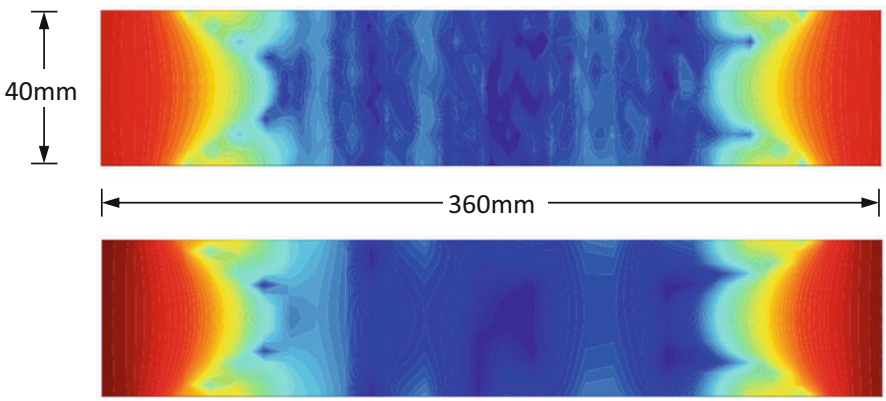


Fig. 1 Field quality maps of a dipole field in a rectangular magnet cross section. Top: interpolation of measurement data. Bottom: reconstruction from a BEM model. The colours ranging from blue to red indicate the deviation from an ideal dipole field, in logarithmic scale

Table 1 Bayesian update and Kálmán update. Quantities $\boldsymbol{v}$ and $\boldsymbol{d}$ are to be understood as random variables with probability distributions $p(\cdot)$. The number of degrees of freedom of the model is denoted by N, and the dimension of measurement data by M

$\boldsymbol{v}$	State vector of BEM model
$\boldsymbol{v} \sim \mathcal{N}(\overline{\boldsymbol{v}}, \boldsymbol{Q})$	$\overline{\boldsymbol{v}} \in \mathbb{R}^N$ mean values
	$\boldsymbol{Q} \in \mathbb{R}^{N\times N}$ covariance matrix, process noise
$\boldsymbol{d}$	measurement data vector
$\boldsymbol{d}\|\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{M}\boldsymbol{v}, \boldsymbol{R})$	$\boldsymbol{M} \in \mathbb{R}^{M\times N}$ discrete measurement operator
	$\boldsymbol{R} \in \mathbb{R}^{M\times M}$ covariance matrix, measurement noise

density is discretized by piecewise linear continuous boundary elements, giving rise to a state vector $\boldsymbol{v}$, with N degrees of freedom.

A hybrid model is established, by estimating the state vector from measurement data. This can be accomplished by *Kálmán filtering* [11]. The relevant quantities are defined in Table 1. We start from a prior of the state $p(\boldsymbol{v})$, and measurement data given the state, $p(\boldsymbol{d}|\boldsymbol{v})$. Then, by *Bayesian update*, we infer the posterior of the state given the measurement data,

$$p(\boldsymbol{v}|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{v})p(\boldsymbol{v})\,. \tag{1}$$

Under normal distribution assumption this can be computed easily explicitly. Then, the Bayesian update turns into a *Kálmán update*, which can be readily expressed in terms of linear algebra operations,

$$\overline{\boldsymbol{v}} \mapsto \overline{\boldsymbol{v}} + \boldsymbol{K}(\boldsymbol{d} - \boldsymbol{M}\overline{\boldsymbol{v}})\,, \tag{2a}$$

$$\boldsymbol{Q} \mapsto (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{M})\,\boldsymbol{Q}\,, \tag{2b}$$

where

$$\boldsymbol{K} := \boldsymbol{Q}\boldsymbol{M}^\top\big(\boldsymbol{M}\boldsymbol{Q}\boldsymbol{M}^\top + \boldsymbol{R}\big)^{-1} \in \mathbb{R}^{N\times N} \tag{3}$$

is the *Kálmán gain matrix*. Matrix $\boldsymbol{M}$ is the measurement matrix. It maps the degrees of freedom of the BEM model to the measured quantities. These are flux density vectors in case of Hall probe measurements, and magnetic fluxes in case of coil-based systems. Technically, this amounts to evaluating the integral operator of the double layer potential, in terms of the discrete model. In actual applications this approach is extended to a box-shaped domain in three dimensions, cf. Fig. 2. The Kálmán update results in a three-step procedure.

1. We select some prior from previous measurements or simulations. In the simplest case, we start from zero, with some estimate for the covariance matrix, i.e. $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$. This is a so-called *smoothing prior*. In fact, the reconstruction of the dipole layer from the measured field boils down to an inverse problem,

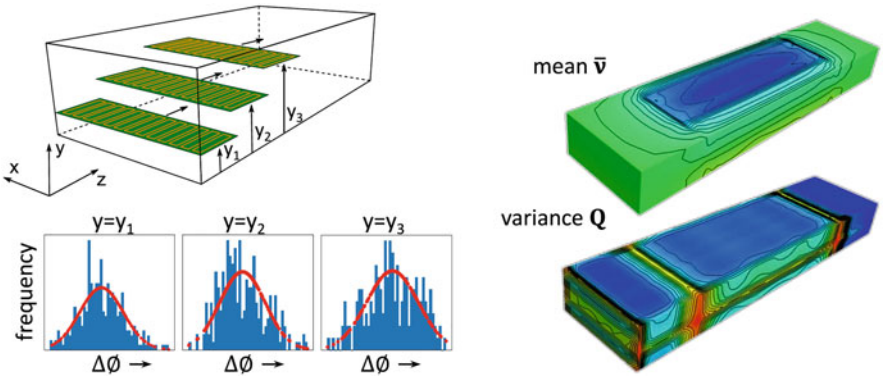


Fig. 2 Hybrid model combining measured data with the BEM. Top left: A translating induction coil consists of multiple single-wire loops in x direction. For fixed y, it measures magnetic flux increments $\Delta\Phi$ between successive trigger points along the z axis. Bottom left: The covariance matrix can be estimated from an ensemble of runs. The figure shows frequency distributions for three exemplary positions. Right: Contour plots of posterior mean and variance field magnitudes

and selecting a smoothing prior of the form $\boldsymbol{Q} = \sigma^2 \boldsymbol{I}$ is related to *Tikhonov regularization* [17, Sect. 2.1.2].

2. The measurement is carried out by a translating induction coil, which consists of multiple single-wire loops. The covariance matrix $\boldsymbol{R}$ is estimated from an ensemble of runs.
3. Finally, the Kálmán update (2) yields the new state vector and its covariance matrix. Since the state is modeled as random variable, we can immediately propagate uncertainties to the quantities of interest, such as magnetic flux density, magnetic vector potential or transfer maps. Moreover, data from different sensors can be combined through this procedure.

3 Data-Driven Field Simulation [18]

This section is about data-driven field simulation for magnetostatic problems. Here, data-driven simulation is meant in the context of simulations directly on the material data. In that manner, data-driven computing bypasses "the empirical material modeling step of conventional computing altogether" [19, p. 81]. Consider Maxwell's equations for magnetostatics,

$$\operatorname{curl} \boldsymbol{H} = \boldsymbol{j}\,, \quad \operatorname{div} \boldsymbol{B} = 0 \quad \text{in } \Omega\,, \tag{4}$$

where $\boldsymbol{H}$ denotes the magnetic field strength, $\boldsymbol{j}$ the imposed source current density, $\boldsymbol{B}$ the magnetic flux density, and Ω the considered domain. Moreover, we assume

suitable boundary conditions on $\Gamma = \partial\Omega$, for instance $\boldsymbol{n} \cdot \boldsymbol{B} = 0$, where $\boldsymbol{n}$ is the normal vector to the boundary.

The phase space of the system is denoted by $\mathscr{Z} := \{z(\boldsymbol{x}) := (\boldsymbol{B}(\boldsymbol{x}), \boldsymbol{H}(\boldsymbol{x}))\}$, $\boldsymbol{x} \in \Omega$. The set of all states that fulfill (4) and the boundary conditions is denoted by $\mathscr{M} \subset \mathscr{Z}$, the set of Maxwell-conforming fields[1]. To uniquely solve equations (4), a relation between $\boldsymbol{B}$ and $\boldsymbol{H}$ is necessary. In a conventional approach, $\boldsymbol{B}$ and $\boldsymbol{H}$ are connected through an empirically determined constitutive law. This law is mostly constructed as a fit with splines or regression techniques through raw measurement data. In contrast to the conventional approach, the data-driven field solver acts directly on the data. The measurement data is collected in a set $\mathscr{D}^* := \{z_i^* := (\boldsymbol{B}_i^*, \boldsymbol{H}_i^*), i = 1, \ldots, N\}$, where N is the number of measurement points. This gives rise to a set of discrete material states $\mathscr{D} := \{z \in \mathscr{Z} \,|\, z(\boldsymbol{x}) \in \mathscr{D}^* \, \forall \boldsymbol{x} \in \Omega\}$. "The material response is not known exactly and, instead, it is imperfectly characterized" [19, p. 95] by the set $\mathscr{D}$.

The solution is given by the states $\mathscr{M} \cap \mathscr{D}$ that fulfill Maxwell's equations, while being compatible with the material states. However, for a finite number of data points, this set is very likely empty, $\mathscr{M} \cap \mathscr{D} = \emptyset$. Therefore, we define the solution $\mathscr{S}$ by the relaxed condition

$$\mathscr{S} := \operatorname{argmin}\{d(z, \mathscr{D}), z \in \mathscr{M}\}, \tag{5}$$

where the distance function

$$d(z, z^*) := \frac{1}{2}\|\boldsymbol{B} - \boldsymbol{B}^*\|_{\tilde{\nu}}^2 + \frac{1}{2}\|\boldsymbol{H} - \boldsymbol{H}^*\|_{\tilde{\mu}}^2 \tag{6}$$

is defined in terms of auxiliary norms $\|\cdot\|_{\tilde{\nu},\tilde{\mu}}$. They do not represent material properties but are rather chosen to improve the convergence of the numerical scheme. The solution of (5) is organized as a fixed point iteration, cf. Fig. 3 right. For any given state $z \in \mathscr{Z}$ a modified FE solver is used to compute the state $z^{\odot} \in \mathscr{M}$ such that $d(z, z^{\odot}) = \min$. The solver is based on a variational principle discussed in [20]. The idea is to solve Ampère's and Faraday's laws exactly and shift the discretization error entirely into the constitutive relation.[2] Given $z^{\odot}$, a discrete optimization selects the closest measurement data points, cf. Fig. 3 left. These so-called active measurement data are associated with a state $z^{\times} \in \mathscr{D}$ such that $d(z^{\times}, z^{\odot}) = \min$. State $z = z^{\times}$ is the starting point for the next iteration. Under convexity assumptions this algorithm converges to the solution of (5). Furthermore, it has been shown in [22] for linear elasticity that an increasing set of measurement data recovers the conventional solution.

[1] We do not delve into regularity considerations or functional analytical frameworks here.

[2] This formulation was proposed on the Compumag Conference 1983 in Genoa. The related variational principle was called "Ligurian", in honor of the Genoa region, and in similarity to "Lagrangian". [21, p. 49].

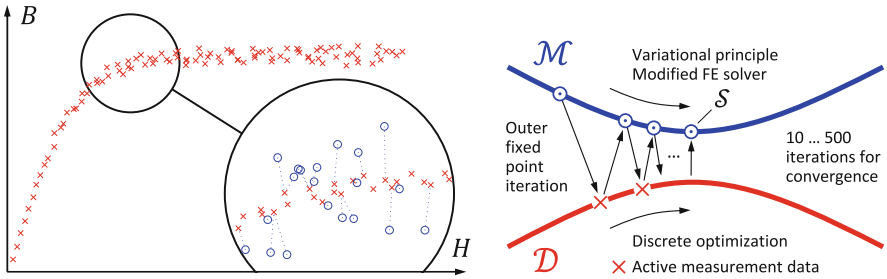


Fig. 3 Iterative data-driven solver. Left: Measured $B(H)$-characteristic. Active measurement data (red crosses) are closest to given field points (blue circles). Right: The outer fixed point iteration combines solutions of a variational principle by a modified FE solver (blue circles) with discrete optimizations that select states associated with active measurement data (red crosses)

In [18], a quadrupole magnet was analyzed in 2D by the proposed method. As a baseline, a standard magnetostatic finite element solution was considered, by discretizing 1/8 of the geometry with 6k piecewise linear elements. The non-linear system was solved by 20 Newton steps. For the novel method, data was created by an equidistant sampling of the given non-linear $B(H)$-characteristic, without adding noise. The relative error of $\boldsymbol{H}$ decreased almost inversely with N. However, depending on $N = 10^2 \ldots 10^4$ the proposed method required $10 \ldots 500$ outer iterations for convergence. As an advantage, the nonlinearity could be naturally included, and exactly known (air) and data-driven (iron) material information could be combined.

4 Bayesian Free-Shape Optimization [23]

Bayesian optimization (BO) is an optimization method to optimize a given function which is expensive to evaluate [24]. It is built upon a hybrid architecture that blends intricate physical models with a Bayesian machine learning technique, such as Gaussian Process (GP) regression. The resulting surrogate models are cheap to evaluate, including derivatives, and keep track of their interpolation uncertainty. The core idea of BO is to successively refine those surrogates in regions of design space that are close to optimal, which are however not known beforehand. Regions with high surrogate uncertainty *might* be optimal even though the mean interpolation says otherwise. Thus, surrogate refinement requires balancing exploration against exploitation during sampling, the so-called bandit problem[3]. There are different strategies for achieving a good balance. One strategy considers the best value

[3] In this model problem, an array of slot machines is considered. The gambler must balance the goal to find the slot machine with the highest gain (exploitation) with the goal to achieve good results on every play (exploration).

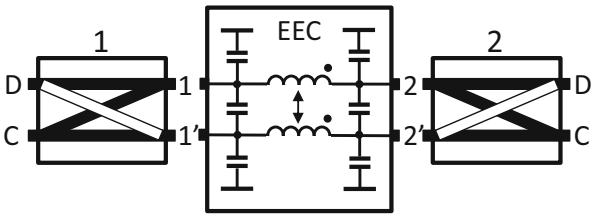


Fig. 4 Schematic for system simulation: extracted equivalent electrical circuit (EEC) of the trace pair, surrounded by mode converters

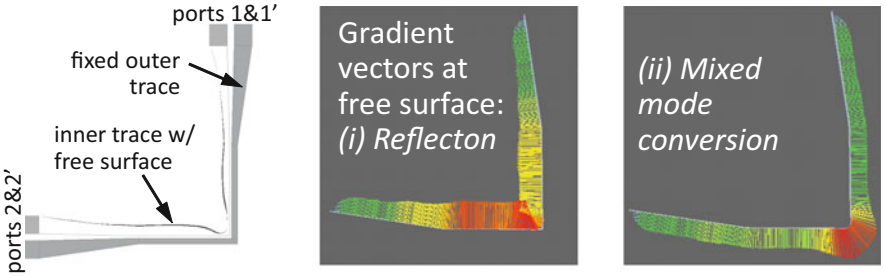


Fig. 5 Trace pair with bend. Left: The outer trace is fixed, the inner trace has a free interior surface. Right: Shape gradient vectors at the free surface for the two objectives (i) and (ii)

achieved so far and computes the *Expected Improvement* (EI). The next sample is taken at the point with the largest EI; this yields yet another optimization problem. The BO algorithm stops if the EI drops below some threshold. The BO approach can be generalized in various ways, such as BO with noise, BO in several dimensions, and BO for several objectives.

As an industrial example we consider BO of a differential trace pair on a printed circuit board. Differential signalling benefits from high immunity against electromagnetic interference and low crosstalk. However, bend discontinuities in transmission lines introduce (i) reflection and (ii) differential-to-common-mode conversion. An optimal design hence requires multi-objective optimization of the geometry. A parametric case was studied in [25], while we aim at *free-shape optimization*. Figure 4 shows a schematic for system simulation. The trace pair with ports 1,1' and 2,2', respectively, is described by an equivalent electrical circuit (EEC). Mode converters admit a separation of differential mode (D) and common mode (C) signal components. The optimization objectives can be stated in terms of S-parameters: [4]

$$\text{(i) reflection } |S_{\mathrm{DD11}}| \stackrel{!}{=} \min; \quad \text{(ii) mixed mode conversion } |S_{\mathrm{CD21}}| \stackrel{!}{=} \min.$$

The geometric setting is depicted in Fig. 5 left. The outer trace is fixed, while the inner trace has a free interior surface. The geometry is described by a finite element

[4] For simplicity evaluated at a fixed frequency of $f = 500$ MHz.

mesh, and the free surface can be re-shaped by mesh morphing. This corresponds to a high-dimensional design space with ≈200 dimensions. This should be put in contrast to the six-dimensional design space that was considered in [25]. The optimization problem is: *Find the Pareto front for the shape of the free surface that minimizes the objectives.*

The ingredients for solving the optimization problem are: finite element electromagnetic field solver, EEC extraction, and adjoint sensitivity analysis. Figure 5 right shows the shape gradient vectors at the free surface for the two objectives (i) and (ii). The two gradient vector fields point in opposite directions, so the objectives are conflicting. However, the gradient fields are not exactly negatives of each other, so there is still subtle room for improvement.

The BO is extended to the multi-objective case as follows. The Pareto front is approached via a sequence of auxiliary optimization problems, each with respect to a certain 2D affine subspace of the high-dimensional design space. This particular affine subspace is spanned by the adjoint-based gradients; it is the subspace of maximum objective variance. For each optimization problem of the sequence, BO learns and optimizes GP surrogate models for the objective functions, restricted to this subspace. Once the intermediate Pareto front is converged in this subspace, new subspaces may be chosen on the intermediate Pareto front. Figure 6 shows the result of this algorithm, after only ≈100 design evaluations. Note that even subtle improvement potentials will be exploited by the hybrid free-shape optimizer.

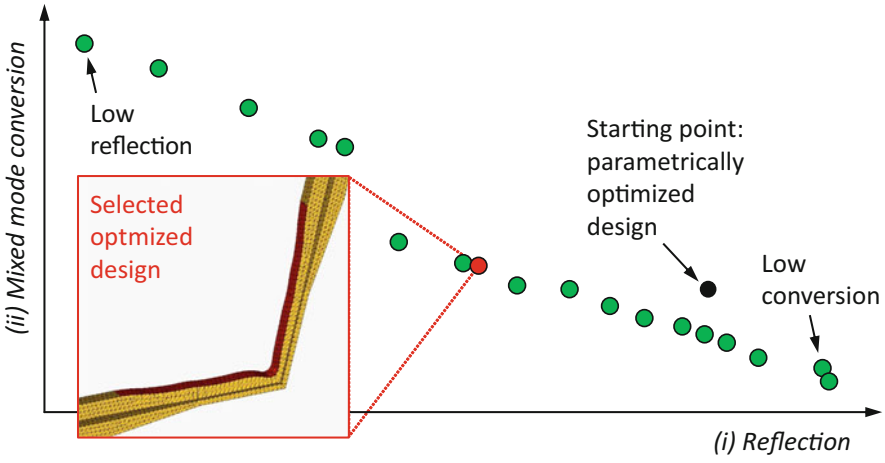


Fig. 6 Converged Pareto front for the trace pair with bend (green dots). The algorithm was started with a parametrically optimized design (black dot). An optimized design from the Pareto front was selected as an example (red dot)

5 Summary and Outlook

We highlighted three examples of hybrid modeling, cf. Table 2. As an outlook, we will discuss hybrid models from other applied mathematics and industrial domains. They will soon find their way into scientific computing in electrical engineering, too.

Physics-Informed Neural Networks [26] This method utilizes a fully-connected neural network (NN) to map a space-time domain to the unknown solution of an initial- and boundary-value problem. The NN is inserted into the governing partial differential equation (PDE) or variational principle and symbolically differentiated. This yields another NN, with modified activation functions but identical parameters, a so-called *physics-informed NN*. No labelled data is required for training. Rather, a combined loss function is minimized. One component is associated with the initial and boundary conditions, the other either with the residual norm or the variational functional of the PDE. The latter component enforces the structure of the physics equation. The solution is mesh-free and analytical. More general network architectures, such as convolutional encoder-decoder NN's are discussed in [27]. Physics-informed NN's benefit from their prior knowledge (also known as *inductive bias*) that helps them overcoming the challenges of generalization and data-efficiency. In fact, such hybrid models require only relatively small training data, typically a few hundred up to a few thousand points [26, p. 688].

Embedding Physics Simulation into Deep Learning [28] This work is motivated by control engineering, in particular by the development of intelligent reinforcement learning agents. "The end result is that we can embed an entire physical simulation environment as a layer in a deep network, enabling agents to both learn the parameters of the environments to match observed behavior and improve control performance via traditional gradient based learning." [28, p. 2]. The main ingredient is an adjoint-based solver, which allows efficient backpropagation of gradients and avoids their tedious computation by finite differences. Then, deep convolutional neural networks can be integrated seamlessly with physics-based models in machine learning platforms such as PyTorch and TensorFlow.

Table 2 Three examples of hybrid modeling

Section	Physics	Data	Approach
2	2D magnetostatics: BEM discretization	Fields measured by Hall probe or moving coil	Synthesis of physics and data by Kálmán update
3	2D magnetostatics: Ampère's & Gauss's law	Measured material data points $\left(\boldsymbol{B}_i^*, \boldsymbol{H}_i^*\right)$	Projection of data into admissible physics manifold
4	Electromagnetic Darwin model	Data sampling from physics model	Bayesian Optimization: GP machine learning

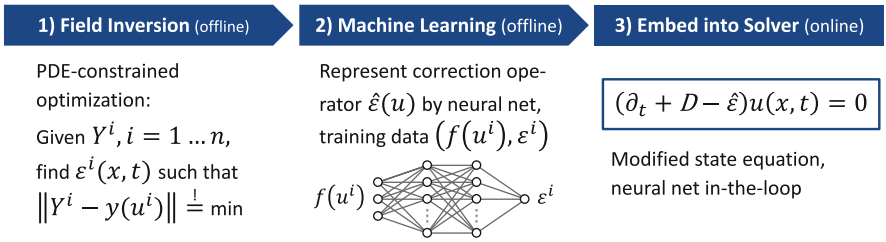


Fig. 7 Field inversion and machine learning. Step (1): Identify the model errors $\varepsilon^i(x,t)$ as defined in (7) for given time series Y^i, by solving PDE-constrained optimization problems. Step (2): Train a NN that describes a mapping from state variables (more precisely: features thereof) to model errors, by using the results from the previous step. Step (3): Include the correction operator $\hat{\varepsilon}$ in the state equation

Field Inversion and Machine Learning (FIML) [29] This method stems from computational fluid dynamics (CFD). For turbulent flows one may either solve Navier-Stokes equations by direct numerical simulation (DNS) or large eddy simulation (LES). This approach is accurate but numerically expensive, since it involves a range of space and time scales. On the other hand, one may use the Reynolds-averaged Navier-Stokes (RANS) method, where turbulence effects are accounted for by phenomenological models rather than first principles. This method is much more efficient but less accurate. With the help of FIML, both approaches can be combined.

Going beyond CFD, on an abstract level, let some system dynamics be governed by a low-fidelity state equation of the form

$$\big(\partial_t + \mathrm{D}\big)u(x,t) = \varepsilon(x,t)\,, \tag{7}$$

where ∂_t is the time derivative, D is the differential operator in space, $u(x,t)$ is the state variable, and $\varepsilon(x,t)$ is the (unknown) model error. We consider a discretized setting. Assume that an observable $y(u)$ is defined by some functional of the state variable, and several observed time series Y^i, $i = 1, \dots, N$ are available, either measured or from high-fidelity simulation. The idea of FIML is to learn a correction operator $\hat{\varepsilon}$ to account for the model error, cf. Fig. 7. Note that the NN does not directly operate on the state variable, but rather on some low-dimensional feature set $f(u)$. Some achievements, limitations and further developments of this method applied to airfoil modeling can be found in [30].

Epilog We have discussed hybrid modeling mainly from a physics-based perspective, where significant advantages could be achieved by joining with data-driven models. Conversely, hybrid modeling is also beneficial from the standpoint of industrial AI. In contrast to consumer AI, industrial AI focuses on smart products and their creation. Such AI should be robust, that is sufficiently tolerant against perturbations, and explainable, that is, the AI function can be made comprehensible

to humans. Industrial AI hence calls for inclusion of domain knowledge by hybrid modeling.

Acknowledgments Support in preparing the examples as well as inspiring discussions with the following colleagues are acknowledged: Armin Galetzka (TU Darmstadt); Andreas Klaedtke, Xiaobai Li, Manuel Schmidt (Bosch Corporate Research); Melih Kandemir, Zico Kolter (Bosch Center for Artificial Intelligence); Melvin Liebsch (CERN).

References

1. M. Walker, Hype cycle for emerging technologies, 2018. Tech. Rep. G00340159, Gartner Research (2018). https://www.gartner.com/en/documents/3885468/hype-cycle-for-emerging-technologies-2018
2. P.V. Coveney, E.R. Dougherty, R.R. Highfield, Big data need big theory too. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **374**(2080), 20160,153 (2016). https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2016.0153
3. W. E, Machine learning: Mathematical theory and scientific applications, in *ICIAM – International Congress on Industrial and Applied Mathematics* (2019). https://web.math.princeton.edu/~weinan/ICIAM.pdf
4. K. Willcox, Predictive data science for physical systems – from model reduction to scientific machine learning (2019), in *ICIAM – International Congress on Industrial and Applied Mathematics* (2019). https://kiwi.oden.utexas.edu/papers/Willcox-Predictive-Data-Science-ICIAM-2019.pdf
5. C.F. Higham, D.J. Higham, Deep learning: an introduction for applied mathematicians. SIAM Rev. **61**(4), 860–891 (2019)
6. Wikipedia contributors: Newton's laws of motion — Wikipedia, the free encyclopedia (2019). https://en.wikipedia.org/w/index.php?title=Newton27s_laws_of_motion&oldid=915580692 [Online. Accessed 26 September 2019]
7. C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006)
8. N. Guarino, D. Oberle, S. Staab, What is an ontology?, in *Handbook on Ontologies* (Springer, New York, 2009), pp. 1–17
9. S. Lee, N. Baker, Basic research needs for scientific machine learning: core technologies for artificial intelligence. Tech. rep., USDOE Office of Science (SC)(United States) (2018). https://www.osti.gov/servlets/purl/1484362
10. R. Swischuk, L. Mainini, B. Peherstorfer, K. Willcox, Projection-based model reduction: Formulations for physics-based machine learning. Comput. Fluids **179**, 704–717 (2019)
11. R.E. Kálmán, A new approach to linear filtering and prediction problems. Trans. ASME–J. Bas. Eng. **82**, 35–45 (1960)
12. A. Stuart, The legacy of Rudolph Kálmán – blending data and mathematical models, in *Boeing Distinguished Colloquia*, Univ. Washington (2019). https://www.sfb1294.de/fileadmin/user_upload/Kalman_Lectures/1st_Kalman_Lecture_2018_Andrew_Stuart.pdf
13. W. Pieczynski, F. Desbouvries, Kálmán filtering using pairwise Gaussian models. in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, vol. 6, pp. VI–57–VI–60 (IEEE, New York, 2003)
14. E. Tonti, *The Mathematical Structure of Classical and Relativistic Physics* (Springer, New York, 2013)
15. E. Tonti, Discrete physics – algebraic formulation of physical fields (2014). http://www.discretephysics.org/en/. [Online; Accessed 26 September 2019]
16. M. Liebsch, S. Russenschuck, S. Kurz, Boundary-element methods for field reconstruction in accelerator magnets. IEEE Trans. Magn. **56**(3), 1–4 (2020)

17. J.M. Bardsley, *Computational Uncertainty Quantification for Inverse Problems*, vol. 19 (SIAM, New York, 2018)
18. H. De Gersem, A. Galetzka, I.G. Ion, D. Loukrezis, U. Römer, Magnetic field simulation with data-driven material modeling (2020). Preprint. arXiv: 2002.03715
19. T. Kirchdoerfer, M. Ortiz, Data-driven computational mechanics. Comput. Methods Appl. Mech. Eng. **304**, 81–101 (2016)
20. J. Rikabi, C. Bryant, E. Freeman, An error-based approach to complementary formulations of static field solutions. Int. J. Numer. Methods Eng. **26**(9), 1963–1987 (1988)
21. B. Trowbridge, Compumag conference – the first 25 years (2001). https://www.compumag.org/wp/wp-content/uploads/2018/07/TwentyFiveYearsOfCompumag.pdf. [Online; Accessed 16 April 2020]
22. S. Conti, S. Müller, M. Ortiz, : Data-driven problems in elasticity. Arch. Ration. Mech. Anal. **229**(1), 79–123 (2018)
23. S. Schuhmacher, A. Klaedtke, C. Keller, W. Ackermann, H. De Gersem, Adjoint technique for sensitivity analysis of coupling factors according to geometric variations. IEEE Trans. Magn. **54**(3), 1–4 (2018)
24. P.I. Frazier, Bayesian optimization, in *Recent Advances in Optimization and Modeling of Contemporary Problems* (INFORMS, Catonsville, 2018), pp. 255–278
25. C. Gazda, I. Couckuyt, H. Rogier, D.V. Ginste, T. Dhaene, Constrained multiobjective optimization of a common-mode suppression filter. IEEE Trans. Electromagn. Compatibil. **54**(3), 704–707 (2012)
26. M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. **378**, 686–707 (2019)
27. Y. Zhu, N. Zabaras, P.S. Koutsourelakis, P. Perdikaris, Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. J. Comput. Phys. **394**, 56–81 (2019)
28. F. de Avila Belbute-Peres, K. Smith, K. Allen, J. Tenenbaum, J.Z. Kolter, End-to-end differentiable physics for learning and control, in *Advances in Neural Information Processing Systems* (2018), pp. 7178–7189
29. E.J. Parish, K. Duraisamy, A paradigm for data-driven predictive modeling using field inversion and machine learning. J. Comput. Phys. **305**, 758–774 (2016)
30. J.R. Holland, J.D. Baeder, K. Duraisamy, Towards integrated field inversion and machine learning with embedded neural networks for RANS modeling, in *AIAA Scitech 2019 Forum* (2019), p. 1884

Machine Learning for Initial Value Problems of Parameter-Dependent Dynamical Systems

Roland Pulch and Maha Youssef

Abstract We consider initial value problems of nonlinear dynamical systems, which include physical parameters. A quantity of interest depending on the solution is observed. A discretisation yields the trajectories of the quantity of interest in many time points. We examine the mapping from the set of parameters to the discrete values of the trajectories. An evaluation of this mapping requires to solve an initial value problem. Alternatively, we determine an approximation, where the evaluation requires low computation work, using a concept of machine learning. We employ feedforward neural networks, which are fitted to data from samples of the trajectories. Results of numerical computations are presented for a test example modelling an electric circuit.

1 Introduction

We examine initial value problems of nonlinear dynamical systems consisting of ordinary differential equations (ODEs) or differential-algebraic equations (DAEs). The systems include physical parameters, which vary in a predetermined bounded domain. A quantity of interest (QoI) is defined depending on the solution of the dynamical system. Hence there is a mapping from the parameters onto the trajectories of the QoI in the time domain. Our aim consists in the determination of an approximation of this mapping, which can be evaluated with a low computational effort. These approximations can be applied as surrogate models in uncertainty quantification, see [7], for example.

Methods using polynomials and their orthogonal bases yield surrogate models of parametric problems, see [9, 10]. Alternatively, we employ an approach of machine learning using artificial neural networks (NNs), see [2, 3], to construct an approximation. In [11, 12], proper orthogonal decomposition (POD) is applied

R. Pulch (✉) · M. Youssef
Universität Greifswald, Institute of Mathematics and Computer Science, Greifswald, Germany
e-mail: roland.pulch@uni-greifswald.de; maha.youssef@uni-greifswald.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_26

to data of solutions of parametric partial differential equations to obtain a reduced basis. Consequently, a mapping between low-dimensional spaces is approximated by NNs. In contrast, we discretise in time and use the data from the trajectories of the QoI in many time points. A mapping from a low-dimensional parameter space to a high-dimensional space is approximated by an NN now.

We apply feedforward NNs for this approximation. The determination of an NN requires a training procedure using data. We obtain the data of the trajectories by solving initial value problems of the dynamical systems for samples of the parameters. The fitting of an NN represents a (nonlinear) optimisation problem. In an NN, the number of neurons in a hidden layer is typically larger than the number of neurons in the input layer or the output layer. Since the number of outputs is large in our case, we also have high numbers of neurons in the hidden layers.

Finally, we demonstrate numerical results for a test example, which is a DAE model of an electric circuit. The trajectories associated to some parameter samples are illustrated. We show statistics of the approximation errors.

2 Parameter-Dependent Dynamical Systems

Let parameters $\mathbf{p} \in \Pi \subset \mathbb{R}^q$ be given. We consider a nonlinear dynamical system of the form

$$\vec{M}(\mathbf{p})\dot{\vec{x}}(t, \mathbf{p}) = \vec{f}(t, \vec{x}(t, \mathbf{p}), \mathbf{p}). \tag{1}$$

The mass matrix $\vec{M} : \Pi \to \mathbb{R}^{n \times n}$ and the right-hand side $\vec{f} : [t_0, t_\mathrm{f}] \times \mathbb{R}^n \times \Pi \to \mathbb{R}^n$ include the parameters. Thus the solution $\vec{x} : [t_0, t_\mathrm{f}] \times \Pi \to \mathbb{R}^n$ depends both on time and the parameters. If the mass matrix is non-singular, then (1) represents a system of ODEs. If the mass matrix is singular, then a system of DAEs is given. We examine initial value problems (IVPs)

$$\vec{x}(t_0, \mathbf{p}) = \vec{x}_0(\mathbf{p}). \tag{2}$$

In the case of DAEs, the initial values have to be consistent, see [5]. Consistent initial values often depend on the parameters. We define a QoI $y : [t_0, t_\mathrm{f}] \times \Pi \to \mathbb{R}^n$ by a function $g : \mathbb{R}^n \to \mathbb{R}$ via

$$y(t, \mathbf{p}) = g(\vec{x}(t, \mathbf{p})). \tag{3}$$

Each selection of the parameters yields a trajectory of the QoI in the time domain. We obtain the mapping

$$\mathbf{p} \mapsto \{(t, y(t, \mathbf{p})) \ : \ t \in [t_0, t_\mathrm{f}]\} \tag{4}$$

for any $\mathbf{p} \in \Pi$. Our aim is to construct an approximation of the mapping (4), which can be evaluated cheap, in particular, without solving IVPs (1), (2) any more.

The following approach can also be used for boundary value problems (BVPs) of dynamical systems, because we only include the trajectories of the QoI in the method. The trajectories are computed from either IVPs or BVPs.

3 Time Discretisation

We discretise the trajectories of the QoI (3) in the time domain $[t_0, t_\mathrm{f}]$. Let

$$t_0 < t_1 < t_2 < \cdots < t_{m-1} < t_m \leq t_\mathrm{f}. \tag{5}$$

Equidistant time points can be used. We consider the mapping $\Theta : \Pi \to \mathbb{R}^m$

$$\Theta : \Pi \to \mathbb{R}^m, \quad \Theta(\mathbf{p}) = \begin{pmatrix} y(t_1, \mathbf{p}) \\ \vdots \\ y(t_m, \mathbf{p}) \end{pmatrix}. \tag{6}$$

Each evaluation of (6) requires to solve an IVP (1), (2) followed by the extraction of the QoI (3). The IVPs of the dynamical systems are solved by numerical methods, see [4, 5], like Runge-Kutta schemes and linear multistep methods. The methods yield approximations of the solution in discrete time points, which are typically determined by a local error control. Thus these time points are not identical to our choice (5). Nevertheless, we obtain the solution in the points (5) by an interpolation or a dense output in time.

Stiff systems of ODEs and all DAEs require implicit methods in the time integration. Therein, a nonlinear system of algebraic equations has to be solved in each time step. Thus the computational effort becomes large. Our goal is to determine an approximation of the mapping (6), whose evaluation is cheap.

4 Machine Learning

We arrange an artificial NN, see [3], to approximate the mapping (6). An NN consists of an input layer, an output layer and additional hidden layers. Figure 1 illustrates the schematic of an NN. Let N_j be the number of neurons in the jth layer for $j = 0, 1, \ldots, J$. Therein, $j = 0$ and $j = J$ represent the input layer and the output layer, respectively. Thus there are $J - 1$ hidden layers. It holds that $N_0 = q$ and $N_J = m$ and thus $N_J \gg N_0$ in our problem.

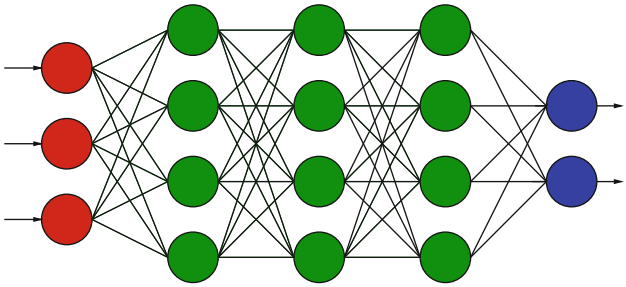


Fig. 1 Artificial NN with input layer (red), hidden layers (green), and output layer (blue)

The mathematical model $\Psi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_J}$ of an NN consists in a chain of operators

$$\Psi = \vec{T}_J \circ \rho \circ \vec{T}_{J-1} \circ \rho \circ \vec{T}_{J-2} \circ \cdots \circ \rho \circ \vec{T}_2 \circ \rho \circ \vec{T}_1. \tag{7}$$

Each operator $\vec{T}_j : \mathbb{R}^{N_{j-1}} \to \mathbb{R}^{N_j}$ is an affine-linear function

$$\vec{T}_j(\vec{z}) = \vec{A}_j \vec{z} + \vec{b}_j$$

including a matrix $\vec{A}_j \in \mathbb{R}^{N_j \times N_{j-1}}$, a vector $\vec{b}_j \in \mathbb{R}^{N_j}$, and the input $\vec{z} \in \mathbb{R}^{N_{j-1}}$. In the context of machine learning, the entries of $\vec{A}_j$ and $\vec{b}_j$ are denominated as weights and biases, respectively. The operator ρ is a nonlinear transfer function $\rho : \mathbb{R} \to \mathbb{R}$ (also called activation function). Typical choices are, for example, the hyperbolic tangent sigmoid function

$$\rho(x) = \tfrac{2}{1+\mathrm{e}^{-2x}} - 1 \tag{8}$$

and the hard-limit function

$$\rho(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } x \geq 0. \end{cases} \tag{9}$$

In (7), the function ρ is applied to vectors in each component separately.

In the fitting of an NN, the ideal is to minimise the distances $\Theta(\mathbf{p}) - \Psi(\mathbf{p})$ for any $\mathbf{p} \in \Pi$. The degrees of freedom are the weights and biases, i.e., $(\vec{A}_j, \vec{b}_j)_{j=1}^{J}$. Since a nonlinear optimisation problem appears, iterative methods are required to obtain numerical solutions.

In practise, the fitting involves three sample sets for training, validation, and test:

$$\begin{aligned} \mathcal{S}_{\text{train}} &= \{\mathbf{p}_1, \ldots, \mathbf{p}_k\} \subset \Pi \\ \mathcal{S}_{\text{valid}} &= \{\vec{q}_1, \ldots, \vec{q}_{k'}\} \subset \Pi \\ \mathcal{S}_{\text{test}} &= \{\vec{r}_1, \ldots, \vec{r}_{k''}\} \subset \Pi. \end{aligned} \tag{10}$$

For example, random samples can be chosen, where a uniform probability distribution is assumed in the parameter domain Π. The minimisation is based on the differences $\Theta(\mathbf{p}_i) - \Psi(\mathbf{p}_i)$ for parameter tuples $\mathbf{p}_i$ from the training set. The (vector-valued) differences are measured using the mean squared error or the mean absolute error. The error measure decreases monotone for the parameters in the training set due to the minimisation. The validation set is included to prevent an overfitting. If the error measure of the validation set increases, then the training is stopped and the best previous case is put out. The test set is not involved in the minimisation at all. Hence this set allows for an estimate of the quality of the trained NN.

5 Numerical Results for Test Example

All numerical computations were performed within the software MATLAB [8] using the Deep Learning Toolbox.

We investigate an electric circuit introduced in [1], which is illustrated by Fig. 2. This circuit performs a voltage doubling for specific choices of parameters and input voltage. A mathematical modelling yields a nonlinear system of DAEs (1) with $n = 3$ equations for the three unknown node voltages presented in [1]:

$$\begin{aligned} C_1\dot{x}_1 &= -\tfrac{x_1}{R_2} + F(-(x_1 + x_3)) \\ C_2\dot{x}_2 &= -\tfrac{1}{R_1}(x_2 + x_3 + u_{\text{in}}) \\ 0 &= -\tfrac{1}{R_1}(x_2 + x_3 + u_{\text{in}}) + F(-(x_1 + x_3)) - F(x_3). \end{aligned} \tag{11}$$

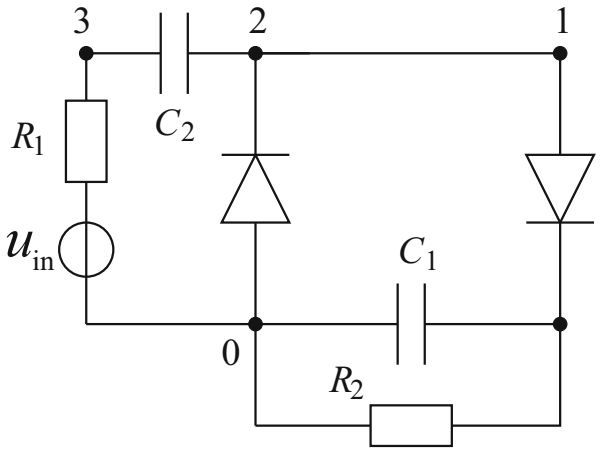


Fig. 2 Diagram of electric circuit

The current–voltage relation of the diodes reads as $F(u) = \gamma(\exp(\delta u) - 1)$. We use the constant parameters $\gamma = 4.067 \cdot 10^{-8}$ and $\delta = 5.634 \cdot 10^{-2}$ given in [6]. The differential index of the DAE system (11) is one. We choose the second node voltage as QoI (3).

We consider variations in four physical parameters: the two capacitances and the two resistances. The ranges $C_j \in [2 \cdot 10^{-9}, 3 \cdot 10^{-9}]$ for $j = 1, 2$, $R_1 \in [10^6, 2 \cdot 10^6]$, $R_2 \in [10^8, 2 \cdot 10^8]$ form the parameter domain $\Pi \subset \mathbb{R}^4$.

As input voltage, we supply the harmonic oscillation

$$u_{\text{in}}(t) = A \sin\left(\tfrac{2\pi}{T} t\right)$$

with amplitude $A = 500$ and period $T = 0.1$. The total time interval of our simulations is $[t_0, t_{\text{f}}] = [0, 0.5]$. The initial values (2) are set to zero, which represents a consistent case in this example. The backward differentiation formulas (BDF), see [4], yield the numerical solutions of the IVPs. High accuracy requests are imposed in the local error control with relative tolerance $\varepsilon_{\text{rel}} = 10^{-4}$ and absolute tolerance $\varepsilon_{\text{abs}} = 10^{-6}$. The error control generates approximations on a non-uniform grid in time. We extract the trajectories of the QoI in $m = 200$ equidistant time points $t_\ell = \ell \Delta t$ for $\ell = 1, \ldots, m$ with $\Delta t = \frac{t_{\text{f}} - t_0}{m}$ by interpolation. The order of accuracy coincides for both the uniform grid and non-uniform grid. The associated error of the time integration is negligible in comparison to the approximation error of the NNs below. Figure 3 gives an impression of the variability within the trajectories of the QoI for our parameter domain.

We select the number of samples as $k = k' = k'' = 500$ in the sets (10). Often the validation set and the test set are chosen smaller than the training set due to a restricted amount of data. In contrast, we are able to use larger sets, since a high number of trajectories can be produced by numerical simulations. In particular, a large test set provides reliable statistics in the error analysis. A pseudo random number generator yields the parameter samples in the multidimensional cuboid Π. Our NNs include two hidden layers with 400 neurons in each layer.

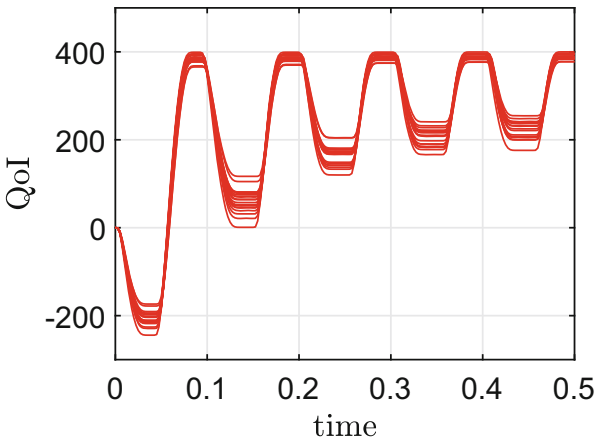


Fig. 3 Twenty trajectories of the QoI for different parameter samples

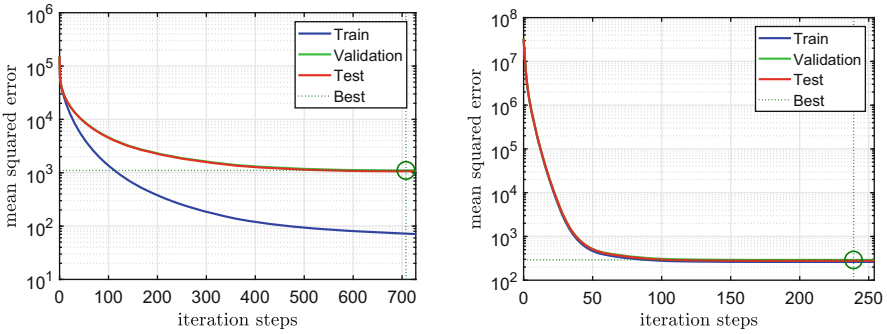


Fig. 4 Mean squared errors during the fitting of the two NNs in the iterative minimisation (The green line of the validation set is mostly located behind the red line of the test set.)

Table 1 Number of iteration steps and mean squared error (MSE) of test set for different training methods in NNs with hard-limit transfer function (i) and purely linear transfer function (ii)

	(i)		(ii)	
	Steps	MSE	Steps	MSE
Conjugate gradient method	728	1067.5	254	277.21
One-step secant method	1696	1062.2	328	277.24
Gradient descent method	10000	1323.1	1413	277.74

Using more hidden layers or more neurons did not improve the results significantly. We investigate two NNs, which differ only in the choice of the transfer function:

(i) hard-limit transfer function (9),
(ii) purely linear transfer function.

In the training, a conjugate gradient backpropagation method iteratively solves the minimisation problem. Figure 4 shows the performance of the training procedure. In the case of the hard-limit transfer function, the training is stopped at the 728th iteration step, because the error of the validation set increases slightly. In the case of the linear transfer function, the training is terminated at the 254th iteration step due to a too small step size. These two NNs are used in the following error analysis.

In addition, we tried two other backpropagation techniques in the training of the NNs: a one-step secant method (quasi Newton method) and a gradient descent method with momentum and adaptive learning rate. More information on all three methods can be found in [2], for example. Table 1 demonstrates the number of iteration steps (until a termination criterion applies) as well as the final mean squared error of the test set for the three techniques. We observe that the conjugate gradient method exhibits the best performance.

Figure 5 illustrates several trajectories of the test set. A comparison of the exact trajectories from the time integration and the approximations from the two NNs is shown. An interesting property is that NN (i) with the nonlinear transfer function

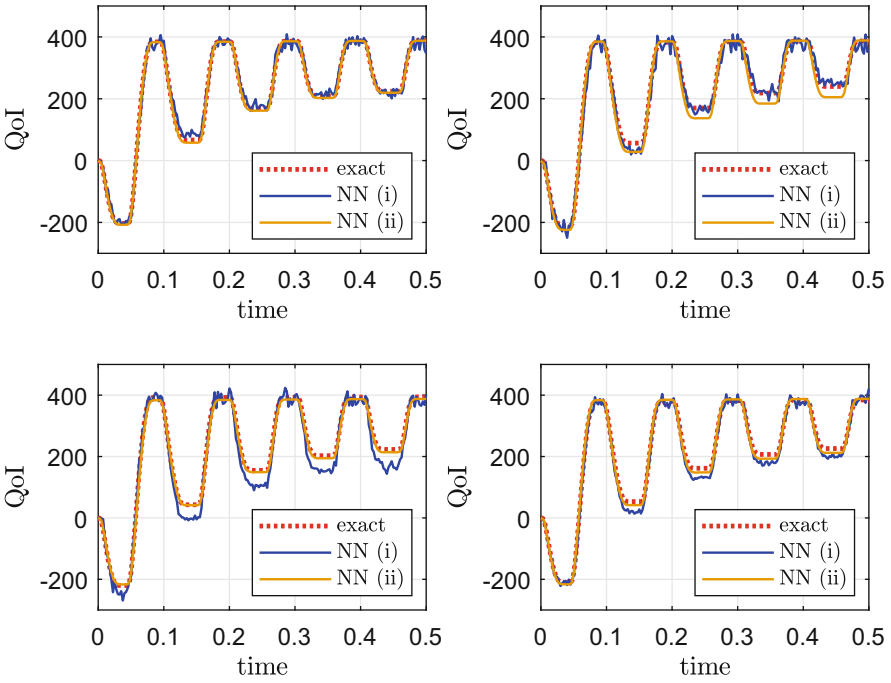


Fig. 5 Trajectories of the QoI for some samples from the test set

exhibits an oscillatory behaviour of the approximations in time, whereas NN (ii) with the linear transfer function generates more smooth approximations in time. The incorrect behaviour is not a transient effect, because all discrete time points (5) are treated equally in the NNs, i.e., without the consideration of their ordering in time. We also tried other nonlinear transfer functions (like (8)), which still caused oscillations in time.

Finally, we quantify the relative errors of the approximations for each sample trajectory using a discrete $\mathscr{L}^1$-norm in time. The error associated to the ith parameter sample reads as

$$E_i = \frac{t_\mathrm{f} - t_0}{m} \sum_{\ell=1}^{m} \frac{|\tilde{y}(t_\ell, \mathbf{p}_i) - y(t_\ell, \mathbf{p}_i)|}{|y(t_\ell, \mathbf{p}_i)|}, \tag{12}$$

where y is the original value from the time integration and $\tilde{y}$ denotes the approximation from an NN. The initial value is not included due to its value zero. The statistics of the errors are depicted in Table 2. We discuss the resulting mean values. In NN (i), a smaller mean error is achieved in the training set, whereas the other two sets show larger errors in comparison to NN (ii). Moreover, the mean errors are balanced for all three sets in NN (ii). This behaviour is in agreement to the performance of the training demonstrated by Fig. 4.

Table 2 Mean value and standard deviation of relative errors in discrete $\mathscr{L}^1$-norm, see (12), for the three parameter sets within the two trained NNs

	Mean	
	NN (i)	NN (ii)
Training set	0.044	0.086
Validation set	0.130	0.086
Test set	0.120	0.079

	St.dev.	
	NN (i)	NN (ii)
Training set	0.135	0.228
Validation set	0.210	0.158
Test set	0.155	0.146

6 Conclusions

We arranged a mapping from a set of parameters to discrete values of a QoI obtained from IVPs of dynamical systems. We approximated this mapping by artificial NNs. A test example was investigated, where two NNs were trained. In both NNs, the quality of the approximations is moderate with respect to the mean values of the errors. However, the NN including a linear transfer function yielded more smooth discretised trajectories in time, whereas the NNs with nonlinear transfer functions produced incorrect oscillations in time.

References

1. B. Barz, E. Suschke, *Numerische Behandlung eines Algebro-Differentialgleichungssystems. RZ-Mitteilungen*, vol. 7 (Humboldt-Universität, Berlin, 1994)
2. K.-L. Du, M.N.S. Swamy, *Neural Networks and Statistical Learning* (Springer, London, 2014)
3. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, London, 2017)
4. E. Hairer, S.P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, 2nd edn. (Springer, Berlin, 1993)
5. E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Equations*, 2nd edn. (Springer, Berlin, 1996)
6. W. Kampowsky, P. Rentrop, W. Schmidt, Classification and numerical solution of electric circuits. Surv. Math. Ind. **2**, 23–65 (1992)
7. O.P. Le Maître, O.M. Knio, *Spectral Methods for Uncertainty Quantification* (Springer, Dordrecht, 2010)
8. *MATLAB, version 9.7.0.1190202 (R2019b)* (The Mathworks Inc., Natick, MA, 2019)
9. R. Pulch, Polynomial chaos for the computation of failure probabilities in periodic problems, in *Scientific Computing in Electrical Engineering SCEE 2008* ed. by J. Roos, L. Costa. Mathematics in Industry, vol. 14 (Springer, Berlin, 2010), pp. 191–198
10. T.J. Sullivan, *Introduction to Uncertainty Quantification* (Springer, Cham, 2015)
11. Q. Wang, J.S. Hesthaven, D. Ray, Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem. J. Comput. Phys. **384**, 289–307 (2019)
12. J. Yu, J.S. Hesthaven, Flowfield reconstruction method using artificial neural network. AIAAJ **57**, 1–17 (2019)

Modeling of Thermoelectric Generator via Parametric Model Order Reduction Based on Modified Matrix Interpolation

Ananya Roy, Gunasheela Sadashivaiah, Chengdong Yuan, M. Nabi, and Tamara Bechtold

Abstract In this paper, finite element modeling of a simplified human tissue model consisting of muscle, fat and skin layers is carried out. A thermoelectric generator is placed in the fat layer and functioned as a power supply for electrically active implants. As the finite element method produces a large number of ordinary differential equations, model order reduction becomes the only way out of the computational complexity. In this work, the height of the thermocouples, which has a large influence on the generated power, is considered as a geometrical parameter. A modified matrix interpolation based parametric model order reduction method is used to construct the reduced order model valid for an arbitrary parameter value.

1 Introduction

During the last years with the development of technology, energy harvesting systems have become a popular area of research. They ensure longevity, eco-friendly operation, low maintenance and have a wide range of applications from aircraft, biosensors [1] to telemetry systems [2] etc. Here presented, thermoelectric generator (TEG) acts as an alternative source of energy to provide stable power to electrically active implants [3]. Choosing the proper geometry for the TEG is very important aspect [4]. To come up with an adequate design, the impact of geometrical param-

A. Roy (✉) · M. Nabi
Indian Institute of Technology, Delhi, New Delhi, India
e-mail: ananya.roy@ee.iitd.ac.in; mnabi@ee.iitd.ac.in

G. Sadashivaiah
University of Rostock, Rostock, Germany
e-mail: gunasheela.sadashivaiah@uni-rostock.de

C. Yuan · T. Bechtold
University of Rostock, Rostock, Germany

Jade University of Applied Sciences, Wilhelmshaven, Germany
e-mail: chengdong.yuan@jade-hs.de; tamara.bechtold@uni-rostock.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_27

eters needs to be analyzed. It has been seen that the thermoelectric performance of the TEG is dependant on the height of the thermocouples [5]. With change in height of the thermocouples, temperature difference between the top and the bottom level changes and so the generated power. We wish to investigate this influence via mathematical method of model order reduction (MOR).

The TEG is modelled as distributed parameter system via partial differential equations. Finite element analysis converts these partial differential equations into large-scale ordinary differential equation systems, which solution is computationally costly. MOR derives the low-dimensional approximation of the higher order original system [6, 7]. Furthermore, during design optimization, the system needs to be simulated repeatedly for different values of geometrical parameters. If these parameters can be preserved within the reduced models, then the full-scale system must not be repeatedly synthesized and reduced at each parameter value. This idea gives rise to parametric model order reduction (pMOR). In this paper, parametric modeling of a TEG is carried out with modified matrix interpolation based pMOR method. The parameter considered is the height of thermocouples.

2 Model Description

The human body is a thermal energy source. When the surrounding temperature varies, the body temperature varies between 23 °C (at the skin surface) and 37 °C (in the body core). Implantable TEG utilizes the temperature difference in the body and generates electrical power. It is made of an array of thermocouples, and each thermocouple consists of a p-type and n-type Bismuth Telluride. In this work, the TEG model contains 8×8 thermocouple legs with cross-sectional area 1 mm^2, which are electrically connected in series through copper interconnects. The thermocouples are thermally connected in parallel between two 17.5×17.5 mm^2 ceramic plates. A schematic of TEG is shown in Fig. 1 together with the Peltier height (the parameter of interest).

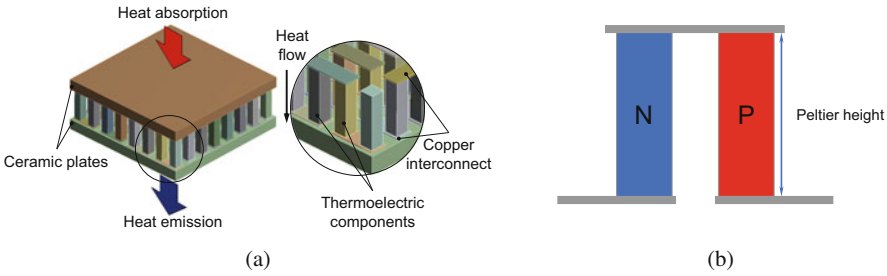


Fig. 1 (**a**) Schematic of a TEG; (**b**) schematic of a thermocouple

Based on Seebeck effect, electrons and holes in the thermocouples start moving when a temperature gradient occurs. As a result, thermal energy is being converted into electrical energy, which can be utilized to power electrical implants. The voltage (V) generated by the TEG is given by:

$$V = n \cdot \Delta T \cdot (\alpha_1 - \alpha_2) \tag{1}$$

where, ΔT is the temperature difference between the top level and bottom level of the TEG, n is the number of thermocouples, $\alpha_{1,2}$ are the Seebeck coefficients of the thermocouple legs.

Here a simplified human tissue model is considered to study the behavior of the TEG inside human tissue. The human tissue model consists of muscle, fat, and skin layers as shown in Fig. 2. The TEG is surrounded with a 40×40 mm^2 housing made of Teflon and placed within the fat layer, as maximum temperature difference occurs there [4].

The material properties of various parts in TEG and different human tissue are shown in Tables 1 and 2.

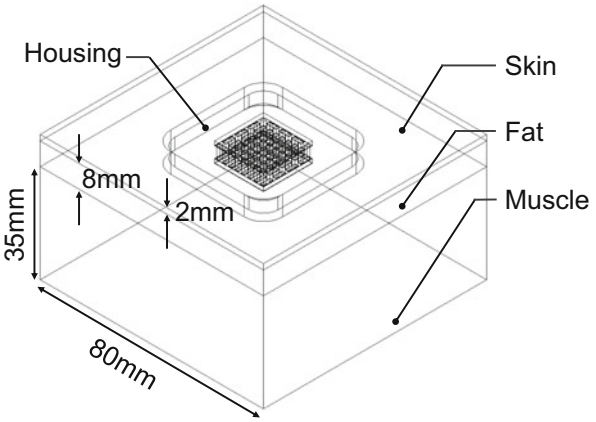


Fig. 2 Schematic of a TEG embedded in the fat layer of the tissue model

Table 1 Material properties of various parts of TEG

	Density (kg/m^3)	Specific heat (J/kg/K)	Thermal conductivity (W/m/K)	Seebeck coefficient (μV/K)
Housing	2250	1000	0.25	–
Ceramic plates	3720	880	25	–
p-type thermocouple leg	7700	90	1.6	200
n-type thermocouple leg	7700	90	1.6	−200

Table 2 Material properties of various tissue types

	Density (kg/m^3)	Specific heat (J/kg/K)	Thermal conductivity (W/m/K)	Perfusion rate (1/s)	Metabolic heat (W/m^3)
Muscle	1090.4	3421.2	0.4949	3.37e−4	498.52
Fat	911	2348.3	0.2115	3.01e−4	279.8
Skin	1109	3390.5	0.3722	9.05e−4	841.57

The heat conduction in the human tissue is described by the Pennes' bioheat model [8] expressed as:

$$\nabla(\kappa \nabla T) + \underbrace{\rho_b c_b \omega (T_a - T)}_{Q_b(T)} + Q_m = \rho c \frac{\partial T}{\partial t} \tag{2}$$

where T is the resulting temperature field and κ, ρ and c are the thermal conductivity, density and specific heat of the tissue, respectively. The heat generation rates provided by metabolism and perfusion are described by Q_m and $Q_b(T)$. The density and specific heat of blood are expressed as $\rho_b = 1049.75$ kg/m^3 and $c_b = 3617$ J/kg/K. ω is the blood perfusion rate in different tissue layers. Blood temperature $T_a = 37$ °C is set as temperature boundary condition at bottom surface of the tissue model. Note that the temperature dependent perfusion effect $Q_b(T)$ can be applied as the 'convection-type' effect as introduced in [9]. The value of the metabolic heat generation rates Q_m in different tissue layers are introduced in Table 2.

The heat is dissipated by convection at the skin surface as the external heat loss effect:

$$q_{conv} = h \cdot (T - T_{amb}) \tag{3}$$

where q_{conv} is the heat flux normal to the boudary skin surface. The heat transfer coefficient is expressed by h and the ambient temperature by T_{amb}. The steady state solution of the system (2), (3) is taken as the initial condition for the transient thermal simulation.

The finite element discretization of Eq. (2) with convection boundary condition (3) leads to a following large-scale system of ordinary differential equations:

$$\begin{aligned} \mathbf{E}\dot{T}(t) &= \mathbf{A}T(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}T(t) \end{aligned} \tag{4}$$

where, $E, A \in \mathbb{R}^{n \times n}$ are the heat capacity and heat conductivity matrices, respectively, $B \in \mathbb{R}^{n \times m}$ is the input distribution matrix and $C \in \mathbb{R}^{p \times n}$ is the user-defined output matrix. In this work, the order of the model, $n \approx 4 \times 10^4$ changes with the Peltier height, is very large, and $T(t) \in \mathbb{R}^n$ is the state vector of unknown nodal temperatures.

3 Parametric Model Order Reduction with Matrix Interpolation

The performance of the TEG is influenced by the height of the thermocouple. In Fig. 3, it can be seen that if the height changes, the temperature difference between the top level and the bottom level of the thermocouple changes, as well as the generated voltage. Hence the length of the thermocouple is considered as a geometric parameter here.

Ideally, the pMOR should be able to cope with an arbitrary number of parameters and allow for situations in which the matrix dependence on parameters can not be expressed analytically. One of such matrix-interpolation-based pMOR schemes has been introduced in [10] and later refined in [11]. This refined scheme is applied in the modeling of TEG.

The system is sampled at k different values of parameter of interest. For each value of parameter a single large-scale finite element model is generated as:

$$\begin{aligned} \mathbf{E}_i \dot{T}_i(t) &= \mathbf{A}_i T_i(t) + \mathbf{B}_i u(t) \\ y_i(t) &= \mathbf{C}_i T_i(t) \end{aligned} \tag{5}$$

k locally reduced order models (ROMs) are generated by projecting each large-scale system onto a lower order subspace. We have used one sided Arnoldi algorithm [12], which generates a transformation matrix $V_i \in \mathbb{R}^{n_i \times r}$, where n_i is the order of each large-scale system and $r \ll n_i$ is the order of the corresponding ROM.

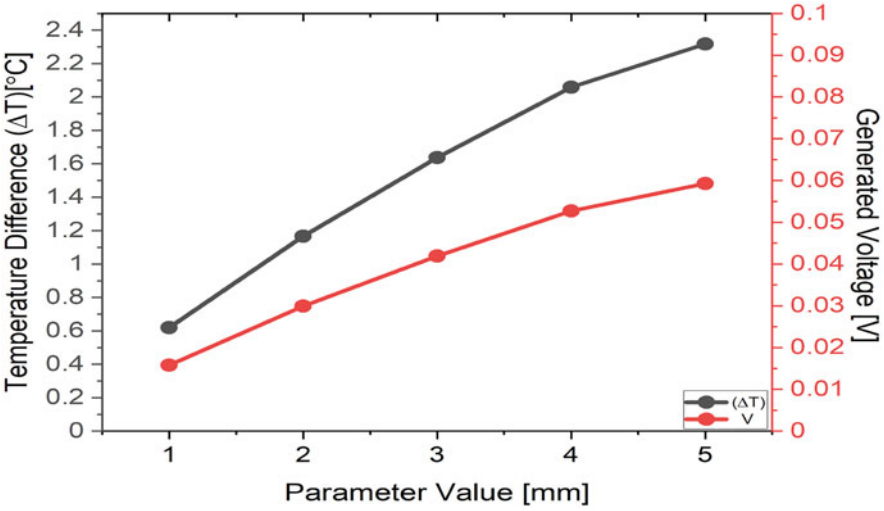


Fig. 3 Effect of peltier height of TEG

On the basis of the method introduced in [10] and [11], the globally reduced models after applying modified matrix interpolation method, can be written as:

$$\overbrace{\mathbf{M}_i \mathbf{E}_{r,i} \mathbf{T}_i^{-1}}^{\mathbf{E}^*_{r,i}} \dot{T}_r^*(t) = \overbrace{\mathbf{M}_i \mathbf{A}_{r,i} \mathbf{T}_i^{-1}}^{\mathbf{A}^*_{r,i}} T_r^*(t) + \overbrace{\mathbf{M}_i \mathbf{B}_{r,i}}^{\mathbf{B}^*_{r,i}} u(t) \tag{6}$$

$$y(t) = \underbrace{\mathbf{C}_{r,i} \mathbf{T}_i^{-1}}_{\mathbf{C}^*_{r,i}} T_r^*(t)$$

where $\mathbf{M}_i \in \mathbb{R}^{r \times r}$ and $\mathbf{T}_i \in \mathbb{R}^{r \times r}$ are transformation matrices which should be chosen appropriately.

Once the globally reduced models ($\mathbf{E}^*_{r,i}, \mathbf{A}^*_{r,i}, \mathbf{B}^*_{r,i}, \mathbf{C}^*_{r,i}$) at all the discrete points $i = 1, 2, \ldots, k$ are obtained, reduce model at any parameter value p can be obtained by using a weighted interpolation of the matrices of these local models as:

$$\mathbf{E}_r \, \dot{T}_r^*(t) = \mathbf{A}_r \, T_r^*(t) + \mathbf{B}_r \, u(t) \tag{7}$$

$$y(t) = \mathbf{C}_r \, T_r^*(t)$$

where, $\mathbf{E}_r = \sum_{i=1}^{k} w_i \mathbf{E}^*_{r,i}$, $\mathbf{A}_r = \sum_{i=1}^{k} w_i \mathbf{A}^*_{r,i}$, $\mathbf{B}_r = \sum_{i=1}^{k} w_i \mathbf{B}^*_{r,i}$, $\mathbf{C}_r = \sum_{i=1}^{k} w_i \mathbf{C}^*_{r,i}$ with $\sum_{i=1}^{k} w_i = 1$.

4 Simulation Results

In this work, a simplified human tissue model consisting of muscle, fat, and skin layers is considered. The TEG is placed within the fat layer. A geometrical parameter, the height of the thermocouple is varied from 3.65 mm to 3.95 mm and discretized at 3.65, 3.75, 3.85, 3.95 mm. Large-scale finite element models are generated at these discrete points by using ANSYS Mechanical [13]. Subsequently, the corresponding ROMs of order 31, are generated by using "Model Reduction inside ANSYS" [14]. Utilizing these ROMs through modified matrix interpolation based pMOR algorithm, a global reduced order model is generated.

To verify the proposed method, an intermediate point is chosen at $p = 3.8$ mm. A global reduced model is interpolated at this point and compared to the full-scale model of order $n = 44{,}942$, with 3.8 mm Peltier height. To study the influence of the convection boundary condition as mentioned in (3), an initial state of the TEG is obtained with heat transfer coefficient $h = 8.8$ W/m^2/K through steady state simulation. Afterwards, a transient simulation with heat transfer coefficient $h = 11.18$ W/m^2/K is carried out for 7000 s. The ambient temperature is set as constant $T_a = 25$ °C. The transient thermal response at the top and the bottom of the thermocouples, at intermediate point $p = 3.8$ mm, is shown in Fig. 4. For ease of measurement, we have calculated the average temperature of the top and bottom surfaces of the thermocouples. A comparative analysis of relative errors of average temperatures are shown in Fig. 5. It can be seen that the relative error at the

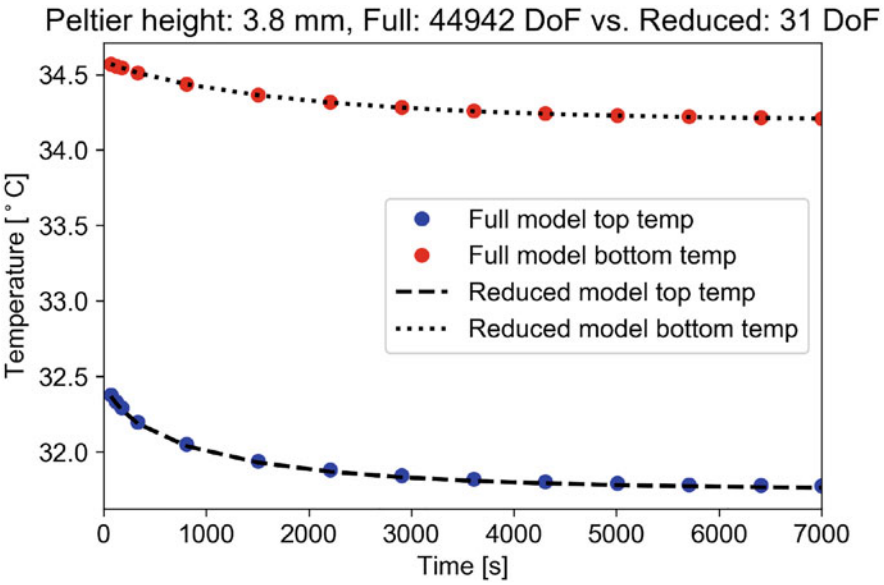


Fig. 4 Transient thermal response at selected top and bottom nodes of thermocouple at intermediate point

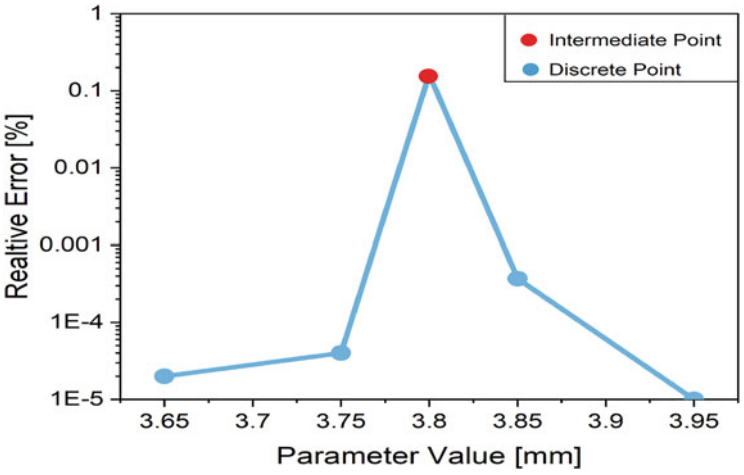


Fig. 5 Relative error at different parameter values

intermediate point is around 0.1624%, which is accurate enough for the problem at hand. Furthermore, the reduced models obtained at every discrete points, produce results with still higher accuracy. This is expected, because the reduced models are obtained at these very discrete points, while the model at the intermediate point is calculated through interpolation.

5 Conclusions and Outlook

In this paper, the potential design optimization strategy, based on modified matrix interpolation pMOR, for a human TEG has been investigated. As the thermoelectric performance of the device is highly affected by its Peltier height, we have chosen this height as a parameter of interest. Modified matrix interpolation based pMOR is applied to reduce computational complexity. A reduced model valid for an arbitrary Peltier height is generated through this method. Numerical simulations of the original large-scale model and its interpolated surrogate prove the efficacy of the proposed method.

In the future work, we will incorporate the cross-sectional area of thermocouple as another parameter and perform multiple-parameter model order reduction to get an optimal design of the TEG.

Acknowledgments Financial support of the CRC 1270 ELAINE (Electrically Active Implants) is acknowledged.

References

1. M. Koplow, A. Chen, D. Steingart, P.K. Wright, J.W. Evans, Thick film thermoelectric energy harvesting systems for biomedical applications, in *Proceedings of the 5th International Summer School Symposium on Medical Devices and Biosensors* (2008), pp. 322–325
2. S. Dalola, V. Ferrari, M. Guizzetti, D. Marioli, E. Sardini, M. Serpelloni, A. Taroni, Autonomous sensor system with power harvesting for telemetric temperature measurements of pipes. IEEE Trans. Instrum. Measure. **58**(5), 1471–1478 (2009)
3. Y.W. Chong, W. Ismail, K. Ko, C.Y. Lee, Energy harvesting for wearable devices: a review. IEEE Sensors J. **19**(20), 9047–9062 (2019)
4. O. Jadhav, C.D. Yuan, D. Hohlfeld, T. Bechtold, Design of a thermoelectric generator for electrical active implants, in *MikroSystemTechnik Congress* (2017), pp. 1–4
5. B. Jang, S. Han, J.Y. Kim, Optimal design for micro-thermoelectric generators using finite element analysis. Microelectron. Eng. **88**(5), 775–778 (2011)
6. W.H.A. Schilders, H.A. Van der Vorst, J. Rommes, *Model Order Reduction: Theory, Research Aspects and Applications*, vol. 13 (Springer, Berlin, 2008)
7. B. Lohmann, B. Salimbahrami, Introduction to Krylov subspace methods in model order reduction, in *Methods and Applications in Automation* (2000), pp. 1–13
8. C.K. Charny, Mathematical models of bioheat transfer, in *Advances in Heat Transfer*, vol. 22 (Elsevier, Amsterdam, 1992), pp. 19–155
9. C.D. Yuan, S. Kreß, G. Sadashivaiah, E.B. Rudnyi, D. Hohlfeld, T. Bechtold, Towards efficient design optimization of a miniaturized thermoelectric generator for electrically active implants via model order reduction and submodeling technique. Int. J. Numer. Methods Biomed. Eng. **36**(4) (2020). https://doi.org/10.1002/cnm.3311
10. H. Panzer, J. Mohring, R. Eid, B. Lohmann, Parametric Model Order Reduction by Matrix Interpolation. Automatisierung-stechnik Methoden und Anwendungen der Steuerungs-, Regelungs- und Informationstechnik **58**(8), 475–484. ISSN (Print) 0178–2312, August 2010. https://doi.org/10.1524/auto.2010.0863
11. A. Roy, M. Nabi, Efficient simulation of electro-thermal micro-gripper using PMOR, in *Indian Control Conference (ICC)*, Kanpur, 2018

12. R.W. Freund, Krylov-subspace methods for reduced-order modeling in circuit simulation. J. Comput. Appl. Math. **123**(1–2), 395–421 (2000)
13. Ansys®, Academic Research Mechanical, 2019 R3. ANSYS, Inc.
14. E.B. Rudnyi, Mor for ANSYS, in *System-Level Modelling of MEMS*, ed. by T. Bechtold, G. Schrag, L. Feng. Wiley-VCH Book Series on Advanced Micro and Nanosystems (Wiley-VCH, Weinheim, 2013), pp. 425–438

Nonlinear Model Order Reduction of a Thermal Human Torso Model

Gunasheela Sadashivaiah, Chengdong Yuan, and Tamara Bechtold

Abstract Considering the aging populations in Europe, electrically active implants play a major role in the modern medical field. Main drawback is their limited battery period. Here presented thermoelectric generator, which is based on Seebeck effect, utilizes temperature gradient inside the human body to generate electric voltage. It is aimed to prolong the life-time of implants. Its design optimization relies on an accurate thermal human torso model. In this work, we account for physiologically correct thermal transfer effects such as, internal heat transfer described by Pennes' bioheat equation and external heat transfer due to convection, radiation and sweating. Furthermore, realistic and temperature-dependent material properties are applied to the human tissues and to the components of thermoelectric generator. The goal of this work is to find an efficient low-rank approximation of this complex, nonlinear model by proper orthogonal decomposition and dynamic mode decomposition techniques.

1 Introduction

In the last couple of years, aging of the population is the main concern especially in European countries [1]. Concerning this, various developments in the medical sector were provoked. The development of electrically active implants is a special boon in regeneration therapies and deep brain stimulation to treat movement disorders. Among the various factors affecting the performance of implants, their limited

G. Sadashivaiah (✉)
Institute for Electronic Appliances and Circuits, University of Rostock, Rostock, Germany
e-mail: gunasheela.sadashivaiah@uni-rostock.de

C. Yuan · T. Bechtold
Institute for Electronic Appliances and Circuits, University of Rostock, Rostock, Germany

Department of Engineering, Jade University of Applied Sciences, Wilhelmshaven, Wilhelmshaven, Germany
e-mail: chengdong.yuan@jade-hs.de; tamara.bechtold@jade-hs.de

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_28

battery period is a major drawback. Several energy harvesting technologies have been proposed to prolong the operational time of the implants [2, 3]. One promising device solution is a thermoelectric generator (TEG). It is an energy harvesting device that transforms thermal into electrical energy by Seebeck effect when embedded in the human body [4].

The design optimization of such TEG device and its driving circuitry depends on an accurate, efficient thermal tissue model. The authors in [5] integrated a TEG model within a simplified linear 3D cubic tissue model. Krylov-subspace based model order reduction (MOR) method [6] was successfully employed to obtain the low-rank approximation of the model. In [7], the authors considered a realistic human torso model with temperature-dependent heat transfer effects. To apply the conventional MOR method, linearization techniques were adopted during model reduction. In this work, we consider more realistic material properties, heat transfer effects and boundary conditions in the torso model, which are temperature-dependent. To enable efficient design optimization, we investigate the feasibility of nonlinear MOR methods such as proper orthogonal decomposition (POD) [8, 9] and dynamic mode decomposition (DMD)[10, 11].

2 Case Study

In this section, we present the model of TEG incorporated in the fat tissue in the chest region of the human torso model. The aim of numerical simulations is to find the temperature difference across the TEG.

Figure 1a represents the setup of an electrically active implant and Fig. 1b describes the model of TEG that was constructed in ANSYS®Workbench [12] based on the available commercial TEGs. The geometry consists of top and bottom ceramic plates made of aluminum oxide, with the cross-sectional area of 24.6×24.6 mm^2 and height of 0.565 mm. The junction between two plates encloses an array of 16×16 p-type and n-type thermocouple legs. The legs are made of temperature-dependent bismuth telluride, each with height of 2.27 mm and cross-section of

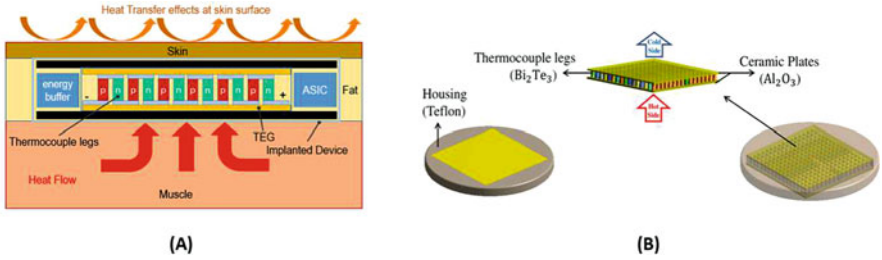


Fig. 1 (**a**) Schematic of a TEG integrated inside the human tissue; (**b**) TEG model with 16×16 thermocouple legs and housing

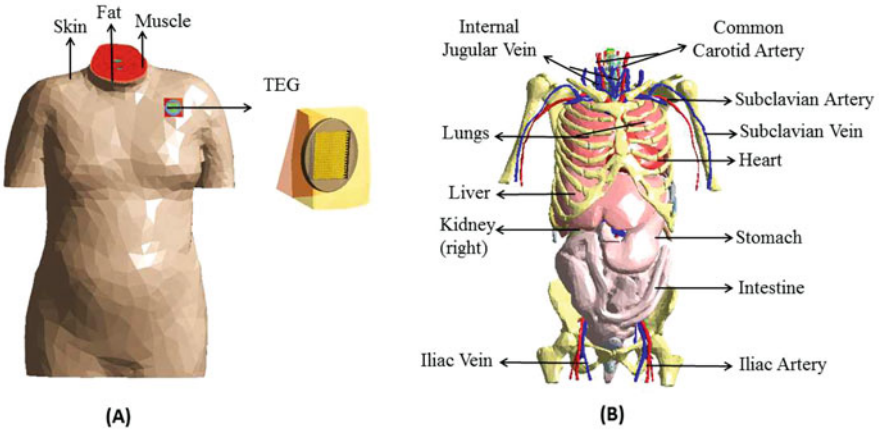


Fig. 2 (**a**) TEG embedded in the fat layer of chest region; (**b**) human torso model with internal organs

0.8×0.8 mm^2. Additionally, TEG is surrounded by a disk-like housing made of Teflon and with radius of 19 mm. The Seebeck voltage produced by the TEG is usually of few microvolts and is given by: $V_{teg} = \frac{n}{2} \cdot \Delta T(\alpha_1 - \alpha_2)$ where, n is the number of thermocouple legs, ΔT is the temperature difference across the thermocouple legs, and α_1, α_2 are the Seebeck coefficients of p and n doped bismuth telluride.

To achieve the realistic temperature distribution effect, the 3D human torso model was modeled based on the segmented magnetic resonance imaging data [13] as shown in Fig. 2a. The model consists of solid internal organs, blood vessels, skeleton and tissue layers as shown in Fig. 2b. The realistic material properties for tissues are applied, as given in [14]. The internal heat transfer in living tissues is given by Pennes' bioheat model [15]:

$$\rho c \frac{\partial T}{\partial t} = \nabla \kappa \nabla T + \underbrace{\rho_b c_b \omega (T_a - T(\vec{r}, t))}_{Q_b(T)} + Q_m, \tag{1}$$

where, ρ, c, and κ are the density, specific heat, and thermal conductivity of tissues respectively. ρ_b, c_b, and ω denote the density, specific heat, and perfusion rate of blood. To maintain the core temperature at 37 °C, heat transfer occurs internally due to metabolic heat generation Q_m, temperature dependent blood perfusion Q_b, and thermal conduction. T_a is the arterial blood temperature and $T(\vec{r}, t)$ is the unknown temperature of the human tissues.

To maintain the natural balance in the body, excess heat is transferred to the ambient environment through the skin surface, which is given by:

$$q_{skin} = \underbrace{h_c(T_{\Gamma_{skin}} - T_{amb})}_{q_{conv}} + \underbrace{\sigma \epsilon (T^4_{\Gamma_{skin}} - T^4_{amb})}_{q_{rad}} + \underbrace{h_e(P_{skin} - P_a)}_{q_{eva}}, \tag{2}$$

where, q_{conv}, q_{rad} and q_{eva} are the convection, radiation, and evaporation effects applied as heat flux inputs to the skin surface in ANSYS. T_{amb} represents the ambient temperature. The variable h_c represents the heat transfer coefficient in W/m^2/K, $\sigma = 5.6705 \times 10^{-8}$ W/m^2/K^4 represents the Stefan-Boltzmann constant and $\epsilon = 0.95$ represents the emissivity. Heat loss due to evaporation occurs mainly in the form of sweating through the skin surface. In the evaporation term P_{skin} and P_a represents the saturated vapour pressure at skin temperature and partial vapour pressure respectively. In accordance with the Lewis relation [16], the evaporation coefficient h_e can be represented in terms of the heat transfer coefficient as:

$$\frac{h_e}{h_c} = 16.5 \ \frac{\text{K}}{\text{kPa}}. \tag{3}$$

The relative humidity is $\phi = \frac{P_a}{P_{sa}}$. Antoine's equation for saturated vapour pressure at skin surface P_{skin} and saturated vapour pressure P_{sa} are defined as:

$$P_{skin} = 0.1\exp(18.956 - \frac{4030.18}{T_{skin} + 235}) \ \text{ in } \ \text{kPa}, \tag{4}$$

$$P_{sa} = 0.1\exp(18.956 - \frac{4030.18}{T_{amb} + 235}) \ \text{ in } \ \text{kPa}. \tag{5}$$

Therefore, the final equation for the evaporation heat loss is represented by:

$$q_{eva} = 1.65 h_c w\{\exp(18.956 - \frac{4030.18}{T_{skin} + 235}) - \phi \cdot \exp(18.956 - \frac{4030.18}{T_{amb} + 235})\}, \tag{6}$$

where, w represents the skin wettedness and its value range between $0.06 - 1$.

The spatial discretization of the model (1) with boundary conditions (2) at the skin surface leads to the following large-scale system of nonlinear ordinary differential equations (ODEs):

$$\sum_N \begin{cases} E \cdot \dot{T}(t) = A(T) \cdot T(t) + \underbrace{B \cdot u(T(t))}_{F(T(t))}, \\ y(t) = C \cdot T(t), \end{cases} \tag{7}$$

where, $A(T) \in \mathbb{R}^{N \times N}$ is the temperature-dependent global heat conductivity matrix and $E \in \mathbb{R}^{N \times N}$ is the constant heat capacity matrix. $B \in \mathbb{R}^{N \times m}$, $C \in \mathbb{R}^{p \times N}$ are the input and output matrices respectively, with m as the number of inputs and p as the number of outputs. $y(t) \in \mathbb{R}^p$ defines the output vector and $u(T(t)) \in \mathbb{R}^m$ is the temperature-dependent load vector. The system is nonlinear due to radiation effect, which corresponds to fourth power of temperature, evaporation effect in the form of sweating and temperature-dependent material properties of the thermocouple

legs. The length of unknown state vector $T(t) \in \mathbb{R}^N$ is $N = 1,340,734$, which defines the dimension of the full system (7). We would like to emphasize that the nonlinear input effects can be linearized, and conventional Krylov-subspace based MOR can be employed. But in our case, the heat conductivity of thermocouple legs is considered temperature-dependent and hence, the nonlinear MOR methods have to be applied.

3 Model Order Reduction

In this work, we employ two different MOR approaches to compute the reduced order model (ROM) of system (7), the proper orthogonal decomposition and the dynamic mode decomposition.

3.1 Proper Orthogonal Decomposition

One of the most common methods for reducing the dimensionality of the nonlinear systems is POD, which is also known as Karhunen-Loève (KL) decomposition or the principal component analysis. The method employs singular value decomposition (SVD) to construct the optimal projection subspace, also called the reduced basis, which captures most of the dynamics of the given data-set [8, 9]. In this work, we compute the projection subspace ϕ_{pod} by employing the POD technique, which is used in conjunction with the Galerkin projection [17] to obtain the ROM:

$$\sum_r \begin{cases} E_r \cdot \dot{T}_r(t) = A_r(T_r) \cdot T_r(t) + \phi_{pod}^T F(\phi_{pod} T_r(t)), \\ y(t) = C_r \cdot T_r(t), \end{cases} \tag{8}$$

where, $E_r = \phi_{pod}^T E \phi_{pod}$, $A_r(T_r) = \phi_{pod}^T A(T) \phi_{pod}$, $C_r = C\phi_{pod}$ are the reduced matrices and accuracy between the full system and ROM is given by $\parallel T(t) - \phi_{pod} T_r(t) \parallel$.

3.2 Dynamic Mode Decomposition

Dynamic mode decomposition is a data-driven method, i. e. it uses the measured data for constructing the ROM. According to Schmid [10], DMD is a special case of Koopman theory [18] and thus we begin with the definition of Koopman operator for the nonlinear system (7), written in symbolic form:

$$\frac{dT(t)}{dt} = N_l\{T(t)\}, \tag{9}$$

where, $T(t) \in \mathscr{M}$, an N-dimensional manifold and N_l is a nonlinear operator. The Koopman operator $\mathscr{K}$, acts on set of observable functions $g : \mathscr{M} \rightarrow \mathbb{C}$ so that: $\mathscr{K} g(T(t)) = g(N_l\{T(t)\})$. Thus, $\mathscr{K}$ is a linear operator that maps nonlinear system in the state space to linear system in the observable space. The method yields the matrix A_n, which is the approximation of the Koopman operator:

$$\frac{d\tilde{T}(t)}{dt} = A_n \tilde{T}(t), \tag{10}$$

where, $\tilde{T}(t) \in \mathbb{R}^N$ is the approximate solution and $A_n \in \mathbb{R}^{N \times N}$ is the matrix, which defines the best-fit linear dynamics using only the measured data from the full system. The main objective of the method is to find the linear approximation operator A_n, so that the true and approximated solution remain close in a least square sense, i.e $\|T(t) - \tilde{T}(t)\|$.

In general, the computed matrix A_n is highly ill-conditioned. The speciality is that, the method predicts the future state of the system by exploiting the low-rank structures, associated from the eigendecomposition of A_n. Therefore, the approximated solution via the associated low-rank structures is given by:

$$\tilde{T}(t) = \sum_{i=1}^{r} b_i \psi_i^{dmd} \exp(\omega_i t), \tag{11}$$

where, ψ_i^{dmd} is the DMD basis of rank r, ω_i is the eigenvalues of the matrix A_n and b_i is the initial condition.

4 Numerical Simulation Results

In this work, all the computations are performed on a PC with an Intel Xeon E5-2680, 2.5 GHz, 128 GB RAM, 4 active cores processor. Initially, steady-state thermal simulation of the model is conducted with $T_{amb} = 25$ °C and $h_c = 3.1$ W/m^2/K. The result of the steady-state simulation is considered as the initial values to conduct the transient thermal simulation for the new value of $h_c = 5.48$ W/m^2/K. For both approaches, snapshot matrix is built out of 20 equidistant snapshots with step size $\Delta t = 350$ s for $t \in [0, 7000]$. The singular values obtained by performing SVD of the snapshot matrix are shown in the Fig. 3a.

Figure 3b and c represents the maximum relative error (%) between the full order model of dimension $N = 1,045,923$ and ROMs of dimension r=3 (POD) and r=4 (DMD). The maximum relative error of the POD approach amounts to 0.051%, and of DMD approach amounts to 0.071%. The runtime for transient thermal simulation of the full system amounts to 5460 s. In both approaches, the time required to construct snapshot matrix and to perform SVD in offline stage is 5460 + 137.5 s,

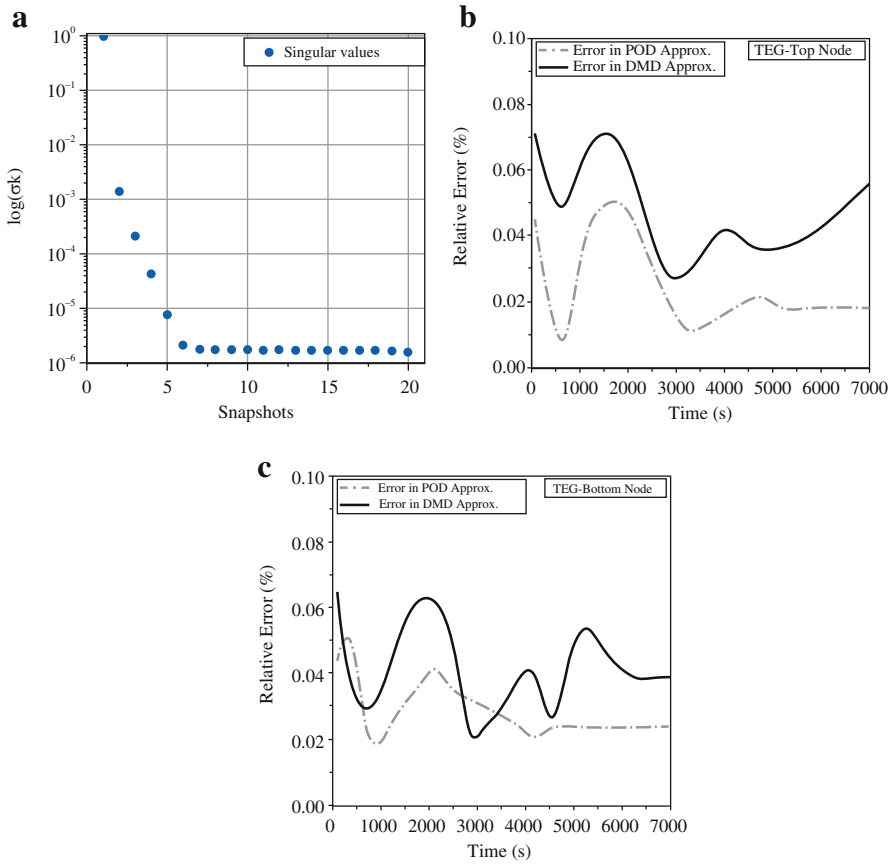


Fig. 3 (**a**) Singular values σ_k of the full human torso model, (**b**) and (**c**) represents the maximum relative error at selected nodes of full and reduced order model

but in online stage POD requires only 63.1 s and DMD only 22.7 s for computation of ROMs.

5 Conclusion and Outlook

In this work, the low-rank approximation of the large-scale nonlinear thermal human torso model was generated via POD and DMD methods. The main advantage of the DMD method over POD is, that it does not require any information about the governing equations of the system to be solved. However, in terms of error convergence rate, POD-based MOR method is superior compared to DMD method due to the fact that DMD modes are not orthogonal. In future, both approaches will

be tested on the parameterized human torso models, with skin wettedness w and ambient temperature T_{amb} as parameters.

References

1. C. Casey, J. Gullo, 2018 Aging readiness and competitiveness report. AARP Int. J. **12**, 14–15 (2019). https://doi.org/10.26419/int.00036.003
2. S. Priya, D.J. Inman, *Energy Harvesting Technologies* (Springer, New York, 2009)
3. M.A. Hannan, M. Saad, S.A. Samad, A. Hussain, Energy harvesting for the implantable biomedical devices: issues and challenges. Biomed. Eng. Online **13**, 79 (2014)
4. P. Miao, P.D. Mitcheson, A.S. Holmes, E.M. Yeatman, T.C. Green, B.H. Stark, Mems inertial power generations for biomedical applications. Microsyst. Technol. **12**(10–11), 1079–1083 (2006)
5. O. Jadhav, C.D. Yuan, D. Hohlfeld, T. Bechtold, Design of a thermoelectric generator for electrical active implants, in *MikroSystemTechnik Congress* (2017), pp. 1–4
6. R.W. Freund, Krylov-subspace methods for reduced-order modeling in circuit simulation. J. Comput. Appl. Math. **123**, 395–421 (2000)
7. C. Yuan, S. Kreß, G. Sadashivaiah, E.B. Rudnyi, D. Hohlfeld, T. Bechtold, Towards efficient design optimization of a miniaturized thermoelectric generator for electrically active implants via model order reduction and submodeling technique. Int. J. Numer. Methods Biomed. Eng. **36**, e3311 (2020)
8. P. René, *Model Reduction via Proper Orthogonal Decomposition* (Springer, Berlin, Heidelberg, 2008), pp. 95–109
9. A. Quarteroni, A. Manzoni, F. Negri, *Reduced Basis Methods for Partial Differential Equations* (Springer, Cham, 2016), pp. 115–140
10. P. Schmid, Dynamic mode decomposition of numerical and experimental data. J. Fluid Mech. **656**, 5–28 (2010)
11. A. Alla, J.N. Kutz, Nonlinear model order reduction via dynamic mode decomposition. SIAM J. Sci. Comput. **39**, 1–20 (2016)
12. Ansys®Academic Research Mechanical, Release 2020 R1, Workbench
13. S. Makarov, G. Noetscher, J. Yanamadala, VHP-Female Datasets. NEVA Electromagnetics, LLC:VHP-Female 2.2 edn. (2015)
14. P.A. Hasgall, G.F. Di, C. Baumgartner, IT'IS database for thermal and electromagnetic parameters of biological tissues, IT'IS Foundation, Switzerland, (2018)
15. H.H. Pennes, Analysis of tissue and arterial blood temperatures in the resting human forearm. J. Appl. Physiol. **1**(2), 93–122 (1948)
16. W.K. Lewis, The evaporation of a liquid into a gas. **44**, 445–446 (1922)
17. P. Benner, S. Gugercin, K. Willcox, A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Rev. **57**, 483–531 (2015)
18. B.O. Koopman, Hamiltonian systems and transformation in Hilbert space. Proc. Natl. Acad. Sci. **17**, 315–318 (1931)

Multi-Level Iterations for Microgrid Control with Automatic Level Choice

Robert Scholz, Armin Nurkanović, Amer Mešanović, Jürgen Gutekunst, Andreas Potschka, Hans Georg Bock, and Ekaterina Kostina

Abstract Microgrids are considered a key technology for the energy transition, but the rising penetration of renewable energy sources is pushing current control approaches to their limits. Nonlinear model predictive control (NMPC) is a promising approach to address this issue, although achieving real-time feasibility with standard schemes is challenging. Therefore, we propose to use the Multi-Level Iteration NMPC scheme with a novel automatic level choice. This allows us to always use the most accurate linearizations available while being real-time feasible, even during strongly transient phases where a fixed level choice may be too slow. We use a realistic-sized microgrid to illustrate the capabilities of this method.

1 Introduction

Microgrids (MG) are small, local electrical networks comprising heterogeneous components, such as generators, storage systems and loads. They are managed autonomously and are operated either as an islanded network or within a connected larger network. This allows the MG to be considered as a single controllable entity in the utility grid. MGs are a promising approach to handle the rising number of renewable energy sources (RES). Their ability to handle unforseen disturbances on a local level is likely to make them a key technology to enable more distributed and heterogeneous networks[1].

Tight operational bounds on frequency and voltage make MG control a challenging task, especially during demanding load scenarios. State-of-the-art methods rely on a hierarchical control structure, using proportional-integral controllers and

R. Scholz (✉) · J. Gutekunst · A. Potschka · H. G. Bock · E. Kostina
Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany
e-mail: robert.scholz@iwr.uni-heidelberg.de

A. Nurkanović · A. Mešanović
Siemens AG, Munich, Germany
e-mail: armin.nurkanovic@siemens.com

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_29

various filters on different levels. However, experience shows that this control paradigm reaches its limits under high penetration of RES [1].

Nonlinear model predictive control (NMPC) is a general model-based methodology to control dynamical processes. Measurements of the process are embedded in an optimal control problem, which is solved repeatedly with respect to an objective function and operational limits. Since NMPC offers a flexible control framework, it appears to be well suited for the control of MGs.

One of the main challenges for NMPC in the field of power engineering is the real-time requirement. Transient electrical dynamics involve high-frequency oscillations, which are costly to simulate, and a high sampling rate is necessary to react to disturbances in time. Therefore, in the literature, NMPC is mainly used on a higher control level, relying on traditional integral or droop controllers to handle the electrical dynamics [2]. To deal with the fast electrical behavior of MGs directly with NMPC, tailored schemes are necessary, like the *Advanced Step Real-Time Iteration* [3].

In this paper, we propose to use the *Multi-Level Iteration* (MLI) scheme for MG control. This approach is based on the well-established *Real-Time Iteration* [4], which eliminates the need to solve the underlying optimal control problems until convergence. Additional computation time is saved by updating the problem linearization only partially in every iteration using cheap update formulas to increase the feedback rates and render NMPC applicable for MG control.

2 Nonlinear Model Predictive Control

NMPC is a general framework to control dynamic processes modeled by differential- algebraic equations (DAE). At each time t_k within a given sequence of sampling points $t_0 < t_1 < \dots$ a feedback signal u_{ξ_k} is computed based on the current system state ξ_k. In the traditional NMPC setting, this is done by solving an optimal control problem (OCP) with fixed time horizon of length T of the form:

$$\min_{x(\cdot),z(\cdot),u(\cdot)} \quad \Phi(x(\cdot), z(\cdot), u(\cdot)) \tag{1a}$$

$$\text{s.t.} \quad \dot{x}(t) = f(x(t), z(t), u(t)), \quad 0 = h(x(t), z(t), u(t)), \tag{1b}$$

$$x(t_k) = \xi_k, \quad t \in [t_k, t_k + T], \tag{1c}$$

$$x^{\text{lo}} \le x(t) \le x^{\text{up}},\ z^{\text{lo}} \le z(t) \le z^{\text{up}},\ u^{\text{lo}} \le u(t) \le u^{\text{up}}. \tag{1d}$$

The differential and algebraic states $x(t)$ and $z(t)$ are subject to the DAE system (1b) with initial value set to the current system state ξ_k (1c). The control inputs of the process are represented by $u(t)$ and the objective is defined by the functional Φ. The NMPC feedback signal applied in the interval $[t_k, t_{k+1})$ is the first part of the solution $u_{\xi_k}(t) = u_k^*(t)$. We discretize the problem with the direct multiple shooting

method introduced by Bock [5] and obtain a finite dimensional, structured nonlinear program (NLP) of the compact form

$$\min_{w} l(w) \qquad \text{s.t.} \qquad b(w) + E\xi_k = 0, \qquad w^{\text{lo}} \leq w \leq w^{\text{up}}. \tag{2}$$

Here l is the discretized objective function (1a), the function b together with the constant matrix E represent the discretized DAE system (1b) with the initial value embedding constraint (1c) and w^{lo} and w^{up} are the lower and upper bounds on states and controls. A sequential quadratic programming (SQP) method is used to solve this NLP, which generates a sequence of primal-dual iterates $(w_j, \lambda_j, \mu_j)_{j\in\mathbb{N}}$ based on the quadratic program (QP)

$$\min_{\Delta w} \frac{1}{2}\Delta w^\top A \Delta w + a^\top \Delta w \quad \text{s.t.} \quad \begin{cases} b(w_j) + E\xi_k + B\Delta w = 0, \\ w^{\text{lo}} \leq \Delta w + w_j \leq w^{\text{up}}. \end{cases} \tag{3}$$

The matrix A is the Hessian (or an approximation thereof) with respect to w of the Lagrangian $\mathscr{L}(w_j, \lambda_j, \mu_j)$ of the NLP (2). The linear objective term is defined by the objective gradient $a = \nabla_w l(w_j)$ and the constraints are linearizations based on $b(w_j)$ and its Jacobian $B = \nabla b(w_j)^\top$. The solution $(\Delta w^{QP}, \lambda^{QP}, \mu^{QP})$ of QP (3) is used to update the primal-dual variables:

$$w_{j+1} = w_j + \Delta w^{QP}, \qquad \lambda_{j+1} = \lambda^{QP}, \qquad \mu_{j+1} = \mu^{QP}. \tag{4}$$

Under mild assumptions, local quadratic convergence of the SQP method is guaranteed. As a consequence, once the iterates are sufficiently close to the true solution, only one iteration per sampling time is sufficient to obtain excellent solution approximations for the NLPs at subsequent sampling times [4]. The Real-Time Iteration, introduced by Diehl [4, 5], allows a further speedup as it exploits that the initial value ξ_k only enters linearly in the QP (3). This means that most of the QP data can be prepared based on the current iterate (w_j, λ_j, μ_j), before ξ_k is known. As soon as ξ_k becomes available, only the QP solution step is necessary to generate the feedback signal.

3 Multi-Level Iterations

Depending on the application, the Real-Time Iteration still requires a high computational effort in every iteration. To set up the QP (3), the constraints, the objective gradient, the constraint Jacobian and the Hessian (corresponding to b, a, B, A in (3)) have to be computed. MLI can reduce this computational effort drastically and thus speed up the feedback process by only updating parts of the QP.

The MLI scheme is based on the fact that Newton-type methods (such as the SQP method described in the previous section) do not require the exact computations of

Table 1 Computations and update formulas for the QP data for the different Levels

	Necessary computations				Update formula for QP data			
Level	$b(w_j)$	$a(w_j)$	$B(w_j)$	$A(w_j, \lambda_j)$	b	a	B	A
D	✓	✓	✓	✓	$b(w_j)$	$a(w_j)$	$B(w_j)$	$A(w_j, \lambda_j)$
C	✓	✓	(✓)[a]	✗	$b(w_j)$	$a(w_j) + (\bar{B}_C^\top - B(w_j)^\top)\lambda_j$	$\bar{B}_C$	$\bar{A}_C$
B	✓	✗	✗	✗	$b(w_j)$	$\bar{a}_B + \bar{A}_B(w_j - \bar{w}_B)$	$\bar{B}_B$	$\bar{A}_B$
A	✗	✗	✗	✗	$\bar{b}_A$	$\bar{a}_A$	$\bar{B}_A$	$\bar{A}_A$

[a] Only the vector-matrix product $\lambda^\top B$ needs to be computed in an adjoint fashion

derivatives in A and B to remain locally convergent. This can be exploited to avoid the expensive evaluation of the Hessian and the Jacobian in every iteration.

Instead, the different components of the QP (3) are updated in four hierarchical levels with descending computational complexity. Each level stores a reference point $(\bar{w}, \bar{\lambda}, \bar{\mu})$ and the corresponding QP data $\bar{b}$, $\bar{a}$, $\bar{B}$ and $\bar{A}$. Every level is working on its own set of iterates, which are independent of the other levels. Table 1 explains which data is computed in each iteration and how the QP data is updated. The convergence is usually analyzed for a fixed system state ξ_k. *Level D* corresponds to a full SQP step and therefore inherits its local quadratic convergence. *Level C* avoids the full Jacobian evaluation, but is still converging to an optimal point of the original NLP (2). *Level B* abandons the sensitivity generation completely and convergences to a feasible point. *Level A* refers to linear MPC, since all QP data is fixed. It provides feedback with the lowest computational effort, but for nonlinear models convergence can not be guaranteed. A detailed description of the levels can be found in [6, 7] and their convergence properties are analyzed in [5].

3.1 Automatic Level Choice

In practice, the presented levels are operated simultaneously. The lower levels are used to give fast feedback and the higher, computationally more expensive levels provide accurate linearizations of the NLP (2). Usually this is done by a sequence of levels, which is fixed in advance. To ensure real-time feasibility, the required computation time must be estimated and the sequence chosen accordingly [6].

This approach turns out to be inflexible, because the evaluation of a level needs to be finished before the next evaluation is scheduled. In order to be real-time feasible, the worst-case computation time needs to be treated. If an adaptive integration method is used, the integration time may vary strongly between the steady state and transients. This leads to an unnecessary conservative scheduling of higher levels, even when the computation time is low and would allow a faster rate. To overcome this issue, we propose to choose the levels automatically online instead. In this method every level is operated in parallel. In the beginning of the simulation, the evaluation of all levels is triggered. As soon as an evaluation is finished, the

corresponding level is marked as ready. When the simulation of the controlled process reaches the next sampling point, the linearization of the highest level, which is marked as ready, is applied to the QP (3). The lower levels are reinitialized with the updated QP data $\bar{b}, \bar{a}, \bar{B}$ and $\bar{A}$, and a new evaluation starts. Since level A includes no reevaluation of the QP data, its computation time is very low and it can always be used as a fallback, if no other level is ready. With this approach, we can always use the most accurate available linearizations whilst at the same time having the guarantee of remaining real-time feasible.

4 Dynamic Microgrid Model

To deal with dynamics on different time scales, MG control is typically organised in three hierarchical levels. *Primary control* relies on local internal controllers of the components and has the goal to stabilize the system. *Secondary control* is responsible for eliminating any steady-state error introduced by the primary control. Long term planning and the incorporation of weather forecast and load prediction is done by the *tertiary control*. Our proposed controller is working on the secondary control level. It receives control signals from the tertiary control and sets the reference values of primary controllers of the individual components. The test MG model comprises not only the physical components, but also the primary control level.

4.1 Scenario and Model Description

We use the model of a test MG which is comprehensively presented in [3]. Here, we give a short overview. The structure of the MG system is depicted in Fig. 1. It comprises two identical diesel generators (DG), a battery (BA), a photovoltaic plant (PV), and a passive PQ-load. The DGs consist of a synchronous generator actuated by a diesel engine with a governor for frequency stabilization (IEEE DEGOV1) and are equipped with an Automatic Voltage Regulator (IEEE AC5A). The setpoints for frequency ω_{ref} and voltage V_{ref} serve as control variables of the MLI-controller. The battery is modeled as a constant DC voltage source connected to an inverter with an internal droop. It is controlled by the setpoints for frequency ω_{BA}^{ref} and voltage V_{BA}^{ref}. The base power of the MG is $S_{grid} = 100\,\text{kVA}$, the nominal power of the generator is $S_{DG} = 325\,\text{kVA}$, and the nominal power of the battery is $S_{BA} = 150\,\text{kVA}$. The complete MG is given as a DAE system of index 1 with 37 differential and 42 algebraic states and 6 control inputs.

To demonstrate the capabilities of the proposed controller, we apply a challenging load scenario. At the beginning, the system is in a steady state and the reference values for the battery are set to $P_{BA}^{ref} = Q_{BA}^{ref} = 0\text{p.u.}$. The generators share the load

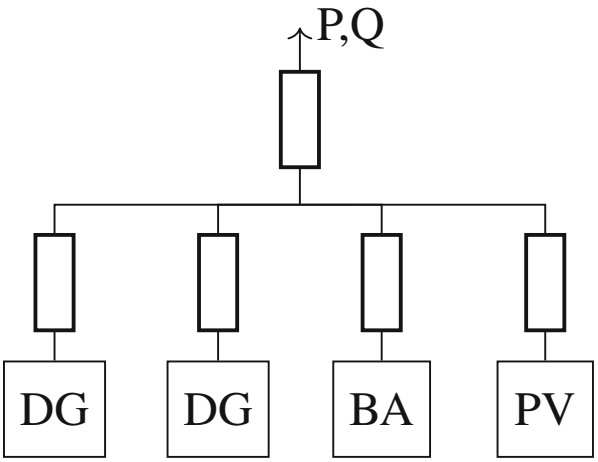


Fig. 1 Topology of the test MG

of $P_{load} = 5$p.u. and $Q_{load} = 1$p.u. equally. After 1 s a sudden unscheduled load step of 40% in active and reactive power takes place, which exceeds the capacity of the generators. To ensure that the operational limits are satisfied, the battery needs to leave the provided reference values and serve the missing load. The simulation has an overall length of 8 s.

In MG control, we need to consider several objectives with different priorities. This is modeled by a continuous least squares objective functional

$$\Phi(x, z, u) = \int_{t_i}^{t_i+T} \|r(x(t), z(t), u(t))\|^2 \, \mathrm{dt} \tag{5}$$

of OCP (1) with a weighted norm and a residual function $r(x, z, u)$. The most important goal is to steer the frequency $\omega(t)$ and voltage at the load $V_{load}(t)$ to the nominal value 1p.u. after a disturbance. During transients, we want to utilize the battery to stabilize frequency and voltage. In steady state, the performance of the battery should follow setpoints P_{BA}^{ref}, Q_{BA}^{ref} from a higher control level, in order to charge or discharge the battery. The generators are supposed to share the remaining load equally. These goals are achieved by tracking terms

$$\begin{aligned} r_1(x, z, u) &= \omega - 1, & r_2(x, z, u) &= V_{\mathrm{load}} - 1, \\ r_3(x, z, u) &= P_{\mathrm{BA}} - P_{\mathrm{BA}}^{ref}, & r_4(x, z, u) &= Q_{\mathrm{BA}} - Q_{\mathrm{BA}}^{ref}, \\ r_5(x, z, u) &= P_1 - P_2, & r_6(x, z, u) &= Q_1 - Q_2. \end{aligned}$$

5 Numerical Results

We discretize OCP (1) with two multiple shooting intervals and the length of the prediction horizon is fixed to $T = 1\,\text{s}$. The length of the first shooting interval corresponds to the sampling time of 100 ms and the second to 900 ms. The numerical simulations are carried out with the NMPC framework MLI [6]. For integration and sensitivity generation, the SolvIND integrator suite is used and the QPs are solved by qpOASES [8].

The continuous least squares objective function (5) enables us to use a Gauß-Newton approximation of the Hessian in QP (3). Besides its favorable numerical properties, its main advantage is, that it relies only on first-order derivatives. Therefore, we do not have to compute second-order derivatives, which is the most costly task when evaluating QP (3).

We compare our proposed MLI-controller with a typical state-of-the-art control setup for small microgrids: The generators are equipped with an integral controller for steady-state error elimination of the frequency with a settling time of approximately 20 s and a sampling time of 100 ms. The voltage setpoint V_{ref} is kept constant during the full simulation time.

In Fig. 2, the performance of the proposed MLI-controller is shown in comparison to the traditional control approach. The MLI-controller steers back the

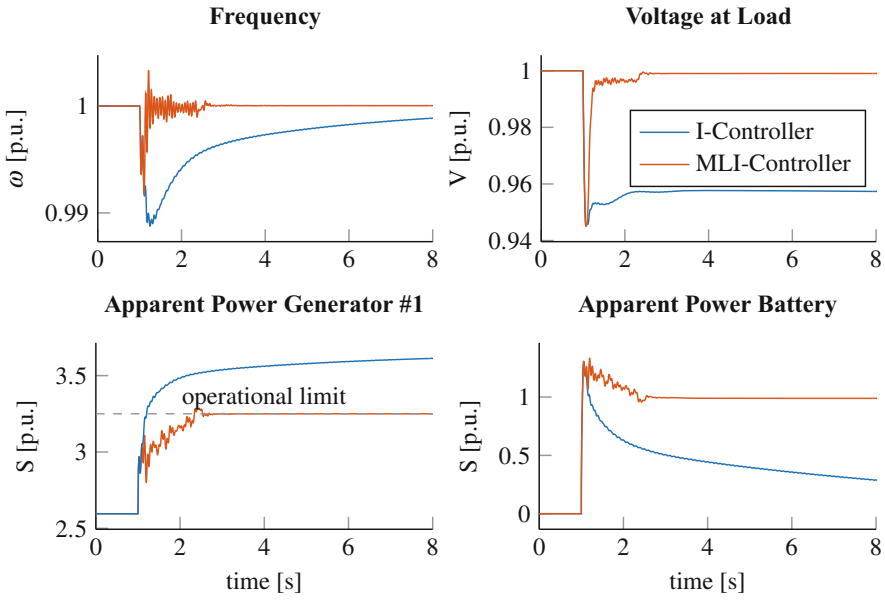


Fig. 2 Control performance of the MLI-controller in comparison to a traditional control approach. In the top row, the frequency and the voltage at the load is depicted. In the bottom row, the apparent power of the generators and the battery is shown. The MLI-controller is able to steer frequency and voltage back to the nominal value faster and with a lower initial drop, while respecting the operational limits

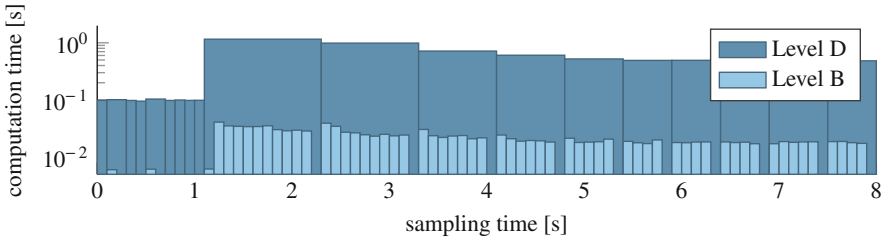


Fig. 3 Computation times and scheduling of Level B and D. The elapsed computation time is depicted by the height of the bars while the width shows in which sampling intervals the computations were performed

frequency faster with a lower initial drop after the unforeseen disturbance. The voltage gets stabilized faster and the steady state offset is eliminated. In the beginning, the battery follows the setpoints and does not contribute in load sharing. After the load drop, the fast reacting dynamics of the battery are used to stabilize the system. Since the overall load exceeds the operational bounds of the generators, the battery temporarily deviates from its reference value and instead serves the necessary additional load. In contrast to this, the integral-controller is not able to obey the operational limits of the generators. If there are no safety measures installed, the generators are overloaded, which may cause physical damage. In Fig. 3, the computation time and the scheduling of the different MLI levels are shown. In the beginning, the system is in a steady state and the computation time is low. After the load jump at $t = 1\,\mathrm{s}$, the system is in a transient phase and the computation time rises sharply, which leads to less level D evaluations. Afterwards, the system gets steered back to a steady state and the computation time decreases. As the evaluation time for level B is always below the sampling time, no level A occurs. Level C is not used, because the Gauß-Newton approximation of the Hessian implies that the difference in computation time between level C and D is low.

6 Conclusion

In this paper a novel MG controller based on Multi-Level Iterations is presented. We used a realistic-sized example MG, modeled with a DAE, to perform numerical experiments and compared the performance with a traditional control approach. In the example shown, the MLI-controller did not only outperform the traditional controller, but also was able to respect operational bounds.

Acknowledgments This research was funded by the German Federal Ministry of Education and Research (BMBF) in the research project MOReNet. (Grant No 05M18VHA)

References

1. M. Ilić, R. Jaddivada, X. Miao, Modeling and analysis methods for assessing stability of microgrids, in *20th IFAC World Congress*, vol. 50, IFAC-PapersOnLine (2017), pp. 5448–5455
2. C. Bordons, F. Garcia-Torres, M.A. Ridao, Model predictive control of microgrids, in *Advances in Industrial Control* (Springer International Publishing, Cham, 2019)
3. A. Nurkanović, A. Mešanović, A. Zanelli, G. Frison, J. Frey, S. Albrecht, M. Diehl, Real-time nonlinear model predictive control for microgrid operation, in *2020 American Control Conference (ACC)* (2020), pp. 4989–4995
4. M. Diehl, Real-Time Optimization for Large Scale Nonlinear Processes. Dissertation, Heidelberg University, 2001
5. H.G. Bock, M. Diehl, E. Kostina, J. Schlöder, Constrained optimal feedback control of systems governed by large differential algebraic equations, in *Real-Time PDE-Constrained Optimization* (SIAM, Philadelphia, 2007), pp. 3–22
6. L. Wirsching, Multi-level iteration schemes with adaptive level choice for nonlinear model predictive control. Dissertation, Heidelberg University, 2018
7. A. Nurkanović, S. Albrecht, M. Diehl, Multi level iterations for economic nonlinear model predictive control, in *Recent Advances in Model Predictive Control: Theory, Algorithms, and Applications*, ed. by T. Faulwasser, M.A. Müller, K. Worthmann (Springer, Cham, 2021), pp. 65–105
8. H.J. Ferreau, C. Kirches, A. Potschka, H.G. Bock, M. Diehl, qpOASES: A parametric active-set algorithm for quadratic programming. Math. Program. Comput. **6**, 327–363 (2014)

Multi-Level Inversion Based on Mesh Decoupling

Benny Shachor, Hadi Hajibeygi, and Domenico Lahaye

Abstract Accurate material characteristics in many real-field engineering applications can only be determined via inverse modeling. This is due to the fact that often times measuring field quantities in the resolution required to model and predict engineering processes are impossible. Despite several advancements, still it remains a challenge as to how to characterise material properties through an inverse modeling approach. Special challenge is driven when the scale of the field is large, while the parameters are expected to be defined at high resolution. Such challenge demands for scalable inverse modeling techniques, clearly beyond the scope of classical single-level approaches. In this paper, we propose a new multi-level approach based on mesh decoupling of the state and design variables. This approach allows for treating the design variables on various scales without comprising the accuracy of the state and adjoint equation solve. The performance of the new method is investigated for estimation of the heterogeneous parameters of one-dimensional and two-dimensional elliptic equations. Results illustrate that the mesh decoupling technique provides a promising framework for solving large-scale heterogeneous systems.

1 Introduction

Inverse modeling is an important step in defining system characteristics when available measurement data are insufficient to fully describe the parameters. Special challenge is imposed on real-life engineering applications, as the scale of the problem is very large while the necessary parameters are expected to be defined on high resolutions. This challenge makes classical inverse modeling approaches

B. Shachor (✉) · H. Hajibeygi
Department of Geoscience and Engineering, TU Delft, Delft, Netherlands
e-mail: H.Hajibeygi@tudelft.nl

D. Lahaye
Delft Institute of Applied Mathematics, TU Delft, Delft, Netherlands
e-mail: D.J.P.Lahaye@tudelft.nl

M. van Beurden et al. (eds.), *Scientific Computing in Electrical Engineering*,
Mathematics in Industry 36, https://doi.org/10.1007/978-3-030-84238-3_30

computationally too expensive to be applicable. Recently, a multiscale inverse strategy has been introduced [3], where the state variable was represented in a coarser resolution, while parameters stayed in the fine scale.

The current research develops a new approach based on the multi-level decoupling technique. Through sequences of parameter restrictions, we first optimise the mismatch between the observation and measurements at the coarsest space. These coarse-scale parameters provide good approximations for the initial parameters at a higher resolution. We perform this procedure until the fine-scale converged parameters are obtained. Note that at any stage, depending on the value of the objective function, one can stop iterations. As such, convergence on the fine-scale resolution is not necessary. This is quite relevant for field applications where only a good approximation of the parameters is acceptable, if they are found efficiently.

2 Problem Description

In this section we subsequently introduce the partial differential equation constrained least squares problems we intend to solve, a classical discrete adjoint solution method and a Newton Trust-Region algorithm [2]. The discrete adjoint method is explained in more detail in [10]. The method is compared with a continuous adjoint method in [6] and the references cited therein. The literature on least squares optimization methods with partial differential equation constraints is very vast. Recent monographs include [1, 5, 9].

Non-linear Least-Squares Problem We intend to solve non-linear least squares problems in which the state variable $u(\underline{x})$ is constrained by a partial differential equation [5, 10]. As state equation we consider the diffusion equation with spatially-varying diffusion coefficient $k(\underline{x})$ and source term $f(\underline{x})$ on a computational domain Ω with homogeneous Dirichlet boundary conditions. For notational convenience we assume Ω to be the unit interval or the init square. This assume is not restrictive for the subsequent arguments we wish to make. Our goal is to recover the spatial variation of $k(\underline{x})$ by solving the state equation for multiple sources $f_j(\underline{x})$ and evaluating the computed solution at various receiver locations. Let N_s and N_r denoted the number of sources and receivers, respectively. The set of differential equations that we solve can then for $1 \leq j \leq N_s$ be written as

$$\nabla \cdot (k(\underline{x}) \nabla u_j) = f_j \text{ on } \Omega. \tag{1}$$

The source function $f(\underline{x})$ is defined as a set of point sources at the source locations. The state equation for $u(\underline{x})$ is solved as many times as the number of sources. At each solve, one source is set to have amplitude one and all other source to amplitude zero. After each solve, the computed solution is evaluated at the receiver locations. After N_s state equation solves, N_s vectors $\vec{u}_j$ of size N_r are available. Each of these vectors depends on $k(\underline{x})$.

A set of measured values at receiver locations is assumed to be available for each source. This set is a set of N_s vectors $\vec{d}_j$ of size N_r. We will work with synthetic data generated by solving the state equation assuming the diffusion equation to be equal to the exact value.

The cost function that we intend to minimize consists of two terms. The first term is the sum over j of the Euclidean norm of the discrepancies $\vec{u}_j - \vec{d}_j$. The second term is a regularization term that penalizes large variations in $k(\underline{x})$. It is proportional to the Euclidean norm of the gradient of $k(\underline{x})$. The second term is multiplied with a weight α before being added to the least squares part of the cost functional. The objective is to find the argument $k(\underline{x})$ that minimizes the cost functional and thus to recover the heterogeneity of the medium that best describe the observations. The set of N_s state equations act as partial differential equation constraints in this minimization problem.

Discrete Adjoint Method We adopt a classical discrete adjoint solution approach. This means that we first discretize the cost functional and the differential equation constraints on Ω. We subsequently differentiate the discrete Lagrangian to obtain the non-linear system that defines the first order critical points. This non-linear system is solved by a Newton method.

We adopt a standard Galerkin finite element method using linear shape functions on segments in one dimension and triangles in two dimensions. We represent the diffusion coefficient by a constant per element. The resulting vector $\vec{k}$ is the discrete set of design variables. In each of the j linear systems for the discrete state $\vec{u}_j$, the coefficient matrix A_j depends on the discrete diffusion coefficient $\vec{k}$. We add to the discretized cost functional the sum over j of the discrete Lagrange parameters $\vec{v}_j$ times the residual vector in $\vec{u}_j$. The discrete Lagrangian obtained depends on the $2\,N_s + 1$ vectors $\vec{u}_j$, $\vec{v}_j$ and $\vec{k}$. First order critical points of the Lagrangian are obtained by setting the derivatives with respect to these vectors equal to zero.

Trust-Region Newton Method The non-linear system that defines the first-order critical points is solved in two steps. First the N_s decoupled linear state and adjoint equations are solved. In solving for the adjoint variables $\vec{v}_j$, the linear algebra of solving for $\vec{u}_j$ is reused. In this paper the LU-factorization of the coefficient matrix $A_j(\vec{k})$ is recycled. Given that the state equation (1) depends on j through the source term $f_j(\underline{x})$ only, only one LU-factorization for all j's is required. The large scale applications of the multi-level inversion that we here propose does however require adopting iterative solution methods as proposed in [3, 4]. In the second step the non-linear system for $\vec{k}$ is solved by trust-region globalized Newton method starting from an initial guess. The gradient vector is computed from the available state and adjoint vectors at negligible computational cost. The Hessian matrix is computed column-wise by finite difference approximation of the gradient vector. This requires N_s additional state and adjoint solves for each component of the vector $\vec{k}$. The Hessian computes therefore dominates the computational cost of the procedure. In future work, this can be alleviated by a Gauss-Newton or a Broyden-Fletcher-Goldfarb-Shanno (BFGS) rank-one iterative approximation of the Hessian.

3 Nested Iteration Based on Mesh Decoupling

In this section we describe a multi-level optimization procedure that exploits a decoupling of the mesh for state and design variables.

Mesh Decoupling Procedure The computational cost of the procedure outlined in the previous section is dominated by the number of design variables given by the number of segments in 1D or triangles in 2D. We wish to accelerate the solution of the optimization problem by reducing the mesh size without lowering the accuracy of the computed solution. We achieve this goal by representing the design variables $\vec{k}$ on a more global scale while at the same time preserving the fine scale resolution for the state $\vec{u}_j$ and adjoint variables $\vec{v}_j$. We therefore introduce two decoupled meshes on the domain of computation Ω.

The first mesh coincides with the classical notion of a mesh to solve for $\vec{u}_j$ and $\vec{v}_j$ given values for $\vec{k}$. The second mesh serves to represent $\vec{k}$. In this work we keep the first mesh fixed at a sufficiently fine level and is denoted by Ω_{fine}. The second mesh is defined independently from the first by a hierarchy denoted by denoted by Ω_k. The coarsest mesh Ω_1 in the design variables $\vec{k}$ has four elements along x in the one-dimensional problem and two elements along x and y in the two-dimensional problem. The finest mesh in $\vec{k}$ is comparable in size to the mesh in $\vec{u}_j$ and $\vec{v}_j$. On each mesh Ω_k, the diffusion coefficient is represented by element-wise constant values. The coefficient values are transferred from the mesh Ω_k to the next finer mesh Ω_{k+1} and from the mesh Ω_k to the mesh Ω_{fine} using injection.

On the coarse meshes Ω_k the gradient and the Hessian have lesser components and are therefore easier to compute. The problem on the coarser mesh has fewer variables and is expected to converge in less iterations.

Components of the Multi-Level Optimization Algorithm The hierarchy of meshes Ω_k allows to define a multi-level solution procedure by as nested iteration. This procedure is started by solving the optimization problem on the mesh Ω_1 using the constant value of 0.5. Having solved the optimization problem on the mesh Ω_{k-1}, the design variables are injected to the finer mesh Ω_k and used as initial guess for the Newton method. The procedure is repeated until reaching the mesh Ω_{fine}. The trust-region method acts as a *smoother* on the meshes Ω_k and Ω_{fine}.

Choice of the Regularization Parameter on Multiple Levels The least-squares problem that we solve is regularized by adding a term that penalizes large variations in the spatial distribution of $\vec{k}$. This spatial variation is measured by the Euclidean norm of the gradient of the design variables $\vec{k}$. This norm is equal to the energy norm of $\vec{k}$ and scales with the mesh width of Ω_k [10]. The factor that weighs the regularization term in the cost functional requires to be adjusted accordingly when traversing the meshes Ω_k in the multi-level algorithm. We employ a numerical scheme such that the energy norm of Ω_k scales with the square of the mesh width of Ω_k. We therefore divide the weighting factor by four in moving from Ω_k to Ω_{k+1} in the multi-level algorithm. On finer meshes Ω_k, the norm of variations in $\vec{k}$ is larger and should be given less weight.

The Algorithm The Multi-Level algorithm is presented below.

Algorithm 4: The multi-level algorithm

Input: Initial guess, α
Output: N_{fine} design variables

```
k_1 := initial guess
for l = 1 : N_level do
    while Trust-Region Newton does not converge do
        Inject k from Ω_k → Ω_fine
        Compute the Cost, Gradient and Hessian functions on Ω_k
    end
    Inject k from Ω_k → Ω_{k+1}
    α = α/4
end
```

4 Numerical Results

In this section we compare the performance of the single-level and multi-level algorithms applied to the one-dimensional and two-dimensional test problems.

The one-dimensional and two-dimensional test problem were taken from [10] and [5], respectively. Details of both test problems are given in Table 1. The assumed exact spatial dependence of the design variables is given by

$$k(x) = -0.2\mathrm{e}^{-72(x-0.45)^2}, \tag{2}$$

in the one-dimensional problem and by

$$\begin{aligned} k(x, y) &= 15, && \text{on } ((x+0.5)^2 + (y+0.5)^2) \leq (0.3^2) \\ k(x, y) &= 300, && \text{on } ((x-0.5)^2 + (y-0.5)^2) \leq (0.3^2) \\ k(x, y) &= 10, && \text{otherwise} \end{aligned} \tag{3}$$

Table 1 One-dimensional and two-dimensional problem definition

	One-dimensional	Two-dimensional
Computational domain	$0 \leq x \leq 1$	$0 \leq x, y \leq 1$
# elements in Ω_{fine}	64	1368
# sources (N_s)	2	5
Sources Locations	$[\frac{1}{3}],[\frac{2}{3}]$	[-0.5,-0.5],[-0.5,0.5],
		[0,0],[0.5,-0.5],[0.5,0.5]
Source Expression	Dirac Delta at the node	Dirac Delta at the node
# Receiver (N_r)	5	5
Observations Locations	[0.1, 0.3 , 0.5, 0.7, 0.9]	Location of the sources

in the two-dimensional problem. In the one-dimensional problem, the mesh Ω_1 is refined four times such that the mesh Ω_5 coincides Ω_{fine}. In the two-dimensional problem, the mesh Ω_1 is refined four times such that each element in Ω_{fine} overlaps with only one element in Ω_5. In the final stage of the multi-level algorithm applied to the two-dimensional problem, the design variable distribution is defined on the unstructured triangular mesh Ω_{fine}.

The converge of the single-level and multi-level optimization algorithm are presented in Figs. 1 and 2 for the one-dimensional and two-dimensional problem, respectively. Both figures show the decrease in the cost-functional vs. a measure of the computational cost. The computational cost of a single PDE solve is given by the number of elements in the mesh Ω_{fine} and is equal in both the single and multi-level algorithm. We therefore adopt the number of PDE solves as a fair metric to compare the single-level and multi-level algorithm. The numerical differentiation to obtain the Hessian is responsible for the bulk of the number of PDE solves. The color in the graphs corresponds to the various levels. In the last stage of the convergence, both the single and multi-level algorithm operate on the same mesh in $\vec{k}$ as indicated by the color coding.

Both Figs. 1 and 2 show that the single-level algorithm converges initially slow and subsequently very fast. The trust-region method brings the Newton algorithm in the region of super-linear convergence around the solution. The multi-level

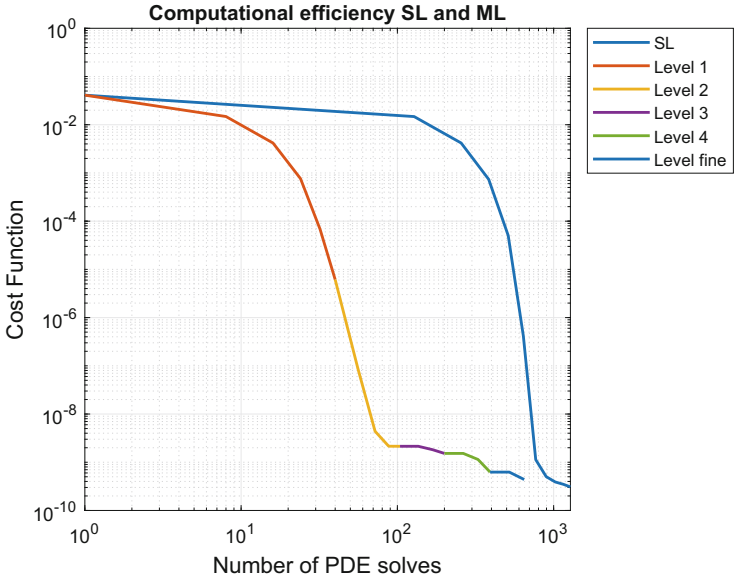


Fig. 1 Comparison of the single-level (SL) and multi-level (ML) algorithm by considering the convergence history in the cost functional as function of the number of PDE solves in the one-dimensional problem. The single-level method converges slowly until reaching the basin of attraction. Subsequently is converges faster. The multi-level method converges super-linearly on each level

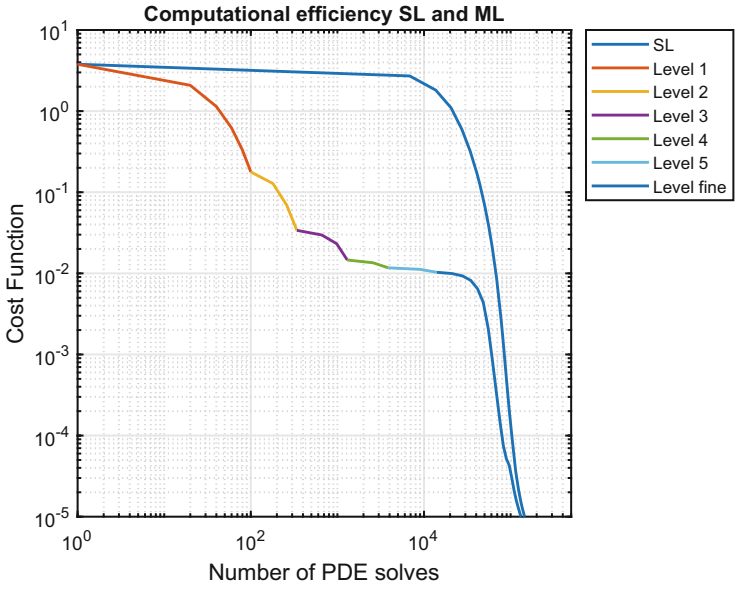


Fig. 2 Comparison of the single-level (SL) and multi-level (ML) algorithm by considering the convergence history in the cost functional as function of the number of PDE solves in the two-dimensional problem. Similar observation as in the one-dimensional problem in Fig. 1

algorithm is seen not to suffer from slow convergence on the finest level. The method is on the contrary seen to significantly reduce the cost functional in a limited number of PDE solves in the coarsest levels in $\vec{k}$. The solution from the coarser level provides a sufficiently good initial approximation for the Newton algorithm to converge superlinearly on a given level. The multi-level algorithm requires considerably less PDE solves to reach solutions with moderate accuracy. However, to reach the final solution, both the single and multi-level algorithm require the same number of PDE solves. This is due to the fact that on the finest level both the single and multi-level algorithm converge very fast. Further research is required to circumvent this issue.

Figure 3 shows the converge of the single-level and multi-level algorithm in terms of the evolution of the design variables $\vec{k}$. The figure shows that the multi-level algorithm avoids premature small scale variations in $\vec{k}$. The figure also shows that the single and multi-level algorithm converge to the same solution. The same message is conveyed in Fig. 4 for the two-dimensional problem.

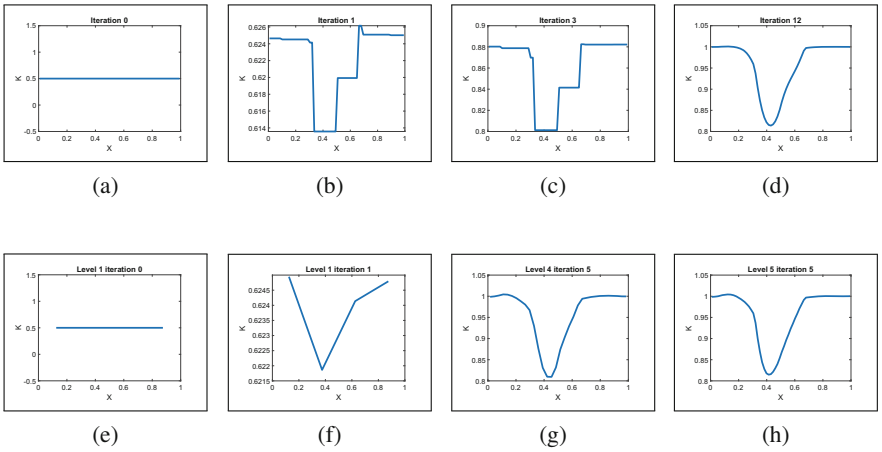


Fig. 3 Comparison of the single-level (SL) and multi-level (ML) algorithm by considering the convergence history in the design variables $\vec{k}$ starting from the same initial guess in the one-dimensional problem. The second column of pictures shows that the ML algorithm avoids premature small scale variations in $\vec{k}$. (**a**) SL, iteration 0. (**b**) SL, iteration 1. (**c**) SL, 384 PDE solves. (**d**) SL, final result. (**e**) ML, level 1, Iteration 0. (**f**) ML, level 1, iteration 1. (**g**) ML, 392 PDE solves. (**h**) ML, final result

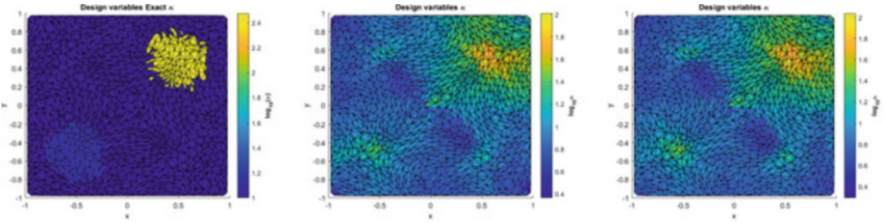


Fig. 4 Assumed exact distribution of the design variables $\vec{k}$ (left) and the solution found by the single-level (middle) and multi-level algorithm (right)

5 Conclusions

We presented a multi-level algorithm to solve non-linear least square problems with a Poisson equation for the state variable as constraint. The algorithm exploits the discretization of the state and adjoint variable on a fixed fine mesh and the representation of the design variables on a mesh hierarchy with varying spatial scale. This decoupling allows to coarsen the space of design variables while at the same time preserving accurate state and adjoint solves. Numerical results show that the multi-level algorithm avoids small scale variations of the design variables in early stages of the algorithms. The multi-level algorithm therefore converges significantly faster than its single-level counterpart on coarse scales in the design variables. The

multi-level algorithm is thus valuable to adopt in often occurring scenarios in which a coarse scale representation of the design variables yields valuable information. The development and results are publicly available at [7] and [8].

References

1. D. Chavent, *Nonlinear Least Squares for Inverse Problems* (Springer, New York,, 2010)
2. A.R. Conn, N.I. Gould, P.L. Toint, *Trust Region Methods* (SIAM, Philadelphia, 2000)
3. R. de Moraes, H. Hajibeygi, J.D. Jansen, A multiscale method for data assimilation. Comput. Geosci. **24**, 425–442 (2020)
4. D. Echeverría, P.W. Hemker, Manifold mapping: a two-level optimization technique. Comput. Vis. Sci. **11**(4–6), 193–206 (2008)
5. E. Haber, *Computational Methods in Geophysical Electromagnetics* (SIAM, Philadelphia, 2014)
6. D. Lahaye, W. Mulckhuyse, Adjoint sensitivity in PDE constrained least squares problems as a multiphysics problem. COMPEL: Int J Comput. Math. Electri. Electron. Eng. **31**(3), 895–903 (2012)
7. B. Shachor, Multi-Level Inversion Based On Mesh Decoupling. Master's thesis, TU Delft, 2019
8. B. Shachor, Multi-Level Inversion Based on Mesh Decoupling for Poisson inverse problems with Dirichlet BC. https://github.com/Bennyshachor/Multi-Level. Cited 12 Jan 2020
9. A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation* (SIAM, Philadelphia, 2005)
10. C.R. Vogel, *Computational Methods for Inverse Problems* (SIAM, Philadelphia, 2002)