TECHNISCHE
UNIVERSITÄT
DARMSTADT

# On structure preserving simulations in nonlinear electromagnetics, electric circuits, and efficient treatment of systems with memory

Dem Fachbereich Mathematik
der Technischen Universität Darmstadt
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr.rer.nat)
genehmigte Dissertation

von

Vsevolod Shashkov, M.Sc.
aus Nischni Nowgorod, Russische Föderation

Referent : Prof. Dr. Herbert Egger
Korreferent : Prof. Dr. Sebastian Schöps
Tag der Einreichung: 18.10.2023
Tag der mündlichen Prüfung: 02.02.2024

Darmstadt
D17

**On structure preserving simulations in nonlinear electromagnetics, electric circuits, and efficient treatment of systems with memory**

Accepted doctoral thesis by Vsevolod Shashkov, M.Sc.

Darmstadt, Technische Universität Darmstadt

Date of thesis defense: February 02, 2024
Year of publication on TUprints: 2024

# Acknowledgments

First of all, I wish to thank my supervisor, Herbert Egger, for his excellent guidance. His door was always open, and he always found time for discussion despite his busy schedule. Discussions with him were inspiring and very productive.

Special thanks go to my colleagues from the numerics group. They always had my back and gave helpful advice. The warm and welcoming atmosphere in the group makes the work very pleasant.

Support by the German Science Foundation (DFG) via TRR 146 (project C3), TRR 154 (project C04), SPP 2256 (project Eg-331/2-1), and GSC-CE is gratefully acknowledged.

Finally, I would like to thank my family and my girlfriend. Their support and care took me through the tough times.

# Abstract

This thesis is dedicated to the modelling and numerical treatment of electromagnetic phenomena governed by (nonlinear) field and circuit equations, which are the fundamental topics in electrical engineering. The main focus is on energy transformation principles and the construction of discretization schemes that preserve these principles. While the numerical treatment of linear problems has been extensively studied in the literature over the years, a systematic treatment of nonlinear problems is not yet fully established.

In the first chapter of the thesis, Maxwell's equations in nonlinear media are discussed. We consider an energy-based modelling approach for material description and present two formulations that lead to systems of certain generalized (port-) Hamiltonian and gradient structures. To preserve the underlying structure, we employ variational techniques based on Galerkin approximations in space and discontinuous- or Petrov-Galerkin methods in time. This approach allows a systematic construction of higher-order schemes based on implicit time-stepping. The discrete energy balance can be derived under relatively general assumptions. For energy-conserving systems, the two approaches enable the construction of dissipative and energy-conserving schemes, ensuring the passivity of the discretizations.

The second chapter is dedicated to electric circuit problems. The state-of-the-art approach to modelling electric circuits is Modified Nodal Analysis (MNA). This formulation leads to differential-algebraic systems with an index of $\nu \leq 2$, which presents challenges in the analysis and numerical treatment. We present an alternative magnetic-oriented nodal analysis formulation (MONA) that results in differential-algebraic systems with an index $\nu \leq 1$, which is much simpler to handle. We demonstrate that both formulations result in finite-dimensional systems of a certain port-Hamiltonian or gradient structure, similar to the field problems. Therefore, variational time-stepping methods can again be utilized to construct passivity-preserving higher-order time-discretization schemes.

In the last chapter, the systems with memory described by a Volterra-integro-differential equation are considered. Such systems arise in the context of dispersive media or reduced order models for field circuit coupling. The numerical treatment of such problems requires an efficient realization of the integral term in an evolutionary manner. After an appropriate discretization, the Volterra integral term can be interpreted as a matrix-vector product with a densely populated matrix. For a sufficiently fine approximation, the size of the system becomes large, leading to storage and complexity issues. We present a fast, oblivious, and evolutionary algorithm based on the $\mathcal{H}^2-$matrix compression technique. The approach can be applied to Volterra integrals of convolution type. Further, it shares some similarities with the fast and oblivious quadrature methods of Schädle, Lopez-Fernandez, and Lubich. The latter can be interpreted as a particular realization of the $\mathcal{H}$-matrix approximation.

# Zusammenfassung

Diese Dissertation widmet sich der Modellierung und numerischen Behandlung von elektromagnetischen Feldern und elektrischen Schaltkreisen, die grundlegende Themen der Elektrotechnik sind. Der Schwerpunkt liegt auf den Energietransformationsprinzipien und der Entwicklung von numerischen Verfahren, die diese erhalten. Während der lineare Fall im Laufe der Jahre gut verstanden wurde, ist eine systematische Behandlung nichtlinearer Probleme noch nicht vollständig etabliert.

Im ersten Kapitel werden die Maxwell-Gleichungen in nichtlinearen Medien betrachtet. Wir verwenden einen energiebasierten Ansatz für die Modelierung der Materialgesetze und präsentieren zwei Formulierungen. Die erste Formulierung führt auf Probleme mit einer verallgemeinerten Hamiltonschen Struktur. Die zweite Formulierung besitzt Gradientenstruktur. Um die Struktur des Problems zu erhalten, verwenden wir Galerkin-Methoden im Raum und discontinuous Galerkin bzw. Petrov-Galerkin-Verfahren in der Zeit. Dadurch ist es möglich, Verfahren höherer Ordnung zu konstruieren, die auf impliziter Zeitintegration basieren. Die diskrete Energiebilanz kann unter vergleichsweise allgemeinen Voraussetzungen hergeleitet werden. Für energieerhaltende Systeme ermöglichen diese beiden Verfahren die Konstruktion von dissipativen und energieerhaltenden Schemata.

Das zweite Kapitel wird Modellierung und Diskretisierung elektrische Schaltkreise diskutiert. Die klassische Methode ist Modified Nodal Analaysis (MNA). Die Formulierung führt zu differenzial-algebraischen Systemen mit einem Index von $\nu \leq 2$, was potenzielle Herausforderungen in der Analyse und Diskretisierung mit sich bringt. Wir präsentieren eine alternative Magnetic Oriented Nodal Analysis (MONA) Formulierung, die zu differenziell-algebraischen Systemen mit einem Index von $\nu \leq 1$ führt und somit die Behandlung vereinfacht. Wir zeigen, dass die MNA- und die MONA-Formulierung zu endlichdimensionalen Systemen mit verallgemeinerten Hamiltonschen und Gradientenstrukturen führen, ähnlich den Feldproblemen. Dies ermöglicht wiederum die Konstruktion von Passivität-erhaltenden Methoden, welche auf den variationellen Zeitintegrationsverfahren basieren.

Im letzten Kapitel werden die Systeme mit Memory betrachtet, die durch eine Volterra-Integro-Differentialgleichung beschrieben werden. Probleme dieser Art entstehen im Kontext dispersiver Materialien oder Feld-Schaltkreis-Kopplungen. Die numerische Behandlung solcher Probleme erfordert eine effiziente Umsetzung des Integralterms auf evolutionäre Weise. Nach einer geeigneten Diskretisierung kann der Volterra-Integralterm als Matrix-Vektor-Produkt mit einer dicht besetzten Matrix interpretiert werden. Für eine ausreichend feine Diskretisierung wird die Größe des Systems in Hinblick auf Speicher und Komplexität problematisch. Wir präsentieren einen schnellen, vergesslichen und evolutionären Algorithmus, der auf der $\mathcal{H}^2-$Matrixkompression basiert. Der Ansatz kann auf Volterra-Integrale vom Faltungstyp angewendet werden. Dieser weist einige Ähnlichkeiten mit den FOCQ-Methoden von Schädle, Lopez-Fernandez und Lubich auf, welche als eine spezifische Realisierung der $\mathcal{H}$-Matrixkompression interpretiert werden kann.

# Contents

# Introduction

This thesis discusses the modeling and numerical treatment of certain types of nonlinear problems in electrical engineering. The main focus is on energy transformation principles and the construction of discretization schemes that preserve these principles.

**Electromagnetic fields.** The first problem under consideration is the electromagnetic wave propagation in nonlinear Kerr-type media, which is a typical application in nonlinear optics. A significant feature of the governing system of partial differential equations is their passivity, i.e., the change in energy is determined by the power inflow through the sources and the dissipation through Ohmic losses. In dielectric media with no sources and losses and no power flow over the boundary, the energy of the system is conserved exactly.

The development of numerical schemes that preserve this property is an active field of research. Several methods have been proposed over the years; see e.g. [2, 18, 74]. However, most of the existing approaches provide low-order approximations only. As we will see, the main difficulty in constructing higher-order methods lies in time integration. Furthermore, the schemes are developed explicitly for Kerr-type media; generalizing them to different types of nonlinear media is not straightforward. In this thesis, we address this issue and discuss strategies for the systematic construction of higher-order discretization schemes in space and time applicable to a relatively general class of nonlinear problems. The key ingredients are to use particular formulations and to utilize variational methods for the overall discreitzation process.

A classical formulation for Maxwell's equations is in terms of the electric and magnetic fields $\mathbf{e}$ and $\mathbf{h}$. Its particular (port-) Hamiltonian structure motivates the use of Galerkin methods in space and discontinuous Galerkin time-stepping schemes, as suggested in [42]. Following this strategy, a systematic construction of passivity-preserving schemes of higher order is possible. However, the resulting schemes have some numerical dissipation. While this effect decreases when decreasing the time step size or increasing the polynomial degree of approximation, numerical dissipation may become an issue for long-time simulation.

As a second attempt, we consider an approach based on the magnetic vector potential $\mathbf{a}$ as a system unknown. The vector potential is a standard tool in low-frequency approximations used to describe magnetic devices and electrical machines. It has been observed that the vector potential formulation for the eddy current problem has a different canonical energy-based structure, which can be understood as a generalization of gradient flow systems. A suitable discretization strategy that preserves the underlying energy balance has been proposed in [43]. This approach employs Galerkin approximation in space and Petrov-Galerkin time-stepping, allowing for the systematic construction of higher-order schemes. These ideas can be extended to Maxwell's wave propagation problem. To this end, we here consider a formulation based on the electric field $\mathbf{e}$ and the magnetic vector potential $\mathbf{a}$. In contrast to the previous strategy, this method preserves the energy balance

exactly, making it perfectly suited for energy-conserving problems. The corresponding results were published in [48].

**Electric circuits.** Simulation of electric circuits is another topic discussed in this thesis. The state-of-the-art approach for circuit modeling is Modified Nodal Analysis (MNA) [62, 69]. This approach uses vectors of electric node potentials and some branch charges as primary unknowns, leading to a differential-algebraic system of equations. It is well understood that under appropriate assumptions on the elements, the governing systems have an index $\nu \leq 2$, where the index depends on the topology of the circuit; see, e.g., [52, 110]. The differential-algebraic nature of MNA systems presents several challenges, and the numerical investigation of DAEs has been greatly inspired by circuit simulations. While index $\nu \leq 1$ systems are relatively easy to handle, the analysis and numerical treatment of index-2 problems are challenging; see, e.g., [66, 81]. In particular, classical implicit methods like the trapezoidal rule or BDF-2 schemes may become unstable in the case of strong nonlinearities [62].

The MNA formulation shares several similarities with the $\mathbf{e}-\mathbf{h}$ formulation of Maxwell's equations; in particular, it has a similar (port-) Hamiltonian structure, which automatically ensures the passivity of the formulation. With this in mind, the systematic construction of passivity-preserving discretization schemes can once again be achieved using discontinuous Galerkin time-stepping. The lowest-order scheme coincides with the implicit Euler method, the state-of-the-art approach for MNA systems [62].

Following the philosophy of the vector potential formulation for Maxwell's equations, we proposed the Magnetic Oriented Nodal Analysis (MONA) formulation for electric circuits [122]. This formulation uses flux linkage potentials as the primary unknowns, which are the integrated quantities corresponding to the electric node potentials used in MNA, similar to how the vector potential is the integrated quantity related to the electric field. Besides the similar modeling viewpoint, the two formulations share the same generalized gradient flow structure, which guarantees the passivity of the MONA formulation. Furthermore, Petrov-Galerkin techniques can again be used for the systematic construction of higher-order schemes that preserve the underlying energy balance of the system.

The main advantage of the MONA formulation over the MNA approach is that it leads to systems of a smaller index. While MNA systems have an index $\nu \leq 2$, the index of MONA systems is $\nu \leq 1$, which drastically simplifies analysis and numerical treatment. The index is again determined by the circuit's topology, with the topological conditions being very similar to those for MNA. In fact, the index of a MONA system is typically smaller by one compared to that of an MNA system in most cases.

A typical application of MONA is in the context of field-circuit coupling. Such multiscale models are often used for the accurate description of power transformers or electrical machines. The fields are typically modeled by the vector potential formulation, while the circuit is described by the MNA approach [117, 137]. Due to the different geometric structures of the two subsystems, the construction of passivity-preserving schemes is not straightforward. However, when MONA is used instead, the coupled system shares a common generalized gradient flow structure. Consequently, Galerkin methods in space and Petrov-Galerkin type time integrators can once again be employed for the systematic construction of higher-order schemes.

Solutions to large-scale field-circuit coupled problems are often computationally expensive. In the typical scenario, the field part of the system drastically dominates the system size and therefore represents the bottleneck of the method. Different techniques can be used to improve computational costs. In particular, multirate methods [117, 118] have often been used in this context. However, it is not yet clear whether the discrete energy balance can be preserved in this case.

**Reduced models.** In the case where the field subsystem is linear, the degrees of freedom associated with the field subsystem can be eliminated in the frequency domain, leading to an equivalent Volterra-integro-differential system with a convolution-type integral term. Similar techniques are often used in the context of dispersive media. The size of the system is essentially determined by the size of the circuit subsystem, which is advantageous. However, solving this system requires an evolutionary evaluation of the convolution integral, which might become expensive for a larger number of time steps. Furthermore, the convolution kernel is given only implicitly via its Laplace transform. In such cases, Convolution Quadrature methods (CQ) [89, 90] provide a suitable discretization strategy. As discussed in [39, 46], an appropriate choice of CQ method allows for the construction of schemes such that the numerical solutions of the reduced Volterra-integro-differential and full coupled formulations coincide. Consequently, the discrete energy balance remains valid after discretization of the reduced system.

As previously mentioned, the evaluation of convolution integrals can become expensive. For $N$ time steps, the naive implementation requires $O(N^2)$ operations and $O(N)$ active memory to store the history of the solutions. An improved approach, called Fast and Oblivious Convolution Quadrature (FOCQ), achieves a complexity of $O(N \log N)$ and requires $O(\log N)$ memory. A further enhancement is based on the key observation that FOCQ can be viewed as a particular realization of the $\mathcal{H}$ matrix realization technique. With this in mind we developed a fast evolutionary and oblivious algorithm based on the $cH^2$ approach, which requires only $O(N)$ operations and $O(\log N)$ memory [40]. This approach is applicable for general Volterra-type integrals, making it useful for a large class of systems with memory.

## Structure of the thesis

The thesis is structured into three chapters, each addressing specific areas: electromagnetic field problems, electric circuits, and systems with memory. Let us briefly overview the content and highlight the main contributions. A detailed introduction to each discussed topic is provided at the beginning of the corresponding chapter.

### Chapter 1: Electromagnetic waves in nonlinear media

The first chapter is dedicated to the simulation of electromagnetic fields. Although the results apply to a large class of electromagnetic problems, we restrict our analysis to nonlinear optics and consider the wave propagation problem in nonlinear dielectric media.

We start by recalling an energy-based approach to modeling material laws, which is the key ingredient in the derivation of power and energy balances. We present two formulations based on fields $\mathbf{e}$ and $\mathbf{h}$ and $\mathbf{e}$ and $\mathbf{a}$, respectively. For each formulation, we discuss its structure and derive the energy transformation principle. We apply variational

methods in space and time, derive the corresponding discrete energy balance, and discuss implementation details. We illustrate theoretical results in a series of numerical experiments and compare the methods to several well-known approaches. This part is strongly based on our publication [48].

Finally, we also show that the presented ideas can be extended to problems with dispersion, which often play an important role in nonlinear optics. We discuss the Kerr-Lorentz model in detail; however, different types of dispersive behavior can be handled analogously. The ideas have been partially presented in [39].

### Chapter 2: Electric circuits

The second chapter of the thesis is devoted to electric circuit simulation. We begin the chapter by providing basic concepts of circuit modeling. We discuss an energy-based concept for modeling circuit elements, which is the essential assumption to obtain power and energy balances. Then, we recall the basic aspects of the Modified Nodal Analysis (MNA) formulation and discuss its (port-) Hamiltonian structure.

We derive the Magnetic Oriented Nodal Analysis (MONA) formulation for electric circuits. We provide a complete index analysis and discuss its generalized gradient flow structure, which ensures passivity and allows the construction of schemes that preserve the underlying energy evolution principle. This part is mostly based on our publication [122]. We also discuss coupling to the field equations in the vector potential formulation and show that the coupled problems again exhibit this canonical structure. We consider stranded and solid conductor models for coupling. Finally, we illustrate the theoretical results in a series of numerical examples.

### Chapter 3: Systems with memory

In the last chapter of the thesis, we discuss the numerical treatment of Volterra-integro-differential systems. We present an evolutionary fast and oblivious algorithm for the approximation of Volterra integrals based on the $\mathcal{H}^2$ technique. We also present an algorithm for Volterra integrals of convolution type, where the kernel is given implicitly in the frequency domain. We discuss the relation of the algorithm to Convolution Quadrature (CQ) methods and its connection to the Fast and Oblivious Convolution Quadrature (FOCQ) algorithm. Finally, we illustrate the theoretical results in a series of examples. The content of this chapter is strongly based on our recent publication [40].

This part of the thesis contains rather abstract results and is not explicitly linked to previously discussed topics and electrical engineering in general. To provide the connection, we consider a magnetic oriented field circuit coupling as one of the examples. Further examples in electrical engineering can be found in our publications [39, 46].

## Publications

The following publications were submitted during the research phase of this thesis and are used as the basis for individual chapters.

### Full list of publications


[46] H. Egger, K. Schmidt, and V. Shashkov. Multistep and Runge–Kutta convolution quadrature methods for coupled dynamical systems. *Journal of Computational and Applied Mathematics*, 387:112618, 2021

[43] H. Egger, O. Habrich, and V. Shashkov. On the energy stable approximation of Hamiltonian and gradient systems. *J. Comput. Meth. Appl. Math.*, 21:335–349, 2021

[47] H. Egger and V. Shashkov. On energy preserving high-order discretizations for nonlinear acoustics. In *Numerical Mathematics and Advanced Applications ENUMATH 2019: European Conference, Egmond aan Zee, The Netherlands, September 30-October 4*, pages 353–361. Springer, 2021

[39] J. Dölz, H. Egger, and V. Shashkov. A convolution quadrature method for Maxwell's equations in dispersive media. In *Scientific Computing in Electrical Engineering: SCEE 2020, Eindhoven, The Netherlands, February 2020*, pages 107–115. Springer, 2021

[40] J. Dölz, H. Egger, and V. Shashkov. A fast and oblivious matrix compression algorithm for volterra integral operators. *Advances in Computational Mathematics*, 47(6):81, 2021

[122] V. Shashkov, I. Cortes Garcia, and H. Egger. MONA—a magnetic oriented nodal analysis for electric circuits. *International Journal of Circuit Theory and Applications*, 2022

[48] H. Egger and V. Shashkov. On higher order passivity preserving schemes for nonlinear Maxwell's equations. *arXiv preprint arXiv:2202.08003*, 2022


### Most relevant publications

The content of this thesis is strongly based on our publications [40, 46, 48, 122]. The results presented there have been detailed, discussed systematically, and put into context, along with extensions. The first chapter also contains an extension to dispersive media, which has not been published yet. In the second chapter, we present a novel magnetic-oriented field-circuit coupling approach, which will be published soon.

# Chapter 1.

# Electromagnetic waves in nonlinear media

Simulation of electromagnetic fields is an active area of research in electrical engineering, which is important for many applications, such as the transmission of radio and optical signals, induction motors, and electrical machines. This chapter discusses the modelling and numerical discretization of problems in high-frequency, high-intensity electrodynamics relevant in nonlinear optics. The main focus of our discussion lies in the modelling of nonlinear materials, the derivation of energy transformation principles, and the construction of discretization schemes that preserve these principles. Before we proceed, let us briefly motivate the topic by a linear case.

### Maxwell's equations in linear media

The mathematical description of electromagnetic phenomena is based on Maxwell's equations [73, 129]. They represent a set of four partial differential equations, namely, Faraday's and Ampere's laws

$$\partial_t \mathbf{b} = -\operatorname{curl} \mathbf{e} \qquad \text{and} \qquad \partial_t \mathbf{d} + \mathbf{j} = \operatorname{curl} \mathbf{h}, \tag{1.1}$$

as well as Gauss's laws $\operatorname{div} \mathbf{b} = 0$ and $\operatorname{div} \mathbf{d} = \rho$ of magneto- and electrostatics. Here $\mathbf{e}$ and $\mathbf{h}$ denote the electric and magnetic fields, $\mathbf{d}$ and $\mathbf{b}$ denote the electric and magnetic fluxes, and $\mathbf{j}$ and $\rho$ represent the electric current density and the charge distribution.

For a complete description, one must further impose material laws, which provide the relations between the fields and the fluxes. In the simplest case of linear isotropic nondispersive materials, these read

$$\mathbf{d} = \epsilon \mathbf{e}, \qquad \mathbf{b} = \mu \mathbf{h}, \qquad \text{and} \qquad \mathbf{j} = \sigma \mathbf{e} + \mathbf{j}_s, \tag{1.2}$$

where constants $\epsilon$ and $\mu$ are the electric permittivity and the magnetic permeability, $\sigma$ is the electric conductivity, and $\mathbf{j}_s$ is the imposed source current. Using these relations, the flux variables can be eliminated, and (1.1) can be written as

$$\mu \partial_t \mathbf{h} = -\operatorname{curl} \mathbf{e}, \tag{1.3}$$

$$\epsilon \partial_t \mathbf{e} = \operatorname{curl} \mathbf{h} - \sigma \mathbf{e} - \mathbf{j}_s. \tag{1.4}$$

### Power balance

The conservation or redistribution of energy is an important principle of a dynamical system, in particular for Maxwell's equations. With the constitutive relations defined as

in (1.2), the electric and magnetic energy densities are given by

$$w_{el}(\mathbf{d}) = \frac{\epsilon^{-1}}{2}|\mathbf{d}|^2 \qquad \text{and} \qquad w_{mag}(\mathbf{b}) = \frac{\mu^{-1}}{2}|\mathbf{b}|^2.$$

The energy densities can also be expressed in terms of the electric and magnetic fields by $w_{el}(\epsilon\mathbf{e}) = \frac{\epsilon}{2}|\mathbf{e}|^2$ and $w_{mag}(\mu\mathbf{h}) = \frac{\mu}{2}|\mathbf{h}|^2$. The change of energy stored in a domain $\Omega$ is then given by

$$\begin{aligned}\frac{d}{dt}\int_\Omega w_{el}(\epsilon\mathbf{e}){+}w_{mag}(\mu\mathbf{h})\; dx &= \int_\Omega \epsilon\partial_t\mathbf{e}\cdot\mathbf{e} + \mu\partial_t\mathbf{h}\cdot\mathbf{h}\; dx \\ &= -\int_\Omega \sigma|\mathbf{e}|^2\; dx - \int_\Omega \mathbf{j}_s\cdot\mathbf{e}\; dx - \int_{\partial\Omega}(\mathbf{e}\times\mathbf{h})\cdot\mathbf{n}\; ds(x),\end{aligned}$$

which is known as *Poynting theorem* [73]. The three terms on the right-hand side of this power balance correspond to the Ohmic losses, the power provided to the system through the power source, and the power flow through the boundary. Integrating in time leads to the corresponding balance of energy, which states that the change of energy is given by the power flow through the boundary and the power source and eddy current losses. With the absence of power flow through the boundary and power source, the system's energy does not increase over time. Hence, the system is *passive* [3, 87]. This relation represents a fundamental physical property of the system, which is relevant for stability analysis and discretization.

### Vector potential formulation

Next to the formulation in terms of fields $\mathbf{e}$ and $\mathbf{h}$, there are many other formulations based on different quantities. In particular, the magnetic vector potential $\mathbf{a}$ is often used to describe electromagnetic phenomena. With the well known relations for vector potential $\mathbf{b} = \operatorname{curl}\mathbf{a}$ and $\mathbf{e} = \partial_t\mathbf{a}$, one can consider the formulation

$$\partial_t\mathbf{a} = -\mathbf{e}, \tag{1.5}$$

$$\epsilon\partial_t\mathbf{e} = \operatorname{curl}(\nu\operatorname{curl}\mathbf{a}) + \sigma\partial_t\mathbf{a} - \mathbf{j}_s, \tag{1.6}$$

where $\nu = \mu^{-1}$ denotes the magnetic reluctivity. This formulation can also be used for the description of electromagnetic fields. The energy densities can then be expressed in terms of the fields $\mathbf{e}$ and $\mathbf{a}$ by $w_{el}(\epsilon\mathbf{e}) = \frac{\epsilon}{2}|\mathbf{e}|^2$ and $w_{mag}(\operatorname{curl}\mathbf{a}) = \frac{\nu}{2}|\operatorname{curl}\mathbf{a}|^2$, and the power balance translates to

$$\begin{aligned}\frac{d}{dt}\int_\Omega w_{el}(\epsilon\mathbf{e}){+}w_{mag}(\operatorname{curl}\mathbf{a})\; dx &= \int_\Omega \epsilon\mathbf{e}\cdot\partial_t\mathbf{e} + \nu\operatorname{curl}\mathbf{a}\cdot\operatorname{curl}\partial_t\mathbf{a}\; dx \\ &= -\int_\Omega \sigma|\partial_t\mathbf{a}|^2\; dx + \int_\Omega \mathbf{j}_s\cdot\partial_t\mathbf{a}\; dx - \int_{\partial\Omega}(\nu\operatorname{curl}\mathbf{a}\times\partial_t\mathbf{a})\cdot\mathbf{n}\; ds(x),\end{aligned}$$

where the three summands again correspond to the power dissipated through eddy current losses, the power supplied by the source, and the power in or outflow through the boundary. Integrating in time yields the corresponding energy balance for this formulation.

From the analytical point of view, both formulations (1.3)–(1.4) and (1.5)–(1.6) are hyperbolic in nature. The existence and uniqueness of solutions corresponding to initial and boundary value problems can be shown by standard semi-group theory. [53, 82].

### Numerical methods

A variety of different discretization schemes have been applied to Maxwell's equations. The state-of-the-art approach for transient simulation is the *finite difference time domain* (FDTD) method; see [130, 144]. This method is based on a finite difference approximation in space and an explicit leap-frog scheme in time. It provides second-order accuracy in space and time for smooth isotropic coefficients and orthogonal grids. Generalizations to different time-stepping strategies and nonlinear materials have been addressed in e.g. [19, 77, 135]. It can be shown that the approximation satisfies a discrete energy balance equation for linear and some nonlinear system types; see e.g. [19, 74].

The *finite integration technique* (FIT) developed in [141] has many similarities with the FDTD, and on rectangular grids with leap-frog as a time-stepping scheme, the two approaches coincide [142, 143]. A severe restriction of the FDTD and FIT methods is that their analysis relies on orthogonal Cartesian grids. The generalization to non-rectangular grids has been addressed in [33, 34, 108, 120], however, the second-order accuracy and even stability may be lost, in general.

For complicated geometries, the *finite element methods* (FEM) provides a more flexible approach. It can be used for the construction of arbitrary high-order schemes. Finite element methods and discontinuous Galerkin schemes for Maxwell's equations have been discussed e.g. in [2, 99, 101]. The efficient realization using *mass lumping* techniques has been studied in [45, 108]. The variational structure of FEM approximation is very advantageous from the energy-based perspective. As we will discuss later, a semi-discrete power balance can be directly derived for Galerkin approximation in space with the same arguments as on the continuous level. For time stepping, well-established leap-frog and higher order Runge-Kutta schemes have been discussed in the literature [2, 8, 16, 18, 70]. While for linear problems the discrete energy balance can be derived for high-order Runge-Kutta schemes, for nonlinear problems this is not possible in general. Some particular results can be found in [2, 18].

### Outline and main contributions

In this chapter, we discuss extensions of the formulations above to nonlinear materials and their numerical treatment. The essential point of our discussion is an appropriate description of energy densities and energy-based modelling of material relations. Then, nonlinear versions of the Poynting theorem and resulting energy balances can be constructed in a similar manner. To preserve the underlying energy balance in the simulations, *structure-preserving* discretization strategies are proposed. As mentioned in the introduction, the strategies utilize the particular structures of the formulations and employ variational methods in space and time [42, 43]. We consider mixed finite element approaches for discretization in space and discontinuous Galerkin and Petrov-Galerkin schemes in time. Using the proposed strategies, systematic construction of arbitrary high-order schemes based on implicit time stepping is possible. Under relatively general restrictions on the materials, the discrete energy balances can then be derived.

Although the presented formalism can be applied to a relatively general class of problems in electromagnetism, we focus on applications motivated by nonlinear optics. We discuss the propagation of high-intensity fields through nonlinear dielectric media of Kerr type,

which has been discussed in many publications; see e.g. [2, 16, 74]. The generalization to electrically conducting materials is straightforward. The nonlinear magnetic problems can be handled by analogy. After that, we discuss the extension of the approaches to materials with dispersion. The general treatment of memory-dependent materials is not yet settled, therefore we only consider a particular Kerr-Lorentz model [19, 103, 126]. We use the auxiliary representation of the dispersion term, which is essential for the analysis. The medium with Debye dispersion can be handled similarly. The contents of this chapter are based on our publications [39, 43, 48].

## 1.1. Constitutive relations

A correct representation of the constitutive relations is an essential ingredient in describing power and energy balances. In the following, we will introduce the constitutive relations and connect them to the energies. Here we discuss only instantaneous material responses. The generalization to memory-dependent materials, which is also a topic our interest, is still subject to ongoing work.

#### Instantaneous electric and magnetic responses

We take an energy-based perspective to modelling the constitutive relations [104, 125] and assume that there exist electric and magnetic energy densities

$$w_{el} : \mathbf{d} \mapsto w_{el}(\mathbf{d}) \in \mathbb{R} \qquad \text{and} \qquad w_{mag} : \mathbf{b} \mapsto w_{mag}(\mathbf{b}) \in \mathbb{R},$$

which are assumed to be sufficiently smooth functions of fluxes $\mathbf{d}$ and $\mathbf{b}$, respectively. Then, the constitutive relations between the electric and magnetic fields and fluxes can be defined using the energy densities by

$$\mathbf{e} = w'_{el}(\mathbf{d}) \qquad \text{and} \qquad \mathbf{h} = w'_{mag}(\mathbf{b}), \tag{1.7}$$

which describe the *instantaneous electric* and *magnetic* responses of the medium. Here $\omega'(\cdot)$ is the gradient of a scalar multivariable function $\omega(\cdot)$. The expressions $w_{el}(\mathbf{d}) = \int_0^{\mathbf{d}} \mathbf{e}\ d\mathbf{d}$ and $w_{mag}(\mathbf{b}) = \int_0^{\mathbf{b}} \mathbf{h}\ d\mathbf{b}$ for the electric and magnetic energy densities can be found in classical textbooks [73, 129] on electromagnetism. The constitutive relations defined by (1.7) justify the well-posedness of these expressions.

#### Inverse relations

Assuming $w_{el}$ and $w_{mag}$ are smooth, strongly convex, and coercive functionals, the inverse relations for the constitutive laws (1.7) are then given by

$$\mathbf{d} = w'_{*,el}(\mathbf{e}) \quad \text{and} \quad \mathbf{b} = w'_{*,mag}(\mathbf{h}) \tag{1.8}$$

where $w_{*,el}$ and $w_{*,mag}$ are convex conjugate functions, also known as *co-energy densities*; see [23, Section 2] and [22, 112]. From this standpoint, the variables $\mathbf{d}$ and $\mathbf{b}$ are also called *energy variables*, while $\mathbf{e}$ and $\mathbf{h}$ are *co-energy variables*. Taking the time derivative of the inverse relations (1.8) yields

$$\partial_t \mathbf{d} = \epsilon(\mathbf{e})\partial_t \mathbf{e} \quad \text{and} \quad \partial_t \mathbf{b} = \mu(\mathbf{h})\partial_t \mathbf{h},$$

where $\epsilon(\mathbf{e}) = w''_{*,el}(\mathbf{e})$ and $\mu(\mathbf{h}) = w''_{*,mag}(\mathbf{h})$ are the *differential permittivity* and *permeability*, respectively.

Using the relation (1.8), the energy densities $w_{el}$ and $w_{mag}$ can be expressed as functions of the co-energy variables $\mathbf{e}$ and $\mathbf{h}$. In the following, we denote the electric and magnetic energy densities as functions of the fields $\mathbf{e}$ and $\mathbf{h}$ with

$$\tilde{w}_{el}(\mathbf{e}) = w_{el}(w'_{*,el}(\mathbf{e})) \qquad \text{and} \qquad \tilde{w}_{mag}(\mathbf{h}) = w_{mag}(w'_{*,mag}(\mathbf{h})), \tag{1.9}$$

respectively. We further conclude that

$$\tilde{w}'_{el}(\mathbf{e}) = \epsilon(\mathbf{e})\mathbf{e} \qquad \text{and} \qquad \tilde{w}'_{mag}(\mathbf{h}) = \mu(\mathbf{h})\mathbf{h}. \tag{1.10}$$

These relations (1.9) and (1.10) provide a connection between the energy densities and incremental permittivity and permeability, and play a fundamental role in the derivation of power balances.

**Example 1.1.1** (linear medium)**.** In the linear case, the energy densities are quadratic functions given by $w_{el}(\mathbf{d}) = \frac{\epsilon^{-1}}{2}|\mathbf{d}|^2$ and $w_{mag}(\mathbf{b}) = \frac{\mu^{-1}}{2}|\mathbf{b}|^2$ with constant permittivity $\epsilon$ and permeability $\mu$. Differentiating these energy densities leads to the well-known constitutive relations $\mathbf{e} = \epsilon^{-1}\mathbf{d}$ and $\mathbf{h} = \mu^{-1}\mathbf{b}$. The corresponding co-energy densities are given by $w_{*,el} = \frac{\epsilon}{2}|\mathbf{e}|^2$ and $w_{*,mag} = \frac{\mu}{2}|\mathbf{h}|^2$, which leads to inverse relations $\mathbf{d} = \epsilon\mathbf{e}$ and $\mathbf{b} = \mu\mathbf{h}$, and $w''_{*,el}(\mathbf{e}) = \epsilon I$ and $w''_{*,mag}(\mathbf{h}) = \mu I$, where $I$ is the identity.

**Example 1.1.2** (Kerr-type medium)**.** As the second example, we consider the nonlinear dielectric response of a Kerr medium, which represents the main focus of this chapter. The constitutive relation for Kerr media is often written in the form

$$\mathbf{d} = \epsilon_0(\tilde{\chi}^{(1)} + \chi^{(3)}|\mathbf{e}|^2)\mathbf{e}, \tag{1.11}$$

with $\tilde{\chi}^{(1)} = 1 + \chi^{(1)}$, where $\chi^{(1)}$ and $\chi^{(3)}$ are the susceptibility components [2, 24]. This expression corresponds to the inverse constitutive relation $\mathbf{d} = w'_{*,el}(\mathbf{e})$, where the co-energy density is now given by the relation $w_{*,el}(\mathbf{e}) = \frac{\epsilon_0}{2}(\tilde{\chi}^{(1)}|\mathbf{e}|^2 + \frac{\chi^{(3)}}{2}|\mathbf{e}|^4)$. Since $w_{*,el}$ is strongly convex and coercive, there exists an inverse relation $\mathbf{e} = w'_{el}(\mathbf{d})$, where $w_{el}(\mathbf{d})$ is the convex conjugate to $w_{*,el}(\mathbf{e})$ and is given by $w_{el}(\mathbf{d}) = \mathbf{d}\cdot\mathbf{e} - w_{*,el}(\mathbf{e})$. Substituting the relation $\mathbf{d} = w'_{*,el}(\mathbf{e})$ brings us to the expression for electric energy density in terms of the co-variable $\mathbf{e}$, namely

$$\tilde{\omega}_{el}(\mathbf{e}) = \frac{\epsilon_0}{2}(\tilde{\chi}^{(1)}|\mathbf{e}|^2 + \frac{3\chi^{(3)}}{2}|\mathbf{e}|^4). \tag{1.12}$$

Hence, the constitutive relations (1.11) and expression for energy density (1.12) are consistent with the energy-based modeling. The differential permittivity is then given by

$$\epsilon(\mathbf{e}) = \epsilon_0(\tilde{\chi}^{(1)} + \chi^{(3)}|\mathbf{e}|^2 + 2\chi^{(3)}\mathbf{e}\mathbf{e}^\top), \tag{1.13}$$

accordingly. These relations play a fundamental role in our analysis below.

#### Ohm's law

For completeness, we also discuss the electric conductivity of the media. The electric field within an electrically conducting material induces an electric current. We assume that the current consists of the source current $\mathbf{j}_s$ and the current $\mathbf{j}_c$ induced by the electric field $\mathbf{e}$, given by *Ohm's law*, namely

$$\mathbf{j} = \mathbf{j}_c + \mathbf{j}_s \qquad \text{with} \qquad \mathbf{j}_c = \sigma(\mathbf{e})\mathbf{e},$$

where $\sigma(\mathbf{e})$ is the *nonlinear conductivity* of the material, which we assume to be symmetric and positive semi-definite. Let us note that the current flow within a conductor generates heat. The amount of heating power is given by $P_{Joule} = \int_\Omega \sigma(\mathbf{e})\mathbf{e} \cdot \mathbf{e}\ dx$. This effect is known as *Joule's heating*, *resistive heating*, or *Ohmic heating*, because of its relation to Ohm's law. The loss of energy to heating is often called *Joule loss* or *Ohmic loss*.

## 1.2. Maxwell's equations in nonlinear dielectric media

In this section, we discuss the propagation of high-frequency electromagnetic fields in nonlinear dielectric media, which is of relevance in nonlinear optics. The governing relations are given by Maxwell's equations

$$\partial_t \mathbf{d} = \operatorname{curl} \mathbf{h} \qquad \text{and} \qquad \partial_t \mathbf{b} = -\operatorname{curl} \mathbf{e}. \tag{1.14}$$

In high-frequency applications, the nonlinearity of a medium response is significant for electric quantities, while the magnetic response can be assumed to be linear. As previously mentioned, we restrict our considerations to a nonlinear electric media of Kerr type, which is the classical example of nonlinear optic media; see e.g. [2, 24, 74]. In the following, we assume the constitutive relations

$$\mathbf{d} = \epsilon_0(\tilde{\chi}^{(1)} + \chi^{(3)}|\mathbf{e}|^2)\mathbf{e} \qquad \text{and} \qquad \mathbf{b} = \mu_0 \mathbf{h}, \tag{1.15}$$

where $\mu_0$ is the permittivity of vacuum and $\tilde{\chi}^{(1)} = 1 + \chi^{(1)}$, where $\chi^{(1)}$ and $\chi^{(3)}$ are electric susceptibility components. Generalization to different types of instantaneous material responses, as well as electrically conducting materials is also possible.

We now present two formulations based on the fields $\mathbf{e}$ and $\mathbf{h}$ and $\mathbf{e}$ and $\mathbf{a}$, respectively. We derive the balances of power and energy and discuss the construction of higher-order schemes, utilizing the frameworks [42, 43] discussed in Section A, which preserve these balances. Additionally, we provide some details on the numerical realization and illustrate the theoretical results using numerical examples.

### 1.2.1. The $\mathbf{e} - \mathbf{h}$ formulation

We start with a formulation in terms of the fields $\mathbf{e}$ and $\mathbf{h}$. Substituting the constitutive relations (1.15) into Faraday's and Amprere's laws (1.14) results in the system

$$\epsilon(\mathbf{e})\partial_t \mathbf{e} = \operatorname{curl} \mathbf{h}, \tag{1.16}$$

$$\mu_0 \partial_t \mathbf{h} = -\operatorname{curl} \mathbf{e}, \tag{1.17}$$

where the differential permittivity $\epsilon(\mathbf{e})$ is given by (1.13). In the scope of this section, we restrict our discussion to a bounded Lipschitz domain denoted by $\Omega$, and impose the perfect magnetic boundary condition

$$\mathbf{n} \times \mathbf{h} = 0 \qquad \text{on } \partial\Omega. \tag{1.18}$$

The analysis below can be easily extended to other types of boundary conditions with minor modifications. Let us further note that conducting media, i.e. $\sigma(\mathbf{e}) \neq 0$, can be handled by analogy.

### Electromagnetic energy

As discussed in Examples 1.1.2 and 1.1.1, the choice of incremental permittivity and permeability corresponds to expressions for electric and magnetic energy densities

$$\tilde{\omega}_{el}(\mathbf{e}) = \frac{\epsilon_0}{2}(\tilde{\chi}^{(1)}|\mathbf{e}|^2 + \frac{3\chi^{(3)}}{2}|\mathbf{e}|^4) \qquad \text{and} \qquad \tilde{\omega}_{mag}(\mathbf{h}) = \frac{\mu_0}{2}|\mathbf{h}|^2,$$

which are here written in terms of the co-energy variables $\mathbf{e}$ and $\mathbf{h}$. Furthermore, we recall the identities

$$\tilde{\omega}'_{el}(\mathbf{e}) = \epsilon(\mathbf{e})\mathbf{e} \qquad \text{and} \qquad \tilde{\omega}'_{mag}(\mathbf{h}) = \mu_0\mathbf{h}, \tag{1.19}$$

which play an important role in the upcoming discussion. With

$$\mathcal{E}(\mathbf{e}, \mathbf{h}) = \int_\Omega \tilde{w}_{el}(\mathbf{e}) + \tilde{w}_{mag}(\mathbf{h})\, dx, \tag{1.20}$$

we denote the corresponding expression for energy in terms of fields $\mathbf{e}$ and $\mathbf{h}$.

### Conservation of energy

Let us first introduce some basic notation. We denote by $L^2(\Omega)$ the space of square-integrable functions and use the abbreviations $\langle \mathbf{v}, \mathbf{w} \rangle = \int_\Omega \mathbf{v} \cdot \mathbf{w}\ dx$ and $\|\mathbf{v}\| = \langle \mathbf{v}, \mathbf{v} \rangle$ for the $L^2(\Omega)$ scalar product and the norm. By $H(\operatorname{curl}, \Omega) = \{\mathbf{w} \in L^2(\Omega) : \operatorname{curl} \mathbf{w} \in L^2(\Omega)\}$ we denotes the subspace of $L^2(\Omega)$ functions with square-integrable curls.

The following lemma summarizes an important variational characterization of the solution, which is the foundation of the variational approach.

**Lemma 1.2.1.** Let $(\mathbf{e}, \mathbf{h})$ be a sufficiently smooth solution of (1.16)–(1.18). Then

$$\langle \epsilon(\mathbf{e}(t))\partial_t \mathbf{e}(t), \mathbf{v} \rangle = \langle \mathbf{h}(t), \operatorname{curl} \mathbf{v} \rangle, \tag{1.21}$$

$$\langle \mu_0 \partial_t \mathbf{h}(t), \mathbf{w} \rangle = -\langle \operatorname{curl} \mathbf{e}(t), \mathbf{w} \rangle, \tag{1.22}$$

for all test functions $\mathbf{v} \in H(\operatorname{curl}, \Omega)$ and $\mathbf{w} \in L^2(\Omega)$ and all $t \geq 0$.

*Proof.* The variational identity (1.22) follows directly by multiplying (1.17) with a test function $\mathbf{w} \in L^2(\Omega)$ and integrating over the domain $\Omega$. For the first identity (1.21), we multiply (1.16) with a test function $\mathbf{v} \in H(\operatorname{curl}, \Omega)$, integrate over the domain $\Omega$, and use integration by parts formula

$$\langle \operatorname{curl} \mathbf{h}, \mathbf{v} \rangle = \langle \mathbf{h}, \operatorname{curl} \mathbf{v} \rangle + \int_{\partial\Omega} \mathbf{n} \times \mathbf{h} \cdot \mathbf{v}\ ds,$$

where the last term vanishes due to the choice of boundary condition (1.18). □

Using this variational characterization of solutions, one can immediately derive a power balance and thus show that the energy of the system is conserved over time.

**Lemma 1.2.2.** Let the energy be given as in (1.20). Then, any smooth solution $(\mathbf{e}, \mathbf{h})$ of the system (1.21)–(1.22) satisfies the power balance

$$\frac{d}{dt}\mathcal{E}(\mathbf{e}(t), \mathbf{h}(t)) = 0, \qquad t > 0.$$

In particular, the system is passive.

*Proof.* Formal differentiation of the energy with respect to time yields

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(\mathbf{e}(t), \mathbf{h}(t)) &= \langle \partial_t \mathbf{e}(t), \tilde{w}'_{el}(\mathbf{e}(t))\rangle + \langle \partial_t \mathbf{h}(t), \tilde{w}'_{mag}(\mathbf{h}(t))\rangle \\ &= \langle \partial_t \mathbf{e}(t), \epsilon(\mathbf{e}(t))\mathbf{e}(t)\rangle + \langle \partial_t \mathbf{h}(t), \mu_0 \mathbf{h}(t)\rangle \\ &= \langle \epsilon(\mathbf{e}(t))\partial_t \mathbf{e}(t), \mathbf{e}(t)\rangle + \langle \mu_0 \partial_t \mathbf{h}(t), \mathbf{h}(t)\rangle = (*).\end{aligned}$$

Here we used the energy relations (1.19). Next, we use the results of Lemma 1.2.1 with $\mathbf{v} = \mathbf{e}(t)$ and $\mathbf{w} = \mathbf{h}(t)$, which are admissible test functions, and obtain

$$(*) = \langle \mathbf{h}(t), \operatorname{curl} \mathbf{e}(t)\rangle - \langle \operatorname{curl} \mathbf{e}(t), \mathbf{h}(t)\rangle = 0.$$

This already yields the desired result. □

#### Structure of the $\mathbf{e} - \mathbf{h}$ formulation

The power balance is a direct consequence of the variational principle (1.21)–(1.22) and the choice $\mathbf{v} = \mathbf{e}(t)$ and $\mathbf{w} = \mathbf{h}(t)$ for the test functions. This principle holds because of the particular *port-Hamiltonian* structure of the problem, namely

$$\mathcal{Q}^*(\mathbf{u})\partial_t \mathbf{u} = -\mathcal{A}(\mathbf{u}), \tag{1.23}$$

$$\mathcal{E}'(\mathbf{u}) = Q(\mathbf{u})\mathbf{u}, \tag{1.24}$$

where $\mathbf{u} = (\mathbf{e}, \mathbf{h})$, the operator $\mathcal{A}(\mathbf{u})$ is defined in the weak sense by $\langle \mathcal{A}(\mathbf{e}, \mathbf{h}), (\mathbf{v}, \mathbf{w})\rangle = \langle \mathbf{h}, \operatorname{curl} \mathbf{v}\rangle - \langle \operatorname{curl} \mathbf{e}, \mathbf{w}\rangle$, the operator $Q(\mathbf{u})$ is given by $Q(\mathbf{e}, \mathbf{h}) = Q(\mathbf{e}, \mathbf{h})^* = \operatorname{diag}(\epsilon(\mathbf{e}), \mu_0)$, and the derivative of the energy is given by $\langle \mathcal{E}'(\mathbf{e}, \mathbf{h}), (\mathbf{v}, \mathbf{w})\rangle = \langle \epsilon(\mathbf{e})\mathbf{e}, \mathbf{v}\rangle + \langle \mu_0 \mathbf{h}, \mathbf{w}\rangle$. The power balance can be alternatively derived at the abstract level

$$\frac{d}{dt}\mathcal{E}(\mathbf{u}(t)) = \langle \partial_t \mathbf{u}, \mathcal{E}'(\mathbf{u})\rangle = \langle \partial_t \mathbf{u}, Q(\mathbf{u})\mathbf{u}\rangle = \langle Q^*(\mathbf{u})\partial_t \mathbf{u}, u\rangle = -\langle \mathcal{A}(\mathbf{u})\mathbf{u}, \mathbf{u}\rangle.$$

The variational discretization strategy for a problem that preserves the energy evolution principle has been proposed in [42]. Some important details are summarized in Appendix A.1. As we will see in Chapter 2, problems of the same structure also arise in electric circuits modelling. We now adopt the strategy and discuss the discretization of the considered problem.

### 1.2.2. Discretization of the $\mathbf{e}-\mathbf{h}$ formulation

Let $W_h \subset H(\mathrm{curl},\Omega)$ and $Q_h \subset L^2(\Omega)$ denote some finite-dimensional subspaces and let $\mathcal{T} = \{t^n : 0 \le n \le N\}$ be a sequence of discrete time steps $t^n = n\tau$ with $\tau = T/N$. With $I^n = [t^{n-1}, t^n]$ we denote the $n$-th time interval and with $P_k(I^n;\mathbb{V})$ we denote the space of polynomials with values in $\mathbb{V}$. By $P_k(\mathcal{T};\mathbb{V})$ we denote the space of piece-wise polynomials, i.e., the functions whose restrictions to any interval $I^n$ lie in $P_k(I^n;\mathbb{V})$. We further use $(*)|_{t^n}$ to abbreviate the evaluation of $(*)$ at time $t = t^n$. We consider the approximation of the problem (1.16)–(1.18) by the following method.

**Problem 1.2.3.** Let initial values $\mathbf{e}_h^0 \in W_h$ and $\mathbf{h}_h^0 \in Q_h$ be given. Then, for $1 \le n \le N$, we seek for $\mathbf{e}_h^n \in P_k(I^n;W_h)$ and $\mathbf{h}_h^n \in P_k(I^n;Q_h)$ such that

$$\int_{I^n} \langle \epsilon(\mathbf{e}_h^n)\partial_t \mathbf{e}_h^n, \mathbf{v}_h\rangle - \langle \mathbf{h}_h^n, \mathrm{curl}\,\mathbf{v}_h\rangle\, dt = \langle \epsilon(\mathbf{e}_h^n)(\mathbf{e}_h^{n-1} - \mathbf{e}_h^n), \mathbf{v}_h\rangle|_{t^{n-1}} \tag{1.25}$$

$$\int_{I^n} \langle \mu_0 \partial_t \mathbf{h}_h^n, \mathbf{w}_h\rangle + \langle \mathrm{curl}\,\mathbf{e}_h^n, \mathbf{w}_h\rangle\, dt = \langle \mu_0(\mathbf{h}_h^{n-1} - \mathbf{h}_h^n), \mathbf{w}_h\rangle|_{t^{n-1}} \tag{1.26}$$

holds for all test functions $\mathbf{v}_h \in P_k(I^n;W_h)$ and $\mathbf{w}_h \in P_k(I^n;Q_h)$.

This scheme is based on a Galerkin approximation in space and a discontinuous method in time. The energy balance provided by this scheme is summarized in the following result.

**Lemma 1.2.4.** Let $(\mathbf{e}_h^n, \mathbf{h}_h^n)_{n\ge 0}$ denote a solution of Problem 1.2.3. Then

$$\mathcal{E}(\mathbf{e}_h^n(t^n), \mathbf{h}_h^n(t^n)) \le \mathcal{E}(\mathbf{e}_h^m(t^m), \mathbf{h}_h^m(t^m)) \tag{1.27}$$

holds for all time steps $0 \le m \le n \le N$.

*Proof.* The following proof is a special case of [42, Theorem 4] which for convenience is presented in Appendix A.1. We here derive the result for the particular problem under investigation. We first consider the case $m = n-1$. The change of energy between two consecutive time steps can be decomposed as follows

$$\begin{aligned}
&\mathcal{E}(\mathbf{e}_h^n(t^n), \mathbf{h}_h^n(t^n)) - \mathcal{E}(\mathbf{e}_h^{n-1}(t^{n-1}), \mathbf{h}_h^{n-1}(t^{n-1})) \\
&\quad = (\mathcal{E}(\mathbf{e}_h^n(t^n), \mathbf{h}_h^n(t^n)) - \mathcal{E}(\mathbf{e}_h^n(t^{n-1}), \mathbf{h}_h^n(t^{n-1}))) \\
&\qquad\qquad + (\mathcal{E}(\mathbf{e}_h^n(t^{n-1}), \mathbf{h}_h^n(t^{n-1})) - \mathcal{E}(\mathbf{e}_h^{n-1}(t^{n-1}), \mathbf{h}_h^{n-1}(t^{n-1}))) \\
&\quad = (i) + (ii).
\end{aligned}$$

Using the fundamental theorem of calculus, we obtain

$$\begin{aligned}
(i) &= \int_{I^n} \frac{d}{dt}\mathcal{E}(\mathbf{e}_h^n(t), \mathbf{h}_h^n(t))\, dt \\
&= \int_{I^n} \langle \partial_t \mathbf{e}_h^n(t), \tilde{\omega}'_{el}(\mathbf{e}_h^n(t))\rangle + \langle \partial_t \mathbf{h}_h^n(t), \tilde{\omega}'_{mag}(\mathbf{h}_h^n(t))\rangle\, dt \\
&= \int_{I^n} \langle \partial_t \mathbf{e}_h^n(t), \epsilon(\mathbf{e}_h^n(t))\mathbf{e}_h^n(t)\rangle + \langle \partial_t \mathbf{h}_h^n(t), \mu_0 \mathbf{h}_h^n(t)\rangle\, dt \\
&= \int_{I^n} \langle \epsilon(\mathbf{e}_h^n(t))\partial_t \mathbf{e}_h^n(t), \mathbf{e}_h^n(t)\rangle + \langle \mu_0 \partial_t \mathbf{h}_h^n(t), \mathbf{h}_h^n(t)\rangle\, dt.
\end{aligned} \tag{1.28}$$

Here we used the relations for energy densities $\tilde{w}'_{el}(\mathbf{e}) = \epsilon(\mathbf{e})\mathbf{e}$ and $\tilde{w}'_{mag}(\mathbf{h}) = \mu_0\mathbf{h}$, which hold for any $\mathbf{e}$ and $\mathbf{h}$. By the variational scheme (1.25)-(1.26) with $\mathbf{v}_h = \mathbf{e}_h^n$ and $\mathbf{w}_h = \mathbf{h}_h^n$, which are admissible test functions, we further conclude

$$\begin{aligned}(i) &= \langle \epsilon(\mathbf{e}_h^n)(\mathbf{e}_h^{n-1} - \mathbf{e}_h^n), \mathbf{e}_h^n\rangle|_{t^{n-1}} + \langle \mu_0(\mathbf{h}_h^{n-1} - \mathbf{h}_h^n), \mathbf{h}_h^n\rangle|_{t^{n-1}} \\ &= -\langle \epsilon(\mathbf{e}_h^n)\mathbf{e}_h^n, (\mathbf{e}_h^n - \mathbf{e}_h^{n-1})\rangle|_{t^{n-1}} - \langle \mu_0\mathbf{h}_h^n, (\mathbf{h}_h^n - \mathbf{h}_h^{n-1})\rangle|_{t^{n-1}}.\end{aligned}$$

We further use the convexity of the energy density to obtain

$$\begin{aligned}(ii) &= \mathcal{E}(\mathbf{e}_h^n(t^{n-1}), \mathbf{h}_h^n(t^{n-1})) - \mathcal{E}(\mathbf{e}_h^{n-1}(t^{n-1}), \mathbf{h}_h^{n-1}(t^{n-1})) \\ &\leq \langle \epsilon(\mathbf{e}_h^n)\mathbf{e}_h^n, \mathbf{e}_h^n - \mathbf{e}_h^{n-1}\rangle|_{t^{n-1}} + \langle \mu_0\mathbf{h}_h^n, \mathbf{h}_h^n - \mathbf{h}_h^{n-1}\rangle|_{t^{n-1}}.\end{aligned}$$

Adding $(i)$ and $(ii)$ together, we directly conclude that $(i) + (ii) \leq 0$, which proves the assertion for $m = n - 1$. The general case $m < n$ follows directly by induction. □

**Remark 1.2.5.** We can only guarantee that the energy does not increase over time, while the analytic problem is energy conserving. The dissipative nature of discontinuous Galerkin and the related implicit Euler and RadauIIA schemes are well known; see e.g. [49]. However, for smooth bounded solutions, we expect the dissipation to become small for sufficiently accurate discretizations, particularly when high-order time-stepping schemes are employed.

**Remark 1.2.6.** The scalar product $\langle \cdot, \cdot \rangle$ can be replaced by $\langle \cdot, \cdot \rangle_h$ resulting from inexact integration using a quadrature rule. Then, the energy balance (1.27) can also be shown for the perturbed energy $\mathcal{E}_h(\cdot, \cdot)$, which is computed using the same quadrature rule. However, the exact evaluation of the time integrals $\int_{I^n} \langle \epsilon(\mathbf{e}_h^n)\partial_t \mathbf{e}_h^n, \mathbf{v}_h\rangle \, dt$ and $\int_{I^n} \langle \mu_0 \partial_t \mathbf{h}_h^n, \mathbf{w}_h\rangle \, dt$ is crucial for the proof of Lemma 1.2.4. Otherwise, the relation (1.28) would not hold in general. The other integral terms can be approximated using a fixed quadrature rule, without losing the energy-conserving property of the discretization.

The presented discretization strategy (1.25)–(1.26) represents a relatively general and flexible approach. Let us make some remarks on possible numerical realizations.

#### Finite element discretization in space

We use mixed finite element methods for the discretization in space. The semi-discretization of problem (1.21)–(1.22) using an appropriate method leads to a system of the form

$$\mathsf{M_e}(\mathsf{e}(t))\partial_t \mathsf{e}(t) = \mathsf{C}\mathsf{h}(t), \tag{1.29}$$

$$\mathsf{M_h}\partial_t \mathsf{h}(t) = -\mathsf{C}^\top \mathsf{e}(t), \tag{1.30}$$

where $\mathsf{e}(t)$ and $\mathsf{h}(t)$ are coefficient vectors, and $\mathsf{C}$, $\mathsf{M_h}$, and $\mathsf{M_e}(\mathsf{e}(t))$ are the system matrices. In our numerical tests, we consider simplified problems in one and two spatial dimensions. For completeness, we now briefly provide an example of a possible realization in 3D.

**Example 1.2.7.** Let $\mathcal{T}_h = \{K\}$ denote a decomposition of the domain $\Omega$ into tetrahedra. We denote by $P_p(K)$ the space of polynomials of degree at most $p$ and by $P_p^h(K)$ we denote

the space of homogeneous polynomials of degree exactly $p$ over the element $K$. Then a suitable pair of discrete subspaces is given by

$$W_h = \{\mathbf{w} \in H(\mathrm{curl}, \Omega) \,:\, \mathbf{w}|_K \in \mathcal{N}_p(K) \text{ for all } K \in \mathcal{T}_h\},$$
$$Q_h = \{\mathbf{q} \in L^2(\Omega) \,:\, \mathbf{q}|_K \in P_p(K)^3 \text{ for all } K \in \mathcal{T}_h\},$$

where $\mathcal{N}_p(K) = P_p(K)^3 \oplus \{\mathbf{x} \times P_p^h(K)^3\}$ denotes the Nedelec space [99, 102]. In the lowest order case a polynomial $\mathbf{w} \in \mathcal{N}_0(K)$ has the form $\mathbf{w}(\mathbf{x}) = \alpha + \mathbf{x} \times \beta$ with constant vectors $\alpha$ and $\beta$. Now let $\{\phi_k\}_k \subset W_h$ and $\{\psi_k\}_k \subset Q_h$ denote sets of basis functions for the finite dimensional subspaces $W_h$ and $Q_h$. The semi-discrete solutions $\mathbf{e}_h(t) \in W_h$ and $\mathbf{h}_h(t) \in Q_h$ can be expanded as $\mathbf{e}_h(t) = \sum_k \mathsf{e}_k(t)\phi_k$ and $\mathbf{h}_h(t) = \sum_k \mathsf{h}_k(t)\psi_k$ where $\{\mathsf{e}_k\}_k$ and $\{\mathsf{h}_k\}_k$ denote the corresponding coefficients. Then, the finite element discretization leads to a system of differential equations (1.29)-(1.30) with

$$(\mathsf{M_e(e)})_{ij} = \langle \epsilon(\sum_l \mathsf{e}_l \phi_l)\phi_j, \phi_i \rangle, \quad (\mathsf{M_h})_{ij} = \langle \mu_0 \psi_j, \psi_i \rangle, \quad \text{and} \quad (\mathsf{C})_{ij} = \langle \mathrm{curl}\, \phi_i, \psi_j \rangle.$$

Since the nonlinear terms consist of polynomial functions, the integration can be carried out by an appropriate quadrature rule, possibly inexact; see Remark 1.2.6. In particular, mass lumping techniques can be considered; see e.g. [44, 108]. For more details on Nedelec spaces and the construction of basis functions, we refer to [17, 100].

### Discontinuous Galerkin discretization in time

Next, we discuss the construction of discontinuous Galerkin time-stepping schemes. Since we consider several problems of the canonical structure (1.23)–(1.24), it is convenient to have a unified implementation for the problems of this type. For this reason, we write the problem (1.29)–(1.30) in a abstract form

$$\mathsf{Q(u}(t))^\top \partial_t \mathsf{u}(t) + \mathsf{A(u}(t)) = 0, \tag{1.31}$$

where $\mathsf{u}(t) = (\mathsf{e}(t), \mathsf{h}(t))$, $\mathsf{A(u)} = \mathsf{Ju}$, and matrices $\mathsf{Q(u}(t))^\top$ and $\mathsf{J}$ are given by

$$\mathsf{Q(u}(t))^\top = \begin{pmatrix} \mathsf{M_e(e}(t)) & 0 \\ 0 & \mathsf{M_h} \end{pmatrix} \quad \text{and} \quad \mathsf{J} = \begin{pmatrix} 0 & -\mathsf{C} \\ \mathsf{C}^\top & 0 \end{pmatrix}.$$

The scheme (1.25)–(1.26) can be equivalently written as follows: For a given $\mathsf{u}^0 \in \mathbb{R}^d$ we seek $\mathsf{u}^n \in P_k(I; \mathbb{R}^d)$, $1 \le n \le N$, such that

$$\int_{I^n} \left[ \mathsf{Q(u}^n(t))^\top \partial_t \mathsf{u}^n(t) + \mathsf{A(u}^n(t)) \right] v(t)\, dt = (\mathsf{Q(u}^n)^\top (\mathsf{u}^{n-1} - \mathsf{u}^n) v)|_{t^0} \tag{1.32}$$

for all $v \in P_k(I^n; \mathbb{R})$. Note that we consider scalar-valued test functions, which is sufficient.

**Example 1.2.8.** Consider the scheme for $k = 0$. For $u \in P_0(\mathcal{T}; \mathbb{R}^d)$, let $u^n \in \mathbb{R}^d$ denote the constant value of $u$ on the interval $I^n$. Then, the numerical scheme (1.32) becomes

$$\mathsf{Q(u}^n)^\top \frac{\mathsf{u}^n - \mathsf{u}^{n-1}}{\tau^n} + \mathsf{A(u}^n) = 0.$$

Therefore, the lowest order scheme corresponds to the implicit Euler method.

**Remarks on implementation**

For the implementation of higher-order schemes, we take a straightforward approach. For simplicity, we now consider only a single time step, i.e. $I = [t^0, t^1]$. We denote the time step by $\tau = t^1 - t^0$ and use the abbreviation $u = u^1 \in P_k(I; \mathbb{R}^d)$. Let $\phi_j(t) \in P_k(I; \mathbb{R})$ denote the basis of the trial space. We now make the expansion assumptions

$$\mathsf{u}(t) = \sum_{j=0}^{k} \mathsf{u}_j \phi_j(t) \quad \text{and} \quad \partial_t \mathsf{u}(t) = \sum_{j=0}^{k} \mathsf{u}_j \partial_t \phi_j(t) \qquad \forall t \in I,$$

where $\mathsf{u}_j \in \mathbb{R}^d$ are vector-valued coefficients. For the evaluation of the integrals, we utilize a numerical quadrature. Now let $(w_\ell, \xi_\ell)$ denote the weights and points of an appropriate quadrature rule and let $\hat{\phi}_j$ denote the basis functions on the reference element $\hat{I} = [0, 1]$. We now make the following denotation

$$\gamma_j^\ell = \hat{\phi}_j(\xi_\ell), \quad \alpha_j^\ell = \partial_t \hat{\phi}_j(\xi_\ell), \quad \beta_j = \hat{\phi}_j(1), \quad \text{and} \quad \theta_j = \hat{\phi}_j(0).$$

With the introduced notation, the numerical solution $\mathsf{u}^1 = \mathsf{u}(t^1)$ is the solution of the algebraic problem

$$\sum_{l=0}^{L} w_\ell \Big[ \mathsf{Q}(\mathsf{U}_\ell)^\top \mathsf{U}'_\ell + \tau \mathsf{A}(\mathsf{U}_\ell) \Big] \gamma_i^\ell + \mathsf{Q}(\mathsf{U}^0)^\top (\mathsf{U}^0 - \mathsf{u}^0) = 0, \quad 0 \le i \le k,$$

$$\mathsf{U}_\ell = \sum_{j=0}^{k} \gamma_j^\ell \mathsf{u}_j, \quad \mathsf{U}'_\ell = \sum_{j=0}^{k} \alpha_j^\ell \mathsf{u}_j, \quad \mathsf{U}^0 = \sum_{j=0}^{k} \theta_j \mathsf{u}_j, \quad \mathsf{u}^1 = \sum_{j=0}^{k} \beta_j \mathsf{u}_j.$$

For the solution of the algebraic system, we use Newton's method. The expression for the Jacobian can be constructed analytically by differentiation of individual terms.

An appropriate choice of a quadrature rule is crucial for our purposes. As mentioned in Remark 1.2.6, it is essential that the energy-related integrals, i.e. the first term in (1.32), are evaluated exactly. Since the permittivity $\epsilon(\cdot)$ is quadratic, the chosen quadrature rule must be exact for polynomials of degree $4k - 1$, where $k$ is the polynomial degree of the approximation. For general problems with non-polynomial nonlinearities, the quadrature has to be chosen such that the integration error becomes insignificant. Some particular cases will be discussed below.

Here, we have used the same quadrature rule for the integration of both terms in (1.32). However, this is not necessary. For the evaluation of the second term, a lower-order, possibly inexact quadrature rule can be used. When a $(k + 1)$-node quadrature rule is employed, Lagrange polynomials at the quadrature nodes can be chosen as the basis functions $\hat{\phi}_j$. This leads to $\gamma_i^\ell = \delta_{i,\ell}$, and we obtain

$$\tau \sum_{l=0}^{L} w_\ell \Big[ \mathsf{A}(\mathsf{U}_\ell) \Big] \gamma_i^\ell = \tau w_i \mathsf{A}(\mathsf{u}_i) \mathsf{u}_i.$$

This slightly simplifies the implementation. Since the term $\mathsf{A}(\mathsf{u}) = \mathsf{J}\mathsf{u}$ is linear, it can be integrated exactly using a $(k + 1)$-node Radau quadrature rule. Therefore, the Radau

quadrature is a convenient choice for linear systems. Furthermore, when the right Radau quadrature is employed, the resulting schemes can be shown to be equivalent to RadauIIA collocation methods; see e.g. [4, 93, 134].

In this thesis, we focus solely on nonlinear systems. Therefore, it is necessary to use a higher-order quadrature for integration. In numerical tests, we employ an appropriate higher-order Gauss quadrature. We use Lagrange interpolation polynomials at the Gauss-Lobatto nodes as the basis, which simplifies the implementation of the jump terms slightly; see e.g. [83, 134]. For further details on the construction of different schemes and their relations to collocation and Runge-Kutta methods, we also refer to [71, 95, 133].

One of the drawbacks of this approach is the numerical dissipation, especially for low-order schemes like implicit Euler, where the dissipation can become significant. This is particularly important for energy-conserving systems like the one we consider, as it violates the underlying physics. To address this issue, we propose a different strategy based on the $\mathbf{e}-\mathbf{a}$ formulation. This approach enables the construction of arbitrarily high-order schemes that unconditionally preserve the energy of the system.

### 1.2.3. The $\mathbf{e}-\mathbf{a}$ formulation

The alternative energy-preserving approach is based on the magnetic vector potential $\mathbf{a}$. The following lemma defines the vector potential and proves its important properties.

**Lemma 1.2.9.** Let $(\mathbf{e}, \mathbf{b})$ be smooth functions satisfying Faraday's law $\partial_t \mathbf{b} = -\operatorname{curl} \mathbf{e}$. Further, let the magnetic vector potential be defined by $\mathbf{a}(t) = \mathbf{a}_0 - \int_0^t \mathbf{e}(s)ds$ with $\operatorname{curl} \mathbf{a}_0 = \mathbf{b}(0)$. Then the following relations hold

$$\mathbf{e}(t) = -\partial_t \mathbf{a}(t) \quad \text{and} \quad \mathbf{b}(t) = \operatorname{curl} \mathbf{a}(t), \qquad \forall t \geq 0.$$

*Proof.* The first equality is trivial. Substituting this equality into Faraday's law leads to $\partial_t \mathbf{b}(t) = -\operatorname{curl} \mathbf{e}(t) = \operatorname{curl} \partial_t \mathbf{a}(t)$. Integrating both sides with respect to time yields

$$\mathbf{b}(t) = \mathbf{b}(0) + \int_0^t \partial_t \mathbf{b}(t)\, dt = \operatorname{curl} \mathbf{a}_0 + \int_0^t \operatorname{curl} \partial_t \mathbf{a}(t)\, dt = \operatorname{curl} \mathbf{a}(t),$$

which completes the proof. □

With the relation $\mathbf{b} = \operatorname{curl} \mathbf{a}$, we can write the constitutive equation $\mathbf{b} = \mu_0 \mathbf{h}$ in the form $\mathbf{h} = \nu_0 \operatorname{curl} \mathbf{a}$, where $\nu_0 = \mu_0^{-1}$ denotes the magnetic reluctivity of vacuum. Substituting this relation into Ampere's law $\partial_t \mathbf{d} = \operatorname{curl} \mathbf{h}$ and using the displacement current representation $\partial_t \mathbf{d} = \epsilon(\mathbf{e})\partial_t \mathbf{e}$, we can reformulate the system (1.16)–(1.17) equivalently as

$$-\epsilon(\mathbf{e})\partial_t \mathbf{a} = \epsilon(\mathbf{e})\mathbf{e}, \tag{1.33}$$

$$\epsilon(\mathbf{e})\partial_t \mathbf{e} = \operatorname{curl}(\nu_0 \operatorname{curl} \mathbf{a}). \tag{1.34}$$

Note that the particular choice of the multiplication factor in the first equation is necessary for the energy-based structure of the problem and will become clear below. The perfect magnetic boundary condition (1.18) then translates to

$$\mathbf{n} \times (\nu_0 \operatorname{curl} \mathbf{a}) = 0 \qquad \text{on } \partial\Omega. \tag{1.35}$$

We call (1.33)–(1.35) the $\mathbf{e}$–$\mathbf{a}$ formulation of our problem.

### Electric and magnetic energy densities

We also rewrite the energy densities in terms of the fields $\mathbf{e}$ and $\mathbf{a}$, which are given by

$$\tilde{w}_{el}(\mathbf{e}) = \frac{\epsilon_0}{2}(\tilde{\chi}^{(1)}|\mathbf{e}|^2 + \frac{3\chi^{(3)}}{2}|\mathbf{e}|^4) \quad \text{and} \quad w_{mag}(\operatorname{curl}\mathbf{a}) = \frac{\nu_0}{2}|\operatorname{curl}\mathbf{a}|^2 d\mathbf{x}. \tag{1.36}$$

The expression for the electric energy density is given as before. The expression for the magnetic energy density (1.36) can be obtained by substituting $\mathbf{b} = \operatorname{curl}\mathbf{a}$ in $\omega_{mag}(\mathbf{b})$ and following Example 1.1.1 with $\nu_0 = \mu_0^{-1}$. With

$$\mathcal{H}(\mathbf{e}, \mathbf{a}) = \int_\Omega \tilde{w}_{el}(\mathbf{e}) + w_{mag}(\operatorname{curl}\mathbf{a})\, dx, \tag{1.37}$$

we denote the energy in terms of system variables $\mathbf{e}$ and $\mathbf{a}$.

### Conservation of electromagnetic energy

In the spirit of the previous section, we start with the variational characterization of the solution, which is used to prove the conservation of energy and builds the foundation for discretization in space, which we consider below.

**Lemma 1.2.10.** Let $(\mathbf{e}, \mathbf{a})$ be a sufficiently smooth solution of (1.33)–(1.35). Then

$$-\langle\epsilon(\mathbf{e}(t))\partial_t\mathbf{a}(t), \mathbf{v}\rangle = \langle\epsilon(\mathbf{e}(t))\partial_t\mathbf{e}(t), \mathbf{v}\rangle, \tag{1.38}$$

$$\langle\epsilon(\mathbf{e}(t))\partial_t\mathbf{e}(t), \mathbf{w}\rangle = -\langle\nu_0 \operatorname{curl}\mathbf{a}(t), \operatorname{curl}\mathbf{w}\rangle, \tag{1.39}$$

for all $\mathbf{v}, \mathbf{w} \in H(\operatorname{curl}, \Omega)$ and $t \geq 0$.

*Proof.* The variational identity (1.38) is a direct consequence of multiplying (1.33) by a test function $\mathbf{v} \in H(\operatorname{curl}, \Omega)$ and integrating in time. The second identity follows from multiplying (1.34) by a test function $\mathbf{w} \in H(\operatorname{curl}, \Omega)$, integrating in space, and using the integration by parts formula

$$\langle\operatorname{curl}(\nu_0 \operatorname{curl}\mathbf{a}), \mathbf{w}\rangle = \langle\nu_0 \operatorname{curl}\mathbf{a}, \operatorname{curl}\mathbf{w}\rangle + \int_{\partial\Omega} \mathbf{n} \times (\nu_0 \operatorname{curl}\mathbf{a}) \cdot \mathbf{w}\, ds,$$

where the last term vanishes due to the choice of boundary condition (1.35). □

Using the variational identities (1.38)–(1.39), we now derive the power balance and prove the energy conservation for this formulation.

**Lemma 1.2.11.** Let the energy be given as in (1.37). Then, any smooth solution $(\mathbf{e}, \mathbf{a})$ of the problem (1.33)–(1.35) satisfies the power balance

$$\frac{d}{dt}\mathcal{H}(\mathbf{e}(t), \mathbf{a}(t)) = 0, \qquad t \geq 0.$$

Therefore, the energy of the system is conserved at all times and the system is passive.

*Proof.* By formal differentiation of the energy functional, we obtain

$$\begin{aligned}\frac{d}{dt}\mathcal{H}(\mathbf{e}(t),\mathbf{a}(t)) &= \langle \tilde{w}'_{el}(\mathbf{e}(t)), \partial_t \mathbf{e}(t)\rangle + \langle w'_{mag}(\operatorname{curl}\mathbf{a}(t)), \operatorname{curl}\partial_t\mathbf{a}(t)\rangle \\ &= \langle \epsilon(\mathbf{e}(t))\mathbf{e}(t), \partial_t\mathbf{e}(t)\rangle + \langle \nu_0 \operatorname{curl}\mathbf{a}(t), \operatorname{curl}\partial_t\mathbf{a}(t)\rangle = (*),\end{aligned}$$

where we used $w'_{el}(\mathbf{e}) = \epsilon(\mathbf{e})\mathbf{e}$ and $w'_{mag}(\mathbf{b}) = \nu_0\mathbf{b}$. Next, we use the result of Lemma 1.2.10 with $\mathbf{v} = \partial_t\mathbf{e}(t)$ and $\mathbf{w} = \partial_t\mathbf{a}(t)$, which are admissible test functions, and obtain

$$(*) = -\langle \epsilon(\mathbf{e}(t))\partial_t\mathbf{a}(t), \partial_t\mathbf{e}(t)\rangle + \langle \epsilon(\mathbf{e}(t))\partial_t\mathbf{e}(t), \partial_t\mathbf{a}(t)\rangle = 0.$$

The conservation of energy follows immediately by integration in time. □

#### Structure of the $\mathbf{e}-\mathbf{a}$ formulation

In the derivation of the power balance, we simply used variational identities (1.38)–(1.39) with test functions $\mathbf{v} = \partial_t\mathbf{e}$ and $\mathbf{w} = \partial_t\mathbf{a}$. This is again a direct consequence of the particular structure of the formulation; while the $\mathbf{e}-\mathbf{h}$ formulation has the generalized port-Hamiltonian structure, the $\mathbf{e}-\mathbf{a}$ formulation (1.33)–(1.34) can be written as an abstract *generalized Gradient* system of the form

$$C(\mathbf{u})\partial_t\mathbf{u} = -\mathcal{H}(\mathbf{u}), \tag{1.40}$$

where $\mathbf{u}(t) = (\mathbf{e}(t), \mathbf{h}(t))$, the operator $C$ is given by

$$C(\mathbf{u}) = \begin{pmatrix} 0 & -\epsilon(\mathbf{e}) \\ \epsilon(\mathbf{e}) & 0 \end{pmatrix},$$

and the energy functional is defined by $\langle \mathcal{H}'(\mathbf{e},\mathbf{a}), (\mathbf{v},\mathbf{w})\rangle = \langle \epsilon(\mathbf{e})\mathbf{e}, \mathbf{v}\rangle + \langle \nu_0 \operatorname{curl}\mathbf{a}, \operatorname{curl}\mathbf{w}\rangle$. For problems of this structure, the power balance can then be directly derived as follows.

$$\frac{d}{dt}\mathcal{H}(\mathbf{u}(t)) = \langle \partial_t\mathbf{u}, \mathcal{H}'(\mathbf{u})\rangle = -\langle \partial_t\mathbf{u}, C(\mathbf{u})\partial_t\mathbf{u}\rangle.$$

A variational discretization strategy for problems of this class that preserves the balance of power has been proposed in [43]. Some important details are summarized in Appendix A.2. We now adopt the strategy and discuss the discretization of the considered problem.

Let us further note that at this point, it should become clear why the specific multiplication factor in (1.16) is chosen. This factor is chosen such that the right-hand side of the system (1.16)–(1.17) corresponds to the derivatives of the energy densities.

### 1.2.4. Discretization of the $\mathbf{e}-\mathbf{a}$ formulation

We denote by $W_h \subset H(\operatorname{curl},\Omega)$ a finite-dimensional subspace and adopt the notation of Section 1.2.2. For the discretization of (1.33)–(1.35), we consider the following approach.

**Problem 1.2.12.** Let initial values $\mathbf{e}_h^0, \mathbf{a}_h^0 \in W_h$ be given. Then, for $1 \le n \le N$ find $\mathbf{e}_h^n, \mathbf{a}_h^n \in P_{k+1}(I^n; W_h)$ with $\mathbf{e}_h^n(t^{n-1}) = \mathbf{e}_h^{n-1}(t^{n-1})$ and $\mathbf{a}_h^n(t^{n-1}) = \mathbf{a}_h^{n-1}(t^{n-1})$, such that

$$-\int_{I^n} \langle \epsilon(\mathbf{e}_h^n)\partial_t\mathbf{a}_h^n, \tilde{\mathbf{v}}_h\rangle\, dt = \int_{I^n} \langle \epsilon(\mathbf{e}_h^n)\mathbf{e}_h^n, \tilde{\mathbf{v}}_h\rangle\, dt, \tag{1.41}$$

$$\int_{I^n} \langle \epsilon(\mathbf{e}_h^n)\partial_t\mathbf{e}_h^n, \tilde{\mathbf{w}}_h\rangle\, dt = \int_{I^n} \langle \nu_0 \operatorname{curl}\mathbf{a}_h^n, \operatorname{curl}\tilde{\mathbf{w}}_h\rangle\, dt, \tag{1.42}$$

for all test functions $\tilde{\mathbf{v}}_h, \tilde{\mathbf{w}}_h \in P_k(\mathcal{T}; W_h)$.

The method (1.41)–(1.42) corresponds to the Galerkin approximation of (1.38)–(1.39) in space together with the Petrov-Galerkin approach in time [72, 95]. Let us emphasize that in contrast to the previous approach, we are now looking for a solution that is continuous in time, and the test functions might be discontinuous and have a smaller by one polynomial degree. Thus, the method is indeed a Petrov-Galerkin approach. The energy-conserving property of this approximation is summarized in the following lemma.

**Lemma 1.2.13.** Let $(\mathbf{e}_h^n, \mathbf{a}_h^n)_{n\geq 0}$ denote a solution of Problem 1.2.12. Then

$$\mathcal{H}(\mathbf{e}_h^n(t^n), \mathbf{a}_h^n(t^n)) = \mathcal{H}(\mathbf{e}_h^m(t^m), \mathbf{a}_h^m(t^m)), \tag{1.43}$$

for $0 \leq m \leq n \leq N$. Thus, the discrete energy is conserved exactly for all time steps.

*Proof.* The proof is a direct consequence of [43, Theorem 2] that has been presented in Appendix A.2. For the convenience of the reader, we show the proof for this particular problem. First, let $m = n - 1$. By the fundamental theorem of calculus, we obtain

$$\begin{aligned}
\mathcal{H}(\mathbf{e}_h^n(t^n), \mathbf{a}_h^n(t^n)) - \mathcal{H}(\mathbf{e}_h^n(t^{n-1}), \mathbf{a}_h^n(t^{n-1})) &= \int_{I^n} \frac{d}{dt} \mathcal{H}(\mathbf{e}_h^n(t), \mathbf{a}_h^n(t))\, dt \\
&= \int_{I^n} \langle \tilde{w}'_{el}(\mathbf{e}_h^n(t)), \partial_t \mathbf{e}_h^n(t) \rangle + \langle w'_{mag}(\operatorname{curl} \mathbf{a}_h^n(t)), \operatorname{curl} \partial_t \mathbf{a}_h^n(t) \rangle\, dt \\
&= \int_{I^n} \langle \epsilon(\mathbf{e}_h^n(t)), \partial_t \mathbf{e}_h^n(t) \rangle + \langle \nu_0 \operatorname{curl} \mathbf{a}_h^n(t), \operatorname{curl} \partial_t \mathbf{a}_h^n(t) \rangle\, dt = (*),
\end{aligned} \tag{1.44}$$

where we used $\tilde{w}'_{el}(\mathbf{e}) = \epsilon(\mathbf{e})\mathbf{e}$ and $w'_{mag}(\mathbf{b}) = \nu_0 \mathbf{b}$, which hold for any argument. Now, we use the scheme (1.41)–(1.42) with $\tilde{\mathbf{v}}_h = \partial_t \mathbf{e}_h^n$ and $\tilde{\mathbf{w}}_h = \partial_t \mathbf{a}_h^n$, which are admissible test functions, and conclude

$$(*) = -\int_{I^n} \langle \epsilon(\mathbf{e}_h^n(t)) \partial_t \mathbf{a}_h^n(t), \partial_t \mathbf{e}_h^n(t) \rangle - \langle \epsilon(\mathbf{e}_h^n(t)) \partial_t \mathbf{e}_h^n(t), \partial_t \mathbf{a}_h^n(t) \rangle\, dt = 0.$$

Due to the continuity of solutions $\mathbf{e}_h^n(t^{n-1}) = \mathbf{e}_h^{n-1}(t^{n-1})$ and $\mathbf{a}_h^n(t^{n-1}) = \mathbf{a}_h^{n-1}(t^{n-1})$, we further conclude that

$$\mathcal{H}(\mathbf{e}_h^n(t^n), \mathbf{a}_h^n(t^n)) = \mathcal{H}(\mathbf{e}_h^n(t^{n-1}), \mathbf{a}_h^n(t^{n-1})) = \mathcal{H}(\mathbf{e}_h^{n-1}(t^{n-1}), \mathbf{a}_h^{n-1}(t^{n-1})),$$

which proves the statement for $m = n - 1$. The case $m < n - 1$ follows by induction. □

**Remark 1.2.14.** Let us emphasize that the scalar product $\langle \cdot, \cdot \rangle$ can be replaced by $\langle \cdot, \cdot \rangle_h$ that comes from inexact integration by a quadrature rule. The energy-conserving principle (1.43) remains valid for the perturbed energy $\mathcal{H}_h(\cdot, \cdot)$, which is computed with the same quadrature rule. The exact evaluation of time integrals associated with the energy term, namely $\int_{I^n} \langle \epsilon(\mathbf{e}_h^n)\mathbf{e}_h^n, \tilde{\mathbf{v}}_h \rangle\, dt$ and $\int_{I^n} \langle \nu_0 \operatorname{curl} \mathbf{a}_h^n, \operatorname{curl} \tilde{\mathbf{w}}_h \rangle\, dt$, is essential for the relation (1.44) to hold. The other time integrals can be approximated by a fixed quadrature rule.

Similarly to the approach of Section 1.2.2, the scheme (1.41)–(1.42) represents a rather general framework. Remarks on a possible numerical implementation are given below.

### Finite element discretization in space

We again use the finite element approach for the Galerkin approximation in space. However, now we use the same spaces in both variables. A semi-discretization of the system (1.33)–(1.34) by an appropriate finite element method leads to a finite-dimensional system of differential equations

$$-\mathsf{M}_{\mathsf{e}}(\mathsf{e}(t))\partial_t \mathsf{a}(t) = \mathsf{M}_{\mathsf{e}}(\mathsf{e}(t))\mathsf{e}(t), \tag{1.45}$$

$$\mathsf{M}_{\mathsf{e}}(\mathsf{e}(t))\partial_t \mathsf{e}(t) = \mathsf{K}_{\nu}\mathsf{a}(t), \tag{1.46}$$

where $\mathsf{e}(t)$ and $\mathsf{a}(t)$ are the coefficient vectors and $\mathsf{M}_{\mathsf{e}}(\mathsf{e}(t))$ and $\mathsf{K}_{\nu}$ are the system matrices.

**Example 1.2.15.** Following the Example 1.2.7, one can again consider discrete subspace

$$W_h = \{\mathbf{w} \in H(\mathrm{curl}, \Omega) \,:\, \mathbf{w}|_K \in \mathcal{N}_p(K) \text{ for all } K \in \mathcal{T}_h\},$$

where $\mathcal{N}_p$ is the Nedelec space. With the appropriate choice of basis $\{\phi_k\}_k \subset W_h$, the discretization of (1.38)–(1.38) leads to the system (1.45)–(1.46) with $(\mathsf{M}_{\mathsf{e}}(\mathsf{e}))_{ij} = \langle \epsilon(\sum_l \mathsf{e}_l \phi_l)\phi_j, \phi_i\rangle$ and $(\mathsf{K}_{\nu})_{ij} = \langle \nu_0 \,\mathrm{curl}\,\phi_j, \mathrm{curl}\,\phi_i\rangle$. The inexact realization of the scalar product using e.g. mass lumping techniques can also be considered; see Remark 1.2.14.

### Petrov-Galerkin discretization in time

In this thesis, we consider several problems of the canonical structure (1.40). Therefore, we again consider a unified implementation strategy. We write the system (1.45)–(1.46) in the abstract form

$$\mathsf{C}(\mathsf{u}(t))\partial_t \mathsf{u}(t) + \mathsf{H}'(\mathsf{u}(t)) = 0, \tag{1.47}$$

where $\mathsf{u}(t) = (\mathsf{e}(t), \mathsf{a}(t))$ is the vector with coefficients, $\mathsf{C}(\mathsf{u}(t))$ is a skew-symmetric matrix

$$\mathsf{C}(\mathsf{u}(t)) = \begin{pmatrix} 0 & \mathsf{M}_{\mathsf{e}}(\mathsf{e}(t)) \\ -\mathsf{M}_{\mathsf{e}}(\mathsf{e}(t)) & 0 \end{pmatrix} \quad \text{and} \quad \mathsf{H}'(\mathsf{u}(t)) = \begin{pmatrix} \mathsf{M}(\mathsf{e}(t))\mathsf{e}(t) \\ \mathsf{K}\mathsf{a}(t) \end{pmatrix}.$$

The time stepping scheme (1.41)–(1.42) can then be formulated as follows: For given $\mathsf{u}^n(t^{n-1})$ we seek for $\mathsf{u}^n \in P_{k+1}(I^n; \mathbb{R}^d)$ such that

$$\int_{I^n} \left[\mathsf{C}(\mathsf{u}^n(t))\partial_t \mathsf{u}^n(t) + \mathsf{H}'(\mathsf{u}^n(t))\right] \tilde{v}(t)\, dt = 0, \tag{1.48}$$

holds for all $\tilde{v} \in P_k(I^n; \mathbb{R})$ and all $1 \le n \le N$. We again use scalar-valued test functions; see e.g. [43, 95]. The relation (1.48) is then a $d-$ dimensional system of equations.

**Example 1.2.16.** Consider the scheme for $k = 0$. In this case, since the test function $\bar{v} \in P_0(I^n; \mathbb{R})$ is constant over the time interval $I^n$, the multiplication with the test function can be neglected. Using the simplified notation $\mathsf{u}^n = \mathsf{u}(t^n)$, the scheme can be written as

$$\int_{I^n} \mathsf{C}(\mathsf{u}(t))\, dt\, \frac{\mathsf{u}^n - \mathsf{u}^{n-1}}{\tau^n} + \int_{I^n} \mathsf{H}'(\mathsf{u}(t))\, dt = 0$$

The method coincides with a *discrete gradient* approach; see e.g. [31, 58].

**Remarks on implementation**

For further discussion on the construction of higher-order schemes, we use a simplified notation. With $I = [t^0, t^1]$ we denote the step of size $\tau$ and write $\mathsf{u} = \mathsf{u}^1 \in P_k(I; \mathbb{R}^d)$. Now let $\psi_j(t) \in P_k(I^n; \mathbb{R})$ denote the basis of the test space. By expanding $\partial_t \mathsf{u} \in P_k(I^n; \mathbb{R})$ with respect to this basis, we obtain

$$\partial_t \mathsf{u}(t) = \sum_{j=0}^{k} \mathsf{U}'_j \psi_j(t) \quad \text{and} \quad \mathsf{u}(t) = \mathsf{u}^0 + \sum_{j=0}^{k} \mathsf{U}'_j \int_0^t \psi_j(\tau)\, d\tau, \qquad \forall t \in I, \tag{1.49}$$

where $\mathsf{U}'_j \in \mathbb{R}^d$, $j = 0, \dots, k$ are the vector valued coefficients. For the evaluation of the integral term in (1.48), we utilize numerical quadrature. Let $(w_\ell, \xi_\ell)$ be weights and points of a quadrature rule and $\hat{\psi}_j$ denote the basis functions on the reference interval $\hat{I} = [0, 1]$. Furthermore, we use the following notation

$$\gamma_j^\ell = \hat{\psi}_j(\xi_\ell), \quad \alpha_j^\ell = \int_0^{\xi_\ell} \hat{\psi}_j(t)\, dt \quad \text{and} \quad \beta_j = \int_0^1 \hat{\psi}_j(t)\, dt.$$

Substituting the ansatz (1.49) in the formulation (1.48), transforming to the reference interval, and replacing the integral with the quadrature rule leads to the formulation

$$\begin{aligned} \sum_{\ell=0}^{L} w_\ell \Big[ \mathsf{C}(\mathsf{U}_\ell) \sum_{j=0}^{k} \gamma_j^\ell \mathsf{U}'_j + \mathsf{H}'(\mathsf{U}_\ell) \Big] \gamma_i^\ell = 0, \quad 0 \le i \le k, \\ \mathsf{U}_\ell = \mathsf{u}^0 + \tau \sum_{j=0}^{k} \alpha_j^\ell \mathsf{U}'_j, \quad \mathsf{u}^1 = \mathsf{u}^0 + \tau \sum_{j=0}^{k} \beta_j \mathsf{U}'_j. \end{aligned} \tag{1.50}$$

The problem represents a nonlinear system of size $(k+1)d$, which can be solved by e.g. Newton method. The expression for the derivative operator can be constructed analytically by derivation of individual terms.

An appropriate choice of a quadrature rule is crucial for our purposes. As mentioned in Remark 1.2.14, it is essential that the energy-related term is integrated exactly. Since the permittivity $\epsilon(\cdot)$ is quadratic, the chosen quadrature rule must be exact for polynomials of degree $4k+3$, where $k+1$ and $k$ are the polynomial orders of the trial and test spaces.

Let us further note that we used the same quadrature for both summands, which is not necessary. Moreover, the inexact realization of the first term often leads to simplified schemes. A particular simplification can be achieved when a $(k+1)-$ node quadrature is chosen. By taking the Lagrange interpolation basis with respect to the quadrature nodes, we conclude that $\gamma_i^\ell = \delta_{i\ell}$ and the first term in (1.50) reads

$$\sum_{\ell=0}^{k} w_\ell \Big[ \mathsf{C}(U_\ell) \sum_{j=0}^{k} \gamma_j^\ell \mathsf{U}'_j \Big] \gamma_i^\ell = w_i \mathsf{C}(\mathsf{U}_i) \mathsf{U}'_i.$$

The resulting scheme is then an *average vector field collocation method* [31, 58, 107]. These methods are perfectly well suited for our purposes. When both terms are approximated in this sense, the scheme becomes a Runge-Kutta collocation method. In particular,

LobattoIIIA schemes can be obtained from the inexact realization of all terms by the corresponding Lobatto quadrature rules.

In this thesis, we utilize the Lagrange interpolation polynomials at Lobtatto nodes as a basis and use a Lobatto quadrature rule of sufficiently higher order for the integration.

### 1.2.5. Numerical illustration

To illustrate the theoretical results of the section and investigate the resulting discretization schemes, we provide two numerical examples. First, we consider a one-dimensional problem, as often done in the related literature; see e.g. [16, 18]. We demonstrate the energy-diminishing and energy-conserving properties of the discretizations, discuss the convergence of the schemes, and compare the methods to FDTD approaches of [94]. In the second example, we consider a transverse magnetic regime. We briefly validate the energy and convergence-related results and compare the efficiency of the schemes to the FDTD approach of [74]. Here and in the following, we consider scaled problems as is often done in related literature. The parameters are dimensionless. The constant corresponding to the nonlinear factor is increased for better visualization of nonlinear effects.

#### 1D optical pulse propagation

The one-dimensional problem is based on the assumptions that the fields are of the form $\mathbf{e} = (e_x, 0, 0)$ and $\mathbf{h} = (0, h_y, 0)$, and depend only on the propagation direction $z$, which leads to $\operatorname{curl} \mathbf{e} = (0, \partial_z e_x, 0)$ and $\operatorname{curl} \mathbf{h} = (-\partial_z h_y, 0, 0)$. We consider the formulation in terms of $e_x$ and $h_y$ as follows

$$\epsilon(e_x)\partial_t e_x = -\partial_z h_y, \quad \mu_0 \partial_t h_y = -\partial_z e_x.$$

By similar considerations, we conclude that $\mathbf{a} = (a_x, 0, 0)$ and depends only on $z$. Then, the formulation in terms of $e_x$ and $a_x$ reads

$$-\epsilon(e_x)\partial_t a_x = \epsilon(e_x)e_x, \quad \epsilon(e_x)\partial_t e_x = -\partial_z(\nu_0 \partial_z a_x).$$

For simplicity, we set $\epsilon_0 = \mu_0 = \nu_0 = \tilde{\chi}^{(1)} = 1$ and $\chi^{(3)} = 0.1$. We choose the initial values $h_y(0) = a_x(0) = 0$ and $e_x(0) = \exp(-100z^2)$. We consider a computational domain $\Omega = [0, 1]$ and assume $h_y = 0$ and $\partial_z a_x = 0$ on $\partial\Omega$, respectively. The snapshots of the propagation of the electric field $e_x$ are illustrated in Figure 1.1. We also plot the solution of the linear problem with $\chi^{(3)} = 0$ with a dashed line to highlight the nonlinear effect. We observe the formation of the characteristic kink-solution, which can be verified at the analytical level; see e.g. [16, 106].

**Discretization in space.** We denote by $\mathcal{T}_h$ an equidistant mesh with grid points $x_i = ih$. Further, we utilize the piecewise polynomial subspaces $W_h = P_{p+1}(\mathcal{T}_h) \cap H^1(\Omega)$ and $Q_h = P_p(\mathcal{T}_h)$ of degree $p+1$ and $p$ for spatial discretization. Moreover, we choose the Lagrange polynomials on Gauss Lobatto Legendre nodes as basis functions and consider an inexact realization of the scalar product $\langle \cdot, \cdot \rangle_h$ resulting from the corresponding Gauss quadrature rule. The same quadrature rule is also used in the energy evaluation. As mentioned in Remark 1.2.6 and Remark 1.2.14, the underlying energy principles remain valid.

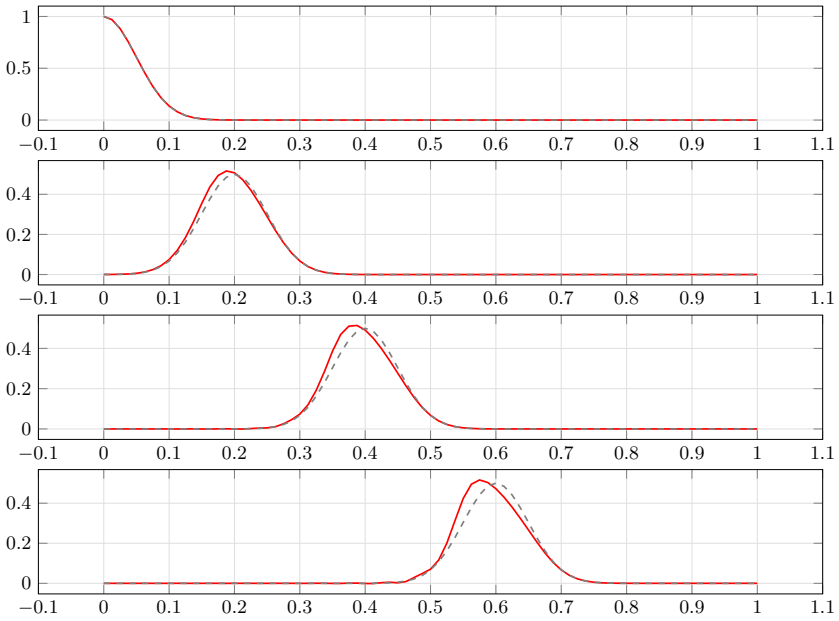


Figure 1.1.: Red solid: Numerical solution $\mathbf{e}(t^n)$ at times $t^n = 0, 0.2, 0.4, 0.6$ obtained by the lowest order scheme (1.41)–(1.42). Black dashed: the corresponding solution of the linear problem, i.e., with $\chi^{(3)} = 0$ for comparison.

**Discretization in time.** For the time integration, we write the semi-discrete problems in abstract forms (1.31) and (1.47). We use Lagrange interpolation polynomials associated with Gauss Lobatto nodes as the basis in both approaches. For details on implementation, we refer to Sections 1.2.2 and 1.2.4. Since the nonlinearity is polynomial, the exact evaluations of time integrals can be achieved by a quadrature rule. For the $\mathbf{e}-\mathbf{h}$ system, the quadrature has to integrate polynomials of degree $4k-1$ exactly, while for the $\mathbf{e}-\mathbf{a}$ approach, the quadrature has to be exact for polynomials of degree $4k+3$. We use the Gauss quadrature with $2k$ and $2k+2$ nodes, respectively. The resulting nonlinear systems are solved by the Newton method with tolerance $10^{-12}$.

**Convergence of the $\mathbf{e}-\mathbf{h}$ scheme.** We applied the scheme (1.25)–(1.26) to the $\mathbf{e}-\mathbf{h}$ formulation with different polynomial degrees in space and time. In Tables 1.1 and 1.2, we summarize the observed errors and convergence rates for polynomial degrees $p = 1, 2, 3$ of approximation in space and degree $k = 0, 1, 2$ of approximation in time. The errors are

| $h$ | $p=1$ | | $p=2$ | | $p=3$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-2}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. |
| 0.05 | 0.248186 | — | 0.387722 | — | 0.417564 | — |
| 0.025 | 0.071272 | 1.80 | 0.003319 | 3.54 | 0.018346 | 4.50 |
| 0.0125 | 0.018438 | 1.95 | 0.000299 | 3.47 | 0.000950 | 4.27 |
| 0.00625 | 0.004641 | 1.99 | 0.000034 | 3.11 | 0.000058 | 4.02 |

Table 1.1.: Convergence in space of the method (1.25)–(1.26) for the $\mathbf{e}-\mathbf{h}$ formulation.

measured by err $= \max_{0\le n\le N} \|e^n_{x,h}(t^n) - e^n_{x,h/2}(t^n)\|_{h/2}$, where $e_{x,h/2}$ denotes the solution

| $\tau$ | $k=0$ | | $k=1$ | | $k=2$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-2}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. |
| 0.025 | 0.257057 | — | 0.280420 | — | 0.550798 | — |
| 0.0125 | 0.171025 | 0.61 | 0.038358 | 2.87 | 0.019199 | 4.84 |
| 0.00625 | 0.100673 | 0.76 | 0.004879 | 2.98 | 0.000610 | 4.97 |
| 0.003125 | 0.054697 | 0.88 | 0.000612 | 3.00 | 0.000019 | 5.00 |

Table 1.2.: Convergence in time of the method (1.25)–(1.26) for the $\mathbf{e}-\mathbf{h}$ formulation.

on the uniformly refined mesh $\mathcal{T}_{h/2}$ with size $h/2$. With $\|\cdot\|_{h/2}$ we denote the approximation of the $L^2$ norm, which is computed by the numerical quadrature rule on the refined mesh $\mathcal{T}_{h/2}$. We observe convergence $O(h^{p+1})$ in space, which is known for linear problems; see [35, 56]. The time discretization errors are computed by $\text{err} = \max_{0\leq n\leq N} \|e^n_{x,h}(t^n) - e^{2n}_{x,h}\|_h$, where $e^{2n}_{x,h}$ denotes a discrete solution on the uniformly refined grid, with $\tilde{\tau} = \tau/2$. We observe super convergence $O(\tau^{2k+1})$ in time. The convergence results coincide with the results for linear problems and, in particular, with related Radau schemes with $s = k+1$ stages; see [93, 134].

**Convergence of the $\mathbf{e}-\mathbf{a}$ scheme.** In Table 1.3 and 1.4, we state the errors obtained by the method (1.41)–(1.42) for different approximation orders $p$ and $k$ in space and time. For error computations, we use the same expressions as previously. We observe the convergence $O(h^{p+1})$ in space and $O(\tau^{2k+2})$ in time. This super convergence in time has been obtained in [10, 95] for different problems and coincides with that of related Lobatto schemes.

| $h$ | $p=1$ | | $p=2$ | | $p=3$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-3}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. |
| 0.05 | 0.412735 | — | 0.297889 | — | 0.277589 | — |
| 0.025 | 0.127333 | 1.70 | 0.022976 | 3.69 | 0.011844 | 4.55 |
| 0.0125 | 0.033235 | 1.94 | 0.002874 | 2.99 | 0.000747 | 3.99 |
| 0.00625 | 0.008372 | 1.99 | 0.000359 | 3.00 | 0.000046 | 3.99 |

Table 1.3.: Convergence in space of the method (1.41)–(1.42) for the $\mathbf{e}-\mathbf{a}$ formulation.

| $\tau$ | $k=0$ | | $k=1$ | | $k=2$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. | err $\times 10^{-4}$ | e.o.c. |
| 0.05 | 0.801343 | — | 0.611080 | — | 0.368882 | — |
| 0.025 | 0.226645 | 1.82 | 0.040060 | 3.93 | 0.006549 | 5.81 |
| 0.0125 | 0.057709 | 1.97 | 0.002538 | 3.98 | 0.000108 | 5.93 |
| 0.00625 | 0.014537 | 1.99 | 0.000160 | 3.98 | 0.000002 | 5.96 |

Table 1.4.: Convergence in time of the method (1.41)–(1.42) for the $\mathbf{e}-\mathbf{a}$ formulation.

**Evolution of energy.** To emphasize the evolution of energy over time, we consider a coarse grid with $h = 0.25$ and $\tau = 0.2$ and increase the nonlinear factor to $\chi^{(3)} = 1$. We consider low order approximations with $p = 1$ and $k = 1$. The evolution of the energy for

the $\mathbf{e}-\mathbf{h}$ approach is depicted with the red line in Figure 1.2. As expected we observe the decay of the energy over time. This is no longer the case for the related 2-stage RadauIIA method, which is illustrated with a red dashed line. For such a coarse discretization the dissipation becomes significant. The strategy based on the $\mathbf{e}-\mathbf{a}$ formulation, on the other hand, preserves the energy up to the error of machine precision $O(10^{-15})$. The evolution of energy is illustrated by the blue line in Figure 1.2. Its inexact realization – the trapezoidal rule, on the other hand, leads to energy growth, as depicted by the blue dashed line. Moreover, the method becomes unstable for larger time steps.

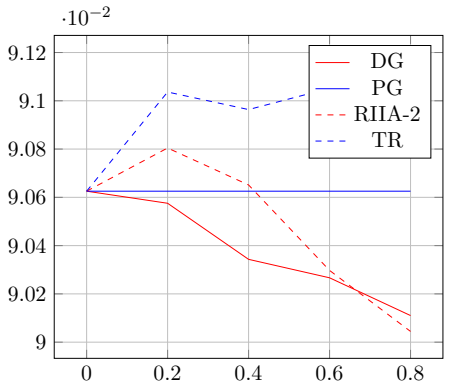



Figure 1.2.: Time evolution of energies for $h = 0.25$, $\tau = 0.2$, $\chi^{(3)} = 0.9$.

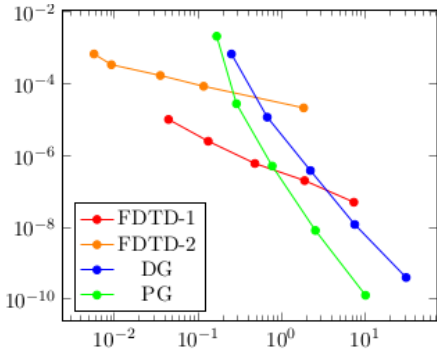



Figure 1.3.: Error - complexity plot. $x$-axis: time, $y$-axis: error

**Computational complexity.** Let us start by mentioning that the complexity of the variational time stepping schemes is comparable to that of the implicit $s = k + 1$ stage Runge-Kutta methods. Both approaches lead to a nonlinear $(k+1)d$-dimensional system, which has to be solved in each time step. The size of the semi-discrete systems for the two approaches is $d = (2N_z - 1)p$ and $d = 2N_z p$, respectively, where $N_z$ is the number of grid points in the spatial discretization. As previously mentioned, we use Newton's method for the solution of nonlinear problems. The accuracy $O(10^{-12})$ is achieved with at most 3 iterations. For the solution of linear problems, we used Matlab's backslash operator, which seems to be of almost linear complexity.

With the blue line in Figure 1.3 we illustrate the relation between the computational time and relative error for the $\mathbf{e}-\mathbf{h}$ approach with $p = 4$ and $k = 2$, which leads to error $\mathrm{err} = \max_{0\leq n\leq N} \|e^n_{x,h}(t^n) - e^{2n}_{x,h/2}(t^n)\|_{h/2} = O(\tau^5 + h^5)$. Here we choose the discretization with $\tau = h/2 = 0.1, 0.05, \ldots, 0.1 \cdot 2^{-6}$. With the green line, we illustrate the results for the $\mathbf{e}-\mathbf{a}$ approach with $p = 5$ and $k = 2$, which leads to an error of $O(\tau^6 + h^6)$. For comparison, with orange and red lines, we illustrate the results for some of the FDTD methods, namely, that of Method-1 and Method-2 in [94]. Let us note that Method-1 is an implicit method and leads to second-order accuracy. Method-2 is, on the other hand, fully explicit and much faster. However, we observe only linear convergence. When high accuracy is desired, the proposed high-order schemes become more efficient. Let us remark that the particular choice $h = 2\tau$ is motivated by the CFL condition, which is necessary for the stability of the FDTD schemes, and is determined numerically.

### 2D Transverse magnetic setting

For the second example, we follow [74] and consider the transverse magnetic regime. We assume that $\mathbf{e} = (0, 0, e_z)$ and $\mathbf{h} = (h_x, h_y, 0)$, where $e_z$, $h_x$, and $h_y$ are functions of $x$ and $y$. In this, we conclude that $\operatorname{curl} \mathbf{e} = (\partial_y e_z, -\partial_x e_z, 0)$ and $\operatorname{curl} \mathbf{h} = (0, 0, \partial_x h_y - \partial_y h_x)$. With $\mathbf{h}_{xy} = (h_x, h_y)$ we denote the $x$ and $y$ component of the vector field $\mathbf{h}$ and with $\mathbf{h}_{xy}^{\perp} = (-h_y, h_x)$ we denote its orthogonal. Then the $\mathbf{e} - \mathbf{h}$ formulation can be written in terms of $e_z$ and $\mathbf{h}_{xy}^{\perp}$ as follows

$$\epsilon(e_z)\partial_t e_z = \operatorname{div} \mathbf{h}_{xy}^{\perp}, \quad \mu_0 \partial_t \mathbf{h}_{xy}^{\perp} = -\nabla e_z.$$

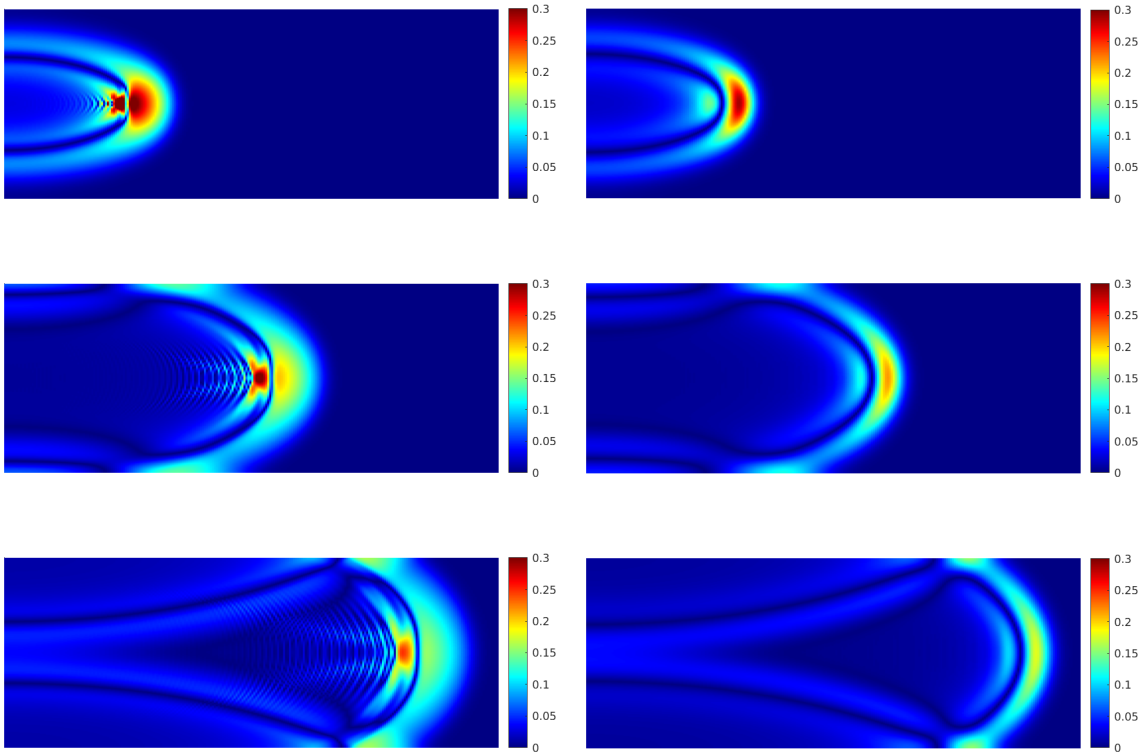
The vector potential then takes the form $\mathbf{a} = (0, 0, a_z)$ with $a_z$ independent of $z$ and we have $\mathbf{h}_{xy}^{\perp} = \nabla a_z$. Then, the $\mathbf{e} - \mathbf{a}$ formulation reads

$$-\epsilon(e_z)\partial_t a_z = \epsilon(e_z)\epsilon_z, \quad \epsilon(e_z)\partial_t \epsilon_z = -\operatorname{div}(\nu_0 \nabla a_z).$$

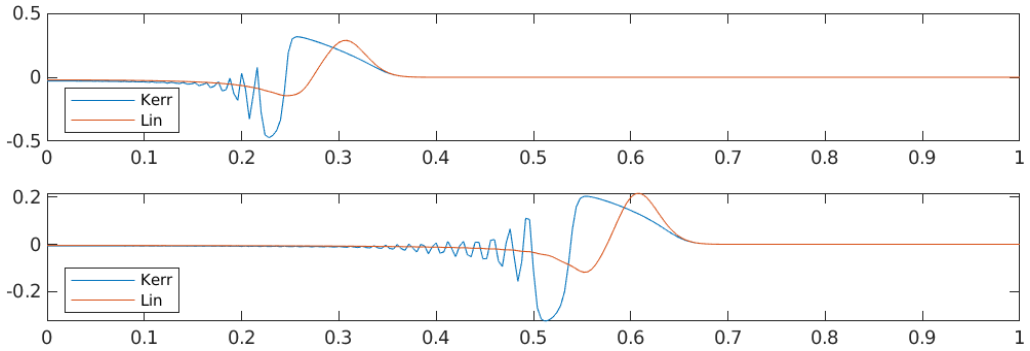
We set $\epsilon_0 = \mu_0 = \nu_0 = \tilde{\chi}^{(1)} = 1$ and consider a bounded domain with $\Omega = [0, 1]^2$. The snapshots of the magnitude of the electric field $|e_z|$ for $\chi^{(3)} = 1$ and initial values $e_z(0) = \exp(-1000x^2 - 100(y - 0.5)^2)$ and $\mathbf{h}_{xy}^{\perp}(0) = 0$ is illustrated in Figure 1.4. For comparison, we also plot the solutions of the linear problem with $\chi^{(3)} = 0$. Here one can observe the so-called self-focusing effect of the Kerr media – the intensity of the pulse becomes highest at the center of the beam. We also plot the values at the horizontal cut through the middle of the domain $y = 0.5$. Similarly to the previous example, we observe the formation of the characteristic kinks. We also observe relatively strong oscillations, related to the discontinuity of the solution.

**Discretization details.** For the finite element approximation in space, we consider a structured tensor grid $\mathcal{T}_{xy} = \mathcal{T}_x \times \mathcal{T}_y$ where both $\mathcal{T}_x$ and $\mathcal{T}_y$ are equidistant meshes in $x$ and $y$ direction. The bases for the discrete subspaces $W_h = P_{p+1} \cap H^1(\Omega)$ and $Q_h = P_p$ are constructed by the product of Lagrange polynomials on Lobatto nodes in each direction. We also use inexact evaluation of the scalar products using the quadrature associated with the Lobatto nodes; see e.g. [36]. The same quadrature is also used for the evaluation of energies. For the time discretization, we use the same implementation as previously.

**Simulation results.** The energy dissipation and energy conservation results were verified for this problem as well. For a sufficiently smooth solution, we also observe $O(\tau^{2k+1} + h^{p+1})$ convergence for the method based on the $\mathbf{e} - \mathbf{h}$ formulation, and $O(\tau^{2k+2} + h^{p+1})$ for the approach based on the $\mathbf{e} - \mathbf{a}$ system. For comparison, we also applied the energy-stable FDTD method of [74], which is second-order accurate. Convergence results for the fourth order $\mathbf{e} - \mathbf{a}$ approach, third order $\mathbf{e} - \mathbf{h}$ method, and the FDTD scheme are illustrated in Table 1.5. We also plot the relations between the error and computational times in Figure 1.5. For these convergence results, we decreased the nonlinear impact by setting $\chi^{(3)} = 0.3$, $e_z(0) = \exp(-100(x - 0.5)^2 - 100(y - 0.5)^2)$, and decreasing simulation time to $T = 0.2$ in order to reduce the impact of discontinuities. We use the same expression as for the one-dimensional case for the evaluation of the error in space and time. The initial time step $\tau$ and grid sizes $h$ are chosen such that the errors in space and time are of the same order, and then refined uniformly. One can again observe that the higher-order schemes become more efficient when higher accuracy is needed. In our experiments, the Newton solver required at most four iterations to achieve the desired accuracy of $O(10^{-12})$, while in the case of FDTD, it required at most two.




(a) Snapshots of the magnitude of the electric field $|e_z(t^n)|$ for nonlinear Kerr media with $\chi^{(3)} = 1$ (left) and for linear media with $\chi^{(3)} = 0$ (right) at time steps $t^n = 0.3, 0.6, 0.9$.




(b) Values of the electric fields $e_z$ at the cut $y = 0.5$ for the nonlinear problem with $\chi^{(3)} = 1$ (red) and for the linear problem with $\chi^{(3)} = 0$ (blue) at time steps $t^n = 0.3, 0.6$.

Figure 1.4.: Behaviour of the electromagnetic field in nonlinear Kerr media with $\chi^{(3)} = 1$ and in linear media with $\chi^{(3)} = 0$ for comparison.

### A note on generalizations

Up to this point, we have considered a specific example of a nonlinear dielectric medium of Kerr-type. However, the methodology is not limited to Kerr media. The ideas presented here can be applied to any other type of nonlinear electric media that satisfies the principles of Section 1.1. The same approach can be extended to electrically conducting materials, where instead of energy conservation, we obtain energy dissipation balances that can be

| FDTD | | | $\mathbf{e}-\mathbf{a}$ method $(p=3,k=1)$ | | | $\mathbf{e}-\mathbf{h}$ method $(p=2,k=1)$ | | |
|---|---|---|---|---|---|---|---|---|
| $h=8\tau$ | err | e.o.c. | $h=4\tau$ | err | e.o.c. | $h=4\tau$ | err | e.o.c. |
| 0.008 | 0.001842 | – | 0.25 | 0.027144 | – | 0.25 | 0.138261 | – |
| 0.004 | 0.000455 | 2.01 | 0.125 | 0.002953 | 3.3 | 0.125 | 0.027322 | 2.24 |
| 0.002 | 0.000149 | 1.91 | 0.0625 | 0.000107 | 4.6 | 0.0625 | 0.002730 | 3.32 |
| 0.001 | 0.000041 | 1.92 | 0.03125 | 0.000006 | 4.1 | 0.03125 | 0.000279 | 3.28 |

Table 1.5.: Convergence in space and time for different methods.

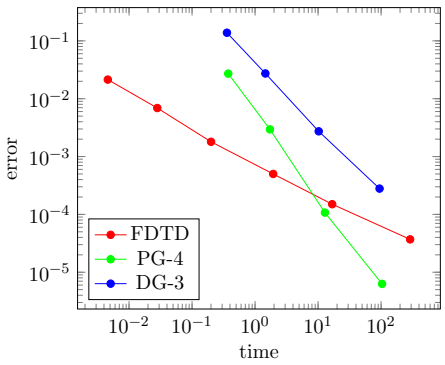



Figure 1.5.: Efficiency comparison of the schemes for 2D transverse magnetic problem. As the benchmark we take the FDTD method introduced in [74].

preserved using the same discretization techniques. Furthermore, the generalization to nonlinear magnetic problems is straightforward, as long as the principles discussed in Section 1.1 remain applicable. The treatment of the vector potential formulation for the nonlinear eddy-current problem is covered in [43] and will be briefly addressed in Chapter 2

Furthermore, the concepts presented in this work can be extended to memory-dependent materials. We stay in the context of nonlinear optical applications and discuss the treatment of problems involving electric dispersion. It is important to note that this topic is still under investigation and has only been explored for a few specific examples. We consider a simple Kerr-Lorentz model where the dispersive component is described by the linear Lorentz oscillator model. By analogy, the Debye model can also be considered. While the generalization to a nonlinear dispersive Kerr-Debye-Lorentz model seems feasible, it has not been extensively studied yet. Further generalizations are yet to be considered.

## 1.3. Nonlinear media with dispersion

We consider the propagation of electromagnetic field in a Kerr-Lorentz media [19, 103, 126]. As previously, the underlying physics are described by Maxwell's equations

$$\partial_t \mathbf{d} = \operatorname{curl} \mathbf{h} \quad \text{and} \quad \partial_t \mathbf{b} = -\operatorname{curl} \mathbf{e}, \tag{1.51}$$

and we assume the constitutive relations

$$\mathbf{b} = \mu_0 \mathbf{h} \quad \text{and} \quad \mathbf{d} = \epsilon_0((\epsilon_\infty + \alpha|\mathbf{e}|^2)\mathbf{e} + \mathbf{p}), \tag{1.52}$$

where $\epsilon_\infty$ denotes the high-frequency limit of the relative permittivity and $\alpha$ is a constant that describes the impact of the nonlinear effect. Furthermore, $\mathbf{p}$ denotes the frequency-dependent part of the polarization, which is characterized by the auxiliary differential equation

$$\partial_{tt}\mathbf{p} + \gamma\partial_t\mathbf{p} + \omega_0^2\mathbf{p} = \omega_p^2\mathbf{e}. \tag{1.53}$$

Here, $\omega_0$ and $\omega_p$ represent the resonance and plasma frequencies, respectively, and $\gamma$ is the damping parameter [84, 126]. It should be noted that these quantities are related through the equation $\omega_p^2 = (\epsilon_\infty - \epsilon_s)\omega_0^2$, where $\epsilon_s$ is the static relative permittivity, and $\gamma = 1/\tau$, with $\tau$ being the relaxation time; see e.g. [18, 131].

**Outline.** By analogy to Section 1.2, we present two formulations as extensions of the $\mathbf{e}-\mathbf{h}$ and $\mathbf{e}-\mathbf{a}$ formulations. We derive the energy balances associated with these formulations and briefly discuss the construction of schemes that preserve these balances. The following results has not been published yet.

### 1.3.1. The e - h formulation for Kerr-Lorentz model

We begin by rewriting equation (1.53) in the first-order form, which is a commonly used practice. Introducing the linear polarization current density $\mathbf{j} = \partial_t\mathbf{p}$, we obtain the system

$$\partial_t\mathbf{p} = \mathbf{j}, \quad \partial_t\mathbf{j} + \gamma\mathbf{j} + \omega_0^2\mathbf{p} = \omega_p^2\mathbf{e}. \tag{1.54}$$

By differentiating the constitutive relation (1.52) with respect to time and using the relation $\mathbf{j} = \partial_t\mathbf{p}$, we obtain the relation for the displacement current

$$\partial_t\mathbf{d} = \tilde{\epsilon}(\mathbf{e})\partial_t\mathbf{e} + \epsilon_0\mathbf{j}, \tag{1.55}$$

where we use the abbreviation $\tilde{\epsilon}(\mathbf{e}) = \epsilon_0(\epsilon_\infty + \alpha|\mathbf{e}|^2\mathbf{e} + 2\alpha\mathbf{e}\mathbf{e}^\top)$. Note that the term $\tilde{\epsilon}(\cdot)$ no longer corresponds to the permittivity in the classical sense but only to its instantaneous part. Substituting the relations (1.54) and (1.55) into Maxwell's equations (1.51) and rearranging the terms leads to the system

$$\tilde{\epsilon}(\mathbf{e})\partial_t\mathbf{e} = \operatorname{curl}\mathbf{h} - \epsilon_0\mathbf{j}, \tag{1.56}$$

$$\mu_0\partial_t\mathbf{h} = -\operatorname{curl}\mathbf{e}, \tag{1.57}$$

$$\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\partial_t\mathbf{p} = \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{j}, \tag{1.58}$$

$$\tfrac{\epsilon_0}{\omega_p^2}\partial_t\mathbf{j} = \epsilon_0\mathbf{e} - \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p} - \tfrac{\epsilon_0\gamma}{\omega_p^2}\mathbf{j}, \tag{1.59}$$

which we call the $\mathbf{e}-\mathbf{h}$ formulation for the Kerr-Lorentz model. The choice of multiplication factors in equations (1.58) and (1.59) is again motivated by the energy structure of the problem. Similarly to the $\mathbf{e}-\mathbf{h}$ formulation of Section 1.2.1, the coefficients on the left-hand side of the system relate to the derivatives of the energy density functions, as will become clear below. Since the factors are constant, their choice is only for illustration purposes. In the following analysis, we again consider the problem in the bounded domain $\Omega$ and impose perfect magnetic boundary conditions

$$\mathbf{n}\times\mathbf{h} = 0 \quad \text{on } \partial\Omega. \tag{1.60}$$

The generalization to other types of boundary conditions is straightforward.

### Electromagnetic energy

Following [18] and using the notation from the previous section, the energy for the Kerr-Lorentz problem (1.56)–(1.59) can be expressed as follows

$$\mathcal{E}(\mathbf{e},\mathbf{h},\mathbf{p},\mathbf{j}) = \int_\Omega \tilde{\omega}_{mag}(\mathbf{h}) + \tilde{\omega}_{el}(\mathbf{e},\mathbf{p},\mathbf{j})\,dx.$$

The magnetic energy density is given by $\tilde{\omega}_{mag}(\mathbf{h}) = \frac{\mu_0}{2}|\mathbf{h}|^2$, as previously, and the electric energy density function for this problem is given by

$$\tilde{\omega}_{el}(\mathbf{e},\mathbf{p},\mathbf{j}) = \tfrac{1}{2}\left(\epsilon_0\epsilon_\infty|\mathbf{e}|^2 + \tfrac{3\epsilon_0\alpha}{2}|\mathbf{e}|^4 + \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}|\mathbf{p}|^2 + \tfrac{\epsilon_0}{\omega_p^2}|\mathbf{j}|^2\right), \tag{1.61}$$

which is a strongly convex function with respect to the argument. The differentiation of the energy densities leads to expressions $\tilde{\omega}'_{mag}(\mathbf{h}) = \mu_0\mathbf{h}$ and $\tilde{\omega}'_{el}(\mathbf{e},\mathbf{p},\mathbf{j}) = (\tilde{\epsilon}(\mathbf{e})\mathbf{e}, \frac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}, \frac{\epsilon_0}{\omega_p^2}\mathbf{j})$, which are important relations for our further analysis.

Let us note that the auxiliary differential equation representation of the polarization term (1.53) is essential for our further analysis. In general, the constitutive relations for dispersive materials are not instantaneous. Therefore, the concept of energy-based modeling as discussed in Section 1.1 cannot be directly applied in this setting. The generalization of this concept to dispersive and memory-dependent media is not yet fully understood and represents a topic for further investigation.

### Power balance

We now again formulate the variational representation of the solution and show that it implies the power balance, which is the basis of our strategy.

**Lemma 1.3.1.** Let $(\mathbf{e},\mathbf{h},\mathbf{p},\mathbf{j})$ be a smooth solution of (1.56)–(1.59). Then, the identities

$$\langle\tilde{\epsilon}(\mathbf{e}(t))\partial_t\mathbf{e}(t),\mathbf{v}\rangle = \langle\mathbf{h}(t),\operatorname{curl}\mathbf{v}\rangle - \langle\epsilon_0\mathbf{j}(t),\mathbf{v}\rangle, \tag{1.62}$$

$$\langle\mu_0\partial_t\mathbf{h}(t),\mathbf{w}\rangle = -\langle\operatorname{curl}\mathbf{e}(t),\mathbf{w}\rangle, \tag{1.63}$$

$$\langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\partial_t\mathbf{p}(t),\mathbf{z}\rangle = \langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{j}(t),\mathbf{z}\rangle, \tag{1.64}$$

$$\langle\tfrac{\epsilon_0}{\omega_p^2}\partial_t\mathbf{j}(t),\mathbf{q}\rangle = \langle\epsilon_0\mathbf{e}(t) - \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}(t) - \tfrac{\epsilon_0\gamma}{\omega_p^2}\mathbf{j}(t),\mathbf{q}\rangle, \tag{1.65}$$

hold for all $\mathbf{v},\mathbf{z},\mathbf{q} \in H(\operatorname{curl},\Omega)$, $\mathbf{w} \in L^2(\Omega)$, and $t \geq 0$. Moreover, it holds

$$\frac{d}{dt}\mathcal{E}(\mathbf{e}(t),\mathbf{h}(t),\mathbf{p}(t),\mathbf{j}(t)) = -\tfrac{\epsilon_0}{\tau\omega_p^2}\|\mathbf{j}(t)\|^2 \leq 0. \tag{1.66}$$

Therefore, the energy does not increase in time, and the formulation is passive.

*Proof.* The proof of the variational equalities is analogous to that of Lemma 1.2.1. We multiply the system by test functions $\mathbf{v},\mathbf{z},\mathbf{q} \in H(\operatorname{curl},\Omega)$ and $\mathbf{w} \in L^2(\Omega)$, and integrate over the domain $\Omega$. The relation (1.62) follows from the integration by parts formula, where the boundary term vanishes due to the choice of the boundary condition (1.60).

The proof of the power balance is based on the same arguments as used in the proof of Lemma 1.2.2. Differentiating the energy with respect to time yields

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(\mathbf{e}(t),\mathbf{h}(t),\mathbf{p}(t),\mathbf{j}(t)) &= \langle\tilde{\epsilon}(\mathbf{e}(t))\mathbf{e}(t),\partial_t\mathbf{e}(t)\rangle + \langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}(t),\partial_t\mathbf{p}(t)\rangle + \langle\tfrac{\epsilon_0}{\omega_p^2}\mathbf{j}(t),\partial_t\mathbf{j}(t)\rangle\\ &= \langle\tilde{\epsilon}(\mathbf{e}(t))\partial_t\mathbf{e}(t),\mathbf{e}(t)\rangle + \langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\partial_t\mathbf{p}(t),\mathbf{p}(t)\rangle + \langle\tfrac{\epsilon_0}{\omega_p^2}\partial_t\mathbf{j}(t),\mathbf{j}(t)\rangle = (*).\end{aligned}$$

Using the variational principle (1.62)–(1.65) with test functions $\mathbf{v} = \mathbf{e}(t)$, $\mathbf{w} = \mathbf{h}(t)$, $\mathbf{z} = \mathbf{p}(t)$, and $\mathbf{q} = \mathbf{j}(t)$, we further obtain

$$\begin{aligned}(*) &= \langle\mathbf{h}(t),\operatorname{curl}\mathbf{e}(t)\rangle - \langle\epsilon_0\mathbf{j}(t),\mathbf{e}(t)\rangle - \langle\operatorname{curl}\mathbf{e}(t),\mathbf{h}(t)\rangle + \langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{j}(t),\mathbf{p}(t)\rangle\\ &\qquad + \langle\epsilon_0\mathbf{e}(t),\mathbf{j}(t)\rangle - \langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}(t),\mathbf{j}(t)\rangle - \langle\tfrac{\epsilon_0\gamma}{\omega_p^2}\mathbf{j}(t),\mathbf{j}(t)\rangle\\ &= -\langle\tfrac{\epsilon_0\gamma}{\omega_p^2}\mathbf{j}(t),\mathbf{j}(t)\rangle = -\tfrac{\epsilon_0\gamma}{\omega_p^2}\|\mathbf{j}(t)\|^2.\end{aligned}$$

Integration in time provides the corresponding energy balance and proves passivity. □

### Structure

Similar to the $\mathbf{e}-\mathbf{h}$ formulation discussed in Section 1.2.1, the key ingredient in derivation of power balance (1.66) is the use of variational equalities (1.62)–(1.65) with the solution $\mathbf{v} = \mathbf{e}(t)$, $\mathbf{w} = \mathbf{h}(t)$, $\mathbf{z} = \mathbf{p}(t)$, and $\mathbf{q} = \mathbf{j}(t)$. Therefore, one can directly conclude that the formulation has the canonical port-Hamiltonian structure

$$\mathcal{Q}^*(\mathbf{u})\partial_t\mathbf{u} = -\mathcal{A}(\mathbf{u}), \qquad \mathcal{E}'(\mathbf{u}) = Q(\mathbf{u})\mathbf{u}.$$

Thus, we may once again employ the framework from [42] for constructing discretization schemes that preserve the passivity of the system, in analogy to Section 1.2.2.

### Discretization

As before, we write $W_h \subset H(\operatorname{curl},\Omega)$ and $Q_h \subset L^2(\Omega)$ for some finite-dimensional subspaces, and we use the notation for the time discretization from Section 1.2.2. For the discretization of problem (1.56)–(1.60), we consider the following approach based on Galerkin approximation in space and discontinuous Galerkin method in time.

**Problem 1.3.2.** Let the initial values $\mathbf{e}_h^0, \mathbf{p}_h^0, \mathbf{j}_h^0 \in W_h$ and $\mathbf{h}_h^0 \in Q_h$ be given. Then for $1 \le n \le N$ find $\mathbf{e}_h^n, \mathbf{p}_h^n, \mathbf{p}_h^n \in P_k(I^n; W_h)$ and $\mathbf{h}_h^n \in P_k(I^n; Q_h)$ such that

$$\int_{I^n}\langle\tilde{\epsilon}(\mathbf{e}_h^n)\partial_t\mathbf{e}_h^n + \epsilon_0\mathbf{j}_h^n,\mathbf{v}_h\rangle - \langle\mathbf{h}_h^n,\operatorname{curl}\mathbf{v}_h\rangle\,dt = \langle\tilde{\epsilon}(\mathbf{e}_h^n)(\mathbf{e}_h^{n-1}-\mathbf{e}_h^n),\mathbf{v}_h\rangle|_{t^{n-1}}, \quad (1.67)$$

$$\int_{I^n}\langle\mu_0\partial_t\mathbf{h}_h^n + \operatorname{curl}\mathbf{e}_h^n,\mathbf{w}_h\rangle\,dt = \langle\mu_0(\mathbf{h}_h^{n-1}-\mathbf{h}_h^n),\mathbf{w}_h\rangle|_{t^{n-1}}, \quad (1.68)$$

$$\int_{I^n}\langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\partial_t\mathbf{p}_h^n - \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{j}_h^n,\mathbf{z}_h\rangle\,dt = \langle\tfrac{\epsilon_0\omega_0^2}{\omega_p^2}(\mathbf{p}_h^{n-1}-\mathbf{p}_h^n),\mathbf{z}_h\rangle|_{t^{n-1}}, \quad (1.69)$$

$$\int_{I^n}\langle\tfrac{\epsilon_0}{\omega_p^2}\partial_t\mathbf{j}_h^n - \epsilon_0\mathbf{e}_h^n + \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}_h^n + \tfrac{\epsilon_0\gamma}{\omega_p^2}\mathbf{j}_h^n,\mathbf{q}_h\rangle\,dt = \langle\tfrac{\epsilon_0}{\omega_p^2}(\mathbf{j}_h^{n-1}-\mathbf{j}_h^n),\mathbf{q}_h\rangle|_{t^{n-1}}, \quad (1.70)$$

holds for all test functions $\mathbf{v}_h, \mathbf{z}_h, \mathbf{q}_h \in P_k(I^n; W_h)$ and $\mathbf{w}_h \in P_k(I^n; Q_h)$.

**Lemma 1.3.3.** Let $(\mathbf{e}_h^n, \mathbf{h}_h^n, \mathbf{p}_h^n, \mathbf{j}_h^n)_{n\geq 0}$ denote the solution of (1.67)–(1.70). Then

$$\mathcal{E}_h^n(t^n) - \mathcal{E}_h^m(t^m) \leq -\tfrac{\epsilon_0 \tau}{\omega_p^2} \int_{t^m}^{t^n} \|\mathbf{j}_h^N(\tau)\|^2 \, d\tau \leq 0,$$

holds for $0 \leq m \leq n \leq N$, where by $\mathbf{j}_h^N \in P_k(\mathcal{T}; W_h)$, $\mathbf{j}_h^N|_{I^n} = \mathbf{j}_h^n$ we denote the global solution in time and we use the abbreviation $\mathcal{E}_h^k(t) = \mathcal{E}(\mathbf{e}_h^k(t), \mathbf{h}_h^k(t), \mathbf{p}_h^k(t), \mathbf{j}_h^k(t))$ for $t \in I^k$.

*Proof.* The proof of the statement is almost identical to that of Lemma 1.2.4. We start by decomposing the evolution of energy after one time step as

$$\mathcal{E}_h^n(t^n) - \mathcal{E}_h^{n-1}(t^{n-1}) = \big(\mathcal{E}_h^n(t^n) - \mathcal{E}_h^n(t^{n-1})\big) + \big(\mathcal{E}_h^n(t^{n-1}) - \mathcal{E}_h^{n-1}(t^{n-1})\big) = (i) + (ii).$$

Using the fundamental theorem of calculus, the first term can be written as

$$\begin{aligned}(i) = \int_{I^n} \frac{d}{dt}\mathcal{E}_h^n(t)\, dt = \int_{I^n} &\langle \tilde{\epsilon}(\mathbf{e}_h^n(t))\partial_t \mathbf{e}_h^n(t), \mathbf{e}_h^n(t)\rangle + \langle \mu_0 \partial_t \mathbf{h}_h^n(t), \mathbf{h}_h^n(t)\rangle \\ &+ \langle \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\partial_t \mathbf{p}_h^n(t), \mathbf{p}_h^n(t)\rangle + \langle \tfrac{\epsilon_0}{\omega_p^2}\partial_t \mathbf{j}_h^n(t), \mathbf{j}_h^n(t)\rangle \, dt.\end{aligned}$$

Using the variational scheme (1.67)–(1.70) with test functions $\mathbf{v}_h = \mathbf{e}_h^n$, $\mathbf{w}_h = \mathbf{h}_h^n$, $\mathbf{z}_h = \mathbf{p}_h^n$, and $\mathbf{q}_h = \mathbf{j}_h^n$, which are admissible test functions, we conclude

$$\begin{aligned}(i) = -\int_{I^n} \langle \tfrac{\epsilon_0\gamma}{\omega_p^2}\mathbf{j}_h^n(t), \mathbf{j}_h^n(t)\rangle \, dt &+ \langle \tilde{\epsilon}(\mathbf{e}_h^n)(\mathbf{e}_h^{n-1} - \mathbf{e}_h^n), \mathbf{e}_h^n\rangle|_{t^{n-1}} + \langle \mu_0(\mathbf{h}_h^{n-1} - \mathbf{h}_h^n), \mathbf{h}_h^n\rangle|_{t^{n-1}} \\ &+ \langle \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}(\mathbf{p}_h^{n-1} - \mathbf{p}_h^n), \mathbf{p}_h^n\rangle|_{t^{n-1}} + \langle \tfrac{\epsilon_0}{\omega_p^2}(\mathbf{j}_h^{n-1} - \mathbf{j}_h^n), \mathbf{j}_h^n\rangle|_{t^{n-1}}.\end{aligned}$$

Lastly, we use the convexity of the energy denisties and obtain the following inequality

$$\begin{aligned}(ii) &\leq \langle \tilde{\epsilon}(\mathbf{e}_h^n)\mathbf{e}_h^n, \mathbf{e}_h^n - \mathbf{e}_h^{n-1}\rangle|_{t^{n-1}} + \langle \mu_0 \mathbf{h}_h^n, \mathbf{h}_h^n - \mathbf{h}_h^{n-1}\rangle|_{t^{n-1}} \\ &\quad + \langle \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}_h^n, \mathbf{p}_h^n - \mathbf{p}_h^{n-1}\rangle|_{t^{n-1}} + \langle \tfrac{\epsilon_0}{\omega_p^2}\mathbf{j}_h^n, \mathbf{j}_h^n - \mathbf{j}_h^{n-1}\rangle|_{t^{n-1}} \\ &= \langle \tilde{\epsilon}(\mathbf{e}_h^n)(\mathbf{e}_h^n - \mathbf{e}_h^{n-1}), \mathbf{e}_h^n\rangle|_{t^{n-1}} + \langle \mu_0(\mathbf{h}_h^n - \mathbf{h}_h^{n-1}), \mathbf{h}_h^n\rangle|_{t^{n-1}} \\ &\quad + \langle \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}(\mathbf{p}_h^n - \mathbf{p}_h^{n-1}), \mathbf{p}_h^n\rangle|_{t^{n-1}} + \langle \tfrac{\epsilon_0}{\omega_p^2}(\mathbf{j}_h^n - \mathbf{j}_h^{n-1}), \mathbf{j}_h^n\rangle|_{t^{n-1}}\end{aligned}$$

Adding $(i)$ and $(ii)$ together proves the statement for $m = n-1$, whereas the case $m < n-1$ follows by induction.

□

### Remark on possible numerical realization

The spatial discretization of the problem (1.62)–(1.65) by an appropriate mixed finite element method leads to the system of differential equations

$$\begin{pmatrix} \mathsf{M_e}(\mathsf{e}) & 0 & 0 & 0 \\ 0 & \mathsf{M_h} & 0 & 0 \\ 0 & 0 & \mathsf{M_p} & 0 \\ 0 & 0 & 0 & \mathsf{M_j} \end{pmatrix} \partial_t \begin{pmatrix} \mathsf{e} \\ \mathsf{h} \\ \mathsf{p} \\ \mathsf{j} \end{pmatrix} = \begin{pmatrix} 0 & \mathsf{C} & 0 & -\mathsf{M}_{\epsilon_0} \\ -\mathsf{C}^\top & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathsf{M_p} \\ \mathsf{M}_{\epsilon_0} & 0 & -\mathsf{M_p} & \mathsf{M_d} \end{pmatrix} \begin{pmatrix} \mathsf{e} \\ \mathsf{h} \\ \mathsf{p} \\ \mathsf{j} \end{pmatrix}$$

for the coefficients vectors $\mathsf{e}(t)$, $\mathsf{h}(t)$, $\mathsf{p}(t)$, and $\mathsf{j}(t)$. With $\mathsf{M}_*$ we denote the mass matrices with different coefficients in accordance to (1.56)–(1.59). The problem can be again written in the abstract form

$$\mathsf{Q}(\mathsf{u}(t))^\top \partial_t \mathsf{u}(t) = -\mathsf{A}(\mathsf{u}(t))$$

with $\mathsf{u} = (\mathsf{e}, \mathsf{h}, \mathsf{p}, \mathsf{j})$ and $\mathsf{Q}(\mathsf{u})$ and $\mathsf{A}(\mathsf{u})$ accordingly. The implementation of the discontinuous Galerkin time-stepping method for Kerr-media, as discussed in Section 1.2.2, can be directly applied in this context.

Next, we present the formulation based on the vector potential and discuss its discretization. The proposed discretization strategy allows the construction of schemes that preserve the energy balance exactly. This might be of a particular advantage in applications where the loss term is neglected, i.e., when $\gamma = 0$, as considered in, for example, [19, 126].

### 1.3.2. The $\mathbf{e} - \mathbf{a}$ formulation for Kerr-Lorentz problem

The construction of the $\mathbf{e}-\mathbf{a}$ formulation works with similar arguments as in Section 1.2.3. By using the relations $\partial_t \mathbf{p} = \mathbf{j}$ and $\mathbf{e} = -\partial_t \mathbf{a}$, equation (1.53) can be written as

$$\partial_t \mathbf{j} + \gamma \partial_t \mathbf{p} + \omega_0^2 \mathbf{p} = -\omega_p^2 \partial_t \mathbf{a}. \tag{1.71}$$

Substituting the relations $\partial_t \mathbf{d} = \tilde{\epsilon}(\mathbf{e})\partial_t \mathbf{e} + \epsilon_0 \partial_t \mathbf{p}$ and $\mathbf{h} = \nu_0 \operatorname{curl} \mathbf{a}$ into Faraday's law (1.51) and using the relation (1.71), we obtain the following formulation

$$-\tilde{\epsilon}(\mathbf{e})\partial_t \mathbf{a} = \tilde{\epsilon}(\mathbf{e})\mathbf{e}, \tag{1.72}$$

$$\tilde{\epsilon}(\mathbf{e})\partial_t \mathbf{e} + \epsilon_0 \partial_t \mathbf{p} = \operatorname{curl} \nu_0 \operatorname{curl} \mathbf{a}, \tag{1.73}$$

$$-\epsilon_0 \partial_t \mathbf{a} - \tfrac{\epsilon_0 \gamma}{\omega_p^2} \partial_t \mathbf{p} - \tfrac{\epsilon_0}{\omega_p^2} \partial_t \mathbf{j} = \tfrac{\epsilon_0 \omega_0^2}{\omega_p^2} \mathbf{p}, \tag{1.74}$$

$$\tfrac{\epsilon_0}{\omega_p^2} \partial_t \mathbf{p} = \tfrac{\epsilon_0}{\omega_p^2} \mathbf{j}. \tag{1.75}$$

The choice of the multiplication factors in (1.74) and (1.75) is again motivated by the energy-based structure of the problem. The factors are chosen such that the terms on the right-hand side of the system correspond to the derivatives of the energy densities. The perfect magnetic boundary condition (1.60) then translates to

$$\mathbf{n} \times \nu_0 \operatorname{curl} \mathbf{a} = 0 \quad \text{on } \partial\Omega. \tag{1.76}$$

Let us mention that this formulation bears a close resemblance to the $\mathbf{a} - \mathbf{d}$ formulation recently proposed in [84]. However, a precise comparative study is yet to be done.

#### Electromagnetic energy and energy balance

By analogy to Section 1.2.3, we now denote the energy in terms of system variables by

$$\mathcal{H}(\mathbf{e}, \mathbf{a}, \mathbf{p}, \mathbf{j}) = \int_\Omega \omega_{mag}(\operatorname{curl} \mathbf{a}) + \tilde{\omega}_{el}(\mathbf{e}, \mathbf{p}, \mathbf{j}) \, dx,$$

where the expression for the electric energy density $\tilde{\omega}_{el}(\mathbf{e}, \mathbf{p}, \mathbf{j})$ is the same as in (1.61), while the magnetic energy density is now given by $\omega_{mag}(\operatorname{curl} \mathbf{a}) = \frac{\nu_0}{2} |\operatorname{curl} \mathbf{a}|^2$. By analogy to previous sections, we now formulate the variational principles for the $\mathbf{e} - \mathbf{a}$ formulation (1.72)–(1.76) and derive the power balance, which is the foundation for the discretization technique.

**Lemma 1.3.4.** Let $(\mathbf{e}, \mathbf{a}, \mathbf{p}, \mathbf{j})$ be a smooth solution of (1.72)–(1.75). Then

$$-\langle \tilde{\epsilon}(\mathbf{e}(t))\partial_t \mathbf{a}(t), \mathbf{v}\rangle = \langle \tilde{\epsilon}(\mathbf{e}(t))\mathbf{e}(t), \mathbf{v}\rangle, \tag{1.77}$$

$$\langle \tilde{\epsilon}(\mathbf{e}(t))\partial_t \mathbf{e}(t) + \epsilon_0 \partial_t \mathbf{p}(t), \mathbf{w}\rangle = \langle \nu_0 \operatorname{curl} \mathbf{a}(t), \operatorname{curl} \mathbf{w}\rangle, \tag{1.78}$$

$$-\langle \epsilon_0 \partial_t \mathbf{a}(t) + \tfrac{\epsilon_0 \gamma}{\omega_p^2}\partial_t \mathbf{p}(t) + \tfrac{\epsilon_0}{\omega_p^2}\partial_t \mathbf{j}(t), \mathbf{z}\rangle = \langle \tfrac{\epsilon_0 \omega_0^2}{\omega_p^2}\mathbf{p}(t), \mathbf{z}\rangle, \tag{1.79}$$

$$\langle \tfrac{\epsilon_0}{\omega_p^2}\partial_t \mathbf{p}(t), \mathbf{q}\rangle = \langle \tfrac{\epsilon_0}{\omega_p^2}\mathbf{j}(t), \mathbf{q}\rangle, \tag{1.80}$$

holds for all $\mathbf{v}, \mathbf{w}, \mathbf{z}, \mathbf{q} \in H(\operatorname{curl}, \Omega)$, $t \geq 0$. Moreover, the following power balance holds

$$\frac{d}{dt}\mathcal{H}(\mathbf{e}(t), \mathbf{a}(t), \mathbf{p}(t), \mathbf{j}(t)) = -\tfrac{\epsilon_0 \gamma}{\omega_p^2}\|\partial_t \mathbf{p}(t)\|^2 \, d\tau \leq 0. \tag{1.81}$$

Therefore, the energy does not increase over time and the formulation is passive.

*Proof.* The proof uses a similar approach as the proofs of Lemma 1.2.10 and Lemma 1.2.11. The variational identities follow directly from multiplying with test functions, integrating over the domain, and utilizing the integration by parts formula, where the boundary term disappears due to the choice of the boundary condition (1.76). For the proof of the second statement, we again use the expression for power

$$\begin{aligned}\frac{d}{dt}\mathcal{H}(\mathbf{e}(t), \mathbf{a}(t), \mathbf{p}(t), \mathbf{j}(t)) &= \langle \tilde{\epsilon}(\mathbf{e}(t))\mathbf{e}(t), \partial_t \mathbf{e}(t)\rangle + \langle \nu_0 \operatorname{curl} \mathbf{a}(t), \operatorname{curl} \partial_t \mathbf{a}(t)\rangle \\ &\quad + \langle \tfrac{\epsilon_0 \omega_0^2}{\omega_p^2}\mathbf{p}(t), \partial_t \mathbf{p}(t)\rangle + \langle \tfrac{\epsilon_0}{\omega_p^2}\mathbf{j}(t), \mathbf{q}\rangle = (*).\end{aligned}$$

Then, using the variational identities (1.77)–(1.80) with $\mathbf{v} = \partial_t \mathbf{e}(t)$, $\mathbf{w} = \partial_t \mathbf{a}$, $\mathbf{z} = \partial_t \mathbf{p}$, and $\mathbf{q} = \partial_t \mathbf{j}$, we conclude

$$\begin{aligned}(*) &= -\langle \tilde{\epsilon}(\mathbf{e}(t))\partial_t \mathbf{a}(t), \partial_t \mathbf{e}(t)\rangle + \langle \tilde{\epsilon}(\mathbf{e}(t))\partial_t \mathbf{e}(t), \partial_t \mathbf{a}(t)\rangle + \langle \epsilon_0 \partial_t \mathbf{p}(t), \partial_t \mathbf{a}\rangle \\ &\quad - \langle \epsilon_0 \partial_t \mathbf{a}(t), \partial_t \mathbf{p}(t)\rangle - \langle \tfrac{\epsilon_0 \gamma}{\omega_p^2}\partial_t \mathbf{p}(t), \partial_t \mathbf{p}(t)\rangle - \langle \tfrac{\epsilon_0}{\omega_p^2}\partial_t \mathbf{j}(t), \partial_t \mathbf{p}(t)\rangle + \langle \tfrac{\epsilon_0}{\omega_p^2}\partial_t \mathbf{p}(t), \partial_t \mathbf{j}(t)\rangle \\ &= -\tfrac{\epsilon_0 \gamma}{\omega_p^2}\|\partial_t \mathbf{p}(t)\|^2.\end{aligned}$$

Finally, integration in time proves the energy decay statement. □

### Structure

The key ingredient in the derivation of the power balance (1.81) is again the use of variational equalities (1.77)–(1.80) with the time derivatives of the solution as test function. There, we can conclude that the problem has the structure

$$C(\mathbf{u})\partial_t \mathbf{u} = -\mathcal{H}'(\mathbf{u})$$

Thus, we may again utilize the framework [43] for the construction of discretization schemes that preserve the energy balance of the system, in analogy to Section 1.2.4.

### Discretization

For the discretization of the system (1.72)–(1.75), we now consider the following method based on the Galerkin discretization in space and the Petrov-Galerkin method in time.

**Problem 1.3.5.** Let the initial values $\mathbf{e}_h^0, \mathbf{a}_h^0, \mathbf{p}_h^0, \mathbf{j}_h^0 \in W_h$ be given. Then for $1 \le n \le N$ find $\mathbf{e}_h^n, \mathbf{a}_h^n, \mathbf{p}_h^n, \mathbf{j}_h^n \in P_{k+1}(I^n; W_h)$ with $\mathbf{e}_h^n(t^{n-1}) = \mathbf{e}_h^{n-1}(t^{n-1})$, $\mathbf{a}_h^n(t^{n-1}) = \mathbf{a}_h^{n-1}(t^{n-1})$, $\mathbf{p}_h^n(t^{n-1}) = \mathbf{p}_h^{n-1}(t^{n-1})$, and $\mathbf{j}_h^n(t^{n-1}) = \mathbf{j}_h^{n-1}(t^{n-1})$ such that

$$-\int_{I^n} \langle \tilde{\epsilon}(\mathbf{e}_h^n)\partial_t \mathbf{a}_h^n, \tilde{\mathbf{v}}_h\rangle \, dt = \int_{I^n} \langle \tilde{\epsilon}(\mathbf{e}_h^n)\mathbf{e}_h^n, \tilde{\mathbf{v}}_h\rangle \, dt, \tag{1.82}$$

$$\int_{I^n} \langle \tilde{\epsilon}(\mathbf{e}_h^n)\partial_t \mathbf{e}_h^n + \epsilon_0 \partial_t \mathbf{p}_h^n, \tilde{\mathbf{w}}_h\rangle \, dt = \int_{I^n} \langle \nu_0 \operatorname{curl} \mathbf{a}_h^n, \operatorname{curl} \tilde{\mathbf{w}}_h\rangle \, dt, \tag{1.83}$$

$$-\int_{I^n} \langle \epsilon_0 \partial_t \mathbf{a}_h^n + \tfrac{\epsilon_0\gamma}{\omega_p^2}\partial_t \mathbf{p}_h^n + \tfrac{\epsilon_0}{\omega_p^2}\partial_t \mathbf{j}_h^n, \tilde{\mathbf{z}}_h\rangle \, dt = \int_{I^n} \langle \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}_h^n, \tilde{\mathbf{z}}_h\rangle \, dt, \tag{1.84}$$

$$\int_{I^n} \langle \tfrac{\epsilon_0}{\omega_p^2}\partial_t \mathbf{p}_h^n, \tilde{\mathbf{q}}_h\rangle \, dt = \int_{I^n} \langle \tfrac{\epsilon_0}{\omega_p^2}\mathbf{j}_h^n, \tilde{\mathbf{q}}_h\rangle \, dt, \tag{1.85}$$

holds for all test functions $\tilde{\mathbf{v}}_h, \tilde{\mathbf{w}}_h, \tilde{\mathbf{z}}_h, \tilde{\mathbf{q}}_h \in P_k(I^n; W_h)$.

**Lemma 1.3.6.** Let $(\mathbf{e}_h^n, \mathbf{a}_h^n, \mathbf{p}_h^n, \mathbf{j}_h^n)$ denote the solution of (1.82)–(1.85). Then

$$\mathcal{H}_h^n(t^n) - \mathcal{H}_h^m(t^m) = -\tfrac{\epsilon_0\gamma}{\omega_p^2}\int_{t^m}^{t^n} \|\partial_t \mathbf{p}_h(\tau)\|^2 \, d\tau \le 0,$$

holds for all $0 \le m \le n$, where $\mathbf{p}_h^N \in P_{k+1}(\mathcal{T}; W_h)$, $\mathbf{p}_h^N|_{I^n} = \mathbf{p}_h^n$ denotes the solution global in time and we use the abbreviation $\mathcal{H}_h^k(t) = \mathcal{H}(\mathbf{e}_h^k(t), \mathbf{a}_h^k(t), \mathbf{p}_h^k(t), \mathbf{j}_h^k(t))$ for $t \in I^k$.

*Proof.* The proof works by analogy to that of Lemma 1.2.13 and that of [43, Theorem 2]. By the fundamental theorem of calculus we obtain

$$\begin{aligned}\mathcal{H}_h^n(t^n) - \mathcal{H}_h^n(t^{n-1}) &= \int_{I^n} \frac{d}{dt}\mathcal{H}_h^n(t)\, dt = \int_{I^n} \langle \tilde{\epsilon}(\mathbf{e}_h^n(t))\mathbf{e}_h^n(t), \partial_t \mathbf{e}_h^n(t)\rangle \\ &\quad + \langle \nu_0 \operatorname{curl} \mathbf{a}_h^n(t), \operatorname{curl} \partial_t \mathbf{a}_h^n(t)\rangle + \langle \tfrac{\epsilon_0\omega_0^2}{\omega_p^2}\mathbf{p}_h^n(t), \partial_t \mathbf{p}_h^n(t)\rangle + \tfrac{\epsilon_0}{\omega_p^2}\langle \mathbf{j}_h^n(t), \partial_t \mathbf{j}_h^n(t)\rangle \, dt = (*).\end{aligned}$$

Next, we use the scheme (1.82)–(1.85) with $\tilde{\mathbf{v}}_h = \partial_t \mathbf{e}_h^n(t)$, $\tilde{\mathbf{w}}_h = \partial_t \mathbf{a}_h^n(t)$, $\tilde{\mathbf{z}}_h = \partial_t \mathbf{p}_h^n(t)$, and $\tilde{\mathbf{q}}_h = \partial_t \mathbf{j}_h^n(t)$, which are admissible test functions, and obtain

$$(*) = -\int_{I^n} \tfrac{\epsilon_0\gamma}{\omega_p^2}\langle \partial_t \mathbf{p}_h^n(t), \partial_t \mathbf{p}_h^n(t)\rangle \, dt = -\tfrac{\epsilon_0\gamma}{\omega_p^2}\int_{I^n} \|\partial_t \mathbf{p}_h^n(t)\|^2.$$

The continuity of the solution at the junctions of the time intervals proves the statement for $m = n-1$. The general case $m < n-1$ follows by induction. □

### Remarks on possible numerical realization

Discretization in space of the problem (1.77)–(1.80) using a suitable finite element method leads to the system of differential equations

$$\begin{pmatrix} 0 & \mathsf{M}_{\mathsf{e}}(\mathsf{e}) & 0 & 0 \\ -\mathsf{M}_{\mathsf{e}}(\mathsf{e}) & 0 & -\mathsf{M}_{\epsilon_0} & 0 \\ 0 & \mathsf{M}_{\epsilon_0} & \mathsf{M}_{\mathsf{d}} & \mathsf{M}_{\mathsf{j}} \\ 0 & 0 & -\mathsf{M}_{\mathsf{j}} & 0 \end{pmatrix} \partial_t \begin{pmatrix} \mathsf{e} \\ \mathsf{a} \\ \mathsf{p} \\ \mathsf{j} \end{pmatrix} = - \begin{pmatrix} \mathsf{M}_{\mathsf{e}}(\mathsf{e}) & 0 & 0 & 0 \\ 0 & \mathsf{K}_{\mathsf{a}} & 0 & 0 \\ 0 & 0 & \mathsf{M}_{\mathsf{p}} & 0 \\ 0 & 0 & 0 & \mathsf{M}_{\mathsf{j}} \end{pmatrix} \begin{pmatrix} \mathsf{e} \\ \mathsf{a} \\ \mathsf{p} \\ \mathsf{j} \end{pmatrix}, \tag{1.86}$$

where $\mathsf{e}(t)$, $\mathsf{a}(t)$, $\mathsf{p}(t)$, and $\mathsf{j}(t)$ are the coefficient vectors and $\mathsf{M}_*$ are the same mass matrices as in Section 1.3.1. This problem can be compactly written in the abstract form

$$(\mathsf{C}(\mathsf{u}(t))\partial_t \mathsf{u}(t) = -\mathsf{H}'(\mathsf{u}(t)),$$

where $\mathsf{u} = (\mathsf{e}, \mathsf{a}, \mathsf{p}, \mathsf{j})$ and $\mathsf{C}(\mathsf{u})$ and $\mathsf{H}'(\mathsf{u})$ are given in accordance to (1.86). Petrov-Galerkin time-stepping can be implemented in the same manner as for the $\mathbf{e} - \mathbf{a}$ formulation for Kerr-media, as discussed in Section 1.2.4.

### 1.3.3. Numerical illustration

In this section, we present a simple one-dimensional example. Similarly to Section 1.2.5, we consider the propagation of the Gaussian pulse and demonstrate the effect of dispersion. Then, we provide the convergence results and briefly discuss further observations.

**Discretization details.** We use the same discretization as for the one-dimensional example discussed in Section 1.2.5. We use Lagrange polynomials on Lobatto nodes as the basis for the semi-discretization, and we again utilize inexact integration in space. For the time-stepping, we use Lagrange polynomials on Lobatto nodes as the basis and we utilize higher order Lobatto quadrature rule for the evaluation of the time integrals.

**Simulation results.** For simplicity, we simply set $\epsilon_0 = \epsilon_\infty = \alpha = \omega_0 = \tau = 1$, $\omega_p = 5$. Snapshots of the electric field are illustrated in Figure 1.6. For comparison, we also plot the solution to the Kerr problem without polarization by the black dashed line. The results are obtained using the lowest order scheme (1.82)–(1.85) with $h = 0.01$ and $\tau = 0.01$. From this figure, one can directly observe the impact of the dispersion. Due to the memory effect, the propagation of the impulse leaves the polarization behind, which slowly oscillates and diminishes through damping. If the loss terms are neglected, it results in the formation of so-called *kink-antikink* solutions [19, 126].

| $h$ | $p=1$ | | $p=2$ | | $p=3$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-2}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. |
| 0.05 | 0.152394 | — | 0.179533 | — | 0.174497 | — |
| 0.025 | 0.039763 | 1.94 | 0.014451 | 3.63 | 0.007996 | 4.44 |
| 0.0125 | 0.010038 | 1.98 | 0.001788 | 3.14 | 0.000483 | 4.04 |
| 0.00625 | 0.002515 | 1.99 | 0.000223 | 3.01 | 0.000031 | 4.01 |

Table 1.6.: Convergence in space of the method based on the $\mathbf{e} - \mathbf{h}$ formulation.

| $\tau$ | $k=0$ | | $k=1$ | | $k=2$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. | err $\times 10^{-4}$ | e.o.c. |
| 0.025 | 0.152881 | — | 0.604931 | — | 0.366292 | — |
| 0.0125 | 0.107546 | 0.51 | 0.082113 | 2.88 | 0.012139 | 4.91 |
| 0.00625 | 0.066889 | 0.68 | 0.010397 | 2.98 | 0.000385 | 4.97 |
| 0.003125 | 0.040212 | 0.73 | 0.001311 | 2.99 | 0.000012 | 4.99 |

Table 1.7.: Convergence in time of the method based on the $\mathbf{e} - \mathbf{h}$ formulation.

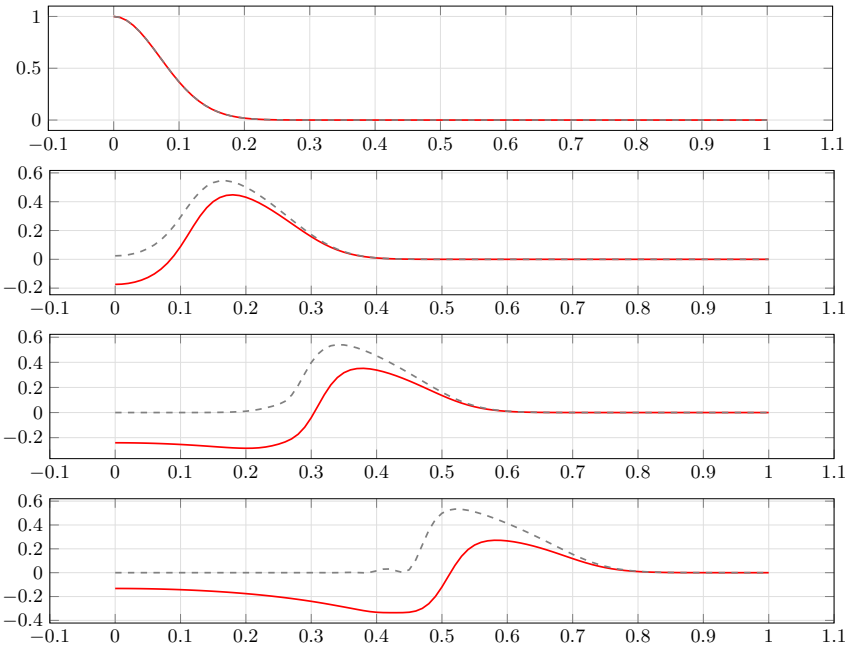


Figure 1.6.: Red solid: Numerical solution $\mathbf{e}(t^n)$ at times $t^n = 0, 0.2, 0.4, 0.6$ obtained by the lowest order scheme (1.82)–(1.85). Black dashed: the corresponding solution of the Kerr problem without dispersion.

| $h$ | $p=1$ | | $p=2$ | | $p=3$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-2}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. |
| 0.05 | 0.202846 | — | 0.211564 | — | 0.175516 | — |
| 0.025 | 0.064344 | 1.65 | 0.022711 | 3.22 | 0.011844 | 3.89 |
| 0.0125 | 0.016893 | 1.92 | 0.002864 | 2.98 | 0.000747 | 3.98 |
| 0.00625 | 0.004260 | 1.98 | 0.000359 | 2.99 | 0.000046 | 3.99 |

Table 1.8.: Convergence in space of the method based on the $\mathbf{e}-\mathbf{a}$ formulation.

| $\tau$ | $k=0$ | | $k=1$ | | $k=2$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. |
| 0.05 | 0.205263 | — | 0.670137 | — | 0.216656 | — |
| 0.025 | 0.055032 | 1.89 | 0.046396 | 3.85 | 0.004088 | 5.73 |
| 0.0125 | 0.013936 | 1.98 | 0.002949 | 3.97 | 0.000066 | 5.94 |
| 0.00625 | 0.003510 | 1.99 | 0.000185 | 3.99 | 0.000001 | 5.98 |

Table 1.9.: Convergence in time of the method based on the $\mathbf{e}-\mathbf{a}$ formulation.

We tested the $\mathbf{e}-\mathbf{h}$ formulation with different polynomial degrees in space and time. In Tables 1.6 and 1.6, we summarize the observed errors in the electric field component and the resulting convergence rates for polynomial degrees $p = 1, 2, 3$ of approximation in space and degrees $k = 0, 1, 2$ of approximation in time. To produce these results, we considered the time interval $T = [0, 0.5]$. The error is measured as in Section 1.2.5. We used the expressions err $= \max_{0\le n\le N} \|e^n_{x,h}(t^n) - e^n_{x,h/2}(t^n)\|_{h/2}$ and err $= \max_{0\le n\le N} \|e^n_{x,h}(t^n) -$

$e^{2n}_{x,h}\|_h$ to evaluate the errors in space and time. Let us recall that with subscript $h/2$ and superscript $2n$, we denote the solutions on the uniformly refined grid in space and time, respectively. We observe the convergence $O(h^{p+1} + \tau^{2k+2})$. In Tables 1.8 and 1.9, we also summarize the error in the electric field for the $\mathbf{e} - \mathbf{a}$ approach. In this case, we observe the convergence $O(h^{p+1} + \tau^{2k+2})$. As expected, these results coincide with the results of Section 1.2.5 and the theoretical results mentioned there.

The theoretical results on energy balance could also be verified. It should be noted that the lowest order $\mathbf{e} - \mathbf{h}$-based scheme with $h = 0.01$ and $\tau = 0.01$ produces a discrepancy between the energy and dissipation of order $O(10^{-4})$, which is the order of the numerical error. For the $\mathbf{e} - \mathbf{a}$ approach with the same discretization parameters, on the other hand, the balance was preserved up to machine precision $O(10^{-16})$.

**A note on comparison.** A comparative study with related schemes is yet to be done. In particular, the two approaches proposed in [19, 84] have not yet been implemented. Although these approaches are capable of higher-order discretization in space, their time stepping is only second-order accurate. Therefore, we anticipate similar results as in Section 1.2.5. We expect our approach to be more efficient if higher accuracy is required.

## 1.4. Summary and outlook

In this chapter, we have discussed the propagation of the nonlinear electromagnetic field in Kerr-type nonlinear media. We presented two formulations and derived energy balances, which were only possible due to the energy-based modelling approach for constitutive relations. To preserve these balances on the discrete level, we presented two variational discretization strategies that can be used to construct arbitrary high-order schemes. We verified our results with two numerical examples in one and two spatial dimensions and compared the efficiency of the approaches to related FDTD schemes. Finally, we demonstrated that the presented approaches can be extended to problems with linear Lorentz-type dispersion and provided an illustrative numerical example in one spatial dimension.

There are many open questions that require further investigation. Firstly, the error analysis and justification of the observed convergence rates are yet to be done. Additionally, the question of efficiency has not been fully explored. It is possible that using different types of Galerkin approximations or choosing different polynomial bases and quadrature rules in space and time may be beneficial for this problem.

A further comparison to other methods represents another topic of further investigation. In particular, the DG-based schemes have not been implemented, for comparison. Furthermore, the recently proposed $\mathbf{a} - \mathbf{d}$ formulation [84] seems interesting. It seems that there is a close resemblance with the presented $\mathbf{e} - \mathbf{a}$ formulation. The comparison of the two formulations and the possible adoption of the presented variational discretization is yet to be covered.

Before we conclude this chapter, it is worth mentioning that the approaches presented here can be applied to a much larger class of electromagnetic field problems. Although we have focused on a specific example of Kerr media, other nonlinear materials with instantaneous responses can be handled in the same way, provided that the constitutive relations

and the energy are consistent, as discussed in Section 1.1. This includes nonlinear magnetic materials. The vector potential formulation for the nonlinear eddy current problem has been discussed in the original publication [43].

Next to the linear Lorentz dispersion, a single- or multi-pole Debye model can be considered by analogy, or a combination of both. The application to general dispersive and memory-dependent materials is not yet settled. In particular, the extension of the energy-based modelling concept is not yet completely clear. It seems that the nonlinear dispersive Kerr-Debye model can also be handled in a similar manner. However, it is not yet completely settled and is subject to further investigation. The nonlinear magnetic problems with hysteresis represent another topic of potential research.

# Chapter 2.

# Electric circuits

Simulation of electric circuits is another fundamental problem in electrical engineering. In this chapter, we discuss the modelling and numerical treatment of nonlinear electric circuits and their coupling to field equations.

### State of the art

The state-of-the-art approach for modeling electric circuits is the *Modified Nodal Analysis* (MNA), introduced in [69]. In its conventional form, the MNA for a circuit consisting of capacitors, inductors, resistors, voltage and current sources reads

$$\begin{aligned} A_C C A_C^\top \partial_t e + A_R G A_R^\top e + A_L i_L + A_V i_V &= -A_I i_{src}, \\ L \partial_t i_L - A_L^\top e &= 0, \\ -A_V^\top e &= -v_{src}, \end{aligned}$$

where $e$ is the vector of electric node potentials, while $i_L$ and $i_V$ are the vectors of branch currents through inductors and voltage sources, respectively. The topology of the circuit is stored in partial incidence matrices $A_X$, while the description of circuit elements is encoded in capacitance $C$, inductance $L$, and conductance $G$ matrices, or matrix valued functions $C(A_C^\top e)$, $L(i_L)$, and $R(A_R^\top e)$ in general. See [32, 62, 110, 111] for an overview.

From a mathematical standpoint, an MNA system is a system of differential-algebraic equations (DAEs) [60, 105]. In fact, research in differential-algebraic systems has been driven for many years by electric circuits, and they can be found as a canonical example in classical DAE books; see e.g. [26, 66, 81]. The classification of differential-algebraic equations is based on the concept of index, which can be seen as a measure of difficulty for solving the DAE – the higher the index, the more issues can appear. In particular, due to the presence of algebraic constraints and possibly *hidden constraints*, the construction of initial values might be difficult; see [26, 30, 66] for further discussions. For the MNA formulation, the index has been extensively studied over the years. It is now well understood that under appropriate assumptions on the circuit elements, the DAE index of an MNA system is $\nu \leq 2$ and depends on the circuit's topology. More precisely,

(a1) if the circuit contains neither loops of voltage nor cutsets of current sources, the MNA system is a regular DAE of index $\nu \leq 2$;

(a2) if the circuit contains neither loops of capacitors and voltage source nor cutsets of inductors and current sources, the index is $\nu \leq 1$ ;

see [52, 61, 132, 136] for details and proofs.

Besides the analytical issues, also the numerical treatment of DAEs is difficult. Because of the algebraic constraints, implicit time stepping schemes have to be used [26, 66, 81]. While passivity on the discrete level can be proven for the implicit Euler method, strict passivity may in general be lost through discretization by standard single or multistep schemes. In the presence of strong nonlinearities, even well-established second-order schemes, like the trapezoidal rule (TR) or BDF-2 method, may become unstable [31]. Hence, low-order time integration schemes are typically used for discretization. In case of stability issues, one switches to the implicit Euler method; we refer to [62, Ch. 10,11] for detailed discussion on this topic. Fortunately, passive discretization can be achieved using variational techniques. The MNA formulation leads to systems of a canonical port-Hamiltonian structure, namely,

$$Q(u)^\top \partial_t u = -\mathcal{A}(u) + f, \quad \mathcal{E}'(u) = Q(u)u, \tag{2.1}$$

where $\mathcal{E}$ is the energy functional. Note that this structure is similar to the structure of the $\mathbf{e} - \mathbf{h}$-based formulations for Maxwell's equations, as discussed in Chapter 1. Therefore, following the framework presented in [42], the discontinuous Galerkin approximation can be employed to construct arbitrarily high-order schemes that unconditionally preserve the passivity of the discretization.

### Magnetic oriented formulation

An alternative *Magnetic Oriented Nodal Analysis* (MONA) formulation for electric circuits was recently introduced in [122]. For circuits containing the same canonical elements, namely capacitors, inductors, resistors, voltage, and current sources, the formulation reads

$$\begin{aligned} A_R G A_R^\top \partial_t \psi + A_C \partial_t q_C + A_V \partial_t q_V &= -A_L L^{-1} A_L^\top \psi - A_I i_{src}, \\ -A_C^\top \partial_t \psi &= -C^{-1} q_C, \\ A_V^\top \partial_t \psi &= -v_{src}, \end{aligned}$$

where $\psi$ is the vector with magnetic vector potentials and $q_C$ and $q_V$ stand for charges across capacitors and voltage sources. The magnetic potential $\psi$ is defined such that

$$\partial_t \psi = e \quad \text{and} \quad \phi_L = A_L^\top \psi$$

hold, from which the electric node potential can be determined through differentiation in time. On the other hand, the access to capacitor charges $q_C$ and inductor fluxes $\phi_L$ is directly provided, which is of an advantage for particular applications; we refer to [62] for discussion on related difficulties for the MNA systems.

A key advantage of the MONA formulation over the MNA is its lower DAE index. It is shown in [52] that under appropriate assumptions on circuit elements, the MONA formulation leads to systems of index $\nu \leq 1$. More precisely,

(b1) if the circuit contains neither loops of voltage sources nor cutsets of current sources, the MONA system is a regular DAE of index $\nu \leq 1$;

(b2) if the circuit contains neither loops of capacitors and/or voltage sources nor cutsets of inductors and current sources, the index is $\nu = 0$.

Note that the regularity condition (b1) is the same as (a1) for the MNA, while the index-0 condition (b2) is very similar to the index-1 condition (a2) for the MNA systems; c.f. [52, 122]. Since MONA systems are of index $\nu \leq 1$, one can directly exclude the presence of hidden constraints, which, in particular, simplifies the construction of consistent initial values. Further, the discretization of index $\nu \leq 1$ DAEs has been much better studied in the literature. In particular, full convergence rates for both differential and algebraic variables can be expected for stiffly accurate time-stepping schemes; see e.g. [26, 66].

From the perspective of energy-based modelling, it is shown that the MONA formulation leads to systems of a certain generalized gradient structure, namely

$$C(\partial_t u)\partial_t u = -\mathcal{H}'(u) + f, \tag{2.2}$$

where $\mathcal{H}$ denotes the energy functional. Let us note that the $\mathbf{e} - \mathbf{a}$-based formulations for Maxwell's equations, as discussed in Chapter 1, lead to problems of a very similar structure. Hence, by following the same variational methodology introduced in [43], the passivity of the problems can be preserved by employing Petrov-Galerkin schemes. Furthermore, for circuits with no power sources and no dissipative elements (resistors), this approach provides energy-conserving discretization schemes.

The magnetic oriented formulation for circuits takes a modelling approach similar to the vector potential formulation for field problems. The two problems share not only the magnetic viewpoint but also have similar geometric structures. As a result, the magnetic oriented ansatz for the field-circuit coupling is not only convenient from a modelling point of view but also leads to problems of the same canonical structure (2.2). Hence, the energy balance preserving discretization for the coupled problems can be constructed with Galerkin discretization in space for the field equations and Petrov-Galerkin time stepping.

### Outline and main contributions

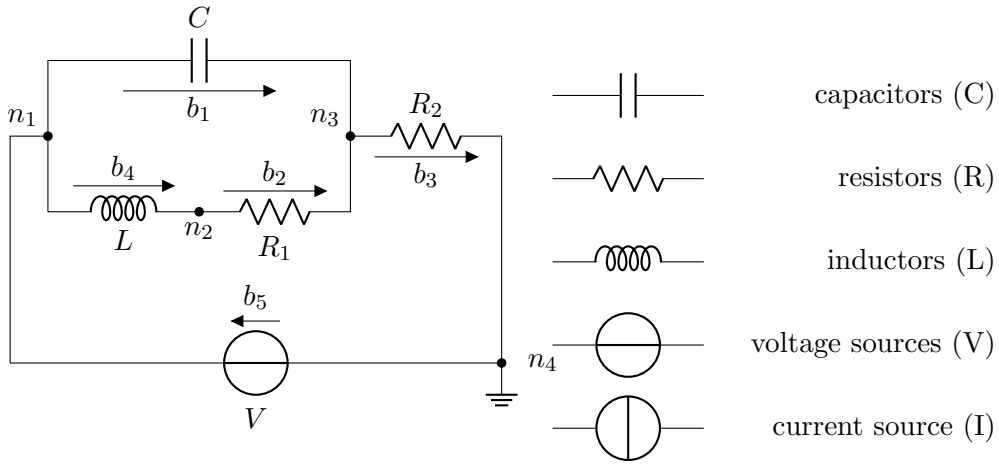
Let us briefly sketch the contents of this chapter and highlight the main contributions. In Section 2.1 we start with discussing the fundamentals of circuit modelling – Kirchhoff's circuit laws and constitutive relations for different element types. We use an energy-based modelling approach for the latter, which is crucial for further analysis. In Section 2.2 we recall the conventional form of the MNA formulation and discuss the standard results on regularity and index characterization. As our first contribution, we show that the MNA systems have the particular port-Hamiltonian structure (2.1), which allows a passivity preserving discretization based on the variational framework [42]. In Section 2.3, we present a novel magnetic oriented formulation (MONA) for electric circuits and provide an index analysis of the resulting DAE systems. This section represents the main contribution of the chapter. We further show that the MONA system has the structure (2.2) and discuss the discretization strategy which guides the construction of schemes preserving the underlying energy balance of the system and, hence, ensuring the passivity of the discretization. In Section 2.4 we briefly discuss the coupling of the proposed MONA system to the field elements described by the magneto-quasistatic model stated in terms of the magnetic vector potential. We show that the coupled problems have again the

canonical structure (2.2), which ensures their passivity and provides the guideline for the energy balance preserving discretization.

The contents of this chapter are based on our publications [43, 46, 122]. The main contribution of this chapter is the MONA formulation for the circuits which has been published in [122]. The magnetic oriented field-circuit coupling approach has not been published yet.

## 2.1. Fundamentals of circuit modeling

An electric circuit is considered a composition of basic electric components like resistors, capacitors, voltage sources etc.; see Figure 2.1a for a simple example. In this section, we introduce the notation, the basic quantities, and the physical laws necessary for the systematic modelling of electric circuits. We start with discussing the interconnection structure of the electric circuit, then we recall the physical balance laws and introduce the mathematical models describing the behavior of individual element types. An energy-based perspective is chosen for the latter. For ease of presentation, we restrict our considerations to circuits consisting of five basic element types: capacitors (C), inductors (L), resistors (R), voltage (V), and current sources (I). The schematic representations of individual element types are illustrated in Figure 2.1b. Further details and extension can be found in e.g. [62, 109, 110].




(a) Example: resonance circuit with damping.

(b) Schematic representation

Figure 2.1.: Example of an electric circuit and schematic representation of its components.

### 2.1.1. Topology of the circuit

The interconnection structure of a circuit is modelled by a finite connected directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{B})$ with nodes $n \in \mathcal{N}$ and branches $b \in \mathcal{B} \subset \mathcal{N} \times \mathcal{N}$. We exclude self-loops by assuming $x \neq y$ for every branch $b = (x, y)$. The topology of the graph is encoded in the

*incidence matrix* $\tilde{A} \in \mathbb{R}^{|\mathbb{N}|\times|\mathcal{B}|}$ defined as

$$\tilde{A}_{kj} = \begin{cases} 1, & \text{if branch } b_j \text{ leaves the node } n_k, \\ -1, & \text{if branch } b_j \text{ enters the node } n_k, \\ 0, & \text{else.} \end{cases}$$

The rows and columns of $\tilde{A}$ contain the connectivity of nodes and branches, respectively. By construction, each column has exactly two nonzero entries, and therefore the sum of all rows is zero, i.e. the rows are linearly dependent. More precisely, it can be shown that $\operatorname{rank}(A) = n_n - 1$ for any connected graph; see e.g. [109, Section 4]. Removing one of the rows from the incidence matrix $\tilde{A}$ results in *reduced incidence matrix* $A$, which consequently has full row rank. This matrix will again be called the incidence matrix in the following. As we will see below, removing one row from the incidence matrix is related to setting the electric potential at the node associated with the row to zero. The zero-potential node is also called the *reference node*. The branches of the graph may be sorted based on the element type. Then the incidence matrix can be decomposed as

$$A = [A_C \,|\, A_R \,|\, A_L \,|\, A_V \,|\, A_I], \tag{2.3}$$

where individual blocks contain all branches of the corresponding element types, e.g. $A_C$ contains all the capacitors, $A_R$ all the resistors, etc.

**Example 2.1.1.** The circuit illustrated in Figure 2.1a consists of four nodes and five branches indicated by $n_i$ and $b_j$ respectively. The underlying graph is then given by $\mathcal{G} = (\mathcal{N}, \mathcal{B})$ where $\mathcal{N} = \{n_1, n_2, n_3, n_4\}$ and $\mathcal{B} = \{b_1, b_2, b_3, b_4, b_5\}$ with $b_1 = (n_1, n_3)$, $b_2 = (n_2, n_3)$, $b_3 = (n_3, n_4)$, $b_4 = (n_1, n_2)$ and $b_5 = (n_4, n_1)$. We consider the ground node $n_4$ as the reference node. The full and reduced incidence matrices are then given by

$$\tilde{A} = \begin{pmatrix} 1 & 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & -1 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad A = \left(\begin{array}{c|cc|c|c} 1 & 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & -1 & 0 \\ -1 & -1 & 1 & 0 & 0 \end{array}\right).$$

Note that $\operatorname{rank}(\tilde{A}) = \operatorname{rank}(A) = 3$ and the branches are sorted in accordance to element types, i.e. $A = [A_C \,|\, A_R \,|\, A_L \,|\, A_V]$ and $A_I$ is empty since there are no current sources present.

### 2.1.2. Kirchhoff's circuit laws

As the next step, we recall the basic balance laws for the electric circuits, which have been postulated in 1845 by the German physicist Robert Gustav Kirchhoff [80]. The laws are stated in terms of branch currents, branch voltages, and node potentials, which are collected in vectors $i$, $v$, and $e$ respectively. By $i_L$, $v_C$, etc. we denote the components of the vectors related to the particular element types.

#### Kirchhoff's current law

To avoid the accumulation of charge in the nodes of the circuit one has to assume that *the sum of all currents going into and out of a node vanishes*; see Figure 2.2a for an

illustration. With the introduced notation, this can be written as $\sum_k A_{jk} i_j = 0$ for all $j$, or alternatively in compact form as

$$Ai = 0, \tag{KCL}$$

where $i$ is the vector of branch currents. Because the reduced incidence matrix is used, the balance of currents for the reference node is not explicitly included in (KCL). However, the relation is implicitly fulfilled as shown in Lemma B.2.1.

#### Kirchhoff's voltage law

We associate an electric potential to every node of the circuit and define voltage as oriented potential difference by

$$v = A^\top e, \tag{KVL}$$

where $e$ is the vector of node potentials and $v$ is the vector of branch currents. As a direct consequence, *the sum of the voltages in every loop vanishes*, which is the second fundamental law postulated by Kirchhoff; see Figure 2.2b for illustration. The last statement is actually equivalent to the existence of a vector with node potentials such that (KVL) holds; see [109, Section 4] or Appendix B.3 for details.


(a) $-i_1 + i_2 - i_3 + i_4 + i_5 = 0$ (b) $v_1 - v_2 + v_3 - v_4 - v_5 = 0$

Figure 2.2.: Simplified illustration of Kirchhoff's current (left) and voltage (right) laws.

### 2.1.3. Constitutive relations

For a convenient description of constitutive relations, we introduce two further quantities associated with the branches – a vector $q$ of *charges* and a vector $\phi$ of *fluxes* which relate to current and voltages as

$$i = \partial_t q \quad \text{and} \quad v = \partial_t \phi, \tag{2.4}$$

respectively. For capacitors and inductors, these quantities have a clear physical meaning, i.e. the entries $q_C$ of $q$ represent the charges on the plates of capacitors, and the entries $\phi_L$ of $\phi$ correspond to magnetic fluxes through the windings of inductors.

### Energy storing elements

Capacitors and inductors correspond to elements that store electric and magnetic energy respectively. Following the approach of the previous chapter, we take an energy-based standpoint and assume to be given appropriate energy functionals

$$\epsilon_C : q_C \mapsto \epsilon_C(q_C) \in \mathbb{R} \quad \text{and} \quad \epsilon_L : \phi_L \mapsto \epsilon_L(\phi_L) \in \mathbb{R},$$

where $q_C$ and $\phi_L$ are the vectors of capacitor charges and inductor fluxes. The constitutive relations for capacitors and inductors are then defined through the relations

$$v_C = \epsilon'_C(q_C) \quad \text{and} \quad i_L = \epsilon'_L(\phi_L), \tag{2.5}$$

where $v_C$ and $i_L$ are the vectors of capacitor voltages and inductor currents, respectively.

**Remark 2.1.2.** Assuming that $\epsilon_C$ and $\epsilon_L$ are smooth, strongly convex, and coercive, the constitutive relations (2.5) can be inverted. The equivalent inverse relations are then given by

$$q_C = \epsilon'_{*,C}(v_C) \quad \text{and} \quad \phi_L = \epsilon'_{*,L}(i_L), \tag{2.6}$$

where $\epsilon_{*,C}$ and $\epsilon_{*,L}$ are convex conjugate functionals often called co-energies; see Section 1.1. From this standpoint, the variables $q_C$ and $\phi_L$ are sometimes called *energy variables* while $v_C$ and $i_L$ are the *co-energy variables*. By differentiating (2.6) and using (2.4) we obtain

$$i_C = C(v_C)\partial_t v_C \quad \text{and} \quad v_L = L(i_L)\partial_t i_L, \tag{2.7}$$

with $C(v_C) = \epsilon''_{*,C}(v_C)$ and $L(i_L) = \epsilon''_{*,L}(i_L)$ representing the *differential* or *incremental capacitance* and *inductance* matrices.

**Example 2.1.3.** For an illustration of the above statements, let us consider quadratic energy functionals $\epsilon_C(q_C) = \frac{1}{2}\|q_C\|^2_{C^{-1}}$ and $\epsilon_L(\phi_L) = \frac{1}{2}\|\phi_L\|^2_{L^{-1}}$ with given symmetric positive definite matrices $C$ and $L$. Then the constitutive relations (2.5) lead to

$$v_C = \epsilon'_C(q_C) = C^{-1}q_C \quad \text{and} \quad i_L = \epsilon'_L(\phi_L) = L^{-1}\phi_L.$$

The corresponding co-energies are then given by the quadratic functionals $\epsilon_{*,C}(v_C) = \frac{1}{2}\|v_C\|^2_C$ and $\epsilon_{*,L}(i_L) = \frac{1}{2}\|i_L\|^2_L$, and the inverse relations (2.6) correspond to

$$q_C = \epsilon'_{*,C}(v_C) = Cv_C \quad \text{and} \quad \phi_L = \epsilon'_{*,L}(i_L) = Li_L.$$

The Hessians $\epsilon''_{*,C}(v_C) = C$ and $\epsilon''_{*,L}(i_L) = L$ are the capacitance and inductance matrices, and the voltage-current relations for capacitors and inductors can then be written in the common forms $i_C = C\partial_t v_C$ and $v_L = L\partial_t i_L$, corresponding to linear elements; see e.g.[62].

### Energy dissipating elements

The resistors correspond to the elements which dissipate the energy. In contrast to energy-storing elements, the voltage-current relation in resistors is algebraic and given by

$$i_R = G(v_R)v_R, \tag{2.8}$$

which is known as Ohm's law. The conductivity matrix $G(v_R)$ is assumed to be symmetric and positive definite. According to Joule's principle, the power dissipated by the currents flowing through the conductors is given by $P_{Joule}(v_R) = \langle i_R, v_R \rangle = \langle G(v_R)v_R, v_R \rangle \geq 0$, since we assume that conductivity matrix $G(v_R)$ is positive definite. Here, $\langle \cdot, \cdot \rangle$ denotes the scalar product.

#### Energy sources

While the capacitors and inductors store the energy and the resistors dissipate the energy, the voltage and current sources act as energy sources and sinks. The constitutive relations for these elements are simply given by

$$v_V = v_{src} \quad \text{and} \quad i_I = i_{src}, \tag{2.9}$$

where $v_{src}$ and $i_{src}$ are assumed to be given, i.e. the source terms are assumed to be independent. More general controlled sources are discussed, e.g. in [52, 62]. The power supplied to or extracted from the system through voltage and current sources is then given by $P_v = \langle v_{src}, i_V \rangle$ and $P_I = \langle i_{src}, v_I \rangle$, where $v_I$ and $i_V$ are the vectors with voltages through current sources and currents through voltage sources.

Having introduced the circuit topology, the basic balance laws, and the constitutive relations for the individual elements, we are now in the position to present and discuss two complete mathematical models for describing the physical behavior of electric circuits.

## 2.2. Modified Nodal Analysis for electric circuits

Kirchhoff's circuit laws and the constitutive equations discussed in the previous section allow a complete description of the dynamical behavior of electric circuits. However, some of the introduced quantities are redundant and can be eliminated to reduce the system and obtain a more compact representation. In this section, we discuss an approach based on electric node potentials and currents across inductors and voltage sources, the so-called *Modified Nodal Analysis (MNA)*. Since its introduction in 1975 by Albert Ruehli and co-workers [69], the approach has become the *de-facto* standard for circuit simulation in industrial applications. Moreover, the modified nodal analysis was also studied intensively in the literature; see e.g. [62, 110]. From the mathematical perspective, the formulation typically leads to a system of differential-algebraic equations (DAE), which provides some challenges for the analytical and numerical treatment, see e.g. [26, 66, 81]. We start this section by deriving the MNA formulation and recalling some basic facts about DAEs and their index, and the application of these results to MNA. Then, we study the particular structure of the systems, prove passivity, and present a passivity-preserving discretization strategy. The latter considerations are strongly based on the energy-based modelling approach presented in Section 2.1.

### 2.2.1. The modified nodal analysis

Decomposition of the incidence matrix $A$ and the current vector $i$ into individual blocks for each element type as in (2.3) allows to state Kirchhoff's current law (KCL) as

$$A_C i_C + A_R i_R + A_L i_L + A_V i_V + A_I i_I = 0. \tag{2.10}$$

In a similar manner, Kirchhoff's voltage law (KVL) can be stated as $v_X = A_X^\top e$ for the different element types $X \in \{C, R, L, V, I\}$. The constitutive relations (2.7) and (2.8) for capacitors, resistors, and inductors can then be phrased in terms of currents and electric potential by

$$i_C = C(A_C^\top e)A_C^\top \partial_t e, \quad i_R = G(A_R^\top e)A_R^\top e, \quad \text{and} \quad A_L^\top e = L(i_L)\partial_t i_L. \tag{2.11}$$

For voltage and current sources (2.9), we similarly have $A_V^\top e = v_{src}$ and $i_I = i_{src}$. By substituting these expressions into Kirchhoff's current law (2.10), we obtain the system

$$\begin{pmatrix} A_C C A_C^\top & 0 & 0 \\ 0 & L & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial_t e \\ \partial_t i_L \\ \partial_t i_V \end{pmatrix} + \begin{pmatrix} A_R G A_R^\top & A_L & A_V \\ -A_L^\top & 0 & 0 \\ -A_V^\top & 0 & 0 \end{pmatrix} \begin{pmatrix} e \\ i_L \\ i_V \end{pmatrix} = \begin{pmatrix} -A_I i_{src} \\ 0 \\ -v_{src} \end{pmatrix} \tag{2.12}$$

with $C = C(A_C^\top e)$, $L = L(i_L)$ and $G = G(A_R^\top e)$ denoting the incremental capacity, inductance, and conductivity matrices, respectively. The system (2.12) is known as the *conventional form* of MNA. Generalizations, like *charge-flux*-based formulations and port-Hamiltonian extensions, and their relation to the conventional form have been discussed in e.g. [62, 116].

### 2.2.2. Index analysis of the MNA

Let us briefly recall some elementary notions used for the analysis of DAEs; for details, see e.g. [26, 81, 110]. An initial value is called *consistent* with a DAE if a (local) solution to the corresponding initial value problem exists. A DAE is called *regular* if, for any choice of consistent initial values, the solution is unique. A further classification of DAEs can be made according to their *index*. Within this thesis, we consider the differentiation index, which essentially corresponds to the number of differentiations necessary to transform a DAE into an equivalent ODE system; see e.g. [26, 81]. An overview of different index concepts and their relations can be found in [96].

Let us return to the MNA system (2.12). The following theorem summarizes the most important facts about its regularity and index characterization; see e.g [52, 109, 137] for proofs and further details. Without further mentioning, we assume that $C(\cdot)$, $L(\cdot)$, $G(\cdot)$ as well as $v_{src}(\cdot)$, $i_{src}(\cdot)$ are smooth functions of their arguments.

**Theorem 2.2.1.** Let $C(v_C)$, $L(i_L)$, $G(v_R)$ be symmetric and positive definite matrices for any admissible argument $v_C$, $i_L$, and $v_R$. Further, assume that

$$N([A_R, A_C, A_V, A_L]^\top) = 0 \quad \text{and} \quad N(A_V) = 0. \tag{A1}$$

Then (2.12) is a regular system of DAEs with index $\nu \leq 2$. If additionally

$$N([A_R, A_C, A_V]^\top) = 0 \quad \text{and} \quad N([A_C, A_V]) = 0 \tag{A2}$$

holds, then the system is of index $\nu \leq 1$. If further

$$N(A_C^\top) = 0 \quad \text{and} \quad \dim(v_{src}) = 0 \tag{A3}$$

holds, then the system has index $\nu = 0$, i.e., it is an ODE.

Here and below we denote by $N(B)$ the nullspace of a matrix $B$ and by $\dim(v)$ the size of a vector $v$. Let us note that the index-1 condition (A2) could actually be slightly relaxed; for details, see [52, 137].

**Remark 2.2.2.** The algebraic conditions (A1)–(A3) are assumptions on the topology of the circuit, and can be interpreted in physical terms as follows:

(A1) the circuit contains neither loops of voltage sources nor cutsets of current sources;

(A2) circuit contains neither loops of capacitors and/or voltage sources nor cutsets of inductors and/or current sources;

(A3) every node in the circuit can be connected to the reference node through a path containing only the capacitors.

Theorem 2.2.1 provides the existence of a unique solution to the MNA system (2.12) for every consistent choice of initial values. Let us note that the construction of consistent initial values for index $\nu = 2$ systems is rather difficult due to presence of *hidden constraints*; for related work on this topic, see e.g. [50, 51, 121], and references therein.

### 2.2.3. Port-Hamiltonian structure and energy balance

We now discuss the basic energy balance provided by the MNA system. To do so, it is convenient to write (2.12) as an abstract port-Hamiltonian system, which is possible because of the energy-based modelling approach discussed in Section 2.1.

#### Port-Hamiltonian structure of the MNA formulation

The total energy in an electric circuit is stored in capacitors and inductors and is given by $\epsilon_C(q_C) + \epsilon_L(\phi_L)$. According to the inverse constitutive relations $q_C = \epsilon'_{*,C}(v_C)$ and $\phi_L = \epsilon'_{*,L}(i_L)$, and the voltage relation $A_C^\top e = v_C$, the energy can be expressed as a function of MNA system variables by $\mathcal{E}(e, i_L, i_V) = \epsilon_C(\epsilon'_{*,C}(A_C^\top e)) + \epsilon_L(\epsilon'_{*,L}(i_L))$. Differentiation of this expression and using (2.4) and (2.11) leads to

$$\begin{aligned} \partial_e \mathcal{E}(e, i_L, i_V) &= A_C C(A_C^\top e) A_C^\top e, \\ \partial_{i_L} \mathcal{E}(e, i_L, i_V) &= L(i_L) i_L, \\ \partial_{i_V} \mathcal{E}(e, i_L, i_V) &= 0, \end{aligned} \tag{2.13}$$

where $C(v_C) = \epsilon''_{*,C}(v_C)$ and $L(i_L) = \epsilon''_{*,L}(i_L)$ are the incremental capacitance and inductance. The conventional MNA formulation (2.12) and the energy relation (2.13) can now be written compactly as

$$Q(u)^\top \partial_t u = -\mathcal{A}(u) + f, \tag{2.14}$$

$$\mathcal{E}'(u) = Q(u)u, \tag{2.15}$$

with state vector $u = (e, i_L, i_V)$, source vector $f = (-A_I i_{src}, 0, -v_{src})$, function $\mathcal{A}$, which takes the form $\mathcal{A}(u) = (R(u) - J)u$, and matrices $Q = Q(u)$, $R = R(u)$, and $J$ given by

$$Q^\top = \begin{pmatrix} A_C C A_C^\top & 0 & 0 \\ 0 & L & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} A_R G A_R^\top & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \text{and} \quad J = \begin{pmatrix} 0 & -A_L & -A_V \\ A_L^\top & 0 & 0 \\ A_V^\top & 0 & 0 \end{pmatrix}.$$

Noting that $C = C(v_C)$, $L = L(i_L)$ and $G = G(v_R)$ were assumed symmetric positive definite, we see that $Q = Q(u)$ and $R = R(u)$ are symmetric positive semi-definite and $J$ is skew-symmetric.

The structure (2.14)-(2.15) of the MNA formulation is therefore similar to that of $\mathbf{e}-\mathbf{h}$-based formulations for Maxwell's equations. The properties of finite dimensional port-Hamiltonian systems have been studied intensively in the literature; see e.g. [42, 97, 139]. As shown in [98], the index of (linear) systems is always $\nu \leq 2$. Furthermore, as briefly discussed in Chapter 1, port-Hamiltonian systems come with a natural balance of power.

#### Energy balance

Following Section 1.2.1 and Appendix A.1, the energy balance for the problems of this structure can be directly derived using simple variational calculus. A solution $u$ of the problem (2.14)-(2.15) satisfies the following power balance

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(u(t)) &= \langle \partial_t u(t), \mathcal{E}'(u(t))\rangle = \langle \partial_t u(t), Q(u(t))u(t)\rangle \\ &= \langle Q(u(t))^\top \partial_t u(t), u(t)\rangle = -\langle \mathcal{A}(u(t)), u(t)\rangle \\ &= -\langle R(u(t))u(t), u(t)\rangle + \langle f(t), u(t)\rangle. \end{aligned} \tag{2.16}$$

In the last step, we substituted $\mathcal{A}(u) = R(u) - J$ and used the fact that $J$ is skew-symmetric and $\langle Jx, x\rangle = 0$. Integrating with respect to time leads to the energy balance

$$\mathcal{E}(u(t)) - \mathcal{E}(u(s)) = -\int_s^t \langle R(u(\tau))u(\tau), u(\tau)\rangle \, d\tau + \int_s^t \langle f(\tau), u(\tau)\rangle \, d\tau.$$

With the assumption that $R(u)$ is symmetric positive semi-definite we obtain

$$\mathcal{E}(u(t)) - \mathcal{E}(u(s)) \leq \int_s^t \langle f(\tau), u(\tau)\rangle \, d\tau,$$

which holds for any input $f$ and the corresponding solution $u$ of (2.14)-(2.15). The energy of the system thus can only grow by power supplied through the inputs, i.e., the system is *passive*; see e.g. [3, 87].

**Remark 2.2.3.** The relation (2.16) can be directly applied to the MNA formulation (2.12), leading to the following power balance

$$\frac{d}{dt}\mathcal{E}(e, i_L, i_V) = -\langle G(A_R^\top e)A_R^\top e, A_R^\top e\rangle - \langle v_{src}, i_V\rangle - \langle i_{src}, A_I^\top e\rangle.$$

This shows, that the MNA formulation together with our assumptions on constitutive equations leads to a passive system.

Utilizing the particular structure of the problem and following the framework [42], we now formulate the discretization strategy based on the discontinuous Galerkin approach that allows the construction of passivity-preserving schemes.

### Passivity preserving discretization

Let $d$ denote the size of the system, i.e., $u(t) \in \mathbb{R}^d$. Let $\mathcal{T} = \{t^n : 0 \le n \le N\}$ be a sequence of discrete time steps $t^n = n\tau$ with $\tau = T/N$. With $I^n = [t^{n-1}, t^n]$ we denote the $n$-th time interval and with $P_k(I^n; \mathbb{V})$ we denote the space of polynomials with values in $\mathbb{V}$. By $P_k(\mathcal{T}; \mathbb{V})$ we denote the space of piece-wise polynomials, i.e., the functions whose restrictions to any interval $I^n$ lie in $P_k(I^n; \mathbb{V})$. We further use $(*)|_{t^n}$ to abbreviate the evaluation of $(*)$ at time $t = t^n$. For the problem (2.14)–(2.15), we now consider the following method.

**Problem 2.2.4.** Let $u^0 = u^0(0)$ be given. For $1 \le n \le N$, find $u^n \in P_k(I^n; \mathbb{R}^d)$ such that

$$\begin{aligned} &\int_{I^n} \langle Q^\top(u^n(t))\partial_t u^n, v(t)\rangle \; dt \; + \; \langle Q^\top(u^n)(u^n - u^{n-1}), v\rangle|_{t^{n-1}} \\ &= \int_{I^n} \langle (J - R(u^n(t)))u^n(t), v(t)\rangle \; dt + \int_{I^n} \langle f(t), v(t)\rangle \; dt, \qquad \forall v \in P_k(I^n; \mathbb{R}^d). \end{aligned} \tag{2.17}$$

The method is a finite-dimensional version of [42, Scheme 4.2] or Scheme (A.8) adapted to the problem structure. Thus, we can directly conclude that its solution satisfies the following dissipation inequality.

**Lemma 2.2.5.** Let $u \in P_k(\mathcal{T}; \mathbb{R})$ be a solution of (2.17). Then

$$\mathcal{E}(u^n(t^n)) - \mathcal{E}(u^m(t^m)) \le -\int_{t^m}^{t^n} \langle R(u(t))u(t), u(t)\rangle \; dt + \int_{t^m}^{t^n} \langle f(t), u(t)\rangle \; dt.$$

*Proof.* The proof of the statement is identical to that of [42, Theorem 4.1], which has been summarized in Lemma A.1.4 in Appendix A.1. □

The scheme applied to the MNA system (2.12) leads to the energy dissipation principle

$$\begin{aligned} &\mathcal{E}(e^n(t^n), i_L^n(t^n), i_V^n(t^n)) - \mathcal{E}(e^m(t^m), i_L^m(t^m), i_V^m(t^m)) \\ &\le -\int_{t^m}^{t^n} \langle G(A_R^\top e)A_R^\top e(t), A_R^\top e(t)\rangle \, dt - \int_{t^m}^{t^n} \langle v_{src}(t), i_V(t)\rangle \, dt - \int_{t^m}^{t^n} \langle i_{src}(t), A_I^\top e(t)\rangle \, dt, \end{aligned}$$

which ensures the discrete passivity for the MNA formulation.

### A note on numerical realization

The numerical realization of the schemes has already been discussed in Section 1.2.2. Let us emphasize that an appropriate choice of the quadrature rule $(w_i, \xi_i)$ is an essential ingredient. In Chapter 1, we discussed problems with quadratic nonlinearities, allowing us to easily determine the necessary exactness degree of the quadrature. In this chapter, we are dealing with non-polynomial nonlinearities. Therefore, the quadrature must be chosen in such a way that the integration error becomes negligible. Our choices are discussed in Section 2.5 below.

## 2.3. Magnetic oriented nodal analysis for electric circuits

After analyzing the MNA formulation and providing the passivity-preserving discretization, we now draw our attention to an alternative *Magnetic Oriented Nodal Analysis (MONA)* formulation, recently introduced in [122]. The MONA formulation is derived from the same physical principles and constitutive relations as the MNA. However, it is written in terms of different quantities, which leads to systems of a smaller index. More precisely:

- the MNA uses electric node potentials and currents as system unknowns and leads to DAEs of index $\nu \leq 2$;
- the MONA formulation is based on magnetic node potentials and charges and leads to systems of index $\nu \leq 1$.

Further, the particular structure of the MONA systems allows the construction of discretization schemes, which preserve the underlying energy balance exactly. Similar to the previous section, we start by deriving the MONA formulation, then analyze the regularity of DAE systems and characterize their index. As the final step, we study the particular structure of the problems, show passivity, and present a structure-preserving discretization strategy.

### 2.3.1. The magnetic oriented nodal analysis

Like in Section 2.1 we introduce vectors $q$, $\phi$, $\psi$ of electric charges, magnetic fluxes, and magnetic node potentials, such that

$$i = \partial_t q, \quad v = \partial_t \phi, \quad \text{and} \quad e = \partial_t \psi;$$

compare with relations (2.4). The Kirchhoff's current law (KCL) can be written as

$$A_C \partial_t q_C + A_R \partial_t q_R + A_L \partial_t q_L + A_V \partial_t q_V + A_I \partial_t q_I = 0.$$

and Kirchhoff's voltage law (KVL) can be written as $\phi_X = A_X^\top \psi$ for different element types $X \in \{C, R, L, V, I\}$. The constitutive relations (2.5) and (2.8) for capacitors, inductors, and resistors can finally be phrased as

$$A_C \partial_t \psi = \epsilon'(q_C), \quad \partial_t q_L = \epsilon_L'(A_L^\top \psi), \quad \text{and} \quad \partial_t q_R = G(A_R^\top \partial_t \psi) A_R^\top \partial_t \psi. \tag{2.18}$$

For voltage and current sources (2.9), we similarly have $A_V^\top \partial_t \psi = v_{src}$ and $\partial_t q_I = i_{src}$. Putting everything together we arrive at

$$\begin{pmatrix} A_R G(A_R^\top \partial_t \psi) A_R^\top & A_C & A_V \\ -A_C^\top & 0 & 0 \\ -A_V^\top & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial_t \psi \\ \partial_t q_C \\ \partial_t q_V \end{pmatrix} = - \begin{pmatrix} A_L \epsilon_L'(A_L^\top \psi) \\ \epsilon_C'(q_C) \\ 0 \end{pmatrix} - \begin{pmatrix} A_I i_{src} \\ 0 \\ v_{src} \end{pmatrix}, \tag{2.19}$$

which we call the MONA formulation; see [122]. Since the two formulations MNA and MONA are based on the same physical principles and constitutive laws, they are equivalent and can be transformed into each other.

### 2.3.2. Analysis of the MONA system

In the following, we summarize some important properties of the MONA system, which were originally stated in [122]. As in the previous section, we assume that the conductivity $G(\cdot)$, the source current $i_{src}(\cdot)$, and source voltage $v_{src}(\cdot)$ are smooth functions of their arguments.

**Theorem 2.3.1** (Regularity and index; Theorem 1 in [122])**.**
Let the conductivity matrix $G(v_R)$ be symmetric positive definite for any $v_R$, and the energy functionals $\epsilon_C(\cdot)$, $\epsilon_L(\cdot)$ be smooth and strictly convex. If

$$N([A_R, A_C, A_V, A_L]^\top) = 0 \qquad \text{and} \qquad N(A_V) = 0, \tag{B1}$$

then the system (2.19) is a regular DAE of index $\nu \leq 1$. If in addition to (B1) also

$$N([A_R, A_C, A_V]^\top) = 0 \qquad \text{and} \qquad N([A_C, A_V]) = 0 \tag{B2}$$

hold, then the index is $\nu = 0$, i.e., (2.19) is an ordinary differential equation.

*Proof.* The proof is adopted from [122]. To simplify the notation, we use the abbreviation $G^\psi = G(A_R^\top \partial_t \psi)$ and consider $v_{src} = 0$ and $i_{src} = 0$ without loss of generality. We start with the second assertion. The leading matrix in (2.19) has the form

$$\begin{pmatrix} K & B^\top \\ -B & 0 \end{pmatrix} \tag{2.20}$$

with $B^\top = [A_C, A_V]$ and $K = A_R G^\psi A_R^\top$. Using assumptions (B1)-(B2), we see that $B^\top$ is surjective and that $K$ is regular on $N(B)$. As a consequence of Brezzi's lemma, see [15, Theorem 3.2], the matrix (2.20) is regular, hence (2.19) can be transformed into an explicit ordinary differential equation. Now we prove the first assertion. If condition (B2) does not hold, we can split the spaces of magnetic potentials $\psi$ and charges $q = (q_C, q_V)$ into

$$\begin{aligned} V_\psi &= N([A_R, A_C, A_V]^\top) \oplus N([A_R, A_C, A_V]^\top)^\perp, \\ V_q &= N([A_C, A_V]) \oplus N([A_C, A_V])^\perp. \end{aligned}$$

We further choose an orthogonal basis for corresponding subspaces and decompose

$$\psi = Q_1\psi_1 + Q_2\psi_2 \qquad \text{and} \quad q = P_1 q_1 + P_2 q_2.$$

such that $[A_R, A_C, A_V]^\top Q_1 = 0$ and $[A_C, A_V] P_1 = 0$, while $[A_R, A_C, A_V]^\top Q_2$ and $[A_C, A_V] P_2$ have trivial nullspaces. Moreover, $Q = [Q_1, Q_2]$ and $P = [P_1, P_2]$ are regular orthogonal matrices. We further decompose the projectors into

$$\begin{pmatrix} q_C \\ q_V \end{pmatrix} = \begin{pmatrix} P_{1,C} \\ P_{1,V} \end{pmatrix} q_1 + \begin{pmatrix} P_{2,C} \\ P_{2,V} \end{pmatrix} q_2.$$

We now multiply the system (2.19) from left by $\operatorname{blkdiag}(Q^\top, P^\top)$ and use the above decompositions for $\psi$ and $q = (q_C, q_V)$, which leads to the equivalent form

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \tilde{K}^\psi & 0 & \tilde{B}^\top \\ 0 & 0 & 0 & 0 \\ 0 & -\tilde{B} & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial_t \psi_1 \\ \partial_t \psi_2 \\ \partial_t q_1 \\ \partial_t q_2 \end{pmatrix} = - \begin{pmatrix} Q_1^\top A_L \epsilon_L'(A_L^\top (Q_1\psi_1 + Q_2\psi_2)) \\ Q_2^\top A_L \epsilon_L'(A_L^\top Q_1\psi_1 + Q_2\psi_2)) \\ P_{1,C}^\top \epsilon_C'(P_{1,C} q_1 + P_{2,C} q_2) \\ P_{2,C}^\top \epsilon_C'(P_{1,C} q_1 + P_{2,C} q_2) \end{pmatrix}$$

with $\tilde{K}^\psi = Q_2^\top A_R G^\psi A_R^\top Q_2$ and $\tilde{B}^\top = (Q_2^\top [A_C, A_V] P_2)$. With a slight rearrangement of variables and equations, the system can then be written compactly as

$$M^\psi \partial_t y = f(y, z), \tag{2.21}$$
$$0 = g(y, z), \tag{2.22}$$

where $y = (\psi_2; q_2)$, $z = (\psi_1; q_1)$, and the matrix in front of the derivative is given by

$$M^\psi = \begin{pmatrix} \tilde{K}^\psi & \tilde{B}^\top \\ -\tilde{B} & 0 \end{pmatrix}.$$

We now show that (2.21)–(2.22), which up to elementary algebraic transformations is equivalent to (2.19), is a Hessenberg system of index-1; see [66]. This requires to verify that $M^\psi$ and the partial Jacobian $g_z(y, z)$ are regular. We first show that the matrix $M^\psi$ is regular. To do so, we note that $N(\tilde{B}^\top) = 0$ by construction. Furthermore, $K^\psi$ is regular on $N(\tilde{B})$, which can be seen as follows:

$$N(\tilde{B}) = N(P_2^\top [A_C, A_V]^\top Q_2) = N([A_C, A_V]^\top Q_2).$$

Since $G^\psi$ is symmetric and positive definite, we have

$$N(\tilde{K}^\psi) \cap N(\tilde{B}) = N(A_R^\top Q_2) \cap N([A_C, A_V^\top] Q_2) = N([A_R, A_C, A_V]^\top Q_2) = 0,$$

by the construction of $Q_2$. Hence, by Brezzi's theorem [15, Theorem 3.2], the matrix $M^\psi$ is regular, as desired, i.e., (2.21) can be transformed to an explicit ODE. In the second step, we show that $g_z(y, z)$ is regular. To do so, we differentiate the algebraic constraints

$$0 = g(y, z) := \begin{pmatrix} Q_1^\top A_L \epsilon_L'(A_L^\top (Q_1 \psi_1 + Q_2 \psi_2)) \\ P_{1,C}^\top \epsilon_C'(P_{1,C} q_1 + P_{2,C} q_2) \end{pmatrix}.$$

By the chain rule, we obtain

$$g_z(y, z) = \begin{pmatrix} Q_1^\top A_L \epsilon_L''(v_L) A_L^\top Q_1 & 0 \\ 0 & P_{1,C}^\top \epsilon_C''(q_C) P_{1,C} \end{pmatrix}, \tag{2.23}$$

where $v_L = A_L^\top (Q_1 \psi_1 + Q_2 \psi_2)$ and $q_C = P_{1,C} q_1 + P_{2,C} q_2$. Since the energies $\epsilon_C$ and $\epsilon_L$ are convex by assumption, the two Hessians $\epsilon_C''(q_C)$ and $\epsilon_L''(\phi_L)$ are positive definite for arbitrary arguments. The regularity of the upper left block in (2.23) follows from $N(A_L^\top Q_1) = 0$, which is a direct consequence of the first condition in (B1). Further, from $N(A_V) = 0$ we can directly deduce that $(0, q_V) \in N(A_C, A_V)$ implies $q_V = 0$. This shows that $P_{1,C}$ is injective, and the lower right block of (2.23) is regular. Therefore the Jacobian $g_z(y, z)$ is regular. In summary, we conclude that (2.21)-(2.22) is a Hessenberg system of index $\nu = 1$, and in particular, a regular DAE. Since only the algebraic equivalence transformations were performed, the result translates to the MONA system (2.19). □

**Remark 2.3.2.** Theorem 2.3.1 guarantees the existence of a unique solution for every choice of consistent initial values. Let us note that the construction of consistent initial values for index $\nu \leq 1$ systems is much simpler than for the index $\nu = 2$, since no hidden constraints are present. From this perspective, the MONA formulation represents a significant improvement over the MNA.

Note that conditions (B1)–(B2) are equivalent to conditions (A1)–(A2) for the MNA formulation, which automatically implicates the equivalent interpretation in terms of topological connectivity as discussed in Remark 2.2.2. The MONA formulation leads to a system of a smaller index compared to the MNA in most cases.

### 2.3.3. Geometric structure and power balance

The fact that both MNA and MONA are based on the same physical principles, allows us to directly deduce the passivity of the latter. However, the underlying structure of MONA differs from that of the MNA; while the MNA systems fit into the port-Hamiltonian framework, the MONA system has a different geometric structure – a generalized gradient flow structure. We now take a closer look at this particular structure, derive the associated power balance, and prove the passivity of the system.

#### Geometric structure of the MONA formulation

Let us recall, that the energy of a circuit consists of electric and magnetic energies stored in capacitors and inductors, respectively. With $\mathcal{H}(\psi, q_C, q_V) = \epsilon_L(A_L^\top \psi) + \epsilon_C(q_C)$ we denote the energy as a function of MONA variables. Differentiating this expression we obtain $\partial_\psi \mathcal{H}(\psi, q_C, q_V) = A_L \epsilon_L'(A_L^\top \psi)$, $\partial_{q_C} \mathcal{H}(\psi, q_C, q_V) = \epsilon_C'(q_C)$, and $\partial_{q_V} \mathcal{H}(\psi, q_C, q_V) = 0$. Abbreviating $u = (\psi, q_C, q_V)$, $\mathcal{H}'(u) = (\partial_\psi \mathcal{H}(u), \partial_{q_C} \mathcal{H}(u), \partial_{q_V} \mathcal{H}(u))$, the MONA system (2.19) can be written compactly as

$$\mathcal{C}(\partial_t u)\partial_t u = -\mathcal{H}'(u) + f \tag{2.24}$$

with state vector $u = (\psi, q_C, q_V)$, source vector $f = (-A_I i_{src}, 0, -v_{src})$, operator $\mathcal{C}$, which can be decomposed into $\mathcal{C}(\partial_t u) = (R(\partial_t u) - J)$, and matrices

$$R(\partial_t u) = \begin{pmatrix} A_R G(A_R^\top \partial_t \psi) A_R^\top & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} 0 & -A_L & -A_V \\ A_L^\top & 0 & 0 \\ A_V^\top & 0 & 0 \end{pmatrix}.$$

Since $G = G(v_C)$ was assumed symmetric and positive definite for any $v_R$, we see that $R(\partial_t x)$ is symmetric and positive semi-definite. Moreover, we observe that $J$ is skew-symmetric.

The structure (2.24) of the MONA formulation is similar to that of $\mathbf{e} - \mathbf{a}$ formulations for Maxwell's equations discussed in Chapter 1. As briefly discussed in Section (1.2.3), the problems of this structure naturally provide the underlying energy balance.

#### Power balance

It is easy to observe that a smooth solution $u$ of (2.24) satisfies the power balance

$$\begin{aligned} \frac{d}{dt}\mathcal{H}(u(t)) = \langle \mathcal{H}'(u(t)), \partial_t u \rangle &= -\langle \mathcal{C}(\partial_t u)\partial_t u(t), \partial_t u(t) \rangle + \langle f(t), \partial_t u(t) \rangle \\ &= -\langle R(\partial_t u(t))\partial_t u(t), \partial_t u(t) \rangle + \langle f(t), \partial_t u(t) \rangle, \end{aligned} \tag{2.25}$$

where in the last step we used $\mathcal{C}(\partial_t u) = (R(\partial_t u(t)) - J)$ and the skew-symmetry of $J$, which leads to $\langle J\partial_t u, \partial_t u\rangle = 0$. Integration of (2.25) with respect to time results in the following energy balance

$$\mathcal{H}(u(t)) - \mathcal{H}(u(s)) = -\int_s^t \langle R(\partial_t u(\tau))\partial_t u(\tau), \partial_t u(\tau)\rangle \, d\tau + \int_s^t \langle f(\tau), \partial_t u(\tau)\rangle \, d\tau. \quad (2.26)$$

With the assumption that $R(\partial_t u)$ is positive semi-definite, we obtain

$$\mathcal{H}(u(t)) - \mathcal{H}(u(s)) \leq \int_s^t \langle f(\tau), \partial_t u(\tau)\rangle \, d\tau,$$

which holds for any input $f$ and the corresponding solution $u$ of (2.24). Hence, the energy of the system can only grow by the energy supplied through the sources, proving the passivity.

**Remark 2.3.3.** The balance (2.25) can be directly applied to the MONA formulation (2.19), which leads to the following power balance

$$\frac{d}{dt}\mathcal{H}(\psi, q_C, q_V) = -\langle G(A_R^\top \partial_t \psi) A_R^\top \partial_t \psi, A_R^\top \partial_t \psi\rangle - \langle v_{src}, \partial_t q_V\rangle - \langle i_{src}, A_I^\top \partial_t q_I\rangle. \quad (2.27)$$

For positive definite conductivity matrices $G(A_R^\top \partial_t \psi)$, the MONA systems are passive.

Utilizing the particular structure of the problem and following the framework presented in [43], we now formulate a discretization strategy based on Petrov-Galerkin time-stepping that allows the construction of schemes preserving the underlying energy balance.

#### Structure preserving discretization

Let $d$ denote the size of the system, i.e., $u(t) \in \mathbb{R}^d$, and recall the notation of Section 2.2.3. We now consider the approximation of the system (2.24) by the following method.

**Problem 2.3.4.** Find $u \in P_{k+1}(\mathcal{T}; \mathbb{R}^d) \cap C([0, T; \mathbb{R}^d])$ with $u(0) = u_0$ and

$$\begin{aligned} &\int_{I^n} \langle (R(\partial_t u(t)) - J)\partial_t u(t), \bar{v}(t)\rangle \, dt \\ &= -\int_{I^n} \langle \mathcal{H}'(u(t)), \bar{v}(t)\rangle \, dt + \int_{I^n} \langle f(t), \bar{v}(t)\rangle \, dt, \end{aligned} \quad (2.28)$$

for all $\bar{v} \in P_k(I^n; \mathbb{R}^d)$ and $1 \leq n \leq N$.

The scheme is a finite-dimensional version of [43, Approach 3.1] and Scheme A.14 adapted to the problem under the investigation. The discrete energy balance is provided in the following lemma.

**Lemma 2.3.5.** Any solution $u$ of Problem (2.28) satisfies

$$\mathcal{H}(u(t^n)) - \mathcal{H}(u(t^m)) = -\int_{t^m}^{t^n} \langle R(\partial_t u(t))\partial_t u(t), \partial_t u(t)\rangle \, dt + \int_{t^m}^{t^n} \langle f(t), \partial_t u(t)\rangle \, dt$$

for all $0 \leq t^m \leq t^n \leq T$.

*Proof.* Proof of the statement is identical to that of [43, Theorem 2], also presented in Lemma A.2.4 in Appendix A.2. □

Thus, for the MONA system (2.19) the energy balance (2.27) is exactly preserved under the proposed discretization. Therefore, the method is particularly well suited for energy-conserving circuits.

#### A note on numerical realization

The approach allows the construction of arbitrary higher-order schemes. Details on a possible realization have already been discussed in Section 1.2.4. The details on the choice of an appropriate quadrature rule for the integration are discussed in Section 2.5

## 2.4. Magnetic oriented formulation for field-circuit coupling

The circuit models discussed in the previous sections are based on simplified descriptions of resistors, capacitors, inductors, and power sources. For more complex circuit elements, electromagnetic field models should be used to obtain an adequate description of their behaviour. In the following, we call these types of elements simply *field elements*. The modelling of circuits containing field elements leads to problems with field-circuit coupling, which have been addressed in a variety of publications; see e.g. [13, 117, 138, 140] and references are given there. We also refer to [13, 37, 55] for the index characterization of coupled problems. In this section, we consider electromagnetic field elements modelled by the eddy current approximation of Maxwell's equations. We introduce a magnetic-oriented formulation for the corresponding field-circuit coupled problems, which is based on the MONA formulation for the circuit and the magnetic vector potential formulation for the field equations. We show that this formulation leads to systems with the canonical structure (2.24). Hence, energy-stable discretization can be achieved by the Petrov-Galerkin technique discussed above. For the coupling between circuit and field quantities, we consider *stranded* and *solid* conductor models discussed in e.g. [14, 119]. The results of this section have not been published yet.

### 2.4.1. Coupling through stranded conductor

Assume that the current is injected into the field element through a stranded conductor. Before we proceed, let us briefly recall some basic details about voltage and current excitation in the field element. We mainly follow [68, 119], where further details are provided.

#### Voltage-current excitation problem

The stranded conductor model is based on the assumption that the current density is a multiple of the current, i.e. $\mathbf{j}_s = i_M \mathbf{j}_0$, where $\mathbf{j}_s$ is the current density within the stranded conductor, $i_M$ denotes the total current, and $\mathbf{j}_0$ is a given *winding function*, also known as *generator current density*. Then, the current excitation problem for a single conductor in a bounded domain $\Omega$ reads

$$\sigma \partial_t \mathbf{a} + \operatorname{curl} \nu(\operatorname{curl}(\mathbf{a})) \operatorname{curl} \mathbf{a} = i_M \mathbf{j}_0, \tag{2.29}$$

where $\mathbf{a}$ is the magnetic vector potential, $\sigma$ is the electric conductivity, $\nu(\cdot)$ is differential reluctance; see [119, 128] for details. The expression

$$\langle \mathbf{j}_0, \partial_t \mathbf{a} \rangle_\Omega + r i_M = v_M \tag{2.30}$$

for the voltage can be derived by comparing the balances of power, as discussed in [68, 128]. With $\langle \cdot, \cdot \rangle_\Omega$ we denote the $L^2$ scalar product $\langle \mathbf{v}, \mathbf{w} \rangle_\Omega = \int_\Omega \mathbf{v} \cdot \mathbf{w} \, d\mathbf{x}$. The quantity $r > 0$ denotes the Ohmic resistance of the stranded conductor. The formulation (2.29)–(2.30) is sometimes called the $\mathbf{a}^*-$ formulation for a stranded conductor; see e.g. [55, 119]. For simplicity, we assume perfect magnetic boundary condition $\nu(\operatorname{curl} \mathbf{a}) \operatorname{curl} \mathbf{a} \times \mathbf{n} = 0$ on $\partial\Omega$.

A Galerkin discretization of (2.29)-(2.30) in space by an appropriate finite element approximation incorporating gauging conditions leads to a system of the form

$$\mathsf{M}_\sigma \partial_t a + \mathsf{K}(a) a = \mathsf{B} i_M, \tag{2.31}$$

$$\mathsf{B}^\top \partial_t a + r i_M = v_M, \tag{2.32}$$

where $a(t)$ is the vector with coefficients with respect to the chosen finite element basis. Since $\sigma = 0$ in the nonconducting region, the mass matrix $M_\sigma$ is singular and positive semi-definite by construction.

We further assume that the differential reluctivity $\nu(\cdot)$ is defined through the energy relation $w'_{mag}(\mathbf{b}) = \nu(\mathbf{b})\mathbf{b}$, where $w_{mag}(\mathbf{b})$ is the magnetic energy density; see Section 1.1. With $\epsilon_M(a) = w_{mag}(\operatorname{curl} \mathbf{a}_h)$ we define the discrete energy as a function of coefficients with respect to the chosen finite element discretization. Then the relation $\epsilon'_M(a) = \mathsf{K}(a)a$ holds by definition of system matrix $\mathsf{K}(\cdot)$.

**Remark 2.4.1.** Modelling of the multi-port field elements, i.e. the elements, where current is injected through several disjoint conductors, leads to the systems of the same form (2.31)–(2.32), where $\mathsf{B}$ and $r$ are matrices and $i_M$ and $v_M$ are vector-valued.

### Field-circuit coupling

Now we consider the circuits consisting of capacitors (C), resistors (R), inductors (L), current (I) and voltage (V) sources, and an additional field element (M) described by (2.31)-(2.32). We denote with $A_M$ the block of the incidence matrix corresponding to the field element. Similarly to Section 2.3.1, we write Kirchhoff's current law (KCL) as

$$A_C \partial_t q_C + A_R \partial_t q_R + A_L \partial_t q_L + A_V \partial_t q_V + A_I \partial_t q_I + A_M \partial_t q_M = 0, \tag{2.33}$$

where $i_X = \partial_t q_X$ for different element types. With Kirchhoff's voltage law $v_X = A_X^\top \partial_t \psi$, the constitutive relations for capacitors, inductors, and resistors are stated as in (2.18), namely,

$$A_C \partial_t \psi = \epsilon'(q_C), \quad \partial_t q_L = \epsilon'_L(A_L^\top \psi), \quad \text{and} \quad \partial_t q_R = G(A_R^\top \partial_t \psi) A_R^\top \partial_t \psi, \tag{2.34}$$

while for voltage and current sources we require $A_V^\top \partial_t \psi = v_{src}$ and $\partial_t q_I = i_{src}$. The system describing the field element is obtained from (2.31)-(2.32) and reads

$$\begin{aligned} \mathsf{M}_\sigma \partial_t a - \mathsf{B} \partial_t q_M &= -\epsilon'_M(a), \\ -A_M^\top \partial_t \psi + \mathsf{B}^\top \partial_t a + r \partial_t q_M &= 0, \end{aligned}$$

where $A_M^\top \partial_t \psi = v_M$ and $\partial_t q_M = i_M$. Putting everything together finally leads to

$$\begin{pmatrix} A_R G A_R^\top & 0 & A_M & A_C & A_V \\ 0 & \mathsf{M}_\sigma & -\mathsf{B} & 0 & 0 \\ -A_M^\top & \mathsf{B}^\top & r & 0 & 0 \\ -A_C^\top & 0 & 0 & 0 & 0 \\ -A_V^\top & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial_t \psi \\ \partial_t a \\ \partial_t q_M \\ \partial_t q_C \\ \partial_t q_V \end{pmatrix} = - \begin{pmatrix} A_L \epsilon_L'(A_L^\top \psi) \\ \epsilon_M'(a) \\ 0 \\ \epsilon_C'(q_C) \\ 0 \end{pmatrix} - \begin{pmatrix} A_I i_{src} \\ 0 \\ 0 \\ 0 \\ v_{src} \end{pmatrix}, \tag{2.35}$$

with $G = G(A_R^\top \partial_t \psi)$ denoting the voltage dependent conductivity matrix for the resistors. Since the formulation is based on magnetic vector potential and magnetic node potentials, we call (2.35) a *magnetic oriented formulation* for field-circuit coupling.

### Geometric structure of the coupled problem

The total energy now consists of electric energy stored in the capacitors and magnetic energy stored in inductors and the field element. With $\mathcal{H}(\psi, a, q_M, q_C, q_V) = \epsilon_L(A_L^\top \psi) + \epsilon_M(a) + \epsilon_C(q_C)$ we denote the energy in terms of the system variables. The coupled system (2.35) has the same geometric structure as MONA, namely

$$(R(\partial_t u) - J)\partial_t u = -\mathcal{H}'(u) + f, \tag{2.36}$$

with state vector $u = (\psi, a, q_M, q_C, q_V)$, source vector $f = (-A_I i_{src}, 0, 0, 0, -v_{src})$ and system matrices

$$R = \begin{pmatrix} A_R G A_R^\top & 0 & 0 & 0 & 0 \\ 0 & \mathsf{M}_\sigma & 0 & 0 & 0 \\ 0 & 0 & r & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} 0 & 0 & A_M & A_C & A_V \\ 0 & 0 & -\mathsf{B} & 0 & 0 \\ -A_M^\top & \mathsf{B}^\top & 0 & 0 & 0 \\ -A_C^\top & 0 & 0 & 0 & 0 \\ -A_V^\top & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{2.37}$$

The skew-symmetry of $J$ is obvious. Assuming that $G = G(A_R^\top \partial_t \psi)$ is symmetric positive definite as before, using that $\mathsf{M}_\sigma$ is symmetric positive semi-definite, and noting that $r > 0$, the matrix $R(\partial_t u)$ can be seen to be symmetric and positive semi-definite. Hence, with the same arguments as used in Section 2.3.3, we obtain the energy balance, as stated in the following corollary.

**Corollary 2.4.2.** Let $u = (\psi, a, q_M, q_C, q_V)$ be a sufficiently smooth solution to (2.36), where $J$ and $R(\partial_t u)$ are given as in (2.37). Then the energy balance (2.26) holds, namely

$$\mathcal{H}(u(t)) - \mathcal{H}(u(s)) = -\int_s^t \langle R(\partial_t u(\tau))\partial_t u(\tau), \partial_t u(\tau)\rangle \, d\tau + \int_s^t \langle f(\tau), \partial_t u(\tau)\rangle \, d\tau,$$

where the dissipative term is now given by

$$\langle R(\partial_t u)\partial_t u, \partial_t u\rangle = \langle G(A_R^\top \partial_t \psi) A_R^\top \partial_t \psi, A_R^\top \partial_t \psi\rangle + \langle r \partial_t q_M, \partial_t q_M\rangle + \langle \mathsf{M}_\sigma \partial_t a, \partial_t a\rangle. \tag{2.38}$$

The three summands in (2.38) correspond to Ohmic losses caused by circuit resistors, losses caused by the resistance of the stranded conductor, and due to eddy currents within the conducting domain of the field element, respectively. With the assumption that $R(\partial_t x)$ is positive semi-definite, we obtain the passivity of the coupled problem (2.35), as discussed in Section 2.3.3.

**Remark 2.4.3.** Because of the particular problem structure (2.36), the construction of schemes, which preserve the underlying energy balance for the coupled problems, can be achieved using the Petrov-Galerkin approach, as discussed in the previous section.

### 2.4.2. Coupling through solid conductor

We now turn to the case where the current is injected into the field element through a solid conductor, which leads to a system with a slightly different structure. We follow [68, Section 5] for modelling voltage and current excitation in the field element. For further details, we also refer to [5, 41, 128].

#### Voltage-current excitation problem

We consider the following model for voltage and current excitation through a solid conductor

$$\sigma \partial_t \mathbf{a} + \operatorname{curl} \nu(\operatorname{curl} \mathbf{a}) \operatorname{curl} \mathbf{a} = v_M \sigma \mathbf{p}, \tag{2.39}$$

$$-\langle \partial_t \mathbf{a}, \mathbf{p} \rangle_\sigma + v_M \|\mathbf{p}\|_\sigma^2 = i_M, \tag{2.40}$$

where $\mathbf{p}$ is a *winding function* representing a normalized electric field distribution in the conductor; see e.g. [119]. We again consider a bounded domain $\Omega$ and assume boundary condition $\nu(\operatorname{curl} \mathbf{a}) \operatorname{curl} \mathbf{a} \times \mathbf{n} = 0$ on $\partial\Omega$.

The Galerkin discretization of (2.39)-(2.40) in space using an appropriate finite element subspace incorporating gauging conditions leads to a system

$$\mathsf{M}_\sigma \partial_t a + \mathsf{K}(a) a = -\mathsf{B} v_M, \tag{2.41}$$

$$\mathsf{B}^\top \partial_t a + \mathsf{P} v_M = i_M. \tag{2.42}$$

Note that for multi-port field elements, $\mathsf{B}$ and $\mathsf{P}$ are matrices and $i_M$ and $v_M$ are vector-valued. System (2.41)-(2.42) looks similar to (2.31)-(2.32), but its analysis requires somewhat different arguments.

#### Field-circuit coupling

Similar to the analysis of the previous section, we use Kirchhoff's voltage law $v_M = A_M^\top \partial_t \psi$ and relation $i_M = \partial_t q_M$ to write the system (2.41)-(2.42) as

$$\begin{aligned} \mathsf{B} A_M^\top \partial_t \psi + \mathsf{M}_\sigma \partial_t a &= -\epsilon_M'(a), \\ \mathsf{P} A_M^\top \partial_t \psi + \mathsf{B}^\top \partial_t a &= \partial_t q_M, \end{aligned}$$

where $\epsilon_M(a)$ is the magnetic energy and $\epsilon_M'(a) = \mathsf{K}(a)a$ holds. Together with Kirchhoff's current law (2.33) and the constitutive relations (2.34) for the circuit components, we obtain the *magnetic oriented formulation*

$$\begin{pmatrix} A_R G A_R^\top + A_M \mathsf{P} A_M^\top & A_M \mathsf{B}^\top & A_C & A_V \\ \mathsf{B} A_M^\top & \mathsf{M}_\sigma & 0 & 0 \\ -A_C^\top & 0 & 0 & 0 \\ -A_V^\top & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial_t \psi \\ \partial_t a \\ \partial_t q_C \\ \partial_t q_V \end{pmatrix} = - \begin{pmatrix} A_L \epsilon_L'(A_L^\top \psi) \\ \epsilon_M'(a) \\ \epsilon_C'(q_C) \\ 0 \end{pmatrix} - \begin{pmatrix} A_I i_{src} \\ 0 \\ 0 \\ v_{src} \end{pmatrix}.$$

Like before, $G = G(A_R^\top \partial_t \psi)$ denotes the voltage-dependent conductivity matrix for the resistive elements. In contrast to the coupling via stranded conductor, the quantity $q_M$ could be eliminated and does not appear in the system.

#### Geometric structure of the coupled problem

The total energy of the field-circuit coupled system is given by $\mathcal{H}(\psi, a, q_C, q_V) = \epsilon_C(q_C) + \epsilon_L(A_L^\top \psi) + \epsilon_M(a)$, and the coupled problem is again of the canonical structure

$$(R(\partial_t u) - J)\partial_t u = -\mathcal{H}'(u) + f, \tag{2.43}$$

with state vector $u = (\psi, a, q_C, q_V)$, source vector $f = (-A_I i_{src}, 0, 0, -v_{src})$, and system matrices

$$R(\partial_t u) = \begin{pmatrix} A_R G(A_R^\top \psi) A_R^\top + A_M \mathsf{P} A_M^\top & A_M \mathsf{B}^\top & 0 & 0 \\ \mathsf{B} A_M^\top & \mathsf{M}_\sigma & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} 0 & 0 & A_C & A_V \\ 0 & 0 & 0 & 0 \\ -A_C^\top & 0 & 0 & 0 \\ -A_V^\top & 0 & 0 & 0 \end{pmatrix}. \tag{2.44}$$

The skew-symmetry of the matrix $J$ immediately follows from its structure. The positive semi-definiteness of $R(\partial_t x)$ follows from positive semi-definiteness of matrix $\tilde{G}$ given by

$$\tilde{G} = \begin{pmatrix} A_M \mathsf{P} A_M^\top & A_M \mathsf{B}^\top \\ \mathsf{B} A_M^\top & \mathsf{M}_\sigma \end{pmatrix},$$

as shown in the following lemma.

**Lemma 2.4.4.** The matrix $\tilde{G}$ is positive semi-definite.

*Proof.* By construction of $\mathsf{M}_\sigma$ we have $\langle \mathsf{M}_\sigma y, y \rangle = \langle \mathbf{y}_h, \mathbf{y}_h \rangle_\sigma$, where $\mathbf{y}_h$ is a function in the finite element subspace associated with the vector of coefficients $y$. In a similar manner we can write $\langle \mathsf{P} z, z \rangle = \langle z\mathbf{p}, z\mathbf{p} \rangle_\sigma$ and $\langle \mathsf{B}^\top y, z \rangle = -\langle \mathbf{y}_h, z\mathbf{p} \rangle_\sigma$ by construction of $\mathsf{P}$ and $\mathsf{B}$, respectively. Therefore, we have

$$\begin{aligned} \langle \tilde{G} y, y \rangle &= \langle \mathsf{M}_\sigma y^1, y^1 \rangle + \langle \mathsf{P} y^2, y^2 \rangle + 2\langle \mathsf{B}^\top y^1, y^2 \rangle \\ &= \langle \mathbf{y}_h^1, \mathbf{y}_h^1 \rangle_\sigma + \langle y^2 \mathbf{p}, y^2 \mathbf{p} \rangle_\sigma - 2\langle \mathbf{y}_h^1, y^2 \mathbf{p} \rangle_\sigma = \| \mathbf{y}_h^1 - y^2 \mathbf{p} \|_\sigma^2 \geq 0, \end{aligned}$$

which proves that $\tilde{G}$ is positive semi-definite. □

**Remark 2.4.5.** In the proof we used that $\mathsf{M}_\sigma$, $\mathsf{B}$, and $\mathsf{P}$ are constructed with respect to the same scalar product $\langle \cdot, \cdot \rangle_\sigma$. This must be taken into account if the inexact integration by e.g. mass lumping is considered.

Since the coupled field-circuit problem is again of the canonical structure (2.43), we can use the arguments of Section 2.3.3 and derive the energy balance for the coupled problem as stated in the following corollary.

**Corollary 2.4.6.** Let $u = (\psi, a, q_C, q_V)$ be a sufficiently smooth solution to (2.43), where $J$ and $R(\partial_t u)$ are given as in (2.44). Then the energy balance (2.26) holds, namely

$$\mathcal{H}(u(t)) - \mathcal{H}(u(s)) = -\int_s^t \langle R(\partial_t u(\tau))\partial_t u(\tau), \partial_t u(\tau)\rangle \, d\tau + \int_s^t \langle f(\tau), \partial_t u(\tau)\rangle \, d\tau,$$

where the dissipative term is now given by

$$\langle R(\partial_t u)\partial_t u, \partial_t u\rangle = \langle G(A_R^\top \partial_t \psi) A_R^\top \partial_t \psi, A_R^\top \partial_t \psi\rangle + \| - \partial_t \mathbf{a}_h + A_M^\top \partial_t \psi \mathbf{p}\|_\sigma^2. \quad (2.45)$$

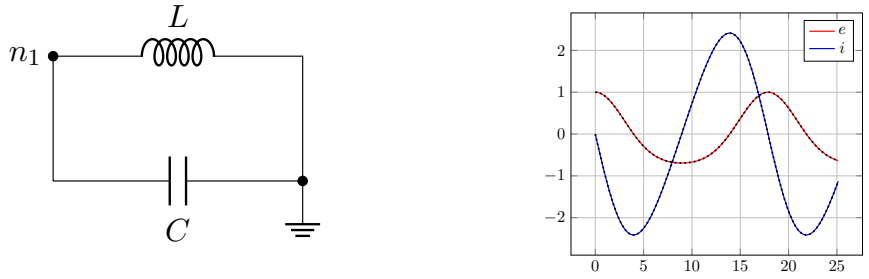
The latter summand in (2.45) corresponds to the approximation of the Joule losses $\|\mathbf{e}\|_\sigma^2$; see [68, 128] for details. The construction of the energy balance preserving scheme can then again be achieved using the Petrov-Galerkin approach.

## 2.5. Numerical illustration

To illustrate some of the discussed aspects, we provide three numerical examples. In the first test case, we consider a nonlinear LC circuit – the most simple example of an energy-preserving system. We illustrate the energy behavior of the presented schemes and investigate their convergence rates. In the second test case, we discuss the particular numerical challenges arising in the index-2 circuits. We illustrate that hidden constraints may cause instabilities when initial values are chosen inappropriately and verify the loss of convergence in algebraic variables. In the third numerical example, we consider a rectifier circuit in which the transformer is modelled by magneto-quasistatic field equations. We illustrate the preservation of energy balance by the proposed discretization when solving the coupled nonlinear field-circuit problem.

### Example 1: Nonlinear LC circuit

Consider a simple circuit consisting of an inductor connected to a capacitor as illustrated in Figure 2.3a. This is a canonical example of an energy-conserving system – the electric energy of capacitors transforms into the magnetic energy of the coil and vice versa while the total energy of the system is conserved.




(a) LC circuit (b) Potential and current

Figure 2.3.: Schematic representation of the circuit and numerical solutions obtained by the proposed discretization schemes for MNA (solid) and MONA (dashed) formulations.

**The MNA formulation.** As the constitutive equation for the capacitor, we consider a nonlinear device model given by $C(v) = c_0(1+\frac{v}{v_0})^{-\gamma}$ with parameters $c_0 = 7.575$, $v_0 = 0.8$, and $\gamma = 0.45$ (slightly modified [1, SMV1235]). For the inductor, we simply assume $L = 1$. The ground node is set to zero potential. With $A_C = 1$ and $A_L = -1$ the MNA system for the LC circuit 2.3a reads

$$C(e)\partial_t e - i = 0, \tag{2.46}$$

$$L\partial_t i + e = 0, \tag{2.47}$$

where $e$ is the potential at the node $n_1$ and $i$ denotes the current through the inductor. This particular problem is of index-0, i.e. is a system of ordinary differential equations. Since there are no algebraic constraints, the choice of initial values is arbitrary. We set

$$e(0) = 1 \quad \text{and} \quad i(0) = 0 \tag{2.48}$$

as an initial condition. Figure 2.3b illustrates the numerical solutions $e(t)$ and $i(t)$ over the time interval $t \in [0, 8\pi]$.

**Results for the MNA system.** For our numerical tests, we used the dG schemes (2.17) with polynomial order $k = 0, 1, 2$. As the basis, we take Lagrange polynomials associated with Gauss Lobatto Legendre nodes. The time integration is performed with Gauss quadrature of a sufficiently high order, such that the integration error becomes insignificant. The nonlinear systems (2.17) in every time step are solved numerically with a tolerance of $10^{-15}$. Table 2.1 illustrates the convergence results for the electric potential $e$. The reported errors are computed by $\text{err} = \max_{0\le n\le N} |e^n(t^n) - \tilde{e}(t^n)|$, where $\tilde{e}$ is a

| $\tau$ | $k=0$ | | $k=1$ | | $k=2$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-1}$ | e.o.c. | err $\times 10^{-3}$ | e.o.c. | err $\times 10^{-6}$ | e.o.c. |
| 1 | 2.244 | – | 1.322 | – | 2.12225 | – |
| 0.5 | 1.199 | 0.90 | 0.173 | 2.92 | 0.06963 | 4.93 |
| 0.25 | 0.542 | 1.15 | 0.022 | 2.98 | 0.00221 | 4.97 |
| 0.125 | 0.184 | 1.55 | 0.003 | 3.15 | 0.00007 | 5.03 |

Table 2.1.: Convergence of the schemes (2.17) applied to (2.46)-(2.47).

numerical solution computed with a sufficiently fine discretization. Similar to the related RadauIIA schemes with $s = k+1$ stages [4], we observe super convergence results $O(\tau^{2k+1})$; see [66]. Similar results also hold for the current $i$.

**The MONA formulation.** The MONA formulation for the LC circuit 2.3a reads

$$\partial_t q = -\epsilon'(L), \tag{2.49}$$

$$-\partial_t \psi = -\epsilon'_C(q), \tag{2.50}$$

where $\psi(t)$ is the magnetic potential at the node $n_1$ and $q_C(t)$ is the capacitor charge. The constitutive laws for the capacitor and the inductor are given by

$$\epsilon'_L(\psi) = L^{-1}\psi \quad \text{and} \quad \epsilon'_C(q) = v_0\left(\left(\tfrac{1-\gamma}{v_0 c_0}q + 1\right)^{\frac{1}{1-\gamma}} - 1\right),$$

respectively. The expression for the capacitor is determined analytically through relations, discussed in Section 2.1.3. It is easy to observe that the MONA formulation leads to a system of ordinary differential equations, hence the initial values can be chosen arbitrarily. For the purpose of comparison to the MNA, we chose $\psi(0) = L^{-1}i(0) = 0$ and $q(0) = \frac{c_0 v_0}{1-\gamma}((1+\frac{1}{v_0})^{1-\gamma} - 1) \approx 22.58$, which correspond to initial values (2.48). The solutions $e = \partial_t \psi$ and $i = \partial_t q$ are depicted by the dashed lines in Figure 2.3b.

**Results for the MONA system.** In the numerical tests, we use the Petrov-Galerkin approach for $k = 1, 2, 3$. We use Lagrange polynomials associated with Gauss Lobatto Legendre nodes as a basis and utilize higher degree Gauss quadrature for integration. The resulting nonlinear systems (2.28) in every time step are solved with a tolerance $10^{-15}$. Table 2.2 illustrates the convergence results for the magnetic potential $\psi$, with the error calculated via $\text{err} = \max_{0\le n\le N} |\psi^n(t^n) - \tilde{\psi}(t^n)|$, where $\tilde{\psi}$ is a numerical solution with a sufficiently small time step. Similar to the related LobattoIIIA methods with $s = k + 1$ stages, we observe the convergence rates $O(\tau^{2k})$; see [66].

| $\tau$ | $k=1$ | | $k=2$ | | $k=3$ | |
|---|---|---|---|---|---|---|
| | err $\times 10^{-2}$ | e.o.c. | err $\times 10^{-5}$ | e.o.c. | err $\times 10^{-8}$ | e.o.c. |
| 1 | 2.249 | – | 5.688 | – | 6.95031 | – |
| 0.5 | 0.576 | 1.96 | 0.359 | 3.98 | 0.11069 | 5.97 |
| 0.25 | 0.138 | 2.06 | 0.022 | 3.99 | 0.00175 | 5.98 |
| 0.125 | 0.028 | 2.32 | 0.001 | 4.08 | 0.00004 | 5.58 |

Table 2.2.: Convergence of the schemes (2.28) applied to (2.49)-(2.50).

The electric node potentials and currents are computed by differentiation in the post-processing. Since a collocation approach is used, the differentiation can be done exactly without a loss of convergence for the differentiated quantities.

**Evolution of the energy.** To illustrate the discrete energy balance and passivity of the two approaches, we consider schemes with a larger time step $\tau = 1$. The red line in Figure 2.4 shows the evolution of the energy for the MNA system (2.46)-(2.47) discretized by the discontinuous Galerkin approach (2.17) with $k = 1$. As expected the energy decays

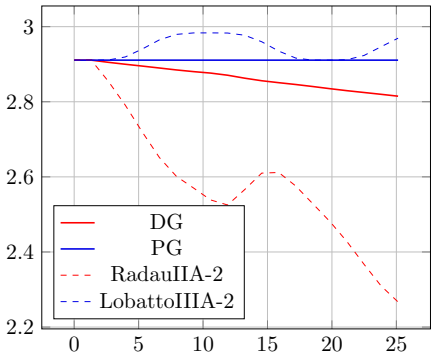



Figure 2.4.: Evolution of energies

over time. This is not the case for discretization by the corresponding RadauIIA scheme

with $s = 2$; the evolution of energy is depicted with a red dashed line in Figure2.4. For the energy calculation, we use the analytic expression

$$\mathcal{E}(e,i) = \frac{Li^2}{2} + \frac{c_0 v_0}{1-\gamma}\left(1+\frac{e}{v_0}\right)^{1-\gamma} e - \frac{c_0 v_0^2}{(1-\gamma)(2-\gamma)}\left(\left(1+\frac{e}{v_0}\right)^{2-\gamma} - 1\right).$$

The evolution of the energy of the MONA system (2.49)-(2.50) discretized by the Petrov-Galerkin approach (2.28) is depicted with a blue line in Figure 2.4. For the calculation of energy, we use the analytic expression

$$\mathcal{H}(\psi,q) = \frac{L^{-1}\psi^2}{2} + \frac{c_0 v_0^2}{2-\gamma}\left(\left(\frac{1-\gamma}{v_0 c_0} q + 1\right)^{\frac{2-\gamma}{1-\gamma}} - 1\right) - v_0 q.$$

As expected, the energy of the system is conserved up to an accumulated error of machine precision $O(10^{-13})$. This is not the case for discretization using the LobattoIIIA scheme.

### Example 2: Linear index-2 circuit; [122, Example 1]

The next example illustrates the impact of a higher index on the numerical solution. We consider the circuit illustrated in Figure 2.5, an example taken from [62, Ch. 10]. Because

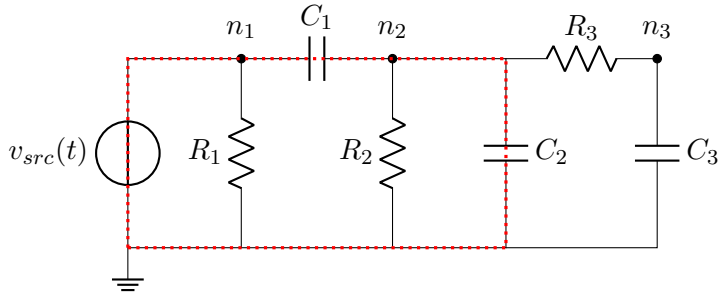



Figure 2.5.: Circuit containing a CV loop from [62, Fig. 10.2].

of the CV loop, depicted by the red dashed line, the conventional MNA leads to a problem with index $\nu = 2$, while the MONA approach results in a system with index $\nu = 1$. The circuit graph consists of four nodes connected by seven branches, corresponding to three capacitors, three resistors, and one voltage source. The ground node is considered to be the reference node. So, three potentials at the nodes $n_1$, $n_2$, and $n_3$ are required for the circuit description. The circuit topology is then encoded in the reduced partial incidence matrices

$$A_C = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{pmatrix}, \quad \text{and} \quad A_V = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

For ease of presentation, we simply set $C_i = 1$ and $R_i = 1$ for $i = 1, 2, 3$, which results in the conductivity and capacitance matrices $G = C = I_3$, where $I_3$ is the identity matrix of

size 3. The conventional MNA formulation then leads to the following system

$$\begin{pmatrix} A_C C A_C^T & 0 \\ 0 & 0 \end{pmatrix} \frac{d}{dt} \begin{pmatrix} e \\ i_V \end{pmatrix} + \begin{pmatrix} A_R G A_R^T & A_V \\ -A_V^T & 0 \end{pmatrix} \begin{pmatrix} e \\ i_V \end{pmatrix} = \begin{pmatrix} 0 \\ -v_{src}(t) \end{pmatrix}, \tag{2.51}$$

while the MONA formulation results in

$$\begin{pmatrix} A_R G A_R^T & A_C & A_V \\ -A_C^T & 0 & 0 \\ -A_V^T & 0 & 0 \end{pmatrix} \frac{d}{dt} \begin{pmatrix} \psi \\ q_C \\ q_V \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & C^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \psi \\ q_C \\ q_V \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -v_{src}(t) \end{pmatrix}.$$

Due to our choice of constitutive equations, both systems are linear and time-invariant.

**Instabilities caused by hidden constraints.** Figure 2.6 illustrates some electric quantities obtained by numerical solution of the two equations by the trapezoidal rule (TR) with a fixed time step $\tau = 0.1$ and for $v_{src}(t) = \sin(\pi t)$. For the simulation, we chose

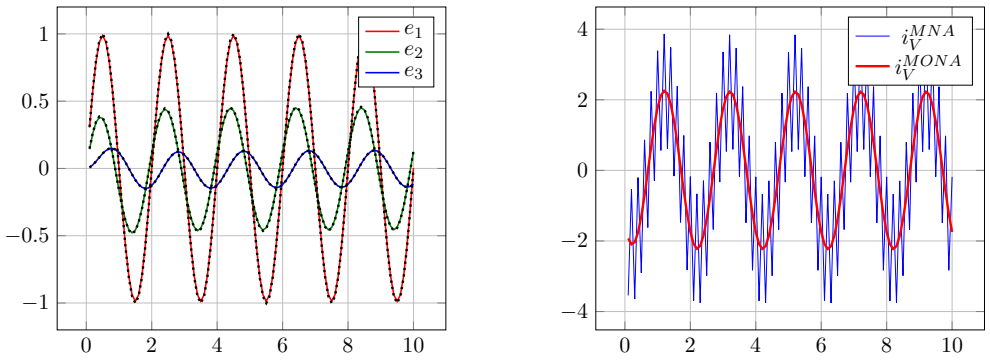



Figure 2.6.: Numerical solutions obtained by TR method applied to the MNA and MONA formulations. Left: potentials (MNA: dotted; MONA: solid); right: current through the voltage source.

trivial initial conditions, which are consistent with the algebraic constraint caused by the voltage source. This suffices to guarantee stability for the index-1 formulation obtained by MONA and, as predicted by the theory, we observe a second-order convergence. The MNA system, on the other hand, has index $\nu = 2$ and an additional hidden constraint arises, which is not satisfied by our choice of initial conditions and causes large oscillations in the algebraic solution component; see Figure 2.6 for illustrative comparison. Let us note that this weak instability could be cured by an appropriate initialization phase by performing the first time step with the implicit Euler method. If we choose the source term $v_{src}(t) = \cos(t)$ inconsistent with the trivial initial values, then the TR-discretization of the MNA formulation leads to strong instabilities which require a longer initialization phase. In contrast, the MONA approach shows a weak instability that can be cured by a single initialization step.

**Convergence reduction for index-2 system.** Loss of convergence (order) is a known issue for index-2 problems; see e.g. [66]. In Table 2.3 we report the convergence rates for the discontinuous Galerkin schemes schemes (2.17) with polynomial degrees $k = 1, 2$

| $\tau$ | $k=1$ | | | | $k=2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $y$ | | $z$ | | $y$ | | $z$ | |
| | err $\times 10^{-3}$ | e.o.c. | err$\times 10^{-1}$ | e.o.c. | err $\times 10^{-5}$ | e.o.c. | err $\times 10^{-2}$ | e.o.c. |
| 1 | 2.1934 | – | 1.7605 | – | 2.6574 | – | 2.7084 | – |
| 0.5 | 0.3009 | 2.87 | 0.5294 | 1.74 | 0.0956 | 4.89 | 0.4195 | 2.69 |
| 0.25 | 0.0395 | 2.93 | 0.1321 | 2.01 | 0.0031 | 4.95 | 0.0544 | 2.95 |
| 0.125 | 0.0045 | 3.14 | 0.0267 | 2.29 | 0.0001 | 5.01 | 0.0061 | 3.15 |

Table 2.3.: Convergence in differential $y$ and algebraic $z$ variables for schemes (2.17) applied to the index-2 system (2.51).

observed in our tests. With $y$ and $z$ we denote the differential and algebraic variable, respectively, which can be determined by appropriate projections. The errors and convergence rates are computed as in the previous example. We observe a super convergence $O(\tau^{2k+1})$ in differential variable $y$, while the algebraic variable $z$ converges with reduced rate $O(\tau^{k+1})$. This result is not surprising since there is an equivalence between the discontinuous Galerkin and RadauIIA schemes for linear circuits [4], and for the latter, the obtained convergence rates are expected; see [66].

### Example 3: Field-circuit coupling

In the last example, we consider a full wave rectifier circuit from [117, Section 6.2] illustrated in Figure 2.7. The transformer is described by the field equations, which leads to a

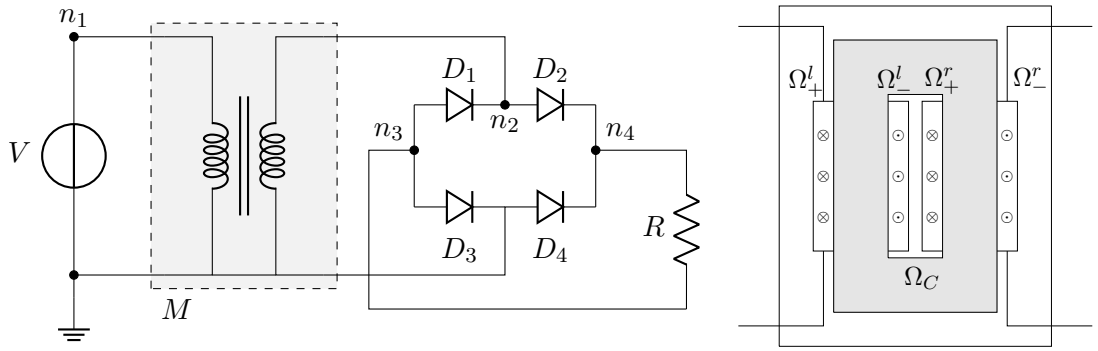



Figure 2.7.: Schematic sketch of a full wave rectifier circuit (left) and geometry of transformer modelled by the field equations (right).

coupled field-circuit problem, discussed in Section 2.4. The topology of the circuit is given by the following partial incidence matrices

$$A_R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & -1 & 1 \end{pmatrix}, \quad A_V = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad A_M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The circuit contains four diodes, which are modelled as nonlinear resistors. We assume nonlinear voltage-current relation $i_D = 2.5\exp(4v_D)v_D$ for diodes and set $R = 1$ for the remaining resistor.

**Description of the field element.** For the description of the transformer we consider a 2D magneto-quasistatic model based on assumptions $\mathbf{j} = (0, 0, j^z)$ and $\mathbf{a} = (0, 0, a^z)$, where $j^z$ and $a^z$ are independent of $z$. The computational domain $\Omega$ is illustrated in Figure 2.7. We set $\sigma = 1$ in $\Omega_C$ and assume $\nu(\mathbf{b}) = \nu_0(1 - \frac{\alpha}{\beta+|\mathbf{b}|})$ in $\Omega$ with $\nu_0 = 1$, $\alpha = 0.5$, and $\beta = 1$. This choice of reluctivity function $\nu(\cdot)$ corresponds to the energy density $\epsilon_{mag}(\mathbf{b}) = \frac{\nu_0^2}{2}(|\mathbf{b}|^2 - \alpha \log(\beta + |\mathbf{b}|^2))$. The current is injected through two stranded conductors with domains $\Omega_{\pm}^l$ and $\Omega_{\pm}^r$. In this setting, the current excitation problem (2.29) simplifies to

$$\sigma \partial_t a^z - \operatorname{div} \nu(\nabla a^z) \nabla a^z = i_l j_{0,l}^z + i_r j_{0,r}^z.$$

The winding functions are defined by $j_{0,l}^z = \chi_{\Omega_+^l} - \chi_{\Omega_-^l}$ and $j_{0,r}^z = \chi_{\Omega_+^r} - \chi_{\Omega_-^r}$ for the left and right conductor, respectively. We neglect the Ohmic resistances of the stranded conductors, which leads to the following expressions for voltages

$$v_l = \langle \partial_t a^z, j_{0,l}^z \rangle \quad \text{and} \quad v_r = \langle \partial_t a^z, j_{0,r}^z \rangle.$$

For the purpose of this section, we use a fixed Galerkin approximation in space by continuous, piece-wise linear finite elements, which leads to the DAE system of the form (2.31)-(2.32). Note, that for this problem in two dimensions, no gauging is necessary.

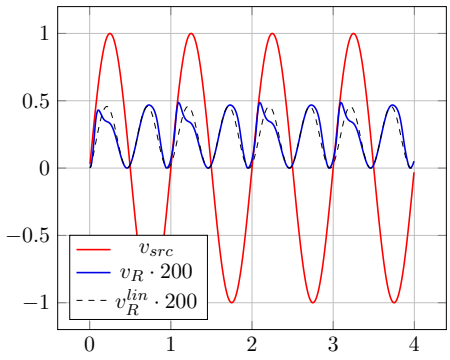
**Simulation results.** Since the circuit contains neither inductors, nor capacitors, nor current sources, the coupled problem (2.35) simplifies to

$$\begin{pmatrix} A_R G A_R^\top & 0 & A_M & A_V \\ 0 & \mathsf{M} & -\mathsf{B} & 0 \\ -A_M^\top & \mathsf{B}^\top & 0 & 0 \\ -A_V^\top & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial_t \psi \\ \partial_t a \\ \partial_t q_M \\ \partial_t q_V \end{pmatrix} = - \begin{pmatrix} 0 \\ \mathsf{K}_\nu(a) a \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ v_{src}(t) \end{pmatrix}.$$

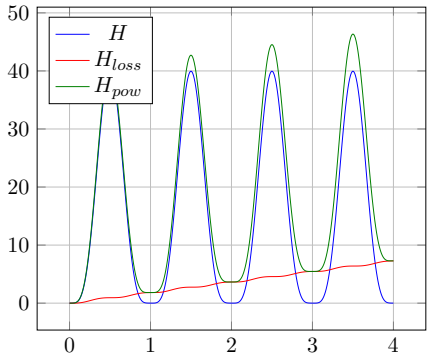
For the time integration, we apply the lowest order Petrov-Galerkin schemes (2.28) with a constant time step $\tau = 0.01$ on the time interval $[0, 4]$. The nonlinear systems in every time step are solved with tolerance $10^{-12}$. The numerical solution for the rectified voltage $v_R(t) = A_R^\top \partial_t \psi$ for the given voltage input $v_{src} = \sin(2\pi t)$ is illustrated in Figure 2.8a. With a dashed line, we plot the solution to the problem with $\nu(\mathbf{b}) = \nu_0$ to highlight the nonlinear effect. Figure 2.8c illustrates the eddy currents $\sigma \partial_t a^z$ inside of the transformer core at the time step $t = 0.15$. Figure 2.8b illustrates the evolution of magnetic energy $H^n = \epsilon_M(a^n(t^n))$ as well as the supplied and dissipated energies given by

$$H_{supp}^n = \sum_{k=1}^{n} \int_{t^{k-1}}^{t^k} v_{src}(t) \partial_t q_V^k \quad \text{and} \quad H_{loss}^n = \sum_{k=1}^{n} \int_{t^{k-1}}^{t^k} \langle M_\sigma a^k, a^k \rangle,$$

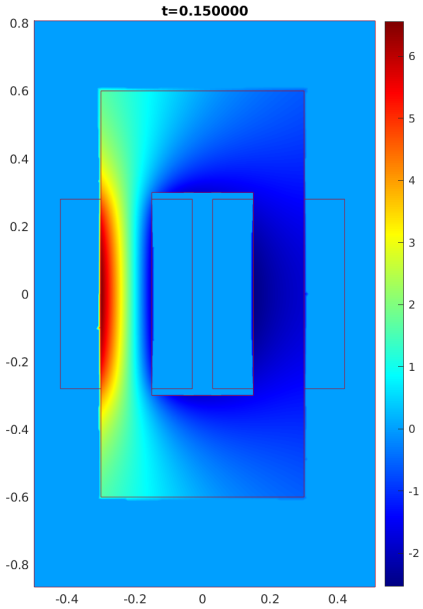
respectively. We list in Table 2.4 the maximal discrepancy in the energy balance $\mathrm{err}_H = \max_n |H^0 + H_{loss}^n - H_{supp}^n - H^n|$ obtained with Petrov-Galerkin and LobattoIIIA time schemes. As expected from theoretical results, for the Petrov-Galerkin approximations, the energy balance $H^n = H^0 + H_{loss}^n - H_{supp}^n$ is conserved up to the tolerance of the nonlinear solver $O(10^{-12})$ and does not depend on the time step width. This is not the case for the related Lobatto schemes, for which the error in energy is correlated to the error of the discretization. Similar results also hold for the coupling through a solid conductor.




(a) Numerical solution for $v_R = A_R^\top \partial_t \psi$.




(b) Evolution of energy.




(c) Eddy current $\sigma \partial_t a^z$ in the transformer core at time step $t = 0.15$.

Figure 2.8.: (a) Numerical solution for $v_R = A_R^\top \partial_t \psi$ for $v_{src} = \sin(2\pi t)$. (b) Evolution of energy $H$, supplied energy $H_{supp}$, and dissipated energy $H_{loss}$. (c) Induced current $\sigma \partial_t a^z$ at the time $t = 0.15$.

| $\tau$ | PG($k=1$) | Lobatto($s=2$) | PG($k=2$) | Lobatto($s=3$) |
|---|---|---|---|---|
| 0.1 | $1.9895{\cdot}10^{-13}$ | $3.523{\cdot}10^{-2}$ | $1.625{\cdot}10^{-13}$ | $2.704{\cdot}10^{-3}$ |
| 0.05 | $1.3567{\cdot}10^{-13}$ | $1.184{\cdot}10^{-2}$ | $2.771{\cdot}10^{-13}$ | $3.382{\cdot}10^{-4}$ |
| 0.025 | $1.8652{\cdot}10^{-13}$ | $3.067{\cdot}10^{-3}$ | $9.037{\cdot}10^{-13}$ | $1.718{\cdot}10^{-5}$ |
| 0.0125 | $6.8923{\cdot}10^{-13}$ | $7.738{\cdot}10^{-4}$ | $7.105{\cdot}10^{-13}$ | $9.891{\cdot}10^{-7}$ |

Table 2.4.: Maximal discrepancy in the energy $\mathrm{err}_H = \max_n |H^0 + H_{loss}^n - H_{supp}^n - H^n|$ obtained by Petrov-Galerkin and LobattoIIIA schemes.

## 2.6. Summary and outlook

We now briefly summarize the main contributions of this chapter and provide an outlook of some open questions for further research. In this chapter, we consider the modelling of electric circuits from an energy-based viewpoint. We show that the conventional MNA formulation leads to systems of a particular port-Hamiltonian structure, which allows the construction of passivity-preserving schemes of arbitrarily high order.

We introduce the MONA formulation, based on magnetic node potentials and charges across capacitors and voltage sources as unknowns. Despite the different modelling per-

spectives, the proposed formulation is suitable for the same general class of circuits. We show that under appropriate assumptions on circuit elements, the MONA formulation leads to a regular DAE system of index $\nu \leq 1$, while the MNA systems have index $\nu \leq 2$. This is a significant advantage of the magnetic-oriented approach. We further show that the MONA formulation leads to a system of a particular geometric structure and discuss the construction of energy-balance-preserving schemes.

Lastly, we formulate the magnetic-oriented formulation for the field-circuit coupling, where the field model is given by the magneto-quasistatic problem in terms of the magnetic vector potential. We show, that under an appropriate Galerkin semi-discretization for the field equations, the coupled problems have the same canonical structure as MONA systems, and the same energy balance-preserving time-stepping strategy can be applied.

The study of the magnetic-oriented approach for circuits is far from complete. The index analysis for the coupled magnetic-oriented problems is missing. Further incorporation of different field models or complicated circuit elements, like transistors and switches is yet to be considered. In particular, the index analysis results of [37] for the circuits containing generalized elements seem to be possible to adopt.

The error analysis of the discussed variational schemes is another topic of future research; the analytical justification for observed super convergence is missing. Further, the efficiency of the schemes has not been optimized. The symbolic analysis techniques can be applied to MNA systems [123, 124], while the extension to MONA seems possible, yet requires further investigation. We further expect that efficient construction of initial values based on topological arguments [50] can be done for MONA systems in a similar manner.

The magnetic-oriented perspective is not necessarily restricted to the nodal analysis. In a similar manner, the *magnetic oriented loop analysis (MOLA)* formulation can be derived. The corresponding analysis of the MOLA formulation is currently in progress.

# Chapter 3.

# Systems with memory

In this chapter, we focus on systems with memory and discuss the numerical treatment of problems that can be modelled by an abstract Volterra-integro-differential equation

$$M(y(t))\partial_t y(t) + N(y(t)) = \int_0^t k(t,s) f(s, y(s))\, ds, \qquad 0 \le t \le T. \tag{3.1}$$

Problems of this form arise in many different applications, such as neural sciences [7], problems with transparent boundary conditions [6, 65, 75, 76], wave propagation problems [6, 39, 65], field-circuit coupling [46], and more; see [27, 28, 90, 114] for an overview.

The main challenge in the numerical treatment of (3.1) arises from the integral term. Its proper realization is the key aspect in the construction of efficient schemes and represents the main topic of this chapter. To keep the reader's attention on the essential parts, we restrict our discussion to a simple Volterra equation of the first kind

$$y(t) = \int_0^t k(t,s) f(s)\, ds, \qquad 0 \le t \le T. \tag{3.2}$$

The discretization strategy presented for (3.2) can then be easily adapted to (3.1). Before we begin, let us briefly highlight the main issues and review related methods.

### The main challenges

The discretization of the integral term (3.2) by an appropriate quadrature rule leads to a matrix-vector multiplication

$$\mathtt{y}_n = (\mathtt{Kf})_n, \qquad 1 \le n \le N, \tag{3.3}$$

with vectors $\mathtt{y}, \mathtt{f} \in \mathbb{R}^N$ and matrix $\mathtt{K} \in \mathbb{R}^{N \times N}$, which is dense and lower block triangular, in general. Since we are interested in the discretization of Volterra-integro-differential problems (3.1), the data $\mathtt{f}_n$ may depend on $\mathtt{y}_n$. Therefore, it is essential that the values $\mathtt{y}_n$ are computed in an evolutionary manner, i.e., where the data $\mathtt{f}_n$ is only required in time steps $m \ge n$. Since $\mathtt{K}$ is a lower triangular matrix, this can be achieved by a textbook multiplication row-by-row, i.e., by traversing the matrix from top to bottom. A naive realization of the matrix-vector product can be done in $O(N^2)$ algebraic operations and requires $O(N)$ active memory to store the history of the solutions $\mathtt{y}_n$. For a sufficiently large $N$, memory consumption represents a major challenge in the simulations. A different realization is based on traversing the matrix $\mathtt{K}$ from left to right. In this case, the entry $\mathtt{f}_n$ is only required in the $n-$th time step and the implementation becomes oblivious. This realization requires $O(N^2)$ arithmetic operations and $O(N)$ active memory to store the partial sums of each row. An essential drawback of this approach is that the number of time steps $N$ has to be fixed a-priori.

### Convolution type integrals

We are particularly interested in problems where the integral in (3.2) is a convolution, i.e.,

$$k(t,s) = k(t-s).$$

In this case, an appropriate discretization leads to an algebraic system (3.3), where the matrix `K` has a block Toeplitz structure. Hence, if the values of `f` and the number of time steps $N$ are fixed, the implementation of the matrix-vector product can be done with complexity $O(N \log N)$ using fast Fourier transform. This realization is not evolutionary, and, it is therefore not suitable for (3.1). An evolutionary version can be realized with $O(N \log^2 N)$ operations; see [67].

Our discussion also involves problems where only the Laplace transform of the convolution kernel $\hat{k}(s)$ is known. In particular, many coupled nonlinear-linear problems can be equivalently formulated as (3.1). Let us consider for example the problem

$$M(y(t))\partial_t y(t) + N(y(t)) = C^\top z, \tag{3.4}$$

$$E\partial_t z + Az = By. \tag{3.5}$$

Assuming the trivial initial condition $z(0) = 0$, we may eliminate the linear part in the frequency domain using the Schur complement technique and obtain a system of the form (3.1), where $f(s, y(s)) = y(s)$ and the convolution kernel $k$ is defined implicitly via its Laplace transform by

$$\hat{k}(s) = C^\top (sE + A)^{-1} B.$$

When the size of the linear system is much larger than the size of the nonlinear system, the formulation (3.1) can have several advantages over (3.4)–(3.5) in the numerical treatment. Such problems arise in e.g. the context of field-circuit coupling; see [46] and Section 3.4.

### Convolution quadrature methods

In the case, where the kernel is given in the frequency domain, the convolution quadrature methods (CQ) introduced in [88, 89, 91] provides a suitable discretization strategy. These methods for integral equations are closely related to the time-stepping schemes for differential equations. In particular situations, one can show that the discrete solution of problem (3.4)–(3.5) by a time stepping scheme and that of (3.1), where the convolution is approximated by the corresponding convolution quadrature, coincide; see [39, 46, 91].

The application of convolution quadrature methods leads to the algebraic problem (3.3). The computation of the entries in the matrix `K` requires $O(N)$ evaluations of $\hat{k}(s)$, which might be computationally expensive. The realization of the discrete convolution requires $O(N)$ active memory and $O(N^2)$ operations if done naively.

An improved fast and oblivious convolution quadrature (FOCQ) approach has been introduced in [85, 92]. The approach provides a fast, evolutionary, and oblivious algorithm of complexity $O(N \log N)$ and uses $O(\log N)$ active memory, and can be understood as a low-rank approximation of matrix `K`; see Section 3.3. Moreover, the algorithm requires only $O(\log N)$ evaluations of $\hat{k}(s)$, which in many applications is the essential factor. The

drawback of the approach is that the number of time steps $N$ has to be known in advance. The fast and oblivious convolution quadrature schemes have been applied in a variety of related problems involving e.g. fractional diffusion [115], impedance and transmission boundary conditions [65], boundary element methods [114], and more.

### Hierarchical approximation methods

An extension of the fast and oblivious CQ algorithm to more general kernels $k(t, s)$ is not directly clear. Matrix approximation methods e.g. fast multipole [54, 59, 113], $\mathcal{H}$- and $\mathcal{H}^2$-matrices [21, 63], multilevel techniques [25, 57] or wavelet algorithms [38], on the other hand, can be applied in this setting. For asymptotically smooth kernels, the matrix $\mathtt{K}$ can be stored efficiently in $O(N \log^\alpha N)$ memory where $\alpha \geq 0$ is some constant. For the convolution-type integrals, the memory requirement is only $O(N)$. If the data $\mathtt{f}_n$ is independent of $\mathtt{y}_n$, the realization of the matrix-vector product can be done with complexity $O(N \log^\alpha N)$ under appropriate smoothness assumptions; see [21, 64] and references within for details. The mentioned realizations are not evolutionary and can not be applied to (3.1) directly.

### Main contributions

We now discuss an algorithm for the realization of Volterra integrals (3.2) or corresponding matrix-vector products (3.3) which has the following important properties, namely, it is

- *evolutionary*: the approximations $\mathtt{y}_n$ can be computed one after another and the number of time steps $N$ and values of $\mathtt{f}$ do not need to be known in advance,
- *oblivious*: the entry $\mathtt{f}_n$ of the right-hand side is only required in the $n-$th step,
- *fast*: the evaluation of all $\mathtt{y}_n$, $1 \leq n \leq N$ requires only $O(N)$ operations, and
- *memory efficient*: the storage of the matrix $\mathtt{K}$ requires only $O(N)$ memory for general kernels and $O(\log N)$ in the case of the convolution. The matrix entities can be computed on the fly, such that only $O(\log N)$ active memory is required to store a compressed history of the data $\mathtt{f}$.

The approach has been published in [40]. The key idea of the method is based on the hierarchical block-wise low-rank approximation for the convolution matrix $\mathtt{K}$ using the polynomial $\mathcal{H}^2-$matrix compression techniques [20, 63]. The accuracy of the approximation can then be guaranteed by well-known approximation results; see [21, 64]. The particular one-dimensional structure of the integration domain allows the explicit characterization of partitioning into blocks in the approximation matrix. This knowledge can be used in the construction of an evolutionary algorithm for matrix-vector multiplication, that traverses the approximation matrix top to bottom with complexity $O(N)$ without the need to fix $N$ in advance.

In the case of convolution kernels, the hierarchical approximation yields compression algorithms for the history of the data $\mathtt{f}$, which reduces memory consumption. The approach shares similar ideas with [6, 11, 75, 76], where a fast multipole expansion was employed to accelerate the *sum of exponentials approach*, or to [78], where a polynomial on growing

time steps was employed for the compression of the data, as well as to [79], where an evolutionary $\mathcal{H}$-matrix approximation with a special low-rank structure was constructed.

We also show that the strategy can be integrated into the convolution quadrature framework [88, 89, 91], where the kernel is accessible via its Laplace transform. The resulting schemes share strong similarities with the fast and oblivious convolution quadrature methods [85, 92]. Moreover, the latter can be understood as a $\mathcal{H}-$matrix approximation with the specific realization of the matrix-vector product.

#### Outline

In Section 3.1 we recall some general approximation results, introduce our basic notation, and state a slightly modified algorithm for the dense evaluation of the Volterra integral operators to illustrate some basic principles that we exploit later on. In Section 3.2 we discuss the partitioning on the domain of integration, the multilevel hierarchy used for the $\mathcal{H}^2$-compression, and the description and analysis of our new algorithm. In Section 3.3 we consider convolution kernels $\hat{k}(s)$ and discuss the relation of our algorithm to Lubich's convolution quadrature and the connections to the fast and oblivious algorithm of [92, 115]. Finally, some numerical results are provided in Section 3.4. To provide a connection to electrical engineering, we consider the application of the approach to field-circuit coupling, similar to [46]. Problems with dispersion can be handled in a similar manner; see [39].

The results of this chapter are based on our publications [39, 40, 46]. Most of the presentation follows closely to [40].

## 3.1. Approximation of Volterra integrals

Let us start by summarizing some necessary information on the discretization of Volterra integral operators. We briefly discuss some general approximation results, introduce the basic notation, and present an algorithm for the uncompressed approximation, which builds the basis for our further discussion. For simplicity, we consider a simple Volterra integral equation of the first kind

$$y(t) = \int_0^t k(t,s) f(s)\, ds, \tag{3.6}$$

with scalar valued functions$y$, $k$, and $f$. The extension to systems of the general form (3.1) is then discussed in Section 3.4.

#### General approximation results

Let $k_h$ and $f_h$ denote suitable approximations of $k$ and $f$. Substituting these approximations in the integral equation (3.6) leads to

$$\widetilde{y}_h(t) = \int_0^t k_h(t,s) f_h(s) ds, \tag{3.7}$$

where $\widetilde{y}_h$ is then the approximation of the solution $y$. The following lemma provides an error bound for the approximation, which is an essential result used in the following.

**Lemma 3.1.1** (see Lemma 1 in [40]). Let $T > 0$, kernels $k, k_h \in L^\infty(0,T;L^r(0,T))$, and $f, f_h \in L^{r'}(0,T)$ be given with $1 \le r, r' \le \infty$ with $1/r + 1/r' = 1$. Furthermore, assume that

$$\|k - k_h\|_{L^\infty(0,T;L^r(0,T))} \le \epsilon \quad \text{and} \quad \|f - f_h\|_{L^{r'}(0,T)} \le \epsilon. \tag{3.8}$$

Then the functions $y$, $\widetilde{y}_h$ defined by (3.6) and (3.7) satisfy

$$\|y - \widetilde{y}_h\|_{L^\infty(0,T)} \le C(\|k\|_{L^r(0,T)} + \|f\|_{L^{r'}(0,T)} + \epsilon)\,\epsilon, \tag{3.9}$$

i.e., the error in the results can be bounded uniformly by the perturbation in the data.

*Proof.* From Hölder's inequality, we can deduce that

$$\begin{aligned} |y(t) - \widetilde{y}_h(t)| &\le \int_0^t |k(t,s)||f(s) - f_h(s)| + |k(t,s) - k_h(t,s)||f_h(s)|ds \\ &\le \|k(t,\cdot)\|_{L^r(0,T)}\|f - f_h\|_{L^{r'}(0,T)} + \|k(t,\cdot) - k_h(t,\cdot)\|_{L^r(0,T)}\|f_h\|_{L^{r'}(0,T)}. \end{aligned}$$

The result then follows by estimating $\|f_h\| \le \|f\| + \|f - f_h\|$, using the estimates for the differences in the data, and taking the supremum over all $0 < t < T$. □

Let us note that the constant $C$ in the estimate (3.9) is independent of $T$. Therefore, it can be used for arbitrarily long time intervals. Using the same arguments, it is also possible to obtain similar estimates can also be obtained for different norms.

The lemma above provides a general approximation result. In the following, we consider a piecewise polynomial approximation, which is often used as a basis in numerical methods for integral, differential, and integro-differential equations.

### 3.1.1. Piecewise polynomial approximation

We consider a uniform grid of the time interval $[0,T]$ with grid points $t^n = nh$, $0 \le n \le N$, where $h$ denotes the constant time step. By $I^n = [t^{n-1}, t^n]$ we denote the $n-$th time interval and we write $\mathcal{T}_h = \{I^n : 1 \le n \le N\}$ for partitioning of the time interval $[0,T]$. We denote by $\mathcal{P}_p(I)$ the space of polynomials of degree at most $p$ over the interval $I$, and we write $\mathcal{P}_{q,q}(I \times J) = \mathcal{P}_q(I) \otimes \mathcal{P}_q(J)$ for the space of polynomials in two variables of degree at most $q$ in each variable. We define piecewise polynomial spaces

$$\begin{aligned} \mathcal{P}_p(\mathcal{T}_h) &= \{f \in L^1(0,T) : f|_{I^n} \in \mathcal{P}_p(I^n)\}, \\ \mathcal{P}_{q,q}(\mathcal{T}_h \times \mathcal{T}_h) &= \{k \in L^1((0,T)\times(0,T)) : k|_{I^m \times I^n} \in \mathcal{P}_{q,q}(I^m \times I^n)\}, \end{aligned}$$

over the grid $\mathcal{T}_h$ and the tensor-product grid $\mathcal{T}_h \times \mathcal{T}_h$.

For sufficiently smooth functions $f$ and $k$, we now consider a picewise polynomial approximations $f_h \in \mathcal{P}_p(\mathcal{T}_h)$ and $k_h \in \mathcal{P}_{q,q}(\mathcal{T}_h \times \mathcal{T}_h)$, which satisfy (3.8) for sufficiently small mesh size $h$. Without any particular knowledge about $f$ and $k$, the use of uniformly partitioned meshes $\mathcal{T}_h$ and $\mathcal{T}_h \times \mathcal{T}_h$ seems plausible. Let us note that for the evaluation of the integral (3.7) only values $k(s,t)$ with $s \le t$ are essential. Therefore, only cells corresponding to $s \le t$ have to be taken into account. The examples of essential cells in the uniformly partitioned grids for the approximation $k_h$ are illustrated in Figure 3.1.

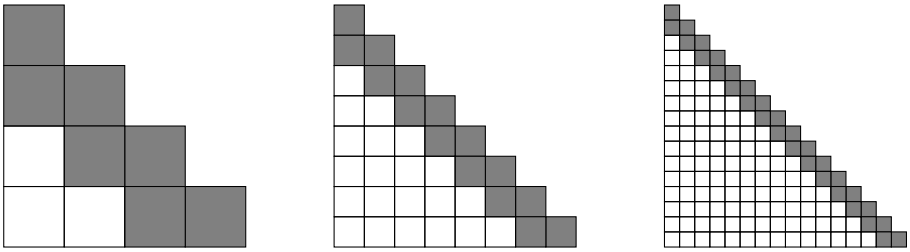


Figure 3.1.: Examples of grids $\mathcal{T}_h \times \mathcal{T}_h$ for approximation of $k$. Only the cells required for approximating $k(t,s)$ for $s \leq t$ are illustrated. The cells near the diagonal are the nearfield cells. They are highlighted in gray and will be treated separately in the following. (See Figure 1 in [40])

We now split the integral (3.7) into two parts corresponding to integrals over the farfield cells and nearfield cells, which are depicted by white and gray in Figure 3.1. We write

$$\widetilde{y}_h(t) = \widetilde{w}_h(t) + \widetilde{z}_h(t),$$

where $\widetilde{w}_h$ denote the contribution from *farfield* and $\widetilde{z}_h$ from the *nearfield* cells. These contributions are treated in a different manner, as discussed in the following passage.

### 3.1.2. Practical realization

With the polynomial approximations $f_h \in \mathcal{P}_p(\mathcal{T}_h)$ and $k_h \in \mathcal{P}_{q,q}(\mathcal{T}_h \times \mathcal{T}_h)$ in (3.7), we can directly conclude that $\widetilde{y}_h \in \mathcal{P}_{p+q+1}(\mathcal{T}_h)$. With the application to Volterra-integro-differential systems (3.1) in mind, it is convenient to look for an approximation of the same degree as the data $f_h$. We replace $\widetilde{y}_h$ in (3.3) by the interpolation $y_h = (P_p\widetilde{y}_h)(t) \in \mathcal{P}_p(\mathcal{T}_h)$ of degree $p$. Following the ideas of collocation schemes, we chose a set of collocation points $t_j^n = t^{n-1} + c_j h$, $j = 0, \ldots, p$ in the interval $I^n$ and use Lagrange interpolation polynomials $\psi_j^n \in \mathcal{P}_p(I^n)$ as the local basis. The approximation $y_h$ is then defined through

$$y_h(t_j^n) = \widetilde{y}_h(t_j^n), \qquad 0 \leq j \leq p. \tag{3.10}$$

We expand the data $f_h$ and the solution $y_h$ locally with respect to this basis as

$$y_h(t) = \sum_{j=0}^{p} y_j^n \psi_j^n(t), \quad f_h(t) = \sum_{j=0}^{p} f_j^n \psi_j^n(t), \quad \text{for } t \in I^n. \tag{3.11}$$

Next, we chose a basis $\{\varphi_i^n\}_{i=0,\ldots,q} \subset \mathcal{P}_q(I^n)$ for the kernel function $k_h \in \mathcal{P}_{q,q}(\mathcal{T}_h \times \mathcal{T}_h)$ and expand it with respect to this basis in each component as follows

$$k_h(s,t) = \sum_{i=0}^{q} \sum_{j=0}^{q} k_{i,j}^{m,n} \varphi_i^m(s) \varphi_j^n(t), \qquad \text{for } s \in I^m,\ t \in I^n. \tag{3.12}$$

Let us note, that we allow the approximations for $y_h$, $f_h$, and $k_h$ to be of different degrees, i.e., $q \neq p$, and therefore, we use two different sets of basis functions.

Since we only consider uniform meshes, it is natural to assume that this basis is invariant under the transformation, i.e., it holds $\varphi_i^n(t - t^n) = \varphi_i^m(t - t^m)$ for all $0 \leq i \leq q$ and all

$1 \leq m, n, \leq N$. The Lagrange basis for $f_h$ and $y_h$ is invariant by construction. These properties are important and will be utilized in the construction of the algorithms below.

For the evaluation of (3.7) at the time $t = t_j^m \in I^m$, we now split the interval $[0, t_j^m]$ into sub-intervals of the mesh and separate the integral into two contributions based on nearfield and farfield cells, as illustrated in Figure 3.1. Using the relation (3.10) we obtain

$$y_h(t_j^m) = \sum_{n=1}^{m-2} \int_{I^n} k_h(t_j^m, s) f_h(s)\, ds + \int_{t_j^{m-2}}^{t_j^m} k_h(t_j^m, s) f_h(s)\, ds, \tag{3.13}$$

where the two summands correspond to the farfield and nearfield contributions respectively. Using the basis representations (3.11) and (3.12) for $y_h$, $f_h$, and $k_h$, the integrals of the farfield contribution can be written as

$$\int_{I^n} k_h(t_j^m, s) f_h(s)\, ds = \sum_{i=0}^{q} \varphi_i^m(t_j^m) \sum_{k=0}^{q} k_{i,k}^{m,n} \sum_{r=0}^{p} \left( \int_{I^n} \varphi_k^n(s) \psi_r^n(s)\, ds \right) f_r^n.$$

For simplicity in notation, we now introduce the matrices $P$ and $Q$ defined as follows

$$P_{j,i} = \varphi_i^m(t_j^m), \qquad Q_{k,r} = \int_{I^n} \varphi_k^n(s) \psi_r^n(s)\, ds, \tag{3.14}$$

Due to the invariance of the bases, we can conclude that these values are independent of the time interval, i.e., the matrices are independent of $m$ and $n$. By $y^m$ we denote the vector with the solutions at the collocation points $y_j^m = y_h(t_j^m)$, $j = 0, \ldots, p$ given by (3.13). By the separation into nearfield and farfield contributions, we now write

$$y^m = w^m + z^m.$$

The evaluation of the farfield contribution $w^m$ can then be compactly written as

$$w^m = P u^m, \qquad u^m = \sum_{n=1}^{m-2} k^{m,n} g^n, \qquad g^n = Q f^n, \tag{3.15}$$

where $k^{m,n}$ is the matrix containing the entries $k_{i,j}^{m,n}$. Alternatively, the expression for $w^m$ can be simplified to $w^m = \sum_{n=0}^{m-2} K^{m,n} f^n$, where $K^{m,n} = P k^{m,n} Q$. The notation (3.15) is used on purpose and will be helpful in the following section. In a similar manner, the nearfield contribution $z^m$ can be expressed by

$$z^m = K^{m,m-1} f^{m-1} + K^{m,m} f^m, \tag{3.16}$$

where the matrices $K^{m,m-1}$, $K^{m,m}$ are constructed by analogy.

### Implementation details

Let $\mathtt{y}$, $\mathtt{f} \in \mathbb{R}^{N(p+1)}$ denote the global vectors by stacking together the contributions $y^m$, $f^n$, respectively. And let $\mathtt{K} \in \mathbb{R}^{N(p+1) \times N(p+1)}$ denote the block lower triangular matrix consisting of blocks $K^{m,n}$, $n \leq m$. The discretization of (3.6) can then be formulated compactly as a matrix-vector product

$$\mathtt{y} = \mathtt{K}\mathtt{f}, \tag{3.17}$$

**Algorithm 1** Evaluation of Volterra integrals for uniform meshes; see Alg. 1 in [40].

**for** $m = 1, \dots, N$ **do**
  $u = 0$
  **for** $n = 1, \dots, m-2$ **do**
    $u = u + k^{m,n} g^n$
  **end for**
  $g^m = Q f^m$
  $w^m = P u$
  $z^m = K^{m,m-1} f^{m-1} + K^{m,m} f^m$
  $y^m = w^m + z^m$
**end for**

as outlined in the Introduction. With the introduced notation, the numerical realization of this vector-matrix product can be done, as summarized in Algorithm 1.

First, let us mention that this realization is evolutionary. The entries $y^m$ are computed successively and only require the knowledge of $f^n$, $n \leq m$. The complexity of the algorithm can be roughly estimated as $O((p+1)^2 N^2)$ resulting from the block-wise matrix multiplication. It is oblivious, but only in the sense that only $f^m$ and $f^{m-1}$ are required in the $m-$th time step. The storage of values $g^n$, $n = 1, \dots, m-1$ is still required for the evaluation of $y^m$. This is a significant issue, which will be resolved in the following section. A rough approximation of the memory consumption adds up to $O((p+1)^2 N^2)$ for the storage of the blocks in the matrix `K` and $O((p+1)N)$ for the values $f^m$ and $g^m$.

Let us emphasize that this algorithm serves a purely educational purpose. The algorithm shares structural similarities with the one developed in the following section, which will aid in highlighting the essential components through direct comparison.

## 3.2. A fast and oblivious algorithm

We now present an algorithm for evaluation of (3.17), which is based on $\mathcal{H}^2-$ compression technique; see e.g. [21, 64]. The presented technique drastically reduces memory consumption and improves the complexity of the matrix-vector product evaluation, which is beneficial for long-time simulations. We start with the introduction of hierarchical meshes, which are the key ingredients of the presented methods.

### 3.2.1. Multilevel partitioning

The basic idea is to use an adaptively coarsening grid for the integration of the farfield contributions. For simplicity, we assume that the number of time steps satisfies $N = 2^L$ for $L \in \mathbb{N}$. Then we denote by $I^{(n;1)} = I^n$ and define the hierarchy of partitions constructed by recursive coarsening into sub-intervals $I^{(n;\ell)}$ given by

$$I^{(n;\ell)} = I^{(2n-1;\ell-1)} \cup I^{(2n;\ell-1)} = \left[ t^{2^{\ell-1}(n-1)}, t^{2^{\ell-1} n} \right], \qquad \ell > 1.$$

The number $\ell$, $1 \leq \ell \leq L-1$ stays for the level of coarsening. By construction we can conclude that the length of $I^{(n;\ell)}$ is $2^{\ell-1} h$, where $h = T/N$. The level $\ell = 1$ is called the

finest level, which corresponds to initial partitioning $I^{(n;1)} = I^n$. The construction of such a hierarchical mesh is illustrated in Figure 3.2.

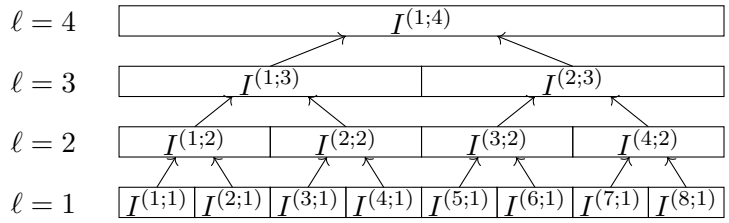



Figure 3.2.: Mesh hierarchy obtained by recursive coarsening of intervals $I^{(n;1)} = I^n$ with maximal coarsening level $L = 3$ and $N = 2^L = 8$ fine grid cells; see [40, Fig. 2].

We now introduce the hierarchical mesh used for the approximation of the kernel $k_h$ by

$$\mathcal{AT}_h = \{I^{(m;\ell)} \times I^{(n;\ell)} : \ell = 1 \text{ with } n \in \{m-1, m\} \text{ or } I^{(m;\ell)} \cap I^{(n;\ell)} = \emptyset \text{ with } I^{(\lceil m/2 \rceil;\ell+1)} \cap I^{(\lceil n/2 \rceil;\ell+1)} \neq \emptyset\},$$

where $\lceil r \rceil$ denotes the smallest integer larger or equal to $r$. Examples of such adaptive meshes $\mathcal{AT}_h$ are illustrated in Figure 3.3. With gray we again highlight the nearfield cells. Let us emphasize, that the coarsening acts only on the farfield cells, the nearfield cells stay untouched. Every element of the adaptive mesh is the square of size $2^{\ell-1}h$, where $\ell = 1, \ldots, L$. This partitioning is the standard choice for the related methods, in particular, $\mathcal{H}-$ and $\mathcal{H}^2-$ matrices; see e.g. [64].

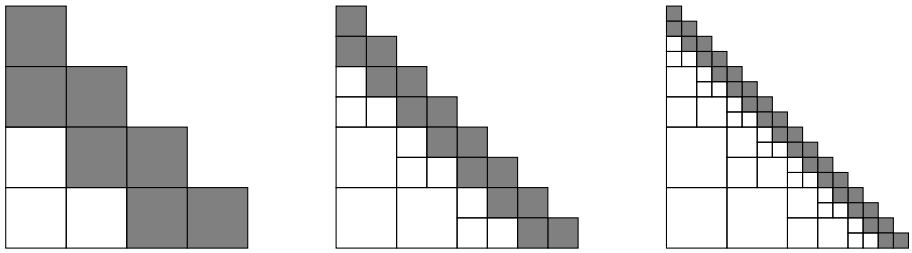


Figure 3.3.: Adaptive hierarchical meshes $\mathcal{AT}_h$ obtained by recursive coarsening of farfield cells in the uniformly refined meshes $\mathcal{T}_h \times \mathcal{T}_h$ illustrated in Figure 3.1; see [40, Fig. 3].

### 3.2.2. Adaptive data-sparse approximation

As the next step, we define by $\mathcal{P}_{q,q}(\mathcal{AT}_h)$ the space of piece-wise polynomials of degree $\leq q$ in each variable over the adaptive mesh $\mathcal{AT}_h$. Since the adaptive hierarchical mesh $\mathcal{AT}_h$ is constructed by coarsening of the uniform grid $\mathcal{T}_h \times \mathcal{T}_h$, we can conclude that

$$\mathcal{P}_{q,q}(\mathcal{AT}_h) \subset \mathcal{P}_{q,q}(\mathcal{T}_h \times \mathcal{T}_h).$$

The key ingredient is to use this subspace $\mathcal{P}_{q,q}(\mathcal{AT}_h)$ in the approximation of (3.13), instead of the space $\mathcal{P}_{q,q}(\mathcal{T}_h \times \mathcal{T}_h)$ on the uniform grid, as discussed in Section 3.1.1.

Considering a coarser grid for the approximation may reflect on the accuracy of the approach. However, the accuracy of the approximation can be preserved for the adaptive approximation under particular assumptions on the kernel. Now let us assume that the kernel $k$ is asymptotically smooth, i.e., there exist constants $c_1, c_2 > 0$, $r \in \mathbb{R}$ such that

$$\left|\partial_t^\alpha \partial_s^\beta k(t,s)\right| \leq c_1 \frac{(\alpha+\beta)!}{c_2^{\alpha+\beta}} (t-s)^{r-\alpha-\beta} \tag{3.18}$$

for all $\alpha, \beta \geq 0$ and all $t \neq s$. As shown in [21, 64], adaptive approximations $k_h \in \mathcal{P}_{q,q}(\mathcal{AT}_h)$ can be constructed for asymptotically smooth kernels, which converge exponentially in $q$ in the farfield. Therefore, the same accuracy can be achieved by adaptive and uniform approximations.

The adaptive approximation requires much fewer degrees of freedom than the uniform approximation. Indeed, it is easy to verify that $\dim(\mathcal{P}_{q,q}(\mathcal{T}_h \times \mathcal{T}_h)) = \mathcal{O}(N^2 q^2)$ while $\dim(\mathcal{P}_{q,q}(\mathcal{AT}_h)) = \mathcal{O}(Nq^2)$. The adaptive hierarchical approximation thus is *data-sparse* and leads to a significant reduction of memory consumption, required for storing the kernel approximation or its matrix representation (3.17).

In the following section, we discuss the appropriate evaluation of the matrix product (3.17) which leads to a significant reduction of complexity and compression of data kept in memory, as outlined in Section 3.1.2. To do so we use the hierarchical basis representation.

### 3.2.3. Multilevel hierarchical basis

In accordance with the multilevel partitioning of the domain, we define the multilevel basis

$$\varphi_i^{(n;\ell)}(t) = \begin{cases} \sum_{j=0}^q A_{i,j}^{(1)} \varphi_j^{(2n-1;\ell-1)}(t), & t \in I^{(2n-1;\ell-1)}, \\ \sum_{j=0}^q A_{i,j}^{(2)} \varphi_j^{(2n;\ell-1)}(t), & t \in I^{(2n;\ell-1)}, \end{cases} \tag{3.19}$$

of the spaces $\mathcal{P}_q(I^{(n;\ell)})$, $\ell > 1$, where $\varphi_i^{(n;1)} = \varphi_i^n$ is the basis on the finest level, as discussed in Section 3.1.2. The recursive construction plays an important role in the following. It is important to note that the coefficients $A_{i,j}^{(1)}$ and $A_{i,j}^{(2)}$ are independent of $n$ and $\ell$, due to the invariance of the basis. Now, in each cell of the domain $\mathcal{AT}_h$ we expand the kernel with respect to this basis as

$$k_h(s,t) = \sum_{i=0}^q \sum_{j=0}^q k_{i,j}^{(m,n;\ell)} \varphi_i^{(m;\ell)}(s) \varphi_j^{(n;\ell)}(t), \qquad (s,t) \in I^{(m;\ell)} \times I^{(n;\ell)}. \tag{3.20}$$

For the evaluation of (3.13), we now split the farfield integration domain into

$$[0, t^{m-2}] = \bigcup_{\ell=1}^{L(m)} \bigcup_{n=1}^{B(m;\ell)} I^{(P(m,n;\ell);\ell)}, \tag{3.21}$$

where $L(m) = \lceil \log_2(m) \rceil - 1$ is the number of levels involved, $B(m;\ell) = \mathrm{bin}(m)_\ell + 1 \in \{1,2\}$ is the number of intervals on each level, and $P(m,n;\ell) = C(m;\ell) - n - 1$ with $C(n;\ell) = \lceil n/2^{\ell-1} \rceil$ corresponds to indices of the intervals on the level $\ell$. With $\mathrm{bin}(m)_\ell$ we denote the $\ell-$th digit from behind of the binary representation of $m$ obtained by Matlab's

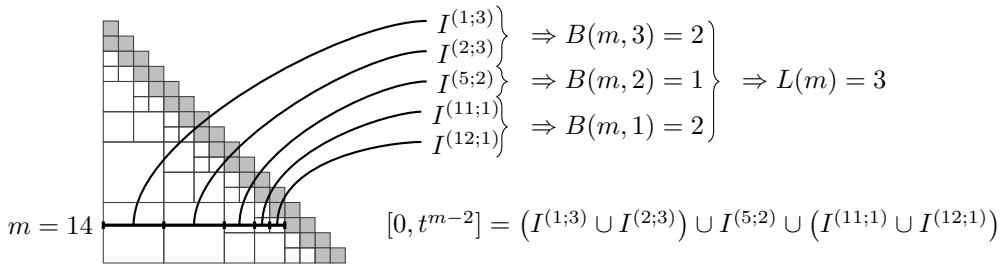



Figure 3.4.: Illustration of partitioning with the description of the underlying quantities $L(m)$, $B(m,\ell)$, and $P(m,n;\ell)$ in (3.21) for $m = 14$; see [40, Fig. 4]

`dec2bin` function. The partitioning (3.21) of the integration domain corresponds to the cells of the mesh $\mathcal{AT}_h$. An example of the splitting is illustrated in Figure 3.4.

By decomposing the farfield integral term in (3.13) with respect to the splitting (3.21), we obtain the expression

$$y_h(t_j^m) = \sum_{\ell=1}^{L(m)} \sum_{n=1}^{B(m;\ell)} \int_{I^{(P(m,n;\ell);\ell)}} k_h(t_j^m, s) f_h(s)\, ds + \int_{t^{m-2}}^{t_j^m} k_h(t_j^m, s) f_h(s)\, ds.$$

Using the expansion (3.20) of the kernel on the interval $I^{(P(m,n;\ell);\ell)}$, the farfield integrals can be expressed as

$$\int_{I^{(P(m,n;\ell);\ell)}} k_h(t_j^m, s) f_h(s)\, ds = \sum_{i=0}^{q} \varphi_i^{(C(n;\ell);\ell)}(t_j^m) \sum_{k=0}^{q} k_{i,k}^{(C(n;\ell),P(m,n;\ell);\ell)} g_k^{(P(m,n;\ell);\ell)},$$

where

$$g_k^{(P(m,n;\ell);\ell)} = \int_{I^{(P(m,n;\ell);\ell)}} \varphi_k^{(P(m,n;\ell);\ell)}(s) f_h(s)\, ds.$$

By the definition of $\varphi^{(n;\ell)}$ as in (3.19), the latter expression satisfies the recursive relation

$$g_k^{(i;\ell)} = \int_{I^{(i;\ell)}} \varphi_k^{(i;\ell)}(s) f_h(s)\, ds = \sum_{j=0}^{q} \big(A_{i,j}^{(1)} g_k^{(2i-1;\ell-1)} + A_{i,j}^{(2)} g_k^{(2i;\ell-1)}\big), \quad \text{for } \ell > 1,$$

and $g_k^{(i;1)} = g_k^i = \sum_{r=0}^{p} Q_{k,r} f_r^i$ for the finest level, as discussed in Section 3.1.1. The evaluation of the relation (3.19) at the time step $t_j^m$ yields

$$\varphi_i^{(C(n;\ell);\ell)}(t_j^m) = \sum_{k=0}^{q} A_{i,k}^{(B(m,\ell-1))} \varphi_k^{(C(n,\ell-1);\ell-1)}(t_j^m),$$

such that we may define intermediate values

$$u_j^{(C(n;\ell);\ell)} = \sum_{i=0}^{q} A_{i,j}^{(B(m;\ell))} u_i^{(C(m,\ell+1);\ell+1)} + \sum_{n=1}^{B(m;\ell)} \sum_{k=0}^{q} k_{i,k}^{(C(n;\ell),P(m,n;\ell);\ell)} g_k^{(P(m,n;\ell);\ell)}.$$

Similarly to (3.15), the integral (3.13) can then be obtained

$$y_j^m = y_h(t_j^m) = \sum_{k=0}^{q} P_{j,k} u_k^{(m;1)} + z^m,$$

where nearfield contributions $z^m$ are as in (3.16) and projections $P_{j,k}$ are given in (3.14).

### Implementation details

The evaluation of the matrix product based on the introduced ideas is summarized in Algorithm 2.

**Algorithm 2** A fast and oblivious evolutionary algorithm; see Alg. 2 in [40].

**for** $m = 1, \dots, N$ **do**
  $L_{\text{coarse}} = 1 + \lfloor \log_2(\texttt{bitxor}(m, m-1)) \rfloor$
  **for** $\ell = L_{\text{coarse}}, \dots, 1$ **do**
    **if** $B(m;\ell) \neq B(m-1;\ell)$ **then**
      $\mathtt{g}^{(2;\ell)} = \mathtt{g}^{(1;\ell)}$
      **if** $\ell > 1$ **then**
        $\mathtt{g}^{(1;\ell)} = A^{(1)}\mathtt{g}^{(1;\ell-1)} + A^{(2)}\mathtt{g}^{(2;\ell-1)}$
      **else**
        $\mathtt{g}^{(1;\ell)} = Q\mathtt{f}^{(2)}$
      **end if**
      Set $(\mathtt{K}_n)_{i,j} = k_{i,j}^{(C(n;\ell),P(m,n;\ell);\ell)}$ for $n \in \{1, B(m;\ell)\}$
      $\mathtt{u}^\ell = \mathtt{K}_1\mathtt{g}^{(1;\ell)}$
      **if** $B(m;\ell) = 2$ **then**
        $\mathtt{u}^\ell = \mathtt{u}^\ell + \mathtt{K}_2\mathtt{g}^{(2;\ell)}$
      **end if**
      $\mathtt{u}^\ell = \mathtt{u}^\ell + \left(A^{(B(m;\ell))}\right)^\top \mathtt{u}^{(\ell+1)}$
    **end if**
  **end for**
  $\mathtt{f}^{(2)} = \mathtt{f}^{(1)}$
  $\mathtt{f}_j^{(1)} = f(t_j^m)$, $j = 0, \dots, p$
  $z^m = K^{m,m-1}\mathtt{f}^{(2)} + K^{m,m}\mathtt{f}^{(1)}$
  $y^m = P\mathtt{u}^{(1)} + z^m$
**end for**

The implementation is done in Matlab. We use the build function $\texttt{bitxor}(a, b)$, which returns the integer generated by a bit-wise xor comparison of the binary representation of $a$ and $b$, to compute the value $L_{\text{coarse}}$ in $O(1)$ complexity. One may set $L_{\text{coarse}} = L(m)$ without any notable difference in computation times. Let us note that at each level $\ell$ only one value $u^{(n;\ell)}$ and two values of $g^{(n;\ell)}$ are required. Furthermore, at most two values of $f^n$ are required at any time step. The required buffers are denoted by $\mathtt{u}^{(\ell)}$, $\mathtt{f}^{(i)}$, and $\mathtt{g}^{(i;\ell)}$, $i = 1, 2$, $\ell \in \mathbb{N}$. For illustration purposes, a time-stepping is summarized in the following example.

**Example 3.2.1.** From the illustration in Figure 3.4 we see that the partitioning of $[0, t^{m-2}]$ for $m = 14$ and $m = 15$, corresponding to the farfield contributions, are given by

$$\begin{aligned}\big[0, t^{12}\big] &= \big(I^{(1;3)} \cup I^{(2;3)}\big) \cup I^{(5;2)} \cup \big(I^{(11;1)} \cup I^{(12;1)}\big),\\ \big[0, t^{13}\big] &= \big(I^{(1;3)} \cup I^{(2;3)}\big) \cup \big(I^{(5;2)} \cup I^{(6;2)}\big) \cup I^{(13;1)},\end{aligned}$$

and that a change in the partitioning structure is only given in the smaller time intervals with coarsening level $\ell \leq 2$. As a consequence, the algorithm only changes variables on level $\ell \leq 2$ in the for loop for $m = 15$, which we would like to elaborate on in the following. At the beginning of the loop, the values of the variables storing the history of the data are

$$\begin{aligned}\mathtt{f}^{(1)} &= f^{14}, & \mathtt{g}^{(1;2)} &= g^{(5;2)},\\ \mathtt{f}^{(2)} &= f^{13}, & \mathtt{g}^{(2;2)} &= g^{(4;2)} \quad \text{(unused)},\\ \mathtt{g}^{(1;1)} &= g^{(12;1)}, & \mathtt{g}^{(1;3)} &= g^{(2;3)},\\ \mathtt{g}^{(2;1)} &= g^{(11;1)}, & \mathtt{g}^{(2;3)} &= g^{(1;3)},\end{aligned}$$

in correspondence to the partitioning of the farfield. Using the recursive relations of the multilevel hierarchy, they are changed efficiently to

$$\begin{aligned}\mathtt{f}^{(1)} &= f^{15}, & \mathtt{g}^{(1;2)} &= g^{(6;2)} = A^{(1)} g^{(11;1)} + A^{(2)} g^{(12;1)},\\ \mathtt{f}^{(2)} &= f^{14} \quad \text{(by copying)}, & \mathtt{g}^{(2;2)} &= g^{(5;2)} \quad \text{(by copying)},\\ \mathtt{g}^{(1;1)} &= g^{(13;1)} = Q f^{13}, & \mathtt{g}^{(1;3)} &= g^{(2;3)} \quad \text{(unchanged)},\\ \mathtt{g}^{(2;1)} &= g^{(12;1)} \quad \text{(by copying)}, & \mathtt{g}^{(2;3)} &= g^{(1;3)} \quad \text{(unchanged)},\end{aligned}$$

during the loop, which corresponds to the decomposition of the farfield at $m = 15$. The intermediate values for computing the farfield contributions at the beginning of the loop are given by

$$\begin{aligned}\mathtt{u}^{(3)} &= K^{(4,1;3)} \mathtt{g}^{(2;3)} + K^{(4,2;3)} \mathtt{g}^{(1;3)},\\ \mathtt{u}^{(2)} &= \big(A^{(1)}\big)^{\top} \mathtt{u}^{(3)} + K^{(7,5;2)} \mathtt{g}^{(1;2)},\\ \mathtt{u}^{(1)} &= \big(A^{(2)}\big)^{\top} \mathtt{u}^{(2)} + K^{(14,11;3)} \mathtt{g}^{(2;3)} + K^{(14,12;3)} \mathtt{g}^{(1;3)},\end{aligned}$$

and, using the multilevel hierarchy again, are changed efficiently to

$$\begin{aligned}\mathtt{u}^{(3)} &= K^{(4,1;3)} \mathtt{g}^{(2;3)} + K^{(4,2;3)} \mathtt{g}^{(1;3)} \quad \text{(unchanged)},\\ \mathtt{u}^{(2)} &= \big(A^{(2)}\big)^{\top} \mathtt{u}^{(3)} + K^{(7,5;2)} \mathtt{g}^{(1;2)},\\ \mathtt{u}^{(1)} &= \big(A^{(1)}\big)^{\top} \mathtt{u}^{(2)} + K^{(14,11;3)} \mathtt{g}^{(2;3)} + K^{(14,12;3)} \mathtt{g}^{(1;3)}.\end{aligned}$$

#### Complexity and memory consumption

We consider the Algorithm 2 for the evaluation of (3.13) with kernel $k_h \in \mathcal{P}_{q,q}(\mathcal{AT}_h)$ and data $f_h \in \mathcal{P}_p(\mathcal{T}_h)$, and with $N = 2^L$ denoting the number of time intervals in $\mathcal{T}_h$. The algorithm (2) represents ah $\mathcal{H}^2-$ vector-matrix product with the rearrangement of the operations. Therefore, complexity and memory consumption can then be directly estimated; see e.g. [21]. For convenience, we summarize the results.

**Lemma 3.2.2** (Lemma 2 in [40])**.** Algorithm 2 can be executed in $\mathcal{O}\big(N(p^2+q^2)\big)$ operations.

*Proof.* The algorithm rearranges the operations of a standard $\mathcal{H}^2$-matrix-vector multiplication without adding any significant operations. We therefore simply estimate the complexity of the corresponding $\mathcal{H}^2$-matrix-vector multiplication. Let us first remark that the computation of $z^m$ in line 21 requires $\mathcal{O}(p^2)$ operations in each time step. Second, on a given level $\ell$, we have to perform $\mathcal{O}(2^\ell)$ applications of $A^{(1)}$ and $A^{(2)}$ in total for obtaining the $g^{(n;\ell)}$ from the ones on level $\ell-1$, see line 7. Similarly, $\mathcal{O}(2^\ell)$ applications of $A^{(B(m;\ell))}$ in line 16 are in total required on level $\ell$ for the computation of the $u^{(n;\ell)}$ and $\mathcal{O}(2^\ell)$ multiplications by $k^{(k,n;\ell)}$ need to be performed in lines 12 and 14. Finally, $\mathcal{O}(N)$ values of $g^n = Qf^n$ and $P\mathtt{u}^{(1)}$ need to be computed in line 9 and line 22. Summing up yields

$$\mathcal{O}(Np^2) + 3\mathcal{O}(q^2)\sum_{\ell=1}^{L}\mathcal{O}(2^{L-\ell}) + 2\mathcal{O}(Npq) = \mathcal{O}(Np^2) + \mathcal{O}(2^L q^2) + \mathcal{O}(Npq),$$

and since $N = 2^L$ Young's inequality yields the assertion. □

**Lemma 3.2.3** (Lemma 3 in [40])**.** The $\mathcal{H}^2$-matrix representation $\mathtt{K}$ of the adaptive hierarchic approximation $k_h \in \mathcal{P}_{q,q}(\mathcal{AT}_h)$ can be stored in $\mathcal{O}\big(N(p^2+q^2)\big)$ memory. If the kernel is of convolution type, i.e., $k(t,s) = k(t-s)$, then the memory cost reduces to $\mathcal{O}\big(p^2 + \log_2(N)q^2\big)\big)$.

*Proof.* The proof for the adaptive approximation is similar to the previous lemma, with the $p^2$-related term arising from the nearfield and the $q^2$-related term from the farfield. For a kernel of convolution type, the hierarchical approximation provides a block Toeplitz structure, such that we only have to store $\mathcal{O}(1)$ coefficient matrices per level for the farfield and $\mathcal{O}(1)$ coefficient matrices for the nearfield. □

Let us finally also remark on the additional memory required during execution.

**Lemma 3.2.4** (Lemma 4 in [40])**.** The active memory required for storing the data history required for 2 is bounded by $\mathcal{O}(q\log_2 N + p)$.

*Proof.* We require $O(1)$ vectors of length $p$ for the nearfield and at most two vectors $g^{(n;\ell)}$ of length $q$ on $L = \log_2(N)$ levels for the farfield contributions. □

The algorithm is executed in an oblivious and evolutionary manner and can therefore be generalized immediately to integro-differential equations of the form (3.1). Furthermore, knowledge of the number of time steps $N$ is not required prior to execution.

The discussed computational approach relies on the explicit knowledge of the kernel. As mentioned in the introduction, a particular interest represents the convolution integrals, where the kernel is given in the frequency domain via its Laplace transform. In the next section, we discuss the extension of the algorithm to the problems of this setting. Our approach is based on convolution quadrature methods [88, 91] and closely related to the fast and oblivious convolution quadrature approach of [115].

## 3.3. Approximation of convolution operators

We now discuss the numerical evaluation of the Volterra integral operators with a kernel of the convolution type

$$k(t,s) = k(t-s),$$

which is implicitly given via its Laplace transform

$$\hat{k}(s) := (\mathcal{L}k)(s) := \int_0^\infty e^{-st} k(t)\, dt, \quad s \in \mathbb{C}.$$

In the context of dynamical systems, the quantity $\hat{k}(s)$ is called the transfer function, $f$ is the input, and $y$ is the output. The access to the kernel can be provided by the inverse Laplace transform

$$k(t) = (\mathcal{L}^{-1}\hat{k})(t) = \frac{1}{2\pi i}\int_\Gamma e^{t\lambda}\hat{k}(\lambda)\, d\lambda, \quad t > 0, \tag{3.22}$$

where $\Gamma$ is a contour in the complex plain connecting $-i\infty$ and $i\infty$; see e.g. [9]. The well-posedness of this relation can be guaranteed under further assumptions on the kernel. We follow [88, 91] and assume that

$$\hat{k}(\lambda) \text{ is analytic in a sector } |\arg(\lambda - c)| < \varphi,\ \frac{\pi}{2} < \varphi < \pi, \tag{3.23}$$

$$\text{and } |\hat{k}(\lambda)| \le M|\lambda|^{-\mu} \text{ for some fixed } M, \mu > 0, \tag{3.24}$$

and the contour $\Gamma$ lies within the analyticity domain of the function $\hat{k}$.

The necessary modification to Algorithm 2 is the use of the inverse Laplace transform (3.22) for the evaluation of the kernel function. In the following lemma we show that a kernel defined via (3.22) with imposed assumptions (3.23) and (3.24) is asymptotically smooth, and, therefore, the adaptive approximation as discussed in Section 3.2 is justified.

**Lemma 3.3.1** (Lemma 5 in [40])**.** Assume that $\hat{k}$ satisfies (3.23) and (3.24). Then $k$ as defined in (3.22) is asymptotically smooth, i.e., it satisfies (3.18) with $c_2 = \sin(\varphi - \pi/2)$.

*Proof.* It is sufficient to consider the case $c = 0$ in (3.23) and $\mu = 1$ in (3.24). Otherwise, we simply transform $\hat{k}(\lambda + c) = \mathcal{L}(e^{-ct}k(t))(\lambda)$ and $k(t) = k_*^{(\mu-1)}(t)$ with $\hat{k}_*(\lambda) := \mathcal{L}(k_*)(\lambda) = |\lambda|^{\mu-1}\hat{k}(\lambda)$ for $\mu \neq 1$. From [9, Theorem 2.6.1], also see [127], we deduce that $k$ has a holomorphic extension into the sector $|\arg(\lambda)| < \varphi - \pi/2$ with $\varphi$ as in (3.23). Thus, the radius of convergence of the Taylor series of $k$ around $t_0 \in (0,\infty)$ is given by $c_2 t_0$, with $c_2 = \sin(\varphi - \pi/2)$ independent of $t_0$. This implies

$$\left|\partial_t^\alpha k(t)\right| \le c_1 \frac{\alpha!}{c_2^\alpha t^\alpha}$$

for some constant $c_1 > 0$. Condition (3.18) then follows by the chain rule. □

For the evaluation of the nearfield contribution, we use the convolution quadrature methods [88, 89, 91]. Let us briefly summarize the basic ideas.

### 3.3.1. Convolution quadrature methods

By substituting the expression for the kernel (3.22) into Volterra integral equation (3.6) with $k(t,s) = k(t-s)$ and changing the order of integration leads to the expression

$$y(t) = \frac{1}{2\pi i}\int_\Gamma \hat{k}(\lambda) \underbrace{\int_0^t e^{(t-s)\lambda} f(s)\, ds}_{=:z(t;0,\lambda)}\, d\lambda, \qquad t \in [0,T]. \tag{3.25}$$

The function $z(t;0,\lambda)$ is the solution of the initial value problem

$$\partial_t z(t;0,\lambda) = \lambda z(t;0,\lambda) + f(t), \qquad z(0;0,\lambda) = 0. \tag{3.26}$$

The key idea of the approach is to discretize this equation by an appropriate method. Multistep methods and Runge-Kutta schemes have been applied in [88, 89, 90].

For illustration, let us consider a simple implicit Euler scheme as presented in [114]. A discretization of problem (3.34) by the implicit Euler time-stepping scheme with uniform step-size $h$, time steps $t^n = nh$, $n \geq 0$, and initial value $z^{-1}(\lambda) = 0$ leads to approximations

$$z^n = h\sum_{\ell=0}^{n} \frac{1}{(1-h\lambda)^{\ell+1}} f(t^{n-\ell}) \approx \int_0^{t^n} e^{\lambda s} f(t^n - s)\, ds. \tag{3.27}$$

Inserting the approximation (3.27) into (3.25) yields

$$y(t^n) \approx y_h(t^n) := \sum_{\ell=0}^{n} \omega_\ell f(t^{n-\ell}) \quad \text{with} \quad \omega_\ell = \left(\frac{h}{2\pi i}\int_\Gamma \frac{\hat{k}(\lambda)}{(1-h\lambda)^{\ell+1}}\, d\lambda\right).$$

Then the discretization of the convolution integral (3.6) can yields to the algebraic problem (3.17) with $\mathtt{y}_m = y_h(t^m)$, $\mathtt{f}_n = f(t^n)$, and $\mathtt{K}_{m,n} = \omega_{m-n}$ for $n \leq m$.

The approximation by higher-order Runge-Kutta collocation methods can be considered in a similar manner. Following [91], application of an $(p+1)-$stages scheme for solution of (3.34) results into

$$y_h(t_j^n) = \sum_{\ell=0}^{n}\sum_{i=0}^{p} \omega_{\ell,ji} f(t_i^{n-\ell})$$

where the quadrature weights $\omega_{\ell,ji}$ are the entries of the matrix $W_\ell$ defined by

$$\sum_{n=0}^{\infty} W_\ell \zeta^n = \hat{k}\Big(\frac{\Delta(\zeta)}{h}\Big) \quad \text{with} \quad \Delta(\zeta) = (A + \frac{\zeta}{1-\zeta} 1 b^\top), \tag{3.28}$$

where $A$ and $b$ are the Runge-Kutta coefficients. Note that with this notation the function $\hat{k}$ takes matrix-valued arguments. This is to be interpreted in terms of the power series and therefore $\hat{k}$ acts on the eigenvalues of the matrix-valued arguments. By choosing $\zeta = \rho e^{i\phi}$ sum in (3.28) becomes the Fourier series, which leads to the explicit formula for the weights

$$W_\ell = \frac{1}{2\pi\rho^n}\int_0^{2\pi} \hat{k}(\Delta(\rho e^{i\phi})/h) e^{-in\phi}\, d\phi.$$

The efficient approximation can then be done using a fast Fourier transform with complexity $O(N \log N)$. As discussed in [88, 91], the weights can be computed with accuracy $O(\epsilon)$ with $L = O(\log \epsilon)N$ points for contour integration and $\log \epsilon = O(h)$. For accuracy $O(\sqrt{\epsilon})$ it is sufficient to use $L = O(N)$ points and $\rho = \sqrt[2N]{\epsilon}$.

The evaluation of the matrix-vector product (3.17) can be realized in $O(N \log N)$ operations and requires $O(N)$ memory to compute and store the weights $W_\ell$ and the intermediate results. As outlined in [67], an evolutionary version of the convolution sum can be realized in $O(N \log^2 N)$ complexity.

Let us note that the collocation time stepping scheme applied (3.34) implicitly utlizes polynomial approximations $f_h$, $y_h$ on each of the intervals $I^n$. Therefore, convolution quadrature methods formally fit into the abstract approximation setting considered in Section 3.1. However, the polynomial basis is not directly visible from the collocation schemes. And thus, it is not directly clear how the matrix $Q$ in 3.14 for evaluation $g^n = Qf^n$ is necessary for the adaptive approximation in the following section. The computation of

$$g_k^n = \int_{I^n} \varphi_k^n(s - t^{n-1}) f_h(s)\, ds$$

is equivalent to solving the differential equation

$$\partial_t \tilde{z}_k(t; t^{n-1}, f_h) = \varphi_k^n(t - t^{n-1}) f_h(t), \qquad \tilde{z}_k(t^{n-1}; t^{n-1}, f_h) = 0, \tag{3.29}$$

up to time $t^n$. Thus, $g^n = \tilde{z}_k(t^n; t^{n-1}, f)$ can be computed from $f^n$ by using the same time-stepping scheme as used for the convolution quadrature applied to (3.29) and does not require any additional evaluations of $f_h$. In fact, the time stepping procedure can be rewritten as $g^n = Qf^n$ where the coefficients of $Q$ can be copied from the coefficients of the underlying time stepping scheme.

### 3.3.2. Adaptive approximation

We proceed by analogy and split the convolution integral into farfield and nearfield contributions as

$$y_h(t^n) = \int_0^{t^{n-2}} k_h(t^n - s) f_h(s)\, ds + \int_{t^{n-2}}^{t^n} k_h(t^n - s) f_h(s)\, ds \tag{3.30}$$

The treatment of the farfield contributions is analog to Section 3.2, where the access to the kernel function $k_h(s)$ is provided by the numerical approximation of (3.22), as discussed below.

#### Numerical inversion of Laplace transform

We follow the approach of [85, 115] and consider a hyperbolic contour of the form

$$\gamma(\theta) = \mu(1 - \sin(\alpha + i\theta)) + \sigma, \qquad \theta \in \mathbb{R}, \tag{3.31}$$

with $0 < \mu$, $0 < \alpha < \pi/2 - \varphi$, and $\sigma \in \mathbb{R}$, such that the contour remains in the sector of analyticity (3.23) of $\hat{k}$. The discretization of the integral (3.22) by the trapezoidal rule with with step $\tau$ yields

$$k(t) \approx \sum_{r=-R}^{R} \frac{i\tau}{2\pi} e^{\gamma(\theta_r)t} \gamma'(\theta_r) \hat{k}(\gamma(\theta_r)), \tag{3.32}$$

with $\theta_r = \tau r$. For the evaluation of $k(t)$ in the interval $t \in [t_{\min}, t_{\max}]$ with chosen $\alpha$ and $\sigma$, the parameters $\tau$ and $\mu$ are chosen as

$$\tau = a_\rho(\rho_{\text{opt}}), \quad \mu = \frac{2\pi\alpha R(1-\rho_{\text{opt}})}{t_{\max} a_\rho(\rho_{\text{opt}})}, \quad \rho_{\text{opt}} = \operatorname*{arg\,min}_{\rho\in(0,1)} \left(\epsilon\epsilon_R(\rho)^{\rho-1} + \epsilon_R(\rho)^\rho\right),$$

where $\epsilon$ is the machine precision and

$$a_\rho(\rho) = \operatorname{acosh}\left(\frac{t_{\max}/t_{\min}}{(1-\rho)\sin(\alpha)}\right), \quad \epsilon_R(\rho) = \exp\left(-\frac{2\pi\alpha R}{a_\rho(\rho)}\right).$$

For more details on the parameter choice and precise error bounds, we refer to [85, 115]. In the numerical tests we chose $\alpha = 3\pi/16$ and $\sigma = 0$.

### 3.3.3. Relation to fast and oblivious convolution quadrature methods

The presented hierarchical approximation is closely related to fast and oblivious convolution quadrature methods introduced in [92, 115]. There, the authors propose a hierarchical partitioning into L-shaped cells as illustrated in Figure 3.5. The convolution integral (3.30) is again decomposed into the nearfield and farfield contributions. The nearfield contributions are computed using the convolution quadrature methods.

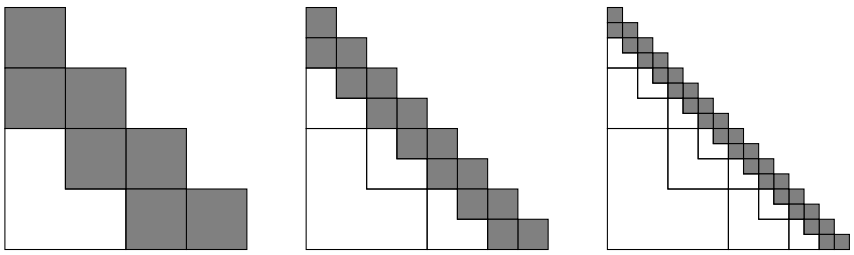


Figure 3.5.: Hierarchical partitions of fast and oblivious convolution quadrature [115]; see [40, Fig. 5].

The farfield part for the entry $y_h(t^m)$ is then based on the partitioning

$$[0, t^{m-2}] = \bigcup_{\ell=1}^{L(m)} \bigcup_{n=1}^{B(m;\ell)} I^{(P(m,n;\ell);\ell)} = \bigcup_{\ell=1}^{L(m)} I^\ell_{\text{FOCQ},m}.$$

Choosing an appropriate contour $\Gamma_\ell$ as in (3.31) and corresponding quadrature points $\theta_r^{(\ell)}$

for each farfield cell and using (3.32) yields an approximation

$$\begin{aligned}
&\int_{I^\ell_{\mathrm{FOCQ},m}} k\big(t^m - s\big) f(s)\, ds \\
&\quad\approx \int_{I^\ell_{\mathrm{FOCQ},m}} \frac{i\tau}{2\pi} \sum_{r=-R}^{R} \hat{k}\big(\gamma(\theta_r^{(\ell)})\big)\gamma'\big(\theta_r^{(\ell)}\big) e^{\gamma(\theta_r^{(\ell)})(t^m - s)} f(s)\, ds \\
&\quad= \frac{i\tau}{2\pi} \sum_{r=-R}^{R} \hat{k}\big(\gamma(\theta_r^{(\ell)})\big)\gamma'\big(\theta_r^{(\ell)}\big) e^{\gamma(\theta_r^{(\ell)})(t^m - b^{(\ell)})} \underbrace{\int_{I^\ell_{\mathrm{FOCQ},n}} e^{\gamma(\theta_r^{(\ell)})(b^{(\ell)} - s)} f(s)\, ds}_{= z(c^{(\ell)}; b^{(\ell)}, \gamma(\theta_r^{(\ell)}))}, \qquad (3.33)
\end{aligned}$$

with $b^{(\ell)} = \min I^\ell_{\mathrm{FOCQ},m}$ and $c^{(\ell)} = \max I^\ell_{\mathrm{FOCQ},m}$. The values $z(c^{(\ell)}; b^{(\ell)}, \gamma(\theta_r^{(\ell)}))$ can be computed by numerically solving the ordinary differential equation

$$\partial_t z(t; b^{(\ell)}, \gamma(\theta_r^{(\ell)})) = \gamma(\theta_r^{(\ell)}) z(t; b^{(\ell)}, \gamma(\theta_r^{(\ell)})) + f(t), \quad z(b^{(\ell)}; b^{(\ell)}, \gamma(\theta_r^{(\ell)})) = 0. \qquad (3.34)$$

Thus, the fast and oblivious convolution quadrature provides an approximation of the convolution matrix by solving an auxiliary set of $(2R+1)L$ differential equations. In order to obtain an oblivious algorithm it is crucial that the solution of each differential equation is updated in each time step, i.e., the compressed convolution matrix must be evaluated from *left to right*; see [92, 115] for details.

The compression approach of the fast and oblivious convolution quadrature can be understood as a low-rank approximation in each of the farfields L-shaped blocks

$$\begin{aligned}
k(t,s) &\approx \sum_{r=-R}^{R} \left( \frac{i\tau}{2\pi} e^{\gamma(\theta_r^{(\ell)})(t - b^{(\ell)})} \hat{k}\big(\gamma\big(\theta_r^{(\ell)}\big)\big)\gamma'\big(\theta_r^{(\ell)}\big) \right) e^{\gamma(\theta_r^{(\ell)})(b^{(\ell)} - s)} \\
&= \sum_{r=-R}^{R} U(t, \theta_r^{(\ell)}) V(s, \theta_r^{(\ell)}).
\end{aligned}$$

Then, the farfield approximation (3.33) can then be written as

$$\begin{aligned}
\int_{I^\ell_{\mathrm{FOCQ},m}} k\big(t^n, s\big) f(s)\, ds &\approx \sum_{r=-R}^{R} U\big(t, \theta_r^{(\ell)}\big) \int_{I^\ell_{\mathrm{FOCQ},m}} V\big(s, \theta_r^{(\ell)}\big) f(s) \mathbf{d}s \\
&= \sum_{r=-R}^{R} U\big(t, \theta_r^{(\ell)}\big) z\big(c^{(\ell)}, b^{(\ell)}, \theta_r^{(\ell)}\big).
\end{aligned}$$

Thus, it can be understood as a realization of a low-rank matrix-vector product by the numerical solution of a differential equation. Furthermore, since the partitioning depicted in Figure 3.5 can easily be refined to an adaptive partitioning as in Figure 3.3, the fast and oblivious convolution quadrature can be interpreted as a particular case of an $\mathcal{H}$-matrix approximation with a specific realization of the $\mathcal{H}$-matrix-vector product; see e.g. [21, 64]

## 3.4. Numerical examples

To illustrate the results we consider two numerical examples. In the first example, we consider the numerical solution of Volterra integral equations stemming from the reformation of an ordinary differential equation by variation of constants formula. In the second example, we consider the discretization of the Volterra-integro-differential equation stemming from the reformulation of the field-circuit problem of Chapter 2.

### Variation of constants formula; see [40, Example 5.1]

In the first example, we consider the initial value problem

$$y'(t) = -2ty(t) + 5\cos(5t), \qquad y(0) = 2. \tag{3.35}$$

By the variation of the constants formula, the solution can be expressed as

$$y(t) = 2e^{-t^2} + 5\int_0^t e^{s^2-t^2}\cos(5s)\,ds. \tag{3.36}$$

Let us note that the integral kernel $k(t,s) = e^{s^2-t^2}$ satisfies the asymptotic smoothness assumption (3.18), which justifies the use Algorithm 2.

We chose the polynomial degree $q = 16$ for the approximation of kernel function and $p = 1, 2, 3$ for data $f(t) = \cos(5t)$ and solution $y$. We chose the right Radau collocation points for the discretization of (3.36) which relates to the solution of (3.35) using RadauIIA schemes. The Radau nodes are the classical choice in the convolution quadrature literature [91, 115]. Furthermore, for the farfield approximation, we construct cells by combining $n_{\min} \times n_{\min} = 16 \times 16$ cells on the finer level, which is the common choice in the related literature.

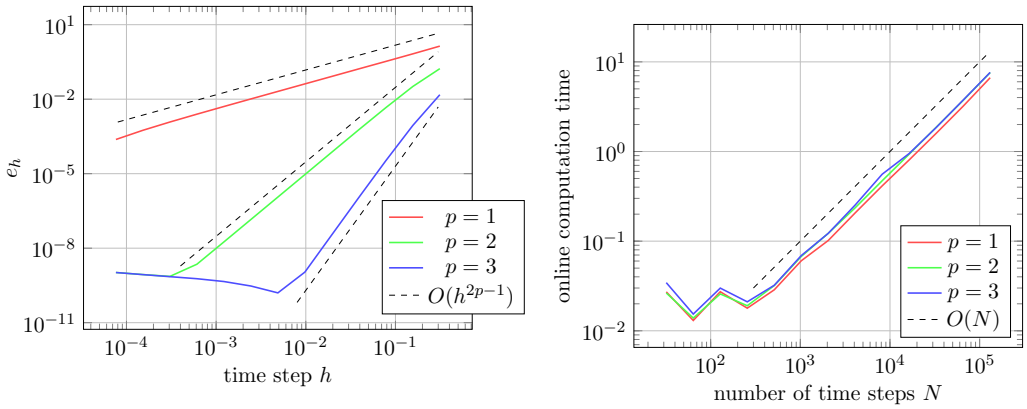



Figure 3.6.: Approximation errors (left) and computation times (right) for the variation of constants formula example of Section 3.4.

The left plot of 3.6 illustrates the convergence rates. We observe the error decay

$$e_h =: \max_{t_i \in \mathcal{T}_h} |y(t_i) - y_h(t_i)| \leq Ch^{2p-1},$$

as expected; see e.g. [27, Chapter 2]. As the reference solution, we chose the approximation with $N = 2^{19}$. On the plot of Figure 3.6 we the relation between the computational time and the number of discretization points. We clearly observe linear convergence, as expected from the theoretical results.

### Field-circuit coupling

In our second experiment, we consider the example of the field circuit coupling model of a rectifier of Section 2.5, as illustrated in Figure 3.7.

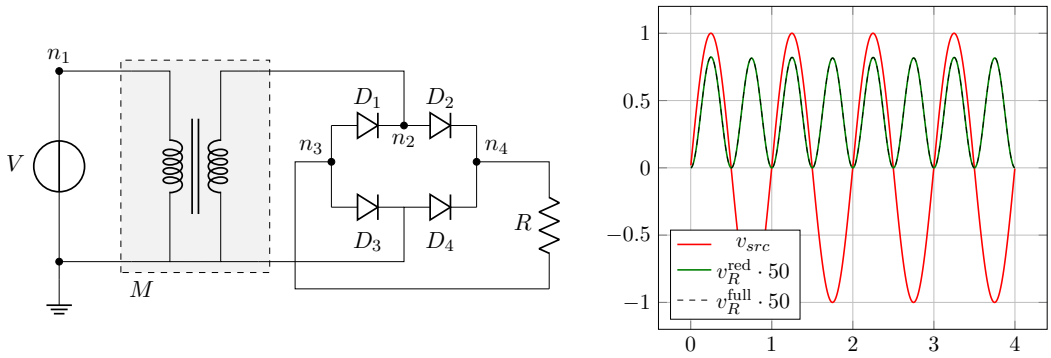



Figure 3.7.: Left: Schematic representation of the rectifier circuit. Right: the input $v_{src}$ and rectified voltage drop at the load $v_R$ at the time interval $[0, 4]$ determined numerically by solving the reduced system (3.42)–(3.43) and of the full system (3.37)–(3.40) for comparison.

**Problem description.** Let us recall that the vector potential formulation for the coupled problem is described by the following set of equations

$$A_R G A_R^\top \partial_t \psi + A_V \partial_t q_V + A_M \partial_t q_M = 0, \tag{3.37}$$

$$-A_V^\top \partial_t \psi = -v_{src}, \tag{3.38}$$

$$\mathsf{M}_\sigma \partial_t a(t) + \mathsf{K}_\nu a(t) - \mathsf{B} \partial_t q_M(t) = 0, \tag{3.39}$$

$$-A_M^\top \partial_t \psi(t) + \mathsf{B}^\top \partial_t a(t) = 0, \tag{3.40}$$

where $G = G(A_R^\top \partial_t \psi)$ is nonlinear. For a precise description of the geometry for the field element, the topology of the circuit, and further details on the model, we refer to Section 2.5. The only change is that we now consider a linear model for the field element, i.e., $\mathsf{K}_\nu$ is constant. In the numerical experiments, we set constant $\nu = 1$ in the complete domain $\Omega$ and $\sigma = 10^3$ in conducting domain $\Omega_c$.

**Volterra integro-differential equation formulation.** By the linearity of the subsystem describing the field element, the magnetic vector potential can be eliminated from the subsystem in the frequency domain. Applying the Laplace transform to (3.39) and (3.40) and rearranging the terms the charge flux relation can be formulated as

$$A_M s \widehat{q_M}(s) = \hat{k}(s) \widehat{\psi}(s), \quad \hat{k}(s) = A_M (\mathsf{B}^\top (s\mathsf{M}_\sigma + \mathsf{K}_\nu)^{-1} \mathsf{B})^{-1} A_M^\top. \tag{3.41}$$

where $\widehat{\psi}$ and $\widehat{q}_M$ denote the Laplace transforms of $\psi$ and $q_M$, respectively. Assuming the trivial initial values, the current-flux relation in the time domain becomes a convolution

$$A_M \partial_t q_M = \int_0^t k(t-s)\psi(s)\,ds,$$

where the kernel $k$ is given in the frequency domain via (3.41). Substituting this expression in (3.37) together with (3.38) leads to a Volterra integro-differential system

$$A_R G A_R^\top \partial_t \psi + A_V \partial_t q_V + \int_0^t k(t-s)\psi(s)\,ds = 0, \tag{3.42}$$

$$-A_V^\top \partial_t \psi = -v_{src}(t). \tag{3.43}$$

In our example, the circuit part has dimension 5 composed of four node potentials and one charge drop across the voltage source. The dimension of the equations for the field element can be arbitrarily large, depending on the accuracy of the semi-discretization. In our simulation, we use discretization in space which results in $a(t) \in \mathbb{R}^{640}$. Hence, the reduced system (3.42)–(3.43) is much smaller than the coupled system (3.37)–(3.40), which reflects in the numerical treatment, as illustrated below.

**Discretization.** We implemented only the second-order method. The circuit equations are discretized by the lowest order Petrov Galerkin scheme; see Example 1.2.16. The treatment of the Volterra term is based on the corresponding collocation points. The nearfield contributions are computed using the corresponding convolution quadrature method. The minimal size of the nearfield is again chosen to be size 16. The right plot in Figure 3.7 illustrates the input voltage $v_{src}(t)$ and the numerical solution to the rectified voltage at the load $v_R^{\mathrm{red}}(t)$. For comparison, we also plot the solution of the full system.

**Convergence results.** We applied the plain convolution quadrature and the presented adaptive methods to this problem. The left plot in Figure 3.8 illustrates the error convergence of the potential $\psi^3$ at the node $n_3$ for different schemes. As expected we observe second-order convergence. Here, we compute the error via $e_h = \max_{t_i \in \mathcal{T}_h} |\psi^3(t_i) - \psi_h^3(t_i)|$, whereas the reference solution we use the numerical solution obtained by a solution of the full system with $N_\infty = 16384$. For these results, we set $L = N$ and $\epsilon = 10^{-8}$ for the computation of the convolution quadrature weights, we chose the polynomial degree $q = 8$ for the approximation of the kernel and use $Lp = 16$ points for the evaluation of inverse Laplace transform.

The observed saturation of convergence is the consequence of the inexact realization of the inverse Laplace transform. The right plot in Figure 3.8 illustrates the discrepancy between the numerical solutions of full and reduced systems on the same grid. As expected, up to the error in the compression and inexact realization of the convolution quadrature weights and inverse Laplace formula, the solutions coincide; see [39, 46]. In particular, we can conclude, that the energy-dissipation balance, as discussed in Section 2.5 remains valid up to an error that is controlled by the discretization parameters.

**Computation times.** The main advantage of the approximation of the reduced system is the gain in efficiency. The convolution quadrature and the presented method require evaluating the transfer function at particular frequencies. This is a costly procedure, which

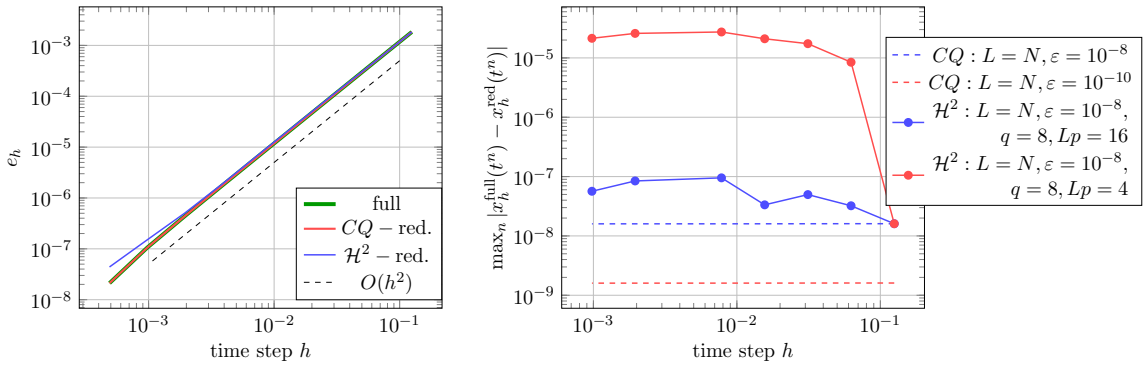



Figure 3.8.: Convergence of the schemes and discrepancy to the numerical solution of the fully coupled problem.

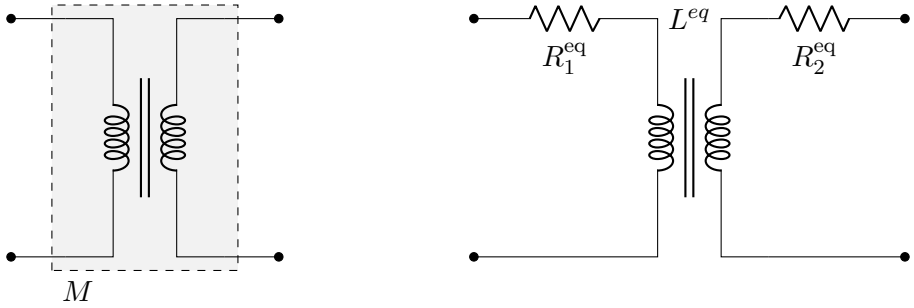



Figure 3.9.: Equivalent circuit.

requires the solution of large-scale problems. For a fixed discretization and number of time steps $N$ this can be done in the pre-processing, independent of the circuit.

We compare the method to the solution of the fully coupled system, to the solution of the reduced system, where the integral term is computed using the convolution quadrature technique, and to the solution of an equivalent circuit problem. For the latter, we consider an approximation of the transformer by a simplified circuit illustrated in Figure 3.9. Replacing the field element with the equivalent circuit leads us to the system of size 7; the two additional degrees of freedom are potential at the nodes between resistors and inductors in the equivalent circuit. The parameters for the circuit $R^{\text{eq}} = \text{diag}(R_1^{\text{eq}}, R_2^{\text{eq}})$ and $L^{\text{eq}} \in \mathbb{R}^{2\times 2}$ are determined by fitting

$$R^{\text{eq}} + sL^{\text{eq}} \approx \mathsf{B}^\top s(s\mathsf{M}\sigma + \mathsf{K}_\nu)^{-1}\mathsf{B},$$

in a least square sense over the range of frequencies $[0, 10^3]$. Let us note that the circuit is a simple approximation, resulting in a modeling error $10^{-1}$, and is intended for illustrative purposes only. Systematic construction of equivalent circuits can be achieved using rational approximations of the transfer function.

The computational times for the approaches are summarized in Table 3.1. For the $\mathcal{H}^2-$ approach, we use the set of parameters as previously. As expected, we observe the complexity $O(N)$. For the convolution quadrature approach, we set $L = N$ and $\epsilon = 10^{-8}$. Here, we use a straightforward implementation of the matrix-vector product, which has

| n | full | equiv | $CQ-$red | | $\mathcal{H}^2-$red | |
|---|---|---|---|---|---|---|
| | | | off | on | off | on |
| 512 | 23.35 | 1.04 | 23.36 | 0.94 | 52.98 | 1.28 |
| 1024 | 49.99 | 1.96 | 46.57 | 1.94 | 80.09 | 2.67 |
| 2048 | 99.77 | 4.08 | 95.24 | 3.83 | 156.28 | 5.66 |
| 4096 | 194.39 | 7.94 | 184.09 | 8.33 | 262.63 | 10.38 |
| 8192 | 389.91 | 15.81 | 468.62 | 32.16 | 387.24 | 18.02 |

Table 3.1.: Computational times for different methods.

the complexity $O(N^2)$. The solutions of the coupled system and of the equivalent circuit problem require $O(N)$ operations. We observe that the online computation times of the presented method are comparable to those for the equivalent circuit.

## 3.5. Summary and outlook

In this section, we discussed the fast and oblivious algorithm for the treatment of Volterra integral operators based on $\mathcal{H}^2-$ matrix compression technique. The algorithm is evolutionary and can therefore be used for the treatment of integro-differential equations. The algorithm can also be extended to Volterra integrals of the convolution type with the kernel given implicitly in the frequency domain. The resulting method perfectly fits into the convolution quadrature framework and is closely related to fast and oblivious convolution quadrature methods.

A precise numerical comparison to fast and oblivious convolution quadrature methods is yet to be done. In particular, the adaptive algorithms [12, 86] are yet to be considered. Comparison to the parallel methods and multi-rate co-simulation techniques [29, 118] is another topic of further research.

# Conclusion

This thesis covers several important problems in electrical engineering. First, we discussed two strategies for passivity-preserving discretization of Maxwell's equations in nonlinear media. The key ingredients for our approach were formulating the problem in a certain (port-) Hamiltonian or generalized gradient flow form and utilizing variational methods in space and time. The construction of higher-order schemes with provable discrete energy balance could be done systematically. The first approach utilizes the $\mathbf{e} - \mathbf{h}$ formulation and allows the construction of dissipative schemes, while the second approach is based on the $\mathbf{e} - \mathbf{a}$ formulation and leads to energy-balance-conserving schemes. Both approaches result in implicit time integrators, which for linear media coincide with certain Runge-Kutta methods.

This methodology applies to a variety of problems in nonlinear electromagnetics. Furthermore, similar ideas could also be transferred to electric circuits. In the second chapter of the thesis, we showed that MNA and MONA have structures similar to the $\mathbf{e} - \mathbf{h}$ and $\mathbf{e} - \mathbf{a}$ formulations, respectively. In fact, the development of the MONA approach was motivated purely by the vector potential formulation for field problems. The structural similarity of the two formulations is advantageous for field-circuit coupling. The coupled system then inherits the common generalized gradient flow structure, which ensures its passivity and allows the construction of energy-balance-preserving schemes. Furthermore, we showed that MONA systems have a lower differential-algebraic index than those of the MNA, drastically simplifying the analysis and numerical treatment. Therefore, the MONA approach seems to be a promising method for circuit simulators.

In the last chapter, we discussed an efficient discretization of Volterra-integro-differential systems, mainly motivated by field-circuit coupling and dispersive media. We presented an efficient evolutionary and oblivious algorithm for the approximation of Volterra integrals based on the $\mathcal{H}^2$ matrix compression technique. This approach is related to well-established CQ methods and represents an improvement over the FOCQ technique. However, it is not restricted to convolution-type integrals, making it more flexible. Using this approach, field-circuit coupled problems could be solved essentially at the cost of solving only the circuit part.

Although the focus of this thesis lies on particular applications, the presented ideas and approaches can be extended to a variety of different problems in electrical engineering and beyond. In particular, the magnetic oriented ansatz to modeling electric circuits represents a fundamentally new idea, opening many possibilities for improvement and further research. In fact, it is a foundation for several publications that are in process.

# Appendix A.

# Variational frameworks

In this thesis we used two variational discretization frameworks [42, 43]. Let us provide a brief summary of the necessary facts. Most of the results are taken from the corresponding publications, where further details are provided.

## A.1. Dissipative framework [42]

Let $\mathbb{H}$ be a Hilbert space and let $\mathbb{V} \subset \mathbb{H}$ and $\mathbb{W} \subset \mathbb{H}$ be two reflexive Banach spaces continuously and densely embedded into $\mathbb{H}$. By identifying $\mathbb{H}$ with its dual space $\mathbb{H}^*$ we obtain the Gelfand triples $\mathbb{V} \subset \mathbb{H} \subset \mathbb{V}^*$ and $\mathbb{W} \subset \mathbb{H} \subset \mathbb{W}^*$ and, therefore, the inclusions $\mathbb{V} \subset \mathbb{W}^*$ and $\mathbb{W} \subset \mathbb{V}^*$ hold. Now let the given energy functional $\mathcal{E} : \mathbb{V} \subset \mathbb{W}^* \to \mathbb{R}$ be differentiable on its domain $\mathrm{dom}(\mathcal{E}) = \{u \in \mathbb{V} : \mathcal{E}(u) < \infty\}$. We assume that the derivative can be written as

$$\mathcal{E}'(u) = Q(u)u, \qquad \forall u \in \mathrm{dom}(\mathcal{E}) \subset \mathbb{V}, \tag{A.1}$$

where $Q(u) : \mathbb{V} \to \mathbb{W}$ is a bounded and linear operator. We now consider the evolution problems of the form

$$Q(u(t))^* \partial_t u(t) = -\mathcal{A}(u(t)), \quad \forall t \geq 0, \tag{A.2}$$

where $Q(u(t))^* : \mathbb{W}^* \to \mathbb{V}^*$ is the dual operator and $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$ is a nonlinear operator. Let us note that the form of $\mathcal{A}$ is not essential for further discussions. Similar results hold for more general systems with $\mathcal{A}(t, u, \partial_t u)$ instead of $\mathcal{A}(u)$. Based on the structural assumptions (A.1)–(A.2) one can immediately derive the power balance and, consequently, the energy balance of the system.

**Lemma A.1.1.** Let $u : [0, T] \to \mathbb{V}$ be sufficiently smooth solution of (A.2). Then

$$\frac{d}{dt}\mathcal{E}(u(t)) = -\langle \mathcal{A}(u(t)), u(t) \rangle_{\mathbb{V}^* \times \mathbb{V}}, \qquad t \geq 0. \tag{A.3}$$

*Proof.* Formal differentiation of the energy functional with respect to time yields

$$\begin{aligned} \frac{d}{dt}\mathcal{E}(u(t)) &= \langle \partial_t u(t), \mathcal{E}'(u(t)) \rangle_{\mathbb{W}^* \times \mathbb{W}} = \langle \partial_t u(t), Q(u(t))u(t) \rangle_{\mathbb{W}^* \times \mathbb{W}} \\ &= \langle Q(u(t))^* \partial_t u(t), u(t) \rangle_{\mathbb{V}^* \times \mathbb{V}} = -\langle \mathcal{A}(u(t)), u(t) \rangle_{\mathbb{V}^* \times \mathbb{V}}, \end{aligned}$$

and concludes the proof. □

Integration in time leads to the corresponding energy balance

$$\mathcal{E}(u(t)) - \mathcal{E}(u(s)) = -\int_s^t \langle \mathcal{A}(u(\tau)), u(\tau)\rangle_{\mathbb{V}^*\times\mathbb{V}}\, d\tau. \tag{A.4}$$

In the following, we discuss the discretization strategy of [42], which respects this energy balance principle. More precisely, the strategy allows the construction of schemes that fulfill (A.4) with "$\leq$" instead of equality. Therefore we call this framework "dissipative".

Let us note, that the proof of the statement is based purely on variational arguments. A solution to the system (A.2) can be characterized by the variational principle

$$\langle Q(u(t))^*\partial_t u(t), v\rangle_{\mathbb{V}^*\times\mathbb{V}} = -\langle \mathcal{A}(u(t)), v\rangle_{\mathbb{V}^*\times\mathbb{V}}, \qquad \forall v \in \mathbb{V},\ t \geq 0. \tag{A.5}$$

The derivation of the power balance relies on this variational principle with $v = u(t)$ as a test function. This is the key ingredient used for the numerical treatment and motivates the use of Galerkin schemes for spatial approximation.

### Galerkin approximation in space

Let $\mathbb{V}_h \subset \mathbb{V}$ denote a closed subspace. The Galerkin approximation of (A.5) in space reads

$$\langle Q(u_h(t))^*\partial_t u_h(t), v_h\rangle_{\mathbb{V}^*\times\mathbb{V}} = -\langle \mathcal{A}(u_h(t)), v_h\rangle_{\mathbb{V}^*\times\mathbb{V}}, \qquad \forall v_h \in \mathbb{V}_h,\ t \geq 0. \tag{A.6}$$

Due to the particular structure of the problem, the power balance (A.7) is then preserved under the approximation, as summarized in the following lemma.

**Lemma A.1.2.** Let $u : [0, T] \to \mathbb{V}_h$ be a smooth solution to (A.6). Then

$$\frac{d}{dt}\mathcal{E}(u_h(t)) = -\langle \mathcal{A}(u_h(t)), u_h(t)\rangle_{\mathbb{V}^*\times\mathbb{V}}, \qquad t \geq 0. \tag{A.7}$$

*Proof.* The proof of Lemma A.1.1 can be directly transferred to the semi-discrete case. □

Integration in time leads to corresponding energy balance

$$\mathcal{E}(u_h(t)) - \mathcal{E}(u_h(s)) = -\int_s^t \langle \mathcal{A}(u_h(\tau)), u_h(\tau)\rangle_{\mathbb{V}^*\times\mathbb{V}}\, d\tau.$$

In other words, the energy balance (A.4) is preserved under a Galerkin semi-discretization. Similar results also hold for specific approximations of individual terms in (A.6). Certain inexact realizations of the duality brackets can also be used.

The result holds for any Galerkin-type approximation. In this thesis, we specifically focus on finite element schemes. However, since we apply this strategy to various types of problems, particularly those with finite dimensions, we do not discuss the details here.

### Discontinuous Galerkin discretization in time

Let $\mathcal{T} = \{t^n : 0 \le n \le N\}$ be a sequence of discrete time steps $t^n = n\tau$ with $\tau = T/N$. With $I^n = [t^{n-1}, t^n]$ we denote the $n$-th time interval and with $P_k(I^n; \mathbb{V})$ we denote the space of polynomials with values in $\mathbb{V}$. By $P_k(\mathcal{T}; \mathbb{V})$ we denote the space of piece-wise polynomials, i.e., the functions whose restrictions to any interval $I^n$ lie in $P_k(I^n; \mathbb{V})$. We further use $(*)|_{t^n}$ to abbreviate the evaluation of $(*)$ at time $t = t^n$. We now consider the time discretization by the following discontinuous Galerkin method.

**Problem A.1.3.** Let $u(0) \in \mathbb{V}$ be given. For $1 \le n \le N$, find $u^n \in P_k(I^n; \mathbb{V})$ such that

$$\begin{aligned} \int_{I^n} \langle Q(u^n(t))^* \partial_t u^n, v(t) \rangle_{\mathbb{V}^* \times \mathbb{V}} \, dt + \langle Q(u^n)^*(u^n - u^{n-1}), v \rangle_{\mathbb{V}^* \times \mathbb{V}}|_{t^{n-1}} \\ = - \int_{I^n} \langle \mathcal{A}(u^n(t)), v(t) \rangle_{\mathbb{V}^* \times \mathbb{V}} \, dt, \qquad \forall v \in P_k(I^n; \mathbb{V}). \end{aligned} \tag{A.8}$$

The main feature of this approach is that the discrete solution satisfies an energy dissipation inequality, assuming the convexity of the energy functional.

**Lemma A.1.4.** Let $u \in P_k(\mathcal{T}; \mathbb{V})$ be solution of the scheme (A.8) and assume $\mathcal{E}(\cdot)$ is convex. Then

$$\mathcal{E}(u^n(t^n)) - \mathcal{E}(u^m(t^m)) \le - \int_{t^m}^{t^n} \langle \mathcal{A}(u(t)), u(t) \rangle_{\mathbb{V}^* \times \mathbb{V}} \, dt.$$

*Proof.* The following proof is identical to the proof of [42, Theorem 4]. We first consider the case $m = n - 1$. By the fundamental theorem of calculus, we conclude

$$\begin{aligned} \mathcal{E}(u^n(t^n)) - \mathcal{E}(u^{n-1}(t^{n-1})) &= \mathcal{E}(u^n(t^n)) - \mathcal{E}(u^n(t^{n-1})) + \mathcal{E}(u^n(t^{n-1})) - \mathcal{E}(u^{n-1}(t^{n-1})) \\ &= \int_{I^n} \frac{d}{dt} \mathcal{E}(u^n(t)) \, dt + \big( \mathcal{E}(u^n(t^{n-1})) - \mathcal{E}(u^{n-1}(t^{n-1})) \big) \\ &= \int_{I^n} \langle \partial_t u^n(t), \mathcal{E}'(u^n(t)) \rangle_{\mathbb{W}^* \times \mathbb{W}} \, dt + \big( \mathcal{E}(u^n(t^{n-1})) - \mathcal{E}(u^{n-1}(t^{n-1})) \big) \\ &= (*) + (**). \end{aligned}$$

Since $\mathcal{E}(\cdot)$ is convex, we can conclude the following inequality for the $(**)$ term

$$\begin{aligned} (**) = \mathcal{E}(u^n(t^{n-1})) - \mathcal{E}(u^{n-1}(t^{n-1})) &\le \langle u^n(t^{n-1}) - u^{n-1}(t^{n-1}), \mathcal{E}'(u^n(t^{n-1})) \rangle_{\mathbb{W}^* \times \mathbb{W}} \\ &= \langle u^n(t^{n-1}) - u^{n-1}(t^{n-1}), Q(u^n(t^{n-1})) u^n(t^{n-1}) \rangle_{\mathbb{W}^* \times \mathbb{W}} \\ &= \langle Q(u^n(t^{n-1}))^* (u^n(t^{n-1}) - u^{n-1}(t^{n-1})), u^n(t^{n-1}) \rangle_{\mathbb{V}^* \times \mathbb{V}}. \end{aligned}$$

Using the scheme (A.8) with test function $v = u^n \in P_k(I^n; \mathbb{V})$ and making use of the energy relation $\mathcal{E}'(u^n) = Q(u^n)u^n$ we further obtain

$$\begin{aligned} (*) &= \int_{I^n} \langle \partial_t u^n(t), Q(u^n(t)) u^n(t) \rangle_{\mathbb{W}^* \times \mathbb{W}} \, dt = \int_{I^n} \langle Q(u^n(t))^* \partial_t u^n(t), u^n(t) \rangle_{\mathbb{V}^* \times \mathbb{V}} \, dt \\ &= - \int_{I^n} \langle \mathcal{A}(u^n(t)), u^n(t) \rangle_{\mathbb{V}^* \times \mathbb{V}} \, dt - \langle Q(u^n(t^{n-1}))^* (u^n(t^{n-1}) - u^{n-1}(t^{n-1})), u^n(t^{n-1}) \rangle_{\mathbb{V}^* \times \mathbb{V}}. \end{aligned}$$

Taking the sum of the two terms $(*)$ and $(**)$ yields

$$(*) + (**) \leq -\int_{I^n} \langle \mathcal{A}(u^n(t)), u^n(t) \rangle_{\mathbb{V}^* \times \mathbb{V}} \, dt,$$

which proves the statement for $m = n - 1$. The case $m < n - 1$ follows by induction. □

Since the underlying structure is preserved under Galerkin approximation in space, the time stepping scheme can be applied to the semi-discrete problem. Hence, the energy dissipation balance can be obtained at the fully discrete level.

## A.2. Conservative framework [43]

The second framework [43] used in this thesis is applicable to problems with a different structure. Consider an abstract evolution problem of the form

$$C(u(t))\partial_t u(t) = -\mathcal{H}'(u(t)), \qquad t \geq 0, \tag{A.9}$$

where $\mathcal{H} : \mathbb{V} \to \mathbb{R}$ is the energy functional with the derivative $\mathcal{H}' : \mathbb{V} \to \mathbb{V}^*$. Here, $\mathbb{V}$ is a Banach space with the duality product $\langle \cdot, \cdot \rangle$. We assume that $C(u(t)) : \mathbb{V} \to \mathbb{V}^*$ is linear and sufficiently smooth and $f(t) \in \mathbb{V}^*$. Let us note that the particular form of $C$ is not essential for the method. Similar results can be obtained with $C(t, \partial_t u, u)$ instead of $C(u)$. Furthermore, we may consider $f(t, \partial_t, u)$ with the same arguments.

Similar to the previously discussed framework, the specific structure of the problem allows for direct access to the power balance, as summarized in the following lemma.

**Lemma A.2.1.** Let $u : [0, T] \to \mathbb{V}$ be sufficiently smooth solution of (A.9). Then

$$\frac{d}{dt}\mathcal{H}(u(t)) = -\langle C(u(t))\partial_t u(t), \partial_t u(t) \rangle + \langle f(t), \partial_t u(t) \rangle, \qquad t \geq 0. \tag{A.10}$$

*Proof.* Formal differentiation of the energy functional with respect to time leads to

$$\frac{d}{dt}\mathcal{H}(u(t)) = \langle \mathcal{H}'(u(t)), \partial_t u(t) \rangle = -\langle C(u(t))\partial_t u(t), \partial_t u(t) \rangle + \langle f(t), \partial_t u(t) \rangle$$

and completes the proof. □

Integration in time yields the corresponding energy balance

$$\mathcal{H}(u(t)) - \mathcal{H}(u(s)) = -\int_s^t \langle C(u(\tau))\partial_t u(\tau), \partial_t u(\tau) \rangle + \langle f(\tau), \partial_t u(\tau) \rangle. \tag{A.11}$$

We now discuss the discretization strategy which allows the construction of schemes that preserve this energy balance exactly. Therefore, we call this framework "conserving".

As in the previous framework, the derivation of the strategy in this framework is based on purely variational arguments. We simply utilize the variational formulation

$$\langle C(u(t))\partial_t u(t), v \rangle = -\langle \mathcal{H}'(u(t)), v \rangle + \langle f(t), v \rangle \tag{A.12}$$

for the problem (A.9) with $v = \partial_t u(t)$. This is the key ingredient for constructing numerical schemes and motivates the use of Galerkin-type schemes for discretization. In contrast to the framework discussed in Section A.1, where the energy balance arises from testing with the solution, in this framework we need to test with the time derivative. This motivates the use of a different variational method for time integration, namely, a certain Petrov-Galerkin approach.

### Galerkin approximation in space

Let $\mathbb{V}_h \subset \mathbb{V}$ be a close subspace. The Galerkin approximation of (A.12) in space reads

$$\langle C(u_h(t))\partial_t u_h(t), v_h\rangle = -\langle \mathcal{H}'(u_h(t)), v_h\rangle + \langle f(t), v_h\rangle, \quad \forall v_h \in \mathbb{V}_h, t \geq 0. \quad \text{(A.13)}$$

The following Lemma shows that the power balance (A.10) holds for the approximation.

**Lemma A.2.2.** Let $[0,T] \to \mathbb{V}_h$ be a smooth solution of (A.13). Then

$$\frac{d}{dt}\mathcal{H}(u_h(t)) = -\langle C(u_h(t))\partial_t u_h(t), \partial_t u_h(t)\rangle + \langle f(t), \partial_t u_h(t)\rangle, \quad t \geq 0.$$

*Proof.* The proof of Lemma A.2.1 translates verbatim since $v_h = \partial_t u_h(t)$ is an admissible test function. □

Integration in time leads to the same energy balance (A.11) for the semi-discretization. Therefore, the energy balance is preserved under the Galerkin approximation in space.

### Petrov-Galerkin discretization in time

Let us recall the notation of Section A.1. For the numerical discretization in time we consider the following Petrov-Galerkin time-stepping strategy.

**Problem A.2.3.** Let $u(0)$ be given. Find $u \in P_{k+1}(\mathcal{T};\mathbb{V}) \cap C([0,T;\mathbb{V}])$ such that

$$\int_{I^n} \langle C(u(t))\partial_t u(t), \bar{v}(t)\rangle \, dt = -\int_{I^n} \langle \mathcal{H}'(u(t)), \bar{v}(t)\rangle \, dt + \int_{I^n} \langle f(t), \bar{v}(t)\rangle \, dt, \quad \text{(A.14)}$$

holds for all $\bar{v} \in P_k(I^n;\mathbb{V})$ and $1 \leq n \leq N$.

Let us note that the trial functions are of polynomial degree $k+1$ and continuous in time, whereas the test functions are of polynomial degree $k$ and can be discontinuous at the junctions. Hence, this method is indeed a Petrov-Galerkin approach.

**Lemma A.2.4.** Let $u \in P_{k+1}(\mathcal{T};\mathbb{V}) \cap C([0,T;\mathbb{V}])$ be a solution of (A.14). Then

$$\mathcal{H}(u(t^n)) - \mathcal{H}(u(t^m)) = -\int_{t^m}^{t^n} \langle C(u(t))\partial_t u(t), \partial_t u(t)\rangle \, dt + \int_{t^m}^{t^n} \langle f(t), \partial_t u(t)\rangle \, dt. \quad \text{(A.15)}$$

*Proof.* The proof is identical to the proof of [43, Theorem 2]. First, we consider $m = n-1$. By the fundamental theorem of calculus, we conclude

$$\mathcal{H}(u(t^n)) - \mathcal{H}(u(t^{n-1})) = \int_{I^n} \frac{d}{dt}\mathcal{H}(u(t)) \, dt = \int_{I^n} \langle \mathcal{H}'(u(t)), \partial_t u^n(t)\rangle \, dt = (*).$$

Using (A.15) with $\bar{v} = \partial_t u^n$, which is an admissible test function, we conclude

$$(*) = \int_{I^n} \langle C(u^n)\partial_t u^n(t), \partial_t u^n(t)\rangle \, dt + \int_{I^n} \langle f(t), \partial_t u^n(t)\rangle \, dt$$

The general case $m < n-1$ follows by induction with continuity of the solution $u$. □

Since the Galerkin approximation in space preserves the underlying structure, the time-stepping method can be applied to the semi-discrete problem. Hence, the fully discrete schemes preserve the energy balance exactly. This property is especially advantageous for energy-conserving systems.

# Appendix B.

# Electric circuits

## B.1. Graphs

**Definition B.1.1.** Graph related definitions:

1. A graph is a set of branches. If branches are oriented, the graph is called oriented. The ends of the branches are called nodes. We denote the graph by $\mathcal{G} = (\mathcal{N}, \mathcal{B})$ with $\mathcal{N}$ the set of nodes and $\mathcal{B}$ the set of branches.
2. A set of branches $\{b_1, \dots b_k\}$ is called a path between node $i$ and $j$ if
   - the two following branches have a common node
   - each node belongs to two branches except for nodes $i$ and $j$ which belong to exactly one path
3. A graph is connected if there exist at least one path between any nodes.
4. A subgraph $\tilde{\mathcal{G}} \subset \mathcal{G}$ is called a loop if $\tilde{\mathcal{G}}$ is connected and each node has exactly two branches
5. A subgraph of connected graph $\tilde{\mathcal{G}} \subset \mathcal{G}$ is called a tree if $\tilde{\mathcal{G}}$ is connected, contains all nodes of $\mathcal{G}$ and has no loops
6. A set of branches $\tilde{\mathcal{B}}$ is called a cutset, if removing it the graph becomes disconnected and adding any branch of $\tilde{\mathcal{B}}$ would again lead to a connected graph

## B.2. Kirchhoff's current law

**Lemma B.2.1.** Let $\tilde{A}$ and $A$ denote full and reduced incidence matrices. Then the Kirchhoff's laws

$$Ai = 0 \quad \text{and} \quad \tilde{A}i = 0$$

are equivalent.

*Proof.* Let $A_i$ denote rows of matrix $\tilde{A}$. Without loss of generality we assume the reduced incidence matrix $A$ is constructed by removing the last row of the full incidence matrix

$$\tilde{A}^\top = (A_1^\top, ..., A_n^\top) \quad \text{and} \quad A^\top = (A_1^\top, ..., A_{n-1}^\top).$$

"$\Leftarrow$": trivial, since $A$ is the upper block of $\tilde{A}$.

"$\Rightarrow$": Since $\sum_{k=1}^{n} A_k = 0$ we write $A_n = -\sum_{k=1}^{n-1}$ and it holds

$$A_n i = \langle -\sum_{k=1}^{n-1} A_k, i\rangle = -\sum_{k=1}^{n-1}\langle A_k, i\rangle = 0$$

since $\langle A_k, i\rangle = 0$ for $k = 1, \ldots, n-1$. □

## B.3. Loop matrix and Kirchhoff's voltage law

Another important topological construct we utilize is the so-called loop matrix. Let $\mathcal{L}$ denote the set of loops of $G$ and let $n_l = |\mathcal{L}|$ denote the number of loops. For the chosen orientation of the loop we define full loop matrix $\tilde{B} \in \mathbb{R}^{n_l \times n_b}$ by

$$\tilde{B}_{ij} = \begin{cases} 1 & \text{if branch } b_j \text{ belongs to loop } l_i \text{ and has the same orientation} \\ -1 & \text{if branch } b_j \text{ belongs to loop } l_i \text{ and has the opposite orientation} \\ 0 & \text{else} \end{cases}$$

The rows of $B$ represent loops, each row contains nonzero entries at the positions corresponding to branches the loop contains. The entry 1 corresponds to the branches with the same direction where $-1$ stays for the opposite direction. The columns of $B$ represent branches. Each column has nonzero entries at the positions corresponding to the loops it is contained. The entry 1 corresponds to the case if the orientation of the loop coincides with the orientation of the branch and $-1$ when the orientation is opposite.

As shown in [109, Section 4], the full loop matrix has linear dependent rows. More precisely, for connected graph holds $\operatorname{rank}(\tilde{B}) = n_b - n_n + 1$. The reduced loop matrix $B$ can be constructed by picking $n_b - n_n + 1$ independent row. Let us further mention that incidence and loop matrices satisfy the following relations

$$\ker \tilde{B} = \operatorname{im}(\tilde{A}^\top) \quad \text{and} \quad \ker B = \operatorname{im}(A^\top) \tag{B.1}$$

as shown in [109, Section 4].

**Example B.3.1.** Consider the circuit illustrated in Firgure 2.1a. The circuit contains three loops

$$\mathcal{L} = \{l_1 = (b_1, b_2, b_4), l_2 = (b_1, b_3, b_5), l_3 = (b_4, b_2, b_3, b_5)\}$$

The full and reduced loop matrices are given by

$$\tilde{B} = \begin{pmatrix} -1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} -1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

and $\operatorname{rank}(\tilde{B}) = \operatorname{rank}(B) = 2$. In this example one can directly observe that the sum of first two rows of $\tilde{B}$ equals to the third row. Further, one can directly verify the orthogonality conditions $\tilde{B}\tilde{A}^\top = 0$ and $BA^\top = 0$

**Remark B.3.2.** To provide an analogy with the field equations, let us remark that the operators $A$ could be interpreted as a divergence, $A^\top$ as a gradient, and $B$ as a curl. Where the relation (B.1) represents the vector identity $\operatorname{curl} \nabla = 0$

With a similar notation, we postulate the voltage law by

$$\sum_{j=1}^{n_b} B_{kj} v_j = 0 \quad \text{for all } k = 1, \ldots, n_l \tag{B.2}$$

where $v_j$ is the voltage over branches $b_j$. Thus, in every loop, the sum of the voltages in the clock direction equals the sum of the voltages in the counter clock direction; see Figure 2.2b for illustration. In vector form, the voltage law can be expressed as

$$Bv = 0 \tag{B.3}$$

where $v \in \mathbb{R}^{n_b}$ is the vector of branch voltages. Here as well, the relation (B.2) corresponds to $\tilde{B}v = 0$ for the full loop matrix $\tilde{B}$, although the relations are equivalent based on similar arguments as for Kirchhoff's current law.

The voltage law (B.3) together with identity (B.1) yield the existence of electric node potential $e \in \mathbb{R}^{n_n - 1}$ such that

$$v = A^\top e$$

and since the rows of the reduced matrix are linearly independent the potential is unique. The electric potential of the reference node, associated with the removed row, is zero.

# Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT in order to correct the grammar and improve the language and readability. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

# Bibliography


[1] https://www.skyworksinc.com/-/media/SkyWorks/Documents/Products/1-100/Varactor_SPICE_Model_AN_200315C.pdf.

[2] A. Aanes and L. Angermann. Energy-stable time-domain finite element methods for the 3D nonlinear Maxwell's equations. *IEEE Photonics J.*, 12:6500415, 2020.

[3] J. Adamy. *Nichtlineare Systeme und Regelungen*. Springer, 2018.

[4] G. Akrivis, C. Makridakis, and R. H. Nochetto. Galerkin and Runge–Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. *Numer. Math.*, 118:429–456, 2011.

[5] A. Alonso Rodríguez and A. Valli. *Eddy current approximation of Maxwell equations*, volume 4 of *MS&A. Modeling, Simulation and Applications*. Springer-Verlag Italia, Milan, 2010. Theory, algorithms and applications.

[6] B. Alpert, L. Greengard, and T. Hagstrom. Rapid evaluation of nonreflecting boundary kernels for time-domain wave propagation. *SIAM Journal on Numerical Analysis*, 37(4):1138–1164, 2000.

[7] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977.

[8] A. Anees and L. Angermann. Time domain finite element method for Maxwell's equations. *IEEE Access*, 7:63852–63867, 2019.

[9] W. Arendt, C. Batty, M. Hieber, and F. Neubrander. *Vector-Valued Laplace Transforms and Cauchy Problems*. Springer Basel, Basel, 2011.

[10] A. K. Aziz and P. Monk. Continuous finite elements in space and time for the heat equation. *Mathematics of Computation*, 52(186):255–274, 1989.

[11] D. Baffet and J. Hesthaven. A Kernel Compression Scheme for Fractional Differential Equations. *SIAM Journal on Numerical Analysis*, 55(2):496–520, Jan. 2017.

[12] L. Banjai and M. Ferrari. Generalized convolution quadrature based on the trapezoidal rule. *arXiv preprint arXiv:2305.11134*, 2023.

[13] A. Bartel, S. Baumanns, and S. Schöps. Structural analysis of electrical circuits including magnetoquasistatic devices. *Appl. Numer. Math.*, 61:1257–1270, 2011.

[14] G. Bedrosian. A new method for coupling finite element field solutions with external circuits and kinematics. *IEEE Transactions on Magnetics*, 29(2):1664–1668, 1993.

[15] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta numerica*, 14:1–137, 2005.

[16] E. Blank. *The Discontinuous Galerkin Method for Maxwell's Equations: Application to Bodies of Revolution and Kerr-Nonlinearities*. PhD thesis, KIT, 2013.

[17] D. Boffi, F. Brezzi, M. Fortin, et al. *Mixed finite element methods and applications*, volume 44. Springer, 2013.

[18] V. A. Bokil, Y. Cheng, Y. Jiang, and F. Li. Energy stable discontinuous Galerkin methods for Maxwell's equations in nonlinear optical media. *J. Comput. Phys.*, 350:420–452, 2017.

[19] V. A. Bokil, Y. Cheng, Y. Jiang, F. Li, and P. Sakkaplangkul. High spatial order energy stable FDTD methods for Maxwell's equations in nonlinear optical media in one dimension. *Journal of Scientific Computing*, 77:330–371, 2018.

[20] S. Börm. Construction of data-sparse hˆ2-matrices by hierarchical compression. *SIAM Journal on Scientific Computing*, 31(3):1820–1839, 2009.

[21] S. Börm. *Efficient Numerical Methods for Non-Local Operators*, volume 14 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010.

[22] A. Bossavit. A course in convex analysis. *Lecture notes. ICM, Warsaw*, 2003.

[23] A. Bossavit. Virtual power principle and maxwell's tensor: which comes first? *COMPEL-The international journal for computation and mathematics in electrical and electronic engineering*, 2011.

[24] R. W. Boyd. *Nonlinear Optics*. Academic Press, 3rd edition, 2008.

[25] A. Brandt and A. Lubrecht. Multilevel matrix multiplication and fast solution of integral equations. *Journal of Computational Physics*, 90(2):348–370, Oct. 1990.

[26] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*. SIAM, 1996.

[27] H. Brunner. *Collocation Methods for Volterra Integral and Related Functional Differential Equations*. Number 15 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, UK ; New York, 2004.

[28] H. Brunner. *Volterra Integral Equations: An Introduction to Theory and Applications*. Cambridge University Press, Cambridge, 2017.

[29] K. Burrage. *Parallel and sequential methods for ordinary differential equations*. Clarendon Press, 1995.

[30] S. L. Campbell. *Singular systems of differential equations*. Pitman Publishing, 1980.

[31] E. Celledoni and E. H. Høiseth. Energy-preserving and passivity-consistent numerical discretization of port-Hamiltonian systems. *arXiv preprint arXiv:1706.08621*, 2017.

[32] L. O. Chua, C. A. Desoer, and E. S. Kuh. *Linear and Nonlinear Circuits*. McGraw-Hill, New York, 1987.

[33] L. Codecasa, B. Kapidani, R. Specogna, and F. Trevisan. Novel FDTD technique over tetrahedral grids for conductive media. *IEEE Transactions on Antennas and Propagation*, 66(10):5387–5396, 2018.

[34] L. Codecasa and M. Politi. Explicit, consistent, and conditionally stable extension of FD-TD to tetrahedral grids by FIT. *IEEE transactions on magnetics*, 44(6):1258–1261, 2008.

[35] G. Cohen. *Higher-Order Numerical Methods for Transient Wave Equations*. Springer, Heidelberg, 2002.

[36] G. Cohen and P. Monk. Gauss point mass lumping schemes for Maxwell's equations. *Numer. Meth. Part. Diff. Equat.*, 14:63–88, 1998.

[37] I. Cortes Garcia, S. Schöps, C. Strohm, and C. Tischendorf. Generalized circuit elements. *arXiv preprint arXiv:1912.05199*, 2019.

[38] W. Dahmen, S. Prössdorf, and R. Schneider. Wavelet approximation methods for pseudodifferential equations II: Matrix compression and fast solution. *Advances in Computational Mathematics*, 1(3):259–335, Oct. 1993.

[39] J. Dölz, H. Egger, and V. Shashkov. A convolution quadrature method for Maxwell's equations in dispersive media. In *Scientific Computing in Electrical Engineering: SCEE 2020, Eindhoven, The Netherlands, February 2020*, pages 107–115. Springer, 2021.

[40] J. Dölz, H. Egger, and V. Shashkov. A fast and oblivious matrix compression algorithm for volterra integral operators. *Advances in Computational Mathematics*, 47(6):81, 2021.

[41] P. Dular. The benefits of nodal and edge elements coupling for discretizing global constraints in dual magnetodynamic formulations. *Journal of computational and applied mathematics*, 168(1-2):165–178, 2004.

[42] H. Egger. Structure preserving approximation of dissipative evolution problems. *Numer. Math.*, 143:85–106, 2019.

[43] H. Egger, O. Habrich, and V. Shashkov. On the energy stable approximation of Hamiltonian and gradient systems. *J. Comput. Meth. Appl. Math.*, 21:335–349, 2021.

[44] H. Egger and B. Radu. A mass-lumped mixed finite element method for Maxwell's equations. In *International Conference on Scientific Computing in Electrical Engineering*, pages 15–24. Springer, 2018.

[45] H. Egger and B. Radu. A second-order finite element method with mass lumping for Maxwell's equations on tetrahedra. *SIAM Journal on Numerical Analysis*, 59(2):864–885, 2021.

[46] H. Egger, K. Schmidt, and V. Shashkov. Multistep and Runge–Kutta convolution quadrature methods for coupled dynamical systems. *Journal of Computational and Applied Mathematics*, 387:112618, 2021.

[47] H. Egger and V. Shashkov. On energy preserving high-order discretizations for nonlinear acoustics. In *Numerical Mathematics and Advanced Applications ENUMATH 2019: European Conference, Egmond aan Zee, The Netherlands, September 30-October 4*, pages 353–361. Springer, 2021.

[48] H. Egger and V. Shashkov. On higher order passivity preserving schemes for nonlinear Maxwell's equations. *arXiv preprint arXiv:2202.08003*, 2022.

[49] D. Estep and A. Stuart. The dynamical behavior of the discontinuous Galerkin method and related difference schemes. *Mathematics of computation*, 71(239):1075–1103, 2002.

[50] D. Estévez Schwarz. A step-by-step approach to compute a consistent initialization for the MNA. *Int. J. Circ. Theor. Appl.*, 30:1–16, 2002.

[51] D. Estévez Schwarz and R. Lamour. The computation of consistent initial values for nonlinear index-2 differential–algebraic equations. *Numerical Algorithms*, 26(1):49–75, 2001.

[52] D. Estévez Schwarz and C. Tischendorf. Structural analysis of electric circuits and consequences for MNA. *Int. J. Circuit Theory Appl.*, 28:131–162, 2000.

[53] L. C. Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

[54] W. Fong and E. Darve. The black-box fast multipole method. *Journal of Computational Physics*, 228(23):8712–8725, Dec. 2009.

[55] I. C. Garcia. *Mathematical Analysis and Simulation of Field Models in Accelerator Circuits*. Springer Nature, 2021.

[56] S. Geevers, W. Mulder, and J. van der Vegt. New higher-order mass-lumped tetrahedral elements for wave propagation modelling. *SIAM J. Sci. Comput.*, 40:A2830–A2857, 2018.

[57] K. Giebermann. Multilevel approximation of boundary integral operators. *Computing*, 67(3):183–207, 2001.

[58] O. Gonzalez. Time integration and discrete Hamiltonian systems. *Journal of Nonlinear Science*, 6:449–467, 1996.

[59] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348, 1987.

[60] E. Griepentrog and R. März. *Differential-algebraic equations and their numerical treatment*. Teubner, Leipzig, 1986.

[61] M. Günther and U. Feldmann. The DAE-index in electric circuit simulation. *Math. Comput. Sim.*, 39:573–582, 1995.

[62] M. Günther, U. Feldmann, and J. ter Maten. Modelling and discretization of circuit problems. In *Handbook of numerical analysis*, volume 13, pages 523–659. 2005.

[63] W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices part I: Introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.

[64] W. Hackbusch. *Hierarchical Matrices: Algorithms and Analysis*. Springer, Heidelberg, 2015.

[65] T. Hagstrom. Radiation boundary conditions for the numerical simulation of waves. *Acta Numerica*, 8:47–106, Jan. 1999.

[66] E. Hairer, C. Lubich, and M. Roche. *The numerical solution of differential-algebraic systems by Runge-Kutta methods*. Springer, 2006.

[67] E. Hairer, C. Lubich, and M. Schlichte. Fast Numerical Solution of Nonlinear Volterra Convolution Equations. *SIAM Journal on Scientific and Statistical Computing*, 6(3):532–541, July 1985.

[68] R. Hiptmair and O. Sterz. Current and voltage excitations for the eddy current model. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, 18(1):1–21, 2005.

[69] C.-W. Ho, A. Ruehli, and P. Brennan. The modified nodal approach to network analysis. *IEEE Trans. Circuits Syst.*, 22:504–509, 1975.

[70] M. Hochbruck and T. Pazur. Implicit Runge–Kutta methods and discontinuous Galerkin discretizations for linear Maxwell's equations. *SIAM Journal on Numerical Analysis*, 53(1):485–507, 2015.

[71] H. Huynh. Discontinuous Galerkin and related methods for ODE.

[72] H. Huynh. Collocation and Galerkin time-stepping methods. In *19th AIAA Computational Fluid Dynamics*, page 4323. 2009.

[73] J. D. Jackson. Electrodynamics. *The Optics Encyclopedia: Basic Foundations and Practical Applications*, 2007.

[74] H. Jia, J. Li, Z. Fang, and M. Li. A new FDTD scheme for Maxwell's equations in Kerr-type nonlinear media. *Numer. Algor.*, 81:223–243, 2019.

[75] S. Jiang and L. Greengard. Fast evaluation of nonreflecting boundary conditions for the Schrödinger equation in one dimension. *Computers & Mathematics with Applications*, 47(6-7):955–966, Mar. 2004.

[76] S. Jiang and L. Greengard. Efficient representation of nonreflecting boundary conditions for the time-dependent Schrödinger equation in two dimensions. *Communications on Pure and Applied Mathematics*, 61(2):261–288, Feb. 2008.

[77] R. M. Joseph and A. Taflove. FDTD Maxwell's equations models for nonlinear electrodynamics and optics. *IEEE Trans. Antenn. Prop.*, 45:364–374, 1997.

[78] S. Kapur, D. Long, and J. Roychowdhury. Efficient time-domain simulation of frequency-dependent elements. In *Proceedings of International Conference on Computer Aided Design*, pages 569–573, San Jose, CA, USA, 1996. IEEE Comput. Soc. Press.

[79] J. Kaye and D. Golez. Low rank compression in the numerical solution of the nonequilibrium Dyson equation. *SciPost Physics*, 10(4):091, 2021.

[80] S. Kirchhoff. Ueber den Durchgang eines elektrischen Stromes durch eine Ebene, insbesondere durch eine kreisförmige. *Annalen der Physik*, 140(4):497–514, 1845.

[81] P. Kunkel and V. Mehrmann. *Differential-algebraic equations*. European Mathematical Society (EMS), Zürich, 2006. Analysis and numerical solution.

[82] R. Leis. *Initial boundary value problems in mathematical physics*. Courier Corporation, 2013.

[83] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. *Publications mathématiques et informatique de Rennes*, (S4):1–40, 1974.

[84] L. Li, M. Lyu, and W. Zheng. An energy-stable finite element method for nonlinear Maxwell's equations. *Journal of Computational Physics*, page 112135, 2023.

[85] M. López-Fernández, C. Palencia, and A. Schädle. A Spectral Order Method for Inverting Sectorial Laplace Transforms. *SIAM Journal on Numerical Analysis*, 44(3):1332–1350, Jan. 2006.

[86] M. Lopez-Fernandez and S. Sauter. Generalized convolution quadrature with variable time stepping. *IMA Journal of Numerical Analysis*, 33(4):1156–1175, 2013.

[87] R. Lozano, B. Brogliato, O. Egeland, and B. Maschke. *Dissipative systems analysis and control: theory and applications*. Springer Science & Business Media, 2013.

[88] C. Lubich. Convolution quadrature and discretized operational calculus. I. *Numerische Mathematik*, 52(2):129–145, Jan. 1988.

[89] C. Lubich. Convolution quadrature and discretized operational calculus. II. *Numerische Mathematik*, 52(4):413–425, July 1988.

[90] C. Lubich. Convolution Quadrature Revisited. *BIT Numerical Mathematics*, 44(3):503–514, Aug. 2004.

[91] C. Lubich and A. Ostermann. Runge-Kutta methods for parabolic equations and convolution quadrature. *Mathematics of Computation*, 60(201):105–105, Jan. 1993.

[92] C. Lubich and A. Schädle. Fast Convolution for Nonreflecting Boundary Conditions. *SIAM Journal on Scientific Computing*, 24(1):161–182, Jan. 2002.

[93] C. Makridakis and R. H. Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numerische Mathematik*, 104(4):489–514, 2006.

[94] I. S. Maksymov, A. A. Sukhorukov, A. V. Lavrinenko, and Y. S. Kivshar. Comparative study of FDTD-adopted numerical algorithms for Kerr nonlinearities. *IEEE Antennas and Wireless Propagation Letters*, 10:143–146, 2011.

[95] G. Matthies and F. Schieweck. Higher order variational time discretizations for nonlinear systems of ordinary differential equations. 2011.

[96] V. Mehrmann. Index concepts for differential-algebraic equations. 2012.

[97] V. Mehrmann and R. Morandin. Structure-preserving discretization for port-Hamiltonian descriptor systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6863–6868. IEEE, 2019.

[98] V. Mehrmann and A. van der Schaft. Differential-algebraic systems with dissipative Hamiltonian structure. *arXiv preprint arXiv:2208.02737*, 2022.

[99] P. Monk. Analysis of a finite element method for Maxwell's equations. *SIAM Journal on Numerical Analysis*, 29(3):714–729, 1992.

[100] P. Monk et al. *Finite element methods for Maxwell's equations*. Oxford University Press, 2003.

[101] P. B. Monk. A mixed method for approximating Maxwell's equations. *SIAM Journal on Numerical Analysis*, 28(6):1610–1634, 1991.

[102] J.-C. Nédélec. Mixed finite elements in R3. *Numerische Mathematik*, 35(3):315–341, 1980.

[103] Z. Peng, V. A. Bokil, Y. Cheng, and F. Li. Asymptotic and positivity preserving methods for Kerr-Debye model with Lorentz dispersion in one dimension. *Journal of Computational Physics*, 402:109101, 2020.

[104] T. Péra, F. Ossart, and T. Waeckerle. Numerical representation for anisotropic materials based on coenergy modeling. *Journal of applied physics*, 73(10):6784–6786, 1993.

[105] L. Petzold. Differential/algebraic equations are not ODE's. *SIAM J. Sci. Stat. Comput.*, 3:367–384, 1982.

[106] M. Pototschnig, J. Niegemann, L. Tkeshelashvili, and K. Busch. Time-domain simulations of the nonlinear Maxwell equations using operator-exponential methods. *IEEE Transactions on Antennas and Propagation*, 57(2):475–483, 2009.

[107] G. Quispel and D. I. McLaren. A new class of energy-preserving numerical integration methods. *Journal of Physics A: Mathematical and Theoretical*, 41(4):045206, 2008.

[108] B. Radu. *Finite element mass lumping for H (div) and H (curl)*. PhD thesis, Dissertation, Darmstadt, Technische Universität Darmstadt, MAGA, 2022.

[109] T. Reis. Mathematical modeling and analysis of nonlinear time-invariant RLC circuits. In *Large-scale networks in engineering and life sciences*, pages 125–198. Springer, 2014.

[110] R. Riaza. *Differential-Algebraic Systems: Analytical Aspects and Circuit Applications*. World Scientific, 2008.

[111] R. Riaza and J. Torres-Ramírez. Non-linear circuit modelling via nodal methods. *Int. J. Circ. Theor. Appl.*, 33:281–305, 2005.

[112] R. T. Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

[113] V. Rokhlin. Rapid solution of integral equations of classical potential theory. *Journal of Computational Physics*, 60(2):187–207, Sept. 1985.

[114] F.-J. Sayas. *Retarded Potentials and Time Domain Boundary Integral Equations*, volume 50 of *Springer Series in Computational Mathematics*. Springer International Publishing, Cham, 2016.

[115] A. Schädle, M. López-Fernández, and C. Lubich. Fast and Oblivious Convolution Quadrature. *SIAM Journal on Scientific Computing*, 28(2):421–438, Jan. 2006.

[116] L. Scholz. The Signature Method for DAEs arising in the modeling of electrical circuits. *J. Comput. Appl. Math.*, 332:107–139, 2018.

[117] S. Schöps. *Multiscale modeling and multirate time-integration of field/circuit coupled problems*. PhD thesis, Universität Wuppertal, Fakultät für Mathematik und Naturwissenschaften ..., 2011.

[118] S. Schops, H. De Gersem, and A. Bartel. A cosimulation framework for multirate time integration of field/circuit coupled problems. *IEEE Transactions on Magnetics*, 46(8):3233–3236, 2010.

[119] S. Schöps, H. De Gersem, and T. Weiland. Winding functions in transient magnetoquasistatic field-circuit coupled simulations. *COMPEL: The international journal for computation and mathematics in electrical and electronic engineering*, 2013.

[120] R. Schuhmann and T. Weiland. A stable interpolation technique for FDTD on non-orthogonal grids. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, 11(6):299–306, 1998.

[121] D. E. Schwarz. Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation. 2000.

[122] V. Shashkov, I. Cortes Garcia, and H. Egger. MONA—a magnetic oriented nodal analysis for electric circuits. *International Journal of Circuit Theory and Applications*, 2022.

[123] G. Shi. An analog circuit analysis method by node elimination. *Analog Integrated Circuits and Signal Processing*, 109(1):247–252, 2021.

[124] G. Shi, S. X.-D. Tan, and E. Tlelo Cuautle. *Advanced Symbolic Analysis for VLSI Systems*. Springer Science+Business Media New York, 2014.

[125] P. P. Silvester and R. P. Gupta. Effective computational models for anisotropic soft BH curves. *IEEE Transactions on Magnetics*, 27(5):3804–3807, 1991.

[126] M. P. Sørensen, G. M. Webb, M. Brio, and J. V. Moloney. Kink shape solutions of the Maxwell-Lorentz system. *Physical Review E*, 71(3):036602, 2005.

[127] M. Sova. The Laplace transform of analytic vector-valued functions (complex conditions). *Casopis pro pestování matematiky*, 104(3):267–280, 1979.

[128] O. Sterz. *Modellierung und Numerik zeitharmonischer Wirbelstromprobleme in 3D*. PhD thesis, 2003.

[129] J. A. Stratton. *Electromagnetic theory*, volume 33. John Wiley & Sons, 2007.

[130] A. Taflove. Application of the finite-difference time-domain method to sinusoidal steady-state electromagnetic-penetration problems. *IEEE Transactions on electromagnetic compatibility*, (3):191–202, 1980.

[131] A. Taflove, S. C. Hagness, and M. Piket-May. Computational electromagnetics: the finite-difference time-domain method. *The Electrical Engineering Handbook*, 3, 2005.

[132] M. Takamatsu and S. Iwata. Index characterization of differential–algebraic equations in hybrid analysis for circuit simulation. *Int. J. Circ. Theor. Appl.*, 38:419–440, 2010.

[133] W. Tang and Y. Sun. Time finite element methods: a unified framework for numerical discretizations of ODEs. *Applied Mathematics and Computation*, 219(4):2158–2179, 2012.

[134] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25. Springer Science & Business Media, 2007.

[135] W. Tierens. Unification of leapfrog and Crank–Nicolson finite difference time domain methods. *SIAM journal on Scientific Computing*, 40(1):A306–A330, 2018.

[136] C. Tischendorf. Topological index calculation of differential-algebraic equations in circuit simulation. *Surveys Math. Indust.*, 8:187–199, 1999.

[137] C. Tischendorf. *Coupled systems of differential algebraic and partial differential equations in circuit and device simulation*. Habilitation, Humboldt-University Berlin, 2003.

[138] I. A. Tsukerman, A. Konrad, G. Meunier, and J. C. Sabonnadiere. Coupled field-circuit problems: trends and accomplishments. *IEEE Transactions on magnetics*, 29(2):1701–1704, 1993.

[139] A. Van Der Schaft, D. Jeltsema, et al. Port-Hamiltonian systems theory: An introductory overview. *Foundations and Trends® in Systems and Control*, 1(2-3):173–378, 2014.

[140] J.-S. Wang. A nodal analysis approach for 2d and 3d magnetic-circuit coupled problems. *IEEE transactions on magnetics*, 32(3):1074–1077, 1996.

[141] T. Weiland. A discretization model for the solution of Maxwell's equations for six-component fields. *Archiv Elektronik und Uebertragungstechnik*, 31:116–120, 1977.

[142] T. Weiland. Time domain electromagnetic field computation with finite difference methods. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, 9(4):295–319, 1996.

[143] T. Weiland. Finite integration method and discrete electromagnetism. In *Computational Electromagnetics: Proceedings of the GAMM Workshop on Computational Electromagnetics, Kiel, Germany, January 26–28, 2001*, pages 183–198. Springer, 2003.

[144] K. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Transactions on antennas and propagation*, 14(3):302–307, 1966.

# Wissenschaftlicher Werdegang

Vsevolod Shashkov

| | |
|---|---|
| 2023 | Promotion Mathematik |
| 2018-2023 | Doktorand, AG Numerilk, TU Darmstadt |
| 2014-2017 | Master of Science in Scientific Computing, TU Berlin |
| 2009 - 2014 | Bachelor of Science in Mathematik, Universität des Saarlandes |
| 2009 | Abitur, Albertus-Magnus-Gymnasium in St.Ingbert |