ADAPTIVE VECTOR FINITE ELEMENT METHODS FOR THE MAXWELL EQUATIONS

The research presented in this thesis was carried out at the Numerical Analysis and Computational Mechanics (NACM) group, Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Applied Mathematics, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands.

This research was supported by the Netherlands Organization of Scientific Research (NWO) under the project number 635.000.011 in the Computational Science Program.

Printed by Wöhrmann Print Service, The Netherlands

ISBN 978-90-365-2508-4

ADAPTIVE VECTOR FINITE ELEMENT METHODS FOR THE MAXWELL EQUATIONS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. W.H.M. Zijm,
on account of the decision of the graduation committee,
to be publicly defended
on Friday, 25 May 2007 at 13.15

by

Davit Harutyunyan
born on 22 June 1979
in Yerevan, Armenia

Dit proefschrift is goedgekeurd door de promotor,
Prof. dr. ir. J. J. W. van der Vegt

en de assistent-promotor,
Dr. M. A. Botchev

This thesis is dedicated
to my beloved parents, brothers
and to my lovely wife

CONTENTS

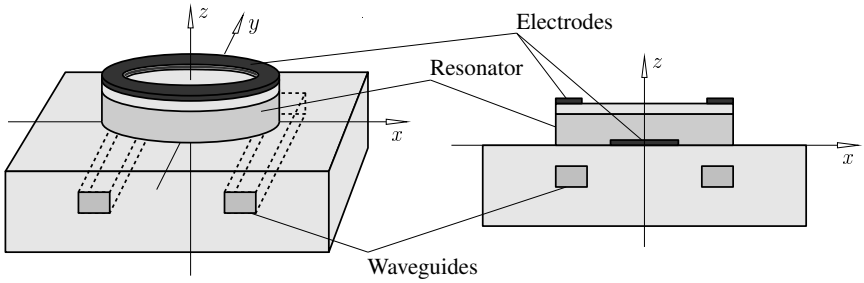
CHAPTER 1

Introduction

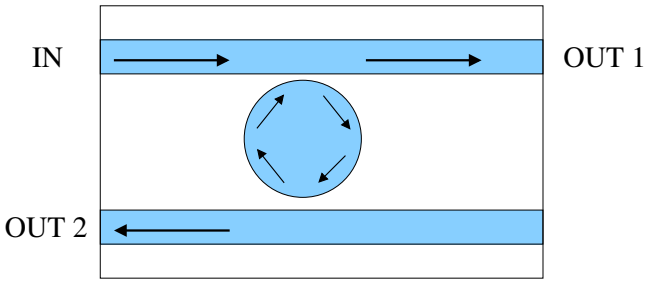
1.1 Motivation

Understanding electromagnetic waves in complicated, often small devices (computer chips, mobile phones, optical switches and other micro-electronic equipment) is a real challenge for engineers. These waves are described by the Maxwell equations which provide an accurate mathematical model for electromagnetic waves. Despite the great progress made in the theoretical understanding of these equations and the development of numerical techniques, there is a clear need for more accurate and efficient numerical methods for the Maxwell equations. In particular, they are of great importance in the design and analysis of micro-electronic devices.

As a specific example let us consider a micro-resonator. This device consists of a circular disk made of high refractive index material and the micro-resonator, which is placed in between two waveguides, see Figure 1.1(a). Light of any wavelength inserted at the upper-left port IN, see Figure 1.1(b), is confined and travels through the upper waveguide until it gets slightly disturbed in its evanescent field by the presence of the resonator. Part of the light is trapped in the disk which subsequently causes interference in the upper and lower waveguide. The local interference can lead, depending on the wavelength, to two different global states. First, in the so-called 'in-resonance' case, complete destructive interference in the upper guide will transfer the light to the lower waveguide where it will exit in the lower-left port OUT 2. In the more generic case, 'put-of-resonance', the interference at the upper guide waveguide is only partially



(a) Resonator



(b) Schematic top-view of a micro-resonator.

Figure 1.1: Schematic view of a micro-resonator (figure provided by M. Hammer).

destructive. The light becomes only slightly disturbed and is largely transmitted to OUT 1.

The wavelength dependent behaviour is the essential property of this device. It makes it possible to filter out light of a specific wavelength. The actual performance of optical devices depends critically on the material properties and on the precise dimensions, where the distance between parts are in the order of fractions of micrometers. Accurate simulation tools are indispensable to design such devices, to find optimal parameter settings, and to determine critical tolerances for actual fabrication.

Another example where the numerical solution of the Maxwell equations is required is the study of electromagnetic wave propagation in caves and tunnels. It is of great practical interest to antenna engineers to understand the behaviour of electromagnetic waves in order to design reliable wireless communication systems in such rough environments. Current wireless radio frequency (RF)

communication systems are not designed to operate reliably in enclosed environments, such as cave-like structures, tunnels or subways. This prohibits the quick deployment of wireless systems in caves and tunnels. If the propagation properties of electromagnetic waves in a tunnel could be better characterized, then a more robust communication system could be designed specifically for operation in such environments. Thus, full electromagnetic wave simulations in this type of environment are very useful to achieve this.

These applications are examples which motivate the research presented in this thesis, which aims at developing new numerical tools for the solution of the three dimensional Maxwell equations and to improve existing methods in various ways.

1.2 Existing methods

To simulate complex electromagnetic wave problems the solution of the Maxwell equations is required. In most real-life applications the analytic solution, i.e. the solution expressed in terms of mathematical formulas, is not readily available. In such situations numerical methods are indispensable tools to solve the Maxwell equations approximately on computers. In this section we give a brief overview of the main approaches to solve the Maxwell equations numerically.

1.2.1 Finite difference time domain method

In the past decades much effort has been put into solving the Maxwell equations with various numerical methods. For a long period of time the famous finite-difference time-domain (FDTD) method, initiated by Yee [109], has been used for discretizing the Maxwell equations, both in space and time. The Yee scheme is fully explicit in its structure and is originally designed for Cartesian regular grids. For an analysis of the Yee and other FDTD schemes we refer to the book of Taflove [99]. For the transverse magnetic (TM) fields, where the components of the magnetic field $\boldsymbol{H}$ and the electric field $\boldsymbol{E}$ satisfy $H_z = E_x = E_y = 0$, the Maxwell equations reduce to

$$\begin{aligned} \partial_t H_x &= -\mu^{-1} \frac{\partial E_z}{\partial y}, \\ \partial_t H_y &= \mu^{-1} \frac{\partial E_z}{\partial x}, \\ \partial_t E_z &= \epsilon^{-1} \left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \right). \end{aligned} \tag{1.1}$$

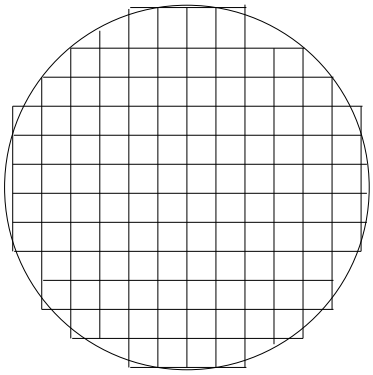


Figure 1.2: An example of a stair-step approximation of a circular cylinder used in many finite difference schemes. Dashed line: domain boundary.

The Yee scheme for (1.1), which uses a staggered mesh, reads:

$$\frac{H_x|_{i,j}^{n+1/2} - H_x|_{i,j}^{n-1/2}}{\Delta t} = -\mu^{-1}\frac{E_z|_{i,j+1/2}^{n} - E_z|_{i,j-1/2}^{n}}{\Delta y},$$
$$\frac{H_y|_{i,j}^{n+1/2} - H_y|_{i,j}^{n-1/2}}{\Delta t} = \mu^{-1}\frac{E_z|_{i+1/2,j}^{n} - E_z|_{i-1/2,j}^{n}}{\Delta x},$$
$$\frac{E_z|_{i,j}^{n+1} - E_z|_{i,j}^{n}}{\Delta t} = \epsilon^{-1}\left(\frac{H_y|_{i+1/2,j}^{n+1/2} - H_y|_{i-1/2,j}^{n+1/2}}{\Delta x} - \frac{H_x|_{i,j+1/2}^{n+1/2} - H_x|_{i,j-1/2}^{n+1/2}}{\Delta y}\right),$$

where $\Delta x, \Delta y$ are the mesh sizes and Δt the time step. The subindices indicate the position in the spatial grid and the superindices show the time level.

Although the Yee scheme is simple in its structure and easy for coding purposes, it has two main disadvantages. One is the lack of flexibility and accuracy in dealing with problems on domains with curved boundaries and inhomogeneous media. On curved boundaries, for instance, a common technique is to use so called stair-step approximations, see Figure 1.2. But in this case the computational grid has to be very fine to approximate the boundary accurately, which means that stair-step approximations are computationally very expensive [61]. The second drawback of the Yee scheme is the time integration method which is explicit. As an explicit scheme it requires restrictions on the time step due to the CFL (Courant-Friedrichs-Levy) condition to guarantee stability of the time

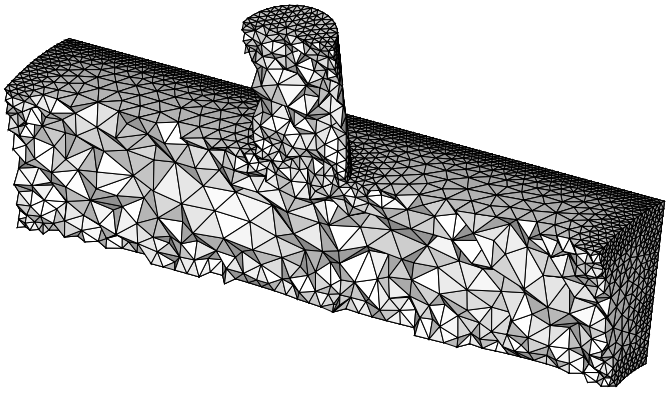


Figure 1.3: An example of tetrahedral finite element mesh of a tube that has a smaller inlet tube connected to the main one.

integration method. Many applications require locally fine meshes to capture important geometrical details or physical phenomena and the time step restriction then can be very severe. Hence many time integration steps are required, which makes the scheme computationally expensive. These two problems can be avoided, respectively, by using finite element methods and unconditionally stable time integration schemes.

1.2.2 Finite element methods

In domains with a complex geometry finite element methods (FEM) are one of the most common techniques for spatial discretizations of partial differential equations (PDE). For a reader unfamiliar with finite element methods we refer to [22, 31] and briefly mention that the main idea of the method is to divide the domain into many small subdomains (elements), see Figure 1.3, and define on each element a number of local basis functions. Then the unknown solution of the PDE, which is transformed into a weak formulation, is approximated by linear combinations of the basis functions on all elements. This method, combined with modern mesh generation techniques, allows scientists to solve partial differential equations on complex domains accurately and use them in mathematical modelling.

The usual Lagrange or node-based finite elements are, however, in many applications not appropriate to represent electromagnetic fields (see e.g. [18], Section 6.3 and [97]). For example, with node-based finite elements it is hard to satisfy the physical conditions at the interfaces between different materials. The reason is that in this case the electromagnetic field $\boldsymbol{E}$ (or $\boldsymbol{B}$) only has continuous tangential (normal) components and discontinuous normal (tangential) components, whereas node-based finite elements enforce full continuity. This generally results in a physically incorrect solution. Another important reason not to use node-based finite elements, is that they do not reflect the underlying geometrical structure of the electromagnetic field at the discrete level. In particular, they do not satisfy the discrete De Rham diagram [17, 57] .

In the last two decades a great deal of work has been done to overcome the problems arising from node-based elements. A major contribution has been made by J.-C. Nédélec [76, 77]. He designed new types of finite elements which describe the electromagnetic field in a better way as compared to existing methods. The Nédélec elements have many attractive properties (e.g. automatic satisfaction of the proper interface conditions between different materials and require less smoothness than standard Lagrangian elements) and are nowadays a common technique in computational electromagnetics.

More information about Nédélec elements (e.g. the construction of these elements, approximation and convergence properties) can be found in [57, 73].
A difficult question in practical situations is, however, the design of an accurate computational mesh for the finite element discretization. This requires knowledge about the error distribution and adaptive algorithms to construct a (nearly) optimal mesh and discretization.

1.3 Objectives

The topics discussed in Section 1.2 have motivated the research documented in this thesis and resulted in the following objectives:

- Construct an accurate and efficient, unconditionally stable time integration scheme for the Maxwell equations.
- Develop an accurate adaptation strategy for finite element discretizations of the Maxwell equations.
- Analyze the Hamiltonian structure of the Maxwell equations and provide a numerical scheme which is both globally and locally energy conservative.

1.3.1 Unconditionally stable schemes

The first objective focuses on the construction of accurate and efficient, unconditionally stable time integration schemes.

Besides the Yee FDTD scheme [109] mentioned in the previous section, there exist many other time stepping schemes for the Maxwell equations [48, 71, 68, 69, 35, 62, 63]. Often the time step in these schemes is restricted either due to stability restrictions or accuracy requirements. In practice, however, one would often like to have a time step free from stability restrictions since on nonuniform finite element meshes or in inhomogeneous media the stability restriction can be much more stringent than the wave resolution requirements. The need for better stability motivated the development of a number of unconditionally stable schemes which proved successful in the finite element framework [48, 71]. Stable time stepping schemes for the Maxwell equations have also been of importance in connection with finite difference spatial discretizations [68, 69, 35, 62, 63]. An unconditionally stable scheme proposed by Gautschi [47] has recently received attention in the literature for the solution of second order highly oscillatory ODE's [60, 59, 54]. In addition to being unconditionally stable, the scheme has excellent wave resolution properties. The time discretization error of this scheme is second order uniformly in the frequencies [59] and this allows to choose time steps larger than the smallest wave length.

In this thesis we show that, using Krylov subspace techniques, the Gautschi cosine scheme can be efficiently implemented for the three-dimensional Maxwell equations discretized in space by Nédélec edge elements. This yields a Gautschi-Krylov cosine scheme which proves to be very competitive, in terms of accuracy and CPU time, to other implicit time-stable schemes for the time integration of the Maxwell equations. Furthermore, the attractive properties of the new scheme are confirmed on several test cases and compared with existing methods.

To achieve high computational efficiency, it is crucial for the new Gautschi-Krylov scheme to properly choose the Krylov subspace dimension every time the action of the matrix function is computed. We propose a new simple strategy for controlling the Krylov subspace dimension which makes the Gautschi cosine scheme an efficient time-integration method for the Maxwell equations.

1.3.2 Adaptive methods

The necessity of adaptive methods arises from the solution of the Maxwell equations when the solution contains structures with limited regularity (by regularity we mean here smoothness of the solution), such as singularities near corners and

non-convex edges. In such situations a finer mesh is required to obtain an accurate solution. The solution of the Maxwell equations on a uniformly finer mesh, especially for three-dimensional applications, will in general require too much computational effort due to the large number of unknowns. These complicated structures can be efficiently modeled using *hp*-adaptive techniques, in which the mesh is locally refined and coarsened (*h*-adaptation) or the polynomial order in individual elements is adjusted (*p*-adaptation). In Figure 1.4 an approximately uniform mesh and its corresponding *h*-adapted mesh near the re-entrant corner in Fichera cube are shown. Examples of *hp*-adaptive techniques applied to the Maxwell equations can be found in e.g. [28, 83, 84]. In simple cases one can predict the regions where the mesh needs to be adapted, but a more general approach requires the use of a posteriori error estimates in which the local error is predicted based on the properties of the numerical solution. General techniques for a posteriori error estimation are discussed in e.g. [3, 43, 49], but providing accurate a posteriori error estimates for the Maxwell equations still poses many problems.

The most common a posteriori error estimators are residual based methods, where the behavior of the (local) error is evaluated based on the estimated (local) residual. For clarity let us consider a simple example of a residual based a posteriori error estimator applied to the Laplace equations. Let u be the solution of the Laplace equation on a domain Ω,

$$-\Delta u = f \quad \text{in } \Omega, \tag{1.2a}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{1.2b}$$

Let $\mathcal{T}$ be a tessellation of Ω with mesh size h and denote by u_h the finite element solution of (1.2) on this mesh. Then an explicit residual based a posteriori error estimator reads

$$\|u - u_h\|^2 + \|\nabla(u - u_h)\|^2 \leq C\mathbf{Q}(h, u_h, f), \tag{1.3}$$

where $\mathbf{Q}$ is a functional depending on the known data h, u_h, f. In many applications (e.g. on anisotropic meshes) the estimate (1.3) is not sharp enough due to large (or unknown) values of the constant C or an improper choice of the functional $\mathbf{Q}$. We encounter the same problems in residual based a posteriori error estimation methods applied to the Maxwell equations. In these methods the error bounds contain in general unknown coefficients, which also depend on the wavenumber in the equations, and frequently result in unsharp estimates.

In this thesis we suggest another approach: the idea of implicit a posteriori error estimates discussed for elliptic partial differential equations in [3] is still

perfectly applicable to the Maxwell equations when properly formulated. In the implicit error estimation approach a local problem is formulated for the error function on each element (or a group of elements) with properly defined approximate boundary conditions, which are solely based on the computed numerical solution (for a detailed description see Chapter 4). Then the local problems formulated for the error function are solved with a properly defined finite element method. This is the main difference with explicit a posteriori error estimators which only use the data provided by the numerical solution.

The success of the implicit a posteriori error estimation technique, however, strongly depends on the proper definition of the boundary conditions and the choice of the basis for the numerical solution of the local problems. We give special attention to cases where the analytic solution is non-smooth and also investigate the problem in computational domains with reentrant corners. In various test cases we verify the performance of the implicit error estimator on cubic elements and compare the results with existing methods.

Of course, cubic elements are not flexible enough for real life problems, because it is nearly impossible to accurately approximate complex domains with cubic elements. Therefore we have extended the implicit error estimation technique to tetrahedral elements, which are very flexible to accurately model complex domains. In Chapter 5 a proper finite element basis is given for the local problems on tetrahedral elements and an adaptive algorithm is developed. The method is tested on various complex domains with reentrant corners and the mesh generation procedure is performed with the Centaur [29] mesh generation package. One of the advantages of the Centaur mesh generator is that it creates adaptive meshes without hanging nodes. Meshes without hanging nodes are desirable for Nédélec type elements, otherwise these elements are not well defined. Another desirable property of this mesh generator is that it avoids the generation of elements with a large dihedral angle, which is important for accuracy requirements.

1.3.3 Energy conservative discretization of the Maxwell equations

Many dynamical systems, i.e. the Maxwell equations and the shallow-water equations, can be written in a Hamiltonian form. Most of them are energy conserving systems, which is directly linked to the Hamiltonian of these equations. The energy conserving properties imply that the change in the energy in a bounded domain is equal to the power supplied to the system through its boundary. In Chapter 6 we exploit the port-Hamiltonian structure of the

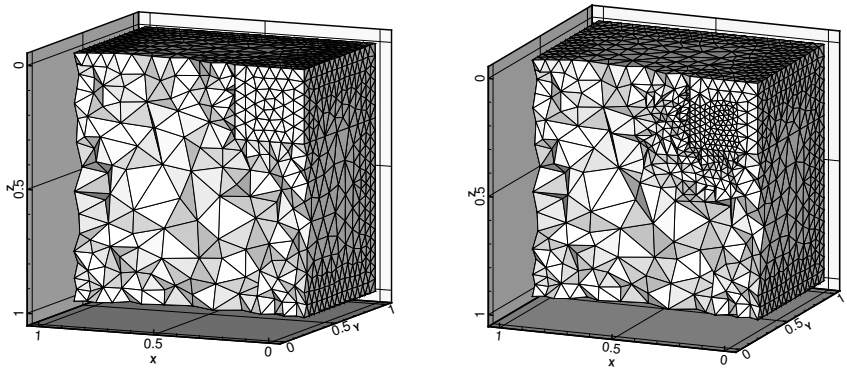


Figure 1.4: Meshes for Fichera cube domain. Left: A cut of the initial mesh. Right: A cut of an adapted mesh.

Maxwell equations and try to design energy conservative numerical algorithms which also provide the correct energy flow through the interfaces between neighboring elements. For this purpose we formulate the Maxwell equations in terms of the electric and magnetic fields and analyze a discretization of the Maxwell equations using Whitney 1-forms, which provide a locally and globally energy conservative scheme at the discrete level. For the time discretization we apply the well known leap-frog symplectic time integration scheme which preserves, in the absence of a source term, the discrete energy exactly. The method is tested for a simple example on meshes with tetrahedral elements.

1.4 Outline of this thesis

This thesis is the result of a four year Ph.D. project carried out at the University of Twente. The main results of this project are presented in the thesis. A short description of each chapter is as follows:

- A brief introduction to the Maxwell equations and some properties of electromagnetic waves are presented in Chapter 2.

- The Gautschi time integration scheme is presented in Chapter 3. A detailed description is given for the matrix-function evaluation which is an essential part of this scheme. The stability and dispersion properties of the Gautschi scheme are investigated in detail and its performance is compared with some existing methods.

- In Chapter 4 an implicit a posteriori error estimation technique is formulated. The well posedness of the local equations for the error function on an element or patch of elements is proven. We verify the method on cubic elements and demonstrate the superior performance of our method compared to some existing methods.

- In Chapter 5, based on the theoretical results from Chapter 4, an implicit a posteriori error estimation technique is applied to the Maxwell equations on meshes with tetrahedral elements. An adaptive algorithm is presented using the Centaur [29] mesh generation package. The performance of the adaptive algorithm is verified on various test cases including non-convex domains.

- In Chapter 6 we first discuss the general theory of the Stokes-Dirac structure and the Hamiltonian formulation for a general class of PDE's is given. As a particular example of these equations we consider the Maxwell equations. The geometrical structure of the Maxwell equations is analyzed at the discrete level by means of Whitney forms. We show that the discretization of the electromagnetic fields with appropriate discrete differential forms preserves many important properties of the physical system, in particular energy conservation.

- In the final chapter conclusions and a brief outlook of the thesis are given.

CHAPTER 2

The Maxwell equations

All the mathematical sciences are founded on relations between physical laws and laws of numbers, so that the aim of exact science is to reduce the problems of nature to the determination of quantities by operations with numbers.

James Clerk Maxwell
(1831-1879)

In 1873 J. C. Maxwell founded the modern theory of electromagnetism with the publication *Treatise on Electricity and Magnetism*, where the equations that now bear his name are formulated. These equations consist of two pairs of coupled partial differential equations describing six fields. Together with the material dependent constitutive relations and boundary conditions the Maxwell equations uniquely define the electromagnetic fields. In practice, there are many types of boundary conditions which are typical for electromagnetic problems, i.e. radiation conditions and perfectly conducting boundary conditions. In this thesis we consider electromagnetic fields in a linear medium. If we consider the

propagation of an electromagnetic field with a single frequency then the Maxwell equations are reduced to their time-harmonic form.

2.1 The Maxwell equations

The macroscopic electromagnetic fields are related by the following Maxwell equations:

$$\partial_t \boldsymbol{D}_s = \nabla \times \boldsymbol{H}_s - \boldsymbol{J}_s, \qquad \text{(Ampère's law)} \tag{2.1a}$$

$$\partial_t \boldsymbol{B}_s = -\nabla \times \boldsymbol{E}_s, \qquad \text{(Faraday's law)} \tag{2.1b}$$

$$\nabla \cdot \boldsymbol{D}_s = \rho_s, \qquad \text{(Gauss's law)} \tag{2.1c}$$

$$\nabla \cdot \boldsymbol{B}_s = 0, \qquad \text{(Gauss's law - magnetic)} \tag{2.1d}$$

where $\boldsymbol{E}_s = (E_x, E_y, E_z)$ and $\boldsymbol{H}_s = (H_x, H_y, H_z)$ ($\boldsymbol{D}_s = (D_x, D_y, D_z)$ and $\boldsymbol{B}_s = (B_x, B_y, B_z)$) are the electric and magnetic fields (respectively, the electric and the magnetic flux densities). Furthermore, $\boldsymbol{J}_s$ and ρ_s denote respectively the electric current and charge density (the latter is a space and time dependent function). The following constitutive relations hold for linear media:

$$\boldsymbol{D}_s = \varepsilon \boldsymbol{E}_s, \quad \boldsymbol{B}_s = \mu \boldsymbol{H}_s, \tag{2.2}$$

$$\boldsymbol{J}_s = \sigma \boldsymbol{E}_s + \boldsymbol{J}_s^{im}, \qquad \text{(Ohm's law)} \tag{2.3}$$

where the dielectric permittivity ε ($=\varepsilon_0 \varepsilon_r$), the conductivity σ, and the magnetic permeability μ ($=\mu_0 \mu_r$) are assumed to be space dependent positive definite tensors. The imposed current density is denoted by $\boldsymbol{J}_s^{im}$. The free space dielectric permittivity and magnetic permeability are defined by ε_0 and μ_0, respectively, and are given by

$$\varepsilon_0 \approx 8.854 \times 10^{-12}\,\mathrm{F/m}\,\text{(Farad per meter)}, \tag{2.4a}$$

$$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{H/m}\,\text{(Henry per meter)}. \tag{2.4b}$$

The dimensionless tensors ε_r and μ_r are material dependent and called relative permittivity and relative permeability, respectively. The conductivity σ is measured in Siemens/meter (S/m) units. The subscript s indicates that the SI units (International System of Units, from the French word *Système International d´Unités*) are used, see Table 2.1.

2.1.1 Material properties

The constitutive parameters ε and μ given in (2.2) define the relations between the electromagnetic fields and are material dependent. For linear and homogeneous media, a more general form for the first equation in (2.2) can be written

Table 2.1: SI units for electromagnetic quantities.

Quantity	Units	Name
Electric field $\boldsymbol{E}_s$	V/m	Volt per meter
Electric flux density $\boldsymbol{D}_s$	C/m^2	Coulomb per square meter
Magnetic field $\boldsymbol{H}_s$	A/m	Ampere per meter
Magnetic flux density $\boldsymbol{B}_s$	T	Tesla
Electric current density $\boldsymbol{J}_s$	A/m^2	Ampere per square meter
Electric charge density ρ_s	C/m^3	Coulomb per cubic meter

as

$$\boldsymbol{D}_s = (\varepsilon_0(1+\chi_e))\boldsymbol{E}_s = \varepsilon_0\varepsilon_r\boldsymbol{E}_s = \varepsilon\boldsymbol{E}_s, \tag{2.5}$$

where χ_e is a dimensionless quantity called electric susceptibility. A medium is called *linear* if χ_e is independent of $\boldsymbol{E}$ and *homogeneous* if χ_e is independent of the space coordinates.
Below we describe several cases.

1. ***Vacuum or free space:*** The relative permittivity ε_r and relative permeability μ_r in free space are one, i.e. $\varepsilon_r = \mu_r = 1$. Then the constitutive relations (2.2) reduce to

   $$\boldsymbol{D}_s = \varepsilon_0\boldsymbol{E}_s, \quad \boldsymbol{B}_s = \mu_0\boldsymbol{H}_s,$$

   where ε_0 and μ_0 are given according to (2.4).

   In free space the conductivity vanishes, $\sigma = 0$. The speed of light in free space, denoted by c_0, is given by $c_0 = \frac{1}{\sqrt{\varepsilon_0\mu_0}}\, m/s$.

2. ***Inhomogeneous, isotropic materials:*** The most commonly occurring case in practice is that various different materials occupy the domain of the electromagnetic field. In this case the domain is called inhomogeneous. If the material properties ε_r and μ_r do not depend on the direction of the electromagnetic fields and the material is linear then the constitutive relations are given by (2.2), where the material properties ε_r and μ_r are positive, bounded, scalar functions of position.

3. ***Inhomogeneous, anisotropic materials:*** In some applications the dielectric permittivity ε and magnetic permeability μ are different for different directions of the electromagnetic fields. For instance, the vector fields $\boldsymbol{D}_s$ and $\boldsymbol{E}_s$ generally will have different directions. In this case the domains are called *anisotropic* and the material properties ε, μ, and σ are 3×3 positive-definite matrix functions of the space coordinates.

Material with zero conductivity is called *dielectric*, and with positive conductivity ($\sigma > 0$) is called a *conductor*. However, in this thesis we will consider only dielectric materials. A detailed description of the electromagnetic theory can be found in [30].

For some materials values of relative permittivities, relative permeabilities and conductivities are given in Table 2.2(a), Table 2.2(b), and Table 2.2(c), respectively.

2.1.2 Dimensionless Maxwell equations

To avoid problems with floating point arithmetic when working with very large numbers, we apply the following space and time scaling:

$$x = \frac{x_s}{L}, \quad t = \frac{c_0}{L} t_s, \tag{2.6}$$

where L is a reference length (expressed in meters), and $c_0 = (\varepsilon_0 \mu_0)^{-1/2} \approx 3 \cdot 10^8$ m/s is the speed of light in vacuum. The scaling for y_s and z_s is done similar to x_s. Furthermore, we normalize the fields as

$$\begin{aligned}
\boldsymbol{E}_s(\mathbf{x}_s, t_s) &= \frac{\tilde{H}_0}{Z_0^{-1}} \boldsymbol{E}(\mathbf{x}, t), & \boldsymbol{H}_s(\mathbf{x}_s, t_s) &= \tilde{H}_0 \boldsymbol{H}(\mathbf{x}, t), \\
\boldsymbol{D}_s(\mathbf{x}_s, t_s) &= \varepsilon_0 \frac{\tilde{H}_0}{Z_0^{-1}} \boldsymbol{D}(\mathbf{x}, t), & \boldsymbol{B}_s(\mathbf{x}_s, t_s) &= \mu_0 \tilde{H}_0 \boldsymbol{B}(\mathbf{x}, t), \\
\boldsymbol{J}_s(\mathbf{x}_s, t_s) &= \frac{\tilde{H}_0}{L} \boldsymbol{J}(\mathbf{x}, t), & \rho_s &= \frac{\tilde{H}_0}{Z_0^{-1}} \rho,
\end{aligned}$$

where $\mathbf{x}_s = (x_s, y_s, z_s)$, $\mathbf{x} = (x, y, z)$, $Z_0 = \sqrt{\mu_0/\varepsilon_0}$ [Ohm] is the free space intrinsic impedance, and $\tilde{H}_0$ is reference magnetic field strength [A/m]. The constitutive relations (2.2) in dimensionless form read:

$$\boldsymbol{D} = \varepsilon_r \boldsymbol{E}, \quad \boldsymbol{B} = \mu_r \boldsymbol{H}. \tag{2.7}$$

The Maxwell equations (2.1) written for scaled quantities yield the following dimensionless form:

$$\partial_t \boldsymbol{D} = \nabla \times \boldsymbol{H} - \boldsymbol{J}, \tag{2.8a}$$

$$\partial_t \boldsymbol{B} = -\nabla \times \boldsymbol{E}, \tag{2.8b}$$

$$\nabla \cdot \boldsymbol{D} = \rho, \tag{2.8c}$$

$$\nabla \cdot \boldsymbol{B} = 0. \tag{2.8d}$$

Table 2.2: Material properties for some materials.

(a) Relative permittivities ε_r.

Material	ε_r	Material	ε_r
Air	1.0	Polyethylene	2.3
Bakelite	5.0	Polystyrene	2.6
Glass	4–10	Porcelain	5.7
Mica	6.0	Rubber	2.3–4.0
Oil	2.3	Soil (dry)	3–4
Paper	2–4	Teflon	2.1
Parafin wax	2.2	Water (distilled)	80
Plexiglass	3.4	Seawater (distilled)	72

(b) Relative permeabilities μ_r.

Material	μ_r	Material	μ_r
Ferromagnetic		*Paramagnetic*	
Nickel	250	Aluminum	1.000021
Cobalt	600	Magnesium	1.000012
Iron (pure)	4.000	Palladium	1.00082
Mumetal	100.000	Titanium	1.00018
Diamagnetic			
Bismuth	0.99983		
Gold	0.99996		
Silver	0.99998		
Copper	0.99999		

(c) Conductivities σ.

Material	σ (S/m)	Material	σ (S/m)
Silver	6.17×10^7	Fresh water	10^{-3}
Copper	5.80×10^7	Distilled water	2×10^{-4}
Gold	4.10×10^7	Dry soil	10^{-5}
Aluminum	3.54×10^7	Transformer oil	10^{-11}
Brass	1.57×10^7	Glass	10^{-12}
Bronze	10^7	Porcelain	2×10^{-13}
Iron	10^7	Rubber	10^{-15}
Seawater	4	Fused quartz	10^{-17}

In the rest of this thesis we consider the Maxwell equations in dimensionless form, unless indicated otherwise.

The divergence conditions (2.8c) and (2.8d) are direct consequences of the fundamental Maxwell equations (2.8a) and (2.8b) provided that the charge conservation law holds:

$$\nabla \cdot \boldsymbol{J} + \frac{\partial \rho}{\partial t} = 0. \tag{2.9}$$

Indeed, if we take the divergence of (2.8a) and (2.8b) and using the relation $\nabla \cdot (\nabla \times \mathbf{F}) = 0$ for a sufficiently smooth vector field $\mathbf{F}$, we obtain

$$\frac{\partial}{\partial t} \nabla \cdot \boldsymbol{B} = \frac{\partial}{\partial t} (\nabla \cdot \boldsymbol{D} - \rho) = 0.$$

Thus if (2.8c) and (2.8d) hold at initial time, they also hold at later time.

2.1.3 Interface conditions

To investigate the interface conditions of electromagnetic fields across two different materials let us first consider the integral form of the Maxwell equations (2.8). We take the surface integral of (2.8a)–(2.8b) over both sides of an open surface S with boundary contour C separating two media and apply the Stokes theorem:

$$\oint_C \boldsymbol{H} \cdot dl = \int_S (\boldsymbol{J} + \partial_t \boldsymbol{D}) \cdot ds, \tag{2.10a}$$

$$\oint_C \boldsymbol{E} \cdot dl = -\int_S \partial_t \boldsymbol{B} \cdot ds. \tag{2.10b}$$

Similarly, we take the volume integral of (2.8c)–(2.8d) over a volume V with a closed surface S and, using the divergence theorem, obtain

$$\oint_S \boldsymbol{D} \cdot ds = \int_V \rho \, dv, \tag{2.10c}$$

$$\oint_S \boldsymbol{B} \cdot ds = 0. \tag{2.10d}$$

Consider the diagram shown in Figure 2.1 for (2.10b) in the limiting case $\Delta h \to 0$. In this case the area of the parallelepiped *abcd* is infinitely small. Because the magnetic flux density $\boldsymbol{B}$ is finite, the right hand side of (2.10b) tends to zero, and thus

$$\oint_C \boldsymbol{E} \cdot dl = \int_a^b \boldsymbol{E}_1 \cdot dl + \int_c^d \boldsymbol{E}_2 \cdot dl = 0, \tag{2.11}$$

where the indices 1 and 2 refer to the fields in the different domains. It immediately follows that $\boldsymbol{E}_1 \cdot \vec{\boldsymbol{ab}} + \boldsymbol{E}_2 \cdot (-\vec{\boldsymbol{ab}}) = 0$. If we denote the unit vector along $\vec{\boldsymbol{ab}}$ by $\boldsymbol{t}_{ab}$, then $\vec{\boldsymbol{ab}} = \boldsymbol{t}_{ab}\Delta w$ and we obtain

$$\boldsymbol{E}_1 \cdot \boldsymbol{t}_{ab} - \boldsymbol{E}_2 \cdot \boldsymbol{t}_{ab} = 0. \tag{2.12}$$

Let us denote by $\boldsymbol{n}$ the normal vector at the interface pointing from Medium 1 into Medium 2. Then there is a vector $\boldsymbol{t}$ from the tangent space of S such that $\boldsymbol{t}_{ab} = \boldsymbol{n} \times \boldsymbol{t}$. Substituting the last relation into (2.12), we obtain

$$(\boldsymbol{n} \times \boldsymbol{E}_1 - \boldsymbol{n} \times \boldsymbol{E}_2) \cdot \boldsymbol{t} = 0. \tag{2.13}$$

Because $\vec{\boldsymbol{ab}}$ is arbitrary, hence $\boldsymbol{t}$ is arbitrary too, we obtain

$$\boldsymbol{n} \times \boldsymbol{E}_1 - \boldsymbol{n} \times \boldsymbol{E}_2 = 0, \tag{2.14}$$

i.e. the tangential component of the electric field is continuous across an interface.

Similarly, we can show that in conducting materials with a finite conductivity coefficient or in dielectric materials the right hand side of (2.10a) is finite, hence it vanishes on an infinitely small parallelepiped *abcd*. Then we obtain

$$\boldsymbol{n} \times \boldsymbol{H}_1 - \boldsymbol{n} \times \boldsymbol{H}_2 = 0, \tag{2.15}$$

i.e. the tangential component of the magnetic field $\boldsymbol{H}$ is continuous across an interface. For the special case of an idealized perfect conductor, where the conductivity $\sigma \to \infty$, a surface current may exist, so that the first surface integral on the right hand side of (2.10a) does not vanish on the infinitely small parallelepiped *abcd*. This means that a surface current $\boldsymbol{j}^s$ can exist on the boundary, normal to the area *abcd*. We conclude that for the case where a surface current may exist, the interface condition for the magnetic field $\boldsymbol{H}$ therefore is

$$\mathbf{n} \times (\boldsymbol{H}_1 - \boldsymbol{H}_2) = \boldsymbol{j}^s,$$

where $\boldsymbol{j}^s$ is the surface current. In most applications the surface current vanishes (i.e. $\boldsymbol{j}^s = 0$).

The interface conditions for the other electromagnetic fields can be obtained in a similar manner.

We summarize the above results and obtain the following interface conditions for the electromagnetic fields on the interface between two different media

$$\mathbf{n} \times (\boldsymbol{E}_1 - \boldsymbol{E}_2) = 0, \tag{2.16a}$$

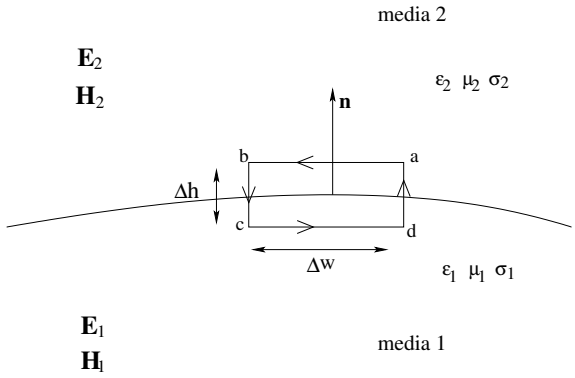


Figure 2.1: Interface between two different media.

$$\mathbf{n} \cdot (\boldsymbol{B}_1 - \boldsymbol{B}_2) = 0, \tag{2.16b}$$
$$\mathbf{n} \times (\boldsymbol{H}_1 - \boldsymbol{H}_2) = \boldsymbol{j}^s, \tag{2.16c}$$
$$\mathbf{n} \cdot (\boldsymbol{D}_1 - \boldsymbol{D}_2) = \rho_S, \tag{2.16d}$$

where ρ_S is the surface charge density. The subscript i, $i = 1, 2$, refers to the restriction of a vector field to medium i. The interface conditions (2.16) in the presence of material discontinuities should be satisfied also in the discrete case. In Chapter 6 we will discuss how to discretize the Maxwell equations in order to preserve these interface conditions.

2.1.4 Second order PDE for the electric field

The Maxwell equations (2.8) can be formulated for the $\boldsymbol{E}$ or $\boldsymbol{D}$ and $\boldsymbol{B}$ or $\boldsymbol{H}$ fields using the constitutive relations (2.7). For example, elimination of the electric flux density $\boldsymbol{D}$ and the magnetic field $\boldsymbol{H}$, yields

$$\partial_t(\varepsilon_r \boldsymbol{E}) = \nabla \times (\mu_r^{-1} \boldsymbol{B}) - \boldsymbol{J}, \tag{2.17}$$
$$\partial_t \boldsymbol{B} = -\nabla \times \boldsymbol{E}. \tag{2.18}$$

A possible drawback of this approach is that one has to work with both fields and construct appropriate finite element approximations for each of the fields. By differentiating (2.17) in time and taking the curl of (2.18), we can eliminate $\boldsymbol{B}$ from the system (2.17)–(2.18) and obtain a second-order hyperbolic partial differential equation for the electric field $\boldsymbol{E}$

$$\partial_{tt}(\varepsilon_r \boldsymbol{E}) + \nabla \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) = -\partial_t \boldsymbol{J}. \tag{2.19}$$

Of course, the choice of keeping $\boldsymbol{E}$ is arbitrary. One can formulate a second order equation for the other fields, too.

2.1.5 Time-harmonic Maxwell equations

When the field quantities in the Maxwell equations are harmonically oscillating functions in time with a single frequency, then the Maxwell equations (2.8) can be reduced to the time-harmonic Maxwell system. In this case we can write the electromagnetic fields in the following form

$$\boldsymbol{E}(x,y,z,t) = \Re(\hat{\boldsymbol{E}}(x,y,z)e^{-i\omega t}), \tag{2.20a}$$
$$\boldsymbol{D}(x,y,z,t) = \Re(\hat{\boldsymbol{D}}(x,y,z)e^{-i\omega t}), \tag{2.20b}$$
$$\boldsymbol{B}(x,y,z,t) = \Re(\hat{\boldsymbol{B}}(x,y,z)e^{-i\omega t}), \tag{2.20c}$$
$$\boldsymbol{H}(x,y,z,t) = \Re(\hat{\boldsymbol{H}}(x,y,z)e^{-i\omega t}), \tag{2.20d}$$

where $i = \sqrt{-1}$ and $\Re(\cdot)$ denotes the real part of the expression in parentheses. The vector phasor $\hat{\boldsymbol{E}}(x,y,z)$ (and similarly the other vector phasors) is a vector field of position, but not of time. It contains information on the direction, magnitude, and phase of the corresponding electromagnetic field. Phasors are in general complex-valued vector fields. Some authors consider the time dependence for time-harmonic waves in the form $e^{i\omega t}$. Of course, this choice is arbitrary and, provided it is used consistently, gives no difficulties. The choice made in (2.20) is standard in the mathematical literature.

It is clear that we should also consider now the current density and charge density as time harmonic, hence we assume

$$\boldsymbol{J}(x,y,z,t) = \Re(\hat{\boldsymbol{J}}(x,y,z)e^{-i\omega t}), \tag{2.20e}$$
$$\rho(x,y,z,t) = \Re(\hat{\rho}(x,y,z)e^{-i\omega t}). \tag{2.20f}$$

After substitution of (2.20) in the Maxwell equations (2.8) we obtain the time-harmonic Maxwell equations, i.e.

$$-i\omega\hat{\boldsymbol{D}} = \nabla\times\hat{\boldsymbol{H}} - \hat{\boldsymbol{J}}, \tag{2.21a}$$
$$-i\omega\hat{\boldsymbol{B}} = -\nabla\times\hat{\boldsymbol{E}}, \tag{2.21b}$$
$$\nabla\cdot\hat{\boldsymbol{D}} = \hat{\boldsymbol{\rho}}, \tag{2.21c}$$
$$\nabla\cdot\hat{\boldsymbol{B}} = 0, \tag{2.21d}$$

where the time-harmonic charge density $\hat{\boldsymbol{\rho}}$ is given via the charge conservation equation (2.9). Similar to the time-dependent case, we can write the Maxwell

equations now as a second-order wave equation

$$\nabla \times (\mu_r^{-1} \nabla \times \hat{\boldsymbol{E}}) - \omega^2 \varepsilon_r \hat{\boldsymbol{E}} = i\omega \hat{\boldsymbol{J}}. \tag{2.22}$$

2.1.6 Boundary conditions

The Maxwell equations described in the previous sections uniquely define the electromagnetic fields in a finite domain if proper boundary conditions are imposed at the boundary of the domain. The boundary conditions usually depend on the specific application. In this thesis for the verification of the numerical results we usually consider perfectly conducting boundary conditions. This important case occurs when a dielectric material is inside of a perfect conductor. Since the electric field $\boldsymbol{E}$ vanishes on a perfect conductor, we obtain the perfectly conducting boundary condition from equation (2.16a)

$$\boldsymbol{n} \times \boldsymbol{E} = 0 \quad \text{on} \quad \Gamma, \tag{2.23}$$

where $\Gamma = \partial\Omega$ is the boundary of the dielectric domain Ω.

More information about different types of boundary conditions can be found in e.g. [73].

CHAPTER 3

The Gautschi time stepping scheme for edge finite element discretizations of the Maxwell equations

For the time integration of edge finite element discretizations of the three-dimensional Maxwell equations, we consider the Gautschi cosine scheme where the action of the matrix function is approximated by a Krylov subspace method. First, for the space-discretized edge finite element Maxwell equations, the dispersion error of this scheme is analyzed in detail and compared to that of two conventional schemes. Second, we show that the scheme can be implemented in such a way that a higher accuracy can be achieved within less computational time (as compared to other implicit schemes). We also analyzed the error made in the Krylov subspace matrix function evaluations. Although the new scheme is unconditionally stable, it is explicit in structure: as an explicit scheme, it requires only the solution of linear systems with the mass matrix.

3.1 Introduction

This chapter deals with the numerical solution of the time dependent Maxwell equations. In particular, we are interested in time integration of the three-dimensional Maxwell equations discretized in space by Nédélec 's edge finite elements [76, 77]. Nédélec 's edge and face elements have a number of attractive properties (as e.g. automatic satisfaction of the proper continuity requirements across the boundaries between different materials) and are a standard tool in the numerical treatment of the Maxwell equations [73]. We emphasize, however, that the time integration techniques presented in this chapter are applicable to

any space-discretized second order wave equation(s).

Many time stepping schemes exist for the time integration of the space-discretized Maxwell equations [35, 48, 62, 63, 68, 69, 71, 109]. Often the time step in these schemes is restricted either due to stability restrictions or accuracy requirements, e.g. to resolve the waves. In practice, however, one often would like to have a step size free from stability restrictions since on nonuniform finite element meshes or in inhomogeneous media this restriction can be much more stringent than the wave resolution requirements. The need for better stability motivated the creation of a number of unconditionally stable schemes which proved successful in the finite element framework [48, 71]. Stable time stepping schemes for the Maxwell equations have been also of importance in connection with finite difference spatial discretizations [35, 62, 63, 68, 69]. A scheme proposed by Gautschi [47] has recently received attention in the literature for the solution of second order highly oscillatory ODE's [54, 59, 60]. This scheme contains a matrix function, is exact for linear equations with constant inhomogeneity and thus unconditionally stable. In each time step the product of a matrix function with a given vector can be computed by Krylov subspace methods [40, 41, 42, 58, 60, 67, 91, 100, 102]. The time error of the scheme is of second order uniformly in the frequencies [59] and this allows to choose time steps larger than the smallest wave length.

In this chapter we show that, using Krylov subspace techniques, the Gautschi cosine scheme can be efficiently implemented for the three-dimensional Maxwell equations discretized in space by edge elements. This yields a Gautschi-Krylov cosine scheme which proves to be very competitive, in terms of accuracy and CPU time, as compared to other implicit time-stable schemes for the time integration of the Maxwell equations.

Several authors study the dispersion properties of the discretized Maxwell equations. For the two-dimensional Maxwell equations discretized with the first order edge finite elements, Monk and Parrot compare dispersion properties of several conventional schemes [75]. A thorough analysis for three-dimensional problems with different boundary conditions on an unstructured tetrahedral meshes is carried out in [74]. For the dispersive properties of the higher order edge elements we refer to the paper of Ainsworth [2]. Dispersion properties of several high order time integration schemes for transient wave equations are considered by Cohen in [32]. In this chapter the attractive properties of the new scheme are confirmed by a dispersion analysis done for the first order edge finite elements. For comparison purposes, the dispersion analysis is also presented for two other schemes, the conventional time-explicit leap-frog scheme and an uncondition-

ally stable scheme of Lee, Lee and Cangellaris often referred to as the Newmark β-scheme (in the sequel, the Newmark scheme) [48, 71].

To achieve high computational efficiency, it is crucial for the new Gautschi-Krylov scheme to properly choose the Krylov subspace dimension every time the action of the matrix function is computed. We propose a new simple strategy for controlling the Krylov subspace dimension.

This chapter is organized as follows: Section 3.2 presents the Maxwell equations and their weak formulation, in Section 3.3 the Gautschi cosine scheme and two other time stepping schemes are described, the Krylov subspace error in the Gautschi-Krylov scheme and the stability of the scheme are analyzed in Section 3.4, and dispersion errors of the three schemes are investigated in Section 3.5. Finally, in the last section we demonstrate numerical results of a comparison of the schemes.

3.2 Maxwell equations

Consider the time-dependent Maxwell equations on a bounded lossless domain $\Omega \subset \mathbb{R}^3$:

$$\partial_t \boldsymbol{D}_s = \nabla \times \boldsymbol{H}_s - \boldsymbol{J}_s, \tag{3.1}$$

$$\partial_t \boldsymbol{B}_s = -\nabla \times \boldsymbol{E}_s, \tag{3.2}$$

$$\nabla \cdot \boldsymbol{D}_s = \rho_s, \tag{3.3}$$

$$\nabla \cdot \boldsymbol{B}_s = 0, \tag{3.4}$$

where $\boldsymbol{E}_s$ and $\boldsymbol{H}_s$ ($\boldsymbol{D}_s$ and $\boldsymbol{B}_s$) are electric and magnetic fields (respectively, the electric and the magnetic flux densities). Furthermore, $\boldsymbol{J}_s$ and ρ_s denote respectively the electric current and charge density (the latter is a space and time dependent function). The subscript s indicates that the SI units are used. Assume that the following boundary and initial conditions are given:

$$(\boldsymbol{n} \times \boldsymbol{E}_s)|_\Gamma = 0, \tag{3.5}$$

$$\boldsymbol{E}_s|_{t_s=0} = \bar{\boldsymbol{E}}_0, \qquad \boldsymbol{H}_s|_{t_s=0} = \bar{\boldsymbol{H}}_0, \tag{3.6}$$

where $\boldsymbol{n}$ is the outward normal vector to the domain boundary $\Gamma = \partial\Omega$. The following constitutive relations hold:

$$\boldsymbol{D}_s = \epsilon \boldsymbol{E}_s, \quad \boldsymbol{B}_s = \mu \boldsymbol{H}_s, \tag{3.7}$$

where the material properties the dielectric permittivity $\epsilon\,(=\epsilon_0\epsilon_r)$ and the magnetic permeability $\mu(=\mu_0\mu_r)$ are assumed to be space dependent tensors. The

free space dielectric permittivity and magnetic permeability are defined by ϵ_0 and μ_0, respectively. The dimensionless tensors ϵ_r and μ_r are material dependent and called relative permittivity and relative permeability, respectively.

3.2.1 Dimensionless Maxwell equations

To avoid problems with floating point arithmetic when working with very large numbers, we apply the following space and time scaling:

$$x = \frac{x_s}{L}, \quad t = \frac{c_0}{L} t_s, \tag{3.8}$$

where L is a reference length (expressed in meters), and $c_0 = (\epsilon_0 \mu_0)^{-1/2} \approx 3 \cdot 10^8$ m/s is the speed of light in vacuum. The scaling for y_s and z_s is done similarly to x_s. Furthermore, we normalize the fields as

$$\boldsymbol{E}_s(\mathbf{x}_s, t_s) = \frac{\tilde{H}_0}{Z_0^{-1}} \boldsymbol{E}(\mathbf{x}, t), \quad \boldsymbol{H}_s(\mathbf{x}_s, t_s) = \tilde{H}_0 \boldsymbol{H}(\boldsymbol{x}, t), \quad \boldsymbol{J}_s(\mathbf{x}_s, t_s) = \frac{\tilde{H}_0}{L} \boldsymbol{J}(\mathbf{x}, t), \tag{3.9}$$

where $\mathbf{x}_s = (x_s, y_s, z_s)$, $\mathbf{x} = (x, y, z)$, $Z_0 = \sqrt{\mu_0/\epsilon_0}$ [Ohm] is the free space intrinsic impedance, and $\tilde{H}_0$ is a reference magnetic field strength [A/m]. Equations (3.1),(3.2) and constitutive relations (3.7) written for the scaled quantities yield the following dimensionless Maxwell equations:

$$\epsilon_r \partial_t \boldsymbol{E} = \nabla \times \boldsymbol{H} - \boldsymbol{J}, \tag{3.10a}$$

$$\mu_r \partial_t \boldsymbol{H} = -\nabla \times \boldsymbol{E}. \tag{3.10b}$$

Since the given boundary conditions are homogeneous, the dimensionless normalization leaves them unchanged.

By differentiating (3.10a) in time and taking curl of (3.10b), we eliminate $\boldsymbol{H}$ from the system (3.10) and obtain a second-order hyperbolic partial differential equation for $\boldsymbol{E}$

$$\epsilon_r \partial_{tt} \boldsymbol{E} + \nabla \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) = -\partial_t \boldsymbol{J}. \tag{3.11}$$

Using (3.10a) we obtain the initial condition for the derivative of $\boldsymbol{E}$:

$$\partial_t \boldsymbol{E}(x, 0) = \epsilon_r^{-1}(-\boldsymbol{J}(\mathbf{x}, 0) + \nabla \times \boldsymbol{H}(\mathbf{x}, 0)). \tag{3.12}$$

3.2.2 Weak formulation and finite element discretization

Defining the space

$$H_0(\mathrm{curl}, \Omega) = \{\boldsymbol{u} \in L_2(\Omega)^3 |\ \nabla \times \boldsymbol{u} \in L_2(\Omega)^3,\ (\boldsymbol{n} \times \boldsymbol{u})|_\Gamma = 0\},$$

we arrive at the following Galerkin weak formulation of (3.11):

Find $\boldsymbol{E} \in H_0(\mathrm{curl}, \Omega)$ such that $\forall\, \boldsymbol{w} \in H_0(\mathrm{curl}, \Omega)$

$$\partial_{tt}(\epsilon_r \boldsymbol{E}, \boldsymbol{w}) + (\mu_r^{-1} \nabla \times \boldsymbol{E}, \nabla \times \boldsymbol{w}) = -(\partial_t \boldsymbol{J}, \boldsymbol{w}). \tag{3.13}$$

Next, we introduce a tessellation of Ω (a hexahedral or tetrahedral mesh) with N internal edges and denote by W_h the space of Nédélec 's first order edge basis functions:

$$W_h = \mathrm{span}\left\{ \boldsymbol{w}_j(x) \mid \text{all internal edges } j = 1, \ldots, N \right\},$$

where each basis function $\boldsymbol{w}_j(x)$ is defined with respect to the edge j as a linear polynomial such that [73, 76]:

$$\alpha_i(\boldsymbol{w}_j) \equiv \int_{\text{edge } i} \boldsymbol{w}_j \cdot \boldsymbol{t}_i \, da = \begin{cases} 0, & \text{if} \quad i \neq j, \\ 1, & \text{if} \quad i = j, \end{cases}$$

where $\alpha_i(\boldsymbol{w}_j)$ are the degrees of freedom associated with the edges and $\boldsymbol{t}_i$ is the unit tangent vector along the edge i. The electric field $\boldsymbol{E}$ is then approximated as

$$\boldsymbol{E} \approx \boldsymbol{E}_h = \sum_{j=1}^{N} e_j(t) \boldsymbol{w}_j.$$

The discretized version of (3.13) then reads:

Find $\boldsymbol{E}_h \in W_h$, such that $\forall\, \boldsymbol{W} \in W_h$

$$\partial_{tt}(\epsilon_r \boldsymbol{E}_h, \boldsymbol{W}) + (\mu_r^{-1} \nabla \times \boldsymbol{E}_h, \nabla \times \boldsymbol{W}) = -(\partial_t \boldsymbol{J}, \boldsymbol{W}). \tag{3.14}$$

Denoting by $\boldsymbol{e}(t)$ a vector function with the entries $e_j(t)$, we can write (3.14) in a matrix form as a system of ordinary differential equations (ODE's)

$$M_\epsilon \boldsymbol{e}'' + A_\mu \boldsymbol{e} = \boldsymbol{j}(t) \tag{3.15}$$

with

$$\begin{aligned} &(M_\epsilon)_{ij} = (\epsilon_r \boldsymbol{w}_i, \boldsymbol{w}_j), && (\boldsymbol{j}(t))_i = -(\partial_t \boldsymbol{J}, \boldsymbol{w}_i), \\ &(A_\mu)_{ij} = (\mu_r^{-1} \nabla \times \boldsymbol{w}_i, \nabla \times \boldsymbol{w}_j). \end{aligned} \tag{3.16}$$

3.3 Time stepping schemes

In this section the Gautschi cosine time-stepping scheme is presented, along with two other conventional time-stepping schemes which we use for comparison with the Gautschi scheme. The first of the two schemes is the explicit staggered leap frog scheme and the second one is an implicit scheme designed for finite element discretizations of the Maxwell equations [48, 71].

3.3.1 Leap frog scheme

The two-step staggered leap frog scheme for the semidiscrete Maxwell equations (3.15) reads

$$M_\epsilon \frac{\boldsymbol{e}^{n+1} - 2\boldsymbol{e}^n + \boldsymbol{e}^{n-1}}{\tau^2} + A_\mu \boldsymbol{e}^n = \boldsymbol{j}^n, \tag{3.17}$$

where τ is the time step size and the superscripts refer to the time levels $t_n = n\tau$. The scheme can be written in the form

$$M_\epsilon \boldsymbol{e}^{n+1} + (\tau^2 A_\mu - 2M_\epsilon)\boldsymbol{e}^n + M_\epsilon \boldsymbol{e}^{n-1} = \tau^2 \boldsymbol{j}^n. \tag{3.18}$$

If the matrices M_ϵ and A_μ are Hermitian, M_ϵ is positive definite and A_μ is positive semidefinite then the leap frog scheme is stable for

$$\tau^2 \leq \frac{4}{\lambda_{\max}},$$

where $\lambda_{\max}$ is the maximum eigenvalue of the matrix $M_\epsilon^{-1} A_\mu$ (see Appendix 3.8.1).

The computational work of the scheme per time step mainly consists of one matrix-vector multiplication with the matrix $M_\epsilon^{-1} A_\mu$. This can be efficiently done with the help of a sparse LU factorization of M_ϵ (see Remark 3.1 in Section 3.3.2).

3.3.2 Gautschi cosine scheme

Reduction of the semidiscrete Maxwell Equations to the normal form

We first transform the ODE system (3.15) into the form

$$\boldsymbol{y}'' + \tilde{A}_{\epsilon,\mu} \boldsymbol{y} = \boldsymbol{f}(t), \tag{3.19}$$

which we call the normal form. Computing a sparse LU factorization of M_ϵ (see Remark 3.1), we obtain

$$M_\epsilon = L_\epsilon U_\epsilon.$$

Note that if ϵ is a symmetric positive definite tensor then the matrix M_ϵ is symmetric positive definite, too, and we can take $U_\epsilon = L_\epsilon^T$ (Cholesky factorization).

It is easy to see that the semidiscrete Maxwell equations (3.15) can be transformed to the form (3.19) with $\tilde{A}_{\epsilon,\mu}$ and $\boldsymbol{y}$ defined in one of the following ways:

$$\tilde{A}_{\epsilon,\mu} = U_\epsilon^{-1} L_\epsilon^{-1} A_\mu, \qquad \boldsymbol{y} = \boldsymbol{e}, \qquad \boldsymbol{f} = U_\epsilon^{-1} L_\epsilon^{-1} \boldsymbol{j}, \tag{3.20}$$

$$\tilde{A}_{\epsilon,\mu} = L_\epsilon^{-1} A_\mu U_\epsilon^{-1}, \qquad \boldsymbol{y} = U_\epsilon \boldsymbol{e}, \qquad \boldsymbol{f} = L_\epsilon^{-1} \boldsymbol{j}, \tag{3.21}$$

$$\tilde{A}_{\epsilon,\mu} = A_\mu U_\epsilon^{-1} L_\epsilon^{-1}, \qquad \boldsymbol{y} = L_\epsilon U_\epsilon \boldsymbol{e}, \qquad \boldsymbol{f} = \boldsymbol{j}, \tag{3.22}$$

where the inverse matrices will normally never be computed explicitly (see Remark 3.1). Since we call (3.19) the normal form of (3.15), the transformations (3.20), (3.21), (3.22) can respectively be called the left, two-sided and right normalizations.

Remark 3.1. *For the used edge finite element discretization a sparse LU (or Cholesky) factorization of the mass matrix can usually be efficiently computed even on fine meshes (at least, if the mesh is not too distorted [90] which is a general requirement for edge finite elements). In practice, matrices L_ϵ^{-1} and U_ϵ^{-1} will usually not be computed explicitly. This would be expensive because the inverses will not be sparse in general. In fact, we will only need to compute the action of the matrices L_ϵ^{-1} and U_ϵ^{-1} on a given vector and this can be done by solving a linear system with L_ϵ or U_ϵ, as is usually done in preconditioning (see e.g. Chapter 13.1 in [103] or Chapter 3.1 in [13]).*

Note that the sparse LU factorization of the mass matrix is also required for explicit schemes. The factorization is performed only once for the complete time integration.

3.3.3 Formulation of Gautschi cosine scheme

Consider the variation of constant formula for the solution of (3.19):

$$\boldsymbol{y}(t+\tau) = \cos(\tau\tilde{A}_{\epsilon,\mu}^{1/2})\boldsymbol{y}(t) + \tilde{A}_{\epsilon,\mu}^{-1/2}\sin(\tau\tilde{A}_{\epsilon,\mu}^{1/2})\boldsymbol{y}'(t) + \int_0^\tau \tilde{A}_{\epsilon,\mu}^{-1/2}\sin((\tau-s)\tilde{A}_{\epsilon,\mu}^{1/2})\boldsymbol{f}(t+s)ds. \tag{3.23}$$

If $\boldsymbol{f} = \text{const}(t)$ then it follows from (3.23) that

$$\boldsymbol{y}(t+\tau) - 2\boldsymbol{y}(t) + \boldsymbol{y}(t-\tau) = \tau^2\psi(\tau^2\tilde{A}_{\epsilon,\mu})(-\tilde{A}_{\epsilon,\mu}\boldsymbol{y}(t) + \boldsymbol{f}), \tag{3.24}$$

where the function ψ is given by

$$\psi(x^2) = 2\frac{1-\cos x}{x^2} = 2\int_0^1 x^{-1}\sin((1-\theta)x)d\theta. \tag{3.25}$$

The Gautschi cosine time stepping scheme [47, 59] for ODE system (3.19) is based on relation (3.24):

$$\boldsymbol{y}^{n+1} - 2\boldsymbol{y}^n + \boldsymbol{y}^{n-1} = \tau^2\psi(\tau^2\tilde{A}_{\epsilon,\mu})(-\tilde{A}_{\epsilon,\mu}\boldsymbol{y}^n + \boldsymbol{f}^n). \tag{3.26}$$

For a complete derivation of the scheme we refer to [59].

Computation of $\psi(\tau^2\tilde{A}_{\epsilon,\mu})\boldsymbol{v}$

Since the matrix $\tilde{A}_{\epsilon,\mu}$ is large and sparse, computation of $\psi(\tau^2\tilde{A}_{\epsilon,\mu})\boldsymbol{v}$ by conventional methods (see e.g. [52], Chapter 11) is hardly feasible. However, the *action* of the matrix function ψ on a given vector at each time step can be efficiently computed by a Krylov subspace method. Algorithms for this have been developed and used in different contexts (we list in the chronological order [102, 40, 67, 91, 41, 58, 42, 60], see also Chapter 11 in the recent book [103]).

Throughout this subsection we denote $A = \tau^2\tilde{A}_{\epsilon,\mu}$, $A \in \mathbb{R}^{N\times N}$. Computation of $\psi(A)\boldsymbol{v}$ for a given vector $\boldsymbol{v}$ is based on the Arnoldi or, when $A = A^*$, on the Lanczos process (see e.g. [103, 92]). The Lanczos process involves the three-term recurrences and is therefore cheaper, especially for large Krylov subspace dimensions m. Since in this case m is not too large we use the Arnoldi process which has better numerical stability properties.

Starting with A and $\boldsymbol{v}$, the Arnoldi process generates after m steps orthonormal vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{m+1}$ (with $\boldsymbol{v}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|$) and a Hessenberg matrix $\bar{H}_m \in \mathbb{R}^{(m+1)\times m}$ such that (see [103, 92])

$$AV_m = V_{m+1}\bar{H}_m, \tag{3.27}$$

where $V_{m+1} \in \mathbb{R}^{N\times m+1}$ is a matrix with column vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{m+1}$ (and, correspondingly, V_m is V_{m+1} with the last column skipped). The vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ span the so-called Krylov subspace $K_m(A, \boldsymbol{v})$:

$$\mathrm{colspan}V_m = K_m(A, \boldsymbol{v}) := \mathrm{span}\{\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{m-1}\boldsymbol{v}\}.$$

Denote by H_m a matrix obtained from $\bar{H}_m$ by deleting its last row. As usually for the Arnoldi process, we expect that for some m

$$AV_m \approx V_m H_m, \tag{3.28}$$

where the approximation improves (but not necessarily monotonically) as m grows, see e.g. [103, 92]. Krylov subspace approximations to $\psi(A)\boldsymbol{v}$ are based on the last relation: since in the Arnoldi process by construction $\boldsymbol{v}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|$ we have

$$\boldsymbol{v} = V_m y, \qquad y = \|\boldsymbol{v}\|e_1,$$

with e_1 being the first canonical basis vector in $\mathbb{R}^m$, and (cf. (3.28))

$$\psi(A)V_m y \approx V_m \psi(H_m)y, \qquad y = \|\boldsymbol{v}\|e_1,$$

so that the action of the matrix function on the given vector $\boldsymbol{v}$ is computed as

$$\psi(A)\boldsymbol{v} \approx \|\boldsymbol{v}\|V_m\psi(H_m)e_1. \tag{3.29}$$

We emphasize that dependence of the orthonormal basis $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ on $\boldsymbol{v}$ is crucial to have a good approximation in (3.29).

In practice m is small (say 20), so that $\psi(H_m)$ in (3.29) can easily be computed by a standard method (see e.g. Chapter 11 in [52] and references therein). In the experiments presented in this chapter, $\psi(H_m)$ was computed with Matlab's built-in functions `sqrtm` and `funm`.

An important question is when to stop the Arnoldi process. One stopping criterion is proposed in [60] and is based on controlling a norm of a generalized residual. Unfortunately, in our experiments this approach appeared to be very sensitive to the given tolerance which had to be tuned for every test problem. For this reason we use another simple strategy: the Arnoldi process was stopped as soon as

$$\left\| \frac{\boldsymbol{y}_{(m)}^{n+1} - \boldsymbol{y}_{(m-1)}^{n+1}}{\boldsymbol{y}_{(m)}^{n+1} - \boldsymbol{y}_{(0)}^{n+1}} \right\|_\infty \leq \texttt{TOL}, \tag{3.30}$$

where $\boldsymbol{y}_{(m)}^{n+1}$ is the numerical solution of the scheme (3.26) obtained with m steps of the Arnoldi process, the division of the vectors is understood elementwise and `TOL` is a tolerance (in all our experiments we used the value $\texttt{TOL} = 10^{-2}$, this value should be chosen according to the relative accuracy required for a specific problem). By $\boldsymbol{y}_{(0)}^{n+1}$ we denote the solution obtained by (3.26) with $\psi(\tau^2 \tilde{A}_{\epsilon,\mu})$ set to the identity matrix (so that no Arnoldi steps are done). Note that $\boldsymbol{y}_{(0)}^{n+1}$ coincides with the solution of the leap frog scheme (cf. (3.17)) and, thus, is a second order time-consistent numerical solution. Stopping criterion (3.30) means that the further increase of the Krylov subspace dimension m leads to no further improvement in the accuracy as compared to the accuracy already obtained with respect to the leap-frog solution $\boldsymbol{y}_{(0)}^{n+1}$. Note that this stopping criterion can be shown to be a controller of the Krylov subspace error (see Section 3.4.2).

The described steps lead to the algorithm for the Gautschi-Krylov time integration scheme presented in Figure 3.1. The analysis of the Krylov subspace error made in the matrix function evaluations and the stability of the new scheme are presented in Section 3.4.

Since the work to compute the matrix function of the small matrix H_m is negligible, the overall computational work of the Gautschi scheme per time step is dominated by $m + 1$ matrix-vector multiplications with the matrix $\tilde{A}_{\epsilon,\mu}$ (m of which are required by the Arnoldi process). This means an increase by a factor of $m + 1$ as compared to the work per time step in the leap frog scheme.

$\boldsymbol{y}^n$ and $\boldsymbol{y}^{n-1}$ are given
$\boldsymbol{v} = \tilde{A}_{\epsilon,\mu}\boldsymbol{y}^n - \boldsymbol{f}^n, \quad \beta = \|\boldsymbol{v}\|_2$
$\boldsymbol{y}_{(0)}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} - \tau^2\boldsymbol{v}$
for $m = 1, \ldots,$
    extend the Krylov basis by one Arnoldi step:
        if$(m = 1)$ then
            $\boldsymbol{v}_1 = \boldsymbol{v}/\beta$
            initialize $\bar{H}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
        else
            extend $\bar{H}_{m-1}$ to $\bar{H}_m$ by adding
            a zero column and a zero row
        endif
        $\boldsymbol{w} = \tau^2\tilde{A}_{\epsilon,\mu}\boldsymbol{v}_m$
        for $i = 1, \ldots, m$
            $h_{i,m} = \boldsymbol{w}^T\boldsymbol{v}_i$
            $\boldsymbol{w} = \boldsymbol{w} - h_{i,m}\boldsymbol{v}_i$
        endfor
        $h_{m+1,m} = \|\boldsymbol{w}\|_2$
        $\boldsymbol{v}_{m+1} = \boldsymbol{w}/h_{m+1,m}$
        $V_{m+1} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m, \boldsymbol{v}_{m+1}]$
    end of Arnoldi step
    compute matrix function $\psi(H_m)$
    $\boldsymbol{u} = V_m[\beta\psi(H_m)e_1]$
    $\boldsymbol{y}_{(m)}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} - \tau^2\boldsymbol{u}$
    exit for-loop if condition (3.30) is fulfilled
endfor
$\boldsymbol{y}^{n+1} = \boldsymbol{y}_{(m)}^{n+1}$

Figure 3.1: The Gautschi scheme with the Krylov subspace matrix function evaluation and adaptive choice of the Krylov dimension.

3.3.4 Newmark scheme

The following scheme proposed by J.-F. Lee, R. Lee, and A. Cangellaris (the Newmark scheme, [71] and [48]) can be applied directly to the semidiscrete Maxwell equations (3.15):

$$M_\epsilon \frac{\boldsymbol{e}^{n+1} - 2\boldsymbol{e}^n + \boldsymbol{e}^{n-1}}{\tau^2} + A_\mu(\frac{1}{4}\boldsymbol{e}^{n-1} + \frac{1}{2}\boldsymbol{e}^n + \frac{1}{4}\boldsymbol{e}^{n+1}) = \boldsymbol{j}^n. \tag{3.31}$$

This scheme can be written in the form

$$(M_\epsilon + \frac{\tau^2}{4}A_\mu)\boldsymbol{e}^{n+1} = \tau^2\boldsymbol{j}^n - (\frac{\tau^2}{2}A_\mu - 2M_\epsilon)\boldsymbol{e}^n - (M_\epsilon + \frac{\tau^2}{4}A_\mu)\boldsymbol{e}^{n-1}, \tag{3.32}$$

revealing that a linear system with matrix $M_\epsilon + \frac{\tau^2}{4}A_\mu$ has to be solved at every time step. For discretizations obtained on relatively coarse grids this can be done by a sparse direct solver, by computing the LU factorization once and reusing it at every time step. If a direct solution is not feasible, a preconditioned Krylov iterative solver can be used.

The Newmark scheme is unconditionally (regardless of the time step τ) stable [71].

3.3.5 One-step formulations of the three schemes

Each of the three schemes described in this section is a two-step scheme (i.e. it requires numerical solutions on both n and $n-1$ time levels to get the next time level solution) but can be written in a one-step form. This is normally done by introducing an auxiliary time derivative variable. These one-step formulations can be used at the first time step where the two-step formulation would have required the normally unknown value of $\boldsymbol{e}^{-1}$.

In the context of the Maxwell equations, a natural way to obtain a one-step formulation of a time integration scheme is to consider the Maxwell equations as the two first order equations. A possible drawback of this approach is that one has to work with both fields and, hence, build up appropriate spatial discretizations for each of the fields. Thus, one of the benefits of treating the Maxwell equations as a second order equation for one of the fields is then lost.

In this section we give the one-step formulations for all schemes. We derive it for the Newmark scheme. The other two one-step formulations can be obtained in a similar way. The formulations are given for an auxiliary variable but directly

applicable to the two first order Maxwell equations, too. Introducing the time-derivative auxiliary variable as

$$\boldsymbol{u}^{n+1/2} = \frac{\boldsymbol{e}^{n+1} - \boldsymbol{e}^n}{\tau}, \tag{3.33}$$

we can write (3.31) as

$$M_\epsilon \frac{\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^{n-1/2}}{\tau} + \frac{1}{2} A_\mu \frac{\boldsymbol{e}^{n-1} + \boldsymbol{e}^n}{2} + \frac{1}{2} A_\mu \frac{\boldsymbol{e}^n + \boldsymbol{e}^{n+1}}{2} = \frac{1}{2}\boldsymbol{j}^n + \frac{1}{2}\boldsymbol{j}^n,$$

or, formally introducing the variable $\boldsymbol{u}^n$, as

$$\begin{aligned} M_\epsilon \frac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1/2}}{\tau/2} + A_\mu \frac{\boldsymbol{e}^{n-1} + \boldsymbol{e}^n}{2} &= \boldsymbol{j}^n, \\ M_\epsilon \frac{\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^n}{\tau/2} + A_\mu \frac{\boldsymbol{e}^n + \boldsymbol{e}^{n+1}}{2} &= \boldsymbol{j}^n. \end{aligned} \tag{3.34}$$

Writing the first half-step update here for the next time level (i.e. replacing n with $n+1$) we have

$$M_\epsilon \frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^{n+1/2}}{\tau/2} + A_\mu \frac{\boldsymbol{e}^n + \boldsymbol{e}^{n+1}}{2} = \boldsymbol{j}^{n+1},$$

which, together with (3.33) and (3.34) leads to the following *one-step formulation of the Newmark scheme*:

$$\begin{aligned} M_\epsilon \frac{\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^n}{\tau/2} + A_\mu \frac{\boldsymbol{e}^n + \boldsymbol{e}^{n+1}}{2} &= \boldsymbol{j}^n, \\ \frac{\boldsymbol{e}^{n+1} - \boldsymbol{e}^n}{\tau} &= \boldsymbol{u}^{n+1/2}, \\ M_\epsilon \frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^{n+1/2}}{\tau/2} + A_\mu \frac{\boldsymbol{e}^n + \boldsymbol{e}^{n+1}}{2} &= \boldsymbol{j}^{n+1}. \end{aligned} \tag{3.35}$$

In this form the sequence of computations for the scheme is not immediately clear and we rewrite it as:

$$\begin{aligned} (M_\epsilon + \frac{\tau^2}{4} A_\mu)\boldsymbol{e}^{n+1} &= \frac{\tau^2}{2}\boldsymbol{j}^n + (M_\epsilon - \frac{\tau^2}{4} A_\mu)\boldsymbol{e}^n + \tau M_\epsilon \boldsymbol{u}^n \\ M_\epsilon \boldsymbol{u}^{n+1} &= \frac{\tau}{2}\boldsymbol{j}^{n+1} - \frac{\tau}{4} A_\mu(\boldsymbol{e}^n + \boldsymbol{e}^{n+1}) + M_\epsilon \frac{\boldsymbol{e}^{n+1} - \boldsymbol{e}^n}{\tau}. \end{aligned}$$

The one-step formulations for the leap frog and the Gautschi scheme can be

obtained along the same lines (see also [59]):

$$\text{One-step leap frog:} \quad \begin{cases} M_\epsilon \dfrac{\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^n}{\tau/2} + A_\mu \boldsymbol{e}^n = \boldsymbol{j}^n, \\ \dfrac{\boldsymbol{e}^{n+1} - \boldsymbol{e}^n}{\tau} = \boldsymbol{u}^{n+1/2}, \\ M_\epsilon \dfrac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^{n+1/2}}{\tau/2} + A_\mu \boldsymbol{e}^{n+1} = \boldsymbol{j}^{n+1}. \end{cases}$$

$$\text{One-step Gautschi:} \quad \begin{cases} \dfrac{\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^n}{\tau/2} = \psi(\tau^2 \tilde{A}_{\epsilon,\mu})(-\tilde{A}_{\epsilon,\mu} \boldsymbol{y}^n + \boldsymbol{f}^n), \\ \dfrac{\boldsymbol{y}^{n+1} - \boldsymbol{y}^n}{\tau} = \boldsymbol{u}^{n+1/2}, \\ \dfrac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^{n+1/2}}{\tau/2} = \psi(\tau^2 \tilde{A}_{\epsilon,\mu})(-\tilde{A}_{\epsilon,\mu} \boldsymbol{y}^{n+1} + \boldsymbol{f}^{n+1}). \end{cases}$$

3.4 Analysis of the Gautschi-Krylov scheme

3.4.1 Krylov subspace approximation error

Theorem 3.2. *Consider the homogeneous ODE system $\boldsymbol{y}'' + A\boldsymbol{y} = 0$. Then, the solution of the Gautschi-Krylov scheme has the form:*

$$\boldsymbol{y}^{n+1} = -\boldsymbol{y}^{n-1} + 2\cos(\tau A^{1/2})\boldsymbol{y}^n + \underbrace{\int_0^\tau A^{-1/2} \sin((\tau - s)A^{1/2})\tilde{\boldsymbol{g}}(s)ds}_{=: \boldsymbol{\delta}^n,\ \textit{Krylov error}}, \tag{3.36}$$

$$\tilde{\boldsymbol{g}}(s) = -\beta h_{m+1,m} s^2 \boldsymbol{v}_{m+1} e_m^T \psi(s^2 H_m) e_1,$$

where τ is the step size, m is the Krylov dimension, $\beta = \|A\boldsymbol{y}^n\|$, $h_{m+1,m}$ is the $(m+1, m)$ entry of the matrix $\bar{H}_m$. The matrices $\bar{H}_m$, H_m, and the vector $\boldsymbol{v}_{m+1}$ are defined in (3.27),(3.28)*, e_1 and e_m are respectively the first and the last canonical basis vectors in $\mathbb{R}^m$, and ψ is given by* (3.25)*.*
For the exact Gautschi scheme (where the matrix function evaluations are done exactly) relation (3.36) *holds with $\boldsymbol{\delta}^n \equiv 0$.*

Proof The proof (inspired by the analysis given in Section 4 of [100]) consists of showing that the solution of the Gautschi-Krylov scheme is the *exact* solution of a perturbed (inhomogeneous) ODE system.

Without loss of generality, we shift for convenience the time variable such that $t_n = 0$, $t_{n+1} = t$ and the Gautschi scheme can be written as

$$\boldsymbol{y}(t) - 2\boldsymbol{y}(0) + \boldsymbol{y}(-t) = -t^2\psi(t^2A)A\boldsymbol{y}(0).$$

Substituting here function ψ as it is defined in (3.25) leads to relation (3.36) with $\boldsymbol{\delta}^n \equiv 0$ which thus indeed holds for the exact Gautschi scheme. In the Gautschi-Krylov scheme the right hand side is computed approximately with Arnoldi or Lanczos process as

$$-t^2\psi(t^2A)A\boldsymbol{y}(0) = -\beta t^2\psi(t^2A)V_m e_1 \approx -\beta t^2 V_m\psi(t^2H_m)e_1,$$

where the matrix V_m is defined in (3.27),(3.28). The Gautschi-Krylov scheme can thus be written as

$$\boldsymbol{y}(t) - 2\boldsymbol{y}(0) + \boldsymbol{y}(-t) = -\beta t^2 V_m\psi(t^2H_m)e_1. \tag{3.37}$$

Denote $(\cdot)' = d(\cdot)/dt$. Since

$$(t^2\psi(t^2H_m))'' = (2H_m^{-1} - 2\cos(tH_m^{1/2})H_m^{-1})'' = 2\cos(tH_m^{1/2}),$$

differentiating equality (3.37) twice with respect to t yields

$$[\boldsymbol{y}(t) + \boldsymbol{y}(-t)]'' = -2\beta V_m\cos(tH_m^{1/2})e_1.$$

We now use the Arnoldi relation (3.27) rewritten as

$$AV_m = V_mH_m + h_{m+1,m}\boldsymbol{v}_{m+1}e_m^T \tag{3.38}$$

and write

$$\begin{aligned}&-2\beta V_m\cos(tH_m^{1/2})e_1 = -2\beta V_mH_mH_m^{-1}\cos(tH_m^{1/2})e_1\\ &= -2\beta(AV_m - h_{m+1,m}\boldsymbol{v}_{m+1}e_m^T)H_m^{-1}\cos(tH_m^{1/2})e_1,\end{aligned}$$

so that

$$[\boldsymbol{y}(t) + \boldsymbol{y}(-t)]'' = -2\beta(AV_m - h_{m+1,m}\boldsymbol{v}_{m+1}e_m^T)H_m^{-1}\cos(tH_m^{1/2})e_1. \tag{3.39}$$

On the other hand, the right hand side of (3.37) can be transformed as

$$\begin{aligned}&-\beta t^2V_m\psi(t^2H_m)e_1 = -2\beta V_m(I - \cos(tH_m^{1/2}))H_m^{-1}e_1\\ &= -2\beta V_mH_m^{-1}e_1 + 2\beta V_m\cos(tH_m^{1/2})H_m^{-1}e_1.\end{aligned} \tag{3.40}$$

Here the term $V_mH_m^{-1}$ reads

$$V_mH_m^{-1} = A^{-1}V_m + h_{m+1,m}A^{-1}\boldsymbol{v}_{m+1}e_m^TH_m^{-1},$$

this follows from the Arnoldi relation (3.38). Substituting the last expression into (3.40) we get the following relation for the right hand side of the Gautschi-Krylov scheme (3.37):

$$-2\beta A^{-1}V_m e_1 - 2\beta h_{m+1,m}A^{-1}\boldsymbol{v}_{m+1}e_m^T H_m^{-1}e_1 + 2\beta V_m \cos(tH_m^{1/2})H_m^{-1}e_1.$$

Note that since the starting vector of the Arnoldi process is $A\boldsymbol{y}(0) = \beta\boldsymbol{v}_1$ (see Figure 3.1 and recall that $\boldsymbol{y}(0) = \boldsymbol{y}^n$), for the first term holds:

$$-2\beta A^{-1}V_m e_1 = -2\beta A^{-1}\boldsymbol{v}_1 = -2A^{-1}A\boldsymbol{y}(0) = -2\boldsymbol{y}(0)$$

and the Gautschi-Krylov scheme thus reads (cf. (3.37))

$$\begin{aligned}\boldsymbol{y}(t) - 2\boldsymbol{y}(0) + \boldsymbol{y}(-t) = &-2\boldsymbol{y}(0) - 2\beta h_{m+1,m}A^{-1}\boldsymbol{v}_{m+1}e_m^T H_m^{-1}e_1\\ &+ 2\beta V_m \cos(tH_m^{1/2})H_m^{-1}e_1.\end{aligned}$$

Here multiplication of both sides with A results in

$$A(\boldsymbol{y}(t) + \boldsymbol{y}(-t)) = -2\beta h_{m+1,m}\boldsymbol{v}_{m+1}e_m^T H_m^{-1}e_1 + 2\beta AV_m \cos(tH_m^{1/2})H_m^{-1}e_1$$

or, taking into account that $\cos(tH_m^{1/2})H_m^{-1} = H_m^{-1}\cos(tH_m^{1/2})$,

$$-2\beta AV_m H_m^{-1}\cos(tH_m^{1/2})e_1 = -A(\boldsymbol{y}(t) + \boldsymbol{y}(-t)) - 2\beta h_{m+1,m}\boldsymbol{v}_{m+1}e_m^T H_m^{-1}e_1.$$

Replacing the first term of the right hand side in (3.39) by the right hand side of the last relation, we obtain

$$\begin{aligned}[\boldsymbol{y}(t) + \boldsymbol{y}(-t)]'' = &- A(\boldsymbol{y}(t) + \boldsymbol{y}(-t)) - 2\beta h_{m+1,m}\boldsymbol{v}_{m+1}e_m^T H_m^{-1}e_1\\ &+ 2\beta h_{m+1,m}\boldsymbol{v}_{m+1}e_m^T H_m^{-1}\cos(tH_m^{1/2})e_1,\end{aligned}$$

and, using (3.25),

$$[\boldsymbol{y}(t) + \boldsymbol{y}(-t)]'' = -A(\boldsymbol{y}(t) + \boldsymbol{y}(-t)) - \underbrace{\beta h_{m+1,m}t^2\boldsymbol{v}_{m+1}e_m^T\psi(t^2H_m)e_1}_{=:\,\tilde{\boldsymbol{g}}(t)}. \tag{3.41}$$

We now can get an analytic expression for $\boldsymbol{u}(t) \equiv \boldsymbol{y}(t) + \boldsymbol{y}(-t)$ by solving the following initial-value problem:

$$\boldsymbol{u}'' = -A\boldsymbol{u} + \tilde{\boldsymbol{g}}(t), \quad \boldsymbol{u}(0) = 2\boldsymbol{y}(0), \quad \boldsymbol{u}'(0) = 0, \tag{3.42}$$

where the initial condition $\boldsymbol{u}'(0) = 0$ holds because function $\boldsymbol{u}(t)$ is even. Applying a variation-of-constants formula to this initial-value problem gives

$$\boldsymbol{u}(t) = \cos(tA^{1/2})\boldsymbol{u}(0) + A^{-1/2}\sin(tA^{1/2})\boldsymbol{u}'(0) + \int_0^t A^{-1/2}\sin((t-s)A^{1/2})\tilde{\boldsymbol{g}}(s)ds,$$

$$\boldsymbol{y}(t) + \boldsymbol{y}(-t) = 2\cos(tA^{1/2})\boldsymbol{y}(0) + \int_0^t A^{-1/2}\sin((t-s)A^{1/2})\tilde{\boldsymbol{g}}(s)ds,$$

which, after changing the time variable back (so that $\boldsymbol{y}(0) = \boldsymbol{y}^n$, $\boldsymbol{y}(\pm t) = \boldsymbol{y}^{n\pm1}$) yields the required relation (3.36). ■

3.4.2 Stopping criterion for the Arnoldi process

The proposed stopping criterion for the Arnoldi process (cf. (3.30)) can be shown to be a controller of the Krylov subspace error specified by (3.36). To see this, we assume that one time step is done with both the Gautschi-Krylov and the exact Gautschi schemes and rewrite (3.36) as

$$\begin{aligned}\boldsymbol{y}_{(m)}^{n+1} &= -\boldsymbol{y}^{n-1} + 2\cos(\tau A^{1/2})\boldsymbol{y}^n + \boldsymbol{\delta}_{(m)}^n,\\ \boldsymbol{y}_{\mathsf{ex}}^{n+1} &= -\boldsymbol{y}^{n-1} + 2\cos(\tau A^{1/2})\boldsymbol{y}^n.\end{aligned}$$

where m is the Krylov subspace dimension, $\boldsymbol{y}_{(m)}^{n+1}$ and $\boldsymbol{y}_{\mathsf{ex}}^{n+1}$ are respectively solutions of the Gautschi-Krylov and the exact Gautschi schemes and the Krylov subspace error $\boldsymbol{\delta}_{(m)}^n$ is given by (3.36):

$$\boldsymbol{\delta}_{(m)}^n = -\beta h_{m+1,m}\int_0^\tau s^2 A^{-1/2}\sin((\tau-s)A^{1/2})\boldsymbol{v}_{m+1}e_m^T\psi(s^2H_m)e_1 ds. \quad (3.43)$$

This expression can not be readily used in practice for the evaluation of $\boldsymbol{\delta}_{(m)}^n$ due to the presence of the term $A^{-1/2}\sin((\tau-s)A^{1/2})\boldsymbol{v}_{m+1}$. Computation of this matrix-vector product with the large matrix A is too expensive and an approximation should be used. This can be done in different ways. For example, one might take several first terms of the following series [46] as an approximation:

$$A^{-1/2}\sin((\tau-s)A^{1/2}) = (\tau-s)I - \frac{1}{3!}(\tau-s)^3A + \frac{1}{5!}(\tau-s)^3A^2 - \dots. \quad (3.44)$$

Note that substituting this relation in (3.43) we could obtain another, more detailed expression for the Krylov subspace error $\boldsymbol{\delta}_{(m)}^n$ (for a similar analysis see Lemma 4.1 in [100]). Instead of (3.44) one might also use some other approximations based, e.g., on Chebyshev polynomials. A more natural and efficient way for estimating the Krylov subspace error is to use the same continued Arnoldi process to get a reference solution (for a different time integration scheme, this was proposed in [100]). More specifically, assume that, in addition to the m steps of the Arnoldi process, another j steps of the process are done. Then

$$\begin{aligned}A^{-1/2}\sin((\tau-s)A^{1/2})\boldsymbol{v}_{m+1} &= A^{-1/2}\sin((\tau-s)A^{1/2})V_{m+j}e_{m+1}^{(m+j)}\\ &\approx V_{m+j}H_{m+j}^{-1/2}\sin((\tau-s)H_{m+j}^{1/2})e_{m+1}^{(m+j)},\end{aligned} \quad (3.45)$$

where $e_{m+1}^{(m+j)}$ is the $(m+1)$th canonical basis vector in $\mathbb{R}^{m+j}$. This approximation is accurate if $|h_{m+j+1,m+j}|$ is small enough (see (3.38) with m replaced by $m+j$). Since $h_{m+j+1,m+j} \approx 0$ implies $\boldsymbol{\delta}_{(m+j)}^n \approx 0$, the solution $\boldsymbol{y}_{(m+j)}^{n+1}$ of the Gautschi-Krylov scheme after $m+j$ steps is then also accurate:

$$\boldsymbol{y}_{(m+j)}^{n+1} \approx \boldsymbol{y}_{\mathsf{ex}}^{n+1}.$$

Hence, the value of $\boldsymbol{\delta}_{(m)}^n$ with approximation (3.45) can be estimated as

$$\boldsymbol{\delta}_{(m)}^n = \boldsymbol{y}_{(m)}^{n+1} - \boldsymbol{y}_{\mathsf{ex}}^{n+1} \approx \boldsymbol{y}_{(m)}^{n+1} - \boldsymbol{y}_{(m+j)}^{n+1}.$$

In the proposed stopping criterion of the Arnoldi process (cf. (3.30)), the difference $\boldsymbol{y}_{(m)}^{n+1} - \boldsymbol{y}_{(m+j)}^{n+1}$ is evaluated in a special relative norm suitable for the time stepping process. The choice $j = 1$ (also made in [100]) is appropriate since in most cases the Arnoldi process for matrix function evaluations exhibits a superlinear convergence [58, 100].

3.4.3 Stability of the Gautschi-Krylov scheme

The original Gautschi scheme (where the matrix function evaluations are performed exactly) is exact for the linear ODE system $\boldsymbol{y}'' + A\boldsymbol{y} = 0$ and hence is trivially stable. To show stability of the Gautschi-Krylov scheme, we follow approach of [60] and consider perturbations $\boldsymbol{\varepsilon}^n \equiv \boldsymbol{y}^n - \boldsymbol{y}_{\mathsf{ex}}^n$ with respect to the solution $\boldsymbol{y}_{\mathsf{ex}}^n$ of the exact Gautschi scheme. Theorem 3.2 states that

$$\begin{aligned} \boldsymbol{y}_{\mathsf{ex}}^{n+1} &= -\boldsymbol{y}_{\mathsf{ex}}^{n-1} + 2\cos(\tau A^{1/2})\boldsymbol{y}_{\mathsf{ex}}^n, \\ \text{or} \qquad \begin{bmatrix} \boldsymbol{y}_{\mathsf{ex}}^{n+1} \\ \boldsymbol{y}_{\mathsf{ex}}^n \end{bmatrix} &= \begin{bmatrix} 2\cos(\tau A^{1/2}) & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{\mathsf{ex}}^n \\ \boldsymbol{y}_{\mathsf{ex}}^{n-1} \end{bmatrix}. \end{aligned} \tag{3.46}$$

Subtracting this relation from (3.36) we arrive at

$$\begin{aligned} \boldsymbol{\varepsilon}^{n+1} &= -\boldsymbol{\varepsilon}^{n-1} + 2\cos(\tau A^{1/2})\boldsymbol{\varepsilon}^n + \boldsymbol{\delta}^n, \\ \text{or} \qquad \begin{bmatrix} \boldsymbol{\varepsilon}^{n+1} \\ \boldsymbol{\varepsilon}^n \end{bmatrix} &= \begin{bmatrix} 2\cos(\tau A^{1/2}) & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon}^n \\ \boldsymbol{\varepsilon}^{n-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\delta}^n \\ 0 \end{bmatrix}. \end{aligned}$$

For $\boldsymbol{\delta}^n \equiv 0$ this recursion coincides with the exact solution recursion (3.46) and thus is stable if and only if the ODE system to be solved is stable. One may understand stability in different ways [86, 50], for instance, we may require that

$$\|G^n\| \le K, \qquad \text{for} \quad n \ge 0, \quad n\tau \le T, \qquad G = \begin{bmatrix} 2\cos(\tau A^{1/2}) & -I \\ I & 0 \end{bmatrix}, \tag{3.47}$$

where K does not depend on τ and T is the final time, for some operator norm $\|\cdot\|$. We now assume that the exact Gautschi scheme is stable in this sense and

thus (3.47) holds true.

Stability of the Gautschi-Krylov scheme follows immediately as it does for perturbed (inhomogeneous) difference schemes (see e.g. [86], Chapter 4 or [50], Section 14). Although the Krylov approximation error $\boldsymbol{\delta}^n$ can formally be made arbitrarily small, the Gautschi-Krylov scheme remains stable even if we allow a linear growth of the norm of $\boldsymbol{\delta}^n$ with respect to the time step τ:

$$\left\| \begin{matrix} \boldsymbol{\delta}^n \\ 0 \end{matrix} \right\| \leq C\tau,$$

with C independent on τ. Denoting

$$\mathcal{E}^n = \begin{bmatrix} \boldsymbol{\varepsilon}^n \\ \boldsymbol{\varepsilon}^{n-1} \end{bmatrix}, \qquad \hat{\boldsymbol{\delta}}^n = \begin{bmatrix} \boldsymbol{\delta}^n \\ \boldsymbol{\delta}^{n-1} \end{bmatrix},$$

one can obtain a standard expression for two-level schemes

$$\mathcal{E}^n = G^n \mathcal{E}^0 + G^{n-1}\hat{\boldsymbol{\delta}}^0 + G^{n-2}\hat{\boldsymbol{\delta}}^1 + \cdots + \hat{\boldsymbol{\delta}}^{n-1},$$

from which the stability estimate follows:

$$\begin{aligned} \|\mathcal{E}^n\| &\leq \|G^n\|\,\|\mathcal{E}^0\| + \|G^{n-1}\|\, n \max_{0\leq i\leq n-1} \|\hat{\boldsymbol{\delta}}^i\| \\ &\leq K\|\mathcal{E}^0\| + KnC\tau \leq K\|\mathcal{E}^0\| + KCT. \end{aligned}$$

3.5 Dispersion Analysis

For PDE's of the wave type dispersion analysis is an important tool to understand the error behavior of the scheme.

In this section we analyze and compare, for the edge finite element spatial discretization on a uniform mesh, the numerical dispersion error for the three schemes introduced in Section 3.3. For the analysis, we make the following two assumptions:

1. Equation (3.11) is given in an infinite source free ($\boldsymbol{J} \equiv 0$) region with periodic boundary conditions:

$$\epsilon_r \partial_{tt} \boldsymbol{E} + \nabla \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) = 0. \tag{3.48}$$

2. μ_r and ϵ_r are constant scalars.

A vector field

$$\boldsymbol{E}(x,y,z,t) = \boldsymbol{E}_0 \exp(i(\mathbf{k}\cdot\mathbf{x} - \omega t)), \quad \text{where } i = \sqrt{-1}, \tag{3.49}$$

is a solution of (3.48) if the dispersion relation

$$\omega^2 = c_r^2 k^2 \tag{3.50}$$

holds, where $\mathbf{k} = (k_1, k_2, k_3)$ is the wave vector, $\mathbf{x} = (x, y, z)$, $k = \|\mathbf{k}\|_2 \equiv \sqrt{k_1^2 + k_2^2 + k_3^2}$ is the wave number, $c_r = 1/(\sqrt{\epsilon_r \mu_r})$ is the scaled speed of light, and ω is the angular frequency.

We consider the finite element discretization of (3.48) on a uniform parallelepiped mesh with elements of size $h \times h \times h$, see Figure 3.2. The angles $\angle DAB$ and $\angle CAB$ are called deformation angles.

Remark 3.3. *To avoid cumbersome expressions, we present many of the formulas for the cubic case $\angle DAB = \angle CAB = 90°$. If a formula is valid only for the cubic elements, this is explicitly reported. However, the whole analysis is valid for the general case and the resulting plots of the dispersion errors are given also for the deformed mesh. Part of computations for the dispersion analysis were done in Maple.*

On this regular mesh the finite element matrices (3.16) take the form $M_\epsilon = \epsilon_r h M$ and $A_\mu = \dfrac{1}{h\mu_r} A$, where the matrices M and A do not depend on the element size h. This results in the following system of ODE's

$$M\boldsymbol{e}'' + \frac{c_r^2}{h^2} A\boldsymbol{e} = 0. \tag{3.51}$$

The time exact dispersion equation is

$$-\omega^2 M\boldsymbol{e} + \frac{c_r^2}{h^2} A\boldsymbol{e} = 0. \tag{3.52}$$

We end up with an eigenvalue problem with large sparse matrices given in (3.51). Since we are working on a uniform mesh, it is possible to reduce the problem size as follows:

The expansion coefficients of the finite element approximation are $e_j(t) = \displaystyle\int_{\text{edge } j} \boldsymbol{E}(\mathbf{x},t)\cdot\boldsymbol{t}_j ds$. If the exact solution of (3.48) is given by (3.49) then for any two parallel edges p and j the expansion coefficients satisfy

$$e_p^{n+q} = \exp\left(i(\mathbf{k}\cdot\Delta_{pj} - \omega q\tau)\right) e_j^n, \tag{3.53}$$

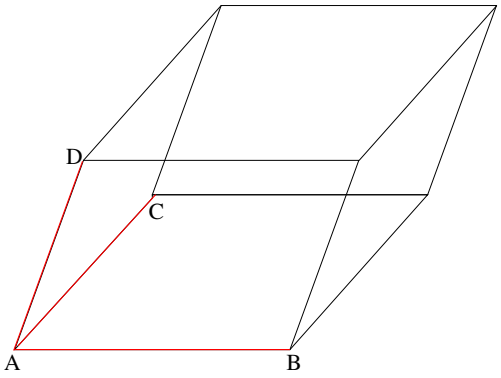


Figure 3.2: Deformed element with deformation angles $\angle CAB$ and $\angle DAB$. The angle $\angle DAC = 90°$.

where the superscript indicates the time level, the subscript indicates the number of the edge to which the coefficient belongs, and Δ_{pj} is a vector from the midpoint of edge p to the midpoint of edge j.

3.5.1 Gautschi method

We analyze the Gautschi scheme under the assumption that the action of the matrix function (3.25) on a given vector can be computed exactly (or very accurately) so that the scheme is exact in time. This assumption is realistic (see Section 3.6.3). Hence, we consider the time-accurate dispersion relation (3.52) for the system (3.51), which gives us the following generalized eigenvalue problem

$$-\omega^2 M\boldsymbol{e}^n + \frac{c_r^2}{h^2} A\boldsymbol{e}^n = 0. \tag{3.54}$$

Denoting $\varphi(\omega) = -\omega^2$ and $\eta = \dfrac{c_r^2}{h^2}$, we have

$$\varphi(\omega) M\boldsymbol{e}^n + \eta A\boldsymbol{e}^n = 0. \tag{3.55}$$

Using the relations (3.53) it is not difficult to see that on a uniform grid the equations (3.55) are the same (up to a constant $\tilde{C}_{pj}$) for parallel edges, i. e. for any two parallel edges p and j holds:

$$\varphi(\omega) M(a_p,:)\boldsymbol{e}^n + \eta A(a_p,:)\boldsymbol{e}^n = \tilde{C}_{pj}(\varphi(\omega) M(a_j,:)\boldsymbol{e}^n + \eta A(a_j,:)\boldsymbol{e}^n) = 0,$$

where $M(a_j,:)$ denotes a_jth row of matrix M, and similarly for A. Therefore it is sufficient to consider the equations corresponding to any three edges a_1, a_2, a_3 among which there are no parallel edges (see Figure 3.3).

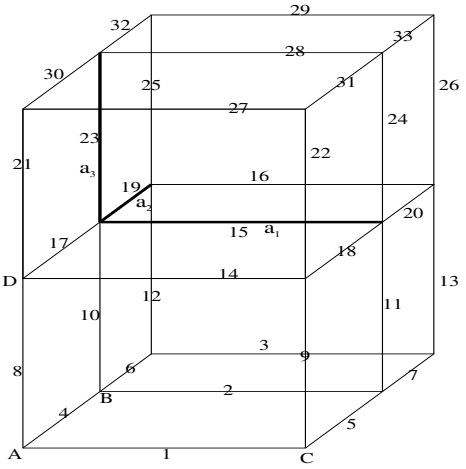


Figure 3.3: Three nonparallel edges a_1, a_2, a_3 and the degrees of freedom (with a local numbering) that appear in equation (3.54) for edge a_1.

Let

$$X(t) = \int_{a_1} \boldsymbol{E}(\mathbf{x},t)\cdot\boldsymbol{t}\,da, \quad Y(t) = \int_{a_2} \boldsymbol{E}(\mathbf{x},t)\cdot\boldsymbol{t}\,da, \quad Z(t) = \int_{a_3} \boldsymbol{E}(\mathbf{x},t)\cdot\boldsymbol{t}\,da,$$

then using (3.53) all the other degrees of freedom (coefficients) in the whole mesh can be expressed in terms of X, Y, Z.
The corresponding equation of edge a_1 is

$$\varphi(\omega)M(a_1,:)\boldsymbol{e}^n + \eta A(a_1,:)\boldsymbol{e}^n = 0. \tag{3.56}$$

The matrices M and A have a sparse structure because in (3.56) coefficients only of those basis functions are present which have nonempty common support with the basis function corresponding to the edge a_1. On a cubic mesh we have

$$\begin{aligned} M(a_1,:)\boldsymbol{e}^n =& \frac{1}{36}(1,4,1,4,16,4,1,4,1)\cdot(\tilde{e}_1,\tilde{e}_2,\tilde{e}_3,\tilde{e}_{14},\tilde{e}_{15},\tilde{e}_{16},\tilde{e}_{27},\tilde{e}_{28},\tilde{e}_{29})^T,\\ A(a_1,:)\boldsymbol{e}^n =& \frac{1}{6}(-2,-2,-2,1,-1,-1,1,1,-1,4,-4,1,-1,-2,16,-2,\\ & 4,-4,-4,4,-1,1,-4,4,-1,1,-2,-2,-2,1,-1,-1,1)\\ & \cdot(\tilde{e}_1,\tilde{e}_2,\tilde{e}_3,\ldots,\tilde{e}_{32},\tilde{e}_{33})^T. \end{aligned} \tag{3.57}$$

Here the tilde sign is used to distinguish the local index with the global index, for example $\tilde{e}_{15} = e_{a_1}$, $\tilde{e}_{19} = e_{a_2}$. Writing the relations similar to (3.56) for edges a_2 and a_3 and using (3.53), we obtain a homogeneous system of equations

$$(\varphi(\omega)F + \eta G)\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = 0. \tag{3.58}$$

On both cubic and deformed meshes the numerical dispersion relation of the Gautschi scheme is

$$\begin{aligned} &\det\left(\varphi(\omega)F + \eta G\right) = 0, \qquad \text{or} \\ &\det\left(-\omega^2 F + \frac{c_r^2}{h^2}G\right) = 0, \end{aligned} \tag{3.59}$$

where the 3×3 matrices F and G depend on the wave vector $\mathbf{k}$ and the mesh size (entries of F and G are specified for the cubic mesh in Appendix 3.8.2). One of the solutions of the dispersion relation is $\omega = 0$, which does not represent anything physical. The other solutions of (3.59) satisfy

$$(\omega_h h)^2 = 18\frac{4-\cos\xi_3\cos\xi_2-\cos\xi_1\cos\xi_2\cos\xi_3-\cos\xi_3\cos\xi_1-\cos\xi_1\cos\xi_2}{(2+\cos\xi_1)(2+\cos\xi_2)(2+\cos\xi_3)}c_r^2, \tag{3.60}$$

where $\xi_i = hk_i$, $i = 1, 2, 3$, and ω_h denotes the numerical angular frequency. The exact phase velocity is given by $c_r = \omega/k$ and the numerical phase velocity is $v = \omega_h/k$. In Figure 3.4 a plot of the phase velocity error is given for cubic elements with $k_3 = 0$. For all the numerical experiments throughout this section we assume that $\epsilon_r = \mu_r = 1$.
Under the assumption $|kh| \ll 1$ the Taylor expansion of (3.60) shows

$$\omega_h = c_r k(1 + \frac{1}{24}\frac{k_1^4 + k_2^4 + k_3^4}{k^2}h^2 + \text{higher order terms}),$$

which means that the dispersion relation for the Gautschi scheme is satisfied up to second order.

3.5.2 Leap frog scheme

Applying relation (3.53) to the leap frog scheme (3.17), we have

$$\begin{aligned} \frac{\boldsymbol{e}^{n+1} - 2\boldsymbol{e}^n + \boldsymbol{e}^{n-1}}{\tau^2} &= \frac{\exp(-i\omega\tau)\boldsymbol{e}^n - 2\boldsymbol{e}^n + \exp(i\omega\tau)\boldsymbol{e}^n}{\tau^2} \\ &= \frac{2(\cos(\omega\tau) - 1)}{\tau^2}\boldsymbol{e}^n. \end{aligned} \tag{3.61}$$

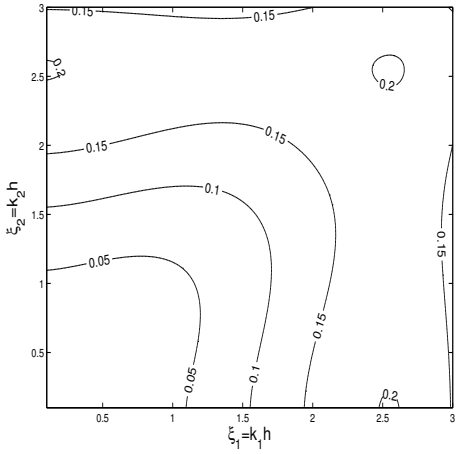


Figure 3.4: The phase velocity error of the Gautschi scheme for cubic elements.

Then the generalized eigenvalue problem of the leap frog scheme is

$$\frac{2(\cos(\omega\tau)-1)}{\tau^2}M\boldsymbol{e}^n + \frac{c_r^2}{h^2}A\boldsymbol{e}^n = 0. \tag{3.62}$$

Introducing $\varphi(\omega) = \dfrac{2(\cos(\omega\tau)-1)}{\tau^2}$ and $\eta = \dfrac{c_r^2}{h^2}$ in (3.55) we obtain the dispersion equation for the leap frog scheme

$$\det\left(\frac{2(\cos(\omega\tau)-1)}{\tau^2}F + \frac{c_r^2}{h^2}G\right) = 0, \tag{3.63}$$

with the 3×3 matrices F and G defined as in (3.59). There are 3 roots, one is zero which is non physical. The solution of (3.63) satisfies (on a cubic mesh)

$$\cos(\omega\tau) = 1 - 2\frac{\chi_1(\tau,h,\mathbf{k})}{\chi_2(\tau,h,\mathbf{k})}, \tag{3.64}$$

where

$$\begin{aligned}\chi_1(\tau,h,\mathbf{k}) &= 9c_r^2\tau^2(4-\cos\xi_1\cos\xi_2\cos\xi_3-\cos\xi_1\cos\xi_2-\cos\xi_2\cos\xi_3-\\ &\qquad -\cos\xi_3\cos\xi_1),\\ \chi_2(\tau,h,\mathbf{k}) &= 2h^2(2+\cos\xi_1)(2+\cos\xi_2)(2+\cos\xi_3),\end{aligned}$$

and $\xi_i = hk_i,\ i=1,2,3.$

According to the exact dispersion relation (3.50), we would like to have only real solutions ω of (3.64). Otherwise, as it is clear from (3.49), the imaginary part of ω will contribute to dissipation of the solution (damping if $Im(\omega) < 0$ or amplification if $Im(\omega) > 0$, see e.g. [106]). The value of ω is real if and only if

$$\left| 1 - 2\frac{\chi_1(\tau, h, \mathbf{k})}{\chi_2(\tau, h, \mathbf{k})} \right| \leq 1,$$

or, equivalently,

$$\frac{c_r \tau}{h} \leq \frac{1}{3} \sqrt{\frac{2(2+\cos\xi_1)(2+\cos\xi_2)(2+\cos\xi_3)}{4 - \cos\xi_1\cos\xi_2\cos\xi_3 - \cos\xi_1\cos\xi_2 - \cos\xi_2\cos\xi_3 - \cos\xi_3\cos\xi_1}}. \tag{3.65}$$

Since it is always true that

$$\sqrt{\frac{2(2+\cos\xi_1)(2+\cos\xi_2)(2+\cos\xi_3)}{4 - \cos\xi_1\cos\xi_2\cos\xi_3 - \cos\xi_1\cos\xi_2 - \cos\xi_2\cos\xi_3 - \cos\xi_3\cos\xi_1}} \geq 1,$$

for the inequality (3.65) to hold true it is sufficient to require that

$$\frac{c_r \tau}{h} \leq \frac{1}{3}, \tag{3.66}$$

which gives stability condition on the uniform mesh. A more general stability condition is given in Appendix 3.8.1.

Under the assumption $|kh| \ll 1$ the Taylor expansion of (3.64) shows

$$\omega_\tau = c_r k (1 + \frac{1}{24} c_r^2 k^2 \tau^2 + \frac{1}{24} \frac{k_1^4 + k_2^4 + k_3^4}{k^2} h^2 + \text{higher order terms}),$$

where ω_τ is the numerical angular frequency. In order to have spatial and temporal error terms of the same order, we should take $\tau = O(h)$. This is a clear disadvantage of leap frog compared to Gautschi.

In Figures 3.5–3.7, the absolute error of the angular frequency for the leap frog scheme is shown in comparison with the Gautschi scheme for different values of the time step τ and deformation angles θ ($\angle DAC = \angle BAC = \theta$, see Figure 3.2). Here, for simplicity, we assume $k_3 = 0$. Note that in all figures the plots of the leap frog scheme become increasingly similar (as τ decreases) to the plot of the time-exact Gautschi scheme. We observe that reduction of the time step beyond 0.002 does not give more accurate results because the spatial error is dominant.

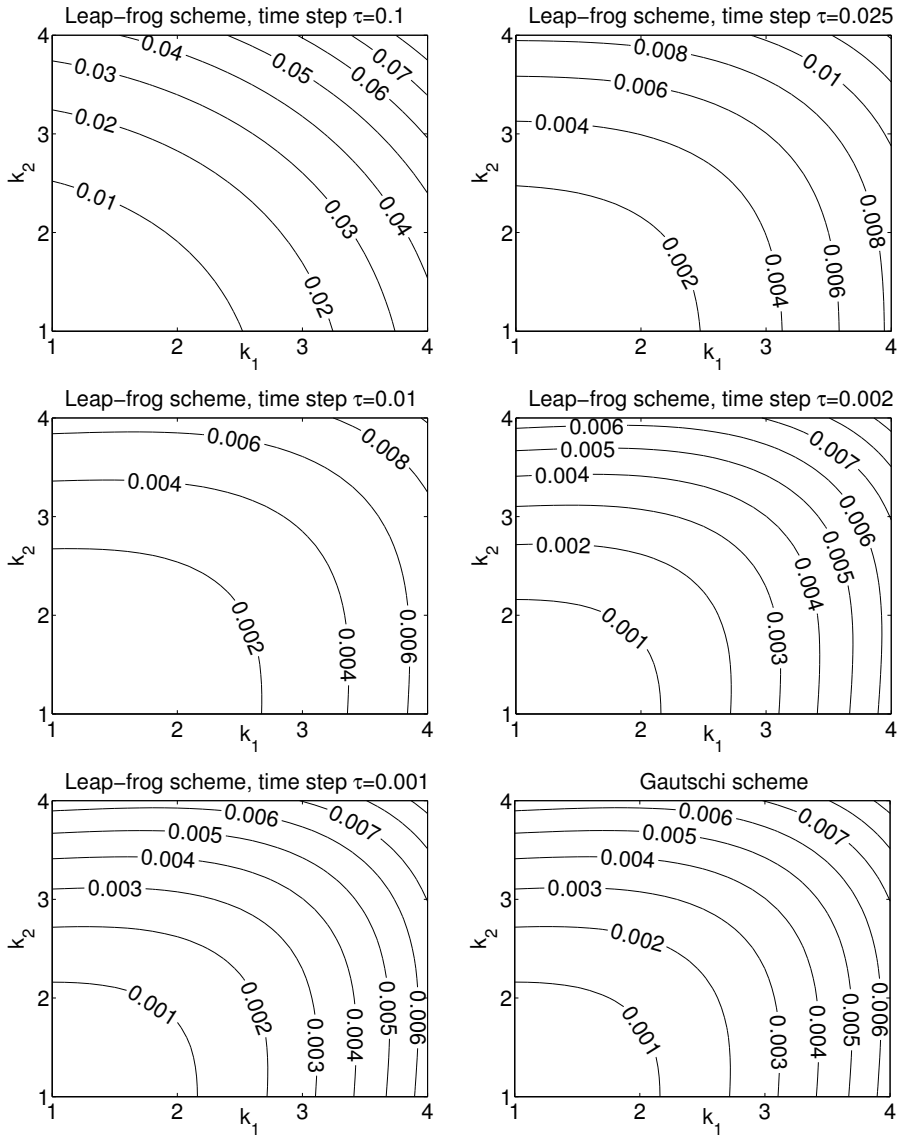


Figure 3.5: Absolute value of the angular frequency errors for the leap frog scheme with different time steps and for the Gautschi scheme, mesh size $h = 1/20$, deformation angle $\theta = \pi/2$.

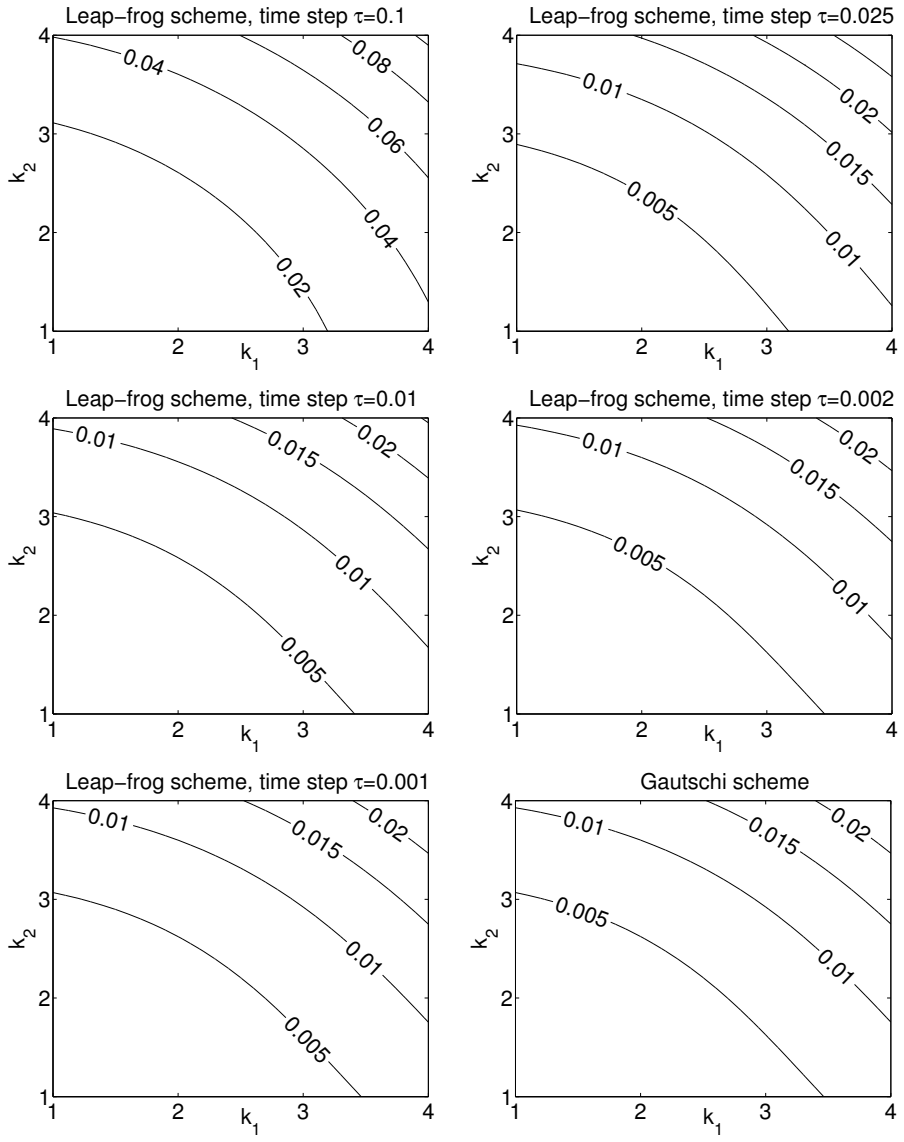


Figure 3.6: Absolute value of the angular frequency errors for the leap frog scheme with different time steps and for the Gautschi scheme, mesh size $h = 1/20$, deformation angle $\theta = \pi/3$.

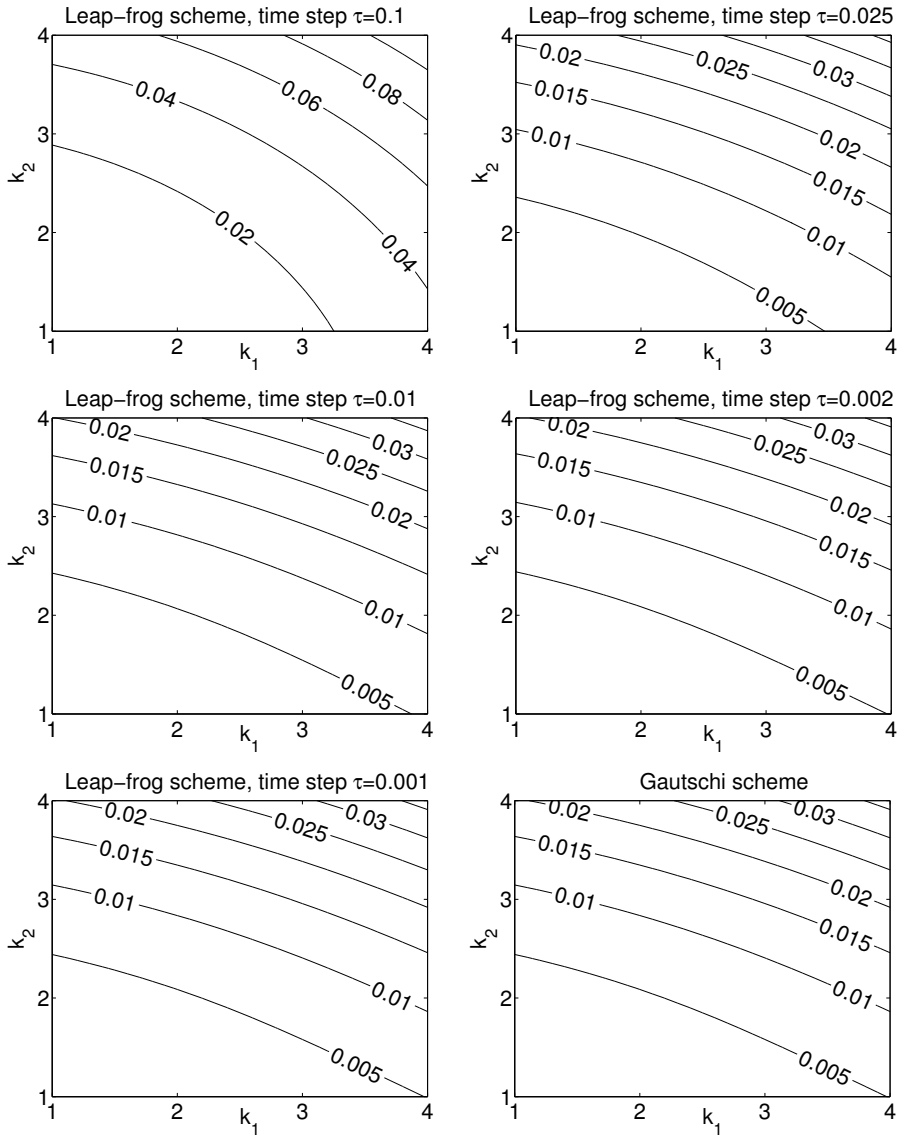


Figure 3.7: Absolute value of the angular frequency errors for the leap frog scheme with different time steps and for the Gautschi scheme, mesh size $h = 1/20$, deformation angle $\theta = \pi/4$.

3.5.3 Newmark scheme

The generalized eigenvalue problem for the Newmark scheme (3.31) is

$$\frac{2(\cos(\omega\tau)-1)}{\tau^2}M\boldsymbol{e}^n+\frac{(\cos(\omega\tau)+1)}{2}\frac{c_r^2}{h^2}A\boldsymbol{e}^n=0.$$

Introducing $\varphi(\omega)=\dfrac{2(\cos(\omega\tau)-1)}{\tau^2}$ and $\eta=\dfrac{\cos(\omega\tau)+1}{2}\dfrac{c_r^2}{h^2}$ in (3.55) we obtain the dispersion equation for the Newmark scheme

$$\det\left(\frac{2(\cos(\omega\tau)-1)}{\tau^2}F+\frac{(\cos(\omega\tau)+1)}{2}\frac{c_r^2}{h^2}G\right)=0, \tag{3.67}$$

where the 3×3 matrices F and G are given as in (3.59). There are 3 roots, one is zero. The solution of (3.67) satisfies (on a cubic mesh)

$$\cos(\omega\tau)=\frac{\chi_2(\tau,h,\mathbf{k})-\chi_1(\tau,h,\mathbf{k})}{\chi_2(\tau,h,\mathbf{k})+\chi_1(\tau,h,\mathbf{k})}, \tag{3.68}$$

where

$$\begin{aligned}\chi_1(\tau,h,\mathbf{k})&=9c_r^2\tau^2(4-\cos\xi_1\cos\xi_2\cos\xi_3-\cos\xi_1\cos\xi_2-\cos\xi_2\cos\xi_3-\\&\qquad-\cos\xi_3\cos\xi_1),\\ \chi_2(\tau,h,\mathbf{k})&=2h^2(2+\cos\xi_1)(2+\cos\xi_2)(2+\cos\xi_3),\end{aligned}$$

and $\xi_i=hk_i,\ i=1,2,3.$

Under the assumption $|kh|\ll 1$ the Taylor expansion of (3.68) shows

$$\begin{aligned}\omega_\tau=c_rk(1-\frac{1}{12}c_r^2k^2\tau^2+\frac{1}{24}\frac{k_1^4+k_2^4+k_3^4}{k^2}h^2\\ +O(h^4)+O(\tau^4)+O(\tau^2h^2)+\text{higher order terms}),\end{aligned} \tag{3.69}$$

where ω_τ denotes the numerical angular frequency. In order to make the spatial and temporal error terms of the same order, we should take $\tau=O(h)$. We note that the dispersion error of the Newmark scheme becomes fourth order accurate if we choose

$$\tau=\sqrt{\frac{1}{2c_r^2}\frac{k_1^4+k_2^4+k_3^4}{k^4}}\,h, \tag{3.70}$$

which can be called *an optimum time step*. We note that (3.69),(3.70) are only valid on a cubic mesh.

In Figures 3.8–3.10, the absolute error of the angular frequency for the Newmark scheme is shown in comparison with the time-accurate Gautschi scheme for different values of time step τ and deformation angles θ ($\angle DAC = \angle BAC = \theta$, see Figure 3.2). Here again we assume for simplicity $k_3 = 0$.
For the Newmark scheme we observe a similar convergence behavior as for the leap frog scheme. Note that the plot for the step size $\tau = 0.025$ in Figure 3.8 differs significantly from the the other plots in the figure due to the increase in the error order observed in (3.69) (cf. (3.70) with $k_3 = 0$ and $k_1 \approx k_2$).

3.6 Numerical experiments

3.6.1 Test problem 1

This test problem is obtained by choosing an arbitrary vector field function $\boldsymbol{E}_{\mathsf{an}}(x, y, z, t)$ satisfying the boundary conditions, projecting it onto the finite element subspace and substituting the projection into the semidiscrete system (3.15). The source function $\boldsymbol{j}(t)$ is then chosen such that the finite element projection of $\boldsymbol{E}_{\mathsf{an}}$ is the exact solution of (3.15). Note that it is important to use the exact solution of the *semidiscrete* system because the difference of this solution with the computed numerical solution represents then *solely* the time error (without the spatial discretization error).

More specifically, we consider the dimensionless Maxwell equations (3.11) in the domain $\Omega = [0, 1] \times [0, 1] \times [0, 1]$ and we take

$$\boldsymbol{E}_{\mathsf{an}}(x, y, z, t) = v(t)\bar{\boldsymbol{E}}(x, y, z).$$

If $\bar{\boldsymbol{e}}$ is the finite element projection of the field $\bar{\boldsymbol{E}}$ then

$$\boldsymbol{e}_{\mathsf{an}}(t) = v(t)\bar{\boldsymbol{e}}$$

is the exact solution of the semidiscrete ODE system (3.15) with

$$\boldsymbol{j}(t) = (v'' M_\epsilon + v A_\mu)\bar{\boldsymbol{e}}.$$

In our experiments we took

$$\epsilon_r = 1, \qquad \mu_r = 1.$$

$$v(t) = \sum_{i=1}^{N_\omega} \cos \omega_i t, \qquad \bar{\boldsymbol{E}}(x, y, z) = \begin{bmatrix} \sin \pi y \, \sin \pi z \\ \sin \pi x \, \sin \pi z \\ \sin \pi x \, \sin \pi y \end{bmatrix}. \tag{3.71}$$

where the values of ω_i are reported later separately for each of the test runs. This test problem is well suited for studying the evolution of the time error, since the exact solution is readily computable for any moment of time t.

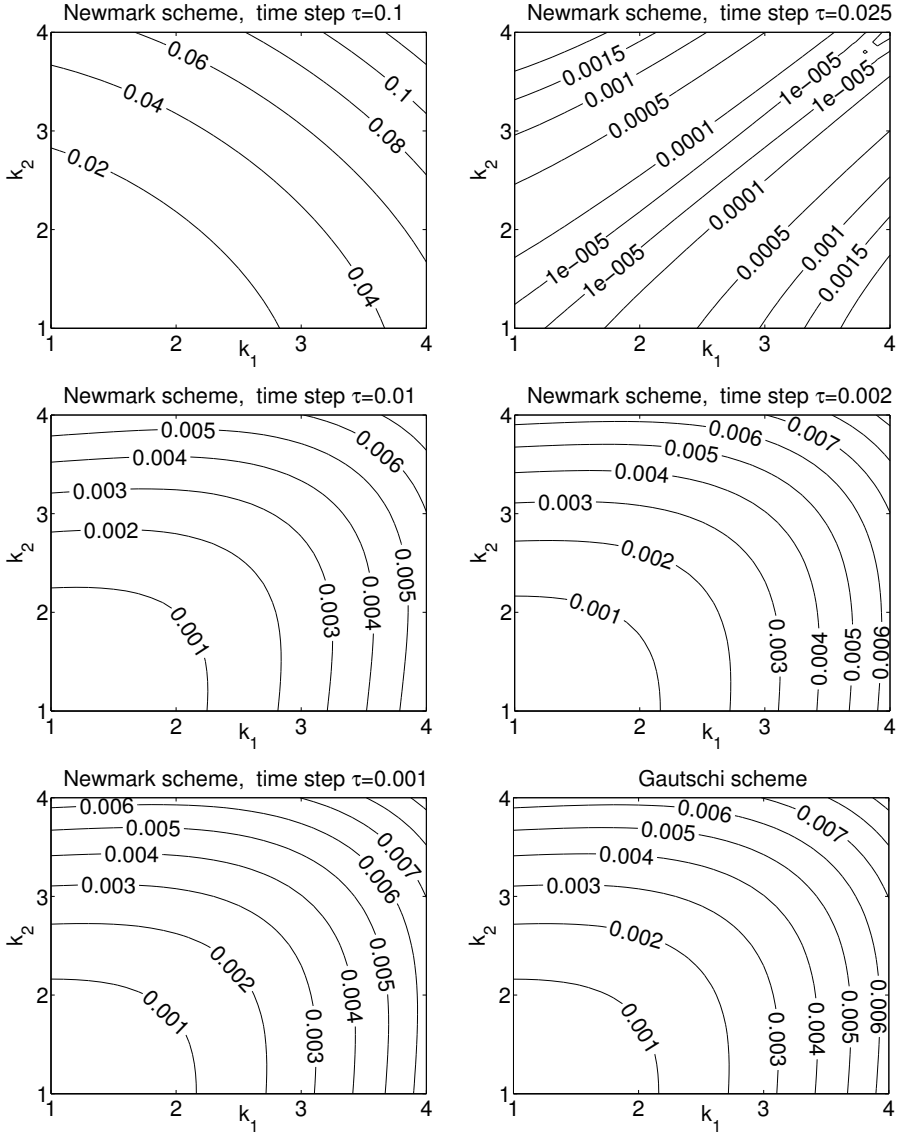


Figure 3.8: Absolute value of the angular frequency errors for the Newmark scheme with different time steps and for the Gautschi scheme, mesh size $h = 1/20$, deformation angle $\theta = \pi/2$. The plot for the time step $\tau = 0.025$ reflects the increase in the error order (cf. (3.70) with $k_3 = 0$ and $k_1 \approx k_2$).

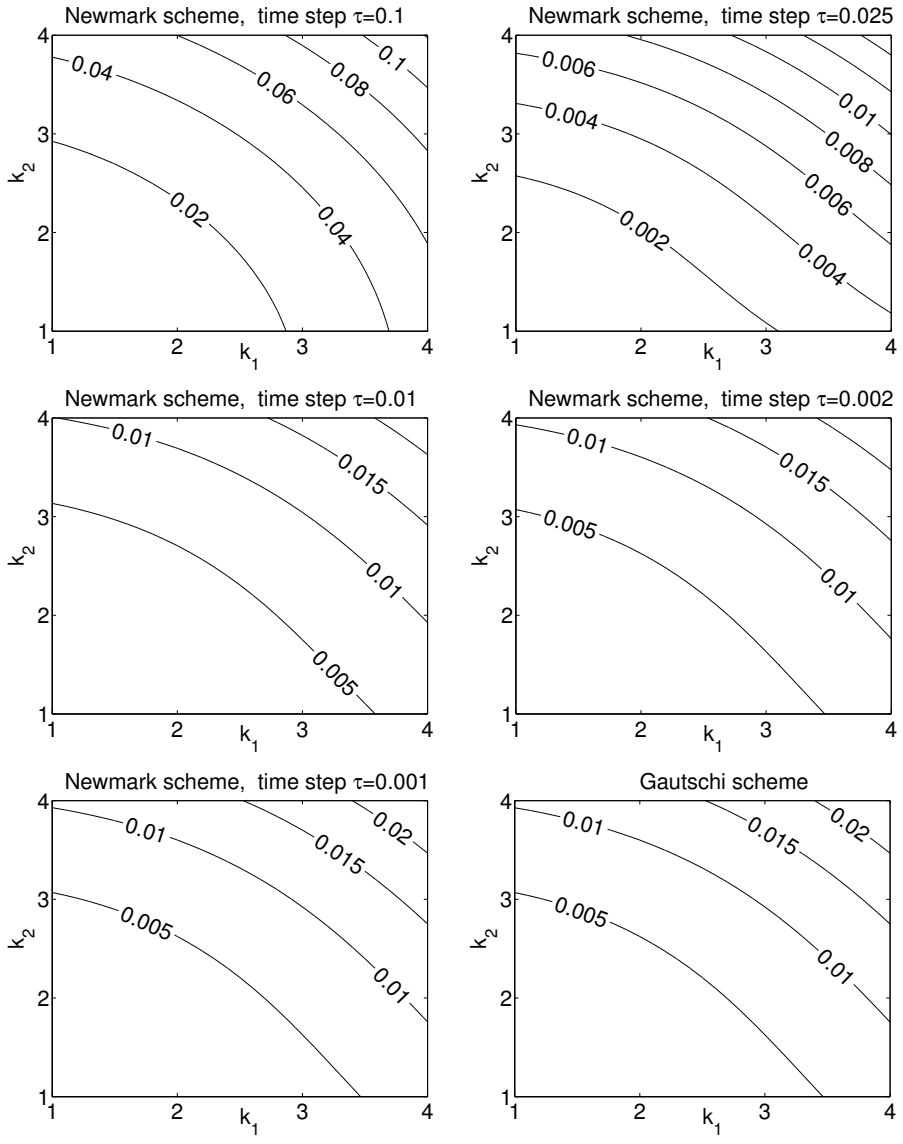


Figure 3.9: Absolute value of the angular frequency errors for the Newmark scheme with different time steps and for the Gautschi scheme, mesh size $h = 1/20$, deformation angle $\theta = \pi/3$.

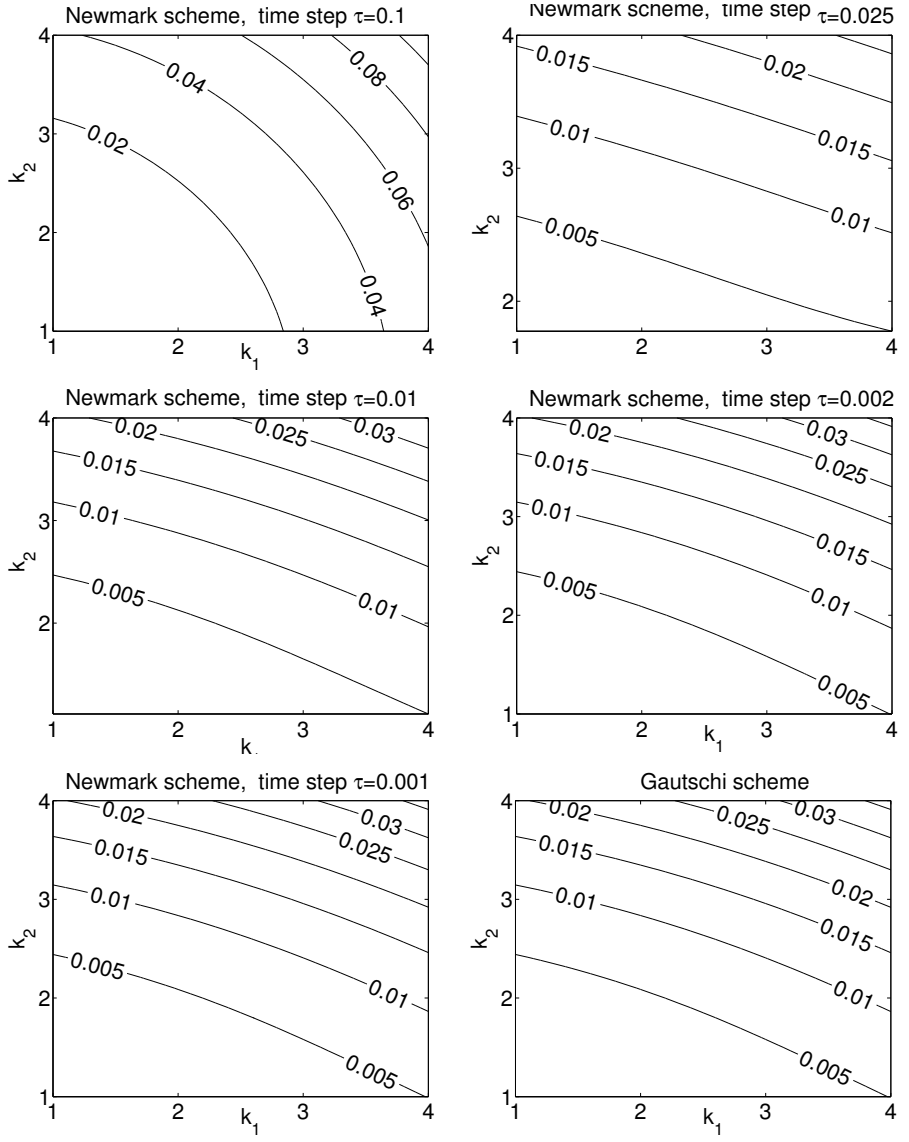


Figure 3.10: Absolute value of the angular frequency errors for the Newmark scheme with different time steps and for the Gautschi scheme, mesh size $h = 1/20$, deformation angle $\theta = \pi/4$.

3.6.2 Test problem 2

This test problem differs from the previous one only by the choice of the exact (reference) solution. The exact solution is obtained by any of the available schemes run with an extremely small time step size τ. With this τ all schemes produce numerical solutions which are practically exact in time but with the same spatial error as the numerical solutions obtained for realistically large τ. Such a testing approach is common in numerical time integration of space-discretized PDE's (see e.g. [96]). This test problem is convenient when one wants to know the error at the final time.

3.6.3 The Krylov subspace dimension and the time error

Here we investigate how the choice of the Krylov subspace dimension in the Gautschi scheme influences its time integration error. We are interested in the evolution of the error in time and therefore use Test problem 1. The frequencies ω_i of the inhomogeneous term $\boldsymbol{j}(t)$ (cf. 3.71) are chosen as

$$\omega_1 = 1, \quad \omega_2 = 10.$$

The results are presented in Figure 3.11. Here, the time error evolution of the Gautschi scheme is shown for different fixed Krylov subspace dimensions m and for the adaptive choice of m based on the condition (3.30). The time integration was done up to the final time $T = 6\frac{2\pi}{\max_i\{\omega_i\}}$ corresponding to the 6 periods of time. The shown error is the Euclidian norm of the difference between the coefficients of the finite element basis expansions of the numerical and the exact solutions.

Inspection of the plots in Figure 3.11 shows that there is a certain value of $m = \tilde{m}$ such that increasing the Krylov subspace dimension beyond $\tilde{m}$ does not lead to any improvement in time accuracy. In other words, even if we compute the action of the matrix function on vectors very accurately the error does not decrease. Thus, for $m \geq \tilde{m}$ we have a scheme where the error caused by the Krylov subspace approximation is negligible as compared to the time error of the exact Gautschi scheme. The adaptive choice of m is able to catch the value of $\tilde{m}$ very accurately: for example, for the upper plot ($14 \times 14 \times 14$ mesh) we can see that $\tilde{m} \approx 4$ whereas the adaptive choice gave values m between 3 and 5.

The typical dimensions of the Krylov subspace, observed in practice, depend on the time step size used. For the step sizes up to a factor two larger than the CFL number (which is the maximal possible step size of the explicit leap frog scheme) the Krylov dimension is usually 2. For larger realistic time step

sizes values of m up to 12 can be observed. The values of m mildly grow as the spatial mesh gets finer.

3.6.4 Computational work

We recall that on the uniform meshes the computational work per time step in the Gautschi scheme is a factor $m+1$ (with m being the Krylov subspace dimension) more than for the leap frog scheme.

On uniform meshes the computational work of the Newmark scheme is difficult to compare explicitly with those of the leap frog and Gautschi schemes. This is because on uniform meshes the sparse LU factorization of the matrix $M_\epsilon + \frac{\tau^2}{4}A_\mu$ in the Newmark scheme is more expensive than that of M_ϵ, since the matrix M_ϵ is sparser due to orthogonality of some basis functions on the cubic elements. This makes the Newmark scheme very expensive on finer meshes as compared to the other two schemes. For this reason the results for the Newmark scheme in this section are shown only for a coarser $10 \times 10 \times 10$ mesh.

On the uniform meshes let us denote the computational work required for the LU factorizations of the matrices $M_\epsilon + \frac{\tau^2}{4}A_\mu$ and M_ϵ as *lu_fac_Newmark* and *lu_fac_lf*, respectively. The computational work for one matrix-vector multiplication with the matrices $M_\epsilon - \frac{\tau^2}{4}A_\mu$ and A_μ involved in the Newmark and leap frog schemes is defined as *mat_vec_Newmark* and *mat_vec_lf*, respectively. The computational work required for the LU solver for the schemes Newmark and leap frog is denoted as *lu_sol_Newmark* and *lu_sol_lf*, respectively.

In contrast to the situation on the uniform meshes, the matrices M_ϵ and $M_\epsilon + \frac{\tau^2}{4}A_\mu$ have the same sparsity structure on unstructured meshes, hence require the same computational work for the LU factorization. Although the computational work per time step in the Gautschi scheme is larger than in the Newmark or the leap frog scheme, the Gautschi scheme appears to be more efficient (see results of Section 3.6.5). Let us define a relative work required for one LU factorization as *lu_fac*, one matrix-vector multiplication as *mat_vec* and one LU solver as *lu_sol*. It is clear that per time step the Newmark and the leap frog schemes require *mat_vec* + *lu_sol* and the Gautschi scheme requires $(m+1)(mat\_vec + lu\_sol)$ operations.

If we denote the required computational work per time step for the cases de-

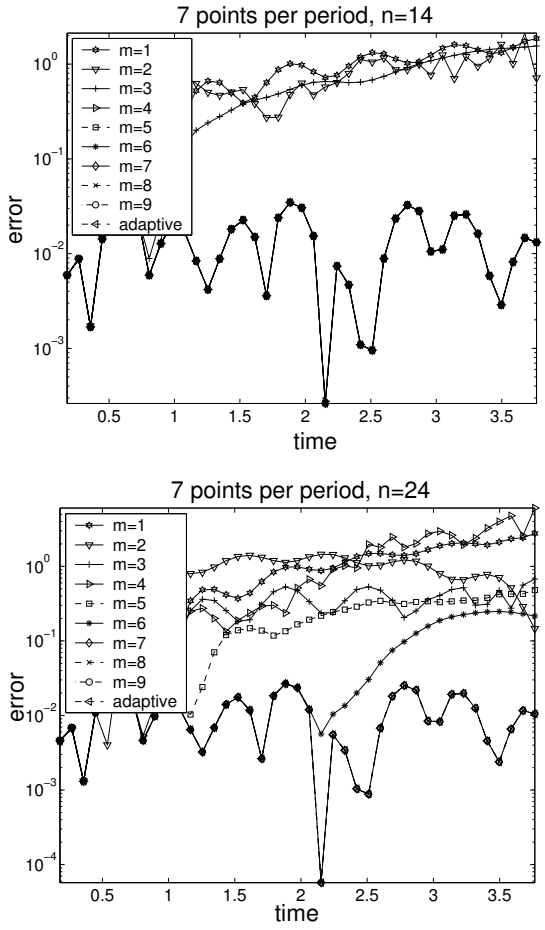


Figure 3.11: Error evolution of the Gautschi scheme for $14 \times 14 \times 14$ (top) and $24 \times 24 \times 24$ (bottom) meshes for different Krylov subspace dimensions m. The step size corresponds to 7 points per time period.

Table 3.1: Computational work for the three schemes.

	uniform mesh	unstructured mesh
Gautschi	$\frac{T}{\tau}(m+1)\cdot \mathtt{Q}_{lf} + lu\_fac\_lf$	$\frac{T}{\tau}(m+1)\cdot \mathtt{Q} + lu\_fac$
Newmark	$\frac{T}{\tau}\cdot \mathtt{Q}_{Newmark} + lu\_fac\_Newmark$	$\frac{T}{\tau}\cdot \mathtt{Q} + lu\_fac$
leap frog	$\frac{T}{\tau}\cdot \mathtt{Q}_{lf} + lu\_fac\_lf$	$\frac{T}{\tau}\cdot \mathtt{Q} + lu\_fac$

scribed above as

$$\begin{aligned} \mathtt{Q}_{lf} &= mat\_vec\_lf + lu\_sol\_lf\,, \\ \mathtt{Q}_{Newmark} &= mat\_vec\_Newmark + lu\_sol\_Newmark, \\ \mathtt{Q} &= mat\_vec + lu\_sol, \end{aligned}$$

then the overall computational work for all the schemes on the uniform and unstructured meshes is given in Table 3.1, where T is the final time and τ is the time step size.
On finer uniform or unstructured meshes the LU factorizations may require too much computational efforts. In this case one could use an iterative solver for the three schemes. In the context of the Arnoldi process used in the Gautschi scheme this would mean that the action of M_ϵ^{-1} is computed by an inner iterative solver. Note that the matrix $M_\epsilon + \frac{\tau^2}{4}A_\mu$ appearing in the Newmark scheme usually requires more iterations of an iterative solver than the well-conditioned mass matrix M_ϵ [45]. Performance of the iterative solvers in all the schemes can be improved by a suitable preconditioning (see [100] for preconditioning of the Krylov subspace matrix function evaluations). On the other hand, the use of approximate implicit schemes [19] or stabilized explicit schemes [105, 95, 96] might be a good option here, too.

3.6.5 Comparisons of the three schemes

We compare now the time stepping errors at the final time and the CPU times of the three schemes presented in Section 3.3. Since we are interested in time errors at the final time, we use Test problem 2. The presented error values are computed as

$$\mathtt{error} = \left\| \frac{\boldsymbol{y}^{\bar{n}} - \boldsymbol{y}^{\bar{n}}_{\text{exact}}}{\boldsymbol{y}^{\bar{n}}_{\text{exact}} + \epsilon_C} \right\|_\infty , \tag{3.72}$$

where the division of the vectors is understood element-wise, $\boldsymbol{y}^{\bar{n}}$ and $\boldsymbol{y}^{\bar{n}}_{\text{exact}}$ are the numerical and the exact (reference) solutions at the final time $T = \bar{n}\tau = 50$, and ϵ_C is the machine epsilon.

Uniform cubic mesh

In the experiments presented in this section, a uniform cubic mesh was used. In the first test, the frequencies ω_i of the source term $\boldsymbol{j}(t)$ were taken to be homogeneously distributed:

$$N_\omega = 101, \qquad w_i \text{ evenly distributed in } [1, 10],\ i = 1, \ldots, 101. \tag{3.73}$$

The results are presented in Figure 3.12. We see that all the schemes clearly exhibit second order time accuracy. The peculiar drop in the error-versus-τ plot of the Newmark scheme, is caused by the increase in the error order observed in (3.69), (3.70).

The nonmonotonicity seen on the error-versus-CPU time plots of the Gautschi scheme are characteristic for the scheme: smaller time step sizes result in reduction of the Krylov dimension m which makes the scheme significantly cheaper. There is, thus, an optimal time step size for which the overall computational work is minimal. As one can see in Figure 3.12, the Gautschi and Newmark schemes lose to the leap frog scheme in performance. This is to be expected since we work on a uniform mesh in a domain with homogeneous ϵ_r and μ_r.

Because of different sparsity patterns of the matrices $M_\epsilon + \frac{\tau^2}{4} A_\mu$ and M_ϵ the plots versus the computational work in Figure 3.13 are presented only for the leap frog and the Gautschi schemes.

Very similar results were obtained for the case where

$$N_\omega = 101, \quad \omega_i \text{ evenly distributed in } [1, 2],\ i = 1, \ldots, 100, \quad \omega_{101} = 10. \tag{3.74}$$

Here all the schemes yield errors which are approximately a factor 10^3 smaller than for the homogeneous distribution of ω_i (3.73). In this the case the error-versus-τ dependence of the Newmark scheme is monotone.

We now present the performance of the Gautschi scheme on a finer mesh $40 \times 40 \times 40$ with higher, as compared to (3.73) and (3.74), frequencies in the source term:

$$\begin{aligned} N_\omega = 103, \qquad & w_i \text{ evenly distributed in } [1, 2],\ i = 1, \ldots, 100, \\ & \omega_{101} = 10, \quad \omega_{102} = 24, \quad \omega_{103} = 25. \end{aligned} \tag{3.75}$$

In Figure 3.14 the errors at the final time are given against the corresponding step sizes and computational work. For this mesh, the sparse LU factorization of the matrix $M_\epsilon + \frac{\tau^2}{4} A_\mu$ in the Newmark scheme is prohibitively expensive and the conjugate gradient iterative solver is used.

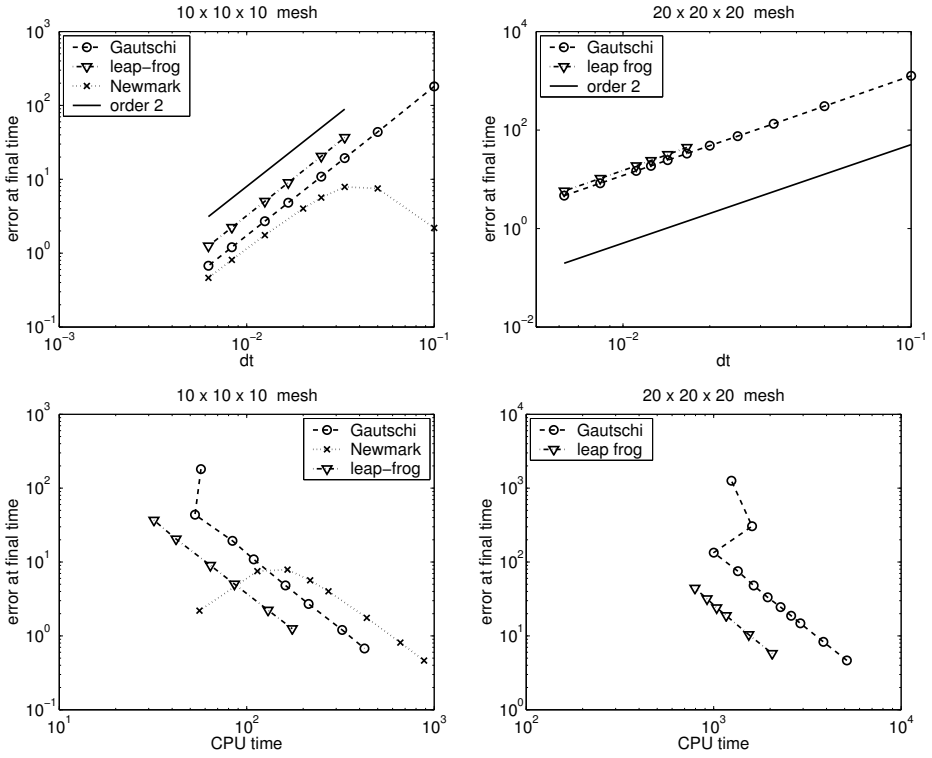


Figure 3.12: Uniform mesh. Errors at the final time against the corresponding step sizes and the required CPU times for the homogeneously distributed frequencies in the source term (cf. (3.73)).

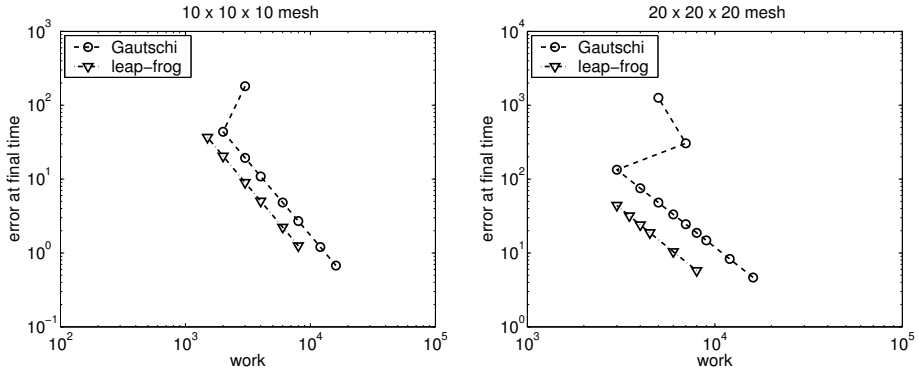


Figure 3.13: Uniform mesh. Errors at the final time against the corresponding computational work for homogeneously distributed frequencies in the source term. The work is measured in the $\mathtt{Q}_{lf}$ units (see Section 3.6.4).

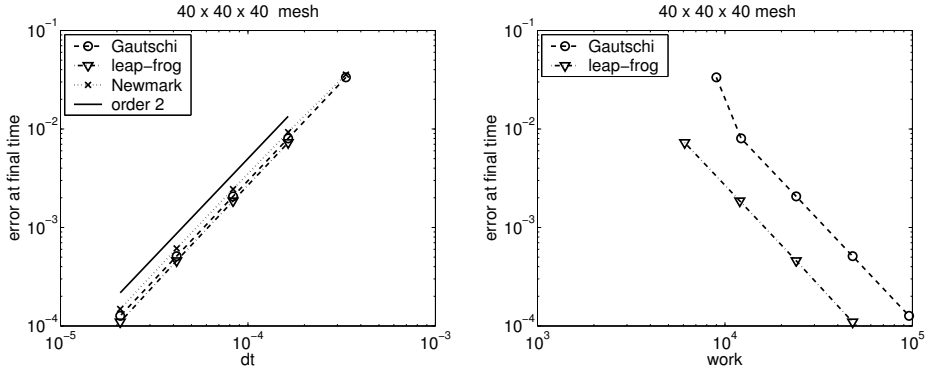


Figure 3.14: Uniform mesh. Errors at the final time against the corresponding step sizes for clustered distribution of frequencies in the source term, see (3.75). The work is measured in the $\mathtt{Q}_{lf}$ units (see Section 3.6.4).

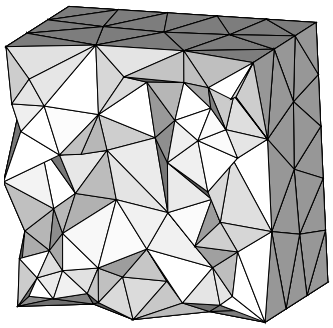


Figure 3.15: A cut of the unstructured mesh used for the experiment.

Unstructured tetrahedral mesh

In this example, Test problem 2 with the homogeneously distributed frequencies in the source term (cf. (3.73)) is solved on a unstructured tetrahedral mesh generated by the Centaur mesh generator. In the mesh used (see Figure 3.15), the ratio between longest and shortest edge is about 17. Although the mesh is rather coarse, the time step of the leap frog scheme is restricted for stability reasons to the relatively small time step 0.0155 (which is approximately a factor two smaller than the stability time step restriction of a uniform mesh with roughly the same number of degrees of freedom).

The results of the experiment are given in Figure 3.16. Note the irregular convergence pattern of the Newmark scheme which is apparently caused by effects of the MATLAB/UMFPACK sparse direct solver used in the scheme (the accuracy of the solver is compromised to retain sparsity in the LU factors).

It is evident that to achieve the same accuracy both the explicit leap frog scheme and the implicit Newmark scheme require much smaller time steps than the Gautschi scheme and their computational times are bigger than that of the Gautschi scheme.

In Figure 3.17 we compare accuracies delivered by the schemes versus required computational work (see Section 3.6.4). It is clear from this figure that on the unstructured mesh the Gautschi scheme appears to be the most efficient.

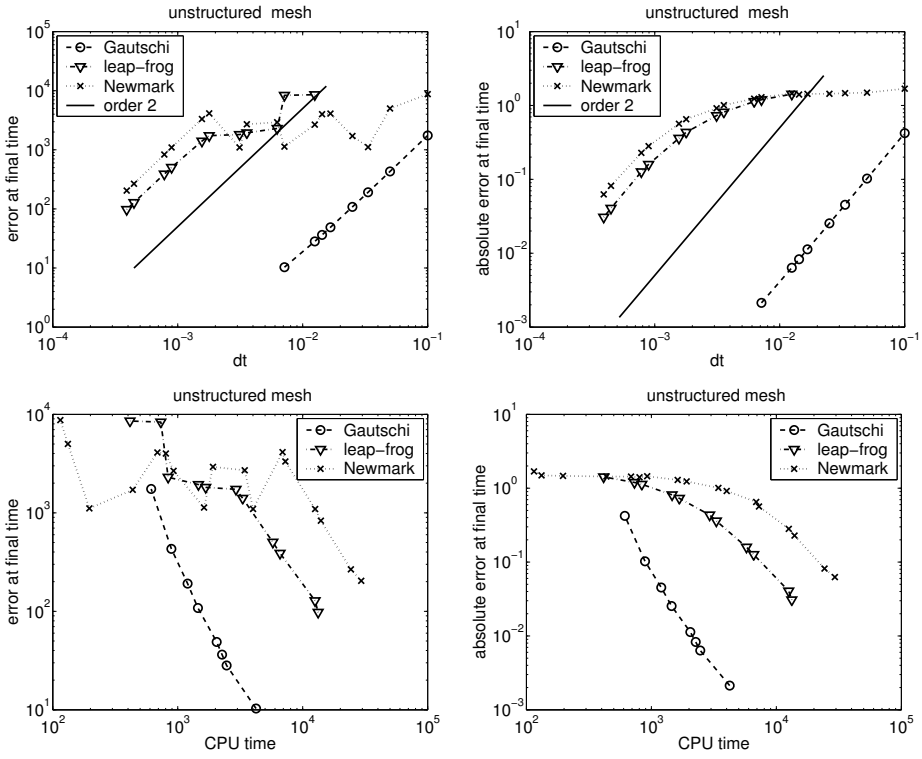


Figure 3.16: Unstructured mesh. Errors at the final time against the corresponding step sizes and the required CPU times for the homogeneously distributed frequencies in the source term (cf. (3.73)). Left plots: the error is measured as in (3.72). Right plots: the error is measured as $\|\boldsymbol{y}^{\bar{n}} - \boldsymbol{y}^{\bar{n}}_{\text{exact}}\|/\|\boldsymbol{y}^{\bar{n}}_{\text{exact}}\|$.

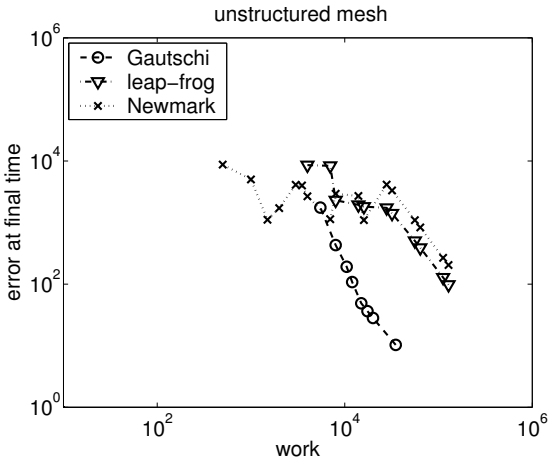


Figure 3.17: Unstructured mesh. Errors at the final time against the corresponding computational work for homogeneously distributed frequencies in the source term. The work is measured in the `Q` units (see Section 3.6.4).

Exactness of the Gautschi scheme for the slowly varying inhomogeneous term

The Gautschi scheme is known to be exact for the constant inhomogeneous term $\boldsymbol{j}(t)$ [47, 59]. To see whether this is the case for our Krylov subspace implementation of the scheme, we take in these two tests (i) zero and (ii) very small values of ω_i:

$$\text{(i)} \quad N_\omega = 1, \qquad \omega_1 = 0, \tag{3.76}$$

$$\text{(ii)} \quad N_\omega = 3, \qquad \omega_1 = 10^{-5},\ \omega_2 = 2.23 \cdot 10^{-5},\ \omega_3 = 8 \cdot 10^{-6}. \tag{3.77}$$

The results obtained on the uniform cubic mesh for zero values of ω_i are presented in Figure 3.18. Similar, practically undistinguishable plots were obtained for the very small frequencies (3.77). Note the superconvergence effects observed for the leap frog and the Newmark schemes on the $10\times10\times10$ mesh: the schemes are almost fourth order accurate. The results clearly show that the Gautschi scheme with adaptive choice of the Krylov subspace dimension is practically exact for these problems.

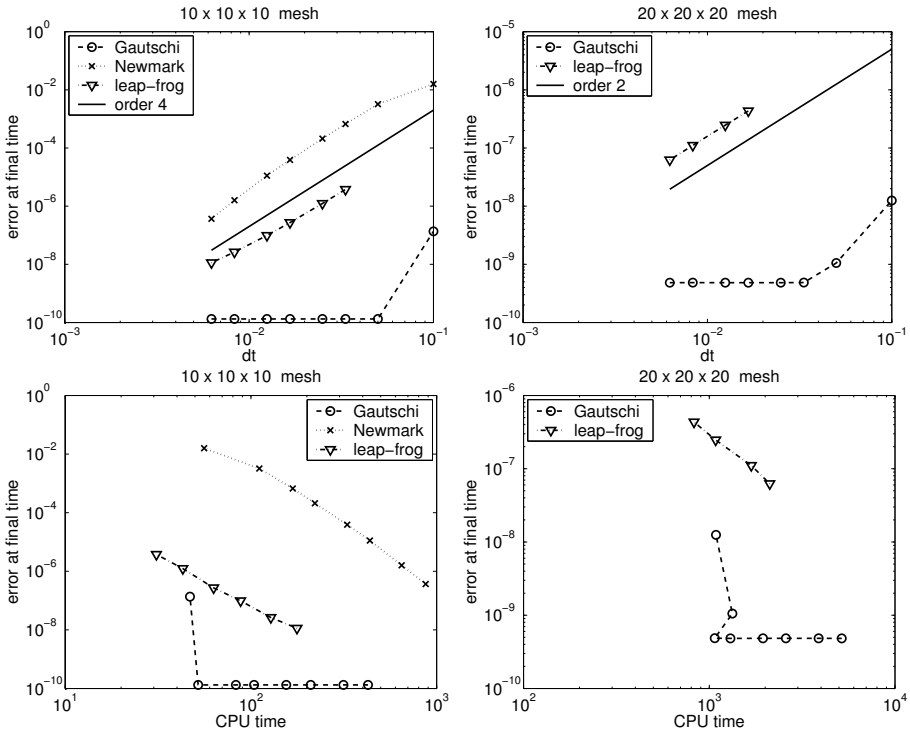


Figure 3.18: Errors at the final time against the corresponding step sizes and the required CPU times for a constant source term (cf. (3.76)).

3.7 Conclusions and suggestions for future research

It is shown that the Gautschi cosine scheme can be efficiently implemented for edge finite element discretizations of the three-dimensional Maxwell equations. The implementation involves a sparse LU (or Cholesky) factorization of the mass matrix which is also required for explicit time stepping schemes and in most cases can be done efficiently. When the direct solution is not feasible the action of the inverse of the mass matrix could also be computed by an iterative solver.

We also proposed a simple strategy for the adaptive choice of the Krylov dimension. This strategy proves to be successful in our experiments, in particular, the error triggered by the Krylov subspace approximation appears negligible to the time error. Moreover, the exactness of the Gautschi scheme for the constant inhomogeneous term was observed in practice for our Gautschi-Krylov implementation. A backward error analysis of the Krylov subspace error was done leading to an explicit formula for the error. This also provided an insight for the stopping criterion used in the Arnoldi process. Furthermore, the stability of the new scheme was proved.

Dispersion analysis presented in this chapter revealed superior properties of the Gautschi scheme as compared to the leap frog and the Newmark scheme.

The presented numerical experiments demonstrate that the Gautschi scheme is more efficient (in terms of the achieved accuracy and the required computational work) than the implicit Newmark scheme. The Gautschi scheme is much more efficient than the explicit leap frog scheme and the Newmark scheme (i) on nonuniform meshes or (ii) when the inhomogeneous source term is a slowly varying function of time.

A relevant future research topic would be an extension of the Gautschi-Krylov scheme to the Maxwell equations with nonzero conductivity terms or absorbing boundary conditions. In both cases the weak formulation (3.14) will contain a first order time derivative. A possible approach here would be to use splitting methods.

It would also be interesting to see how the Gautschi-Krylov scheme performs with the recently developed matrix function preconditioning technique [100].

The presented results indicate that the Gautschi-Krylov scheme is a promising

tool for efficient time integration of the Maxwell equations.

3.8 APPENDIX

3.8.1 Stability of the leap frog scheme

To derive a stability condition for the leap frog scheme we consider the homogeneous case $\boldsymbol{j}(t) = 0$:

$$M_\epsilon \boldsymbol{e}^{n+1} + (\tau^2 A_\mu - 2M_\epsilon)\boldsymbol{e}^n + M_\epsilon \boldsymbol{e}^{n-1} = 0, \tag{3.78}$$

or in its equivalent form

$$\boldsymbol{e}^{n+1} + (\tau^2 M_\epsilon^{-1} A_\mu - 2I)\boldsymbol{e}^n + \boldsymbol{e}^{n-1} = 0. \tag{3.79}$$

In our analysis, we follow the standard approach based on diagonalizing the matrices involved in the scheme (see e.g. [111]). Any solution of (3.79) can be written as

$$\boldsymbol{e}^n = \sum_m \gamma_m^n \boldsymbol{\alpha}_m, \tag{3.80}$$

where $\boldsymbol{\alpha}_m$'s are the eigenvectors corresponding to the eigenvalues (λ_m) of the following eigenvalue problem

$$M_\epsilon^{-1} A_\mu \boldsymbol{x} = \lambda \boldsymbol{x}. \tag{3.81}$$

We assume that matrices M_ϵ and A_μ are Hermitian, M_ϵ is positive definite and A_μ is positive semidefinite. This is guaranteed by the finite element discretization provided that μ and ϵ have corresponding properties. The eigenvalues of (3.81) are then nonnegative. Substitution of (3.80) into (3.79) yields

$$\begin{aligned} &\sum_m \gamma_m^{n+1}\boldsymbol{\alpha}_m + (\tau^2 M_\epsilon^{-1} A_\mu - 2I)\sum_m \gamma_m^n \boldsymbol{\alpha}_m + \sum_m \gamma_m^{n-1}\boldsymbol{\alpha}_m = \\ &\quad = \sum_m \gamma_m^{n+1}\boldsymbol{\alpha}_m + \sum_m \gamma_m^n(\tau^2\lambda_m - 2)\boldsymbol{\alpha}_m + \sum_m \gamma_m^{n-1}\boldsymbol{\alpha}_m = 0. \end{aligned} \tag{3.82}$$

which, due to the linear independence of the $\boldsymbol{\alpha}_m$'s, implies

$$\gamma_m^{n+1} + (\tau^2\lambda_m - 2)\gamma_m^n + \gamma_m^{n-1} = 0, \quad \text{for all } m. \tag{3.83}$$

This recurrence is stable (i. e. $|\gamma_m^n| \leq 1$) if and only if the roots $\nu_{1,2}$ of its characteristic equation

$$\nu^2 + (\tau^2\lambda_m - 2)\nu + 1 = 0 \tag{3.84}$$

do not exceed one in absolute value. The solution of (3.84) is

$$\nu_{1,2} = 1 - \frac{\tau^2}{2}\lambda_m \pm \sqrt{(1 - \frac{\tau^2}{2}\lambda_m)^2 - 1}. \tag{3.85}$$

A straightforward computation shows that the stability condition $|\nu_{1,2}| \leq 1$ is fulfilled if and only if

$$(1 - \frac{\tau^2}{2}\lambda_m)^2 - 1 \leq 0, \tag{3.86}$$

which, together with (3.85), necessarily means that $|\nu_{1,2}| = 1$. The solutions of (3.86) satisfy

$$\tau^2 \leq \frac{4}{\lambda_m}, \quad \text{for all } m, \quad (\lambda_m \geq 0).$$

Then the stability condition for the leap frog scheme is

$$\tau^2 \leq \frac{4}{\lambda_{\max}},$$

where $\lambda_{\max}$ is the maximum eigenvalue of the matrix $M_\epsilon^{-1} A_\mu$.

3.8.2 Dispersion relation matrices F and G

The matrices F and G in (3.59) on a cubic mesh with element size $h \times h \times h$ are given as:
the matrix F is diagonal, with entries

$$\begin{aligned}
F_{11} &= \frac{1}{9}\cos(hk_2)\cos(hk_3) + \frac{2}{9}\cos(k_3 h) + \frac{2}{9}\cos(k_2 h) + \frac{4}{9}, \\
F_{22} &= \frac{1}{9}\cos(hk_1)\cos(hk_3) + \frac{2}{9}\cos(k_3 h) + \frac{2}{9}\cos(k_1 h) + \frac{4}{9}, \\
F_{33} &= \frac{1}{9}\cos(hk_1)\cos(hk_2) + \frac{2}{9}\cos(k_2 h) + \frac{2}{9}\cos(k_1 h) + \frac{4}{9},
\end{aligned}$$

the matrix G is complex Hermitian with entries

$$G = \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ \bar{g}_{12} & g_{22} & g_{23} \\ \bar{g}_{13} & \bar{g}_{23} & g_{33} \end{pmatrix},$$

where $\bar{g}$ denotes the complex conjugate of g and

$$
\begin{aligned}
g_{11} &= \frac{8}{3} - \frac{2}{3}\cos(h(k_2-k_3)) - \frac{2}{3}\cos(hk_2) - \frac{2}{3}\cos(hk_3) - \frac{2}{3}\cos(h(k_2+k_3)),\\
g_{12} &= -\frac{2}{3} + \frac{1}{6}e^{-ih(k_2+k_3)} - \frac{1}{6}e^{-ih(-k_1+k_2+k_3)} - \frac{2}{3}e^{-ih(-k_1+k_2)} +\\
&\quad + \frac{2}{3}e^{-ihk_2} + \frac{1}{6}e^{-ih(-k_1+k_3)} - \frac{1}{6}e^{-ihk_3} - \frac{1}{6}e^{-ih(-k_1+k_2-k_3)} +\\
&\quad + \frac{1}{6}e^{-ih(k_2-k_3)} + \frac{1}{6}e^{ih(k_1+k_3)} - \frac{1}{6}e^{ihk_3} + \frac{2}{3}e^{ik_1h},\\
g_{13} &= -\frac{2}{3} + \frac{1}{6}e^{-ih(k_2+k_3)} - \frac{1}{6}e^{-ih(-k_1+k_2+k_3)} + \frac{1}{6}e^{-ih(-k_1+k_2)} -\\
&\quad - \frac{1}{6}e^{-ihk_2} - \frac{2}{3}e^{-ih(-k_1+k_3)} + \frac{2}{3}e^{-ihk_3} + \frac{1}{6}e^{ih(k_1+k_2)} -\\
&\quad - \frac{1}{6}e^{ih(k_1+k_2-k_3)} + \frac{1}{6}e^{ih(k_2-k_3)} - \frac{1}{6}e^{ihk_2} + \frac{2}{3}e^{ik_1h},\\
g_{22} &= -\frac{2}{3}\cos(k_1h) + \frac{8}{3} - \frac{2}{3}\cos(h(-k_1+k_3)) - \frac{2}{3}\cos(hk_3) - \frac{2}{3}\cos(h(k_1+k_3)),\\
g_{23} &= -\frac{2}{3} - \frac{1}{3}\cos(k_1h) + \frac{1}{6}e^{-ih(-k_1+k_3)} + \frac{2}{3}e^{-ihk_3} + \frac{1}{6}e^{ih(k_1+k_2)} -\\
&\quad - \frac{1}{6}e^{ih(k_1+k_2-k_3)} - \frac{2}{3}e^{ih(k_2-k_3)} + \frac{2}{3}e^{ihk_2} + \frac{1}{6}e^{ih(-k_1+k_2)} -\\
&\quad - \frac{1}{6}e^{ih(-k_1+k_2-k_3)} + \frac{1}{6}e^{-ih(k_1+k_3)},\\
g_{33} &= -\frac{2}{3}\cos(h(k_1+k_2)) + \frac{8}{3} - \frac{2}{3}\cos(k_1h) - \frac{2}{3}\cos(hk_2) - \frac{2}{3}\cos(h(-k_1+k_2)).
\end{aligned}
$$

CHAPTER 4

Implicit a posteriori error estimates for the Maxwell equations

An implicit a posteriori error estimation technique is presented and analyzed for the numerical solution of the time-harmonic Maxwell equations using Nédélec edge elements. For this purpose we define a weak formulation for the error on each element and provide an efficient and accurate numerical solution technique to solve the error equations locally. We investigate the well-posedness of the error equations and also consider the related eigenvalue problem for cubic elements. Numerical results for both smooth and non-smooth problems, including a problem with reentrant corners, show that an accurate prediction is obtained for the local error, and in particular the error distribution, which provides essential information to control an adaptation process. The error estimation technique is also compared with existing methods and provides significantly sharper estimates for a number of reported test cases.

4.1 Introduction

The solution of the Maxwell equations frequently contains structures with limited regularity, such as singularities near corners and non-convex edges. These structures can be efficiently captured using *hp*-adaptive techniques, in which the mesh is locally refined and coarsened (*h*-adaptation) or the polynomial order in individual elements is adjusted (*p*-adaptation). Examples of *hp*-adaptive techniques applied to the Maxwell equations can be found in e.g. [28, 38, 83, 84, 85, 110]. The *hp*-adaptation technique is a promising approach to obtain efficient

numerical algorithms to solve the Maxwell equations, but requires a reasonably accurate estimate of the local error in the numerical solution in order to control the adaptation process. In simple cases one can predict the regions which need to be adapted, but a more general approach requires the use of a posteriori error estimates in which the local error is predicted based on the numerical solution. General techniques for a posteriori error estimation are discussed in e.g. [3, 10, 11, 43, 49, 104], but providing accurate a posteriori error estimates for the Maxwell equations still poses many problems.

In the a posteriori error analysis of the Maxwell equations one encounters two basic problems: the bilinear form of the Maxwell equations is in general not coercive and the analytic solution is not necessarily smooth. Moreover, in real-life situations computations often have to be done in three-dimensional domains of complex geometry (e.g. with reentrant corners) and consisting of different materials (so that the coefficients of the equations are discontinuous). To avoid these difficulties, several studies [15, 16, 80] only investigate a problem defined by a coercive bilinear form. Others, e.g. [73], assume some regularity in the solution of the dual problem.

There are several techniques to obtain a posteriori error bounds for the Maxwell equations. Explicit methods, see e.g. [15, 73], give an error estimate based on the available numerical solution and are relatively easy to implement. The error bounds in explicit methods contain in general unknown coefficients, which also depend on the wave number in the equations, and frequently result in unsharp estimates. Another approach is provided by using a hierarchical basis, see e.g. [3], [12]. This approach has been applied in [16] to the (curl) elliptic Maxwell equations. The analysis of this method is based on some assumptions, such as the saturation assumption ([3], Section 5.2) and the replacement of the bilinear form by an equivalent (localized) bilinear form which ignores coupling terms to obtain a small linear system for the error equations. The validity and effect of these assumptions on the accuracy requires, however, careful attention. Implicit error estimators for the Maxwell equations have been developed in [85], based on the approach in [36], and successfully applied to an *hp* adaptive finite element algorithm for the Maxwell equations in 3D. Lacking a complete analysis, the authors applied an equilibration technique to ensure well-posedness which results in a rather complicated computational procedure.

In this chapter we further investigate the use of implicit error estimators. We follow an approach originally developed for elliptic partial differential equations, see e.g. [3], which when properly formulated, is also applicable to obtain local error estimates for the Maxwell equations. For this purpose, we first define a

weak formulation for the error in each element, which is solved with a finite element method. This is the main difference with explicit a posteriori error estimates which only use the data provided by the numerical solution. We consider the time-harmonic Maxwell equations with perfectly conducting boundaries discretized with Nédélec edge elements, but many ideas can be applied in a more general setting. The main benefit of implicit error estimates is that we do not encounter unknown or very large constants in the a posteriori error estimates. The success of this approach, however, strongly depends on the definition of the boundary conditions for the local error equations and the choice of a proper basis for the numerical solution of the local problems. The latter is achieved by using higher order face and element bubble functions.

The second topic we address are the theoretical properties of the implicit a posteriori error estimation technique. First, we investigate the well posedness of the weak formulation for the local error, in particular in relation to the boundary conditions used in the local error equations. Also, the eigenvalue problem related to the local error equations is investigated in detail. Following the lines in [3] we introduce an error indicator and point out that this provides a lower bound for the exact error and an upper bound for our implicit estimate (up to a constant factor in both cases). These results give the implicit a posteriori error estimation technique a sound theoretical basis and make it possible to avoid some of the involved techniques used in [85, 110].

Instead of giving sharp error estimates we check the preciseness of our method using numerical experiments. We pay special attention to cases where the analytic solution is non-smooth and also investigate the problem in computational domains with reentrant corners. Both the local and global error are estimated in the $H(\mathrm{curl})$ norm and compared with the exact error. We also consider the error distribution, since this is essential information to decide in which areas the mesh has to be adapted. Despite of the expected difficulties we obtain rather precise estimates in each case.

The outline of Chapter 4 is as follows. In Section 4.2 we start with the mathematical formulation and define the finite element discretization. The implicit error estimation technique is formulated and analyzed in Section 4.3. Next, we discuss lower and upper bounds on the error indicator in Section 4.4. The implicit a posteriori error estimation technique is tested numerically on a number of problems of increasing difficulty in Section 4.5. Finally, Section 4.6 contains conclusions and suggestions for future work.

In this chapter we frequently use notations and techniques discussed in the monograph [73]. For a short, self-contained introduction to finite element meth-

ods for the Maxwell equations we refer to [37].

4.2 Mathematical formalization

Consider the time harmonic Maxwell equations for the electric field $\boldsymbol{E} : \Omega \to \mathbb{R}^3$ with perfectly conducting boundary conditions, which are defined as

$$\begin{aligned} \operatorname{curl}\operatorname{curl} \boldsymbol{E} - k^2 \boldsymbol{E} &= \boldsymbol{J} \quad \text{in } \Omega, \\ \boldsymbol{E} \times \boldsymbol{\nu} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where $\Omega \subset \mathbb{R}^3$ is a Lipschitz domain with outward normal vector $\boldsymbol{\nu}$ and $\boldsymbol{J} \in [L_2(\Omega)]^3$ a given source term which is related to the wave number k. Here $k = \frac{\omega}{c}$ with the frequency ω and the speed of light c. Here $\boldsymbol{E} \times \boldsymbol{\nu}$ is defined in a trace sense [24], [73] discussed later.

In the subsequent derivations we will need the following spaces and operators. The Hilbert space corresponding to the Maxwell equation is

$$H(\operatorname{curl}, \Omega) = \{\boldsymbol{u} \in [L_2(\Omega)]^3 : \operatorname{curl} \boldsymbol{u} \in [L_2(\Omega)]^3\},$$

equipped with the curl norm

$$\|\boldsymbol{u}\|_{\operatorname{curl},\Omega} = (\|\boldsymbol{u}\|^2_{[L_2(\Omega)]^3} + \|\operatorname{curl} \boldsymbol{u}\|^2_{[L_2(\Omega)]^3})^{1/2}. \tag{4.2}$$

The differential operators div and curl are understood in a distributional sense. While analyzing (4.1) a subspace of $H(\operatorname{curl}, \Omega)$ is commonly used, namely

$$H_0(\operatorname{curl}, \Omega) = \{\boldsymbol{u} \in H(\operatorname{curl}, \Omega) : \boldsymbol{\nu} \times \boldsymbol{u}|_{\partial\Omega} = 0\}.$$

The above definition of $H_0(\operatorname{curl}, \Omega)$ makes only sense if $\boldsymbol{u}|_{\partial\Omega}$ is well defined and a duality between this trace and the outward normal can be defined. To be more precise we first define for a smooth function $\boldsymbol{v} \in C^\infty(\Omega)$ the operators γ_τ and π_τ with

$$\gamma_\tau \boldsymbol{v} = \boldsymbol{\nu} \times \boldsymbol{v}|_{\partial\Omega} \text{ and } \pi_\tau \boldsymbol{v} = (\boldsymbol{\nu} \times \boldsymbol{v}|_{\partial\Omega}) \times \boldsymbol{\nu}. \tag{4.3}$$

One can prove that γ_τ can be extended to a continuous operator mapping $H(\operatorname{curl}, \Omega)$ into $[H^{-1/2}(\partial\Omega)]^3$. The trace operator π_τ can also be extended to $H(\operatorname{curl}, \Omega)$, however, a natural norm on the range is more involved. For the details we refer to [73], or for a more extensive analysis to [25]. The above definitions and notations will also be used for a subdomain $K \subset \Omega$.

The scalar product in $[L_2(\Omega)]^3$ and $[L_2(K)]^3$ are denoted with $(\cdot, \cdot)$ and $(\cdot, \cdot)_K$, respectively. Similarly, $(\cdot, \cdot)_{\partial\Omega}$ and $(\cdot, \cdot)_{\partial K}$ denote the duality pairing between

the two types of traces on $\partial\Omega$ and ∂K, respectively.

We also recall an appropriate Green's formula: for any $\boldsymbol{u}, \boldsymbol{v} \in H(\mathrm{curl}, \Omega)$ we have the identity

$$(\mathrm{curl}\, \boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{u}, \mathrm{curl}\, \boldsymbol{v}) = (\gamma_\tau \boldsymbol{u}, \pi_\tau \boldsymbol{v})_{\partial\Omega}, \quad \forall\, \boldsymbol{u}, \boldsymbol{v} \in H(\mathrm{curl}, \Omega), \tag{4.4}$$

and a corresponding formula holds on K.

Turning to the variational formulation of (4.1) we introduce the bilinear form

$$B : H(\mathrm{curl}, \Omega) \times H(\mathrm{curl}, \Omega) \to \mathbb{R}$$

with

$$B(\boldsymbol{u}, \boldsymbol{v}) = (\mathrm{curl}\, \boldsymbol{u}, \mathrm{curl}\, \boldsymbol{v}) - k^2(\boldsymbol{u}, \boldsymbol{v}).$$

and B_K will denote the corresponding bilinear form on $H(\mathrm{curl}, K) \times H(\mathrm{curl}, K)$.

Using the above notations the weak formulation of the problem (4.1) is to find $\boldsymbol{E} \in H_0(\mathrm{curl}, \Omega)$ such that

$$(\mathrm{curl}\, \boldsymbol{E}, \mathrm{curl}\, \boldsymbol{v}) - k^2(\boldsymbol{E}, \boldsymbol{v}) = (\boldsymbol{J}, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in H_0(\mathrm{curl}, \Omega). \tag{4.5}$$

4.2.1 Finite elements in $H(\mathbf{curl})$: Edge elements

For the numerical solution of (4.5) we use the $H(\mathrm{curl})$ conforming edge finite element method initiated by Nédélec [76].

The finite element $(\hat{K}, \hat{P}, \hat{\mathcal{A}})$ [22] is defined on the unit cube $\hat{K}$ and an isoparametric mapping ([22], Section 4.7) $D_K : \hat{K} \to K$ is used to define it on a hexahedron K. In general, D_K could be nonlinear but we restrict the analysis to affine maps.

In the numerical experiments we use the lowest order elements

$$\hat{P} = \{\boldsymbol{u} = [u_1, u_2, u_3]^t : u_1 \in Q_{0,1,1}; u_2 \in Q_{1,0,1}; u_3 \in Q_{1,1,0}\}, \tag{4.6}$$

with $Q_{k,l,m}$ the vector space of k, l and m order polynomials with respect to their first, second and third variables, respectively.

It is well known that the covariant transformation preserves line integrals under a change of coordinates [73, 90], so that the basis functions for a given hexahedron K can be defined as

$$\boldsymbol{w}_{j,K}(x, y, z) = \left((\mathrm{d}D_K^{-1})^T \boldsymbol{W}_j^0\right) \circ D_K^{-1}(x, y, z), \tag{4.7}$$

where $\mathrm{d}D_K$ is the Jacobian of the transformation D_K, and $\{\boldsymbol{W}_j^0\}_{j=1}^{12}$ is a basis in 4.6.

Using the transformation in (4.7) one easily computes the curl of the basis functions (see [73], (3.76) and the corresponding statements: Lemma 3.57, Corollary 3.58). A similar transformation for divergence conforming finite elements ([73], (3.77)) is well known and called the Piola transformation ([72], p. 112).

Next, we introduce a hexahedral tessellation $\mathcal{T}_h$ of Ω. The space W_h of Nédélec's edge elements is then defined by

$$W_h = \{\boldsymbol{u}_h \in H(\mathrm{curl}, \Omega) : \boldsymbol{u}_h|_K \in \ \mathrm{span}(w_{j,K}),\ \forall K \in \mathcal{T}_h\},$$

where $\mathrm{span}(w_{j,K})$ denotes the linear hull of $w_{1,K}, w_{2,K}, \ldots, w_{12,K}$ and with this the discretized version of (4.5) reads:

Find $\boldsymbol{E}_h \in W_h$, such that for all $\boldsymbol{W} \in W_h$ the following relation is satisfied:

$$(\mathrm{curl}\, \boldsymbol{E}_h, \mathrm{curl}\, \boldsymbol{W}) - k^2(\boldsymbol{E}_h, \boldsymbol{W}) = (\boldsymbol{J}, \boldsymbol{W}). \tag{4.8}$$

4.3 Implicit error estimation

Providing reliable explicit bounds for the computational error in case of the Maxwell equations is still an unsolved problem due to many difficulties as mentioned in the introduction. The idea of implicit a posteriori error estimation can help to overcome these problems. In this procedure we are not interested in explicit error bounds (depending the data from some existing numerical approximation), we rather formulate a local problem for the error using the available information. This local problem has to be equipped with some meaningful boundary condition and we have to ensure that it is well posed.

4.3.1 Formulation of the local problem

Assume that $\boldsymbol{E}_h$ is a computed numerical solution. Our aim is to estimate the computational error $\boldsymbol{e}_h = (\boldsymbol{E} - \boldsymbol{E}_h)|_K$ on a subdomain K consisting of a set of elements $K \in \mathcal{T}_h$, with $\mathcal{T}_h$ being the finite element tessellation. For this we

state a variational problem for $\boldsymbol{e}_h$ on the subdomain K as follows:

$$\begin{aligned} B_K(\boldsymbol{e}_h, \boldsymbol{v}) &= (\operatorname{curl} \boldsymbol{e}_h, \operatorname{curl} \boldsymbol{v})_K - k^2(\boldsymbol{e}_h, \boldsymbol{v})_K \\ &= (\operatorname{curl} (\boldsymbol{E} - \boldsymbol{E}_h), \operatorname{curl} \boldsymbol{v})_K - k^2(\boldsymbol{E} - \boldsymbol{E}_h, \boldsymbol{v})_K \\ &= (\operatorname{curl} \boldsymbol{E}, \operatorname{curl} \boldsymbol{v})_K - k^2(\boldsymbol{E}, \boldsymbol{v})_K - ((\operatorname{curl} \boldsymbol{E}_h, \operatorname{curl} \boldsymbol{v})_K - k^2(\boldsymbol{E}_h, \boldsymbol{v})_K) \\ &= (\operatorname{curl} \operatorname{curl} \boldsymbol{E}, \boldsymbol{v})_K - (\gamma_\tau \operatorname{curl} \boldsymbol{E}, \pi_\tau \boldsymbol{v})_{\partial K} - k^2(\boldsymbol{E}, \boldsymbol{v})_K - B_K(\boldsymbol{E}_h, \boldsymbol{v})_K \\ &= (\boldsymbol{J}, \boldsymbol{v})_K - (\gamma_\tau \operatorname{curl} \boldsymbol{E}, \pi_\tau \boldsymbol{v})_{\partial K} - B_K(\boldsymbol{E}_h, \boldsymbol{v}), \quad \forall \boldsymbol{v} \in H(\operatorname{curl}, K). \end{aligned} \tag{4.9}$$

In order to get a well defined right hand side we should use an approximation

$$\gamma_\tau \operatorname{curl} \boldsymbol{E} \approx \widehat{\gamma_\tau \operatorname{curl}} \boldsymbol{E} \tag{4.10}$$

on the interelement faces. The quantity $\gamma_\tau \operatorname{curl} \boldsymbol{E}$ will be called the *natural* boundary data from now on. In the literature, the homogeneous natural boundary condition is called the magnetic symmetry wall condition [37]. Introducing this approximation into (4.9) we arrive at the variational problem for the implicit error estimate: Find $\hat{\boldsymbol{e}}_h \in H(\operatorname{curl}, K)$ such that

$$B_K(\hat{\boldsymbol{e}}_h, \boldsymbol{v}) = (\boldsymbol{J}, \boldsymbol{v})_K - (\widehat{\gamma_\tau \operatorname{curl}} \boldsymbol{E}, \pi_\tau \boldsymbol{v})_{\partial K} - B_K(\boldsymbol{E}_h, \boldsymbol{v}), \quad \forall \boldsymbol{v} \in H(\operatorname{curl}, K). \tag{4.11}$$

4.3.2 Numerical solution of the local problem

We will now give a discretized form of the local problem (4.11) which requires a concrete choice for the approximation (4.10) of the natural boundary condition and a finite element basis on the subdomain K.

Approximation of the natural boundary condition

We first specify the approximation in (4.10). For the definition of the boundary condition for the local error equation (4.10) we introduce l_j the common face of the two neighboring elements K and K_j and $\boldsymbol{\nu}_j$ the outward normal on l_j with respect to K. We approximate $\gamma_\tau \operatorname{curl} \boldsymbol{E}$ on l_j with the average of the tangential traces of the numerical approximation $\boldsymbol{E}_h$ on its two sides K and K_j. That is we shall use the approximation

$$\gamma_\tau \operatorname{curl} \boldsymbol{E}|_{l_j} = \boldsymbol{\nu}_j \times \operatorname{curl} \boldsymbol{E}|_{l_j} \approx \frac{1}{2}(\boldsymbol{\nu}_j \times \left[\operatorname{curl} \boldsymbol{E}_h|_K + \operatorname{curl} \boldsymbol{E}_h|_{K_j}\right]) \tag{4.12}$$

which can be straightforwardly implemented.

Choice of the local basis

The local error equation (cf. (4.11)) is solved numerically in a finite-dimensional space which we denote with V_h. As discussed in [3] (see Section 3.4.2 in [3] for several examples), the space V_h has to be selected carefully.

It is known that the finite element solution $\boldsymbol{E}_h$ is a quasioptimal approximation of $\boldsymbol{E}$ within the finite element space W_h. Therefore, it does not make sense to use this space to solve the error equation, we should rather estimate the components of $\hat{\boldsymbol{e}}_h$ which are not present in W_h.

On the other hand, Nédélec type edge elements are related to the electric field strength along the edges. Therefore, to enhance the approximation of the error we should use elements in the local problems which are zero on all edges since the error is mainly non-zero away from the edges. This corresponds to the technique in [23], Section 2.2 for elliptic problems and helps to localize the error equation.

Based on these requirements we use a finite dimensional space which is zero on all edges and for all faces we associate a basis function which is nonzero only on that face. In concrete terms, the finite element space $V_h = \operatorname{span}(\boldsymbol{\phi}_j)$, $j = 1, \ldots, 9$ on each element K is given by

$$\boldsymbol{\phi}_j(x, y, z) = \boldsymbol{\phi}_j^0(\xi, \eta, \zeta) \circ D_K^{-1}(x, y, z), \tag{4.13}$$

and $\hat{V} = \operatorname{span}(\boldsymbol{\phi}_j^0), j = 1, 2, \ldots, 9$, where the face and the element bubble functions on the reference element $\hat{K}$ are

$$\begin{aligned}
&\boldsymbol{\phi}_1^0 = ((1-\xi)(1-\eta)\eta(1-\zeta)\zeta, 0, 0)^T, && \boldsymbol{\phi}_6^0 = (0, 0, (1-\xi)\xi(1-\eta)\eta\zeta)^T,\\
&\boldsymbol{\phi}_2^0 = (\xi(1-\eta)\eta(1-\zeta)\zeta, 0, 0)^T, && \boldsymbol{\phi}_7^0 = ((1-\xi)\xi(1-\eta)\eta(1-\zeta)\zeta, 0, 0)^T,\\
&\boldsymbol{\phi}_3^0 = (0, (1-\xi)\xi(1-\eta)(1-\zeta)\zeta, 0)^T, && \boldsymbol{\phi}_8^0 = (0, (1-\xi)\xi(1-\eta)\eta(1-\zeta)\zeta, 0)^T,\\
&\boldsymbol{\phi}_4^0 = (0, (1-\xi)\xi\eta(1-\zeta)\zeta, 0)^T, && \boldsymbol{\phi}_9^0 = (0, 0, (1-\xi)\xi(1-\eta)\eta(1-\zeta)\zeta)^T\\
&\boldsymbol{\phi}_5^0 = (0, 0, (1-\xi)\xi(1-\eta)\eta(1-\zeta))^T. &&
\end{aligned}$$

Compared to (4.7) the transformation in (4.13) is a minor simplification which results in the same finite element space as the one in (4.7) but makes the computations slightly easier.

The analysis of the local problem given in Appendix 4.7 confirms that this basis results in a well posed problem for the discrete form of the error equation (4.11).

Weak form of the local problem

Using the approximation (4.12) and the local basis V_h we obtain the discrete form of (4.11):

Find $\hat{\boldsymbol{e}}_h \in V_h$ such that $\forall\, \boldsymbol{w} \in V_h$

$$\begin{aligned}(\operatorname{curl} \hat{\boldsymbol{e}}_h, \operatorname{curl} \boldsymbol{w})_K - k^2(\hat{\boldsymbol{e}}_h, \boldsymbol{w})_K = (\boldsymbol{J}, \boldsymbol{w})_K - (\operatorname{curl} \boldsymbol{E}_h, \operatorname{curl} \boldsymbol{w})_K \\ +k^2(\boldsymbol{E}_h, \boldsymbol{w})_K - \frac{1}{2}(\boldsymbol{\nu}_j \times \left(\operatorname{curl} \boldsymbol{E}_h|_K + \operatorname{curl} \boldsymbol{E}_h|_{K_j}\right), \boldsymbol{w})_{\partial K}.\end{aligned} \tag{4.14}$$

4.3.3 Analysis of implicit error estimation

Our objective is to solve the local problems arising in (4.14) for the unknown error term $\hat{\boldsymbol{e}}_h$. In this section, we establish that any reasonable approximation in (4.10) results in a well posed problem in (4.11). Observe that this equation is the variational form of the Maxwell equations equipped with natural boundary data. Note that a similar procedure for elliptic problems results in ill-posed local problems which require some postprocessing to be solvable [3].

Lifting of the local problem

The well-posedness of (4.11) will be investigated for the case of homogeneous natural boundary conditions. To do this we apply the trace lifting $l : \operatorname{Ran}(\gamma_\tau \circ \operatorname{curl}|_{\partial K}) \to H(\operatorname{curl}, K)$ (or equivalently, take an inverse of the tangential trace of the curl operator on $H(\operatorname{curl}, K)$) such that

$$\gamma_\tau(\operatorname{curl} l\boldsymbol{u}) = \boldsymbol{u},\ \forall \boldsymbol{u} \in \operatorname{Ran}(\gamma_\tau \circ \operatorname{curl}|_{\partial K}).$$

Defining now $\bar{\boldsymbol{e}}_h = \hat{\boldsymbol{e}}_h - l(\boldsymbol{\nu}_i \times \operatorname{curl} \hat{\boldsymbol{e}}_h|_{\partial K})$ we can rewrite (4.11) as follows:

$$\begin{aligned}B_K(\bar{\boldsymbol{e}}_h, \boldsymbol{v}) &= B_K(\hat{\boldsymbol{e}}_h, \boldsymbol{v}) - B_K(l(\boldsymbol{\nu}_i \times \operatorname{curl} \hat{\boldsymbol{e}}_h|_{\partial K}), \boldsymbol{v}) \\ &= (\boldsymbol{J}, \boldsymbol{v})_K - (\gamma_\tau \widehat{\operatorname{curl}}\, \boldsymbol{E}, \pi_\tau \boldsymbol{v})_{\partial K} - B_K(\boldsymbol{E}_h, \boldsymbol{v}) \\ &\quad - (\operatorname{curl} \operatorname{curl} l(\gamma_\tau \widehat{\operatorname{curl}}\, \boldsymbol{E}), \boldsymbol{v})_K \\ &\quad + (\gamma_\tau \widehat{\operatorname{curl}}\, \boldsymbol{E}, \pi_\tau \boldsymbol{v})_{\partial K} + k^2(l(\gamma_\tau \widehat{\operatorname{curl}}\, \boldsymbol{E}), \boldsymbol{v}) \\ &= (\boldsymbol{J} - \operatorname{curl} \operatorname{curl} l(\gamma_\tau \widehat{\operatorname{curl}}\, \boldsymbol{E}) + k^2 l(\gamma_\tau \widehat{\operatorname{curl}}\, \boldsymbol{E}), \boldsymbol{v})_K - B_K(\boldsymbol{E}_h, \boldsymbol{v}), \\ &\quad \forall\, \boldsymbol{v} \in H(\operatorname{curl}, K).\end{aligned} \tag{4.15}$$

Observe that on the right hand side we obtained a bounded linear functional of $\boldsymbol{v}$ such that using the Riesz representation theorem we will denote it with

$(\tilde{\boldsymbol{J}}, \boldsymbol{v})_K$. This approach is only necessary for our analysis, we do not need to compute the lifting operator explicitly, since in the finite element procedure the inhomogeneous natural boundary condition can be included in the variational form.

Preliminaries to the well posedness result

In the following two subsections we prove some well posedness results which are formulated in an arbitrary simply connected domain with a Lipschitz boundary. In this way, all of the results can be applied also in the case when local problems will be investigated in the subdomains of the original domain Ω. Before proving the well posedness of the variational problem we state a lemma which will be the cornerstone of our compactness arguments.

Lemma 4.1. *The Hilbert space $H(\mathrm{curl}, \Omega)$ can be decomposed as the direct sum of orthogonal subspaces:*

$$H(\mathrm{curl}, \Omega) = \ker \mathrm{curl} \oplus (\ker \mathrm{curl})^{\perp}, \tag{4.16}$$

and the imbedding of the second component into $[L_2(\Omega)]^3$ is compact.

Proof Using the decomposition theorem in [34], p. 216, and the imbedding in Theorem 2.8 in [5] a standard argument gives the statement of the lemma. For a complete proof we refer to [65]. ■

Well posedness of the local problem

We can state now the well-posedness of (4.15) and prove the following:

Lemma 4.2. *Assume that k is not a Maxwell eigenvalue in the sense that the problem: Find $\boldsymbol{u} \in H(\mathrm{curl}, \Omega)$ such that*

$$B(\boldsymbol{u}, \boldsymbol{v}) = 0, \quad \forall\, \boldsymbol{v} \in H(\mathrm{curl}, \Omega)$$

has only the trivial ($\boldsymbol{u} = 0$) solution. Then the variational problem: Find $\boldsymbol{u} \in H(\mathrm{curl}, \Omega)$ such that

$$B(\boldsymbol{u}, \boldsymbol{v}) = (\tilde{\boldsymbol{J}}, \boldsymbol{v}), \quad \forall\, \boldsymbol{v} \in H(\mathrm{curl}, \Omega) \tag{4.17}$$

has a unique solution for all $\tilde{\boldsymbol{J}} \in [L_2(\Omega)]^3$.

Proof The proof can be carried out using Lemma 4.1 and the technique in [73]. For a complete proof, we refer to [65]. ■

Lemma 4.3. *Assume that for the solution* $\boldsymbol{u}$ *of (4.17)* $\operatorname{curl}\operatorname{curl}\boldsymbol{u} \in [L_2(\Omega)]^3$ *holds. Then* $\boldsymbol{u}$ *solves the Maxwell equation equipped with the natural boundary condition:*

$$\begin{aligned} \operatorname{curl}\operatorname{curl}\boldsymbol{u} - k^2\boldsymbol{u} &= \tilde{\boldsymbol{J}} \quad \text{in } \Omega, \\ \boldsymbol{\nu} \times \operatorname{curl}\boldsymbol{u} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.18}$$

Proof Using the assumption in this lemma and the Green theorem (4.4) for the curl operator we obtain that

$$0 = (\operatorname{curl}\boldsymbol{u}, \operatorname{curl}\boldsymbol{v}) - k^2(\boldsymbol{u}, \boldsymbol{v}) - (\tilde{\boldsymbol{J}}, \boldsymbol{v}) = (\operatorname{curl}\operatorname{curl}\boldsymbol{u}, \boldsymbol{v}) - k^2(\boldsymbol{u}, \boldsymbol{v}) - (\tilde{\boldsymbol{J}}, \boldsymbol{v}).$$

holds for every $\boldsymbol{v} \in \mathcal{D}(\Omega)$ with $\mathcal{D}(\Omega) = \{\boldsymbol{v} \in C^\infty(\Omega) : \boldsymbol{v} \text{ has compact support}\}$. Since the embedding $\mathcal{D}(\Omega) \subset H(\operatorname{curl}, \Omega)$ is dense, $(\operatorname{curl}\operatorname{curl}\boldsymbol{u} - k^2\boldsymbol{u} - \tilde{\boldsymbol{J}}, \boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in H(\operatorname{curl}, \Omega)$ $\operatorname{curl}\operatorname{curl}\boldsymbol{u} - k^2\boldsymbol{u} = \tilde{\boldsymbol{J}}$ holds in $[L_2(\Omega)]^3$.

Using the Green theorem again we obtain that for all $\boldsymbol{v} \in H(\operatorname{curl}, \Omega)$

$$\begin{aligned} 0 &= (\operatorname{curl}\operatorname{curl}\boldsymbol{u} - k^2\boldsymbol{u} - \tilde{\boldsymbol{J}}, \boldsymbol{v}) \\ &= (\operatorname{curl}\boldsymbol{u}, \operatorname{curl}\boldsymbol{v}) + (\gamma_\tau \operatorname{curl}\boldsymbol{u}, \pi_\tau \boldsymbol{v})_{\partial\Omega} - k^2(\boldsymbol{u}, \boldsymbol{v}) - (\tilde{\boldsymbol{J}}, \boldsymbol{v}) \\ &= (\gamma_\tau \operatorname{curl}\boldsymbol{u}, \pi_\tau \boldsymbol{v})_{\partial\Omega}. \end{aligned} \tag{4.19}$$

In this way, it also holds for any $\boldsymbol{v} \in [H^1(\Omega)]^3$, therefore, by the surjectivity of the trace mapping $H^1(\Omega) \to H^{1/2}(\partial\Omega)$ we obtain that $\gamma_\tau \operatorname{curl}\boldsymbol{u} = 0$ in the $[H^{-1/2}(\partial\Omega)]^3$ sense. ■

Remarks:
1. Indeed, the dual space of the π_τ map of $H(\operatorname{curl}, \Omega)$ is the kernel space of the natural boundary trace $\gamma_\tau \circ \operatorname{curl}$. For the details, see [25].
2. As far as we consider a finite element scheme with piecewise polynomial basis functions, the eigenvalue problem in the weak sense (discussed in Lemma 4.3) is also equivalent with the eigenvalue problem for the original equation (4.1).

4.3.4 The eigenvalue problem for the time-harmonic Maxwell equations with natural boundary conditions

The well posedness of the local problems for the error can be guaranteed if we ensure that k is not a Maxwell eigenvalue of these problems.

More specifically, recall that in Section 2, the weak form (4.9) for the local error equation equipped with the natural boundary condition (4.12) results in the variational problem (4.15) which is well posed by Lemma 4.2 if and only if k is not an eigenvalue of the appropriate boundary value problem (4.18). In this

section, we determine the eigenvalues belonging to cubic subdomains. In this way, for any k (given in the original problem (4.1)) we will be able to choose the subdomain K in the tessellation such that the boundary value problem (4.9) (or even (4.11)) on K will be well posed.

First, we reduce the eigenvalue problem such that we can apply some techniques and results which are available for the Laplacian operator.

Lemma 4.4. *If $k \neq 0$ is a Maxwell eigenvalue of the differential operator on the left hand side of (4.18), then its eigenfunction $\boldsymbol{u}$ is in the subspace $\boldsymbol{u} \in (\ker \operatorname{curl})^{\perp}$ and solves the following Helmholtz equation:*

$$\begin{aligned} \Delta \boldsymbol{u} - k^2 \boldsymbol{u} &= 0 \quad \text{in } \Omega, \\ \boldsymbol{\nu} \times \operatorname{curl} \boldsymbol{u} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.20}$$

where the operator Δ is defined componentwise: for $\boldsymbol{v} : \Omega \to \mathbb{R}^3$ with $\boldsymbol{v} = (v_1, v_2, v_3)$, $\Delta \boldsymbol{v} := (\Delta v_1, \Delta v_2, \Delta v_3)$.

Proof Assume that $\boldsymbol{u} = \boldsymbol{u}_1 + \boldsymbol{u}_2$ (according to the decomposition (4.16)) is an eigenfunction of (4.18). Then

$$\operatorname{curl}\operatorname{curl}(\boldsymbol{u}_1 + \boldsymbol{u}_2) - k^2(\boldsymbol{u}_1 + \boldsymbol{u}_2) = \operatorname{curl}\operatorname{curl} \boldsymbol{u}_2 - k^2(\boldsymbol{u}_1 + \boldsymbol{u}_2) = 0. \tag{4.21}$$

Note that the boundary condition $\boldsymbol{\nu} \times \operatorname{curl} \boldsymbol{u} = 0$ implies that $\boldsymbol{\nu} \times \operatorname{curl} \boldsymbol{u}_2 = 0$ and therefore, taking the $[L_2(\Omega)]^3$ scalar product of both sides with $\boldsymbol{u}_1$ and using the orthogonality in (4.16) we obtain

$$\begin{aligned} 0 &= (\operatorname{curl}\operatorname{curl} \boldsymbol{u}_2 - k^2(\boldsymbol{u}_1 + \boldsymbol{u}_2), \boldsymbol{u}_1) \\ &= (\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_1) + (\gamma_\tau \operatorname{curl} \boldsymbol{u}_2, \pi_\tau \boldsymbol{u}_1)_{\partial\Omega} - k^2(\boldsymbol{u}_1, \boldsymbol{u}_1) = -k^2 \|\boldsymbol{u}_1\|^2. \end{aligned}$$

This means that $\boldsymbol{u}_1 = 0$ and using the relation $\operatorname{curlcurl}\boldsymbol{u}_2 = -\Delta \boldsymbol{u}_2 + \operatorname{graddiv}\boldsymbol{u}_2 = -\Delta \boldsymbol{u}_2$ in (4.21) we obtain the statement in the lemma. ■

In the following we use the conditions arising from the fact that $\boldsymbol{u} \in (\ker \operatorname{curl})^{\perp} \subset H_0(\operatorname{div} 0, \Omega)$ (see proof of Lemma 1) and the boundary condition in (4.20):

$$\operatorname{div} \boldsymbol{u} = 0 \quad \text{in } \Omega, \tag{4.22}$$

$$\boldsymbol{u} \cdot \boldsymbol{\nu} = 0 \quad \text{on } \partial\Omega, \tag{4.23}$$

$$\boldsymbol{\nu} \times \operatorname{curl} \boldsymbol{u} = 0 \quad \text{on } \partial\Omega. \tag{4.24}$$

Accordingly, our objective is to find the eigenvalues and eigenfunctions of the

following operator $\mathcal{L}$:

$$\begin{aligned} &\mathcal{L} : [L_2(\Omega)]^3 \to [L_2(\Omega)]^3, \\ \operatorname{Dom} \mathcal{L} &= \{\boldsymbol{u} \in H_0(\operatorname{div} 0, \Omega) : \Delta \boldsymbol{u} \in [L_2(\Omega)]^3 \text{ and } \boldsymbol{\nu} \times \operatorname{curl} \boldsymbol{u} = 0 \quad \text{on } \partial\Omega\}, \\ &\mathcal{L}\boldsymbol{u} = \Delta \boldsymbol{u}. \end{aligned} \tag{4.25}$$

4.3.5 Eigenvalues in a rectangular domain

In this subsection, we investigate the eigenvalue problem on the cube $\Omega = (0, \pi) \times (0, \pi) \times (0, \pi)$. Applying a linear transformation for the eigenfunctions on Ω allows us to solve the eigenvalue problem on rectangular domains.

This result makes it possible to choose the subdomains in the finite element tessellation in such a way that the local problem for the error is well posed on each subdomain.

We present only the results in this section, the proofs are provided in [65].

Theorem 4.5. *The eigenfunctions of $\mathcal{L}$ defined in* (4.25) *are given by:*

$$\boldsymbol{u}(x_1, x_2, x_3) = \begin{pmatrix} C_1 \sin k_1 x_1 \cos k_2 x_2 \cos k_3 x_3 \\ C_2 \cos k_1 x_1 \sin k_2 x_2 \cos k_3 x_3 \\ C_3 \cos k_1 x_1 \cos k_2 x_2 \sin k_3 x_3 \end{pmatrix}, \tag{4.26}$$

for any $k_1, k_2, k_3 \in \mathbb{N}$ with

$$k_1 C_1 + k_2 C_2 + k_3 C_3 = 0 \quad \textit{and} \quad C_1, C_2, C_3 \in \mathbb{R}. \tag{4.27}$$

Based on Theorem 4.5 we can give the eigenfunctions and eigenvalues of the Maxwell eigenvalue problem with homogeneous natural boundary conditions

$$\begin{aligned} \operatorname{curl} \operatorname{curl} \boldsymbol{v} &= k^2 \boldsymbol{v} \quad \text{in } B_{a,b,c}, \\ \boldsymbol{\nu} \times \operatorname{curl} \boldsymbol{v} &= 0 \quad \text{on } \partial B_{a,b,c}, \end{aligned} \tag{4.28}$$

where $B_{a,b,c}$ is a rectangular domain with edge lengths a, b and c, respectively.

Theorem 4.6. *The eigenfunctions of the Maxwell equation (4.28) are*

$$\boldsymbol{v}(x, y, z) = \begin{pmatrix} C_1 \sin \frac{k_1 \pi}{a} x \cos \frac{k_2 \pi}{b} y \cos \frac{k_3 \pi}{c} z \\ C_2 \cos \frac{k_1 \pi}{a} x \sin \frac{k_2 \pi}{b} y \cos \frac{k_3 \pi}{c} z \\ C_3 \cos \frac{k_1 \pi}{a} x \cos \frac{k_2 \pi}{b} y \sin \frac{k_3 \pi}{c} z \end{pmatrix},$$

for any $k_1, k_2, k_3 \in \mathbb{N}$ *with*

$$\frac{k_1\pi}{a}C_1 + \frac{k_2\pi}{b}C_2 + \frac{k_3\pi}{c}C_3 = 0 \quad \textit{and} \quad C_1, C_2, C_3 \in \mathbb{R}$$

and the appropriate eigenvalues are $k^2 = \left(\frac{k_1\pi}{a}\right)^2 + \left(\frac{k_2\pi}{b}\right)^2 + \left(\frac{k_3\pi}{c}\right)^2$ *with* $k_1, k_2, k_3 \in \mathbb{N}$ *arbitrary.*

4.4 Implicit error estimate as a lower bound of the error

If the a posteriori error estimates have to be used in an adaptation procedure we also have to investigate a lower bound for the exact error. This will ensure that we do not get a pessimistic overestimate of the actual error when the mesh size is reduced. For the estimates in this section we will define an *error indicator* η_K on K. Our analysis consists of two steps: first, we point out that the implicit a posteriori error estimate $\hat{\boldsymbol{e}}_h$ discussed in this chapter provides a lower bound estimate of η_K (up to a certain factor). Second, we verify that the error indicator can also be used as a lower bound (up to a certain factor and some computable remainders) of the exact error on a patch of the subdomain K. The proof is based on the approach in [3]. While the second step is a rather straightforward modification of the original proof for elliptic problems, the first one needs a more careful analysis since the bilinear form B in the variational problem is not coercive.

Since the mappings D_K are affine, the subdomains in the tesselation $\mathcal{T}_h$ consist of parallelepipeds. Moreover, we assume that the mesh is non-degenerated, i.e. the ratio of the diameter of elements and their minimal edge length $\frac{\operatorname{diam} K}{\min |e_K|}$ is bounded. An important consequence of this assumption is that there are constants $K_1, K_2 \in \mathbb{R}^+$ such that for any K with $\max |e_k| < h$ we have

$$K_1 h \leq \min |\operatorname{eig} \mathrm{d}D_K| \leq \max |\operatorname{eig} \mathrm{d}D_K| \leq K_2 h, \tag{4.29}$$

where eig denotes the spectrum. Since the mapping D_K is affine, $\mathrm{d}D_K$ is the linear part of D_K.

For solving the local problem for the error we use the finite dimensional space V_h on K. This choice should lead to a well posed local problem. A necessary condition for this is given in Lemma 4.12.

In the finite element discretization we use the notation $\boldsymbol{\nu} \times \cdot$ instead of γ_τ, and similarly, for functions $\boldsymbol{v} \in [H^1(K)]^3$ we may omit the operator π_τ on the

boundary making a closer link to the numerical procedure. For the consecutive computations we also recall a Green's formula according to (4.4) which states that for all $\boldsymbol{u} \in H(\mathrm{curl}, \Omega)$ and $\boldsymbol{v} \in [H^1(\Omega)]^3$ the following identity holds (Theorem 3.29 in [73]):

$$(\mathrm{curl}\, \boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{u}, \mathrm{curl}\, \boldsymbol{v}) = (\boldsymbol{\nu} \times \boldsymbol{u}, \boldsymbol{v})_{\partial\Omega}. \tag{4.30}$$

Using (4.9) and the approximation (4.12) we obtain the weak form for the error estimate $\hat{\boldsymbol{e}}_h$ on the bubble function space V_h (introduced in Section 4.3.2) as follows: Find an $\hat{\boldsymbol{e}}_h \in V_h$, such that $\forall \boldsymbol{v} \in V_h$ the following relation is satisfied:

$$\begin{aligned}
B_K(\hat{\boldsymbol{e}}_h, \boldsymbol{v}) &= (\boldsymbol{J}, \boldsymbol{v})_K - B_K(\boldsymbol{E}_h, \boldsymbol{v}) - \frac{1}{2}\sum_j (\boldsymbol{\nu}_j \times \left[\mathrm{curl}\, \boldsymbol{E}_h|_K + \mathrm{curl}\, \boldsymbol{E}_h|_{K_j}\right], \boldsymbol{v})_{l_j} \\
&= (\boldsymbol{J}, \boldsymbol{v})_K - (\mathrm{curl}\,\mathrm{curl}\, \boldsymbol{E}_h, \boldsymbol{v})_K + k^2(\boldsymbol{E}_h, \boldsymbol{v})_K + (\boldsymbol{\nu} \times \mathrm{curl}\, \boldsymbol{E}_h, \boldsymbol{v})_{\partial K} \\
&- \frac{1}{2}\sum_j (\boldsymbol{\nu}_j \times \left[\mathrm{curl}\, \boldsymbol{E}_h|_K + \mathrm{curl}\, \boldsymbol{E}_h|_{K_j}\right], \boldsymbol{v})_{l_j} \\
&= (\boldsymbol{J}, \boldsymbol{v})_K - (\mathrm{curl}\,\mathrm{curl}\, \boldsymbol{E}_h, \boldsymbol{v})_K + k^2(\boldsymbol{E}_h, \boldsymbol{v})_K \\
&+ \frac{1}{2}\sum_j (\boldsymbol{\nu}_j \times \left[\mathrm{curl}\, \boldsymbol{E}_h|_K - \mathrm{curl}\, \boldsymbol{E}_h|_{K_j}\right], \boldsymbol{v})_{l_j} \\
&= (\boldsymbol{r}_K, \boldsymbol{v})_K + \sum_j (\mathbf{R}_{l_j}, \boldsymbol{v})_{l_j},
\end{aligned} \tag{4.31}$$

where we have used the notations

$$\boldsymbol{r}_K = \boldsymbol{J} - \mathrm{curl}\,\mathrm{curl}\, \boldsymbol{E}_h + k^2 \boldsymbol{E}_h \quad \text{in } K$$

for the residual within the subdomain K and

$$\mathbf{R}_{l_j} = \frac{1}{2}(\boldsymbol{\nu}_j \times \left[\mathrm{curl}\, \boldsymbol{E}_h|_K - \mathrm{curl}\, \boldsymbol{E}_h|_{K_j}\right])$$

for the tangential jump of the curl at l_j within the subdomain K and $\hat{\boldsymbol{e}}_h$ denotes the desired implicit error estimate. Note that (4.31) gives a variational form for $\hat{\boldsymbol{e}}_h$ which includes the approximation in (4.12). In the following we drop the subscript for the residual and the tangential jump, respectively, the localization will be shown when taking the norm (or some bilinear map) of these quantities.

Using the above quantities $\boldsymbol{r}$ and $\mathbf{R}$ we define η_K as error indicator as follows:

$$\eta_K = (h^2 \|\boldsymbol{r}\|^2_{[L_2(K)]^3} + h\|\mathbf{R}\|^2_{[L_2(\partial K)]^3})^{\frac{1}{2}}.$$

4.4.1 Bubble functions

For the analysis we use the bubble function technique outlined in [3] and recall some basic definitions:

Definition 4.7. *Let* $\hat{\Psi} : \hat{K} \to \mathbb{R}$ *be given by*

$$\hat{\Psi}(\xi, \eta, \zeta) = \xi(1-\xi)\eta(1-\eta)\zeta(1-\zeta)$$

and $\Psi : K \to \mathbb{R}$ *defined using the isoparametric mapping* D_K*. Similarly, for a given* $\hat{\boldsymbol{p}} \in \hat{P}$ *where* $\hat{P} \subset [\mathcal{P}^1(\hat{K})]^3$*, with* $\mathcal{P}^1(\hat{K})$ *a finite dimensional space of polynomials, an appropriate* $\boldsymbol{p} \in P \subset [\mathcal{P}^1(K)]^3$ *can be defined by* $\boldsymbol{p}(\boldsymbol{x}) = \hat{\boldsymbol{p}}(D_K^{-1}(\boldsymbol{x}))$*.*

In the error estimation process we substitute $\boldsymbol{v} \in P$ with $\Psi\boldsymbol{v}$ on K, where $\Psi\boldsymbol{v}$ can be extended by continuity to ∂K. The following lemmas ensure that the multiplication with Ψ does not influence the magnitude of the L_2 or the curl norm compared to that of $\boldsymbol{v}$. The proofs are based on scaling arguments.

Lemma 4.8. *Consider a non-degenerate family* $\mathcal{T}_h$ *of parallelepiped meshes on* Ω*. Then there exists a positive constant* C *such that for all subdomains* $K \subset \mathcal{T}_h$ *and* $\boldsymbol{p} \in P$

$$C^{-1}\|\boldsymbol{p}\|^2_{[L_2(K)]^3} \leq (\Psi\boldsymbol{p}, \boldsymbol{p})_K \leq C\|\boldsymbol{p}\|^2_{[L_2(K)]^3} \tag{4.32}$$

and

$$C^{-1}\|\boldsymbol{p}\|^2_{[L_2(K)]^3} \leq \|\Psi\boldsymbol{p}\|^2_{[L_2(K)]^3} + h^2\|\operatorname{curl}\Psi\boldsymbol{p}\|^2_{[L_2(K)]^3} \leq C\|\boldsymbol{p}\|^2_{[L_2(K)]^3}, \tag{4.33}$$

with h *the diameter of* K*.*

Definition 4.9. *According to [3] p. 24 we define the face bubble functions* $\hat{\Phi}_j : \hat{K} \to \mathbb{R}, j = 1, \ldots, 6$*, which vanish on all edges of* $\hat{K}$ *and all faces* $\hat{l}_m$*,* $m = 1, \ldots, 6$*,* $m \neq j$*.*
Using the mapping $D_K : \hat{K} \to K$ *we define* $\Phi_j, j = 1, \ldots, 6$*, the bubble functions associated with the faces* l_j *of* $K \in \mathcal{T}_h$*.*

We localize a face bubble function Φ_i, associated to the face l_i, by restricting it to the subdomain $\tilde{K}_i = K \cup l_i \cup K_i$, where l_i is the common face of K and K_i. For the consecutive estimates we again provide some inequalities:

Lemma 4.10. *Let us consider a non-degenerate family* $\mathcal{T}_h$ *of parallepiped elements on* Ω*. Then there exists a positive constant* C *such that for all subdomains* $K \subset \mathcal{T}_h$*, functions* $\boldsymbol{p} \in P$ *(where* P *is a fixed finite dimensional subspace of* $[L_2(K)]^3$*) and faces* $l_i, i = 1, \ldots, 6$ *the following inequalities hold:*

$$C^{-1}\|\boldsymbol{p}\|^2_{[L_2(l_i)]^3} \leq \int_{l_i} \Phi_i\boldsymbol{p} \cdot \boldsymbol{p} \leq C\|\boldsymbol{p}\|^2_{[L_2(l_i)]^3}, \tag{4.34}$$

$$\|\Phi_i \boldsymbol{p}\|^2_{[L_2(\tilde{K}_i)]^3} \leq Ch\|\boldsymbol{p}\|^2_{[L_2(l_i)]^3} \tag{4.35}$$

$$\|\Phi_i \boldsymbol{p}\|^2_{[L_2(l_i)]^3} \leq C\|\boldsymbol{p}\|^2_{[L_2(l_i)]^3} \tag{4.36}$$

$$\|\Phi_i \boldsymbol{p}\|^2_{\mathrm{curl},\tilde{K}_i} \leq Ch^{-1}\|\boldsymbol{p}\|^2_{[L_2(l_i)]^3}. \tag{4.37}$$

Proof The proof can be carried out again using scaling arguments similarly to Theorem 2.4 in [3] using that the mesh is non-degenerate. ■

We also state a lemma on the comparison of the different norms in the finite element spaces when they are given on a scale of cubes.

Lemma 4.11. *For all subdomains $K \subset \mathcal{T}_h$ (with the faces $l_i, i = 1, \ldots, 6$) and for all $\boldsymbol{v} \in V_h$ we have the estimates*

$$\|\boldsymbol{v}_k\|_{[L_2(K)]^3} \leq Ch\|\mathrm{curl}\, \boldsymbol{v}_k\|_{[L_2(K)]^3} \tag{4.38}$$

and

$$\|\boldsymbol{v}_k\|_{[L_2(l_i)]^3} \leq Ch^{\frac{1}{2}}\|\mathrm{curl}\, \boldsymbol{v}_k\|_{[L_2(K)]^3} \tag{4.39}$$

with a constant C which does not depend on h.

Proof The proof can be carried out with a change of variables according to D_K. ■

4.4.2 Lower bound for the computational error in terms of the residuals

When the bilinear form B is restricted to $\hat{V} \times \hat{V}$ we may identify it with the stiffness matrix B_1, which is given as $B_1 = B_{1,\mathrm{curl}} - k^2 B_{1,0}$ such that the (i,j)th entries $i, j = 1, 2, \ldots, n$ are as follows:

$$B_{1,\mathrm{curl}}[i][j] = (\mathrm{curl}\, \hat{\Phi}^*_i, \mathrm{curl}\, \hat{\Phi}^*_j)_{\hat{K}}, \qquad B_{1,0}[i][j] = (\hat{\Phi}^*_i, \hat{\Phi}^*_j)_{\hat{K}}$$

with a basis $\{\hat{\Phi}^*_i\}_{i=1}^n$ of $\hat{V}$. $B_{1,K}$ denotes the appropriate mass matrix on K which can be again decomposed as

$$B_{1,K} = B_{1,\mathrm{curl},K} - k^2 B_{1,0,K},$$

where the components are defined as

$$B_{1,\mathrm{curl},K}(\Phi^*_i, \Phi^*_j) = (\mathrm{curl}\, \Phi^*_i, \mathrm{curl}\, \Phi^*_j)_K, \quad \text{and} \quad B_{1,0,K}(\Phi^*_i, \Phi^*_j) = (\Phi^*_i, \Phi^*_j)_K$$

with an appropriate basis $\{\Phi^*_i\}_{i=1}^n$ of V_h.

We can state now the following lemma, which is central in our analysis.

Lemma 4.12. *Assume that $\mathcal{T}_h$ is a cubic mesh, then for a sufficiently fine mesh the bilinear form B_K satisfies the discrete inf-sup condition uniformly on $V_h \times V_h$, namely there is positive h_0 and a constant $C_0 > 0$ such that for any $0 < h < h_0$*

$$C_0 \sup_{\boldsymbol{v} \in V_h} \frac{B_K(\boldsymbol{u}, \boldsymbol{v})}{\|\boldsymbol{v}\|_{\mathrm{curl},K}} \geq \|\boldsymbol{u}\|_{\mathrm{curl},K}, \quad \forall\, \boldsymbol{u} \in V_h, \tag{4.40}$$

where K is the cube with edge length h and C_0 does not depend on h.

The proof of Lemma 4.12 is given in Appendix 4.7.

Lemma 4.13. *If we choose the bubble function space V_h and consider a cubic tesselation then the bilinear form for the error on $K = (0, h)^3$ reads as $B_{1,K} = \frac{1}{h} B_{1,curl} - k^2 h B_{1,0}$ and the mesh size h_0 can be taken as*

$$h_0 = \frac{20(16 - \sqrt{246})}{\sqrt{2k^2 + 1}(16 + \sqrt{246})}$$

in Lemma 4.12.

It is easy to see that B_K is continuous on the whole of $H(\mathrm{curl}, K) \times H(\mathrm{curl}, K)$ as stated in the following lemma.

Lemma 4.14. *For the bilinear form B_K we have the continuity estimate:*

$$|B_K(\boldsymbol{u}, \boldsymbol{v})| \leq \sqrt{2}(1 + k^2)\|\boldsymbol{u}\|_{\mathrm{curl},K}\|\boldsymbol{v}\|_{\mathrm{curl},K} \quad \forall\, \boldsymbol{u}, \boldsymbol{v} \in H(\mathrm{curl}, K). \tag{4.41}$$

Proof We prove the lemma with a straightforward computation as follows:

$$\begin{aligned} |B_K(\boldsymbol{u}, \boldsymbol{v})|^2 &= |(\mathrm{curl}\, \boldsymbol{u}, \mathrm{curl}\, \boldsymbol{v})_K - k^2(\boldsymbol{u}, \boldsymbol{v})_K|^2 \\ &\leq 2(\mathrm{curl}\, \boldsymbol{u}, \mathrm{curl}\, \boldsymbol{v})_K^2 + 2k^4(\boldsymbol{u}, \boldsymbol{v})_K^2 \\ &\leq 2\|\mathrm{curl}\, \boldsymbol{u}\|^2_{[L_2(K)]^3}\|\mathrm{curl}\, \boldsymbol{v}\|^2_{[L_2(K)]^3} + 2k^4\|\boldsymbol{u}\|^2_{[L_2(K)]^3}\|\boldsymbol{v}\|^2_{[L_2(K)]^3} \\ &\leq 2(1 + k^4)\|\boldsymbol{u}\|^2_{\mathrm{curl},K}\|\boldsymbol{v}\|^2_{\mathrm{curl},K}. \end{aligned}$$

Taking the square roots on both sides we obtain the estimate in the lemma. ■

We now compare the error indicator η_K with the implicit error estimate $\hat{\boldsymbol{e}}_h$ obtained from the weak form (4.31) and state the following lemma.

Lemma 4.15. *There is a constant C_1 independently from the mesh parameter h such that*

$$\|\hat{\boldsymbol{e}}_h\|_{\mathrm{curl},K} \leq C_1 \eta_K, \tag{4.42}$$

where $\hat{\boldsymbol{e}}_h$ is the implicit error estimate on the subdomain K.

Proof Using the weak form (4.31) we obtain

$$\begin{aligned}\|\hat{\boldsymbol{e}}_h\|_{\mathrm{curl},K} &\le C_0 \sup_{\boldsymbol{v}\in V_h} \frac{B_K(\hat{\boldsymbol{e}}_h,\boldsymbol{v})}{\|\boldsymbol{v}\|_{\mathrm{curl},K}} = C_0 \sup_{\boldsymbol{v}\in V_h} \frac{(\boldsymbol{r},\boldsymbol{v})_K + (\mathbf{R},\boldsymbol{v})_{\partial K}}{\|\boldsymbol{v}\|_{\mathrm{curl},K}} \\ &\le C_0 \sup_{\boldsymbol{v}\in V_h} \frac{1}{\|\boldsymbol{v}\|_{\mathrm{curl},K}} \big(\|\boldsymbol{r}\|_{[L_2(K)]^3}\|\boldsymbol{v}\|_{[L_2(K)]^3} + \\ &\|\mathbf{R}\|_{[L_2(\partial K)]^3}\|\boldsymbol{v}\|_{[L_2(\partial K)]^3}\big) \\ &\le C\cdot C_0 h\|\boldsymbol{r}\|_{[L_2(K)]^3} + h^{\frac{1}{2}}\|\mathbf{R}\|_{[L_2(\partial K)]^3} \le C\cdot C_0\sqrt{2}\eta_K\end{aligned} \tag{4.43}$$

In the first inequality we used (4.40), then the weak formulation (4.31) followed by the Cauchy-Schwarz inequality. Finally, we applied the estimates (4.38) and (4.39) and a basic inequality. ■

Note that the error estimate $\hat{\boldsymbol{e}}_h$ in Lemma 4.15 gives the exact error (according to the weak form (4.31)) assuming that the boundary condition (4.12) is exact. In the following we will compute an approximation of $\hat{\boldsymbol{e}}_h$ in the finite element space V_h.

The upper estimate of the error indicator η_K will be obtained using the bubble function technique [3]. We first provide a variational form for the exact error $\boldsymbol{e}_h$ based on the third line of (4.9) and using the notations of the previous sections.

$$\begin{aligned}B(\boldsymbol{e}_h,\boldsymbol{v}) &= (\boldsymbol{J},\boldsymbol{v}) - ((\operatorname{curl}\boldsymbol{E}_h,\operatorname{curl}\boldsymbol{v}) - k^2(\boldsymbol{E}_h,\boldsymbol{v})) \\ &= \sum_{K\in\mathcal{T}_h}\{(\boldsymbol{J},\boldsymbol{v})_K - (\operatorname{curl}\operatorname{curl}\boldsymbol{E}_h - k^2\boldsymbol{E}_h,\boldsymbol{v})_K \\ &+ \sum_j (\boldsymbol{\nu}_j\times\operatorname{curl}\boldsymbol{E}_h,\pi_\tau\boldsymbol{v})_{l_j}\} \\ &= \sum_{K\in\mathcal{T}_h}(\boldsymbol{r},\boldsymbol{v})_K + \sum_\gamma(\mathbf{R},\boldsymbol{v})_\gamma \quad \forall\,\boldsymbol{v}\in H(\operatorname{curl},\Omega),\end{aligned} \tag{4.44}$$

where the last sum is taken over all of the interelement faces γ inside of Ω. To obtain the second line in (4.44) we used the perfect conducting boundary condition on $\partial\Omega$, while the final expression was obtained by summation of the components of a given face from the both sides. This expression can also be related with (4.31); on the whole domain Ω, the variational form for the exact and the estimated error coincide since the boundary conditions are given.

We will choose $\boldsymbol{v}$ in (4.44) using the bubble function technique such that the boundary integral vanishes, which will result in a lower bound for the error on each subdomain depending only on the element residual $\boldsymbol{r}$ in the subdomain K.

It is also important that the choice for $\boldsymbol{v}$ is suitable for the estimates (4.32)-(4.33), which only apply in a finite dimensional space. In light of this, we denote with $\bar{\boldsymbol{r}}$ the elementwise interpolation of the residual $\boldsymbol{r}$ using the function space V_h and choose $\boldsymbol{v} = \Psi_K \bar{\boldsymbol{r}}$ on each subdomain K and zero elsewhere. Inserting this choice for $\boldsymbol{v}$ into (4.44) gives that

$$B(\boldsymbol{e}_h, \Psi_K \bar{\boldsymbol{r}}) = (\boldsymbol{r}, \Psi_K \bar{\boldsymbol{r}})_K. \tag{4.45}$$

In the following estimates we use C for a generic constant independent of the mesh size h and frequency k, which can be different in each formula. Using (4.45), and inequalities (4.32), (4.33) for $h \leq 1$ we obtain the following estimate:

$$\begin{aligned} \|\bar{\boldsymbol{r}}\|^2_{[L_2(K)]^3} &\leq C(\Psi_K \bar{\boldsymbol{r}}, \bar{\boldsymbol{r}})_K = C\left((\Psi_K \bar{\boldsymbol{r}}, \bar{\boldsymbol{r}} - \boldsymbol{r})_K + B(\boldsymbol{e}_h, \Psi_K \bar{\boldsymbol{r}})\right) \\ &\leq C(\|\Psi_K \bar{\boldsymbol{r}}\|_{[L_2(K)]^3} \|\bar{\boldsymbol{r}} - \boldsymbol{r}\|_{[L_2(K)]^3} \\ &+ (1 + k^2)\|\boldsymbol{e}_h\|_{\mathrm{curl},K} \|\Psi_K \bar{\boldsymbol{r}}\|_{\mathrm{curl},K}) \\ &\leq C(\|\bar{\boldsymbol{r}}\|_{[L_2(K)]^3} \|\bar{\boldsymbol{r}} - \boldsymbol{r}\|_{[L_2(K)]^3} \\ &+ (1 + k^2)h^{-1}\|\boldsymbol{e}_h\|_{\mathrm{curl},K}) \|\bar{\boldsymbol{r}}\|_{[L_2(K)]^3}. \end{aligned} \tag{4.46}$$

Dividing by $\|\bar{\boldsymbol{r}}\|_{[L_2(K_i)]^3}$ and using the triangle inequality we finally obtain that

$$\begin{aligned} \|\boldsymbol{r}\|_{[L_2(K)]^3} &\leq \|\bar{\boldsymbol{r}}\|_{[L_2(K)]^3} + \|\boldsymbol{r} - \bar{\boldsymbol{r}}\|_{[L_2(K)]^3} \\ &\leq C(\|\bar{\boldsymbol{r}} - \boldsymbol{r}\|_{[L_2(K)]^3} + (1 + k^2)h^{-1}\|\boldsymbol{e}_h\|_{\mathrm{curl},K}). \end{aligned} \tag{4.47}$$

We proceed similarly for the boundary jumps and denote with $\bar{\mathbf{R}}$ the approximation of the boundary jump using the trace of V_h on the element boundaries, which is then defined on the interelement faces l_i. The error arising from these terms can be localized on $\tilde{K}$ by the choice:

$$\boldsymbol{v} = \Phi_l \bar{\mathbf{R}},$$

associated to the face l as in Lemma 4.10 which is again extended (preserving the continuity) to be zero outside of $\tilde{K}$. This leads us to the identity

$$(\Phi_l \bar{\mathbf{R}}, \mathbf{R})_l = B_{\tilde{K}}(\boldsymbol{e}_h, \Phi_l \bar{\mathbf{R}}) - (\Phi_l \bar{\mathbf{R}}, \boldsymbol{r})_{\tilde{K}}. \tag{4.48}$$

Using then (4.48), the Cauchy–Schwarz inequality and inequalities (4.34), (4.35),

(4.36), and (4.37) we obtain the following estimate

$$\begin{aligned}\|\bar{\mathbf{R}}\|^2_{[L_2(l)]^3} &\leq C(\Phi_l\bar{\mathbf{R}},\bar{\mathbf{R}})_l = C(\Phi_l\bar{\mathbf{R}},\bar{\mathbf{R}}-\mathbf{R})_l + C(\Phi_l\bar{\mathbf{R}},\mathbf{R})_l\\ &= C((\Phi_l\bar{\mathbf{R}},\bar{\mathbf{R}}-\mathbf{R})_l + B_{\tilde{K}}(\boldsymbol{e}_h,\Phi_l\bar{\mathbf{R}}) - (\Phi_l\bar{\mathbf{R}},\boldsymbol{r})_{\tilde{K}})\\ &\leq C(\|\Phi_l\bar{\mathbf{R}}\|_{[L_2(l)]^3}\|\bar{\mathbf{R}}-\mathbf{R}\|_{[L_2(l)]^3}\\ &\quad + \sqrt{2}(1+k^2)\|\boldsymbol{e}_h\|_{\mathrm{curl},\tilde{K}}\|\Phi_l\bar{\mathbf{R}}\|_{\mathrm{curl},\tilde{K}} + \|\Phi_l\bar{\mathbf{R}}\|_{[L_2(\tilde{K})]^3}\|\boldsymbol{r}\|_{[L_2(\tilde{K})]^3})\\ &\leq C(\|\bar{\mathbf{R}}\|_{[L_2(l)]^3}\|\bar{\mathbf{R}}-\mathbf{R}\|_{[L_2(l)]^3}\\ &\quad + h^{-\frac{1}{2}}(1+k^2)\|\boldsymbol{e}_h\|_{\mathrm{curl},\tilde{K}}\|\bar{\mathbf{R}}\|_{[L_2(l)]^3} + h^{\frac{1}{2}}\|\bar{\mathbf{R}}\|_{[L_2(l)]^3}\|\boldsymbol{r}\|_{[L_2(\tilde{K})]^3}),\end{aligned} \tag{4.49}$$

which after division by $\|\bar{\mathbf{R}}\|_{[L_2(l)]^3}$, yields

$$\|\bar{\mathbf{R}}\|_{[L_2(l)]^3} \leq C(\|\bar{\mathbf{R}}-\mathbf{R}\|_{[L_2(l)]^3} + h^{-\frac{1}{2}}(1+k^2)\|\boldsymbol{e}_h\|_{\mathrm{curl},\tilde{K}} + h^{\frac{1}{2}}\|\boldsymbol{r}\|_{[L_2(\tilde{K})]^3}). \tag{4.50}$$

Finally, adding $\|\bar{\mathbf{R}}-\mathbf{R}\|_{[L_2(l)]^3}$ to both sides results in the estimate

$$\begin{aligned}\|\mathbf{R}\|_{[L_2(l)]^3} &\leq \|\bar{\mathbf{R}}\|_{[L_2(l)]^3} + \|\mathbf{R}-\bar{\mathbf{R}}\|_{[L_2(l)]^3}\\ &\leq C(\|\bar{\mathbf{R}}-\mathbf{R}\|_{[L_2(l)]^3} + h^{-\frac{1}{2}}(1+k^2)\|\boldsymbol{e}_h\|_{\mathrm{curl},\tilde{K}} + h^{\frac{1}{2}}\|\boldsymbol{r}\|_{[L_2(\tilde{K})]^3}).\end{aligned} \tag{4.51}$$

Using (4.47) we obtain that

$$\begin{aligned}\|\mathbf{R}\|_{[L_2(l)]^3} &\leq C(h^{-\frac{1}{2}}(1+k^2)\|\boldsymbol{e}_h\|_{\mathrm{curl},\tilde{K}}\\ &\quad + h^{\frac{1}{2}}\|\bar{\boldsymbol{r}}-\boldsymbol{r}\|_{[L_2(\tilde{K})]^3} + \|\bar{\mathbf{R}}-\mathbf{R}\|_{[L_2(l)]^3}).\end{aligned} \tag{4.52}$$

Taking the square of (4.52) and (4.47) respectively, we obtain:

$$\begin{aligned}\|\mathbf{R}\|^2_{[L_2(l)]^3} &\leq C(h^{-1}(1+k^2)^2\|\boldsymbol{e}_h\|^2_{\mathrm{curl},\tilde{K}}\\ &\quad + h\|\bar{\boldsymbol{r}}-\boldsymbol{r}\|^2_{[L_2(\tilde{K})]^3} + \|\bar{\mathbf{R}}-\mathbf{R}\|^2_{[L_2(l)]^3}).\end{aligned} \tag{4.53}$$

and

$$\|\boldsymbol{r}\|^2_{[L_2(K)]^3} \leq C(\|\bar{\boldsymbol{r}}-\boldsymbol{r}\|^2_{[L_2(K)]^3} + (1+k^2)^2h^{-2}\|\boldsymbol{e}_h\|^2_{\mathrm{curl},K}). \tag{4.54}$$

Using the obvious equality

$$\|\mathbf{R}\|^2_{[L_2(\partial K)]^3} = \sum_{j=1}^{6}\|\mathbf{R}\|^2_{[L_2(l_j)]^3}$$

we sum up (4.53) for all faces l_j of K and multiplying $\|\boldsymbol{r}\|^2_{[L_2(K)]^3}$ with h^2 and $\|\mathbf{R}\|^2_{[L_2(\partial K)]^3}$ with h, respectively. Using the definition of η_K we finally get the

estimate:

$$\begin{aligned}\frac{1}{C}\eta_K^2 \;&\leq h^2\|\bar{\boldsymbol{r}}-\boldsymbol{r}\|^2_{[L_2(\tilde{K})]^3} + h\sum_{l\subset\partial K}\|\bar{\mathbf{R}}-\mathbf{R}\|^2_{[L_2(l)]^3} \\ &\quad + (1+k^2)^2\|\boldsymbol{e}_h\|^2_{\mathrm{curl},\tilde{K}}.\end{aligned} \tag{4.55}$$

Summarizing, we obtained that the error indicator η_K provides a lower bound for the exact error on the patch $\tilde{K}$ plus some computable remainders (arising from interpolation errors).

$$\eta_K^2 \leq C((1+k^2)^2\|\boldsymbol{e}_h\|^2_{\mathrm{curl},\tilde{K}} + h^2\|\bar{\boldsymbol{r}}-\boldsymbol{r}\|^2_{[L_2(\tilde{K})]^3} + h\|\bar{\mathbf{R}}-\mathbf{R}\|^2_{[L_2(\partial K)]^3}). \tag{4.56}$$

Using in addition Lemma 4.15 we state the main result of this section:

Theorem 4.16. *The implicit a posteriori error estimate $\hat{\boldsymbol{e}}_h$ can be used as a lower bound for the exact error with respect to the curl norm as follows:*

$$\begin{aligned}\|\hat{\boldsymbol{e}}_h\|_{\mathrm{curl},K} \leq C_1\eta_K \leq C((1+k^2)^2\|\boldsymbol{e}_h\|^2_{\mathrm{curl},\bar{K}} + h^2\|\bar{\boldsymbol{r}}-\boldsymbol{r}\|^2_{[L_2(\bar{K})]^3} \\ + h\|\bar{\mathbf{R}}-\mathbf{R}\|^2_{[L_2(\partial K)]^3})^{1/2}.\end{aligned} \tag{4.57}$$

Proof We get the desired result immediately using the estimates (4.42) and (4.56). ■

Remarks

1. Up to the estimate (4.56) we kept track of the k-dependence in the estimates.
2. If the local divergence free property of the estimate $\hat{\boldsymbol{e}}_h$ is desirable (for example, to ensure the equivalence of some norms in the error estimates [36]), one should enforce this condition by projecting to a divergence free basis. Although the finite element space that we used (see Section 4.2.1) consists of second order elements as well, the choice of V_h should be done according to the above requirement, when $\boldsymbol{E}_h$ is obtained using a higher order Nédélec space.
3. Another special situation occurs, if curl $\boldsymbol{E}_h = 0$ then $\mathbf{R}_{l_j} = 0$. In this way, one expects that the result in Theorem 4.16 can not be sharpened in the sense that neither $\hat{\boldsymbol{e}}_h$ nor η_K will provide an upper bound for the error. In this case the Helmholtz decomposition (see Lemma 4.5 in [73]) of $\boldsymbol{E}_h$ consists of only a gradient which can be non-smooth. For the smoothness of the components in the Helmholtz decomposition we refer to [5], Remark 2.16 and Theorem 2.17. This is in a good agreement with the fact that in the proof of upper bounds in residual based error estimation techniques one needs the regularity of solutions ([3], Section 2.2 and Section 3.2.3). This can fail for the present solution $\boldsymbol{E}$, see the test case in Section 5.1.2.
4. The remainder terms in Theorem 4.16 can make the estimate unsharp when $\|\bar{\boldsymbol{r}}-\boldsymbol{r}\|_{[L_2(K)]^3}$ and $\|\bar{\mathbf{R}}-\mathbf{R}\|_{[L_2(\partial K)]^3}$ are of the same order as the residuals

$\|\boldsymbol{r}\|_{[L_2(K)]^3}$ and $\|\mathbf{R}\|_{[L_2(\partial K)]^3}$, respectively. This can happen if the right hand side $\boldsymbol{J}$ is non-smooth or if we take a more general type of Maxwell equation with discontinuous material coefficients. Then in an adaptive refinement technique we should generate $\mathcal{T}_h$ in such a way that the solution in the subdomains is smooth. If this is not possible (e.g. if we want to avoid the use of curvilinear hexahedra) then some extra refinement should be performed in this critical region.

4.5 Numerical results

In this section, we demonstrate the performance of the implicit error estimator (4.11) applied to the time harmonic Maxwell equations. We consider the Maxwell equations on a domain Ω which is taken to be a cubic domain or a so-called Fichera cube, see Figure 4.9.

In order to be able to evaluate the discretization errors, we pick up a vector field $\boldsymbol{E}$, substitute it into the Maxwell equations and choose the source term $\boldsymbol{J}$ such that $\boldsymbol{E}$ is the solution.

Recall that $\boldsymbol{E}_h$ denotes the numerical solution of the Maxwell equations (4.1) obtained by using the edge finite elements given in Section 4.2.1. In the rest of this section the elements of the tessellation $\mathcal{T}_h$ are cubes with size $h \times h \times h$.

We verify the performance of the implicit error estimator for the Maxwell equations on three different test cases. The local problems (4.11) are solved by using the numerical model discussed in Section 4.3.2.

Several aspects determine the usefulness of an a posteriori error estimator:

- The error estimator has to be able to find those areas in the domain where the finite element solution has a large error, since this information is for local mesh adaptation.

- The error estimator should be close in magnitude to the real error, both locally and globally.

We check the performance of the implicit error estimator in the following way. First, we check if the estimator provides the right type of error distribution in the domain. Secondly, the magnitude of the global error estimate and its convergence under mesh refinement are compared with the exact error.

Define the exact error δ_K and the implicit *local error estimate* $\hat{\delta}_K$ on element K as

$$\delta_K = \|\boldsymbol{E} - \boldsymbol{E}_h\|_{\mathrm{curl},K}, \quad \hat{\delta}_K = \|\hat{\boldsymbol{e}}_h\|_{\mathrm{curl},K}. \tag{4.58}$$

The exact global error δ and the implicit global error estimate δ_h are obtained by summing up the local contributions

$$\delta = \left(\sum_{K \in \mathcal{T}_h} \delta_K^2 \right)^{1/2}, \quad \delta_h = \left(\sum_{K \in \mathcal{T}_h} \hat{\delta}_K^2 \right)^{1/2}. \tag{4.59}$$

4.5.1 Smooth solution

The first test case we consider are the Maxwell equations (4.1) on the domain $\Omega = (0,1)^3$ with the given source term $\boldsymbol{J}$ defined as

$$\boldsymbol{J}(x,y,z) = (\pi^2(p^2+m^2) - 1) \begin{pmatrix} \sin(\pi p y)\sin(\pi m z) \\ \sin(\pi p z)\sin(\pi m x) \\ \sin(\pi p x)\sin(\pi m y) \end{pmatrix}, \tag{4.60}$$

and which have a smooth exact solution

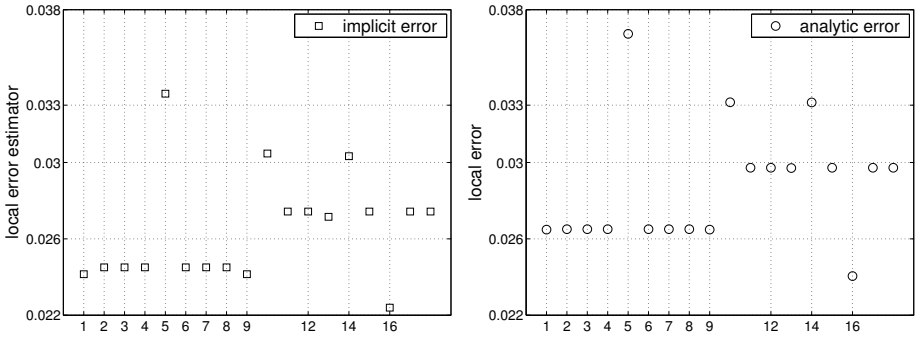
$$\boldsymbol{E}(x,y,z) = \begin{pmatrix} \sin(\pi p y)\sin(\pi m z) \\ \sin(\pi p z)\sin(\pi m x) \\ \sin(\pi p x)\sin(\pi m y) \end{pmatrix} \tag{4.61}$$

with $p, m \in \mathbb{N}$.

In Figures 4.1 and 4.2 we plot the local errors (4.58) obtained with the implicit error estimator and the exact error on a representative set of elements. The error distribution diagram for the case $p = 1, m = 1$ is given in Figure 4.1 and for the case $p = 5, m = 1$ the results are shown in Figure 4.2. The locations of some elements where the error is computed in the mesh with mesh size $h = \frac{1}{16}$ are shown in Figure 4.3. The labels on the horizontal axis in Figures 4.1 and 4.2 refer to the element numbers shown in Figure 4.3.
The error distribution obtained with the implicit error estimator shows a good agreement with the exact results. In the case $p = m = 1$, where the analytic solution has only one period in the domain, the error distribution is very close to the exact one. For the case $p = 5, m = 1$, where the analytic solution is more oscillatory, we observe that in some elements the distribution is slightly different, but the scheme is still able to detect subdomains with relatively large errors.

The rate of convergence and a global estimate of the error (4.59) for the case $p = m = 1$ are given in Figure 4.4. It shows the same convergence behavior

Elements with size $h \times h \times h$, where $h = \dfrac{1}{8}$.



Elements with size $h \times h \times h$, where $h = \dfrac{1}{16}$.

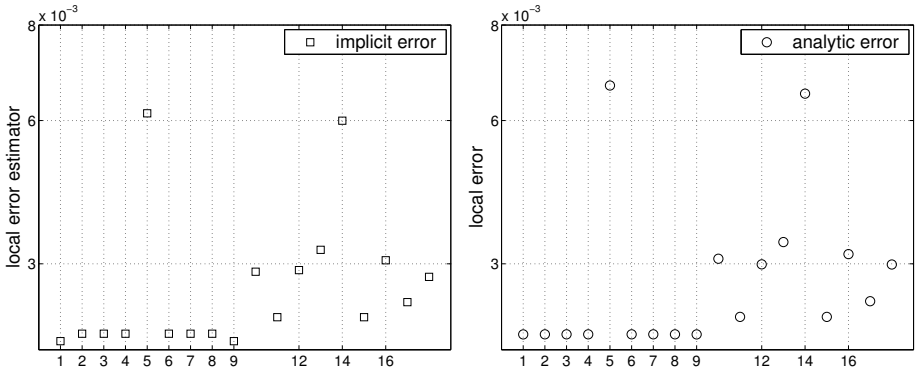
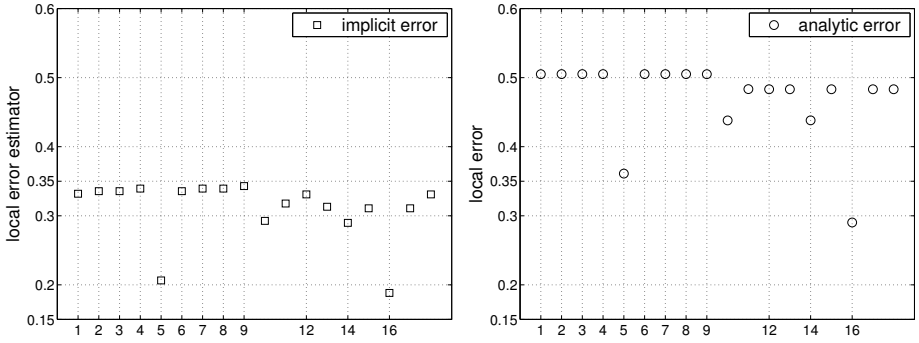


Figure 4.1: Error distribution in the $H(\mathrm{curl})$ norm for the smooth test case with $p = m = 1$.

Elements with size $h \times h \times h$, where $h = \frac{1}{8}$.



Elements with size $h \times h \times h$, where $h = \frac{1}{16}$.

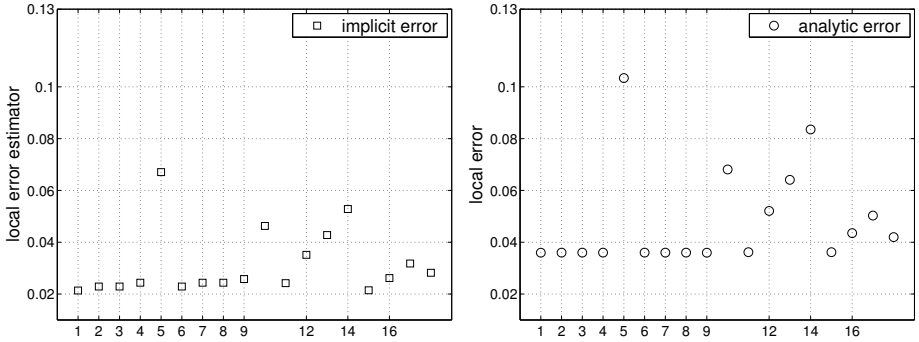


Figure 4.2: Error distribution in the $H(\mathrm{curl})$ norm for the smooth test case with $p = 5$, $m = 1$.

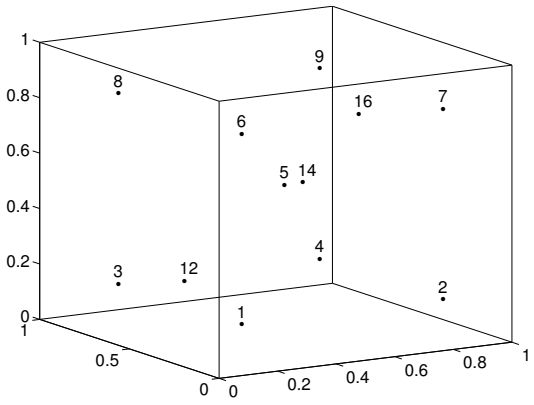


Figure 4.3: Location of some of the elements where the implicit error estimation was conducted (smooth test case) on a mesh with $h = \frac{1}{16}$.

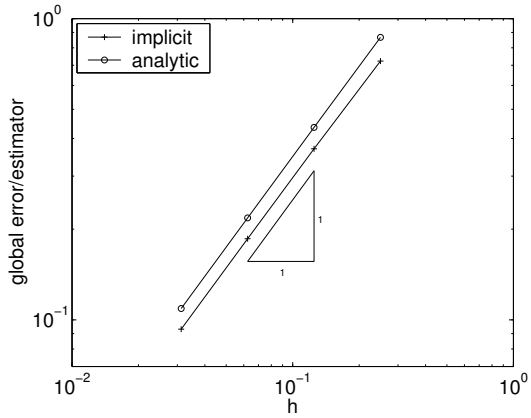


Figure 4.4: The global error estimate and the exact global error in the $H(\mathrm{curl})$ norm versus the mesh size h for smooth test case with $p = m = 1$.

under mesh refinement as the analytic error. Also, the predicted error magnitude is close to the true error.

4.5.2 Test case with singularities in the solution

Let us consider the domain $\Omega = (-1,1)^3$ and the function

$$f : \Omega \to \mathbb{R} \text{ with } f = \max\{|x|, |y|, |z|\}.$$

Define $\boldsymbol{E} : \Omega \to \mathbb{R}$ as $\boldsymbol{E} := -\nabla f(x,y,z)$. Then $\boldsymbol{E}$ solves the following Maxwell problem:

$$\begin{aligned} \operatorname{curl}\operatorname{curl} \boldsymbol{E} - \boldsymbol{E} &= \nabla f \quad \text{in } \Omega, \\ \boldsymbol{E} \times \boldsymbol{\nu} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.62}$$

In this example the right hand side function is in $[L_2(\Omega)]^3$ but the exact solution is not smooth, it is not even in $[H^{\frac{1}{2}}(\Omega)]^3$. Therefore, theoretically we can not guarantee even 1/2 order of convergence for the finite element solution. Numerically we have observed almost 1/2 order convergence in the $H(\mathrm{curl})$ norm, see Figure 4.7. For a similar example we refer to [15], where ∇f was smooth and the bilinear form B_K remained coercive. However, compared to the results given in [15], we could improve the accuracy of the estimator using the implicit error estimation technique (see also Section 4.5.4).

The error distribution computed on different elements for different values of the mesh size is depicted in Figure 4.5 and the plot of the global error estimator is given in Figure 4.7. The location of the elements on a mesh with $h = \frac{1}{8}$ is depicted in Figure 4.6. The labels on the horizontal axis in Figure 4.5 refer to the elements shown in Figure 4.6. We observe that the implicit error estimator provides the same type of error distribution as the exact error and also the estimates are close to the exact values. The convergence rate of the implicit error estimator is of the same order as the exact error.

4.5.3 Fichera cube

In this subsection we analyze the method on the Fichera cube $\Omega = (-1,1)^3 \setminus [-1,0]^3$. The solution on this domain has corner and edge singularities and can serve as a difficult test case. The boundary conditions and the source term in (4.1) are chosen such that the exact solution is $\boldsymbol{E} = \operatorname{grad}(r^{2/3}\sin(\frac{2}{3}\phi))$ in spherical coordinates, with $r = \sqrt{x^2+y^2+z^2}$, $\phi = \arccos\dfrac{r}{z}$. It is clear that $\boldsymbol{E}$ does not belong to $[H^1(\Omega)]^3$.

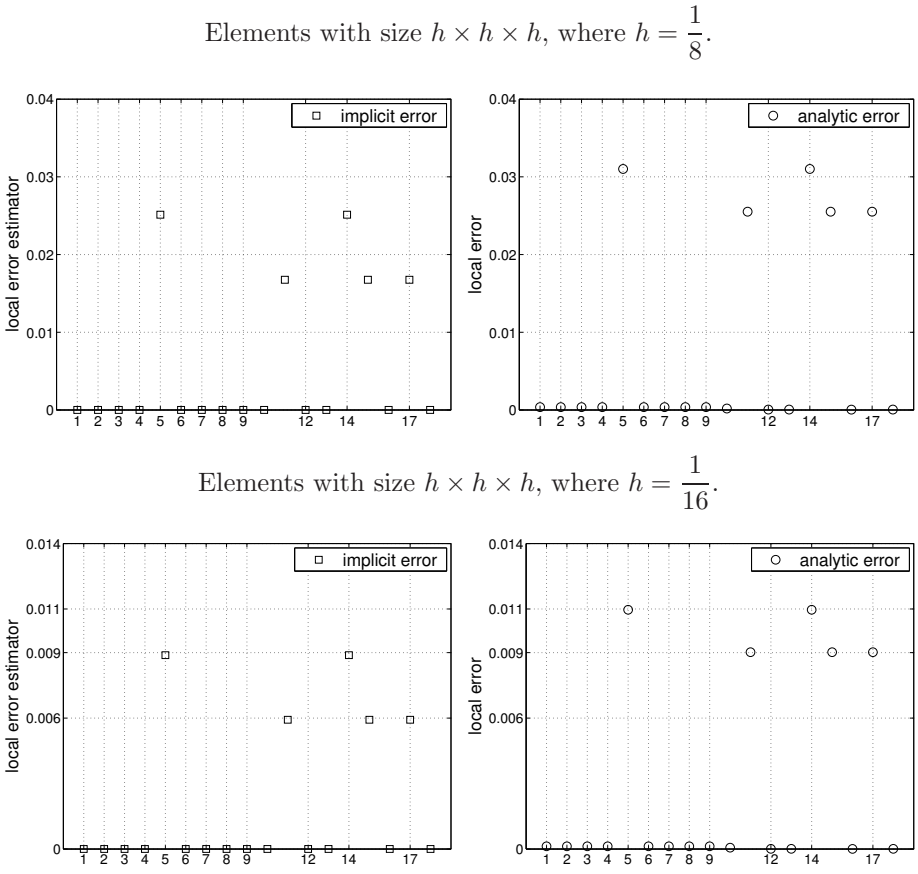


Figure 4.5: Error distribution in the $H(\mathrm{curl})$ norm for the singular test case.

The error distribution diagram is given in Figure 4.8. The large errors correspond to those elements which are close to the Fichera corner, located in the point $(0, 0, 0)$, see Figure 4.9. The plot of the global error estimate is given in Figure 4.10. As in the previous test cases, we observe a good agreement between the implicit error estimator and the exact errors, both in the error distribution and in the numerical values. The implicit error estimation is clearly capable of providing a rather accurate error estimate for a range of smooth and non-smooth flows, but even more important for an adaptation algorithm, it gives a clear indication of those regions where the error is large. The numerical results also show that the implicit error estimates are always bounded by the true error, which was proven in Theorem 4.16.

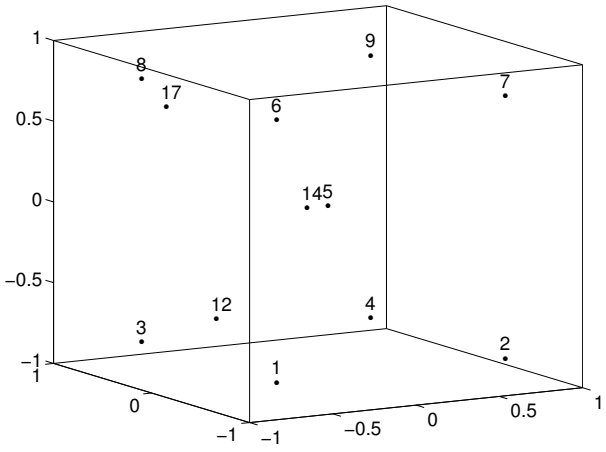


Figure 4.6: Location of some of the elements where the implicit error estimation was conducted (singular test case, Section 4.5.2) on a mesh with $h = \frac{1}{8}$.

4.5.4 Comparisons with some existing schemes

In [15] Beck, Hiptmair et al. consider the following elliptic boundary value problem

$$\begin{aligned} \text{curl } (\chi \text{curl } \boldsymbol{E}) + \beta \boldsymbol{E} &= \boldsymbol{J} \quad \text{in } \Omega, \\ \boldsymbol{E} \times \boldsymbol{\nu} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.63}$$

where χ and β are given positive functions on Ω.
We apply our implicit error estimator to (4.63) and compare the results with those given in [15] and [16].

For comparison purposes we consider the first example given in [15] and [16] on the domain $\Omega = (0,1)^3$

In this example the parameter χ is set to one and different values of β are taken into account. The exact solution is rather smooth and is given by $\boldsymbol{E} = (0, 0, \sin(\pi x))$. Roughly speaking, the system (4.63) reduces to (4.1) if we choose $\beta = -k^2$ and no other changes were necessary to the algorithm discussed for (4.1). Note, the bilinear form for this problem is coercive, contrary to the bilinear form (4.5) discussed in this chapter which is indefinite.

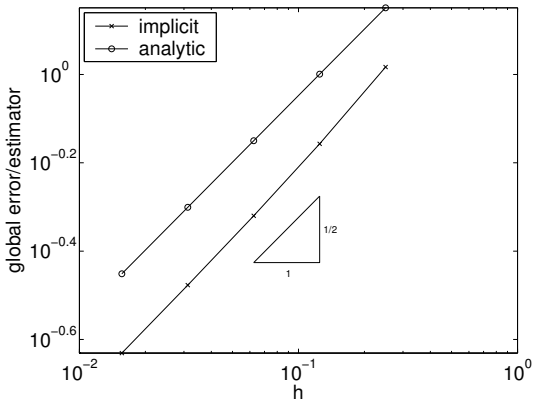
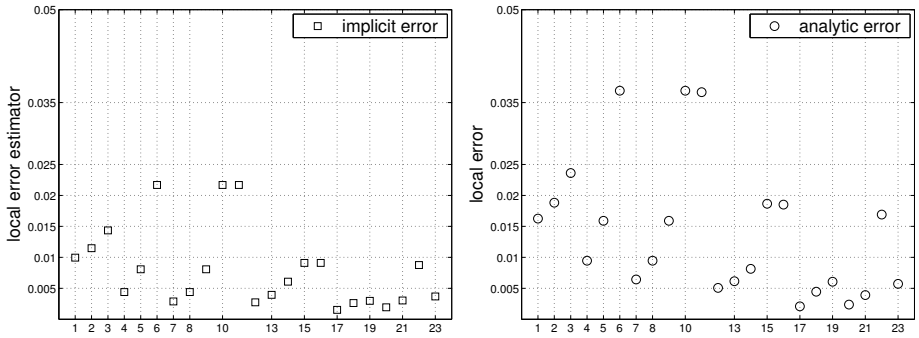


Figure 4.7: The global error estimate and the exact global error in the $H(\mathrm{curl})$ norm for the singular test case (Section 4.5.2).

For the finite element solution in the first example in [15] the authors start with a coarse grid (level 0) consisting of 6 tetrahedrons, which is refined uniformly up to five levels. In [15] also an adaptive strategy has been presented for other test cases, see Experiments 6-8 therein. In [16] a hierarchical type (implicit) error estimator is applied using preconditioning for solving the global problems for the error.

We make comparisons in terms of the effectivity index $\varepsilon_h := \dfrac{\delta_h}{\delta}$, which gives the ratio between the estimated and the true global error, where δ and δ_h are given by (4.59). This quantity merely reflects the quality of the global estimate, while we are mainly interested in local error estimates. The comparison table of the effectivity of the estimators given in [15], [16] and the implicit error estimator developed in this chapter are listed in Table 4.1. In comparison to the results given in [15] and [16] the estimates obtained with the implicit error estimator given by (4.11) are nearly insensitive to the value of β.
The above index is not capable to indicate the correlation between the distribution of the estimated and the exact error, which influences the effectivity of an adaptive technique. Therefore, we investigate a second quantity used for the comparison, the so-called "*fraction of incorrect decisions*", denoted by $\mu^{(1)}$. This measures how much the refinement controlled by the estimator differs from the refinement based upon an "ideal" estimator. The indicator $\mu^{(1)}$ is defined using the following sets:

Elements with size $h \times h \times h$, where $h = \dfrac{1}{4}$.



Elements with size $h \times h \times h$, where $h = \dfrac{1}{8}$.

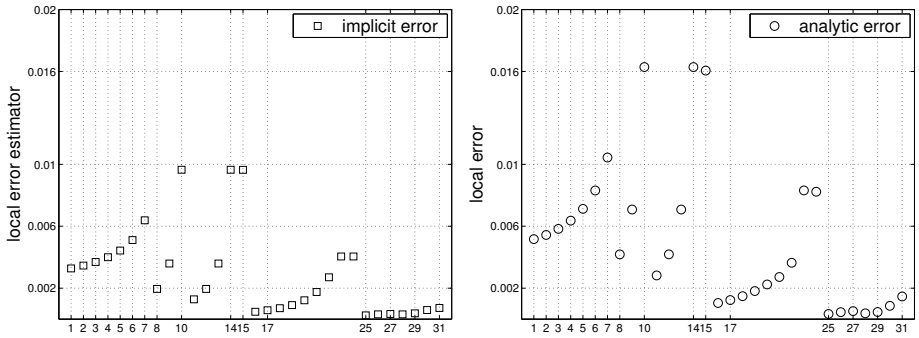


Figure 4.8: Error distribution in the $H(\mathrm{curl})$ norm on the Fichera cube.

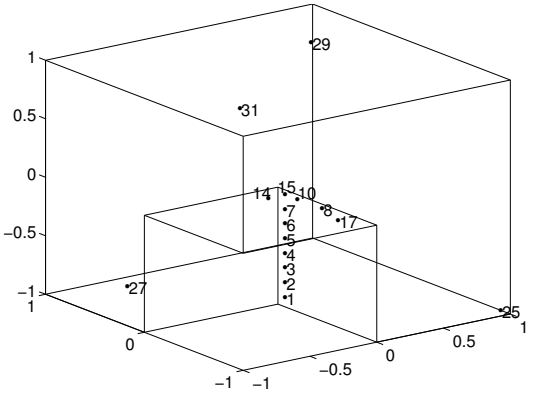


Figure 4.9: Location of some of the elements on the Fichera cube where the implicit error estimation was conducted on a mesh with $h = \frac{1}{8}$.

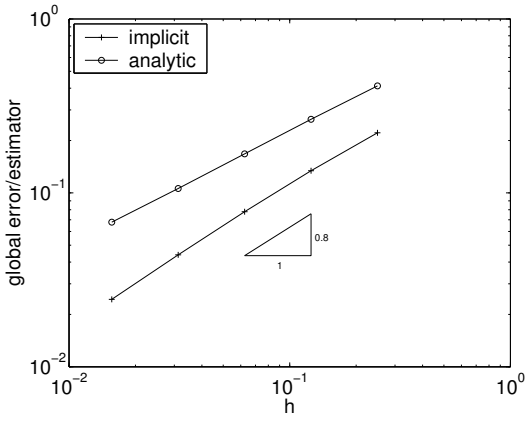


Figure 4.10: The global error estimate and the exact global error in the $H(\mathrm{curl})$ norm on the Fichera cube.

Table 4.1: Comparison of the effectivity indices.

β \ Level	0	1	2	3	4	5
10^{-4}	4.05	8.05	8.18	8.24	8.27	8.29
10^{-2}	4.05	8.05	8.17	8.23	8.27	8.29
1	3.01	7.64	7.78	7.84	7.87	7.89
10^{2}	2.29	4.27	4.70	4.95	5.20	5.26
10^{4}	2.33	4.23	4.66	4.86	4.95	5.00

Effectivity index ε_h for the error estimator given in [15].

β \ Level	0	1	2	3	4	5
10^{-4}	0.55	0.82	0.92	0.96	0.98	0.99
10^{-2}	0.55	0.82	0.92	0.96	0.98	0.99
1	0.56	0.83	0.92	0.95	0.97	0.98
10^{2}	0.71	0.87	0.92	0.91	0.90	0.90
10^{4}	0.72	0.88	0.93	0.94	0.94	0.93

Effectivity index ε_h for the Gauss-Seidel-based hierarchical error estimator given in [16].

β \ h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
10^{-4}	0.67	0.67	0.67	0.67	0.67
10^{-2}	0.67	0.67	0.67	0.67	0.67
1	0.67	0.67	0.67	0.67	0.67
10^{2}	0.63	0.65	0.67	0.67	0.67
10^{4}	0.44	0.37	0.40	0.53	0.63

Effectivity index ε_h for the implicit error estimator given by (4.11).

The set of elements marked for refinement by the error estimator are defined as

$$\hat{A} := \left\{ K \in \mathcal{T}_h : \hat{\delta}_K^2 > \sigma \frac{\delta_h^2}{n_K} \right\},$$

where $\sigma = 0.95$ and n_K denotes the number of elements in the tessellation $\mathcal{T}_h$, and the set of elements that should have been marked

$$A := \left\{ K \in \mathcal{T}_h : \delta_K^2 > \sigma \frac{\delta^2}{n_K} \right\}.$$

Table 4.2: Comparison of the "*incorrect decisions*".

β \ Level	0	1	2	3	4	5
10^{-4}	0.33	0.17	0.12	0.1	0.1	0.085
10^{-2}	0.33	0.17	0.12	0.1	0.1	0.086
1	0.33	0.25	0.18	0.14	0.13	0.13
10^{2}	0	0.42	0.088	0.14	0.15	0.16
10^{4}	0	0.44	0.11	0.15	0.16	0.16

Fraction of incorrect decisions $\mu^{(1)}$ for the error estimator given in [15].

β \ h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
10^{-4}	0	0	0	0	0
10^{-2}	0	0	0	0	0
1	0	0	0	0	0
10^{2}	0	0	0	0	0
10^{4}	0.25	0.31	0.07	0.00	0.03

Fraction of incorrect decisions $\mu^{(1)}$ for the implicit error estimator given by (4.11).

Then the indicator $\mu^{(1)}$ is defined as

$$\mu^{(1)} := \frac{1}{n_K} \# \left\{ (A \cap \hat{A}^c) \cup (A^c \cap \hat{A}) \right\}, \tag{4.64}$$

where for any set $S \subset \mathcal{T}_h$ the compliment with respect to $\mathcal{T}_h$ is denoted by S^c. In [15] satisfactory performance of the estimator means that $\mu^{(1)}$ stays bounded below 1 as refinement proceeds. The results are given in Table 4.2. For the implicit estimator this parameter is close to 0, which shows a much better performance than that given in [15]. In other words it means that the implicit error estimator developed in this chapter is able to find almost all elements which need refinement. The error indicator from [15] gives between 8.5% and 16% error ($\mu^{(1)} \cdot 100\%$) on the finest mesh, see Table 4.2.

4.6 Conclusions and further works

In Chapter 4 we have developed and analyzed an implicit a posteriori error estimation technique for the time harmonic Maxwell equations. The algorithm is well suited both for cases where the bilinear form is coercive and for the more complicated indefinite case. A nice feature of the implicit error estimator

is that no unknown constants appear. The algorithm is tested on a number of increasingly complicated test cases and the results show that it gives an accurate prediction of the error distribution and the local and global error. Also, in comparison with other a posteriori error estimation techniques [15, 80], for all considered tests it gives a sharper estimate of the error and its distribution.
In future work we will apply the implicit a posteriori error estimator in an adaptive algorithm and also consider different types of elements including the effect of mesh deformation on hexahedral elements.

4.7 Appendix

For the proof of Lemma 4.12 we recall some notations and results from linear algebra.

If the symmetric matrix A is positive semidefinite, we shortly write it as $A \geq 0$, while $A > 0$ is used for positive definite symmetric matrices. We use the fact that the maximal eigenvalue $\lambda_{A,\max}$ of a symmetric matrix $A \in \mathbb{R}^{n\times n}$ can be characterized as

$$\lambda_{A,\max} = \max_{\substack{\boldsymbol{u}\in\mathbb{R}^n \\ |\boldsymbol{u}|=1}} (A\boldsymbol{u}, \boldsymbol{u}) \tag{4.65}$$

and similarly, for the minimal eigenvalue $\lambda_{A,\min}$ of A

$$\lambda_{A,\min} = \min_{\substack{\boldsymbol{u}\in\mathbb{R}^n \\ |\boldsymbol{u}|=1}} (A\boldsymbol{u}, \boldsymbol{u}), \tag{4.66}$$

where $(\cdot,\cdot)$ yields the standard scalar product in $\mathbb{R}^n$. We will use also the notation $\lambda^+_{A,\min} > 0$ for the minimal nonzero eigenvalue of a positive semidefinite matrix A.

For the proof of Lemma 4.12 we need the following linear algebraic estimate.

Lemma 4.17. *Assume that $k \in \mathbb{R}$ and the symmetric matrices $A \geq 0, B > 0$ in $\mathbb{R}^{n\times n}$ are given such that $\ker A$ is an invariant space of B. Then, there are constants $c, h_0 \in \mathbb{R}^+$ such that for any h with $0 < h < h_0$ and any $\boldsymbol{u} \in \mathbb{R}^n$ there is a $\boldsymbol{v} \in \mathbb{R}^n$ such that*

$$|((\frac{1}{h^2}A - k^2B)\boldsymbol{u}, \boldsymbol{v})|^2 \geq c((\frac{1}{h^2}A + B)\boldsymbol{u}, \boldsymbol{u})((\frac{1}{h^2}A + B)\boldsymbol{v}, \boldsymbol{v}). \tag{4.67}$$

Proof In the proof we assume that $k \geq 1$, the remaining case can be handled in the same way; we have only to consider k^2B instead of B.

According to the assumptions every $\boldsymbol{u} \in \mathbb{R}^n$ can be decomposed as $\boldsymbol{u} = \boldsymbol{u}_1 + \boldsymbol{u}_2$, where $\boldsymbol{u}_1 \in \ker A$ and $\boldsymbol{u}_2 \perp \ker A$. Note that according to the assumptions, $B\boldsymbol{u}_1 \in \ker A$ and therefore, $(B\boldsymbol{u}_1, \boldsymbol{v}_2) = 0$ for all $\boldsymbol{v}_2 \perp \ker A$.

We distinguish two cases:

Case 1. $\boldsymbol{u}_1 = 0$. Then we can choose $\boldsymbol{v} = \boldsymbol{u} = \boldsymbol{u}_2$ and h_0 such that

$$k^2 \lambda_{B,\max} \leq \frac{1}{2h_0^2} \lambda^+_{A,\min}. \tag{4.68}$$

In this way, the characterizations in (4.65) and (4.66) imply that for any $0 < h < h_0$ the eigenvalues of the matrix $\frac{1}{h^2}A - k^2 B$ will be positive and their minimum is at least $\frac{1}{2h_0}\lambda^+_{A,\min}$. With these the left hand side of (4.67) can be estimated as

$$|((\frac{1}{h^2}A - k^2 B)\boldsymbol{u}_2, \boldsymbol{u}_2)|^2 \geq (\frac{1}{2h_0^2}\lambda^+_{A,\min}|\boldsymbol{u}_2|^2)^2.$$

The scalar products on the right hand side of (4.67) can be also estimated (using (4.68)) as follows:

$$\begin{aligned}((\frac{1}{h^2}A + B)\boldsymbol{u}, \boldsymbol{u})((\frac{1}{h^2}A + B)\boldsymbol{v}, \boldsymbol{v}) &= ((\frac{1}{h^2}A + B)\boldsymbol{u}_2, \boldsymbol{u}_2)((\frac{1}{h^2}A + B)\boldsymbol{u}_2, \boldsymbol{u}_2) \\ &\leq (\frac{2}{h_0^2}\lambda_{A,\max}|\boldsymbol{u}_2|^2)^2.\end{aligned}$$

Therefore, the choice $c = \left(\frac{1}{4}\frac{\lambda^+_{A,\min}}{\lambda_{A,\max}}\right)^2$ is appropriate in the first case.

Case 2. $\boldsymbol{u}_1 \neq 0$. We choose now $\boldsymbol{v} = \boldsymbol{u}_1 - \boldsymbol{u}_2$. The left hand side of (4.67) can be then rewritten as

$$\begin{aligned}&|((\frac{1}{h^2}A - k^2 B)\boldsymbol{u}, \boldsymbol{v})|^2 \\ &= |((\frac{1}{h^2}A - k^2 B)(\boldsymbol{u}_1 + \boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2)|^2 \\ &= |((\frac{1}{h^2}A - k^2 B)\boldsymbol{u}_1, \boldsymbol{u}_1) - ((\frac{1}{h^2}A - k^2 B)\boldsymbol{u}_2, \boldsymbol{u}_2)|^2 \\ &= |-k^2(B\boldsymbol{u}_1, \boldsymbol{u}_1) - ((\frac{1}{h^2}A - k^2 B)\boldsymbol{u}_2, \boldsymbol{u}_2)|^2 \\ &= k^4|(B\boldsymbol{u}_1, \boldsymbol{u}_1)|^2 + 2k^2(B\boldsymbol{u}_1, \boldsymbol{u}_1)((\frac{1}{h^2}A - k^2 B)\boldsymbol{u}_2, \boldsymbol{u}_2) \\ &+ |((\frac{1}{h^2}A - k^2 B)\boldsymbol{u}_2, \boldsymbol{u}_2)|^2.\end{aligned} \tag{4.69}$$

The right hand side of (4.67) can be simplified according to the following equality:

$$\begin{aligned}
&((\frac{1}{h^2}A+B)\boldsymbol{u},\boldsymbol{u})((\frac{1}{h^2}A+B)\boldsymbol{v},\boldsymbol{v})\\
&=((\frac{1}{h^2}A+B)(\boldsymbol{u}_1+\boldsymbol{u}_2),\boldsymbol{u}_1+\boldsymbol{u}_2)((\frac{1}{h^2}A+B)(\boldsymbol{u}_1-\boldsymbol{u}_2),\boldsymbol{u}_1-\boldsymbol{u}_2)\\
&=((B\boldsymbol{u}_1,\boldsymbol{u}_1)+((\frac{1}{h^2}A+B)\boldsymbol{u}_2,\boldsymbol{u}_2))^2\\
&=|(B\boldsymbol{u}_1,\boldsymbol{u}_1)|^2+2(B\boldsymbol{u}_1,\boldsymbol{u}_1)((\frac{1}{h^2}A+B)\boldsymbol{u}_2,\boldsymbol{u}_2)+|((\frac{1}{h^2}A+B)\boldsymbol{u}_2,\boldsymbol{u}_2)|^2.
\end{aligned} \tag{4.70}$$

Comparing (4.69) and (4.70), using the fact that $k \geq 1$, the positive definiteness of B and choosing $c = \frac{1}{2}$, then a sufficient condition for the inequality to hold is that we need an h_0 such that for any h with $0 < h < h_0$ and for all $\mathbb{R}^n \ni \boldsymbol{u}_2 \perp \ker A$ the following inequality holds:

$$((\frac{1}{h^2}A-k^2B)\boldsymbol{u}_2,\boldsymbol{u}_2) \geq \frac{1}{2}((\frac{1}{h^2}A+B)\boldsymbol{u}_2,\boldsymbol{u}_2).$$

We need in fact that the matrix $\frac{1}{2h^2}A-(k^2+\frac{1}{2})B$ is positive semidefinite. According to *Case 1* this holds whenever h_0 is chosen according to the inequality

$$(k^2+\frac{1}{2})\lambda_{B,\max} \leq \frac{1}{2h_0^2}\lambda^+_{A,\min}. \tag{4.71}$$

Comparing the estimates in (4.68) and (4.71) gives that the choice

$$h_0^2 \leq \frac{1}{2k^2+1}\frac{\lambda^+_{A,\min}}{\lambda_{B,\max}} \tag{4.72}$$

is sufficient in both cases. ■

Using this result we can prove Lemma 4.12.

Proof of Lemma 4.12. Using the notations introduced in Section 4.4.2 and a simple change of variables we obtain that the mass matrix on a cube K with edge length h can be written as

$$B_{1,K} = \frac{1}{h}B_{1,\mathrm{curl}} - k^2hB_{1,0}$$

and accordingly, for $\boldsymbol{u} = \sum_{i=1}^n u_i\boldsymbol{\phi}_i \in V_h$ and $\boldsymbol{v} = \sum_{i=1}^n v_i\boldsymbol{\phi}_i \in V_h$

$$B_K(\boldsymbol{u},\boldsymbol{v}) = ((\frac{1}{h}B_{1,\mathrm{curl}} - k^2hB_{1,0})(u_1,u_2,\ldots,u_n)^T,(v_1,v_2,\ldots,v_n)^T) \tag{4.73}$$

In the same way,

$$\|\boldsymbol{u}\|^2_{\mathrm{curl},K} = ((\frac{1}{h}B_{1,\mathrm{curl}} + hB_{1,0})(u_1, u_2, \ldots, u_n)^T, (u_1, u_2, \ldots, u_n)^T). \quad (4.74)$$

Substituting (4.73) and (4.74) into the desired inequality (4.40) we have to find positive constants h_0 and a $C_0 > 0$ such that for any $0 < h < h_0$ and any $\boldsymbol{u} \in \mathbb{R}^n$ there is a $\boldsymbol{v} \in \mathbb{R}^n$ such that

$$C_0((\frac{1}{h}B_{1,\mathrm{curl}} - k^2 hB_{1,0})\boldsymbol{u}, \boldsymbol{v})^2 \geq ((\frac{1}{h}B_{1,\mathrm{curl}} + hB_{1,0})\boldsymbol{u}, \boldsymbol{u})((\frac{1}{h}B_{1,\mathrm{curl}} + hB_{1,0})\boldsymbol{v}, \boldsymbol{v}). \quad (4.75)$$

Dividing both sides with h^2, we can apply Lemma 4.17 with $A = B_{1,\mathrm{curl}}$ and $B = B_{1,0}$. For this, we have to check that the conditions in Lemma 4.17 hold. A lenghty computation gives that

$$B_{1,\mathrm{curl}} = \frac{1}{1080}\begin{pmatrix} 8 & 4 & -1 & 1 & -1 & 1 & 2 & 0 & 0 \\ 4 & 8 & 1 & -1 & 1 & -1 & 2 & 0 & 0 \\ -1 & 1 & 8 & 4 & -1 & 1 & 0 & 2 & 0 \\ 1 & -1 & 4 & 8 & 1 & -1 & 0 & 2 & 0 \\ -1 & 1 & -1 & 1 & 8 & 4 & 0 & 0 & 2 \\ 1 & -1 & 1 & -1 & 4 & 8 & 0 & 0 & 2 \\ 2 & 2 & 0 & 0 & 0 & 0 & 4/5 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 & 0 & 4/5 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 & 4/5 \end{pmatrix}$$

and

$$B_{1,0} = \frac{1}{1080}\begin{pmatrix} 2/5 & 1/5 & 0 & 0 & 0 & 0 & 1/10 & 0 & 0 \\ 1/5 & 2/5 & 0 & 0 & 0 & 0 & 1/10 & 0 & 0 \\ 0 & 0 & 2/5 & 1/5 & 0 & 0 & 0 & 1/10 & 0 \\ 0 & 0 & 1/5 & 2/5 & 0 & 0 & 0 & 1/10 & 0 \\ 0 & 0 & 0 & 0 & 2/5 & 1/5 & 0 & 0 & 1/10 \\ 0 & 0 & 0 & 0 & 1/5 & 2/5 & 0 & 0 & 1/10 \\ 1/10 & 1/10 & 0 & 0 & 0 & 0 & 1/25 & 0 & 0 \\ 0 & 0 & 1/10 & 1/10 & 0 & 0 & 0 & 1/25 & 0 \\ 0 & 0 & 0 & 0 & 1/10 & 1/10 & 0 & 0 & 1/25 \end{pmatrix}.$$

By definition, for any vector $\boldsymbol{u} = (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n$

$$(B_{1,\mathrm{curl}}\boldsymbol{u}, \boldsymbol{u}) = (\mathrm{curl} \sum_{i=1}^{9} u_i \boldsymbol{\phi}_i^0, \mathrm{curl} \sum_{i=1}^{9} u_i \boldsymbol{\phi}_i^0)_{[L_2(\hat{K})]^3} = \|\mathrm{curl} \sum_{i=1}^{9} u_i \boldsymbol{\phi}_i^0\|^2_{[L_2(\hat{K})]^3}, \quad (4.76)$$

where the basis $\{\phi_i^0\}_{i=1}^9$ is defined in (4.3.2). Similarly,

$$(B_{1,0}\boldsymbol{u},\boldsymbol{u}) = (\sum_{i=1}^{9} u_i\boldsymbol{\phi}_i^0, \sum_{i=1}^{9} u_i\boldsymbol{\phi}_i^0)_{[L_2(\hat{K})]^3} = \|\sum_{i=1}^{9} u_i\boldsymbol{\phi}_i^0\|^2_{[L_2(\hat{K})]^3},$$

which show that $B_{1,\mathrm{curl}} \geq 0$ and $B_{1,0} > 0$. Moreover (4.76) gives that for any nonzero vector $\boldsymbol{u} \in \mathbf{R}^n$ the condition $(B_{1,\mathrm{curl}}\boldsymbol{u},\boldsymbol{u}) = 0$ can only be fulfilled if curl $\sum_{i=1}^9 u_i\boldsymbol{\phi}_i^0 = \mathbf{0}$.

In order to determine such a linear combination, observe that an arbitrary element $\hat{\Phi}$ in the finite dimensional bubble function space $\hat{V}$ given in (4.3.2) can be written as

$$\hat{\Phi}(\xi,\eta,\zeta) = \begin{pmatrix} (a_1\xi^2 + b_1\xi + c_1)\eta(1-\eta)\zeta(1-\zeta) \\ \xi(1-\xi)(a_2\eta^2 + b_2\eta + c_2)\zeta(1-\zeta) \\ \xi(1-\xi)\eta(1-\eta)(a_3\zeta^2 + b_3\zeta + c_3) \end{pmatrix}$$

and accordingly,

$$\begin{aligned} &\mathrm{curl}\,\hat{\Phi}(\xi,\eta,\zeta) \\ &= \begin{pmatrix} \xi(1-\xi)(1-2\eta)(a_3\zeta^2 + b_3\zeta + c_3) - \xi(1-\xi)(a_2\eta^2 + b_2\eta + c_2)(1-2\zeta) \\ (a_1\xi^2 + b_1\xi + c_1)\eta(1-\eta)(1-2\zeta) - (1-2\xi)\eta(1-\eta)(a_3\zeta^2 + b_3\zeta + c_3) \\ (1-2\xi)(a_2\eta^2 + b_2\eta + c_2)\zeta(1-\zeta) - (a_1\xi^2 + b_1\xi + c_1)(1-2\eta)\zeta(1-\zeta) \end{pmatrix}. \end{aligned}$$

This can be zero on $\hat{K}$ if the following equalities hold:

$$\begin{aligned} (1-2\eta)(a_3\zeta^2 + b_3\zeta + c_3) &= (a_2\eta^2 + b_2\eta + c_2)(1-2\zeta), \\ (a_1\xi^2 + b_1\xi + c_1)(1-2\zeta) &= (1-2\xi)(a_3\zeta^2 + b_3\zeta + c_3), \\ (1-2\xi)(a_2\eta^2 + b_2\eta + c_2) &= (a_1\xi^2 + b_1\xi + c_1)(1-2\eta). \end{aligned}$$

Here using the first equation we obtain that

$$a_2 = a_3 = 0,\ \frac{b_2}{c_2} = \frac{b_3}{c_3} = -2,\ b_2 = b_3,\ c_2 = c_3$$

or, alternatively, $a_2 = a_3 = b_2 = c_2 = b_3 = c_3 = 0$. Similarly, the second one gives that

$$a_3 = a_1 = 0,\ \frac{b_3}{c_3} = \frac{b_1}{c_1} = -2,\ b_3 = b_1,\ c_3 = c_1$$

or, alternatively, $a_1 = a_3 = b_3 = c_3 = b_1 = c_1 = 0$. The third equation is trivially satisfied with these coefficients. We may assume that $c_1 = 1$ and in this way, $c_1 = c_2 = c_3 = 1$ and $b_1 = b_2 = b_3 = -2$. This gives the eigenfunction

$$(\xi,\eta,\zeta) \to \begin{pmatrix} (1-2\xi)\eta(1-\eta)\zeta(1-\zeta) \\ \xi(1-\xi)(1-2\eta)\zeta(1-\zeta) \\ \xi(1-\xi)\eta(1-\eta)(1-2\zeta) \end{pmatrix} \tag{4.77}$$

which is a linear combination $\sum_{i=1}^{9} u_i \Phi_i^0$, with $\boldsymbol{u} = (1,-1,1,-1,1,-1,0,0,0)^T$. This gives that $\ker B_{1,\mathrm{curl}} = \mathrm{span}\,(1,-1,1,-1,1,-1,0,0,0)$, which is an invariant space of $B_{1,0}$ since $B_{1,0}(\ker B_{1,\mathrm{curl}}) = \ker B_{1,\mathrm{curl}}$. In this way, the conditions in Lemma 4.17 are fullfilled for $A = B_{1,\mathrm{curl}}$ and $B = B_{1,0}$, which gives the statement in Lemma 4.12. ■

Proof of Lemma 4.13. Symbolic calculations give that $\lambda^+_{B_{1,\mathrm{curl}},\min} = \frac{4}{675} - \frac{\sqrt{246}}{2700}$ and $\lambda_{B_{1,0},\max} = \frac{1}{3375} + \frac{\sqrt{246}}{54000}$.

Using the above numerical results and substituting them into (4.72) we obtain the statement of Lemma 4.13. ■

Remarks 1. In the proof of Lemma 4.12 we determined that the kernel of the curl operator in the bubble function space is the one dimensional subspace generated by the function in (4.77). This coincides with the subspace of the bubble function space which consists of discrete gradients, namely (4.77) is the gradient of the function given by

$$(\xi,\eta,\zeta) \to (\xi(1-\xi),\eta(1-\eta),\zeta(1-\zeta)).$$

2. The reason why we did not include k into the matrix B in Lemma 4.17 is that we wanted to demonstrate the dependence of the mesh size parameter h_0 on k.
3. An easy calculation gives that for $K = (0,h)^3$ the condition number of $B_{1,K}$ is proportional to h^{-2}. However, this does not harm the solution of the local problem due to Lemma 4.12.

CHAPTER 5

Adaptive finite element techniques for the Maxwell equations using implicit a posteriori error estimates

For the adaptive solution of the Maxwell equations with Nédélec edge finite element methods on three-dimensional domains, we consider an implicit a posteriori error estimation technique. On each element of the tessellation an equation for the error is formulated and solved with a properly chosen local finite element basis. The discrete bilinear form of the local problems is shown to satisfy an inf-sup condition which ensures the well posedness of the error equations. An adaptive solution algorithm is developed based on the obtained error estimates. The performance of the method is tested on various problems including non-convex domains with non-smooth boundaries. The numerical results show that the estimated error, computed by the implicit a posteriori error estimation technique, correlates well with the actual error. On the meshes generated adaptively with the help of the error estimator, a higher accuracy is achieved than on globally refined meshes. Moreover, the rate of the error convergence on the locally adapted meshes is faster than that on the globally refined meshes.

5.1 Introduction

In many real life problems (for example scattering problems, optical fibers, design of antennas) it becomes increasingly important to solve the full set of Maxwell equations on complex three-dimensional domains. Due to the complexity of the domains the solution of the Maxwell equations frequently has limited regularity, such as singularities near corners and non-convex edges [33],

and efficient solution methods require adaptive techniques in order to capture detailed structures.

A posteriori error estimation techniques to control the adaptation process in finite element methods have become popular tools for the numerical solution of partial differential equations, see e.g. [3, 10, 11, 78, 104], and are also important for the Maxwell equations. A crucial requirement for a posteriori error estimation techniques is that they provide an accurate estimate of the error throughout the finite element mesh. The a posteriori error estimate is then used to generate meshes locally finer in areas where the mesh resolution is not sufficient to achieve the required accuracy. For wave type problems, in particular for the Maxwell equations this is, however, only possible if we use a sufficiently fine mesh compared to the wave length. In this case, the major part of the computational error arises from the boundary singularities. Otherwise, the pollution effect can make the a posteriori error estimates unreliable [7] and a further careful analysis is needed to estimate the pollution error separately [8]. There are basically two types of a posteriori error estimation methods, namely explicit and implicit techniques.

Explicit error estimation techniques provide an upper bound for the local error residual based on the numerical solution (see e.g. [9, 10, 104]), but generally contain an unknown constant and as such are frequently not sharp and do not provide computable error bounds. There are several techniques to obtain explicit bounds for the unknown constant term (see e.g. [27]), but in most applications the estimates are somewhat pessimistic, hence the resulting estimators tend to be unrealistic and fail to detect the more subtle nuances of the specific problem. Several applications of adaptive methods with an explicit error estimation technique for the Maxwell equations can be found in [15, 20, 28, 83, 84].

Implicit error estimators seek to avoid these disadvantages by retaining the structure of the original equation as far as possible. The idea of implicit a posteriori error estimates is to formulate local problems for the error function, either over a single element or over a small patch of elements, with suitably chosen boundary conditions and then solve them with an appropriate finite element method [1, 3]. This technique can provide reliable estimates, but one has to solve additional, small boundary value problems. Beyond the standard elliptic case it has been applied for flow problems in two-dimensional domains [70] and for the Maxwell equations with a coercive bilinear form [16]. The numerical experiments in [21, 81] for the time harmonic Maxwell equations suggest the implicit error estimation technique as a promising approach. Moreover, in [36, 85] equilibration techniques have been applied in case of higher order elements, but

a precise analysis of this method is still lacking.

In [65, 66], see also Chapter 4, we developed an implicit a posteriori error estimation technique for the time harmonic Maxwell equations on a cubic mesh and proved well posedness of the local problems (without any post-processing) with suitably chosen boundary conditions. We also pointed out that this gives a lower bound for the analytic error.

The main goal of this chapter is to apply the implicit error estimation technique in an adaptive mesh refinement algorithm. We perform the adaptation on a tetrahedral mesh, which requires some modifications in the analysis compared to Chapter 4 and [65, 66]. As a natural choice for the finite element spaces we use Nédélec first order edge basis functions. Then we define a weak formulation for the error in each element, which is solved with a finite element method. The local problems formulated for the error are solved with second order Nédélec elements without the linear part. The use of higher order elements to solve the local error equations is essential to obtain a good approximation of the true error and also reduces the pollution effect discussed in [7, 8]. In various test cases (on non-convex domains with singular solutions) we verify the performance of the implicit error estimation technique. Provided that the mesh resolution is fine enough we show that the method is capable of detecting regions with a relatively large error and, based on this information and using an adaptive mesh generation technique, we are able to achieve a smaller error on adaptively generated meshes than on globally refined meshes. Also, the reduction of the error using the adaptation procedure based on the implicit error estimation technique is faster than that on globally refined meshes.

An important issue for adaptive methods is how to adapt a mesh while maintaining mesh quality. In particular, it is important to choose a selection algorithm for the subdomains where finer elements are needed. Here we would like to mention that there is no best algorithm for marking elements and several options are discussed in Section 5.6. For more information about refinement strategies we refer to [4, 39, 89, 98, 14]. In all our numerical experiments we use the Centaur mesh generator [29] with so called source based mesh generation (see Section 5.6) depending on the selection of a fixed fraction of elements for mesh adaptation. This approach tries to make the local mesh finer in specified regions while preserving the high quality of the mesh. One of the advantages of the Centaur mesh generator is that it creates adaptive meshes without hanging nodes. Meshes without hanging nodes are necessary for Nédélec type elements, otherwise these elements are not well defined. Another desirable property of this mesh generator is that it avoids elements with large dihedral angles, which

is important for accuracy requirements.

This chapter is organized as follows: in Section 5.2 we present the Maxwell equations, their weak formulation and define the finite element discretization. Section 5.3 describes the implicit error estimation technique with a properly chosen local finite element space. The inf-sup condition for the local error formulation is proven in Section 5.4 using a Poincaré type inequality (Lemma 5.5). A similar result for quasi uniform subdomains is available in [53], Lemma 4.1. We also investigate the frequency dependence of the parameters in the estimates. An efficiency analysis of the method is presented in Section 5.5. In Section 5.6 we discuss several adaptation strategies. The performance of the implicit error estimation technique is investigated for various test cases including non-convex domains in Section 5.7. Finally, conclusions are drawn in Section 5.8.

5.2 Mathematical formalization

Consider the time harmonic Maxwell equations for the electric field $\boldsymbol{E}:\Omega\rightarrow\ \mathbb{R}^3$ with perfectly conducting boundary conditions:

$$\operatorname{curl}\operatorname{curl}\boldsymbol{E}-k^2\boldsymbol{E}=\boldsymbol{J}\quad\text{in }\Omega, \tag{5.1a}$$

$$\boldsymbol{E}\times\boldsymbol{\nu}=0\quad\text{on }\partial\Omega, \tag{5.1b}$$

where $\Omega\subset\mathbb{R}^3$ is a Lipschitz domain with outward normal vector $\boldsymbol{\nu}$ and $\boldsymbol{J}\in[L_2(\Omega)]^3$ a given source function. The wave number k relates to the frequency ω and the velocity of the wave propagation c as $k=\frac{\omega}{c}$. The velocity of wave propagation is given as $c=\frac{1}{\sqrt{\varepsilon\mu}}$, where the dielectric permittivity $\varepsilon=\varepsilon_0\varepsilon_r$ and the magnetic permeability $\mu=\mu_0\mu_r$ are material properties. The free space dielectric permittivity and magnetic permeability are defined by $\varepsilon_0=\frac{1}{36\pi}10^{-9}\,\mathrm{Fm}^{-1}$ and $\mu_0=4\pi10^{-7}\,\mathrm{Hm}^{-1}$, respectively [73]. The dimensionless parameters ε_r and μ_r are material dependent and called relative permittivity and relative permeability, respectively.

In this chapter we consider the dimensionless Maxwell equations to avoid problems with floating point arithmetic when working with very large numbers. How to make the Maxwell equations dimensionless is explained in e.g. [73].

In the subsequent derivations we will need the following Hilbert space corresponding to the Maxwell equations

$$H(\operatorname{curl},\Omega)=\{\boldsymbol{u}\in[L_2(\Omega)]^3:\operatorname{curl}\boldsymbol{u}\in[L_2(\Omega)]^3\},$$

which is equipped with the curl norm

$$\|\boldsymbol{u}\|_{\mathrm{curl},\Omega} = (\|\boldsymbol{u}\|^2_{[L_2(\Omega)]^3} + \|\mathrm{curl}\,\boldsymbol{u}\|^2_{[L_2(\Omega)]^3})^{1/2}. \tag{5.2}$$

The differential operator curl is understood in a distributional sense. While analyzing (5.1), usually a subspace of $H(\mathrm{curl},\Omega)$ is used, namely

$$H_0(\mathrm{curl},\Omega) = \{\boldsymbol{u} \in H(\mathrm{curl},\Omega) : \boldsymbol{\nu} \times \boldsymbol{u}|_{\partial\Omega} = 0\},$$

where $\boldsymbol{\nu} \times \boldsymbol{u}|_{\partial\Omega}$ denotes the extension of the tangential trace to non smooth functions [73].

For the weak formulation of (5.1) we introduce the following bilinear form

$$B : H(\mathrm{curl},\Omega) \times H(\mathrm{curl},\Omega) \to \mathbb{R}$$

with

$$B(\boldsymbol{u},\boldsymbol{v}) = (\mathrm{curl}\,\boldsymbol{u}, \mathrm{curl}\,\boldsymbol{v}) - k^2(\boldsymbol{u},\boldsymbol{v}).$$

Similarly, the bilinear form B_K is defined in the same way but now on the subdomain $K \subset \Omega$ (instead of Ω). For the corresponding $[L_2(\Omega)]^3$ scalar product on the domain K and at its boundary ∂K the notations $(\cdot,\cdot)_K$ and $(\cdot,\cdot)_{\partial K}$ are used, respectively. In the same way, the curl norm on K is defined by

$$\|\boldsymbol{u}\|_{\mathrm{curl},K} = (\|\boldsymbol{u}\|^2_{[L_2(K)]^3} + \|\mathrm{curl}\,\boldsymbol{u}\|^2_{[L_2(K)]^3})^{1/2}.$$

Using the above notations the weak formulation of the time harmonic Maxwell equations (5.1) is: for a given source function $\boldsymbol{J}$, find $\boldsymbol{E} \in H_0(\mathrm{curl},\Omega)$ such that for all $\boldsymbol{v} \in H_0(\mathrm{curl},\Omega)$ the following relation is satisfied

$$B(\boldsymbol{E},\boldsymbol{v}) = (\boldsymbol{J},\boldsymbol{v}). \tag{5.3}$$

5.2.1 Finite elements in $H(\mathrm{curl})$: First order edge elements

For the numerical solution of (5.3) we use the $H(\mathrm{curl})$ conforming edge finite element method proposed by Nédélec [76] for tetrahedral elements.

It is convenient to define the finite elements first on a reference element, which in our case is a tetrahedron $\hat{K}$ with nodes $\hat{X}_1, \hat{X}_2, \hat{X}_3, \hat{X}_4$, see Figure 5.1, where

$$\hat{X}_1 = (0,0,0),\ \hat{X}_2 = (1,0,0),\ \hat{X}_3 = (0,1,0),\ \hat{X}_4 = (0,0,1).$$

The first order Nédélec elements are defined on the reference element $\hat{K}$ as

$$\boldsymbol{W}^0_i = (L_{i_1}\nabla L_{i_2} - L_{i_2}\nabla L_{i_1})l_i, \qquad i = 1,\ldots,6,$$

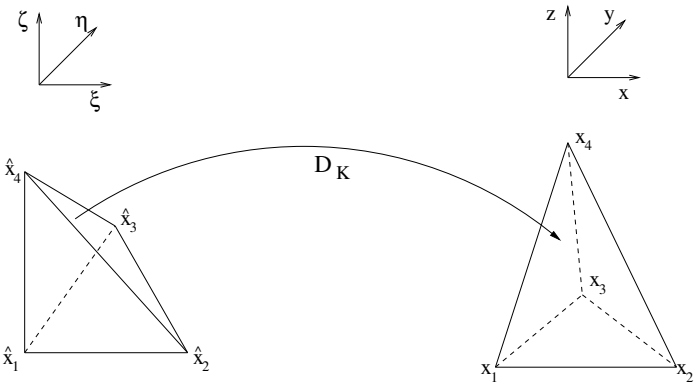


Figure 5.1: The reference tetrahedron (left) and tetrahedron in physical space (right).

where L_j is the Lagrange basis function corresponding to node j of $\hat{K}$, l_i the length of edge i, and i the edge number associated with the nodes i_1 and i_2 (see Table 5.1). In more explicit form this basis reads

$$\begin{aligned}
\boldsymbol{W}_1^0 &= (1-\eta-\zeta,\xi,\xi)^T, & \boldsymbol{W}_2^0 &= (\eta,1-\xi-\zeta,\eta)^T,\\
\boldsymbol{W}_3^0 &= (\zeta,\zeta,1-\xi-\eta)^T, & \boldsymbol{W}_4^0 &= \sqrt{2}\,(-\eta,\xi,0)^T,\\
\boldsymbol{W}_5^0 &= \sqrt{2}\,(\zeta,0,-\xi)^T, & \boldsymbol{W}_6^0 &= \sqrt{2}\,(0,-\zeta,\eta)^T,
\end{aligned}$$

with (ξ,η,ζ) the local coordinates on $\hat{K}$.

A detailed construction of Nédélec basis functions can be found, for example, in [73]. Next, we introduce a tetrahedral tessellation $\mathcal{T}_h$ of Ω with N elements and N_e edges. The basis defined on the reference element $\hat{K}$ can be transformed to an arbitrary tetrahedron $K \in \mathcal{T}_h$ using the isoparametric mapping

$$D_K : (\xi,\eta,\zeta) \in \hat{K} \mapsto (x,y,z) = \sum_{i=1}^{4} X_i L_i(\xi,\eta,\zeta) \in K, \tag{5.4}$$

provided that this mapping is a diffeomorphism. Here $X_i = (x_i, y_i, z_i)$ denote the nodes of K. We numerate the nodes in $\hat{K}$ and K such that $X_i = D_K(\hat{X}_i)$. It is well known that the covariant transformation preserves line integrals under a change of coordinates [73, 90], so that the basis functions for a given tetrahedron K can be defined as

$$\boldsymbol{w}_j(x,y,z) = (\mathrm{d}D_K^{-1})^T \boldsymbol{W}_j^0(\xi,\eta,\zeta), \quad j = 1,\ldots,6, \tag{5.5}$$

Table 5.1: Edge and face enumeration.

Edge #	Node i_1	Node i_2
1	1	2
2	1	3
3	1	4
4	2	3
5	4	2
6	3	4

Face #	Node i_1	Node i_2	Node i_3
1	2	3	4
2	1	3	4
3	1	2	4
4	1	2	3

where $\mathrm{d}D_K$ is the Jacobian of the transformation D_K.

We denote by $W_h \subset H(\mathrm{curl}, \Omega)$ the space of Nédélec first order edge basis functions:

$$W_h = \mathrm{span}\left\{\boldsymbol{w}_j(x,y,z) \mid \text{all edges } j = 1, \ldots, N_e \text{ in } \mathcal{T}_h\right\},$$

where each basis function $\boldsymbol{w}_j(x,y,z)$ is defined with respect to edge j according to (5.5). The discretized version of (5.3) reads:

For given source function $\boldsymbol{J}$, find $\boldsymbol{E}_h \in W_h$, such that for all $\boldsymbol{W} \in W_h$ the following relation is satisfied

$$B(\boldsymbol{E}_h, \boldsymbol{W}) = (\boldsymbol{J}, \boldsymbol{W}). \tag{5.6}$$

5.3 Implicit error estimation

In this section we formulate the implicit error estimation method to estimate the error in each element of the domain. Also, appropriate local basis functions and boundary conditions are considered for the numerical solution of the local problems.

5.3.1 Formulation of the local error equation

Assume that $\boldsymbol{E}_h$ is a numerical solution computed using first order Nédélec elements. We aim to estimate the computational error $\boldsymbol{e}_h = (\boldsymbol{E} - \boldsymbol{E}_h)|_K$ on an element $K \in \mathcal{T}_h$, with $\mathcal{T}_h$ the finite element tessellation. For this we state a variational problem for the local error (see [65, 66]) on element K as follows:

Find $\boldsymbol{e}_h \in H(\mathrm{curl}, K)$ such that for all $\boldsymbol{v} \in H(\mathrm{curl}, K)$ the following relation is satisfied

$$\begin{aligned} B_K(\boldsymbol{e}_h, \boldsymbol{v}) &= (\mathrm{curl}\, \boldsymbol{e}_h, \mathrm{curl}\, \boldsymbol{v})_K - k^2(\boldsymbol{e}_h, \boldsymbol{v})_K \\ &= (\mathrm{curl}\, (\boldsymbol{E} - \boldsymbol{E}_h), \mathrm{curl}\, \boldsymbol{v})_K - k^2(\boldsymbol{E} - \boldsymbol{E}_h, \boldsymbol{v})_K \qquad (5.7) \\ &= (\mathrm{curl}\, \boldsymbol{E}, \mathrm{curl}\, \boldsymbol{v})_K - k^2(\boldsymbol{E}, \boldsymbol{v})_K - ((\mathrm{curl}\, \boldsymbol{E}_h, \mathrm{curl}\, \boldsymbol{v})_K - k^2(\boldsymbol{E}_h, \boldsymbol{v})_K) \\ &= (\mathrm{curl}\, \mathrm{curl}\, \boldsymbol{E}, \boldsymbol{v})_K - (\boldsymbol{\nu} \times \mathrm{curl}\, \boldsymbol{E}, \boldsymbol{v})_{\partial K} - k^2(\boldsymbol{E}, \boldsymbol{v})_K - B_K(\boldsymbol{E}_h, \boldsymbol{v})_K \\ &= (\boldsymbol{J}, \boldsymbol{v})_K - (\boldsymbol{\nu} \times \mathrm{curl}\, \boldsymbol{E}, \boldsymbol{v})_{\partial K} - B_K(\boldsymbol{E}_h, \boldsymbol{v}), \end{aligned}$$

where a Green's identity is applied in the fourth line and (5.1a) is used in the last line. In order to get a computable right hand side in (5.7) we use the approximation

$$\boldsymbol{\nu} \times \mathrm{curl}\, \boldsymbol{E} \approx \boldsymbol{\nu} \times \widehat{\mathrm{curl}\, \boldsymbol{E}} \quad \text{on interelement faces,} \qquad (5.8)$$

instead of using the unknown exact value $\boldsymbol{\nu} \times \mathrm{curl}\, \boldsymbol{E}$. The quantity $\boldsymbol{\nu} \times \mathrm{curl}\, \boldsymbol{E}$ will henceforth be called the *natural* boundary data. The following variational problem for the error on element K can now be formulated:

For a given source function $\boldsymbol{J}$ and numerical solution $\boldsymbol{E}_h$, find $\hat{\boldsymbol{e}}_h \in H(\mathrm{curl}, K)$ such that for all $\boldsymbol{v} \in H(\mathrm{curl}, K)$ the following relation is satisfied

$$B_K(\hat{\boldsymbol{e}}_h, \boldsymbol{v}) = (\boldsymbol{J}, \boldsymbol{v})_K - (\boldsymbol{\nu} \times \widehat{\mathrm{curl}\, \boldsymbol{E}}, \boldsymbol{v})_{\partial K} - B_K(\boldsymbol{E}_h, \boldsymbol{v}). \qquad (5.9)$$

5.3.2 Numerical solution of the local error equation

We will now give a discretized form of the local problem (5.9) which requires a specific choice for the approximation (5.8) of the natural boundary conditions and the finite element basis on element K.

Approximation of the natural boundary conditions

We first specify the approximation in (5.8). For the definition of the boundary conditions for the local error equation (5.7) we introduce f_j, the common face of the two neighboring elements K and K_j, and $\boldsymbol{\nu}_j$ the outward normal on f_j with respect to K. We approximate $\boldsymbol{\nu} \times \mathrm{curl}\, \boldsymbol{E}$ on f_j with the average of the tangential traces of the numerical approximation $\boldsymbol{E}_h$ on its two sides K and K_j. That is we shall use the approximation

$$\boldsymbol{\nu}_j \times \mathrm{curl}\, \boldsymbol{E}|_{f_j} \approx \frac{1}{2}(\boldsymbol{\nu}_j \times \left[\mathrm{curl}\, \boldsymbol{E}_h|_{\partial K \cap f_j} + \mathrm{curl}\, \boldsymbol{E}_h|_{\partial K_j \cap f_j}\right]), \qquad (5.10)$$

which can be straightforwardly implemented. It remains to supply the boundary conditions for (5.7) on elements which have a boundary face connected to the boundary of Ω.

Suppose that element K intersects with a portion of the boundary of the domain Ω where perfectly conducting boundary conditions are imposed. The appropriate boundary condition for the local error equation (5.7) is then

$$\boldsymbol{\nu} \times \hat{\boldsymbol{e}}_h = 0 \text{ on } \partial K \cap \partial\Omega. \tag{5.11}$$

Here, it is assumed that the finite element approximation has been constructed so that the perfectly conducting boundary conditions are satisfied exactly, for details see [3].

Choice of the local basis

As discussed in [3], Section 3.4.2, the finite dimensional space used to discretize the local error equations (5.9) has to be selected carefully. In case of elliptic boundary value problems a different local basis is considered in [3] for the solution of the local error equations. It is advocated there that the use of different basis functions than those used for the original problem results in a better approximation of the error. For the Maxwell equations it is also beneficial to use higher order polynomials for the error equation which is explained by the fact that the dominant term in the error is associated with polynomials of a degree which is one order higher than used to approximate the field, see [21, 81, 85]. In our numerical experiments we observe similar phenomena. If we use first order Nédélec elements to solve the local problems then the computed error does not describe the true error and leads to a non-physical solution. If we use the full second order Nédélec elements again the obtained results are poor, see Section 5.7.4. This is due to the linear part present in the basis. Therefore, as a basis for the solution of the local error equations, we use the second order Nédélec edge basis functions with the linear basis functions removed.

Again, the basis functions for the local problem are first defined on a reference tetrahedron and then with the covariant transformation (5.5) transformed to the physical elements. There are eight face based basis functions defined as

$$\begin{aligned}
\boldsymbol{\phi}_1^0 &= L_2L_3\nabla L_4 - L_2L_4\nabla L_3, & \boldsymbol{\phi}_2^0 &= L_2L_3\nabla L_4 - L_3L_4\nabla L_2,\\
\boldsymbol{\phi}_3^0 &= L_1L_3\nabla L_4 - L_1L_4\nabla L_3, & \boldsymbol{\phi}_4^0 &= L_1L_3\nabla L_4 - L_3L_4\nabla L_1,\\
\boldsymbol{\phi}_5^0 &= L_1L_2\nabla L_4 - L_1L_4\nabla L_2, & \boldsymbol{\phi}_6^0 &= L_1L_2\nabla L_4 - L_2L_4\nabla L_1,\\
\boldsymbol{\phi}_7^0 &= L_1L_2\nabla L_3 - L_1L_3\nabla L_2, & \boldsymbol{\phi}_8^0 &= L_1L_2\nabla L_3 - L_2L_3\nabla L_1,
\end{aligned}$$

or, in more explicit form,

$$\begin{aligned}
\boldsymbol{\phi}_1^0 &= (0, -\xi\zeta, \xi\eta)^T, & \boldsymbol{\phi}_2^0 &= (-\eta\zeta, 0, \xi\eta)^T, \\
\boldsymbol{\phi}_3^0 &= (0, -(1-\xi-\eta-\zeta)\zeta, (1-\xi-\eta-\zeta)\eta)^T, & \boldsymbol{\phi}_4^0 &= (\eta\zeta, \eta\zeta, (1-\xi-\eta)\eta)^T, \\
\boldsymbol{\phi}_5^0 &= (-(1-\xi-\eta-\zeta)\zeta, 0, (1-\xi-\eta-\zeta)\xi)^T, & \boldsymbol{\phi}_6^0 &= (\xi\zeta, \xi\zeta, (1-\xi-\eta)\xi)^T, \\
\boldsymbol{\phi}_7^0 &= (\eta(1-\xi-\eta-\zeta), \xi(1-\xi-\eta-\zeta), 0)^T, & \boldsymbol{\phi}_8^0 &= (\xi\eta, (1-\xi-\zeta)\xi, \xi\eta)^T.
\end{aligned}$$

These basis functions are transformed to a tetrahedron $K \in \mathcal{T}_h$ with the covariant transformation as

$$\boldsymbol{\phi}_j(x, y, z) = (\mathrm{d}D_K^{-1})^T \boldsymbol{\phi}_j^0(\xi, \eta, \zeta), \quad j = 1, \ldots, 8, \tag{5.12}$$

with D_K the transformation defined in (5.4). This reduced finite element space on an element K is denoted by $\mathcal{N}_2^2(K)$:

$$\mathcal{N}_2^2(K) = \mathrm{span}\{\boldsymbol{\phi}_j\}_{j=1,\ldots,8}.$$

For more details on the construction of second order Nédélec elements we refer to [73, 93].

Weak form of the local error equation

Using approximation (5.10) and the local basis $\mathcal{N}_2^2(K)$ we obtain the discrete form of the local error equation (5.9):

For a given source function $\boldsymbol{J}$ and numerical solution $\boldsymbol{E}_h$, find $\hat{\boldsymbol{e}}_h \in \mathcal{N}_2^2(K)$ such that for all $\boldsymbol{w} \in \mathcal{N}_2^2(K)$ the following relation is satisfied

$$\begin{aligned}
(\mathrm{curl}\, \hat{\boldsymbol{e}}_h, \mathrm{curl}\, \boldsymbol{w})_K - k^2(\hat{\boldsymbol{e}}_h, \boldsymbol{w})_K &= (\boldsymbol{J}, \boldsymbol{w})_K - (\mathrm{curl}\, \boldsymbol{E}_h, \mathrm{curl}\, \boldsymbol{w})_K \\
+k^2(\boldsymbol{E}_h, \boldsymbol{w})_K - \frac{1}{2}(\boldsymbol{\nu}_j \times \left(\mathrm{curl}\, \boldsymbol{E}_h|_K + \mathrm{curl}\, \boldsymbol{E}_h|_{K_j}\right), \boldsymbol{w})_{\partial K}.
\end{aligned} \tag{5.13}$$

5.3.3 Properties of the local error estimator

We investigate the existence and uniqueness of the local error estimate and state that it provides a lower bound (up to a constant) for the exact error $\boldsymbol{e}_h$.

Well posedness of the local error equation

Using a lifting operator we can associate an $\bar{\boldsymbol{e}}_h$ to $\hat{\boldsymbol{e}}_h$ and define a function $\hat{\boldsymbol{J}}_K \in [L_2(\Omega)]^3$ such that the well posedness of (5.9) is equivalent with that of the variational problem:

Find an $\bar{\boldsymbol{e}}_h \in H(\mathrm{curl}, K)$ such that for all $\boldsymbol{v} \in H(\mathrm{curl}, K)$ the following relation is satisfied

$$B_K(\bar{\boldsymbol{e}}_h, \boldsymbol{v}) = (\hat{\boldsymbol{J}}_K, \boldsymbol{v}). \tag{5.14}$$

For the details we refer to Section 4.3.3 or [65, 66], Section 3.3.1.

The well posedness of (5.14) is stated in the following:

Lemma 5.1. *Assume that k is not a Maxwell eigenvalue on K in the sense that only $\boldsymbol{u} = 0 \in H(\mathrm{curl}, K)$ satisfies the relation*

$$B_K(\boldsymbol{u}, \boldsymbol{v}) = 0, \ \forall \boldsymbol{v} \in H(\mathrm{curl}, K).$$

Then the variational problem (5.14) *has a unique solution.*

For the proof we refer to Section 4.3.3 or [65, 66], Section 3.3.3.

In order to apply Lemma 5.1 we need to ensure that k is not a Maxwell eigenvalue on K for all kind of tetrahedra arising in the finite element tessellation $\mathcal{T}_h$. Instead of performing a detailed analysis for this, we rather ensure well posedness for the discretized problems in (5.14) by proving an inf-sup condition, which is discussed in Section 5.4.

Efficiency of the local error estimate

We state that the error estimate $\hat{\boldsymbol{e}}_h$ is efficient which means that it is bounded by the analytic error plus higher order terms (for a precise definition see [26]). For this we use the notations:

$$\boldsymbol{r}_K = \boldsymbol{J} - \mathrm{curl}\,\mathrm{curl}\,\boldsymbol{E}_h + k^2 \boldsymbol{E}_h \quad \text{in } K$$

for the residual within the subdomain K and

$$\mathbf{R}_{l_j} = \frac{1}{2}(\boldsymbol{\nu}_j \times \left[\mathrm{curl}\,\boldsymbol{E}_h|_K - \mathrm{curl}\,\boldsymbol{E}_h|_{K_j}\right])$$

for the tangential jump of the curl at the common face l_j of K and K_j. We also introduce $\bar{\boldsymbol{r}}$ as the approximation of $\boldsymbol{r}$ in the finite element space $\mathcal{N}_2^2(K)$. Similarly, $\bar{\mathbf{R}}$ denotes the approximation of $\mathbf{R}$ on ∂K with the trace of functions in $\mathcal{N}_2^2(K)$ and the patch $\tilde{K}$ of K is defined as follows:

$$\tilde{K} = \{\cup K_i : K_i \in \mathcal{T}_h, \overline{K} \cap \overline{K}_i \neq \emptyset\}.$$

Theorem 5.2. *If* $\mathrm{diam}\, K = h < \frac{h_*}{k}$ *for some positive constant h_* then the error estimate $\hat{\boldsymbol{e}}_h$ is efficient,*

$$\|\hat{\boldsymbol{e}}_h\|^2_{\mathrm{curl},K} \leq C((1+k^2)^2 \|\boldsymbol{e}_h\|^2_{\mathrm{curl},\tilde{K}} + h^2 \|\bar{\boldsymbol{r}} - \boldsymbol{r}\|^2_{[L_2(K)]^3} + h\|\bar{\mathbf{R}} - \mathbf{R}\|^2_{[L_2(\partial K)]^3}), \tag{5.15}$$

where C does not depend on h.

The proof is postponed to Section 5.5.

5.4 Inf-sup condition for the implicit error estimator

In this section we show that the computations using the implicit error estimation technique are stable in the sense that the local matrices in the bilinear form B_K in (5.9) remain uniformly well conditioned. Equivalently, we prove that they satisfy the inf-sup condition uniformly.

Theorem 5.3. *The bilinear form $B_K : \mathcal{N}_2^2(K) \times \mathcal{N}_2^2(K) \to \mathbb{R}$ satisfies the inf-sup condition uniformly in K; namely there is a positive constant h_0 such that for any non-degenerate $K \in \mathcal{T}_h$ with $\operatorname{diam} K < h_0$ and any $\boldsymbol{u} \in \mathcal{N}_2^2(K)$*

$$\sup_{\boldsymbol{v} \in \mathcal{N}_2^2(K)} \frac{B_K(\boldsymbol{u}, \boldsymbol{v})}{\|\boldsymbol{v}\|_{\operatorname{curl},K}} \geq \min\{\frac{1}{2}, k^2\} \|\boldsymbol{u}\|_{\operatorname{curl},K}.$$

To prove this theorem we first give the explicit expression of the bilinear form B_K in terms of the original basis functions. Using (5.12) we obtain that for any $\boldsymbol{v} = \sum_{i=1}^{8} v_i \boldsymbol{\phi}_i \in \mathcal{N}_2^2(K)$

$$\begin{aligned} (\boldsymbol{v}, \boldsymbol{v})_K &= (\sum_{i=1}^{8} v_i \boldsymbol{\phi}_i, \sum_{j=1}^{8} v_j \boldsymbol{\phi}_j)_K \\ &= |\det \mathrm{d}D_K| ((\mathrm{d}D_K^{-1})^T \sum_{i=1}^{8} v_i \boldsymbol{\phi}_i^0, (\mathrm{d}D_K^{-1})^T \sum_{j=1}^{8} v_j \boldsymbol{\phi}_j^0)_{\hat{K}}. \end{aligned} \tag{5.16}$$

Using (5.12) one can easily prove (see [73], Corollary 3.58) that

$$\operatorname{curl}_{x,y,z} \boldsymbol{\phi}_j = \frac{1}{\det \mathrm{d}D_K} \mathrm{d}D_K \operatorname{curl}_{\xi,\eta,\zeta} \boldsymbol{\phi}_j^0, \quad j = 1, 2, \ldots, 8. \tag{5.17}$$

Therefore,

$$\begin{aligned} (\operatorname{curl} \boldsymbol{v}, \operatorname{curl} \boldsymbol{v})_K &= (\operatorname{curl}_{x,y,z} \sum_{i=1}^{8} v_i \boldsymbol{\phi}_i, \operatorname{curl}_{x,y,z} \sum_{j=1}^{8} v_j \boldsymbol{\phi}_j)_K \\ &= \frac{1}{|\det \mathrm{d}D_K|} (\mathrm{d}D_K \operatorname{curl}_{\xi,\eta,\zeta} \sum_{i=1}^{8} v_i \boldsymbol{\phi}_i^0, \mathrm{d}D_K \operatorname{curl}_{\xi,\eta,\zeta} \sum_{j=1}^{8} v_j \boldsymbol{\phi}_j^0)_{\hat{K}}. \end{aligned} \tag{5.18}$$

We also use the following two lemmas:

Lemma 5.4. *Let us denote with $\mathcal{T}_h^1$ a non-degenerate family of tetrahedra such that* $\det \mathrm{d}D_K = 1$ *for all* $K \in \mathcal{T}_h^1$. *Then there is a compact set* $\mathcal{D} \subset \mathbb{R}^{3\times 3}$ *such that* $0 \notin \mathcal{D}$ *and* $\mathrm{d}D_K \in \mathcal{D}$ *for all* $K \in \mathcal{T}_h^1$.

Proof We use the notion of the spectral norm which is given for an arbitrary matrix $D \in \mathbb{R}^{n\times n}$ as

$$\|D\|_{\mathrm{sp}} = \sup_{|\xi|=1} |D\xi|. \tag{5.19}$$

We can establish the lemma if we prove that there are positive constants C_1, C_2 such that for any $K \in \mathcal{T}_h^1$ the following inequality holds:

$$C_1 \le \|\mathrm{d}D_K\|_{\mathrm{sp}} \le C_2. \tag{5.20}$$

Then the set

$$\mathcal{D} = \overline{\{\mathrm{d}D_K : K \in \mathcal{T}_h^1\}}$$

is closed, and bounded with respect to the spectral norm (which is equivalent with any norm in $\mathbb{R}^{3\times 3}$), moreover, the condition $\det \mathrm{d}D_K = 1, \forall\ K \in \mathcal{T}_h^1$ implies that $0 \notin \mathcal{D}$.

Note that the condition $\det \mathrm{d}D_K = 1$ implies that the volume of any tetrahedron K is the same as that of $\hat{K}$.

Indirectly, assume first that there is no constant C_2 in (5.20), i.e. there is a sequence K_n such that $\|\mathrm{d}D_{K_n}\|_{\mathrm{sp}} > n$. Then according to Lemma 5.10 in [73]

$$n < \|\mathrm{d}D_{K_n}\|_{\mathrm{sp}} \le \frac{h_{K_n}}{\rho_{\hat{K}}},$$

where $h_{K_n} = \operatorname{diam} K_n$ and $\rho_{\hat{K}}$ denotes the radius of the largest ball contained in the reference element $\hat{K}$. Then $\lim_{n\to\infty} h_{K_n} = \infty$ while the condition on the volume implies that ρ_{K_n} remains bounded. This contradicts to the non-degenerate property of the meshes.

Assume now that there is no positive lower bound C_1 in (5.20). Then according to (5.19) there is a sequence K_n such that

$$\max |\mathrm{eig}\, \mathrm{d}D_{K_n}| \le \|\mathrm{d}D_{K_n}\|_{\mathrm{sp}} = \sup_{|\xi|=1} |\mathrm{d}D_{K_n}\xi| < \frac{1}{n}, \tag{5.21}$$

where eig denotes an eigenvalue. Since the determinant is the product of the eigenvalues, this is again a contradiction. ■

Lemma 5.5. *For **any** $C > 0$ there is a positive number $h_0 \in \mathbb{R}^+$ such that for every non-degenerate tetrahedron $K \in \mathcal{T}_h$ with $\operatorname{diam} K \le h_0$ and $\boldsymbol{u}_2 \in \mathcal{N}_2^2(K)$ with $\boldsymbol{u}_2 \perp \ker \operatorname{curl}$ the following inequality holds:*

$$(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K \ge C(\boldsymbol{u}_2, \boldsymbol{u}_2)_K. \tag{5.22}$$

Proof: First, we decompose the transformation D_K as follows: $D_K = D_K \tilde{D}_K^{-1} \circ \tilde{D}_K$, where

$$\tilde{D}_K = \frac{1}{\sqrt[3]{\det \mathrm{d}D_K}} D_K : \hat{K} \to \tilde{K} \tag{5.23}$$

and

$$D_K \tilde{D}_K^{-1} = \sqrt[3]{\det \mathrm{d}D_K} I : \tilde{K} \to K, \tag{5.24}$$

where the corresponding matrices are denoted with $\mathrm{d}D_K$, $\mathrm{d}\tilde{D}_K$ and $\mathrm{d}D_K \mathrm{d}\tilde{D}_K^{-1}$, respectively, and $\det \mathrm{d}\tilde{D}_K = 1$.

For a function $\boldsymbol{u}_2 = \sum_{i=1}^{8} u_{2,i} \boldsymbol{\phi}_i^0 \in \{\operatorname{span} \boldsymbol{\phi}_j^0\}_{j=1,\dots,8}$ we define

$$\tilde{\boldsymbol{u}}_2 : \tilde{K} \to \mathbb{R}^3 \quad \text{with} \quad \tilde{\boldsymbol{u}}_2 = \sum_{i=1}^{8} u_{2,i} \tilde{\boldsymbol{\phi}}_i,$$

where the basis functions $\tilde{\boldsymbol{\phi}}_i : \tilde{K} \to \mathbb{R}^3$ $(i = 1, 2, \dots, 8)$ are defined using (5.12) with the transformation $\tilde{D}_K$ instead of D_K. Using (5.16) for the linear mapping $D_K \tilde{D}_K^{-1}$ we obtain that

$$(\boldsymbol{u}_2, \boldsymbol{u}_2)_K = |\det \mathrm{d}D_K| \frac{1}{(\sqrt[3]{\det \mathrm{d}D_K})^2} (\tilde{\boldsymbol{u}}_2, \tilde{\boldsymbol{u}}_2)_{\tilde{K}} \tag{5.25}$$

and using (5.18) gives that

$$(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K = \frac{1}{|\det \mathrm{d}D_K|} (\sqrt[3]{\det \mathrm{d}D_K})^2 (\operatorname{curl} \tilde{\boldsymbol{u}}_2, \operatorname{curl} \tilde{\boldsymbol{u}}_2)_{\tilde{K}}. \tag{5.26}$$

Using then (5.16) and (5.18) and the transformation formula (5.12) we obtain that for $\boldsymbol{u}_2 = \sum_{i=1}^{8} u_{2,i} \boldsymbol{\phi}_i^0 \in \{\operatorname{span} \boldsymbol{\phi}_j^0\}_{j=1,\dots,8} \cap \ker \operatorname{curl}^\perp$ (which can be identified with the coefficients $u_{2,i}$) and $\mathrm{d}\tilde{D}_K \in \mathbb{R}^{3\times 3}$ the mapping of type $\mathbb{R}^8 \times \mathbb{R}^{3\times 3} \to \mathbb{R}$ defined by

$$[\boldsymbol{u}_2, \mathrm{d}\tilde{D}_K] \to \frac{(\operatorname{curl} \tilde{\boldsymbol{u}}_2, \operatorname{curl} \tilde{\boldsymbol{u}}_2)_{\tilde{K}}}{(\tilde{\boldsymbol{u}}_2, \tilde{\boldsymbol{u}}_2)_{\tilde{K}}} \tag{5.27}$$

is a continuous function of type $\mathbb{R}^8 \times \mathbb{R}^{3\times 3} \to \mathbb{R}^+$. We may assume that it is given only on the unit sphere of $\mathbb{R}^8$, since $\lambda \boldsymbol{u}_2$ and $\boldsymbol{u}_2$ result in the same values in (5.27). In this way, the mapping in (5.27) is given on a compact set,

see Lemma 5.4. Therefore its infimum equals to its minimum, which should be positive. Using this with the relations in (5.25) and (5.26) we obtain that for any $\mathrm{d}D_K \in \mathbb{R}^{3\times 3}$ and $(u_{2,1}, u_{2,2}, \ldots, u_{2,8}) \in \mathbb{R}^8$

$$0 < \tilde{c} \leq \frac{(\operatorname{curl} \tilde{\boldsymbol{u}}_2, \operatorname{curl} \tilde{\boldsymbol{u}}_2)_{\tilde{K}}}{(\tilde{\boldsymbol{u}}_2, \tilde{\boldsymbol{u}}_2)_{\tilde{K}}} = (\sqrt[3]{\det \mathrm{d}D_K})^2 \frac{(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K}{(\boldsymbol{u}_2, \boldsymbol{u}_2)_K}. \tag{5.28}$$

Obviously, $(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K \geq \dfrac{\tilde{c}}{(\sqrt[3]{\det \mathrm{d}D_K})^2}(\boldsymbol{u}_2, \boldsymbol{u}_2)_K$, and $\det \mathrm{d}D_K \to 0$ as the diameter of K converges to zero, then for some h_0 we will have $\dfrac{\tilde{c}}{(\sqrt[3]{\det \mathrm{d}D_K})^2} \geq C$ in (5.22), which proves the lemma. ■

Proof of Theorem 5.3: Decompose $\boldsymbol{u} \in \mathcal{N}_2^2(K)$ as $\boldsymbol{u} = \boldsymbol{u}_1 + \boldsymbol{u}_2$, where $\operatorname{curl}\boldsymbol{u}_1 = 0$ and $\boldsymbol{u}_2 \perp \ker \operatorname{curl}$. Then, for a given $\boldsymbol{u}$ choose $\boldsymbol{v} = \boldsymbol{u}_1 - \boldsymbol{u}_2$ and with this

$$\begin{aligned} |B_K(\boldsymbol{u}, \boldsymbol{v})| &= |-(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K - k^2(\boldsymbol{u}_1 + \boldsymbol{u}_2, \boldsymbol{u}_1 - \boldsymbol{u}_2)_K| \\ &= |(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K - k^2(\boldsymbol{u}_2, \boldsymbol{u}_2)_K + k^2(\boldsymbol{u}_1, \boldsymbol{u}_1)_K|. \end{aligned} \tag{5.29}$$

On the other hand,

$$\begin{aligned} \|\boldsymbol{u}\|_{\operatorname{curl},K}\|\boldsymbol{v}\|_{\operatorname{curl},K} &= \|\boldsymbol{u}_1 + \boldsymbol{u}_2\|_{\operatorname{curl},K}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{\operatorname{curl},K} \\ &= (\|\boldsymbol{u}_1\|^2_{\operatorname{curl},K} + \|\boldsymbol{u}_2\|^2_{\operatorname{curl},K})^{\frac{1}{2}}(\|\boldsymbol{u}_1\|^2_{\operatorname{curl},K} + \|\boldsymbol{u}_2\|^2_{\operatorname{curl},K})^{\frac{1}{2}} \\ &= \|\boldsymbol{u}_1\|^2_{\operatorname{curl},K} + \|\boldsymbol{u}_2\|^2_{\operatorname{curl},K} \\ &= (\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K + (\boldsymbol{u}_2, \boldsymbol{u}_2)_K + (\boldsymbol{u}_1, \boldsymbol{u}_1)_K. \end{aligned} \tag{5.30}$$

Using Lemma 5.5 with $C = 2k^2 + 1$ there is an $h_0 > 0$ such that for any K with $\operatorname{diam} K < h_0$

$$\begin{aligned} &(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K - k^2(\boldsymbol{u}_2, \boldsymbol{u}_2)_K \\ &\geq \frac{1}{2}(\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K + (k^2 + \frac{1}{2})(\boldsymbol{u}_2, \boldsymbol{u}_2)_K - k^2(\boldsymbol{u}_2, \boldsymbol{u}_2)_K \\ &= \frac{1}{2}((\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K + (\boldsymbol{u}_2, \boldsymbol{u}_2)_K). \end{aligned}$$

Inserting this into (5.29) and using (5.30) we obtain that for every K, with $\operatorname{diam} K < h_0$, that

$$\begin{aligned} |B_K(\boldsymbol{u}, \boldsymbol{v})| &\geq (\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K - k^2(\boldsymbol{u}_2, \boldsymbol{u}_2)_K + k^2(\boldsymbol{u}_1, \boldsymbol{u}_1)_K \\ &\geq \min\{\frac{1}{2}, k^2\}((\operatorname{curl} \boldsymbol{u}_2, \operatorname{curl} \boldsymbol{u}_2)_K + (\boldsymbol{u}_2, \boldsymbol{u}_2)_K + (\boldsymbol{u}_1, \boldsymbol{u}_1)_K) \\ &= \min\{\frac{1}{2}, k^2\}\|\boldsymbol{u}\|_{\operatorname{curl},K}\|\boldsymbol{v}\|_{\operatorname{curl},K}. \end{aligned}$$

Summarized, there is a $h_0 > 0$ such that for any non-degenerate tetrahedron K with $\operatorname{diam} K < h_0$ and for an arbitrary $\boldsymbol{u} \in \mathcal{N}_2^2(K)$ one can find $\boldsymbol{v} \in \mathcal{N}_2^2(K)$ such that

$$|B_K(\boldsymbol{u}, \boldsymbol{v})| \geq \min\{\frac{1}{2}, k^2\} \|\boldsymbol{u}\|_{\operatorname{curl},K} \|\boldsymbol{v}\|_{\operatorname{curl},K}.$$

Dividing both sides with $\|\boldsymbol{v}\|_{\operatorname{curl},K}$ gives the statement of the theorem. ■

5.4.1 Dependence of the estimates on the wave number

We can sharpen the result of Theorem 5.3 further and compute the dependence of the critical mesh size h_0 on the wavenumber k. Accordingly, we use the notation $B_{K,\alpha}$ for the bilinear form on $H(\operatorname{curl}, K) \times H(\operatorname{curl}, K)$ with

$$B_{K,\alpha}(\boldsymbol{u}, \boldsymbol{v}) = (\operatorname{curl} \boldsymbol{u}, \operatorname{curl} \boldsymbol{v})_K - (\alpha k)^2 (\boldsymbol{u}, \boldsymbol{v})_K,$$

where $\alpha > 1$ is a given parameter.

Lemma 5.6. *Assume that Theorem 5.3 holds for the wavenumber k with the constant h_0. Then for any $K \in \mathcal{T}_h$ with $\operatorname{diam} K < \frac{1}{\alpha} h_0$, any $\alpha > 1$ and any $\boldsymbol{u} \in \mathcal{N}_2^2(K)$ we have for the wave number αk the inf-sup condition*

$$\sup_{\boldsymbol{v} \in \mathcal{N}_2^2(K)} \frac{B_{K,\alpha k}(\boldsymbol{u}, \boldsymbol{v})}{\|\boldsymbol{v}\|_{\operatorname{curl},K}} \geq \min\{\frac{1}{2}, k^2\} \|\boldsymbol{u}\|_{\operatorname{curl},K} \tag{5.31}$$

Proof We use (5.16) and (5.18) in the case when $\tilde{K}$ is a tetrahedron with $\operatorname{diam} \tilde{K} < h_0$ and $D_\alpha : \tilde{K} \to K$ is defined by $D_\alpha = \frac{1}{\alpha} I$.

$$\begin{aligned}
&\sup_{\boldsymbol{v} \in \mathcal{N}_2^2(K)} \frac{B_{K,\alpha}(\boldsymbol{u}, \boldsymbol{v})}{\|\boldsymbol{v}\|_{\operatorname{curl},K}} = \sup_{\boldsymbol{v} \in \mathcal{N}_2^2(K)} \frac{(\operatorname{curl} \boldsymbol{u}, \operatorname{curl} \boldsymbol{v})_K - (\alpha k)^2 (\boldsymbol{u}, \boldsymbol{v})_K}{\|\boldsymbol{v}\|_{\operatorname{curl},K}} \\
&= \sup_{\tilde{\boldsymbol{v}} \in \mathcal{N}_2^2(\tilde{K})} \frac{\alpha (\operatorname{curl} \tilde{\boldsymbol{u}}, \operatorname{curl} \tilde{\boldsymbol{v}})_{\tilde{K}} - \frac{1}{\alpha} (\alpha k)^2 (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})_{\tilde{K}}}{\sqrt{\alpha (\operatorname{curl} \tilde{\boldsymbol{v}}, \operatorname{curl} \tilde{\boldsymbol{v}})_{\tilde{K}} + \frac{1}{\alpha} (\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{v}})_{\tilde{K}}}} \\
&\geq \sqrt{\alpha} \sup_{\tilde{\boldsymbol{v}} \in \mathcal{N}_2^2(\tilde{K})} \frac{(\operatorname{curl} \tilde{\boldsymbol{u}}, \operatorname{curl} \tilde{\boldsymbol{v}})_{\tilde{K}} - k^2 (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})_{\tilde{K}}}{\|\tilde{\boldsymbol{v}}\|_{\operatorname{curl},\tilde{K}}} \\
&\geq \min\{\frac{1}{2}, k^2\} \sqrt{\alpha} \sqrt{\|\operatorname{curl} \tilde{\boldsymbol{u}}\|^2_{[L_2(\tilde{K})]^3} + \|\tilde{\boldsymbol{u}}\|^2_{[L_2(\tilde{K})]^3}} \\
&= \min\{\frac{1}{2}, k^2\} \sqrt{\alpha} \sqrt{\frac{1}{\alpha} \|\operatorname{curl} \boldsymbol{u}\|^2_{[L_2(K)]^3} + \alpha \|\boldsymbol{u}\|^2_{[L_2(K)]^3}} \\
&= \min\{\frac{1}{2}, k^2\} \sqrt{\|\operatorname{curl} \boldsymbol{u}\|^2_{[L_2(K)]^3} + \alpha^2 \|\boldsymbol{u}\|^2_{[L_2(K)]^3}} \geq \min\{\frac{1}{2}, k^2\} \|\boldsymbol{u}\|_{\operatorname{curl},K},
\end{aligned}$$

where (5.16) and (5.18) were applied in the second line, Theorem 5.3 in the fourth line and again (5.16) and (5.18) in the fifth line. ■

Lemma 5.6 shows that for the inf-sup condition we only need that kh is smaller than some positive constant.

Using Lemma 5.6 and some results in [66] we can prove the k-dependence of the constant C on the right hand side of (5.15).

Proof of Theorem 5.2: According to Theorem 4.16

$$\|\hat{\boldsymbol{e}}_h\|^2_{\mathrm{curl},K} \leq C_1^2 C_2((1+k^2)^2\|\hat{\boldsymbol{e}}_h\|^2_{\mathrm{curl},K} + h^2\|\bar{\boldsymbol{r}} - \boldsymbol{r}\|^2_{[L_2(K)]^3} + h\|\bar{\mathbf{R}} - \mathbf{R}\|^2_{[L_2(K)]^3}),$$

where C_2 does not depend on h and k, and $C_1 = \frac{1}{\min\{\frac{1}{2}, k^2\}}$ is the inverse of the constant in the inf-sup condition. For wave numbers $k \geq \sqrt{\frac{1}{2}}$ we obtain $C_1^2 = 4$.

The proof was carried out for rectangular elements in Chapter 4, see also [66], but it is applicable also for tetrahedral elements if Section 4.4 is changed accordingly. This requires a standard bubble function technique which only uses the non-degenerate property of the mesh. We omit this straightforward but lengthy analysis. ■

5.5 Computational costs

For numerical simulations it is not only important to have an accurate but also an efficient algorithm. We show that an adaptation technique based on the implicit a posteriori error estimator developed in this chapter is more efficient than the global refinement technique. In all numerical experiments the linear systems associated with (5.6) are solved with the MINRES iterative solver using diagonal preconditioning, see e.g. [52, 92, 103].

The computational work of the adaptive finite element solution on each mesh, denoted by $\texttt{mesh}_i$, consists of the following steps:

- E flops for the element-wise computation of the implicit error estimate on the mesh $\texttt{mesh}_i$,
- Q_a flops for the solution of the linear system for the Maxwell equations on the adapted mesh using MINRES,
- M flops for the mesh generation to obtain the next adapted mesh $\texttt{mesh}_{i+1}$.

The overall computational work associated with the implicit a posteriori error estimator is:

$$\mathsf{W}_c = \mathsf{Q}_a + \mathsf{E} + \mathsf{M}.$$

The required computational work on the globally refined mesh of $\texttt{mesh}_i$ is denoted by W_f flops. It mainly consists of the work required for the solution of the linear system resulting from the discretization of the Maxwell equations on this globally refined mesh. The aim is to compare the two quantities W_c and W_f.

In all our numerical experiments we have observed that for each tetrahedral finite element mesh the following relation between the number of elements (N) and the number of the edges (N_e) holds:

$$1.2 \cdot N \approx N_e.$$

Provided that this relation holds and that after global refinement of mesh $\texttt{mesh}_i$ each element in the mesh is subdivided into 8 elements, the size of the matrix on the globally refined mesh is 8 times larger than on the mesh $\texttt{mesh}_i$. As was observed experimentally, the computational work required for solving a linear system of equations of size $M \times M$ with the MINRES iterative solver is approximately $C_o \cdot M^2$, with the constant C_o independent of the finite element mesh. This means that the computational work required to solve the linear system of equations on the globally refined mesh is approximately 8^2 times more than on the mesh $\texttt{mesh}_i$. If we define the computational work required for the solution of the linear system on the mesh $\texttt{mesh}_i$ by Q_0 using MINRES, then the claimed work on the globally refined mesh is

$$\mathsf{W}_f \approx 64 \cdot \mathsf{Q}_0.$$

For the implicit a posteriori error estimation on each element we have to solve a small linear system of equations. The Gauss elimination method for an $n \times n$ matrix with partial pivoting requires approximately $\frac{2}{3} \cdot n^3$ operations. The size of the local matrices used for the implicit error estimation is 8×8, therefore the required computational cost for solving the local linear systems is approximately $\frac{2}{3} \cdot 8^3$ per element. Hence we obtain

$$\mathsf{E} \approx \frac{2}{3} \cdot 8^3 \cdot N \approx 341 \cdot N.$$

Based on the adaptation algorithm we allow a growth in the number of degrees of freedom with a factor of at most 3.5, see equation (5.37), which means that the size of the matrix for the finite element discretization of the Maxwell equations on the adapted mesh $\texttt{mesh}_{i+1}$ is at most 3.5 times larger than on the previous mesh $\texttt{mesh}_i$. It follows that the solution of the linear system on the adapted mesh with MINRES will therefore require at most $3.5^2 = 12.25$ times more operations as compared to the operations on the mesh $\texttt{mesh}_i$, i.e.

$$\mathsf{Q}_a \approx 12.25 \cdot \mathsf{Q}_0.$$

If we summarize the above information we obtain

$$\begin{aligned} \mathsf{W}_c &\approx 12.25 \cdot \mathsf{Q}_0 + 341 \cdot N + \mathsf{M} \\ &\approx 12.25 \cdot \mathsf{Q}_0 + 284 \cdot N_e + \mathsf{M}, \end{aligned}$$

where the dimension of the matrix on the coarse mesh is $N_e \times N_e$. The computational work required for solving the linear system of equations with the MINRES iterative solver on the mesh $\texttt{mesh}_i$ is approximately $\mathsf{Q}_0 \approx C_o \cdot N_e^2$, then we obtain

$$\mathsf{W}_c \approx 12.25 \cdot C_o \cdot N_e^2 + 284 \cdot N_e + \mathsf{M}.$$

It is realistic to assume that $\mathsf{M} \ll C_o \cdot N_e^2$, and, hence, the dominant term in the last expression is the computational work required for the solution of the linear system resulting from the finite element discretization of the Maxwell equations. Comparing the obtained estimates for W_c and W_f, we can easily see that the computational work on the globally refined mesh requires approximately $\frac{64}{12.25} \approx 5.2$ times more work than for the adaptive finite element solution.

In Section 5.7.1 the computational work in terms of CPU time of the adaptive and global refinement algorithms are compared for the solution of the Maxwell equations on a cylindrical domain.

5.6 Adaptive mesh generation

In this section we describe how to use the implicit a posteriori error estimation technique in real applications. Let us define the exact error δ_K, which is unknown in practice, and the implicit *local error estimate* (*indicator*) $\hat{\delta}_K$ on element K by

$$\delta_K = \|\boldsymbol{E} - \boldsymbol{E}_h\|_{\mathrm{curl},K}, \quad \hat{\delta}_K = \|\hat{\boldsymbol{e}}_h\|_{\mathrm{curl},K}. \tag{5.32}$$

Recall that $\boldsymbol{E}_h$ denotes the numerical solution of the Maxwell equations (5.1) obtained by using the first order edge finite elements (see Section 5.2.1) and that $\hat{\boldsymbol{e}}_h$ denotes the computed error with the implicit error estimator, defined in (5.13), and solved with the help of the finite element space $\mathcal{N}_2^2(K)$ (see Section 5.3.2).

The exact global error δ and the implicit global error estimate δ_h can be obtained as

$$\delta = \left(\sum_{K \in \mathcal{T}_h} \delta_K^2 \right)^{1/2}, \quad \delta_h = \left(\sum_{K \in \mathcal{T}_h} \hat{\delta}_K^2 \right)^{1/2}. \tag{5.33}$$

Accordingly, if we sum up the terms terms in (5.15), we obtain

$$\delta_h \leq C(1+k^2)\delta,$$

where C does not depend on h and k.

For a given tolerance TOL we aim to construct a mesh $\mathcal{T}_h$ such that

$$\delta_h < \texttt{TOL}. \tag{5.34}$$

There are several adaptation strategies to achieve this.

Strategy 1. In this strategy, proposed in [44], the algorithm tries to equidistribute the local error over all elements of $\mathcal{T}_h$. Thus, we insist that for all elements K in the tessellation $\mathcal{T}_h$ the condition $\hat{\delta}_K \approx \dfrac{\texttt{TOL}}{\sqrt{N}}$ is satisfied, where N denotes the total number of elements in the tessellation. Element K in the mesh $\mathcal{T}_h$ is marked for refinement if

$$\hat{\delta}_K > \frac{\texttt{TOL}}{\sqrt{N}}.$$

Strategy 2. This algorithm is based on an area-weighted tolerance approach. For the given element K denote by V_K its volume. Then element K is marked for refinement if

$$\hat{\delta}_K > \texttt{TOL}\sqrt{\frac{V_K}{V_\Omega}},$$

where V_Ω is the volume of the domain Ω. This strategy coincides with Strategy 1 if all the elements in the tessellation have the same volume.

Strategy 3. An alternative strategy for error balancing is to refine the element K where the computed error estimate $\hat{\delta}_K$ exceeds a certain fraction of the total (or maximum) estimated error [44].

Strategy 4. One can also choose to refine a given percentage of the elements whose error indicator is the largest.

In [94] the authors study several adaptation strategies, such as fixed threshold, error equidistribution and error density equidistribution strategies, but the fixed fraction Strategy 4 appears to be the most useful, because in their experiments the other strategies can lead to an unacceptable decrease in the error reduction rate or even to a stagnation or oscillatory behavior in the error reduction.

It is also argued in [44], that Strategy 4 is preferable compared to the other algorithms. Therefore in the rest of this chapter mesh adaptation Strategy 4 is used in all numerical experiments.

5.7 Numerical results

In this section we demonstrate the performance of the implicit error estimator (5.9) applied to the time harmonic Maxwell equations on a domain $\Omega \subset \mathbb{R}^3$. In this section we choose the wavenumbers such that we can get rid of the major part of the pollution error [7]. Moreover, we expect that Theorem 5.2 provides an accurate lower estimate of the error.

A good a posteriori error estimator should possess the following properties:

- The error estimator should be able to find those areas in the domain where the finite element solution has a large error.
- The error estimator should have a magnitude close to the real error, both locally and globally.

We verify the performance of the implicit error estimator for the Maxwell equations on five different test cases and define the effectivity index as

$$\varepsilon_h = \frac{\delta_h}{\delta}. \tag{5.35}$$

This quantity merely reflects the quality of the global error estimate but is useful to get an impression on the performance of the adaptive algorithm. For any adaptive algorithm the local behavior of the error is, however, one of the most important factors, therefore we evaluate the quality of the local error estimation by computing the correlation coefficient between $\{\delta_i\}_{i=1}^N$ and $\{\hat{\delta}_i\}_{i=1}^N$, where $\delta_i \equiv \delta_{K_i}$ and $\hat{\delta}_i \equiv \hat{\delta}_{K_i}$ are defined in (5.32).

Whenever the exact error δ_K is available we compute the correlation coefficient [108] between the exact and estimated error as

$$r = \frac{N \sum_{j=1}^{N} \delta_j \hat{\delta}_j - \left(\sum_{j=1}^{N} \delta_j \right) \left(\sum_{j=1}^{N} \hat{\delta}_j \right)}{\sqrt{\left(N \sum_{j=1}^{N} \delta_j^2 - (\sum_{j=1}^{N} \delta_j)^2 \right) \left(N \sum_{j=1}^{N} \hat{\delta}_j^2 - (\sum_{j=1}^{N} \hat{\delta}_j)^2 \right)}}. \tag{5.36}$$

There is a strong correlation between $\{\delta_i\}_{i=1}^N$ and $\{\hat{\delta}_i\}_{i=1}^N$ if $r \geq 0.7$

In the experiments described in this section, the initial mesh is denoted by $\texttt{mesh}_0$, and the subsequent adapted meshes are denoted by $\texttt{mesh}_i$, $i = 1, 2, \ldots$.

For adaptation we use the Centaur mesh generator [29] with the so called source based mesh generation technique. In this method regions where the mesh generator should create finer elements are called sources, which in our case are taken as spheres.

We organize the mesh adaptation as follows:

1. Initialize $i = 0$ and N_{smax}.
2. Solve problem (5.1) on $\texttt{mesh}_i$ and compute the implicit error estimate. Stop if the error satisfies (5.34).
3. If the local error is almost homogeneously distributed over the elements then stop the adaptation procedure and apply global refinement. Set $i = i + 1$ and move to (2). Otherwise
4. Mark $q\%$ of the elements with the largest error in the current mesh $\texttt{mesh}_i$ for adaptation. Based on these marked elements generate at most N_{smax} sources.
5. Based on the created sources generate a new mesh and $i = i + 1$, then move to (2).

Based on the created source information, a new mesh is generated by Centaur such that

$$1.5 \leq \frac{N_{i+1}^{\text{dof}}}{N_i^{\text{dof}}} \leq 3.5, \tag{5.37}$$

where N_i^{dof} is the number of degrees of freedom (DOF) in $\texttt{mesh}_i$. Algorithm 1 describes the mesh adaptation procedure in detail.

The value of q can vary between $1\% - 20\%$ and is highly dependent on the mesh generation algorithm. In all our numerical experiments we have chosen $q = 1$ and $N_{\text{smax}} = 15$. The small value of q is explained by the fact that the mesh generator Centaur creates meshes of high quality (no hanging nodes, no large dihedral angles in an element). A larger value of q would result in a huge increase in the number of elements compared to the previous mesh and would not satisfy condition (5.37).

For the adaptation procedure it is also useful to have a lower bound for the exact error. In Chapter 4 a lower bound for the exact error is provided in terms of the implicit error estimate. This lower bound ensures that the resulting error estimate is not a pessimistic overestimate of the exact error when the mesh size is reduced.

Algorithm 1 Algorithm to create sources for the mesh generation algorithm.

1: $N_{\text{smax}} = 15$ and $N_s = 0$
2: Reorder the elements according to their corresponding error in descending order. $N_m = [\frac{N}{100}]q$ - number of marked elements, N - number of elements in the mesh
3: **for** $i = 1, \dots N_m$ **do**
4: **if** $N_s = 0$ **then**
5: create a source with a center located in the barycenter of element i with radius $r = \max(r_s, r_i)$, where $r_s = \alpha \cdot L$ with L being the domain size and r_i the radius of the circumsphere of element i. The parameter α depends on the mesh generator and in all our numerical experiments we choose $\alpha = 0.08$.
6: $N_s = 1$
7: **else**
8: **for** $j = 1, \dots, N_s$ **do**
9: **if** the barycenter of element i is inside source N_j **then**
10: do nothing, exit loop 8, go to loop 3
11: **else**
12: create a new source as described in step 5
13: $N_s = N_s + 1$
14: **end if**
15: **if** $N_s = N_{\text{smax}}$ **then**
16: STOP the algorithm
17: **end if**
18: **end for**
19: **end if**
20: **end for**

5.7.1 Cylindrical domain

In this subsection we test the adaptation method by solving the Maxwell equations on a section of a cylindrical domain shown in Figure 5.2 and defined as:

$$\Omega = \{(x, y, z) = (r\cos(\phi), r\sin(\phi), z) \in \mathbb{R}^3 : 0 < r < 1,\ 0 < \phi < 3\pi/2,\ 0 < z < 1\},$$

with the wave number $k = 1$.

The solution of this problem has corner and edge singularities and can serve as a suitable test case. The adaptation algorithm should be able to detect this singular behavior and result in a denser mesh around the singularities.

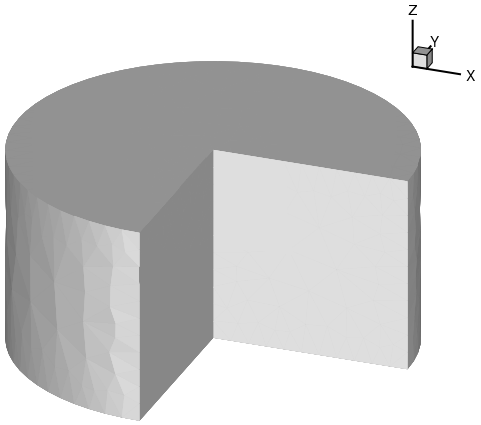


Figure 5.2: Section of the cylindrical domain.

Cylindrical domain with perfectly conducting boundary conditions

In order to be able to evaluate the true discretization errors we first choose a test problem with a known analytical solution. We pick up a vector field $\boldsymbol{E} = [E_1, E_2, E_3]$, substitute it into the first equation of (5.1) and obtain the corresponding right hand side function $\boldsymbol{J}$ and boundary conditions.
This test case is described in [79]. The exact solution of (5.1) is taken as

$$\boldsymbol{E} = z(1-z)(1-r^2)\nabla w, \quad \text{where} \quad w = r^{\frac{2}{3}}\sin(\frac{2}{3}\phi). \tag{5.38}$$

More specifically

$$E_1 = \frac{2}{3}z(1-z)(1-x^2-y^2)\frac{\sin(\frac{2}{3}\arctan\frac{y}{x})x - \cos(\frac{2}{3}\arctan\frac{y}{x})y}{(x^2+y^2)^{\frac{2}{3}}},$$

$$E_2 = \frac{2}{3}z(1-z)(1-x^2-y^2)\frac{\sin(\frac{2}{3}\arctan\frac{y}{x})y + \cos(\frac{2}{3}\arctan\frac{y}{x})x}{(x^2+y^2)^{\frac{2}{3}}},$$

$$E_3 = 0.$$

This function $\boldsymbol{E}$ has a typical singular behavior along the z axis and does not belong to $H^1(\Omega)$. For a discussion of its regularity see [79].

For comparison purposes we also show the convergence of the error on globally refined meshes where the error is computed both using the implicit error estimator and the analytic expression. The lines corresponding to the locally

Table 5.2: Implicit error estimate δ_h, analytic error δ, effectivity index ε_h and correlation coefficient r on the cylindrical domain with perfectly conducting boundary conditions, see Section 5.7.1.

	# edges	# elements	δ_h	δ	ε_h	r
$\texttt{mesh}_0$	1231	981	0.3503	0.2038	1.71	0.57
$\texttt{mesh}_1$	2828	2259	0.1758	0.1268	1.38	0.80
$\texttt{mesh}_2$	10541	8607	0.1090	0.0787	1.38	0.78
$\texttt{mesh}_3$	17700	14550	0.0991	0.0708	1.39	0.80
$\texttt{mesh}_4$	44247	36826	0.0695	0.0518	1.34	0.80

and globally refined meshes are labelled with a subscript loc and glob, respectively. The numerical results and convergence plots are given in Table 5.2 and Figure 5.3. It is clear from Figure 5.3 that adapted meshes, constructed by the implicit error estimator, result in a smaller error than the globally refined meshes with the same number of degrees of freedom. It is also important to note that as the refinement procedure is continued the effectivity index remains constant $\varepsilon \approx 1.3$ and is close to one, which indicates that the error obtained from the implicit error estimator is a good approximation of the true error. The correlation coefficients in Table 5.2 indicate strong correlation, which means that the local error predicted by the implicit a posteriori error estimation method is very similar to the exact error distribution, see Figure 5.4. On the left hand side of Figure 5.5 a contour plot of the implicit error estimate on the fourth adapted mesh is given. The elements with larger error are mainly concentrated near the singularity line along the z axis. The right hand side plot shows the corresponding adapted mesh where, as we expected, the finer elements are created along the singularity axis z.

Cylindrical domain with non-homogeneous tangential boundary conditions

In (5.38), the factor $z(1-z)(1-r^2)$ in front of ∇w was used to satisfy the perfectly conducting boundary conditions and appears to play a regularizing role. In the following test case we solve the Maxwell equations with a non-homogeneous tangential condition on the boundary of Ω, where the same domain is used as in the previous example with the exact solution of the form

$$E = z\nabla w, \tag{5.39}$$

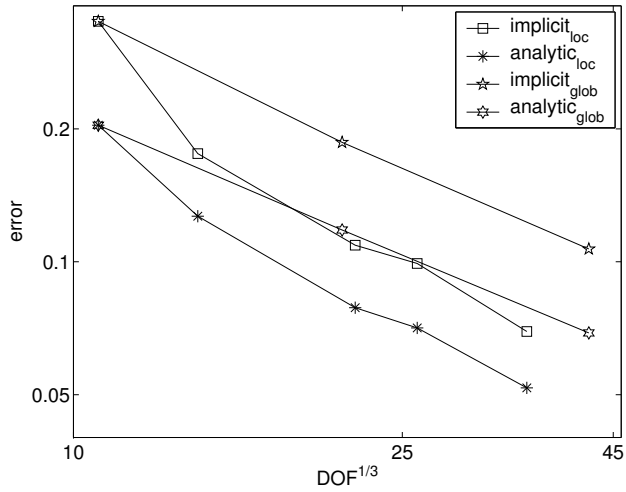


Figure 5.3: Convergence plot in *loglog* scale for the cylindrical domain test case with perfectly conducting boundary conditions, see Section 5.7.1.

with w defined in (5.38). This function, as well as its curl, have the same regularity as in the previous example [79].

The numerical results are given in Table 5.3 and the corresponding convergence diagrams are shown in Figure 5.6. The sequence of meshes used in this experiment are shown in Figure 5.8. We observe the same type of convergence for the implicit error estimator and the exact error as in the previous test case 5.7.1. We note that as the refinement procedure is continued the effectivity index remains constant $\varepsilon \approx 0.8$ which confirms the robustness of the method. The correlation coefficient is also within the range of strong correlation which indicates a good prediction of the local error behavior. The local error distribution diagram on the final mesh (see Figure 5.9) is given in Figure 5.7. It clearly shows that the local error distribution of both schemes has the same behavior throughout the mesh. In Figure 5.9 a contour plot of the implicit error estimate on the final mesh is given. As expected, the elements with larger error are concentrated near the singularity line along the z axis.

To verify the work estimates discussed in Section 5.5 we plot in Figure 5.10 the exact global error δ versus the CPU time, both on globally and adaptively refined meshes. It clearly shows that the adaptive algorithm is computationally

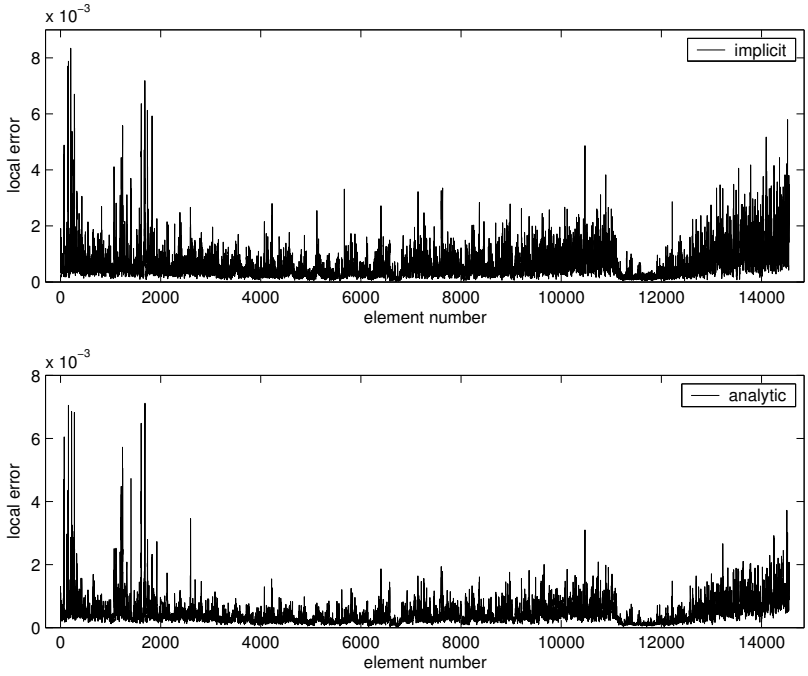


Figure 5.4: Element-wise error distribution of the implicit error estimate and the exact error on the fourth adapted mesh in the cylindrical domain with perfectly conducting boundary conditions, see Section 5.7.1.

Table 5.3: Implicit error estimate δ_h, analytic error δ, effectivity index ε_h and correlation coefficient r on the cylindrical domain with non-homogeneous tangential boundary conditions, see Section 5.7.1.

	# edges	# elements	δ_h	δ	ε_h	r
$\texttt{mesh}_0$	1231	981	0.2081	0.2209	0.91	0.68
$\texttt{mesh}_1$	5219	4287	0.1286	0.1449	0.87	0.72
$\texttt{mesh}_2$	10967	9018	0.0960	0.1156	0.83	0.78
$\texttt{mesh}_3$	15277	12542	0.0922	0.1068	0.86	0.79
$\texttt{mesh}_4$	26861	24853	0.0784	0.0936	0.83	0.80

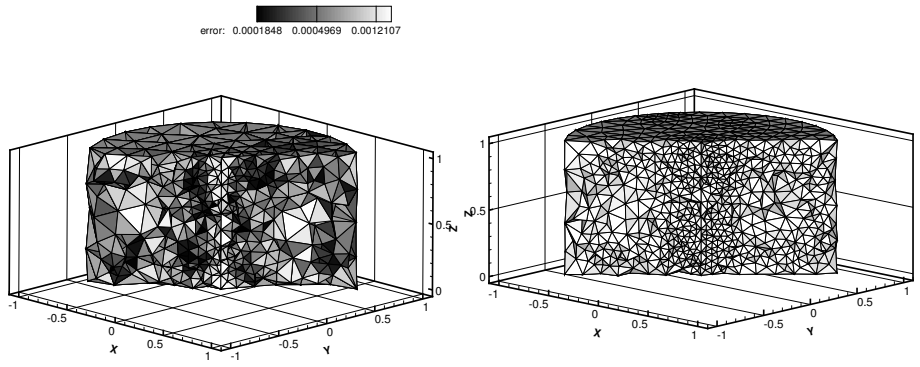


Figure 5.5: Distribution of the implicit error estimate on the fourth adapted mesh (left) and the resulting adapted finite element mesh (right) in the cylindrical domain (cross section with x=y) with perfectly conducting boundary conditions, see Section 5.7.1

more efficient than using globally refined meshes.

5.7.2 Fichera cube

The next test problem we consider are the Maxwell equations defined on a Fichera cube $\Omega = (-1,1)^3 \setminus [-1,0]^3$, with the wave number $k = 1$.

Fichera corner with non-homogeneous tangential boundary conditions

In this test $\boldsymbol{E} = \mathrm{grad}(r^{2/3}\sin(\frac{2}{3}t))$, with $r = \sqrt{x^2+y^2+z^2}$, $t = \arccos(\frac{xyz}{r})$. More specifically

$$E_1 = -\frac{2}{3}\frac{(z^3y+zy^3)\cos(\frac{2}{3}\arccos(\frac{xyz}{\sqrt{x^2+y^2+z^2}}))}{\sqrt{x^2+y^2+z^2-x^2y^2z^2}(x^2+y^2+z^2)^{4/3}} + \frac{2}{3}\frac{\sin(\frac{2}{3}\arccos(\frac{xyz}{\sqrt{x^2+y^2+z^2}}))x}{(x^2+y^2+z^2)^{2/3}},$$

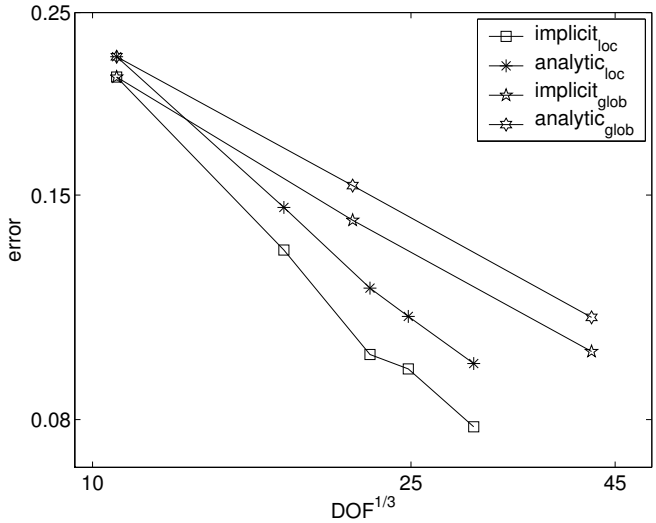


Figure 5.6: Convergence plot in the *loglog* scale for the cylindrical domain test case with non-homogeneous tangential boundary conditions, see Section 5.7.1.

$$E_2 = -\frac{2}{3}\frac{(zx^3+xz^3)\cos(\frac{2}{3}\arccos(\frac{xyz}{\sqrt{x^2+y^2+z^2}}))}{\sqrt{x^2+y^2+z^2-x^2y^2z^2}(x^2+y^2+z^2)^{4/3}} + \frac{2}{3}\frac{\sin(\frac{2}{3}\arccos(\frac{xyz}{\sqrt{x^2+y^2+z^2}}))y}{(x^2+y^2+z^2)^{2/3}},$$

$$E_3 = -\frac{2}{3}\frac{(yx^3+xy^3)\cos(\frac{2}{3}\arccos(\frac{xyz}{\sqrt{x^2+y^2+z^2}}))}{\sqrt{x^2+y^2+z^2-x^2y^2z^2}(x^2+y^2+z^2)^{4/3}} + \frac{2}{3}\frac{\sin(\frac{2}{3}\arccos(\frac{xyz}{\sqrt{x^2+y^2+z^2}}))z}{(x^2+y^2+z^2)^{2/3}}.$$

This vector field has a singular behavior near the origin and it is clear that $\boldsymbol{E}$ does not belong to $[H^1(\Omega)]^3$.

In Table 5.4 the numerical results are given and the corresponding convergence plots of the errors are shown in Figure 5.11. We observe that the error in the adaptive algorithm requires a smaller number of degrees of freedom, when the

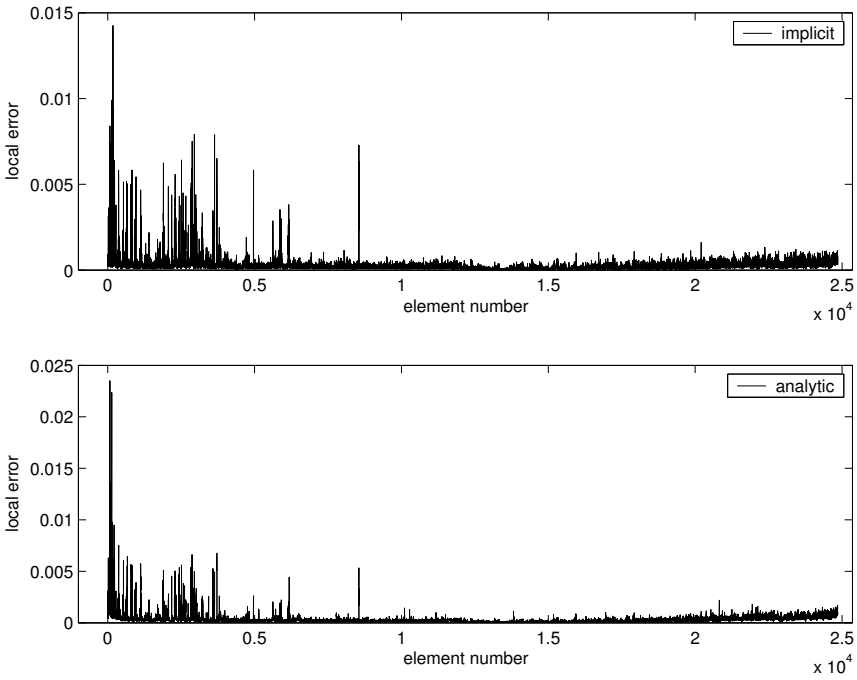


Figure 5.7: Element-wise error distribution of the implicit error estimate and the exact error on the fifth adapted mesh in the cylindrical domain with non-homogeneous tangential boundary conditions, see Section 5.7.1.

implicit error estimation method is used to control the adaptation process, than for the globally refined meshes. During the mesh adaptation procedure the effectivity index remains constant, which means that the error behavior of the implicit error estimation technique is similar to that of the analytic error except for a scaling factor. The correlation coefficients again indicate a strong correlation which means that the local error behavior of the implicit a posteriori error estimation method is very similar to the exact error. In Figure 5.12 a plot of the local error on the third adapted mesh, both for the implicit error estimate and the exact error, versus the element number is given. It also shows a clear correspondence between the local error predicted by the implicit a posteriori error estimation technique and the exact error. In the left hand side of Figure 5.13 a contour plot of the implicit error estimate on the third adapted mesh is given. The elements with larger error are mostly concentrated near the Fichera cor-

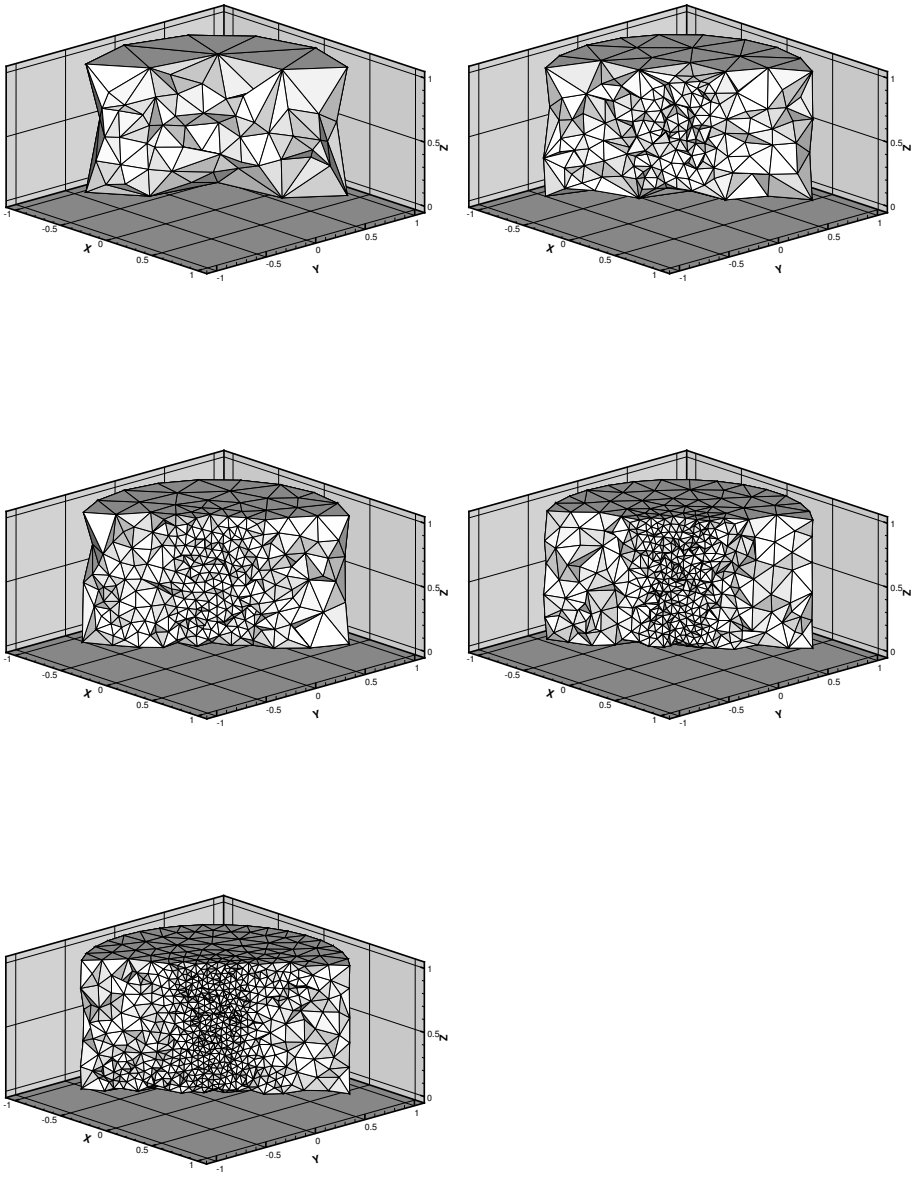


Figure 5.8: Sequence of tetrahedral meshes based on the implicit error estimator used on the cylindrical domain with non-homogeneous tangential boundary conditions, see Section 5.7.1. Cross section with $x = y$.

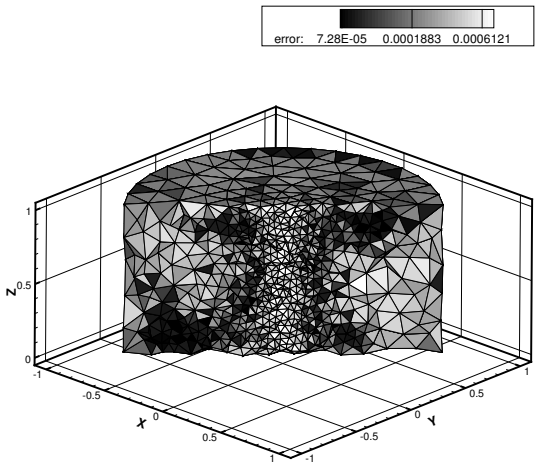


Figure 5.9: Distribution of the implicit error estimate on the fifth adapted mesh used on the cylindrical domain with non-homogeneous tangential boundary conditions, see Section 5.7.1. Cross section with $x = y$.

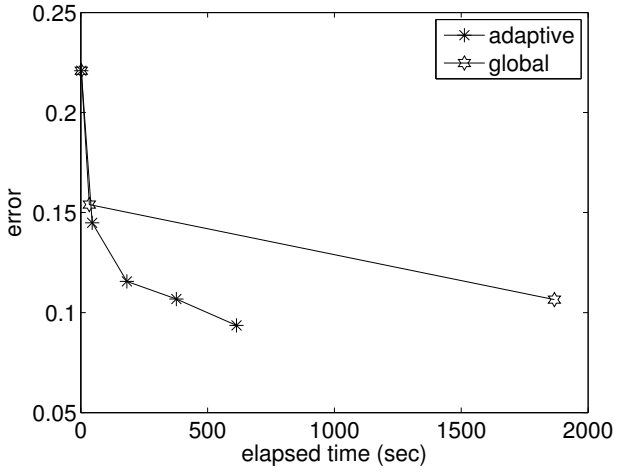


Figure 5.10: Error versus CPU time on the cylindrical domain with non-homogeneous tangential boundary conditions, see Section 5.7.1.

Table 5.4: Implicit error estimate δ_h, analytic error δ, effectivity index ε_h and correlation coefficient r on the Fichera cube with non-homogeneous tangential boundary conditions, see Section 5.7.2.

	# edges	# elements	δ_h	δ	ε_h	r
$\texttt{mesh}_0$	930	710	0.1115	0.5558	0.20	0.70
$\texttt{mesh}_1$	3377	2716	0.0665	0.3972	0.16	0.82
$\texttt{mesh}_2$	9285	7588	0.0238	0.2436	0.096	0.74
$\texttt{mesh}_3$	14923	12293	0.0124	0.1880	0.066	0.80
$\texttt{mesh}_4$	30816	25642	0.0098	0.1485	0.066	0.72

ner. The right hand side plot shows the corresponding adapted mesh where, as we expected, the smaller elements are located near the Fichera corner and its neighborhood.

Note: The fact that the implicit error estimation technique predicts a significantly smaller error in this test case than the exact error can be explained by the fact that the exact solution is curl free. In this case when the curl of the numerical solution is "nearly" zero, then the lower bound for the exact error provided by Theorem 4.16 in Chapter 4 reduces to a pessimistic estimate for the true error.

Fichera corner with perfectly conducting boundary conditions

In this test problem we consider the Maxwell equations on the same Fichera cube but now with the perfectly conducting boundary conditions and a given right hand side function

$$\boldsymbol{J} = \frac{1}{d^2} e^{-\frac{(x-\alpha)^2+(y-\alpha)^2+(z-\alpha)^2}{d^2}} \begin{pmatrix} \cos(\pi(y-\alpha))\cos(\pi(z-\alpha)) \\ \cos(\pi(z-\alpha))\cos(\pi(x-\alpha)) \\ \cos(\pi(x-\alpha))\cos(\pi(y-\alpha)) \end{pmatrix},$$

where $d = 0.5$, $\alpha = 0.25$.

For this problem the exact analytic solution is unknown, therefore the numerical results are presented only for the implicit error estimator, see Table 5.5 and Figure 5.14.

It is clear that the adapted scheme using the implicit error estimation technique produces a smaller error for the same number of degrees of freedom as compared

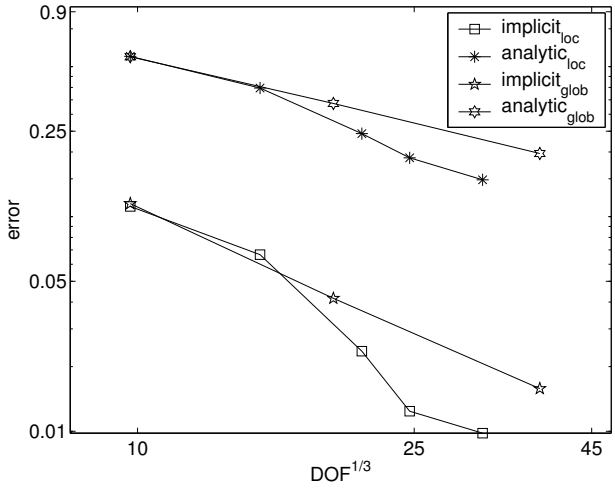


Figure 5.11: Convergence plot in *loglog* scale for the Fichera cube test case with non-homogeneous tangential boundary conditions, see Section 5.7.2.

to the error obtained on the globally refined meshes. The rate of convergence of the implicit error estimator is also higher than that on the globally refined meshes.

The large correlation coefficients observed in all our numerical experiments (of course, except the last one) indicate that the error distribution predicted by the implicit error estimator is very similar to the error distribution of the exact error. This important property is obtained thanks to the proper choice of the local basis used for the finite element solution of (5.9), which will be discussed in Section 5.7.4.

5.7.3 Cylindrical domain with high wave number

It is a well known problem that for wave type equations with high wave numbers the finite element solution provides a good approximation only under certain restrictions on the finite element mesh size, see e.g. [7, 8]. For more details we refer to [64] where for a range of numerical experiments the performance of a finite element scheme is demonstrated for the 1-dimensional Helmholtz equation with high wave numbers.

In this section we investigate the performance of the implicit error estimation

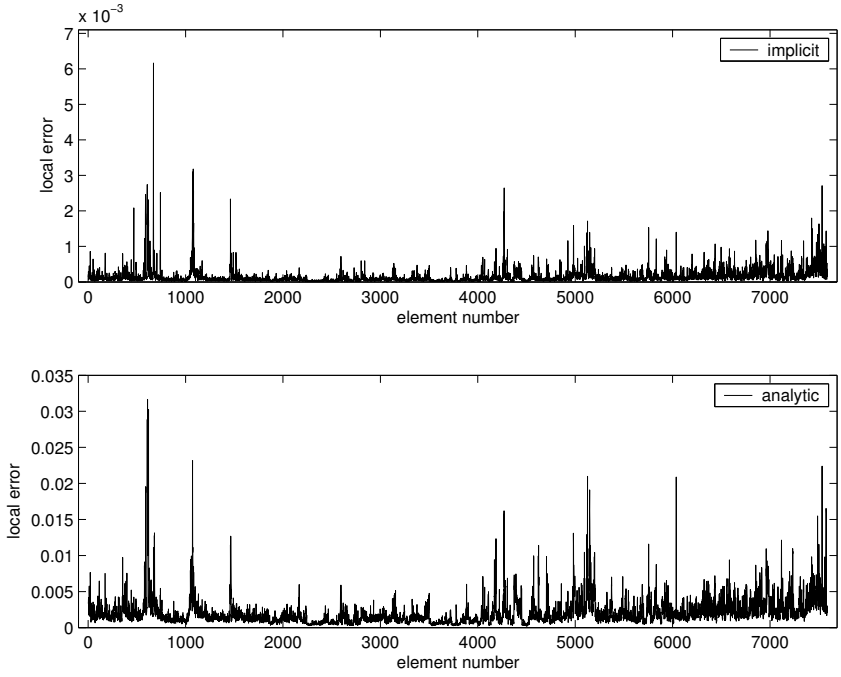


Figure 5.12: Element-wise error distribution of the implicit error estimate and the exact error on the third adapted mesh used for the Fichera domain with non-homogeneous tangential boundary conditions, see Section 5.7.2.

Table 5.5: Implicit error estimate δ_h on the Fichera cube with perfectly conducting boundary conditions, see Section 5.7.2.

	# edges	# elements	δ_h
$\texttt{mesh}_0$	898	683	0.3586
$\texttt{mesh}_1$	2874	2247	0.2410
$\texttt{mesh}_2$	8574	6939	0.1584
$\texttt{mesh}_3$	29689	24497	0.1302
$\texttt{mesh}_4$	62575	51969	0.0943

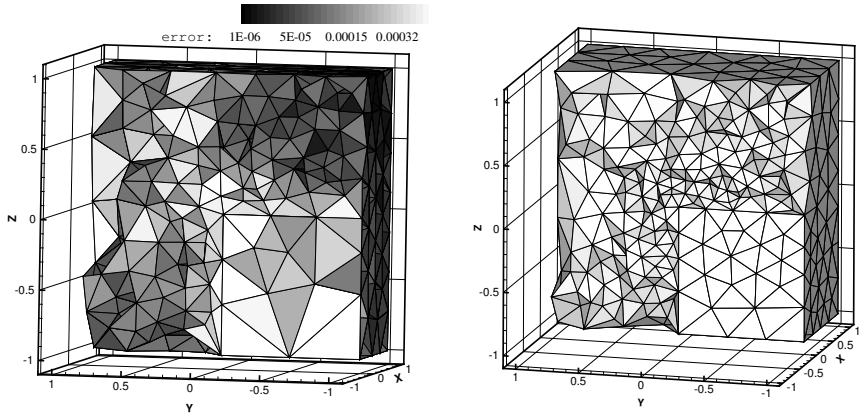


Figure 5.13: Distribution of the implicit error estimate on the third adapted mesh (left) and the adapted finite element mesh of the fourth adapted mesh (right) used on the Fichera domain with non-homogeneous tangential boundary conditions, see Section 5.7.2. Cross section with $x = 0$.

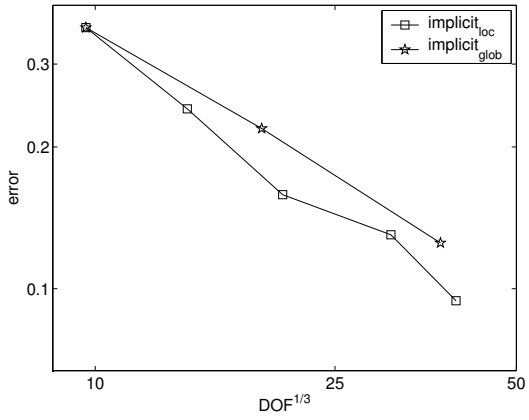


Figure 5.14: Convergence plot in *loglog* scale for the Fichera cube test case with perfectly conducting boundary conditions, see Section 5.7.2.

method developed in this chapter for the Maxwell equations with a high wave number provided that the mesh contains a reasonable number of elements per wave length, as indicated in Lemma 5.6. Otherwise, the mesh is too coarse to represent the waves properly and the results will also be influenced by a significant pollution error. Here we will only evaluate the performance of the implicit a posteriori error estimator. The wave number decides the minimum number of elements per wave length which determines the mesh size for a given domain and is strongly influenced by the computer capacity. Moreover, on very fine meshes with a high wave numbers one needs to apply special techniques for the solution of the linear systems which are, however, beyond the scope of this chapter.

Let us consider the same cylindrical domain as in Section 5.7.1 with the exact solution given by (5.39). The wave number is chosen to be $k = 7$ so that we have two wavelengths (λ) in the domain:

$$\lambda = \frac{2\pi}{7} \approx 0.9.$$

We will demonstrate the performance of the implicit a posteriori error estimation method on a sufficiently fine mesh and will show that the estimator is able to detect the regions with a larger error.

Let us define the mesh size h on a finite element mesh to be the length of the longest edge in the domain. The finite element mesh, constructed for this example has 118602 tetrahedrons and 146943 edges in the domain with the mesh size $h = 0.12$. The solution of the Maxwell equations and the application of the implicit error estimation method on this mesh produced the following results for the implicit error estimate, analytic error, effectivity index and correlation coefficient, respectively:

$$\delta_h = 0.0974,\ \delta = 0.1619,\ \epsilon_h = 0.60,\ r = 0.71. \tag{5.40}$$

The effectivity index, correlation coefficient and the error distribution diagram, shown in Figure 5.15, show that the implicit error estimation technique is able to detect on elements with a relatively large error for a wave number $k = 7$ (2 wavelengths in the domain). This shows that the adaptive algorithm is also applicable for larger values of the wave number k, but this requires computational meshes which are significantly larger than used in the test cases discussed in this chapter and are beyond the present capabilities of our computers.

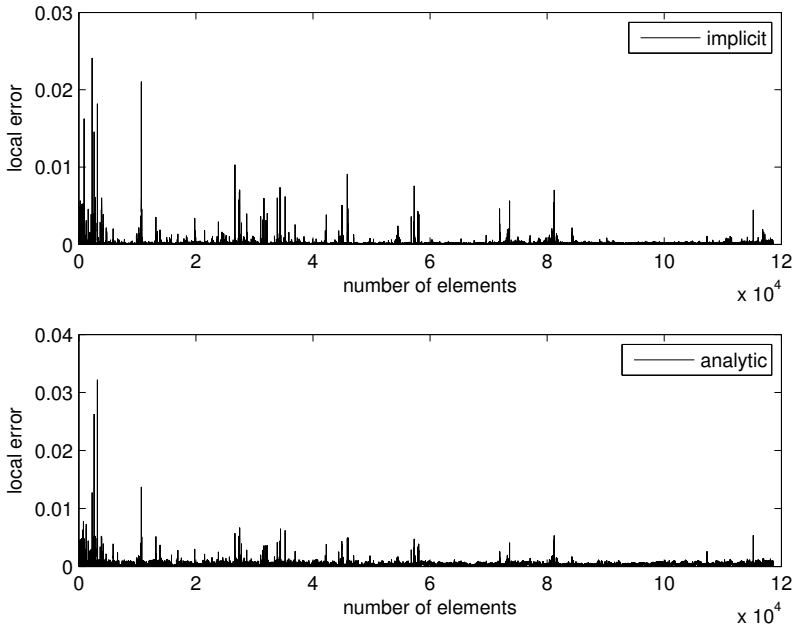


Figure 5.15: Element-wise error distribution of the implicit error estimate and the exact error on finite element mesh in the cylindrical domain with a wave number $k = 7$.

5.7.4 Influence of the local basis on the implicit a posteriori error estimator

As is discussed in the previous sections, an improper choice of the local basis used for the solution of (5.9) may result in a poor approximation of the exact error. We would like to mention that for some simple test cases (not described in this chapter) we have also implemented the implicit error estimation technique with first order Nédélec elements as a local basis for (5.9). The obtained error distribution diagrams of this implicit error estimation method did, however, not describe the true error very well. Here we discuss the performance of the implicit error estimation method on the test case described in Section 5.7.1 but now using the full second order Nédélec basis [93] for the solution of (5.9). Compared to the basis used in the previous section we only add the linear part of the second order Nédélec basis functions which results in a total of 20

Table 5.6: Implicit error estimate δ_h, analytic error δ, effectivity index ε_h and correlation coefficient r on the cylindrical domain with non-homogeneous tangential boundary conditions, see Section 5.7.1. For the solution of (5.9) the full second order Nédélec basis is used.

	# edges	# elements	δ_h	δ	ε_h	r
$\texttt{mesh}_0$	1231	981	3.3382	0.2209	15.11	0.59
$\texttt{mesh}_1$	5219	4287	4.2651	0.1449	29.41	0.59
$\texttt{mesh}_2$	10967	9018	4.6682	0.1156	40.38	0.65
$\texttt{mesh}_3$	15277	12542	5.4128	0.1068	50.66	0.65
$\texttt{mesh}_4$	26861	24853	6.0435	0.0936	64.49	0.69

basis functions per element. This increases the computational work required for the implicit error estimation with $\frac{20^3}{8^3} = 15.625$ times more than for the basis functions used in our experiments, but also has a negative effect on the accuracy.

In Table 5.6 the numerical results of the implicit error estimation using the full Nédélec second order basis in (5.9) are given.

The results from Table 5.6 show that the global error obtained with the implicit error estimation method is now far from the exact error which results in large numbers for the effectivity index. Moreover, on finer meshes the error of the implicit estimator does not converge towards the real error, although the method produced moderate correlation coefficients. This example also shows that both the effectivity index and the correlation coefficient are important factors to judge the quality of the error estimator.

5.8 Conclusions

We discussed an adaptive finite element method using tetrahedral Nédélec elements applied to the Maxwell equations on three-dimensional domains. The adaptation is based on the implicit error estimation technique described in this chapter. We show that the local problems defined for the error equation are well posed. The local problems are solved with a finite element method using second order Nédélec elements without the linear basis functions. The method is tested on various examples including non-convex domains and the results show a good prediction of the true error, both locally and globally. Based on the theoretical

analysis and the numerical results we conclude that the implicit error estimation technique is a powerful method for the adaptive solution of the Maxwell equations. We have also proposed a mesh adaptation algorithm suitable for the Centaur mesh generation package. The algorithm creates adaptive meshes without a drastic increase in the number of elements and generates high quality meshes, i.e. without hanging nodes and no large dihedral angles in an element.

An interesting topic for future work will be the implementation of the implicit error estimation method for the Maxwell equations with higher order Nédélec elements. In that case an important challenge will be to find a suitable well defined local basis for the error equation.

CHAPTER 6

Hamiltonian structure of the Maxwell equations and compatible finite element discretizations

Energy conservation is a desirable property for numerical discretizations of the Maxwell equations. The Stokes-Dirac structure considers the Maxwell equations from a geometrical point of view using differential forms and provides a good starting point to derive energy conservative discretizations. A central issue to achieve this is the choice of variables. We show that Nédélec edge elements for the electric and magnetic field, together with a leap-frog time integration scheme, result in an energy conservative discretization of the Maxwell equations. An important aspect of this discretization is that it also ensures the proper transfer of energy between neighboring elements, which is a central topic in the so called port-Hamiltonian formulation of the Maxwell equations.

6.1 Introduction

In many real life problems it is important to understand the behavior of electromagnetic waves which are described by the Maxwell equations. The mathematical foundation of the Maxwell equations is well developed and depending on the problem one can formulate them in several ways. In particular, the Hamiltonian formulation provides much insight into the geometrical structure of the Maxwell equations. In this chapter we consider the three-dimensional time-dependent Maxwell equations and focus on numerical discretizations which preserve the mathematical structure of the Maxwell equations as much as possible, in particular their energy conserving properties.

In [101] a general class of PDE's is considered in terms of differential forms which provides the right tool to analyze the geometrical structure of these equations. An important issue in this context is to obtain a formulation which ensures that the energy transfer through the interface between two interconnected domains takes place in the correct physical way. For this purpose the notion of Dirac structure is introduced which is defined on certain spaces of differential forms on the spatial domain of the system and its boundary. Since the definition of the Dirac structure is based on Stokes' theorem it is called Stokes-Dirac structure. Its construction emphasizes the geometrical content of the physical variables involved by identifying them as differential k-forms for appropriate values of k. The Stokes-Dirac structure relates to a power-conserving property, namely the change of the interior energy is equal to the power supplied to the system through its boundary. Due to this property it provides the essential framework to obtain a so called port-Hamiltonian formulation [101].

The aim of this chapter is to solve the Maxwell equations numerically while satisfying as many properties of the Stokes-Dirac structure at the discrete level as possible. For the numerical solution of the Maxwell equations we employ finite element methods which have proven to be a very useful technique to represent electromagnetic fields at a discrete level. In many applications the usual Lagrangian or node-based finite elements are, however, not appropriate to represent the electromagnetic fields (see e.g. [18], Section 6.3, [97]) and they can result in physically incorrect solutions.

In the last two decades a great deal of work has been done to overcome the problems arising from node-based elements. An important step forward was made by J.-C. Nédélec [76, 77] who designed a new type of finite elements which describe the electromagnetic field in a better way compared to existing methods. The Nédélec elements have many attractive properties (e.g. automatic satisfaction of the proper continuity requirements across the boundary between two different materials) and are more closely related to the geometrical structure of the Maxwell equations than Lagrangian elements. Nowadays, they are a common technique in computational electromagnetics.

In order to preserve the Stokes-Dirac structure of the Maxwell equations as much as possible at the discrete level we formulate the Maxwell equations in terms of the electric and magnetic field fields and use Nédélec edge finite elements for the spatial discretization of these fields. We show that by this choice of basis functions the energy transfer through the interface of two neighboring elements takes place in a correct way. For the time discretization we apply the

leap-frog symplectic time integration scheme which has nice geometrical properties, e.g. it conserves the discrete energy exactly in the absence of a source term.

This chapter is organized as follows: In Section 6.2 the notion of Dirac structure is highlighted and Section 6.3 provides the main framework for the Stokes-Dirac structure and port-Hamiltonian formulation for a general class of PDE's. In Section 6.4, a particular case of the Stokes-Dirac structure is considered which reduces to the Maxwell equations. Then, in Sections 6.5 and 6.6, the variational problem and the appropriate (discrete) functional spaces are considered for the Maxwell equations. In Sections 6.7 and 6.8 we analyze the numerical properties of the discrete scheme. Section 6.9 presents two ways to discretize the Hodge operator which is used to compute the electric and magnetic flux densities. In the last section the numerical method is verified on a model problem using an unstructured tetrahedral finite element mesh.

6.2 Dirac structures

Many physical phenomena are described by energy conserving systems of partial differential equations (PDE). Some of these equations fit into the concept of a Dirac structure where the PDE's are considered from a geometrical point of view. Dirac structures are a generalization of symplectic and Poisson structures [101] and formalize power-conserving interconnections, thereby allowing the Hamiltonian formulation of interconnected and constrained mechanical and electrical systems.

In this section we briefly introduce some definitions and properties of Dirac structures. For a detailed analysis we refer to [101].

Let $\mathcal{F}$ and $\mathcal{E}$ be real vector spaces whose elements are labelled as $\boldsymbol{f}, \boldsymbol{\varrho}$, respectively. We call $\mathcal{F}$ the space of *flows*, and $\mathcal{E}$ the space of *efforts*. The product space $\mathcal{P} := \mathcal{F} \times \mathcal{E}$ is assumed to be endowed with a scalar pairing $< \cdot | \cdot >: \mathcal{F} \times \mathcal{E} \to \mathbb{R}$, formalizing the notion of *power*.

Definition 6.1. *A map $< \cdot | \cdot >: \mathcal{P} \to \mathbb{R}$ is called a* ***scalar pairing*** *if it is linear in each argument and it is non-degenerate, that is, if $< \boldsymbol{\varrho} | \boldsymbol{f} >= 0,\ \forall \boldsymbol{\varrho} \in \mathcal{E}$ then $\boldsymbol{f} = 0$ and if $< \boldsymbol{\varrho} | \boldsymbol{f} >= 0,\ \forall \boldsymbol{f} \in \mathcal{F}$ then $\boldsymbol{\varrho} = 0$.*

Definition 6.2. *The space $\mathcal{P}$ is called a multi-dimensional (power) port, $p = (\boldsymbol{f}, \boldsymbol{\varrho})$ is called the vector of* ***port variables****. On $\mathcal{P}$ there exists a bilinear form $<<, >>:\ \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined as*

$$<< (\boldsymbol{f}^1, \boldsymbol{\varrho}^1), (\boldsymbol{f}^2, \boldsymbol{\varrho}^2) >>:=< \boldsymbol{\varrho}^1 | \boldsymbol{f}^2 > + < \boldsymbol{\varrho}^2 | \boldsymbol{f}^1 > .$$

Definition 6.3. *A subspace $\mathcal{D} \subset \mathcal{P}$ is called a* ***Dirac structure*** *if $\mathcal{D} = \mathcal{D}^{\perp}$, where $\perp$ denotes the orthogonal complement with respect to the bilinear form $<< \cdot, \cdot >>$.*

A direct implication of Definition 6.3 is that if $\mathcal{D}$ is a Dirac structure, then

$$0 =<< (\boldsymbol{f}, \boldsymbol{\varrho}), (\boldsymbol{f}, \boldsymbol{\varrho}) >>= 2 < \boldsymbol{\varrho} | \boldsymbol{f} >, \ \forall (\boldsymbol{f}, \boldsymbol{\varrho}) \in \mathcal{D}, \tag{6.1}$$

meaning that, if $(\boldsymbol{\varrho}, \boldsymbol{f})$ is a pair of power variables, then the condition $(\boldsymbol{f}, \boldsymbol{\varrho}) \in \mathcal{D}$ implies power conservation $< \boldsymbol{\varrho} | \boldsymbol{f} >= 0$ of the Dirac structure.

This power conserving property (6.1) of a Dirac structure is the starting point for a geometrical formulation of many power-conserving physical systems.

6.3 Stokes-Dirac structures

Let Z be an n dimensional smooth domain in $\mathbb{R}^n$, with smooth $n-1$ dimensional boundary ∂Z, representing the space of spatial variables.

Denote by $\mathcal{DF}^k(Z)$ the space of differential k-forms on Z, $k = 1, 2, \ldots, n$, and by $\mathcal{DF}^k(\partial Z)$ denote the space of differential k-forms on ∂Z, $k = 1, 2, \ldots, n-1$. For more information about differential forms we refer to [17, 57, 55]. There exists a natural pairing between the linear spaces $\mathcal{DF}^k(Z)$ and $\mathcal{DF}^{n-k}(Z)$ and is given by

$$< \boldsymbol{\beta} | \boldsymbol{\alpha} >:= \int_Z \boldsymbol{\beta} \wedge \boldsymbol{\alpha}, \tag{6.2}$$

where $\boldsymbol{\beta} \in \mathcal{DF}^k(Z)$, $\boldsymbol{\alpha} \in \mathcal{DF}^{n-k}(Z)$ and $\wedge$ denotes the usual wedge product of differential forms [55]. The pairing (6.2) is non-degenerate in the sense of Definition 6.1.

Similarly, there exists a pairing between $\mathcal{DF}^k(\partial Z)$ and $\mathcal{DF}^{n-k-1}(\partial Z)$ given by

$$< \boldsymbol{\beta} | \boldsymbol{\alpha} >:= \int_{\partial Z} \boldsymbol{\beta} \wedge \boldsymbol{\alpha}, \tag{6.3}$$

where $\boldsymbol{\beta} \in \mathcal{DF}^k(\partial Z)$, $\boldsymbol{\alpha} \in \mathcal{DF}^{n-k-1}(\partial Z)$.
Next, we define the following linear spaces

$$\mathcal{F}_{p,q} := \mathcal{DF}^p(Z) \times \mathcal{DF}^q(Z) \times \mathcal{DF}^{n-p}(\partial Z), \tag{6.4a}$$

$$\mathcal{E}_{p,q} := \mathcal{DF}^{n-p}(Z) \times \mathcal{DF}^{n-q}(Z) \times \mathcal{DF}^{n-q}(\partial Z), \tag{6.4b}$$

where $p + q = n + 1, \ p, q \in \mathbb{N}$.

Then the pairings (6.2) and (6.3) yield a non-degenerate pairing between the spaces $\mathcal{F}_{p,q}$ and $\mathcal{E}_{p,q}$. Symmetrization of this pairing yields the following bilinear form on $\mathcal{F}_{p,q} \times \mathcal{E}_{p,q}$:

$$<< (\boldsymbol{f}_p^1, \boldsymbol{f}_q^1, \boldsymbol{f}_b^1, \boldsymbol{e}_p^1, \boldsymbol{e}_q^1, \boldsymbol{e}_b^1), (\boldsymbol{f}_p^2, \boldsymbol{f}_q^2, \boldsymbol{f}_b^2, \boldsymbol{e}_p^2, \boldsymbol{e}_q^2, \boldsymbol{e}_b^2) >>:= \tag{6.5}$$
$$\int_Z \left(\boldsymbol{e}_q^2 \wedge \boldsymbol{f}_q^1 + \boldsymbol{e}_p^2 \wedge \boldsymbol{f}_p^1 + \boldsymbol{e}_q^1 \wedge \boldsymbol{f}_q^2 + \boldsymbol{e}_p^1 \wedge \boldsymbol{f}_p^2\right) + \int_{\partial Z} \left(\boldsymbol{e}_b^1 \wedge \boldsymbol{f}_b^2 + \boldsymbol{e}_b^2 \wedge \boldsymbol{f}_b^1\right),$$

where

$$\begin{aligned} &\boldsymbol{f}_p^i \in \mathcal{DF}^p(Z), && \boldsymbol{f}_q^i \in \mathcal{DF}^q(Z), \\ &\boldsymbol{e}_p^i \in \mathcal{DF}^{n-p}(Z), && \boldsymbol{e}_q^i \in \mathcal{DF}^{n-q}(Z), \\ &\boldsymbol{f}_b^i \in \mathcal{DF}^{n-p}(\partial Z), && \boldsymbol{e}_b^i \in \mathcal{DF}^{n-q}(\partial Z), \quad i = 1, 2. \end{aligned}$$

The spaces $\mathcal{DF}^p(Z)$ and $\mathcal{DF}^q(Z)$ of differential forms will represent the energy variables of two different physical energy domains interacting with each other, while $\mathcal{DF}^{n-p}(\partial Z)$ and $\mathcal{DF}^{n-q}(\partial Z)$ will denote the boundary variables whose wedge product represents the boundary energy flow.

Example 6.4. *For the Maxwell equations in $\mathbb{R}^3$ we have $n = 3$ and $p = q = 2$, hence $\mathcal{DF}^p(Z) = \mathcal{DF}^2(Z)$ and $\mathcal{DF}^q(Z) = \mathcal{DF}^2(Z)$, being the space of the electric flux $\boldsymbol{d}$ and magnetic flux $\boldsymbol{b}$ densities, respectively. The space of differential 1-forms $\mathcal{DF}^1(Z)$ is the correct space for the electric $\boldsymbol{e}$ and magnetic $\boldsymbol{h}$ fields, and $\mathcal{DF}^1(\partial Z)$ denotes the electric $\boldsymbol{e}|_{\partial Z}$ and magnetic $\boldsymbol{h}|_{\partial Z}$ fields at the boundary. Their wedge product is the Poynting vector $\boldsymbol{e} \wedge \boldsymbol{h}$.*

Throughout this chapter lowercase bold face letters refer to differential forms and uppercase bold face letters refer to vector fields (proxies).

Theorem 6.5 (van der Schaft and Maschke [101]). *Consider $\mathcal{F}_{p,q}$ and $\mathcal{E}_{p,q}$ as given in (6.4) with $p + q = n + 1$, and bilinear form $<< \cdot|\cdot >>$ given according to (6.5). Define the following linear subspace $\mathcal{D}$ of $\mathcal{F}_{p,q} \times \mathcal{E}_{p,q}$ as*

$$\mathcal{D} = \left\{ (\boldsymbol{f}_p, \boldsymbol{f}_q, \boldsymbol{f}_b, \boldsymbol{e}_p, \boldsymbol{e}_q, \boldsymbol{e}_b) \in \mathcal{F}_{p,q} \times \mathcal{E}_{p,q} \middle| \begin{bmatrix} \boldsymbol{f}_p \\ \boldsymbol{f}_q \end{bmatrix} = \begin{bmatrix} 0 & (-1)^r \mathrm{d} \\ \mathrm{d} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_p \\ \boldsymbol{e}_q \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} \boldsymbol{f}_b \\ \boldsymbol{e}_b \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -(-1)^{n-q} \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_p|_{\partial Z} \\ \boldsymbol{e}_q|_{\partial Z} \end{bmatrix} \right\}, \tag{6.6}$$

where $|_{\partial Z}$ denotes restriction to the boundary ∂Z, $r = pq + 1$ and d denotes the exterior derivative on differential forms. Then $\mathcal{D} = \mathcal{D}^\perp$, that is, $\mathcal{D}$ is a Dirac structure.

A natural question is whether a Dirac structure defined on two domains with a common boundary will be also Dirac structure on their union. The following remark is very useful in the construction of a finite element discretization and provides the correct physical condition to connect two neighboring domains (elements).

Remark 6.6 ([101]). *The spatial compositionality properties of the Stokes-Dirac structure immediately follow from its definition. Let Z_1 and Z_2 be two n-dimensional domains with boundaries ∂Z_1 and ∂Z_2, respectively, and share a common face $\Gamma = \partial Z_1 \cap \partial Z_2$. Then the Stokes-Dirac structures $\mathcal{D}_1$ and $\mathcal{D}_2$ defined on Z_1 and Z_2, respectively, compose to the Stokes-Dirac structure on the domain $Z_1 \cup Z_2$ with boundary $(\partial Z_1/\Gamma) \cup (\partial Z_2/\Gamma)$ if we equate on Γ the boundary variables $\boldsymbol{f}_b^1$ (corresponding to $\mathcal{D}_1$) with $-\boldsymbol{f}_b^2$ (corresponding to $\mathcal{D}_2$), or if we reverse the orientation. The minus sign ensures that the power flowing into Z_1 via Γ is equal to the power flowing out of Z_2 via Γ. Besides, we have to equate on Γ the boundary variables $\boldsymbol{e}_b^1$ (corresponding to $\mathcal{D}_1$) with $\boldsymbol{e}_b^2$ (corresponding to $\mathcal{D}_2$).*

6.3.1 Distributed-parameter port-Hamiltonian systems

The Hamiltonian formulation of physical systems strongly depends on the stored energy in the system and describes the energy evolution with respect to the incoming and dissipated energy in the domain. There is a vast literature available on Hamiltonian formulations of physical systems and we will only dwell upon this subject briefly.

We follow the approach given in [101] to define the Hamiltonian of the system (6.6). On a given domain Z let us consider a Hamiltonian density $\mathcal{H}$ (energy per volume element)

$$\mathcal{H} : \mathcal{DF}^p(Z) \times \mathcal{DF}^q(Z) \times Z \to \mathcal{DF}^n(Z). \tag{6.7}$$

Then the Hamiltonian is defined as

$$\mathrm{H} = \int_Z \mathcal{H}, \tag{6.8}$$

which is equal to the total energy.

Let $\boldsymbol{x}_p$, $\Delta\boldsymbol{x}_p \in \mathcal{DF}^p(Z)$ and $\boldsymbol{x}_q$, $\Delta\boldsymbol{x}_q \in \mathcal{DF}^q(Z)$, then under a weak smoothness

condition on $\mathcal{H}$ we have

$$\begin{aligned}\mathrm{H}(\boldsymbol{x}_p+\Delta\boldsymbol{x}_p,\boldsymbol{x}_q+\Delta\boldsymbol{x}_q) &= \int_Z \mathcal{H}(\boldsymbol{x}_p+\Delta\boldsymbol{x}_p,\boldsymbol{x}_q+\Delta\boldsymbol{x}_q,z) \\ &= \int_Z \mathcal{H}(\boldsymbol{x}_p,\boldsymbol{x}_q,z)+\int_Z \delta_p\mathrm{H}\wedge\Delta\boldsymbol{x}_p+\delta_q\mathrm{H}\wedge\Delta\boldsymbol{x}_q \\ &\quad + \text{higher order terms in } \Delta\boldsymbol{x}_p,\Delta\boldsymbol{x}_q,\end{aligned} \tag{6.9}$$

for certain differential forms $\delta_p\mathrm{H}\in\mathcal{DF}^{n-p}(Z)$, $\delta_q\mathrm{H}\in\mathcal{DF}^{n-q}(Z)$. Then the total energy satisfies

$$\frac{d\mathrm{H}}{dt}=\int_Z \delta_p\mathrm{H}\wedge\frac{\partial\boldsymbol{x}_p}{\partial t}+\delta_q\mathrm{H}\wedge\frac{\partial\boldsymbol{x}_q}{\partial t}. \tag{6.10}$$

The differential forms $\frac{\partial\boldsymbol{x}_p}{\partial t}$ and $\frac{\partial\boldsymbol{x}_q}{\partial t}$ are the velocities of the energy variables $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$, respectively, and are connected to the Stokes-Dirac structure by setting $f_p=-\frac{\partial\boldsymbol{x}_p}{\partial t}$, $f_q=-\frac{\partial\boldsymbol{x}_q}{\partial t}$ and $\boldsymbol{\varrho}_p=\delta_p\mathrm{H}$, $\boldsymbol{\varrho}_q=\delta_q\mathrm{H}$.

In order to indicate that the boundary variables are playing the role of connection variables, which link the system to other systems and whose product represents power, these models are called port-Hamiltonian systems.

Definition 6.7 ([101]). *The distributed-parameter* ***port-Hamiltonian*** *system in an n-dimensional domain Z, state space $\mathcal{DF}^p(Z)\times\mathcal{DF}^q(Z)$ (with $p+q=n+1$), Stokes-Dirac structure given by (6.6), and Hamiltonian* H*, is given as*

$$\begin{bmatrix}\boldsymbol{f}_p\\ \boldsymbol{f}_q\end{bmatrix}=\begin{bmatrix}0 & (-1)^r\mathrm{d}\\ \mathrm{d} & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\varrho}_p\\ \boldsymbol{\varrho}_q\end{bmatrix},\begin{bmatrix}\boldsymbol{f}_b\\ \boldsymbol{\varrho}_b\end{bmatrix}=\begin{bmatrix}1 & 0\\ 0 & -(-1)^{n-q}\end{bmatrix}\begin{bmatrix}\boldsymbol{\varrho}_p|_{\partial Z}\\ \boldsymbol{\varrho}_q|_{\partial Z}\end{bmatrix}, \tag{6.11}$$

where $r=pq+1$ and

$$\boldsymbol{f}_p=-\frac{\partial\boldsymbol{x}_p}{\partial t},\quad \boldsymbol{f}_q=-\frac{\partial\boldsymbol{x}_q}{\partial t},\quad \boldsymbol{\varrho}_p=\delta_{\boldsymbol{x}_p}\mathrm{H},\quad \boldsymbol{\varrho}_q=\delta_{\boldsymbol{x}_q}\mathrm{H}.$$

The variables $(\boldsymbol{x}_p,\boldsymbol{x}_q)$ are called state variables.

A direct consequence of the Stokes-Dirac structure is:

Corollary 6.8. *By the power conserving property (6.1) of a Dirac structure it immediately follows that for any $(\boldsymbol{f}_p,\boldsymbol{f}_q,\boldsymbol{f}_b,\boldsymbol{\varrho}_p,\boldsymbol{\varrho}_q,\boldsymbol{\varrho}_b)$ in the Stokes-Dirac structure $\mathcal{D}$ the following power conserving relation holds true:*

$$\int_Z \boldsymbol{\varrho}_p\wedge\boldsymbol{f}_p+\boldsymbol{\varrho}_q\wedge\boldsymbol{f}_q+\int_{\partial Z}\boldsymbol{\varrho}_b\wedge\boldsymbol{f}_b=0. \tag{6.12}$$

Using the notations from Definition 6.7 and the power-conserving property (6.12) we obtain

Proposition 6.9. *Consider the distributed-parameter system (6.11) with the total energy* H*, given by* (6.8) *and stored in the domain* Z*, then*

$$\frac{d\mathrm{H}}{dt} = \int_{\partial Z} \boldsymbol{\varrho}_b \wedge \boldsymbol{f}_b, \tag{6.13}$$

expressing that the increase of energy in the domain Z *is equal to the power supplied to the system through the boundary* ∂Z*.*

For a detailed description of the model in the presence of external forces or source terms we refer to [101].

6.4 Port-Hamiltonian formulation of the Maxwell equations

In the previous sections we have discussed a general framework for energy conserving physical systems. Next, we will show that the Maxwell equations also fit into this framework, thus providing an energy conserving physical system.

If we take in Definition 6.7, the space dimension $n = 3$ and $p = q = 2$, we obtain the following set of equations:

$$\begin{bmatrix} \boldsymbol{f}_p \\ \boldsymbol{f}_q \end{bmatrix} = \begin{bmatrix} 0 & -\mathrm{d} \\ \mathrm{d} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\varrho}_p \\ \boldsymbol{\varrho}_q \end{bmatrix}, \begin{bmatrix} \boldsymbol{f}_b \\ \boldsymbol{\varrho}_b \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\varrho}_p|_{\partial Z} \\ \boldsymbol{\varrho}_q|_{\partial Z} \end{bmatrix}. \tag{6.14}$$

If we associate the flow variables (see Example 6.4) with

$$\boldsymbol{f}_q := -\frac{\partial \boldsymbol{b}}{\partial t}, \ \boldsymbol{f}_p := -\frac{\partial \boldsymbol{d}}{\partial t},$$

and the effort variables with

$$\boldsymbol{\varrho}_q := \boldsymbol{h}, \ \boldsymbol{\varrho}_p := \boldsymbol{e}, \quad \text{where} \quad \boldsymbol{e} = \delta_{\boldsymbol{d}}\mathrm{H}, \ \ \boldsymbol{h} = \delta_{\boldsymbol{b}}\mathrm{H},$$

with the Hamiltonian of the system H defined on a domain $Z \subset \mathbb{R}^3$ as

$$\mathrm{H} = \frac{1}{2}\int_Z (\boldsymbol{d} \wedge \boldsymbol{e} + \boldsymbol{b} \wedge \boldsymbol{h}), \tag{6.15}$$

and the state variables

$$(\boldsymbol{x}_q, \boldsymbol{x}_p) := (\boldsymbol{b}, \boldsymbol{d}),$$

then it is straightforward to show that the port-Hamiltonian structure given by (6.14) describes the Maxwell equations without source term, i.e.

$$\begin{bmatrix} -\dfrac{\partial \boldsymbol{d}}{\partial t} \\ -\dfrac{\partial \boldsymbol{b}}{\partial t} \end{bmatrix} = \begin{bmatrix} 0 & -\mathrm{d} \\ \mathrm{d} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{e} \\ \boldsymbol{h} \end{bmatrix}, \begin{bmatrix} \boldsymbol{f}_b \\ \boldsymbol{e}_b \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{e}|_{\partial Z} \\ \boldsymbol{h}|_{\partial Z} \end{bmatrix}. \tag{6.16}$$

In the presence of a source term we have to modify the first matrix equation in (6.16) into

$$\begin{bmatrix} -\dfrac{\partial \boldsymbol{d}}{\partial t} \\ -\dfrac{\partial \boldsymbol{b}}{\partial t} \end{bmatrix} = \begin{bmatrix} 0 & -\mathrm{d} \\ \mathrm{d} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{e} \\ \boldsymbol{h} \end{bmatrix} + \begin{bmatrix} \boldsymbol{j} \\ 0 \end{bmatrix}, \tag{6.17}$$

with the source term $\boldsymbol{j} \in \mathcal{DF}^2(Z)$.

In the presence of a source term $\boldsymbol{j} \in \mathcal{DF}^2(Z)$ the Maxwell equations can be written as

$$\frac{\partial \boldsymbol{b}}{\partial t} = -\mathrm{d}\boldsymbol{e}, \qquad \text{on } Z, \tag{6.18a}$$

$$\frac{\partial \boldsymbol{d}}{\partial t} = \mathrm{d}\boldsymbol{h} - \boldsymbol{j}, \qquad \text{on } Z. \tag{6.18b}$$

In view of the energy conserving properties of port-Hamiltonian systems the energy balance relation for the Maxwell equations is

$$\frac{d\mathrm{H}}{dt} = -\int_{\partial Z} \boldsymbol{e} \wedge \boldsymbol{h} - \int_Z \boldsymbol{j} \wedge \boldsymbol{e}, \tag{6.19}$$

which is known as the Poynting theorem.

To complete the Maxwell equations we consider only linear materials, therefore the following constitutive relations must hold:

$$\boldsymbol{d} = \varepsilon \star \boldsymbol{e}, \quad \boldsymbol{b} = \mu \star \boldsymbol{h}, \quad \text{on } Z, \tag{6.20}$$

where the dielectric permittivity $\varepsilon \in \mathbb{R}^{3\times3}$ and the magnetic permeability $\mu \in \mathbb{R}^{3\times3}$ are assumed to be space dependent positive definite tensors. The Hodge star operator is denoted by $\star : \mathcal{DF}^p(Z) \to \mathcal{DF}^{3-p}(Z)$.

6.4.1 The Maxwell equations with perfectly conducting boundary conditions

The Maxwell equations (6.18) together with the constitutive relations (6.20) uniquely define the electromagnetic fields if proper boundary conditions are

given.

As a model problem we consider the Maxwell equations on a bounded domain $\Omega \subset \mathbb{R}^3$, i.e.

$$\frac{\partial \boldsymbol{b}}{\partial t} = -\mathrm{d}\boldsymbol{e}, \qquad \text{on } \Omega, \tag{6.21a}$$

$$\frac{\partial \boldsymbol{d}}{\partial t} = \mathrm{d}\boldsymbol{h} - \boldsymbol{j}, \qquad \text{on } \Omega, \tag{6.21b}$$

together with the constitutive relations

$$\boldsymbol{d} = \varepsilon \star \boldsymbol{e}, \quad \boldsymbol{b} = \mu \star \boldsymbol{h}, \quad \text{on } \Omega. \tag{6.22}$$

The Maxwell equations (6.21) are complemented with perfectly conducting boundary conditions, written as

$$\mathbf{t}_\Gamma(\boldsymbol{e}) = 0, \qquad \text{on } \Gamma = \partial\Omega, \tag{6.23}$$

where $\mathbf{t}_\Gamma$ denotes the trace operator of 1-forms on the boundary of the domain Ω, for details see [57]. In terms of vector proxies the boundary condition (6.23) is equivalent to

$$\boldsymbol{n} \times \boldsymbol{E} \times \boldsymbol{n} = 0, \quad \text{on } \Gamma = \partial\Omega, \tag{6.24}$$

where $\boldsymbol{E}$ is the vector proxy of the differential form $\boldsymbol{e}$ and $\boldsymbol{n}$ is the outward unit normal vector at the boundary of Ω.

Remark 6.10. *The energy balance relation (6.19) for the Maxwell equations (6.21a)–(6.21b) with perfectly conducting boundary conditions (6.23) on the whole domain ($Z = \Omega$) reduces to the following form*

$$\frac{d\mathrm{H}}{dt} = -\int_\Omega \boldsymbol{j} \wedge \boldsymbol{e}. \tag{6.25}$$

Remark 6.11. *In view of Remark 6.6, the spatial compositionality properties of the Stokes-Dirac structure of the Maxwell equations are preserved if on the interface between two non-overlapping domains the traces of the 1-forms $\boldsymbol{e}$ and $\boldsymbol{h}$ are continuous. One notes that this requirement is equivalent to the physical condition that the tangential component of the electric and magnetic fields across the interface between two subdomains must be continuous.*

Later we will elaborate on this issue and see how to construct a numerical scheme which can preserve the spatial compositionality properties of the Stokes-Dirac structure of the Maxwell equations at the discrete level.

6.5 Variational formulation

To derive the variational formulation of the Maxwell equations (6.21)–(6.22) we will follow the approach of Hiptmair, see [57]. First, let us define the following bilinear forms,

$$a_\varepsilon(\boldsymbol{u},\boldsymbol{v})_\Omega = \int_\Omega \varepsilon \Upsilon_1 \boldsymbol{u} \cdot \Upsilon_1 \boldsymbol{v} \mathrm{d}\boldsymbol{x}, \qquad a_{1/\mu}(\boldsymbol{u},\boldsymbol{v})_\Omega = \int_\Omega \mu^{-1} \Upsilon_2 \boldsymbol{u} \cdot \Upsilon_2 \boldsymbol{v} \mathrm{d}\boldsymbol{x}, \tag{6.26a}$$

$$a_{1/\varepsilon}(\boldsymbol{u},\boldsymbol{v})_\Omega = \int_\Omega \varepsilon^{-1} \Upsilon_2 \boldsymbol{u} \cdot \Upsilon_2 \boldsymbol{v} \mathrm{d}\boldsymbol{x}, \qquad a_\mu(\boldsymbol{u},\boldsymbol{v})_\Omega = \int_\Omega \mu \Upsilon_1 \boldsymbol{u} \cdot \Upsilon_1 \boldsymbol{v} \mathrm{d}\boldsymbol{x}, \tag{6.26b}$$

where $\boldsymbol{u}$, $\boldsymbol{v}$ are forms of appropriate degree and Υ_i, with $i = 1, 2$, being the vector proxy of the corresponding form.

Let us denote by $\mathrm{H}_{\mathrm{el}}(\boldsymbol{e})$ the energy contained in the electric field within a bounded domain $\Omega \subset \mathbb{R}^3$. Since we consider only linear materials, $\mathrm{H}_{\mathrm{el}}(\boldsymbol{e})$ is a quadratic form, which arises from a symmetric positive definite linear form a_ε by $\mathrm{H}_{\mathrm{el}}(\boldsymbol{v}) = \frac{1}{2} a_\varepsilon(\boldsymbol{v},\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathcal{DF}^1(\Omega)$. Then the electric flux density $\boldsymbol{d}$ has to satisfy

$$\int_\Omega \boldsymbol{d} \wedge \boldsymbol{e}' = a_\varepsilon(\boldsymbol{e},\boldsymbol{e}'), \quad \forall \boldsymbol{e}' \in \mathcal{DF}^1(\Omega). \tag{6.27}$$

Similarly, the magnetic flux density $\boldsymbol{b}$ possesses the magnetic energy $\mathrm{H}_{\mathrm{mag}}$ on Ω. It is related to a symmetric, positive definite linear form $a_{1/\mu}$ by $\mathrm{H}_{\mathrm{mag}}(\boldsymbol{v}) = \frac{1}{2} a_{1/\mu}(\boldsymbol{v},\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathcal{DF}^2(\Omega)$. Then the magnetic field $\boldsymbol{h}$ has to fulfill

$$\int_\Omega \boldsymbol{h} \wedge \boldsymbol{b}' = a_{1/\mu}(\boldsymbol{b},\boldsymbol{b}'), \quad \forall \boldsymbol{b}' \in \mathcal{DF}^2(\Omega). \tag{6.28}$$

Since the exterior product introduces a non-degenerate pairing, (6.27) and (6.28) assign energies also to the fields $\boldsymbol{h}$ and $\boldsymbol{d}$. Thus we may introduce symmetric, positive definite linear forms a_μ and $a_{1/\varepsilon}$ as

$$\int_\Omega \boldsymbol{b} \wedge \boldsymbol{h}' = a_\mu(\boldsymbol{h},\boldsymbol{h}'), \quad \forall \boldsymbol{h}' \in \mathcal{DF}^1(\Omega), \tag{6.29}$$

and

$$\int_\Omega \boldsymbol{e} \wedge \boldsymbol{d}' = a_{1/\varepsilon}(\boldsymbol{d},\boldsymbol{d}'), \quad \forall \boldsymbol{d}' \in \mathcal{DF}^2(\Omega). \tag{6.30}$$

The material laws in variational form (6.27)–(6.30) can be combined with the topological laws (6.21a)–(6.21b) and lead to the natural weak formulation of the Maxwell equations.

Testing equation (6.21a) with a 1-form $\boldsymbol{h}' \in \mathcal{DF}^1(\Omega)$ yields

$$\int_\Omega \frac{\partial \boldsymbol{b}}{\partial t} \wedge \boldsymbol{h}' = -\int_\Omega \mathrm{d}\boldsymbol{e} \wedge \boldsymbol{h}'. \tag{6.31}$$

Application of Stokes' integration by parts formula gives the following weak formulation of (6.21a):

$$\int_\Omega \frac{\partial \boldsymbol{b}}{\partial t} \wedge \boldsymbol{h}' = -\int_\Omega \boldsymbol{e} \wedge \mathrm{d}\boldsymbol{h}' - \int_{\partial\Omega} \boldsymbol{e} \wedge \boldsymbol{h}', \tag{6.32}$$

which using (6.29) yields

$$a_\mu(\frac{\partial \boldsymbol{h}}{\partial t}, \boldsymbol{h}') = -\int_\Omega \boldsymbol{e} \wedge \mathrm{d}\boldsymbol{h}' - \int_{\partial\Omega} \boldsymbol{e} \wedge \boldsymbol{h}'. \tag{6.33}$$

In the same way testing equation (6.21b) with a 1-form $\boldsymbol{e}' \in \mathcal{DF}^1(\Omega)$ and using (6.27), we obtain following weak formulation of (6.21b):

$$a_\varepsilon(\frac{\partial \boldsymbol{e}}{\partial t}, \boldsymbol{e}') = \int_\Omega \mathrm{d}\boldsymbol{h} \wedge \boldsymbol{e}' - \int_\Omega \boldsymbol{j} \wedge \boldsymbol{e}'. \tag{6.34}$$

With the perfectly conducting boundary condition (6.23) the variational formulation of the Maxwell equations reads:

Find an $\boldsymbol{e} \in \mathcal{DF}^1(\Omega)$ and $\boldsymbol{h} \in \mathcal{DF}^1(\Omega)$ such that for all $\boldsymbol{e}', \boldsymbol{h}' \in \mathcal{DF}^1(\Omega)$ the following relations holds true:

$$a_\mu(\frac{\partial \boldsymbol{h}}{\partial t}, \boldsymbol{h}') = -\int_\Omega \boldsymbol{e} \wedge \mathrm{d}\boldsymbol{h}', \tag{6.35a}$$

$$a_\varepsilon(\frac{\partial \boldsymbol{e}}{\partial t}, \boldsymbol{e}') = \int_\Omega \mathrm{d}\boldsymbol{h} \wedge \boldsymbol{e}' - \int_\Omega \boldsymbol{j} \wedge \boldsymbol{e}'. \tag{6.35b}$$

In the rest of this chapter we consider the numerical solution of (6.35). Our aim is to design numerical schemes which can preserve the port-Hamiltonian structure at the discrete level as much as possible.

In view of (6.27) and (6.29), the energy (6.8) of the Maxwell equations with Hamiltonian (6.15) in the domain Ω can be written as

$$\mathrm{H} = \frac{1}{2}\left(a_\varepsilon(\boldsymbol{e}, \boldsymbol{e})_\Omega + a_\mu(\boldsymbol{h}, \boldsymbol{h})_\Omega\right). \tag{6.36}$$

6.6 Function spaces

The following Hilbert spaces are very important function spaces for the characterization of the Maxwell equations.

The Hilbert space $H(\mathrm{curl}, \Omega)$ is defined as

$$H(\mathrm{curl}, \Omega) = \{\boldsymbol{u} \in [L_2(\Omega)]^3 : \mathrm{curl}\, \boldsymbol{u} \in [L_2(\Omega)]^3\},$$

and the Hilbert space $H(\mathrm{div}, \Omega)$ is defined as

$$H(\mathrm{div}, \Omega) = \{\boldsymbol{u} \in [L_2(\Omega)]^3 : \mathrm{div}\, \boldsymbol{u} \in L_2(\Omega)\}.$$

The differential operators curl and div are understood in a distributional sense.

6.6.1 Discrete differential forms

In this section we will briefly discuss the discrete counterparts of differential forms and show how they can be used to discretize the Maxwell equations. First we introduce some definitions.

Definition 6.12 (Tessellation). *A finite set of oriented subdomains of Ω is called a tessellation, and denoted by $\mathcal{T} = \{K\}$, if*

1. *$\overline{\Omega} = \bigcup_{K \in \mathcal{T}} \overline{K}$, where for any domain K its closure is denoted by $\overline{K}$,*
2. *for each $K \in \mathcal{T}$, K is an open set,*
3. *if K_1 and K_2 are distinct elements of $\mathcal{T}$, then $K_1 \cap K_2 = \varnothing$,*
4. *each K is a Lipschitz domain*

Definition 6.13. *A tessellation $\mathcal{T}$ of a domain Ω is called **conforming** if there are no hanging nodes in the tessellation. This means that for any two neighboring elements $K_1, K_2 \in \mathcal{T}$, such that $\bar{K}_1 \cap \bar{K}_2 \neq \varnothing$, any node of the two elements belonging to their common face $\bar{K}_1 \cap \bar{K}_2$ necessarily coincides with some node of the other element, see Figure 6.1.*

We restrict ourselves in this chapter to conforming tessellations with tetrahedral elements which are very flexible for approximating complex geometries.

On the tetrahedral elements we define the following Whitney forms [107]. For a given differential m-form $\boldsymbol{u}$ with $(0 \leq m \leq 3)$ its first order vector proxy $\Upsilon_m \boldsymbol{u}$

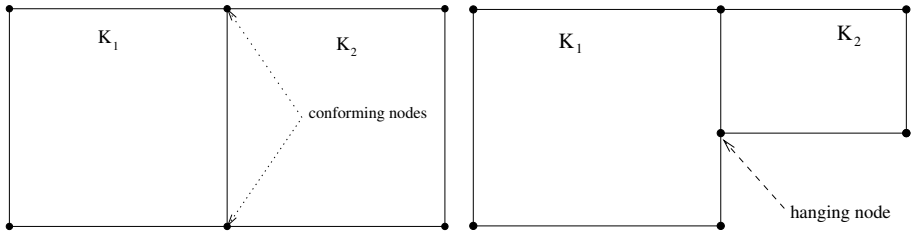


Figure 6.1: Two dimensional example of conforming (left) and non-conforming (right) elements.

on a given tetrahedron T reads

$$\Upsilon_0 \boldsymbol{u}_{(i)} = \lambda_i, \quad i = 1,2,3,4, \tag{6.37a}$$

$$\Upsilon_1 \boldsymbol{u}_{(i,j)} = \lambda_i \nabla\lambda_j - \lambda_j \nabla\lambda_i, \quad i<j,\ i,j = 1,2,3,4, \tag{6.37b}$$

$$\Upsilon_2 \boldsymbol{u}_{(i,j,k)} = 2(\lambda_i \nabla\lambda_j \times \nabla\lambda_k + \lambda_j \nabla\lambda_k \times \nabla\lambda_i + \lambda_k \nabla\lambda_i \times \nabla\lambda_j), \quad i<j<k,\ i,j,k = 1,2,3,4, \tag{6.37c}$$

$$\Upsilon_3 \boldsymbol{u}_{(0,1,2,3)} = 1/\mathrm{Vol}(T), \tag{6.37d}$$

where λ_i $(i = 1,2,3,4)$ are the barycentric coordinates of the tetrahedron T and $\mathrm{Vol}(T)$ is the volume of the element T. Higher order Whitney 1-forms are given in Appendix 6.11.

We will show that the Whitney forms fit in a very natural way into the definition of classical finite elements due to Ciarlet [31], see e.g. [6, 57, 88],

Definition 6.14 (Ciarlet). *Let*

- $K \subset \mathbb{R}^3$ *be any domain (e.g. tetrahedron, hexahedron or prism),*
- $\mathcal{P}_K$ *be a finite-dimensional space of functions defined on* K*,*
- Σ_K *be a set of linear functionals defined on* $\mathcal{P}_K$*. These linear functionals are called the degrees of freedom.*

Then $(K, \mathcal{P}_K, \Sigma_K)$ *is called a finite element.*

The degrees of freedom Σ_K should be defined in such a way that for a given value it uniquely defines a function from $\mathcal{P}_K$.

Definition 6.15. *The finite element* $(K, \mathcal{P}_K, \Sigma_K)$ *is said to be unisolvent if specifying a value for each of the degrees of freedom in* Σ_K *uniquely determines a function in* $\mathcal{P}_K$*.*

The following definition is also important:

Definition 6.16. *Let $\mathcal{W}$ be a space of functions. The finite element $(K, \mathcal{P}_K, \Sigma_K)$ is said to be $\mathcal{W}$ conforming if the corresponding global finite element space is a subspace of $\mathcal{W}$.*

Table 6.1 summarizes the local spaces and degrees of freedom (DOF) for vector proxies of tetrahedral elements with first order Whitney forms. It is clear that the integrals of Whitney elements over l-facets ($l = 0, 1, 2, 3$) are linear functionals on $W^l(T)$ and serve as degrees of freedom.

Table 6.1: Local spaces and degrees of freedom for first order vector proxies Whitney forms on a tetrahedron T with vertices a_0, a_1, a_2, a_3. Vertex indices have to be distinct.

Local space	Local DOF
$W^0(T) = \{\boldsymbol{x} \mapsto \mathbf{a} \cdot \boldsymbol{x} + b,\ \mathbf{a} \in \mathbb{R}^3,\ b \in \mathbb{R}\}$	$u \mapsto u(a_i)$
$W^1(T) = \{\boldsymbol{x} \mapsto \mathbf{a} \times \boldsymbol{x} + \mathbf{b},\ \mathbf{a}, \boldsymbol{b} \in \mathbb{R}^3\}$	$u \mapsto \int_{(a_i,a_j)} \boldsymbol{u} \cdot \mathrm{d}s$
$W^2(T) = \{\boldsymbol{x} \mapsto a\boldsymbol{x} + \mathbf{b},\ a \in \mathbb{R},\ \boldsymbol{b} \in \mathbb{R}^3\}$	$u \mapsto \int_{(a_i,a_j,a_k)} \boldsymbol{u} \cdot \boldsymbol{n}\mathrm{d}S$
$W^2(T) = \{\boldsymbol{x} \mapsto a,\ a \in \mathbb{R}\}$	$u \mapsto \int_T u\mathrm{d}\boldsymbol{x}$

On a given tessellation $\mathcal{T}$ of a domain Ω the global finite element space of Whitney l-forms is denoted by $W_h^l(\Omega)$, $l = 0, 1, 2, 3$, such that $W_h^l(\Omega)|_T = W^l(T)$, see Table 6.1.

It is observed that the Whitney 1- and 2-forms coincide with Nédélec edge and face finite element basis functions [76, 77], respectively.

Let us emphasize that the Whitney 1-forms are $H(\mathrm{curl}, \Omega)$ conforming, i.e. $W_h^1(\Omega) \subset H(\mathrm{curl}, \Omega)$. Therefore $W_h^1(\Omega)$ is a suitable function space for a conforming discretization of the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields.

In the rest of this chapter we only consider first order Whitney 1-forms for the discretization of $\boldsymbol{E}$ and $\boldsymbol{H}$ fields. Then we have

$$W_h^1(\Omega) = \mathrm{span}\{\boldsymbol{w}_j :\ j = 1, 2, \ldots, N_e\}, \tag{6.38}$$

where $\boldsymbol{w}_i$ is the first order edge basis function corresponding to the i-th edge and N_e denotes the number of the edges in the tessellation $\mathcal{T}$.

Remark 6.17. *One can easily use higher order Whitney 1-forms for the discretization of the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields. In this case all the formulas below remain*

unchanged and one only needs to replace the space $W_h^1(\Omega)$ with the corresponding higher order space of Whitney 1-forms.

The space of discrete differential 1-forms on a tessellation $\mathcal{T}$ of the domain Ω is denoted by $\mathcal{DF}_h^1(\Omega) \subset \mathcal{DF}^1(\Omega)$ such that the vector proxy of any discrete differential 1-form belongs to the space of Whitney 1-forms, i.e.

$$\Upsilon_1(\mathcal{DF}_h^1(\Omega)) = W_h^1(\Omega).$$

The basis function of $\mathcal{DF}_h^1(\Omega)$ corresponding the ith basis function of $W_h^1(\Omega)$ is denoted by ω_i such that $\Upsilon_1(\omega_i) = \boldsymbol{w}_i$ and

$$\mathcal{DF}_h^1(\Omega) = \text{span}\{\omega_j\}. \tag{6.39}$$

We refer to a recent article [6] for a detailed construction of discrete differential forms of higher order, in particular for $\mathcal{DF}_h^1(\Omega)$.

6.7 The Maxwell equations for E-H fields

On a given tessellation $\mathcal{T}$ of the domain Ω the discretized variational formulation (6.35) reads:

Find an $\boldsymbol{e}_h \in \mathcal{DF}_h^1(\Omega)$ and $\boldsymbol{h}_h \in \mathcal{DF}_h^1(\Omega)$ such that for all $\boldsymbol{e}', \boldsymbol{h}' \in \mathcal{DF}_h^1(\Omega)$ the following relations hold true:

$$a_\mu(\frac{\partial \boldsymbol{h}_h}{\partial t}, \boldsymbol{h}')_\Omega = -\int_\Omega \boldsymbol{e}_h \wedge \mathrm{d}\boldsymbol{h}', \tag{6.40a}$$

$$a_\varepsilon(\frac{\partial \boldsymbol{e}_h}{\partial t}, \boldsymbol{e}')_\Omega = \int_\Omega \mathrm{d}\boldsymbol{h}_h \wedge \boldsymbol{e}' - \int_\Omega \boldsymbol{j} \wedge \boldsymbol{e}'. \tag{6.40b}$$

Now let us analyze the properties of the discrete system (6.40) and check how well this discrete system fits into the framework of the port-Hamiltonian formulation discussed in the previous sections.

6.7.1 Energy conservation

If we replace $\boldsymbol{h}'$ with $\boldsymbol{h}_h$ and $\boldsymbol{e}'$ with $\boldsymbol{e}_h$ in (6.40) and sum up, we obtain that the discretized energy H_h corresponding to (6.36) and given by

$$\mathrm{H}_h = \frac{1}{2}\left(a_\varepsilon(\boldsymbol{e}_h, \boldsymbol{e}_h)_\Omega + a_\mu(\boldsymbol{h}_h, \boldsymbol{h}_h)_\Omega\right), \tag{6.41}$$

satisfies the energy balance relation (6.25), i.e.

$$\frac{d\mathrm{H}_h}{dt} = -\int_\Omega \boldsymbol{j} \wedge \boldsymbol{e}_h. \tag{6.42}$$

The system (6.40) reads in matrix form:

$$M_\mu \partial_t h = -Ke \tag{6.43a}$$
$$M_\varepsilon \partial_t e = K^T h - \bar{j}, \tag{6.43b}$$

where h and e are the expansion coefficients of the discretized fields $\boldsymbol{h}_h$ and $\boldsymbol{e}_h$, respectively, and

$$M_\mu(i,j) = a_\mu(\omega_i, \omega_j)_\Omega, \qquad M_\varepsilon(i,j) = a_\varepsilon(\omega_i, \omega_j)_\Omega, \tag{6.44}$$
$$K(i,j) = \int_\Omega \omega_i \wedge \mathrm{d}\omega_j. \qquad \bar{j}(i) = \int_\Omega \boldsymbol{j} \wedge \omega_j. \tag{6.45}$$

We note that the system (6.43) is written in such a form that if we apply a symplectic time integrator [56, 87] with $\boldsymbol{j} = 0$, then the time-discrete energy of the numerical solution of (6.43) oscillates around the time-continuous energy, which is defined as $\frac{1}{2}(e^T M_\varepsilon e + h^T M_\mu h)$. We conclude that in the absence of a source term $\boldsymbol{j} = 0$, the discretization of the Maxwell equations in the variational form (6.40) using edge elements for the spatial discretization and a symplectic integrator for the discretization in time, results in an energy conserving numerical scheme on the domain Ω, which is one of the properties of the Hamiltonian formulation, see (6.13).

6.7.2 Port-Hamiltonian structure of the Maxwell equations

One of the main features of the Stokes-Dirac structure is its spatial compositionality property, see Remark 6.6, which provides the physical conditions to obtain the correct energy transfer between two neighboring domains with a common interface.

The finite element method applied to the Maxwell equations results in a discrete solution, therefore we need to check if the spatial compositionality properties are also satisfied at the discrete level to ensure that the energy transfer through internal faces takes place in the correct way, i.e. the energy flowing into the element Z_1 through a common interface Γ_{12} from a neighboring element Z_2 should be equal to the energy coming out of the element Z_2 via the same interface.

If we replace $\boldsymbol{h}$, $\boldsymbol{h}'$ with $\boldsymbol{h}_h$ and $\boldsymbol{e}$, $\boldsymbol{e}'$ with $\boldsymbol{e}_h$ in (6.33) and (6.35b), respectively, but applied on element K and use the relation $\mathcal{DF}_h^1(K) \subset \mathcal{DF}^1(K)$, then the

numerical solution satisfies on each element $K \in \mathcal{T}$ the following relations

$$a_\mu(\frac{\partial \boldsymbol{h}_h}{\partial t}, \boldsymbol{h}_h)_K = -\int_K \boldsymbol{e}_h \wedge \mathrm{d}\boldsymbol{h}_h - \int_{\partial K} \boldsymbol{e}_h \wedge \boldsymbol{h}_h, \tag{6.46a}$$

$$a_\varepsilon(\frac{\partial \boldsymbol{e}_h}{\partial t}, \boldsymbol{e}_h)_K = \int_K \mathrm{d}\boldsymbol{h}_h \wedge \boldsymbol{e}_h - \int_K \boldsymbol{j} \wedge \boldsymbol{e}_h. \tag{6.46b}$$

For the further analysis we need to consider how the energy transfer takes place through the internal faces. Let us consider two neighboring elements K_1, $K_2 \in \mathcal{T}$ with a common boundary $\Gamma_{12} = \bar{K}_1 \cap \bar{K}_2$. If we denote by $\boldsymbol{e}_h^i$ and $\boldsymbol{h}_h^i$ the restriction of the numerical solutions $\boldsymbol{e}_h$ and $\boldsymbol{h}_h$ of (6.40) on K_i, with $i = 1, 2$, then the energy $\int_{\Gamma_{12}} \boldsymbol{e}_h^1 \wedge \boldsymbol{h}_h^1$ flowing into the element K_1 through Γ_{12} from element K_2 should be equal to the energy $-\int_{\Gamma_{12}} \boldsymbol{e}_h^2 \wedge \boldsymbol{h}_h^2$ coming out of element K_2 via the interface Γ_{12}. In order to have the correct energy transfer through this interface the numerical solution $\boldsymbol{e}_h$ and $\boldsymbol{h}_h$ should therefore satisfy:

$$\int_{\Gamma_{12}} \boldsymbol{e}_h^1 \wedge \boldsymbol{h}_h^1 = -\int_{\Gamma_{12}} \boldsymbol{e}_h^2 \wedge \boldsymbol{h}_h^2. \tag{6.47}$$

To show that relation (6.47) holds true, we write it in its equivalent form in terms of vector proxies, then (6.47) follows straightforwardly.

If we denote by $\boldsymbol{E}_h$ and $\boldsymbol{H}_h$ the vector proxies of $\boldsymbol{e}_h$ and $\boldsymbol{h}_h$, respectively, and $\boldsymbol{E}_h^i$ and $\boldsymbol{H}_h^i$ are the restriction of $\boldsymbol{E}_h$ and $\boldsymbol{H}_h$ on K_i, and $\boldsymbol{n}_i$ is the outward unit normal (note that $\boldsymbol{n}_1 = -\boldsymbol{n}_2$) at the interface of K_i, with $i = 1, 2$, then relation (6.47) is equivalent to

$$\int_{\Gamma_{12}} \boldsymbol{n}_1 \cdot (\boldsymbol{E}_h^1 \times \boldsymbol{H}_h^1) = -\int_{\Gamma_{12}} \boldsymbol{n}_2 \cdot (\boldsymbol{E}_h^2 \times \boldsymbol{H}_h^2). \tag{6.48}$$

We have

$$\begin{aligned}
&\int_{\Gamma_{12}} \boldsymbol{n}_1 \cdot (\boldsymbol{E}_h^1 \times \boldsymbol{H}_h^1) + \int_{\Gamma_{12}} \boldsymbol{n}_2 \cdot (\boldsymbol{E}_h^2 \times \boldsymbol{H}_h^2) = \\
&= \int_{\Gamma_{12}} \boldsymbol{n}_1 \cdot (\boldsymbol{E}_h^1 \times \boldsymbol{H}_h^1 - \boldsymbol{E}_h^2 \times \boldsymbol{H}_h^2) \\
&= \int_{\Gamma_{12}} (\boldsymbol{n}_1 \times \boldsymbol{E}_h^1) \cdot \boldsymbol{H}_h^1 - (\boldsymbol{n}_1 \times \boldsymbol{E}_h^2) \cdot \boldsymbol{H}_h^2 \\
&\text{(because edge elements are used } \boldsymbol{n}_1 \times \boldsymbol{E}_h^1 = \boldsymbol{n}_1 \times \boldsymbol{E}_h^2) \\
&= \int_{\Gamma_{12}} (\boldsymbol{n}_1 \times \boldsymbol{E}_h^1) \cdot (\boldsymbol{H}_h^1 - \boldsymbol{H}_h^2) \\
&= \int_{\Gamma_{12}} \boldsymbol{E}_h^1 \cdot (\boldsymbol{H}_h^1 \times \boldsymbol{n}_1 - \boldsymbol{H}_h^2 \times \boldsymbol{n}_1)
\end{aligned}$$

$$= 0.$$

(because edge elements are used $\boldsymbol{H}_h^1 \times \boldsymbol{n}_1 = \boldsymbol{H}_h^2 \times \boldsymbol{n}_1$)

If we sum up the equations (6.46a) and (6.46b) we obtain

$$a_\mu(\frac{\partial \boldsymbol{h}_h}{\partial t}, \boldsymbol{h}_h)_K + a_\varepsilon(\frac{\partial \boldsymbol{e}_h}{\partial t}, \boldsymbol{e}_h)_K = -\int_{\partial K} \boldsymbol{e}_h \wedge \boldsymbol{h}_h - \int_K \boldsymbol{j} \wedge \boldsymbol{e}_h. \tag{6.49}$$

The energy stored in each element K is defined as

$$\mathrm{H}_K = \frac{1}{2}\left(a_\mu(\boldsymbol{h}_h, \boldsymbol{h}_h)_K + a_\varepsilon(\boldsymbol{e}_h, \boldsymbol{e}_h)_K\right). \tag{6.50}$$

Then (6.49) can be written as

$$\frac{d\mathrm{H}_K}{dt} + \int_{\partial K} \boldsymbol{e}_h \wedge \boldsymbol{h}_h + \int_K \boldsymbol{j} \wedge \boldsymbol{e}_h = 0. \tag{6.51}$$

We note that (6.51) is the discrete counterpart of the energy balance relation (6.19) for the Maxwell equations on an arbitrary element $K \in \mathcal{T}$.

If we sum up the equations in (6.51) for all elements $K \in \mathcal{T}$ and using the relation (6.47), we obtain that all internal boundary contributions in (6.51) sum up to zero, hence

$$\frac{d\mathrm{H}_h}{dt} + \int_{\partial\Omega} \boldsymbol{e}_h \wedge \boldsymbol{h}_h + \int_\Omega \boldsymbol{j} \wedge \boldsymbol{e}_h = 0, \tag{6.52}$$

where H_h is the total energy defined in (6.41). With the perfectly conducting boundary conditions we have $\int_{\partial\Omega} \boldsymbol{e}_h \wedge \boldsymbol{h}_h = 0$, hence

$$\frac{d\mathrm{H}_h}{dt} + \int_\Omega \boldsymbol{j} \wedge \boldsymbol{e}_h = 0. \tag{6.53}$$

The relation (6.53) is a discrete counterpart of the energy balance relation given by (6.25). It also shows that we can discretize the Maxwell equations, given by (6.40), on each subdomain K and, since the interface relations are also satisfied at the discrete level, we can assemble the subdomains to obtain a solution on the complete domain which is also energy conservative.

6.8 Leap-frog time discretization

For the time discretization of system (6.43) we use the symplectic leap-frog scheme, i.e.

$$M_\mu \frac{h^{n+1/2} - h^{n-1/2}}{\Delta t} = -Ke^n \tag{6.54a}$$

$$M_\varepsilon \frac{e^{n+1} - e^n}{\Delta t} = K^T h^{n+1/2} - \bar{j}^{n+1/2}, \tag{6.54b}$$

where Δt is the time step and the superscripts refer to the time level. This scheme is conditionally stable, see for example [90], with the stability condition

$$\Delta t \leq \frac{2}{\sqrt{\max(\psi)}} =: \text{CFL}, \tag{6.55}$$

where ψ is an eigenvalue of the amplification matrix $M_\varepsilon^{-1} K^T M_\mu^{-1} K$.

For the leap-frog scheme (6.54) in the absence of the source term $\bar{j}$ the following discrete energy $\mathrm{H}_h^n = (e^n)^T M_\varepsilon e^n + (h^{n-1/2})^T M_\mu h^{n+1/2}$ for (6.43) is conserved exactly [82], i.e.

$$\mathrm{H}_h^n = \text{const}, \quad n = 1, 2, \ldots, n-1, \tag{6.56}$$

where the superscript n refers to the time level.

6.9 Computation of B and D fields

To complete the numerical solution of the Maxwell equations we have to calculate the magnetic flux density $\boldsymbol{B}$ and the electric flux density $\boldsymbol{D}$ based on the computed numerical solutions $\boldsymbol{E}_h$ and $\boldsymbol{H}_h$. Below we present two approaches: one is based on the discretization of Faraday's and Ampere's law, see (6.21a) and (6.21b), respectively, and provides a globally divergence free magnetic flux density $\boldsymbol{B}_h$. The second approach is based on the variational formulation of the constitutive relations (6.27)–(6.30), and requires each time step the solution of a sparse linear system representing the global mass matrix associated with the 2-forms.

6.9.1 Globally divergence free B and D fields

The Maxwell equations (6.21) involve two conservation laws which we have not been addressed yet. Provided that the initial magnetic flux density is divergence free, then the magnetic flux density $\boldsymbol{B}$ satisfies the following divergence constraint:

$$\nabla \cdot \boldsymbol{B} = 0, \quad \text{on} \quad \Omega. \tag{6.57a}$$

Similarly, the electric flux density $\boldsymbol{D}$ satisfies

$$\nabla \cdot \boldsymbol{D} = \rho, \quad \text{on} \quad \Omega, \tag{6.57b}$$

where ρ is the charge density.

In some applications violation of the divergence constraints at the discrete level results in non-physical solutions. After the numerical solution of (6.40) for the $\boldsymbol{E}_h$ and $\boldsymbol{H}_h$ fields we aim to construct the discrete counterpart of the magnetic flux density, denoted by $\boldsymbol{B}_h$, and the electric flux density, denoted by $\boldsymbol{D}_h$, such that

$$\nabla \cdot \boldsymbol{B}_h = 0, \qquad \text{on} \quad \Omega, \tag{6.58a}$$
$$\nabla \cdot \boldsymbol{D}_h = 0, \qquad \text{on} \quad \Omega. \tag{6.58b}$$

Here we assume for simplicity that $\rho = 0$.

We follow the approach presented in [51] where the method is applied to the time-harmonic Maxwell equations.

Thanks to the fact that the Whitney elements satisfy the discrete De Rham diagram, i.e.

$$W_h^0(\Omega) \xrightarrow{\nabla} W_h^1(\Omega) \xrightarrow{\nabla\times} W_h^2(\Omega) \xrightarrow{\nabla\cdot} W_h^3(\Omega)$$

we immediately obtain that $\nabla \times \boldsymbol{E}_h \in W_h^2(\Omega)$ is a discrete 2-form.
From the Maxwell equations we have

$$\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E}, \tag{6.59}$$

and discretize (6.59) as

$$\frac{\boldsymbol{B}_h^{n+1} - \boldsymbol{B}_h^n}{\Delta t} = -\nabla \times \boldsymbol{E}_h^n. \tag{6.60}$$

Then the discrete magnetic flux density is computed by

$$\boldsymbol{B}_h^{n+1} = -\Delta t \nabla \times \boldsymbol{E}_h^n + \boldsymbol{B}_h^n, \tag{6.61}$$

or

$$\boldsymbol{B}_h^{n+1} = -\sum_{m=0}^{n} \Delta t \nabla \times \boldsymbol{E}_h^m + \boldsymbol{B}^0(x, y, z, t), \tag{6.62}$$

where $\boldsymbol{B}^0(x, y, z, t)$ is the initial magnetic flux density which should be divergence free.

Then it is straightforward to see, using the discrete De Rham diagram, that the computed magnetic flux density $\boldsymbol{B}_h^n$ is also divergence free in the whole domain Ω, i.e.

$$\nabla \cdot \boldsymbol{B}_h^{n+1} = -\sum_{m=0}^{n} \Delta t \underbrace{\nabla \cdot (\nabla \times \boldsymbol{E}_h^m)}_{=0,\text{ on } \Omega} + \nabla \cdot \boldsymbol{B}^0(x,y,z,t) \tag{6.63}$$

$$= \nabla \cdot \boldsymbol{B}^0(x,y,z,t) = 0. \tag{6.64}$$

This approach is, however, not very practical since we lose one order of accuracy in $\boldsymbol{B}_h$ when computing the curl of $\boldsymbol{E}_h$. In particular, for first order Whitney elements this did not result in a convergent scheme.

6.9.2 Discrete Hodge operator

Let us note that the Whitney 2-forms are $H(\mathrm{div},\Omega)$ conforming, i.e. $W_h^2(\Omega) \subset H(\mathrm{div},\Omega)$. Therefore $W_h^2(\Omega)$ is a suitable function space for a conforming discretization of the $\boldsymbol{B}$ and $\boldsymbol{D}$ fields. The discrete counterpart of the Hodge operator, used to define the material laws, then can be used in the following way to obtain the electric and magnetic flux density. We expand the fields $\boldsymbol{e}, \boldsymbol{h}, \boldsymbol{d}, \boldsymbol{b}$ with their appropriate discrete Whitney forms and introduce these into the variational form (6.28) and (6.30) of the material laws, using the appropriate Whitney forms also for the test functions. Then we obtain matrix equations for the expansion coefficients:

$$M_{1/\mu} b = Ch, \tag{6.65a}$$

$$M_{1/\varepsilon} d = Ce, \tag{6.65b}$$

where the matrices $M_{1/\mu}$ and $M_{1/\varepsilon}$ are real symmetric positive definite and are associated with Whitney 2-forms, i.e.

$$M_{1/\mu}(i,j) = a_{1/\mu}(\eta_i, \eta_j)_\Omega, \quad \text{and} \quad M_{1/\varepsilon}(i,j) = a_{1/\varepsilon}(\eta_i, \eta_j)_\Omega,$$

where η_i's are the proxies of discrete Whitney 2-forms given by (6.37c). The coupling matrix C involves the exterior product of Whitney 1- and 2- forms and is not regular, even not square, in general, i.e.

$$C(i,j) = \int_\Omega \eta_i \wedge \omega_j.$$

6.10 Numerical experiments

In this section we check the convergence properties of the numerical scheme (6.54) on an unstructured tetrahedral finite element mesh on a simple model

problem.

For this purpose let us consider the Maxwell equations (6.21)-(6.22) on the unit cube $\Omega = (0,1)^3$ with the perfectly conducting boundary conditions

$$\boldsymbol{n} \times \boldsymbol{E} = 0, \quad \text{on} \quad \Gamma = \partial\Omega.$$

The source term $\boldsymbol{J}$ is chosen such that the exact solution of the problem is given by

$$\boldsymbol{E}(\boldsymbol{x},t) = \sin(\omega t) \begin{bmatrix} \sin(\pi y)\sin(\pi z) \\ \sin(\pi x)\sin(\pi z) \\ \sin(\pi x)\sin(\pi y) \end{bmatrix}, \tag{6.66}$$

$$\boldsymbol{B}(\boldsymbol{x},t) = \frac{\cos(\omega t)}{\omega} \begin{bmatrix} \pi\sin(\pi x)\big(\cos(\pi y) - \cos(\pi z)\big) \\ \pi\sin(\pi y)\big(\cos(\pi z) - \cos(\pi x)\big) \\ \pi\sin(\pi z)\big(\cos(\pi x) - \cos(\pi y)\big) \end{bmatrix}. \tag{6.67}$$

Using the constitutive relations (6.22) we obtain

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}, \quad \boldsymbol{H} = \frac{1}{\mu}\boldsymbol{B}. \tag{6.68}$$

Finally, substituting (6.68) into the Ampere law (6.21b) and choosing the source term $\boldsymbol{J}$ accordingly, results in

$$\boldsymbol{J} := \nabla \times \boldsymbol{H} - \partial_t \boldsymbol{D} = \cos(\omega t)\Big(\frac{2\pi^2}{\mu\omega} - \varepsilon\omega\Big) \begin{bmatrix} \sin(\pi y)\sin(\pi z) \\ \sin(\pi x)\sin(\pi z) \\ \sin(\pi x)\sin(\pi y) \end{bmatrix}. \tag{6.69}$$

The following values are used for this experiment: $\varepsilon = \mu = \omega = 1$.

To demonstrate the performance of the scheme, we start with an initial tetrahedral mesh, denoted by $\texttt{mesh}_1$ (54 elements and 105 edges), generated by the Centaur mesh generator [29]. The second mesh $\texttt{mesh}_2$ (432 elements and 660 edges) is obtained by subdividing a tetrahedron of $\texttt{mesh}_1$ into 8 smaller tetrahedrons (we call this procedure global refinement). Similarly, by global refinement, we obtain $\texttt{mesh}_3$ (3456 elements and 4632 edges) and $\texttt{mesh}_4$ (27648 elements and 34608 edges), see Figure 6.6. On a given mesh we define the mesh size h to be the longest edge in the mesh.

The system of space discretized Maxwell equations (6.43) is integrated numerically using leap-frog scheme (6.54). The system is solved for the time interval $[0,T]$, with $T = 10$. The time step Δt is chosen depending on the finite element

mesh according to the CFL restriction (6.55). More specifically, we perform the stability analysis on the coarsest mesh $\texttt{mesh}_1$ and compute the CFL time step restriction. The time step for $\texttt{mesh}_1$ is then chosen to be $\Delta t_1 \approx \frac{\text{CFL}}{2}$, more specifically $\Delta t_1 = 0.125$. The time steps for the meshes $\texttt{mesh}_i$, i=2,3,4, are chosen as $\Delta t_i = \frac{\Delta t_1}{2^i}$, $i = 2, 3, 4$.

For a given vector field $\boldsymbol{F}$ and its numerical approximation $\boldsymbol{F}_h$ we compute the error in the following norms:

$$\|\boldsymbol{F} - \boldsymbol{F}_h\|_{([L_2(\Omega\times[0,T])]^3)} := \sqrt{\int_0^T \int_\Omega |\boldsymbol{F} - \boldsymbol{F}_h|^2}, \tag{6.70}$$

$$\|\boldsymbol{F}(t_n) - \boldsymbol{F}_h^n\|_{[L_2(\Omega)]^3} := \sqrt{\int_\Omega |\boldsymbol{F} - \boldsymbol{F}_h^n|^2}, \tag{6.71}$$

where n is the time level and $t_n = n\Delta t$. For the computation of the integrals in space we use a four point Gauss quadrature rule and in time we apply the Simpson quadrature rule.

A convergence diagram of the error for the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields measured according to (6.70) is given in Figure 6.2 in the *loglog* scale showing that the error decreases proportionably to h^2. For the other convergence diagrams we will also use the *loglog* scale.

An error convergence diagram for the $\boldsymbol{E}$ field evaluated at different time levels is given in Figure 6.3 and an error convergence diagram for the $\boldsymbol{H}$ field evaluated at the final time is given in Figure 6.5a. These results also show that second order accuracy is obtained when using first order Whitney elements.

To compute the $\boldsymbol{B}$ and $\boldsymbol{D}$ fields we apply postprocessing, i.e. we use the relations (6.65) to compute these fields. The electric flux density $\boldsymbol{D}$ is computed by solving a linear system given in (6.65b). Similarly, the magnetic flux density $\boldsymbol{B}$ is computed via the relation (6.65a).

Error plots for the $\boldsymbol{D}$ field evaluated at different time levels is given in Figure 6.4 and an error plot for the $\boldsymbol{B}$ field evaluated at the final time is given in Figure 6.5b.

We note that the postprocessing algorithm (6.65) for discretization of the Hodge operator produces approximations for the electric flux density $\boldsymbol{D}$ and magnetic flux density $\boldsymbol{B}$ which have the same order of accuracy as we obtained for the electric $\boldsymbol{E}$ and magnetic $\boldsymbol{H}$ fields, respectively. The small differences in all the

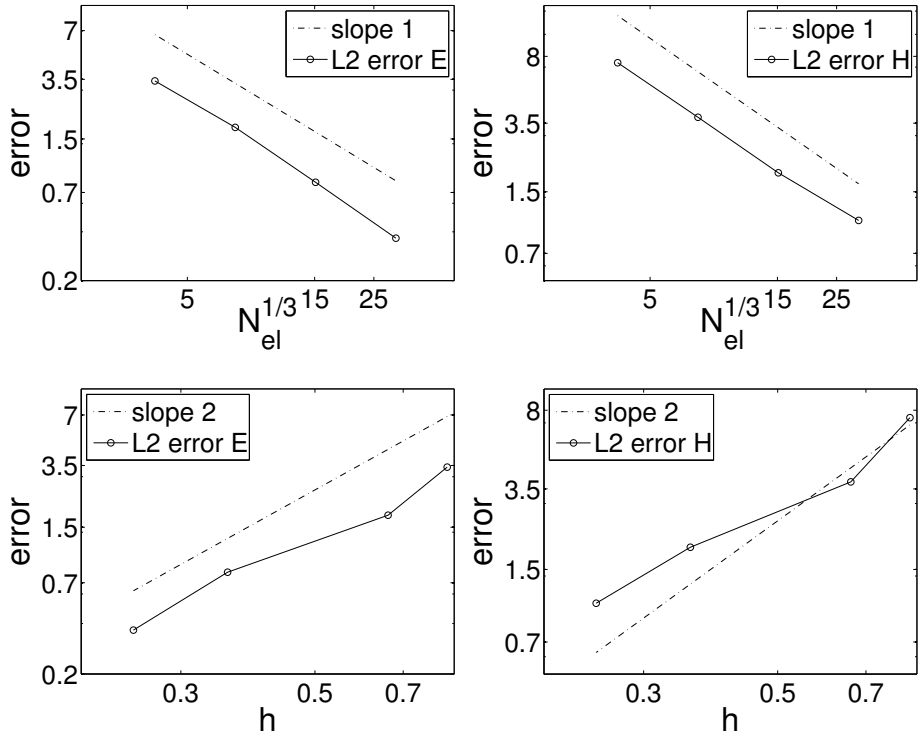


Figure 6.2: Top: Error in the electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{H}$ in the $[L_2(\Omega \times [0,T])]^3$ norm (6.70) versus number of elements N_{el}. Bottom: Error versus mesh size.

convergence diagrams can be explained by the fact that the computations are performed on unstructured meshes.

6.10.1 Energy conservation

In order to verify the discrete energy conservation given by (6.56) we apply the leap-frog time integration scheme on the second finite element mesh denoted as $\texttt{mesh}_2$ with the source term $\bar{j} = 0$ and the initial conditions for $\boldsymbol{e}^0$ and $\boldsymbol{h}^0$ are taken to be vectors with all the elements 1 and 0.5, respectively.

In Figure 6.7 we plot the error $\text{abs}(\text{H}_h^n - \text{H}_h^0)$ for all the time levels. It clearly shows that the discrete energy remains constant up to 12 digits for all time

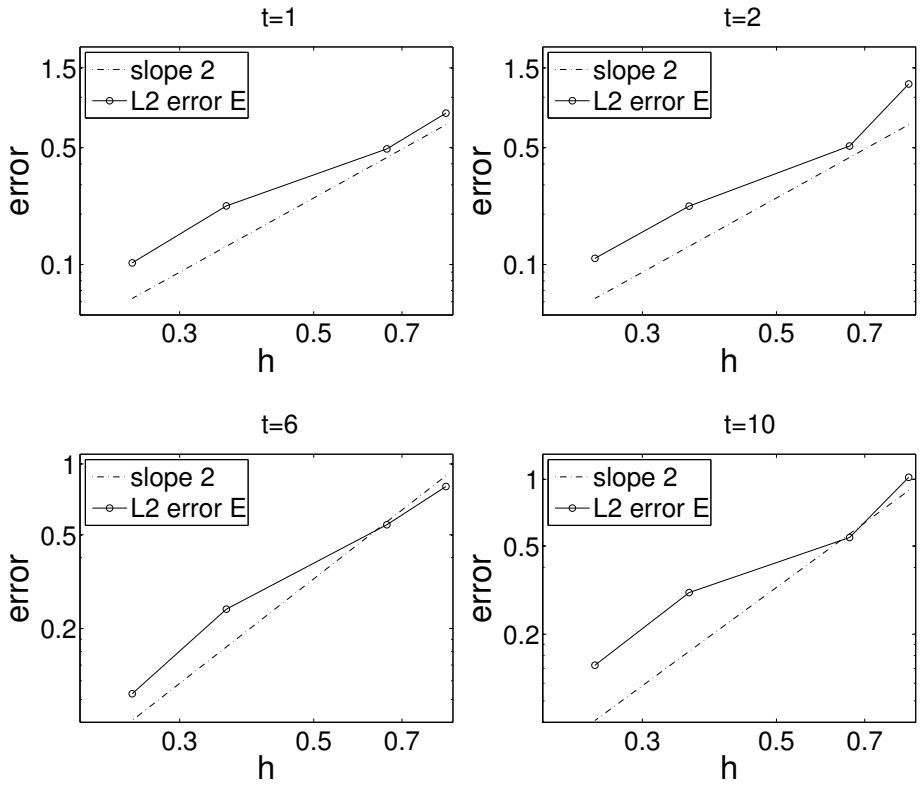


Figure 6.3: Error of the electric field $\boldsymbol{E}$ in the $[L_2(\Omega)]^3$ norm (6.71) versus mesh size evaluated at different time levels.

levels.

6.11 Appendix

6.11.1 Higher order Whitney elements

A detailed derivation of higher order Whitney forms with their corresponding properties on tetrahedral and hexahedral elements is given in [57]. It is out of the scope of this chapter to go through the construction of these Whitney forms, we rather give the vector proxies of second and third order Whitney 1-forms, see for example [93]. These forms can be used directly in the numerical

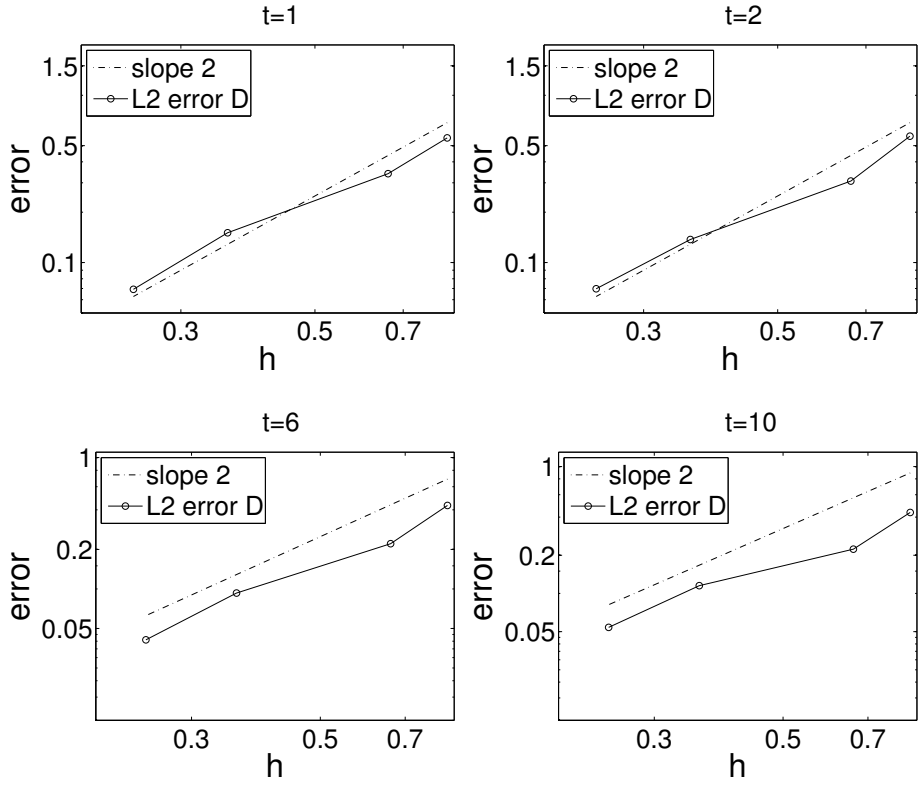


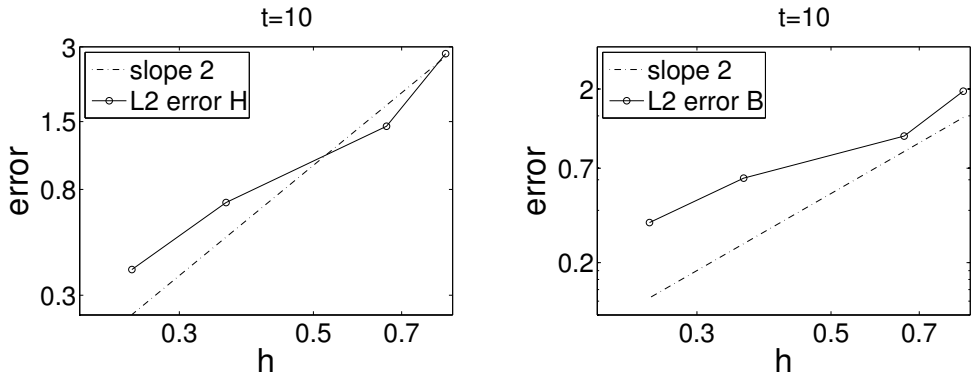
Figure 6.4: Error of the electric flux density $\boldsymbol{D}$ in the $[L_2(\Omega)]^3$ norm (6.71) versus mesh size evaluated at different time levels.

discretization discussed in this chapter to obtain a higher order accurate energy conserving discretization of the Maxwell equations.

There are 20 second order Whitney 1-forms on a given tetrahedron:

12 edge based
$$\lambda_i \nabla \lambda_j, \quad \text{for all } i \neq j, \tag{6.72}$$

8 face based
$$\lambda_i \lambda_j \nabla \lambda_k - \lambda_i \lambda_k \nabla \lambda_j, \tag{6.73}$$
$$\lambda_i \lambda_j \nabla \lambda_k - \lambda_j \lambda_k \nabla \lambda_i \quad \text{for all } i < j < k. \tag{6.74}$$



(a) (b)

Figure 6.5: Error of the magnetic field $\boldsymbol{H}$ (left) and the magnetic flux density $\boldsymbol{B}$ (right) in the $[L_2(\Omega)]^3$ norm (6.71) versus mesh size evaluated at the final time.

The third order Whitney 1-forms on a given tetrahedron are:

18 edge based
$$\lambda_i(2\lambda_i - 1)\nabla\lambda_j, \quad \text{for all } i \neq j, \tag{6.75}$$
$$\lambda_i\lambda_j(\nabla\lambda_i - \nabla\lambda_j), \quad \text{for all } i < j, \tag{6.76}$$

24 face based
$$\lambda_i(2\lambda_i - 1)(\lambda_j\nabla\lambda_k - \lambda_k\nabla\lambda_j), \tag{6.77}$$
$$\lambda_i^2(\lambda_j\nabla\lambda_k - \lambda_k\nabla\lambda_j), \quad \text{for all } i \neq j \neq k, \tag{6.78}$$

3 cell based
$$\lambda_1\lambda_2\lambda_3\nabla\lambda_4 - \lambda_2\lambda_3\lambda_4\nabla\lambda_1, \tag{6.79}$$
$$\lambda_1\lambda_2\lambda_4\nabla\lambda_3 - \lambda_2\lambda_3\lambda_4\nabla\lambda_1,$$
$$\lambda_1\lambda_3\lambda_4\nabla\lambda_2 - \lambda_2\lambda_3\lambda_4\nabla\lambda_1.$$

Another derivation of higher order Whitney forms is given in [87].

6.12 Conclusions

In this chapter we have considered the discretization of the Maxwell equations from a geometrical point of view. For this we used the formulation of the equations in terms of differential forms which are a very convenient and natural

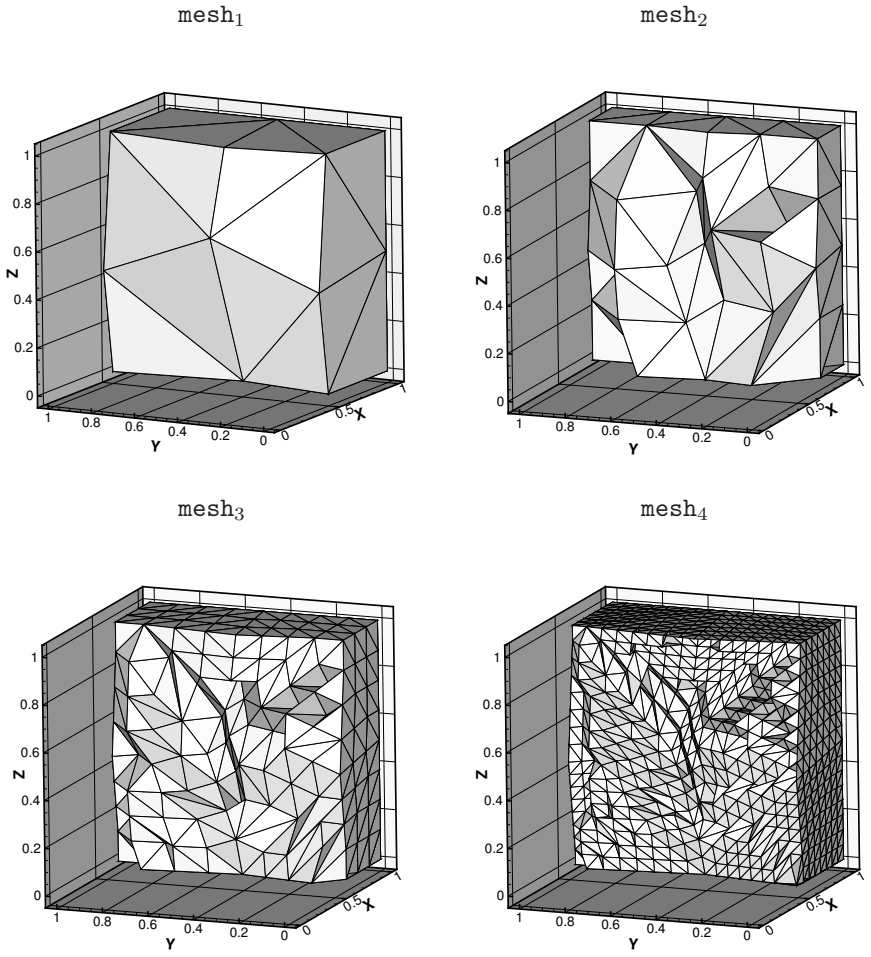


Figure 6.6: Sequence of finite element tetrahedral meshes used in the computations.

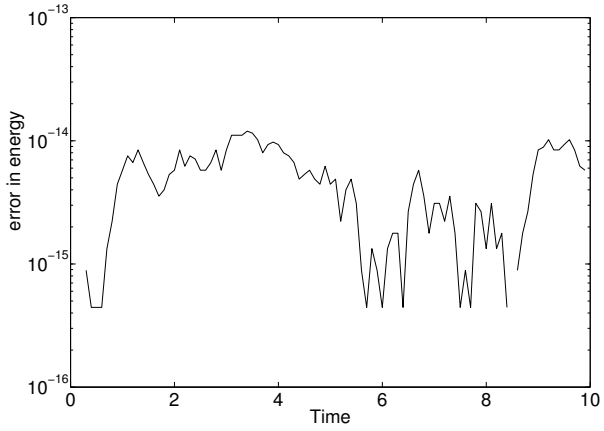


Figure 6.7: Time evolution of the discrete energy error.

way to describe the Maxwell equations. We show that the spatial discretization of the Maxwell equations with the Whitney 1-froms formulated for the electric and magnetic fields preserves the correct energy transfer through the interfaces of neighboring elements. For the time integration we use the symplectic leap-frog scheme which results in an energy conservative discretization of the time-dependent Maxwell equations. We investigated two postprocessing algorithms for the discretization of the Hodge operator to obtain the electric and magnetic flux densities. The direct discretization of the Hodge operator turned out to be the most accurate procedure. Finally, the numerical method is verified on a simple test case with tetrahedral elements, where second order error convergence and energy conservation are shown for first order Whitney elements.

CHAPTER 7

Conclusions

The aim of this thesis is to develop efficient and accurate methods for the numerical solution of the three dimensional Maxwell equations. In many real life problems one needs to solve the time-dependent Maxwell equations on a complicated domain where unstructured finite element meshes are necessary to approximate the geometry accurately. In such situations classical time integration methods usually require a small time step to obtain a stable numerical scheme. In order to relax this severe restriction on the time step the Gautschi time integration scheme is considered in Chapter 3. This scheme involves a special matrix function which provides excellent wave resolution properties and unconditional stability of the scheme. To efficiently compute this expression the Krylov subspace method is applied and we propose a simple algorithm as stopping criterion on the Krylov subspace dimension. The dispersion and dissipation properties of the Gautschi scheme are analyzed on cubic elements and compared with classical time integration methods. The Gautschi time integration method was verified on several test cases and we show that it results in a more efficient scheme on unstructured meshes than several well known explicit time integration methods (e.g. leap-frog) which suffer from a severe time step restriction.

In many physical applications the solution of the Maxwell equations contains singularities for instance on non-convex domains. Adaptive methods are an indispensable tool to solve these problems efficiently. Therefore an accurate error estimator is required. Residual based a posteriori error estimators contain unknown coefficients and are generally not sharp for the Maxwell equations. A

better alternative is provided by implicit error estimators. In Chapter 4 an overview of the method and a theoretical analysis is given. The well-posedness of the local variational problems formulated for the error function is ensured as the wave number is not an eigenvalue of the local problems. For cubic elements we calculated the eigenvalues of the local bilinear form analytically. The method is verified on several test problems with cubic elements and the results obtained with the implicit error estimator show a very good prediction of the true error, even for non-smooth solutions and non-convex domains.

Based on the theoretical background developed for the implicit error estimator we have developed in Chapter 5 an h-adaptive refinement method using tetrahedral elements for the three-dimensional Maxwell equations. A proper finite element basis is given for the solution of the local problems for the error function. The method is verified on various examples with non-convex domains and the results show a good prediction of the true error, both locally and globally. We have proposed a mesh adaptation algorithm suitable for the Centaur mesh generation package. The adaptation algorithm creates meshes without a drastic increase in the number of elements and generates high quality meshes, i.e. without hanging nodes and large dihedral angles in an element.

In Chapter 6 we consider the Maxwell equations from a geometrical point of view. First, we provide some theoretical background on the port-Hamiltonian formulation of these equations, which basically describes the energy conserving properties of the system. Then we show that the discretization of the Maxwell equations formulated for the electric and magnetic fields using edge finite elements obeys some important properties of the port-Hamiltonian structure of the system. In particular, we show that at the discrete level the energy transfer through the interelement boundaries takes place in the correct physical way. We considered two algorithms for computing the electric and magnetic flux density, $\boldsymbol{D}$ and $\boldsymbol{B}$, respectively. The first algorithm is rather simple and provides a globally divergence free solution for the $\boldsymbol{B}$ field. With this method one loses, however, one order of accuracy and the method is not suitable for first order elements. The second algorithm is based on the projection of the Hodge operator on the appropriate discrete spaces. Although this approach is computationally more expensive it provides the same order of approximation for $\boldsymbol{D}$ and $\boldsymbol{B}$ as we obtained for the electric and magnetic fields. The method is verified on a simple test case with unstructured tetrahedral meshes.

BIBLIOGRAPHY

[1] M. Ainsworth. The influence and selection of subspaces for a posteriori error estimators. *Numerische Mathematik*, 73:399–418, 1996.

[2] M. Ainsworth. Dispersive properties of high order Nédélec/edge element approximation of time-harmonic Maxwell equations. *Phil. Trans. Roy. Soc. Series A*, 362(1816):471–493, 2004.

[3] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics. Wiley-Interscience [John Wiley & Sons], New York, 2000.

[4] M. Ainsworth and B. Senior. An adaptive refinement strategy for *hp*–finite element computations. *Applied Numerical Mathematics*, 26:165–178, 1998.

[5] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault. Vector potentials in three-dimensional non-smooth domains. *Math. Methods Appl. Sci.*, 21(9):823–864, 1998.

[6] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numerica*, 15:1–155, 2006.

[7] I. Babuška, F. Ihlenburg, T. Strouboulis, and S. K. Gangaraj. A posteriori error estimation for finite element solutions of Helmholtz' equation. I. The quality of local indicators and estimators. *Internat. J. Numer. Methods Engrg.*, 40(18):3443–3462, 1997.

[8] I. Babuška, F. Ihlenburg, T. Strouboulis, and S. K. Gangaraj. A posteriori error estimation for finite element solutions of Helmholtz' equation. II.

Estimation of the pollution error. *Internat. J. Numer. Methods Engrg.*, 40(21):3883–3900, 1997.

[9] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.

[10] I. Babuška and T. Strouboulis. *The finite element method and its reliability.* Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 2001.

[11] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.

[12] R. E. Bank and R. K. Smith. A posteriori error estimates based on hierarchical bases. *SIAM J. Numer. Anal.*, 30(4):921–935, 1993.

[13] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. A. van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods.* SIAM, Philadelphia, PA, 1994. Available at URL http:// www. netlib .org/ templates/.

[14] T. J. Barth. A posteriori error estimation and mesh adaptivity for finite volume and finite element methods. *Springer series Lecture Notes in Computational Science and Engineering*, 41, 2004.

[15] R. Beck, R. Hiptmair, R. H. W. Hoppe, and B. Wohlmuth. Residual based a posteriori error estimators for eddy current computation. *M2AN Math. Model. Numer. Anal.*, 34(1):159–182, 2000.

[16] R. Beck, R. Hiptmair, and B. Wohlmuth. Hierarchical error estimator for eddy current computation. In *Numerical mathematics and advanced applications (Jyväskylä, 1999)*, pages 110–120. World Sci. Publishing, River Edge, NJ, 2000.

[17] P. Bochev. A discourse on variational and geometric aspects of stability of discretizations. In *33rd Computational Fluid Dynamics Course - Novel methods for solving convection dominated systems*, 2003-05. Von Karman Institute for Fluid Dynamics, March 2003.

[18] A. Bossavit. *Computational electromagnetism. Variational formulations, complementarity, edge elements.* Electromagnetism. Academic Press Inc., San Diego, CA, 1998.

[19] M. A. Botchev, G. L. G. Sleijpen, and H. A. van der Vorst. Stability control for approximate implicit time stepping schemes with minimum residual iterations. *Appl. Numer. Math.*, 31(3):239–253, 1999.

[20] M. M. Botha and D. B. Davidson. An explicit a posteriori error indicator for electromagnetic, finite element-boundary integral analysis. *IEEE Transactions on antennas and propagation*, 53(11):3717–3725, 2005.

[21] M. M. Botha and D. B. Davidson. The implicit, element residual method for a posteriori error estimation in FE-BI analysis. *IEEE Transactions on antennas and propagation*, 54(1):255–258, 2006.

[22] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.

[23] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo. $b = \int g$. *Comput. Methods Appl. Mech. Engrg.*, 145(3-4):329–339, 1997.

[24] A. Buffa and P. Ciarlet, Jr. On traces for functional spaces related to Maxwell's equations. II. Hodge decompositions on the boundary of Lipschitz polyhedra and applications. *Math. Methods Appl. Sci.*, 24(1):31–48, 2001.

[25] A. Buffa, M. Costabel, and D. Sheen. On traces for $\mathbf{H}(\mathbf{curl}, \Omega)$ in Lipschitz domains. *J. Math. Anal. Appl.*, 276(2):845–867, 2002.

[26] C. Carstensen. Some remarks on the history and future of averaging techniques in a posteriori finite element error analysis. *ZAMM Z. Angew. Math. Mech.*, 84(1):3–21, 2004.

[27] C. Carstensen and S. A. Funken. Fully reliable localized error control in the FEM. *SIAM J. Sci. Comput.*, 21(4):1465–1484 (electronic), 1999/00.

[28] W. Cecot, W. Rachowicz, and L. Demkowicz. An *hp*-adaptive finite element method for electromagnetics. III. A three-dimensional infinite element for Maxwell's equations. *Internat. J. Numer. Methods Engrg.*, 57(7):899–921, 2003.

[29] CentaurSoft. Web address: http://www.centaursoft.com/.

[30] D. K. Cheng. *Field and wave electromagnetics*. Addison-Wesley Series in Electrical engineering. Addison-Wesley Publishing Company, 1992.

[31] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.

[32] G. C. Cohen. *Higher-order numerical methods for transient wave equations*. Springer, 2002.

[33] M. Costabel, M. Dauge, and S. Nicaise. Singularities of Maxwell interface problems. *M2AN Math. Model. Numer. Anal.*, 33(3):627–649, 1999.

[34] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 3.* Springer-Verlag, Berlin, 1990. Spectral theory and applications, With the collaboration of Michel Artola and Michel Cessenat, Translated from the French by John C. Amson.

[35] H. De Raedt, K. Michielsen, J. S. Kole, and M. T. Figge. One-step finite-difference time-domain algorithm to solve the Maxwell equations. *Phys. Rev. E*, 67:056706, 2003.

[36] L. Demkowicz. A posteriori error analysis for steady-state Maxwell's equations. In *Advances in adaptive computational methods in mechanics (Cachan, 1997)*, volume 47 of *Stud. Appl. Mech.*, pages 513–526. Elsevier, Amsterdam, 1998.

[37] L. Demkowicz. Finite element methods for maxwell equations. In *Encyclopedia of Computational Mechanics*, pages 723–737. Jonh Wiley & Sons, Chicester, 2004.

[38] L. Demkowicz and L. Vardapetyan. Modeling of electromagnetic absorption/scattering problems using *hp*-adaptive finite elements. *Comput. Methods Appl. Mech. Engrg.*, 152(1-2):103–124, 1998. Symposium on Advances in Computational Mechanics, Vol. 5 (Austin, TX, 1997).

[39] A. Díaz-Morcillo, L. Nuño, J. V. Balbastre, and D. Sánchez-Hernández. Adaptive mesh refinement in electromagnetic problems. In *9th International Meshing Roundtable*, pages 147–155. Sandia National Laboratories, 2000.

[40] V. L. Druskin and L. A. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *U.S.S.R. Comput. Maths. Math. Phys.*, 29(6):112–121, 1989.

[41] V. L. Druskin and L. A. Knizhnerman. Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic. *Numer. Lin. Alg. Appl.*, 2:205–217, 1995.

[42] V. L. Druskin and L. A. Knizhnerman. Extended Krylov subspaces: approximation of the matrix square root and related functions. *SIAM J. Matrix Anal. Appl.*, 19(3):755–771 (electronic), 1998.

[43] A. Ern and J.-L. Guermond. *Theory and practice of finite elements.* Springer-Verlag, New York, 2004.

[44] A. Ern and L.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences.* Springer, 2003.

[45] A. Fisher, R. N. Rieban, G. H. Rodrigue, and D. A. White. A generalized mass lumping technique for vector finite-element solutions of the time-dependent Maxwell equations. *IEEE Transections on Antennas and Propagation*, 53(9):2900–2910, 2005.

[46] F. R. Gantmacher. *The theory of matrices. Vol. 1.* AMS Chelsea Publishing, Providence, RI, 1998. Translated from the Russian by K. A. Hirsch, Reprint of the 1959 translation.

[47] W. Gautschi. Numerical integration of ordinary differential equations based on trigonometric polynomials. *Numer. Math*, 3:381–397, 1961.

[48] S. D. Gedney and U. Navsariwala. An unconditionally stable finite element time-domain solution of the vector wave equation. *IEEE Microwave and Guided Wave Letters*, 5(10):332–334, Oct. 1995.

[49] M. B. Giles and E. Süli. Adjoint methods. *Acta Numerica*, 11:145–236, 2002.

[50] S. K. Godunov and V. S. Ryabenkii. *Difference Schemes. An Introduction to the Underlying Theory.* Elsevier Science, 1987.

[51] N. Golias, T. Tsiboukis, and A. Bossavit. Constitutive inconsistency: rigorous solution of Maxwell equations based on a dual approach. *IEEE Transactions on Magnetics*, 30(5):3586–3589, September 1994.

[52] G. H. Golub and C. F. Van Loan. *Matrix computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[53] J. Gopalakrishnan and J. E. Pasciak. Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell equations. *Math. Comp.*, 72(241):1–15 (electronic), 2003.

[54] V. Grimm. On error bounds for the Gautschi-type exponential integrator applied to oscillatory second-order differential equations. *Numer. Math.*, 100:71–89, 2005.

[55] V. Guillemin and A. Pollack. *Differential topology.* Prentice-Hall Inc., Englewood Cliffs, N.J., 1974.

[56] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*. Springer-Verlag, Berlin, 2002. Structure-preserving algorithms for ordinary differential equations.

[57] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numerica*, 11:237–339, 2002.

[58] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 34(5):1911–1925, Oct. 1997.

[59] M. Hochbruck and C. Lubich. A Gautschi-type method for oscillatory second-order differential equations. *Numer. Math.*, 83:403–426, 1999.

[60] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19(5):1552–1574, 1998.

[61] R. Holland, P. C. Vaughn, and L. C. Wilson. Finite-Volume Time-Domain (FVTD) Techniques for EM Scattering. *IEEE Transactions on Electromagnetic Compatibility*, 33(4), 1991.

[62] R. Horváth. Uniform treatment of numerical time-integrations of the Maxwell equation. In W. H. A. Schilders, J. W. ter Maten, and S. H. M. J. Houben, editors, *Proc. of the Conference "Scientific Computing in Electrical Engineering" SCEE-2002, Eindhoven*, Springer Series Mathematics in Industry, pages 231–239. Springer Verlag, 2004.

[63] R. Horváth, I. Faragó, and W. Schilders. Investigation of numerical time-integrations of maxwell's equations using the staggered grid spatial discretization. *Int. J. Numer. Model.*, 18:149–169, 2005.

[64] F. Ihlenburg and I. Babuska. Finite element solution to the Helmholtz equation with high wave number. part 1. the h-version of the FEM. Technical report, Maryland University College Park Inst for Physical Science and Technology, November 1993. URL: http://handle.dtic.mil/100.2/ADA277396.

[65] F. Izsák, D. Harutyunyan, and J. J. W. van der Vegt. *Implicit a posteriori error estimates for the Maxwell equations*. Internal report, Dept. of Appl. Math., University of Twente, Netherlands, 2007. http://eprints.eemcs.utwente.nl/8448/.

[66] F. Izsák, D. Harutyunyan, and J. J. W. van der Vegt. Implicit a posteriori error estimates for the Maxwell equations. *Math. Comp.*, accepted, 2007.

[67] L. A. Knizhnerman. Calculation of functions of unsymmetric matrices using Arnoldi's method. *U.S.S.R. Comput. Maths. Math. Phys.*, 31(1):1–9, 1991.

[68] J. S. Kole, M. T. Figge, and H. De Raedt. Unconditionally stable algorithms to solve the time-dependent Maxwell equations. *Phys. Rev. E*, 64:066705, 2001.

[69] J. S. Kole, M. T. Figge, and H. De Raedt. Higher-order unconditionally stable algorithms to solve the time-dependent Maxwell equations. *Phys. Rev. E*, 65:066705, 2002.

[70] J. O. L. Demkowicz and T. Strouboulis. Adaptive finite elements for flow problems with moving boundaries. part I: Variational principles and a posteriori estimates. *Computer Methods in Applied Mechanics and Engineering*, 46:217–251, 1984.

[71] J.-F. Lee, R. Lee, and A. Cangellaris. Time-domain finite-element methods. *IEEE transactions on antennas and propagation*, 45(3):430–442, 1997.

[72] J. E. Marsden and T. J. R. Hughes. *Mathematical foundations of elasticity*. Dover Publications Inc., New York, 1994. Corrected reprint of the 1983 original.

[73] P. Monk. *Finite Element Methods for Maxwell's Equations*. Oxford University Press, 2003.

[74] P. Monk. A simple proof of convergence for an edge element discretization of Maxwell's equations. In *Computational electromagnetics (Kiel, 2001)*, volume 28 of *Lect. Notes Comput. Sci. Eng.*, pages 127–141. Springer, Berlin, 2003.

[75] P. B. Monk and A. K. Parrott. A dispersion analysis of finite element methods for Maxwell's equations. *SIAM J. Sci. Comput.*, 15(4):916–937, 1994.

[76] J.-C. Nédélec. Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 35(3):315–341, 1980.

[77] J.-C. Nédélec. A new family of mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 50(1):57–81, 1986.

[78] P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation*, volume 33 of *Studies in Mathematics and its Applications*. Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.

[79] S. Nicaise. Edge elements on anisotropic meshes and approximation of the Maxwell equations. *SIAM J. Numer. Anal.*, 39(3):784–816, 2001.

[80] S. Nicaise and E. Creusé. A posteriori error estimation for the heterogeneous Maxwell equations on isotropic and anisotropic meshes. *Calcolo*, 40(4):249–271, 2003.

[81] Ü. Pekel and R. Lee. An a posteriori error reduction scheme for the three-dimensional finite element solution of Maxwell's equations. *IEEE Transactions on Microwave Theory and Techniques*, 43(2):421–427, 1995.

[82] S. Piperno. Symplectic local time-stepping in non-dissipative DGTD methods applied to wave propagation problems. *Rapport de recherche de l'INRIA*, (5643), 2005.

[83] W. Rachowicz and L. Demkowicz. An *hp*-adaptive finite element method for electromagnetics. part I: data structure and constrained approximation. *Comput. Methods Appl. Mech. Engrg.*, 187(1-2):307–337, 2000.

[84] W. Rachowicz and L. Demkowicz. An *hp*-adaptive finite element method for electromagnetics. II. A 3D implementation. *Internat. J. Numer. Methods Engrg.*, 53(1):147–180, 2002. *p* and *hp* finite element methods: mathematics and engineering practice (St. Louis, MO, 2000).

[85] W. Rachowicz and A. Zdunek. An *hp*-adaptive finite element method for scattering problems in computational electromagnetics. *International Journal for Numerical Methods in Engineering*, 62(9):1226–1249, 2005.

[86] R. D. Richtmyer and K. W. Morton. *Difference methods for initial-value problems*. Interscience Publishers John Wiley & Sons, Inc., New York-London-Sydney, 1967. Second edition.

[87] R. Rieben, D. White, and G. Rodrigue. High-order symplectic integration methods for finite element solutions to time dependent Maxwell equations. *IEEE Trans. Antennas and Propagation*, 52(8):2190–2195, 2004.

[88] R. N. Rieben, G. H. Rodrigue, and D. A. White. A high order mixed vector finite element method for solving the time dependent maxwell equations on unstructured grids. *J. Comput. Phys.*, 204(2):490–519, 2005.

[89] B. Rivière and M. F. Wheeler. A posteriori error estimates and mesh adaptation strategy for discontinuous Galerkin methods applied to diffusion problems. *Computers and Mathematics with Applications*, 46(1):141–163, 2003.

[90] G. Rodrigue and D. White. A vector finite element time-domain method for solving Maxwell's equations on unstructured hexahedral grids. *SIAM Journal on Scientific Computing*, 23(3):683–706, 2001.

[91] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 29(1):209–228, 1992.

[92] Y. Saad. Iterative methods for sparse linear systems. Book out of print, 2000. Available at URL http:// www-users .cs .umn .edu/ ~saad /books.html.

[93] J. S. Savage and A. F. Peterson. Higher-order vecotr finite elements for tetrahedral cells. *IEEE Transections on Microwave Theory and Techniques*, 44(6), June 1996.

[94] N. Sczygiol and A. Nagórka. An *hp*-adaptive finite element method for the solidification problem based on explicit residual error. In *Proceedings of Computer Methods in Mechanics*, Wilsa, Poland, 3-6 June 2003.

[95] A. S. Shvedov and V. T. Zhukov. Explicit iterative difference schemes for parabolic equations. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 13(2):133–148, 1998.

[96] B. P. Sommeijer, L. F. Shampine, and J. G. Verwer. RKC: An explicit solver for parabolic PDEs. *J. Comput. Appl. Math.*, 88:315–326, 1997.

[97] D. Sun, J. Manges, X. Yuan, and Z. Cendes. Spurious modes in finite-element methods. *Antennas and Propagation Magazine, IEEE*, 37(5):12–24, 1995.

[98] S. Sun and M. F. Wheeler. Mesh adaptation strategies for discontinuous Galerkin methods applied to reactive transport problems. In H.-W. Chu, S. M, and B. Sanchez, editors, *Proceedings of the International Conference on Computing, Communication and Control Technologies*, pages 223–228, August 14-17, 2004.

[99] A. Taflove and S. C. Hagness. *Computational electrodynamics: the finite-difference time-domain method*. Artech House Inc., second edition, 2000.

[100] J. van den Eshof and M. Hochbruck. Preconditioning Lanczos approximations to the matrix exponential. Preprint of the Dept. of Mathematics, Heinrich Heine University, Duesseldorf, Germany, March, 2004, March 2004. To appear in SIAM J. Sci. Comput.

[101] A. J. van der Schaft and B. M. Maschke. Hamiltonian formulation of distributed-parameter systems with boundary energy flow. *Journal of Geometry and Physics*, 42:166–194, 2002.

[102] H. A. van der Vorst. An iterative solution method for solving $f(A)x = b$, using using Krylov subspace information obtained for the symmetric positive definite matrix A. *J. Comput. Appl. Math.*, 18:249–263, 1987.

[103] H. A. van der Vorst. *Iterative Krylov methods for large linear systems.* Cambridge University Press, 2003.

[104] R. Verfürth. *A review of a posteriori estimation and adaptive mesh-refinement techniques.* Advances in Numerical Mathematics. Wiley-Teubner, New York – Stuttgart, 1996.

[105] J. G. Verwer. Explicit Runge–Kutta methods for parabolic partial differential equations. *Appl. Num. Math.*, 22:359–379, 1996.

[106] P. Wesseling. *Principles of Computational Fluid Dynamics.* Springer, 2001.

[107] H. Whitney. *Geometric integration theory.* Princeton University Press, 1957.

[108] Wikipedia. Correlation — Wikipedia, the free encyclopedia, 2006. [Online; accessed 02-August-2006], http://en.wikipedia.org/wiki/Correlation.

[109] K. S. Yee. Numerical solution of initial boundary value problems involving Maxwells equations in isotropic media. *IEEE Trans. Antennas Propagat.*, 14(3):302–307, March 1966.

[110] A. Zdunek and W. Rachowicz. Application of the *hp* adaptive finite element method to the three-dimensional scattering of time-harmonic electromagnetic waves. In *Proceedings of the European Conference on Computational Mechanics*, Cracow, Poland, June 26-29 2001.

[111] O. C. Zienkiewicz and I. Morgan. *Finite elements and approximations.* John Wiley & Sons, 1983.

Acknowledgments

First of all I would like to mention my promoter Prof. Jaap van der Vegt who gave me this opportunity to pursue my Ph.D. at the University of Twente. His vast knowledge on the subject of numerical mathematics and our regular meetings helped me to overcome many fundamental problems that I faced during my research.

I would like to express my gratitude to my daily supervisor Dr. Mike Botchev. I am indebted to him for his daily support, stimulating advices, patience in correcting my articles. During these years, I have learned a lot from him. His constant encouragement and inspiration helped me to complete my Ph.D. thesis successfully.

The fourth and fifth chapters of this thesis is a result of collaborative work with my colleague and friend Dr. Ferenc Izsak. It was a great pleasure to work with him. Apart from work, we had a lot of fun outside university. I can never forget our social dinners and hot discussions on various subjects.

Let me take the opportunity to mention that NACM group is one of the best places that I have ever worked. I would like to thank everybody in the group for creating such a nice and friendly atmosphere. Especially, I express my sincere thanks to our secretary Marielle, who was ready to help me in every aspect of my life in Twente. I have to mention also my officemates Jaqueline, Henry, Pablo, Monika, Bert, Vijaya, Domokos, Lee, Alyona and Remco for making our office a charming place. I also thank my colleagues Kiran, Sena, Vita, Hadi, Natanael, Joris, Lars, Arek, Chris, Federik and Diana for being good colleagues and interesting companies.

I want to thank Vijaya Ambati for his support on C++ and for introducing me to PETSc package. Besides work, we had unforgettable times together. I will

always remember those nice evenings with a cup of vodka, which promoted us to many hot discussions on various topics. Here, I have to also thank Ram who always used to take part with us. Thank you both for sharing such a wonderful time with me. Many credits goes to Chris Klaij, who helped me to overcome various Fortran problems and provided a very nice thesis style. I deeply appreciate the contribution of Joris who translated the summary into Dutch: the Samenvatting. My work under Linux computer environment would never have been easy without the support of Enno Oosterhuis from helpdesk. Thank you Enno, you were always there to solve my problems as quickly as possible.

I would like to thank my committee members Prof. W. H. A. Schilders, Prof. K. Vuik, Dr. H. J. Zwart, Prof. B. Koren and Prof. S. A. van Gils for taking their valuable time to read my thesis. Your comments and corrections have further improved the thesis.

I have to mention the Armenian community in the Netherlands. Frequent contacts with them was very valuable for me and I felt as being in a small Armenia. In particular, I would like to specially thank Abrahamyans family from Nijmegen. I will never forget their hospitality and friendship through these years. My Armenian friends in Enschede, Karo and family of Norayr made my life in Enschede more entertaining and I will always remember the nice days with them. All those marvellous meetings with my friend Arayik Nalbandyan and his family are still fresh in my mind. My sincere gratitude to my friend and housemate Armen and his girlfriend Rima. It was a great pleasure to share the apartment for two years with such an intelligent friend and with whom I could share and exchange many interesting ideas.

During my master study in Nijmegen, I met Kamyar and Gantumur, and we soon became friends. I am very thankful to their constant support to me whenever I needed. All those indelible days during our master study and Ph.D. period are unforgettable.

I am also very much indebted to my family (mother, father and two brothers). It is their love and devotion which supported me to overcome many obstacles in my life.

Last but not least, I would like to thank my wife Astghik for her love, patience and support.

Davit Harutyunyan

Enschede, 2007

Summary

The increasing demand to understand the behaviour of electromagnetic waves in many real life problems requires solution of the Maxwell equations. In most cases the exact solution of the Maxwell equations is not available, hence numerical methods are indispensable tool to solve them numerically using different methods. In this thesis we consider the Maxwell equation in time and frequency domain, and for the space discretization we always apply Nédélec elements which are correct basis to describe the electromagnetic waves.

For time dependent problems many classical time integration schemes on unstructured meshes require very severe restriction on the time step, hence many time steps are required to solve the problem. To relax the time restrictions while preserving accuracy requirements a new Gautschi time integration scheme is applied to the Maxwell equations. The scheme involves a matrix-function evaluation, which is computed by Krylov subspace methods. We develop a very simple approach which adaptively chooses the dimension of the Krylov subspace. We prove that the scheme is unconditionally stable and allows to choose time steps larger than the smallest wave length.

In many problems the solution of the Maxwell equations contains structures with limited regularity, such as singularities near corners and non-convex edges. These complicated structures can be efficiently captured by adaptive methods, where the finite element mesh is locally coarsened or refined. For the time-harmonic Maxwell equations we develop an implicit a posteriori error estimation technique. The idea of the method is to solve for the error function an additional boundary value problem which provides a computable error. Then

this data can be easily used to identify the regions with larger error where refinement is required. The method is based on sound theoretical analysis and various complicated numerical experiments come to validate the technique.

The original Maxwell equations represent a coupled set of partial differential equations which can be formulated as a Hamiltonian system. The main characteristic of Hamiltonian dynamics is its energy conserving properties. In this thesis we show that discretization of the Maxwell equations for the electric and magnetic fields with the edge elements preserves correct energy transfer through the interfaces of neighboring elements. For time integrations the symplectic leap-frog scheme is applied which has a discrete energy conserving properties.

Samenvatting

De groeiende behoefte om het gedrag van elektromagnetische golven, die van belang zijn in veel practijk problemen, te begrijpen, vereist het oplossen van de Maxwell vergelijkingen. In de meeste gevallen is een exacte oplossing van de Maxwell vergelijkingen niet beschikbaar, daarom zijn numerieke oplostechnieken een onmisbaar gereedschap om discrete oplossingen te vinden. In deze thesis beschouwen we de Maxwell vergelijking in het tijds- en frequentiedomein. Voor de ruimtelijke discretisatie worden steeds Nédélec elementen toegepast. Deze vormen een correcte basis om electromagnetische golven te beschrijven.

Wanneer klassieke tijds integratie schemas op ongestructureerde roosters worden benut om tijdsafhankelijke problemen op te lossen, is er een strikte beperking op de maximale tijdstap. Hierdoor zijn er veel tijdstappen nodig om het probleem op te lossen. Om deze tijdstapbeperking te versoepelen en toch dezelfde nauwkeurigheid te behouden, is een nieuw Gautschi tijdsintegratieschema toegepast op de Maxwell vergelijkingen. Het schema vereist een matrixfunctie--evaluatie, om deze te berekenen worden Krylov subspace methodes toegepast. We ontwikkelen een eenvoudige aanpak om de dimensie van de Krylov subspace adaptief te bepalen. We tonen aan dat het schema onconditioneel stabiel is, en tijdstappen toelaat die groter zijn dan de kleinste golflengte.

In veel problemen bevat de oplossing van de Maxwell vergelijking structuren met beperkte regulariteit, zoals singulariteiten in de buurt van hoeken en niet-convexe randen. Deze ingewikkelde structuren kunnen efficient gevangen worden met behulp van adaptieve methodes. Hierbij wordt het eindige elementen rooster lokaal vergrofd of verfijnd. Voor de tijdsharmonische Maxwell vergeli-

jkingen ontwikkelen we een impliciete a posteriori foutafschattingsmethode. De gedachte achter de methode is om voor de errorfunctie een extra randwaardeprobleem op te lossen, hierdoor is de fout uit te rekenen. Met deze informatie kan eenvoudig worden bepaald in welke gebieden roosterverfijning nodig is. De methode is gebaseerd op betrouwbare theoretische analyse en de techniek is gevalideerd middels verscheidene gecompliceerde numerieke experimenten.

De originele Maxwell vergelijkingen omvatten een gekoppelde set partiele differentiaalvergelijkingen, die ook als Hamiltoniaans systeem kunnen worden geformuleerd. De voornaamste eigenschap van Hamiltoniaanse dynamica is de energie behoudendheid. In deze thesis laten we zien dat discretisatie van de Maxwell vergelijkingen, voor het elektrisch en magnetisch veld, met behulp van edge elementen, de juiste energie overdracht met naburige elementen behoud. Voor de tijdsintegratie wordt een symplectic leap-frog schema toegepast, dit geeft energie behoud op discreet niveau.